

*Springer Series in Statistics*

Ludwig Fahrmeir  
Gerhard Tutz

Multivariate  
Statistical  
Modelling Based  
on Generalized  
Linear Models

Second Edition



Springer

# Springer Series in Statistics

*Advisors:*

P. Bickel, P. Diggle, S. Fienberg, K. Krickeberg,  
I. Olkin, N. Wermuth, S. Zeger

# Springer Series in Statistics

---

- Andersen/Borgan/Gill/Keiding*: Statistical Models Based on Counting Processes.
- Atkinson/Riani*: Robust Diagnostic Regression Analysis.
- Berger*: Statistical Decision Theory and Bayesian Analysis, 2nd edition.
- Bolfarine/Zacks*: Prediction Theory for Finite Populations.
- Borg/Groenen*: Modern Multidimensional Scaling: Theory and Applications
- Brockwell/Davis*: Time Series: Theory and Methods, 2nd edition.
- Chen/Shao/Ibrahim*: Monte Carlo Methods in Bayesian Computation.
- David/Edwards*: Annotated Readings in the History of Statistics.
- Devroye/Lugosi*: Combinatorial Methods in Density Estimation.
- Efronovich*: Nonparametric Curve Estimation: Methods, Theory, and Applications.
- Fahrmeir/Tutz*: Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd edition.
- Farebrother*: Fitting Linear Relationships: A History of the Calculus of Observations 1750-1900.
- Federer*: Statistical Design and Analysis for Intercropping Experiments, Volume I: Two Crops.
- Federer*: Statistical Design and Analysis for Intercropping Experiments, Volume II: Three or More Crops.
- Fienberg/Hoaglin/Kruskal/Tanur (Eds.)*: A Statistical Model: Frederick Mosteller's Contributions to Statistics, Science and Public Policy.
- Fisher/Sen*: The Collected Works of Wassily Hoeffding.
- Glaz/Naus/Wallenstein*: Scan Statistics.
- Good*: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, 2nd edition.
- Gouriéroux*: ARCH Models and Financial Applications.
- Grandell*: Aspects of Risk Theory.
- Haberman*: Advanced Statistics, Volume I: Description of Populations.
- Hall*: The Bootstrap and Edgeworth Expansion.
- Härdle*: Smoothing Techniques: With Implementation in S.
- Hart*: Nonparametric Smoothing and Lack-of-Fit Tests.
- Hartigan*: Bayes Theory.
- Hedayat/Sloane/Stufken*: Orthogonal Arrays: Theory and Applications.
- Heyde*: Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation.
- Huet/Bouvier/Gruet/Jolivet*: Statistical Tools for Nonlinear Regression: A Practical Guide with S-PLUS Examples.
- Kolen/Brennan*: Test Equating: Methods and Practices.
- Kotz/Johnson (Eds.)*: Breakthroughs in Statistics Volume I.
- Kotz/Johnson (Eds.)*: Breakthroughs in Statistics Volume II.
- Kotz/Johnson (Eds.)*: Breakthroughs in Statistics Volume III.
- Küchler/Sørensen*: Exponential Families of Stochastic Processes.
- Le Cam*: Asymptotic Methods in Statistical Decision Theory.
- Le Cam/Yang*: Asymptotics in Statistics: Some Basic Concepts, 2nd edition.
- Longford*: Models for Uncertainty in Educational Testing.

(continued after index)

Ludwig Fahrmeir  
Gerhard Tutz

# Multivariate Statistical Modelling Based on Generalized Linear Models

Second Edition

With contributions from Wolfgang Hennevogl

With 51 Figures



Springer

Ludwig Fahrmeir  
Department of Statistics  
University of Munich  
Ludwigstrasse 33  
D-80539 München  
Germany

Gerhard Tutz  
Department of Statistics  
University of Munich  
Akademiestr. 1  
D-80799 München  
Germany

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

Multivariate statistical modelling based on generalized linear models / Ludwig Fahrmeir, Gerhard Tutz.—2nd ed.  
p. cm. — (Springer series in statistics)  
Includes bibliographical references and index.  
ISBN 978-1-4419-2900-6 ISBN 978-1-4757-3454-6 (eBook)  
DOI 10.1007/978-1-4757-3454-6  
1. Multivariate analysis. 2. Linear models (Statistics). I. Tutz, Gerhard.  
II. Title. III. Series.  
QA278.F34 2001  
519.5'38—dc21

00-052275

Printed on acid-free paper.

© 2001, 1994 Springer Science+Business Media New York  
Originally published by Springer-Verlag New York, Inc in 2001  
Softcover reprint of the hardcover 2nd edition 2001

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher Springer Science+Business Media, LLC,  
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.

Production managed by Allan Abrams; manufacturing supervised by Jerome Basma.  
Photocomposed copy prepared from the authors' LaTeX files.

9 8 7 6 5 4 3 2 1

ISBN 978-1-4419-2900-6

SPIN 10789428

# Preface to the Second Edition

Since our first edition of this book, many developments in statistical modelling based on generalized linear models have been published, and our primary aim is to bring the book up to date. Naturally, the choice of these recent developments reflects our own teaching and research interests.

The new organization parallels that of the first edition. We try to motivate and illustrate concepts with examples using real data, and most data sets are available on [http://www.stat.uni-muenchen.de/welcome\\_e.html](http://www.stat.uni-muenchen.de/welcome_e.html), with a link to *data archive*. We could not treat all recent developments in the main text, and in such cases we point to references at the end of each chapter.

Many changes will be found in several sections, especially with those connected to Bayesian concepts. For example, the treatment of marginal models in Chapter 3 is now current and state-of-the-art. The coverage of nonparametric and semiparametric generalized regression in Chapter 5 is completely rewritten with a shift of emphasis to linear bases, as well as new sections on local smoothing approaches and Bayesian inference. Chapter 6 now incorporates developments in parametric modelling of both time series and longitudinal data. Additionally, random effect models in Chapter 7 now cover nonparametric maximum likelihood and a new section on fully Bayesian approaches. The modifications and extensions in Chapter 8 reflect the rapid development in state space and hidden Markov models. Monte Carlo techniques, in particular MCMC, are described in greater detail in the text and in a new appendix. New sections in Chapter 8 extend the main ideas from state space models with time series or longitudinal data into models that can accommodate spatial and spatio-temporal data.

We are grateful to Brian Marx for his many helpful suggestions and comments. Tom Santner is also gratefully acknowledged for his careful reading and corrections. Our thanks are further extended to Michael Schindler for the typesetting and to Jochen Einbeck, Ludwig Heigenhauser, Andrea Hennerfeind, and Stefan Lang for computational assistance.

The support of the Deutsche Forschungsgemeinschaft in the form of the *Sonderforschungsbereich Analysis of Discrete Structures* has helped this second edition into its final form.

München, Germany  
November 23, 2000

Ludwig Fahrmeir  
Gerhard Tutz

# Preface to the First Edition

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 GLMs 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 GLMs: 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 GLMs 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, loglinear models for contingency tables, which can be treated as special cases of GLMs, are totally omitted. The estimation approach 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 A.3, A.4, and A.5, 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 skillful 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, Germany  
February 2, 1994

Ludwig Fahrmeir  
Gerhard Tutz  
Wolfgang Hennevogl

Since the first printing of this book, several very readable books and other important works have been published. We have included some remarks at the end of each chapter.

München and Berlin, Germany  
May 28, 1997

Ludwig Fahrmeir  
Gerhard Tutz

# Contents

Preface to the Second Edition .....	v
Preface to the First Edition .....	vii
List of Examples .....	xvii
List of Figures .....	xxi
List of Tables .....	xxv
<b>1. Introduction .....</b>	<b>1</b>
1.1 Outline and Examples .....	2
1.2 Remarks on Notation .....	13
1.3 Notes and Further Reading .....	14
<b>2. Modelling and Analysis of Cross-Sectional Data: A Review of Univariate Generalized Linear Models .....</b>	<b>15</b>
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 .....	24
2.1.4 Models for Binary and Binomial Responses .....	24
Linear Probability Model .....	25
Probit Model .....	26
Logit Model .....	26
Complementary Log-Log Model .....	26
Complementary Log-Model .....	26
Binary Models as Threshold Models of Latent Linear Models .....	29
Parameter Interpretation .....	29
Overdispersion .....	35
2.1.5 Models for Count Data .....	36
Log-linear Poisson Model .....	36

	Linear Poisson Model . . . . .	36
2.2	Likelihood Inference . . . . .	38
2.2.1	Maximum Likelihood Estimation . . . . .	38
	Log-likelihood, Score Function and Information Matrix	39
	Numerical Computation of the MLE by	
	Iterative Methods . . . . .	41
	Uniqueness and Existence of MLEs* . . . . .	43
	Asymptotic Properties . . . . .	44
	Discussion of Regularity Assumptions*	46
	Additional Scale or Overdispersion Parameter . . . . .	47
2.2.2	Hypothesis Testing and Goodness-of-Fit Statistics . . . . .	47
	Goodness-of-Fit Statistics . . . . .	50
2.3	Some Extensions . . . . .	55
2.3.1	Quasi-likelihood Models . . . . .	55
	Basic Models . . . . .	55
	Variance Functions with Unknown Parameters . . . . .	58
	Nonconstant Dispersion Parameter . . . . .	59
2.3.2	Bayesian Models . . . . .	60
2.3.3	Nonlinear and Nonexponential Family Regression Models* . . . . .	65
2.4	Notes and Further Reading . . . . .	67
3.	<b>Models for Multicategorical Responses: Multivariate Extensions of Generalized Linear Models</b> . . . . .	69
3.1	Multicategorical Response Models . . . . .	70
3.1.1	Multinomial Distribution . . . . .	70
3.1.2	Data . . . . .	71
3.1.3	The Multivariate Model . . . . .	72
3.1.4	Multivariate Generalized Linear Models . . . . .	75
3.2	Models for Nominal Responses . . . . .	77
3.2.1	The Principle of Maximum Random Utility . . . . .	77
3.2.2	Modelling of Explanatory Variables: Choice of Design Matrix . . . . .	79
3.3	Models for Ordinal Responses . . . . .	81
3.3.1	Cumulative Models: The Threshold Approach . . . . .	83
	Cumulative Logistic Model or Proportional Odds Model . . . . .	83
	Grouped Cox Model or Proportional Hazards Model .	86
	Extreme Maximal-value Distribution Model . . . . .	86
3.3.2	Extended Versions of Cumulative Models . . . . .	87
3.3.3	Link Functions and Design Matrices for Cumulative Models . . . . .	88
3.3.4	Sequential Models . . . . .	92
	Generalized Sequential Models . . . . .	95
	Link Functions of Sequential Models . . . . .	98

3.3.5	Strict Stochastic Ordering*	99
3.3.6	Two-Step Models	100
	Link Function and Design Matrix for Two-Step Models	102
3.3.7	Alternative Approaches	103
3.4	Statistical Inference	105
3.4.1	Maximum Likelihood Estimation	105
	Numerical Computation	107
3.4.2	Testing and Goodness-of-Fit	107
	Testing of Linear Hypotheses	107
	Goodness-of-Fit Statistics	107
3.4.3	Power-Divergence Family*	109
	Asymptotic Properties under Classical “Fixed Cells” Assumptions	111
	Sparseness and “Increasing-Cells” Asymptotics	112
3.5	Multivariate Models for Correlated Responses	112
3.5.1	Conditional Models	114
	Asymmetric Models	114
	Symmetric Models	116
3.5.2	Marginal Models	119
	Marginal Models for Correlated Univariate Responses	120
	The Generalized Estimating Approach for Statistical Inference	123
	Marginal Models for Correlated Categorical Responses	129
	Likelihood-based Inference for Marginal Models	135
3.6	Notes and Further Reading	136
	Bayesian Inference	136
4.	Selecting and Checking Models	139
4.1	Variable Selection	139
4.1.1	Selection Criteria	140
4.1.2	Selection Procedures	142
	All-Subsets Selection	142
	Stepwise Backward and Forward Selection	143
4.2	Diagnostics	145
4.2.1	Diagnostic Tools for the Classical Linear Model	146
4.2.2	Generalized Hat Matrix	147
4.2.3	Residuals and Goodness-of-Fit Statistics	151
4.2.4	Case Deletion	156
4.3	General Tests for Misspecification*	161
4.3.1	Estimation under Model Misspecification	162
4.3.2	Hausman-type Tests	165
	Hausman Tests	165
	Information Matrix Test	166

4.3.3	Tests for Nonnested Hypotheses . . . . .	167
	Tests Based on Artificial Nesting . . . . .	168
	Generalized Wald and Score Tests . . . . .	168
4.4	Notes and Further Reading . . . . .	170
	Bayesian Model Determination . . . . .	170
	Robust Estimates . . . . .	172
	Model Tests Against Smooth Alternatives . . . . .	172
5.	<b>Semi- and Nonparametric Approaches to Regression Analysis</b> . . . . .	173
5.1	Smoothing Techniques for Continuous Responses . . . . .	174
5.1.1	Regression Splines and Other Basis Functions . . . . .	174
	Regression Splines . . . . .	176
	Other Basis Functions . . . . .	178
	Regularization . . . . .	179
5.1.2	Smoothing Splines . . . . .	181
5.1.3	Local Estimators . . . . .	183
	Simple Neighborhood Smoothers . . . . .	183
	Local Regression . . . . .	184
	Bias-Variance Trade-off . . . . .	187
	Relation to Other Smoothers . . . . .	189
5.1.4	Selection of Smoothing Parameters . . . . .	190
5.2	Smoothing for Non-Gaussian Data . . . . .	193
5.2.1	Basis Function Approach . . . . .	193
	Fisher Scoring for Penalized Likelihood* . . . . .	194
5.2.2	Penalization and Spline Smoothing . . . . .	195
	Fisher Scoring for Generalized Spline Smoothing* . . . . .	196
	Choice of Smoothing Parameter . . . . .	197
5.2.3	Localizing Generalized Linear Models . . . . .	198
	Local Fitting by Weighted Scoring . . . . .	201
5.3	Modelling with Multiple Covariates . . . . .	202
5.3.1	Modelling Approaches . . . . .	207
	Generalized Additive Models . . . . .	207
	Partially Linear Models . . . . .	208
	Varying-Coefficient Models . . . . .	208
	Projection Pursuit Regression . . . . .	209
	Basis Function Approach . . . . .	210
5.3.2	Estimation Concepts . . . . .	213
	Backfitting Algorithm for Generalized Additive Models . . . . .	213
	Backfitting with Spline Functions . . . . .	217
	Choice of Smoothing Parameter . . . . .	220
	Partial Linear Models . . . . .	220
5.4	Semiparametric Bayesian Inference for Generalized Regression . . . . .	221

5.4.1	Gaussian Responses .....	221
	Smoothness Priors Approaches .....	221
	Basis Function Approaches .....	227
	Models with Multiple Covariates .....	228
5.4.2	Non-Gaussian Responses .....	231
	Latent Variable Models for Categorical Responses .....	234
5.5	Notes and Further Reading .....	239
<b>6.</b>	<b>Fixed Parameter Models for Time Series and Longitudinal Data .....</b>	<b>241</b>
6.1	Time Series .....	242
6.1.1	Conditional Models .....	242
	Generalized Autoregressive Models .....	242
	Quasi-Likelihood Models and Generalized Autoregression Moving Average Models .....	246
6.1.2	Statistical Inference for Conditional Models .....	249
6.1.3	Marginal Models .....	255
	Estimation of Marginal Models .....	258
6.2	Longitudinal Data .....	260
6.2.1	Conditional Models .....	261
	Generalized Autoregressive Models, Quasi-Likelihood Models .....	261
	Statistical Inference .....	262
	Transition Models .....	264
	Subject-specific Approaches and Conditional Likelihood .....	264
6.2.2	Marginal Models .....	267
	Statistical Inference .....	268
6.2.3	Generalized Additive Models for Longitudinal Data .....	274
6.3	Notes and Further Reading .....	278
<b>7.</b>	<b>Random Effects Models .....</b>	<b>283</b>
7.1	Linear Random Effects Models for Normal Data .....	285
7.1.1	Two-stage Random Effects Models .....	285
	Random Intercepts .....	286
	Random Slopes .....	287
	Multilevel Models .....	288
7.1.2	Statistical Inference .....	289
	Known Variance-Covariance Components .....	289
	Unknown Variance-Covariance Components .....	289
	Derivation of the EM algorithm* .....	291
7.2	Random Effects in Generalized Linear Models .....	292
	Generalized Linear Models with Random Effects .....	293
	Examples .....	294
7.3	Estimation Based on Posterior Modes .....	298

7.3.1	Known Variance-Covariance Components . . . . .	298
7.3.2	Unknown Variance-Covariance Components . . . . .	299
7.3.3	Algorithmic Details* . . . . .	300
	Fisher Scoring for Given Variance-Covariance Components . . . . .	300
	EM Type Algorithm . . . . .	302
7.4	Estimation by Integration Techniques . . . . .	303
7.4.1	Maximum Likelihood Estimation of Fixed Parameters	303
	Direct Maximization Using Fitting Techniques for GLMs . . . . .	305
	Nonparametric Maximum Likelihood for Finite Mix- tures . . . . .	308
7.4.2	Posterior Mean Estimation of Random Effects . . .	310
7.4.3	Indirect Maximization Based on the EM Algorithm* .	311
7.4.4	Algorithmic Details for Posterior Mean Estimation* .	315
7.5	Examples . . . . .	318
7.6	Bayesian Mixed Models . . . . .	321
	Bayesian Generalized Mixed Models . . . . .	321
	Generalized Additive Mixed Models . . . . .	322
7.7	Marginal Estimation Approach to Random Effects Models .	325
7.8	Notes and Further Reading . . . . .	328
8.	<b>State Space and Hidden Markov Models . . . . .</b>	331
8.1	Linear State Space Models and the Kalman Filter . . . .	332
8.1.1	Linear State Space Models . . . . .	332
8.1.2	Statistical Inference . . . . .	337
	Linear Kalman Filtering and Smoothing . . . . .	338
	Kalman Filtering and Smoothing as Posterior Mode Estimation* . . . . .	340
	Unknown Hyperparameters . . . . .	342
	EM Algorithm for Estimating Hyperparameters* .	343
8.2	Non-Normal and Nonlinear State Space Models . . . .	345
8.2.1	Dynamic Generalized Linear Models . . . . .	345
	Categorical Time Series . . . . .	347
8.2.2	Nonlinear and Nonexponential Family Models* . . . .	349
8.3	Non-Normal Filtering and Smoothing . . . . .	350
8.3.1	Posterior Mode Estimation . . . . .	351
	Generalized Extended Kalman Filter and Smoother* .	352
	Gauss-Newton and Fisher-Scoring Filtering and Smoothing* . . . . .	354
	Estimation of Hyperparameters* . . . . .	356
	Some Applications . . . . .	356
8.3.2	Markov Chain Monte Carlo and Integration-based Approaches . . . . .	361
	MCMC Inference . . . . .	362

	Integration-based Approaches . . . . .	365
8.4	Longitudinal Data . . . . .	369
8.4.1	State Space Modelling of Longitudinal Data . . . . .	369
8.4.2	Inference For Dynamic Generalized Linear Mixed Models . . . . .	372
8.5	Spatial and Spatio-temporal Data . . . . .	376
8.6	Notes and Further Reading . . . . .	383
<b>9.</b>	<b>Survival Models . . . . .</b>	<b>385</b>
9.1	Models for Continuous Time . . . . .	385
9.1.1	Basic Models . . . . .	385
	Exponential Distribution . . . . .	386
	Weibull Distribution . . . . .	387
	Piecewise Exponential Model . . . . .	388
9.1.2	Parametric Regression Models . . . . .	388
	Location-Scale Models for $\log T$ . . . . .	388
	Proportional Hazards Models . . . . .	389
	Linear Transformation Models and Binary Regression Models . . . . .	390
9.1.3	Censoring . . . . .	391
	Random Censoring . . . . .	391
	Type I Censoring . . . . .	392
9.1.4	Estimation . . . . .	393
	Exponential Model . . . . .	394
	Weibull Model . . . . .	394
	Piecewise Exponential Model . . . . .	395
9.2	Models for Discrete Time . . . . .	396
9.2.1	Life Table Estimates . . . . .	397
9.2.2	Parametric Regression Models . . . . .	400
	The Grouped Proportional Hazards Model . . . . .	400
	A Generalized Version: The Model of Aranda-Ordaz .	402
	The Logistic Model . . . . .	403
	Sequential Model and Parameterization of the Baseline Hazard . . . . .	403
9.2.3	Maximum Likelihood Estimation . . . . .	404
9.2.4	Time-varying Covariates . . . . .	408
	Internal Covariates* . . . . .	411
	Maximum Likelihood Estimation* . . . . .	412
9.3	Discrete Models for Multiple Modes of Failure . . . . .	414
9.3.1	Basic Models . . . . .	414
9.3.2	Maximum Likelihood Estimation . . . . .	417
9.4	Smoothing in Discrete Survival Analysis . . . . .	420
9.4.1	Smoothing Life Table Estimates . . . . .	420
9.4.2	Smoothing with Covariates . . . . .	422
9.4.3	Dynamic Discrete-Time Survival Models . . . . .	423

Posterior Mode Smoothing .....	423
Fully Bayesian Inference via MCMC .....	425
9.5 Remarks and Further Reading .....	429
<b>A.</b> .....	433
A.1 Exponential Families and Generalized Linear Models .....	433
A.2 Basic Ideas for Asymptotics .....	437
A.3 EM Algorithm .....	442
A.4 Numerical Integration .....	443
A.5 Monte Carlo Methods .....	449
<b>B. Software for Fitting Generalized Linear Models and Extensions</b> .....	455
Bibliography .....	467
Author Index .....	505
Subject Index .....	512

# List of Examples

1.1	Caesarian birth study .....	2
1.2	Credit-scoring .....	2
1.3	Cellular differentiation .....	3
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 .....	10
1.10	Ohio children .....	11
1.11	Duration of unemployment .....	13
2.1	Caesarian birth study .....	30
2.2	Credit-scoring .....	31
2.3	Cellular differentiation .....	37
2.4	Caesarian birth study (Example 2.1, continued) .....	51
2.5	Credit-scoring (Example 2.2, continued) .....	53
2.6	Cellular differentiation (Example 2.3, continued) .....	53
2.7	Cellular differentiation (Examples 2.3, 2.6, continued) .....	59
3.1	Caesarian birth study .....	73
3.2	Breathing test results .....	81
3.3	Job expectation .....	81
3.4	Breathing test results (Example 3.2, continued) .....	89
3.5	Job expectation (Example 3.3, continued) .....	91
3.6	Tonsil size .....	93
3.7	Tonsil size (Example 3.6, continued) .....	96
3.8	Breathing test results (Examples 3.2 and 3.4, continued) .....	96
3.9	Rheumatoid arthritis .....	100
3.10	Rheumatoid arthritis (Example 3.9, continued) .....	101
3.11	Caesarian birth study (Example 3.1, continued) .....	108
3.12	Reported happiness .....	115
3.13	Visual impairment .....	127

3.14	Forest damage .....	131
4.1	Credit-scoring (Examples 2.2, 2.5, continued) .....	144
4.2	Vaso constriction .....	148
4.3	Job expectation (Examples 3.3, 3.5, continued) .....	149
4.4	Vaso constriction (Example 4.2, continued) .....	154
4.5	Job expectation (Examples 3.3, 3.5, 4.3, continued) .....	156
4.6	Vaso constriction (Examples 4.2, 4.4, continued) .....	160
4.7	Job expectation (Examples 3.3, 3.5, 4.3, 4.5, continued) .....	161
4.8	Credit-scoring (Examples 2.2, 2.5, 4.1, continued) .....	170
5.1	Motorcycle data .....	174
5.2	Rainfall data .....	197
5.3	Short-term unemployment .....	202
5.4	Rainfall data (Example 5.2, continued) .....	202
5.5	Vaso constriction data (Examples 4.2, 4.4, continued) .....	204
5.6	Chronic bronchitis .....	210
5.7	Women's role in society .....	211
5.8	Vaso constriction data (Example 5.5, continued) .....	220
5.9	Rental rates .....	230
5.10	Credit scoring revisited .....	236
6.1	Polio incidence in the United States .....	252
6.2	Polio incidence in the United States (Example 6.1, continued) .....	259
6.3	IFO business test .....	265
6.4	Ohio children .....	271
6.5	A longitudinal study on forest damage .....	275
7.1	Ohio children data (Example 6.4, continued) .....	292
7.2	Bitterness of white wines .....	292
7.3	Ohio children data (Example 7.1, continued) .....	318
7.4	Bitterness of white wines (Example 7.2, continued) .....	319
7.5	Longitudinal study on forest damage (Example 6.5, continued) .....	323
8.1	Rainfall data (Example 5.2, continued) .....	356
8.2	Advertising data .....	358
8.3	Phone calls .....	360
8.4	Rainfall data (Example 8.1, continued) .....	367
8.5	Weekly incidence of AHC .....	368
8.6	Business test (Example 6.3, continued) .....	373
8.7	Rents for flats (Example 5.9, continued) .....	381

9.1	Duration of unemployment .....	399
9.2	Duration of unemployment (Example 9.1, continued) .....	407
9.3	Duration of unemployment (Example 9.2, continued) .....	418
9.4	Duration of unemployment: A spatio-temporal analysis .....	425
9.5	Head and neck cancer .....	428

# List of Figures

1.1	Number of occurrences of rainfall in the Tokyo area for each calendar day 1983-1984. ....	9
1.2	Monthly number of polio cases in the United States from 1970 to 1983. ....	10
2.1	The gamma distribution: $G(\mu = 1, \nu)$ . ....	24
2.2	Response functions for binary responses. ....	27
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. ....	49
3.1	Densities of the latent response for two subpopulations with different values of $x$ (logistic, extreme minimal-value, extreme maximal-value distributions). ....	85
3.2	Frequency distribution of damage classes. ....	132
4.1	Index plot of $h_{ii}$ for vaso constriction data. ....	150
4.2	Index plot of $\text{tr}(H_{ii})$ and $(H_{ii})$ for grouped job expectation data. ....	151
4.3	Index plot of $r_i^P, r_{i,s}^P$ , and $r_i^D$ for vaso constriction data. ....	155
4.4	$N(0, 1)$ -probability plot of $r_{i,s}^P$ for vaso constriction data. ....	155
4.5	Relative frequencies and response curve of the fitted cumulative logistic mode. ....	157
4.6	Index plot of $(r_{i,s}^P)'(r_{i,s}^P)$ for grouped job expectation data. ....	158
4.7	$\chi^2(2)$ -probability plot of $(r_{i,s}^P)'(r_{i,s}^P)$ for grouped job expectation data. ....	159
4.8	Index plot of $c_{i,1}$ and $c_i$ for vaso constriction data. ....	160
4.9	Index plot of $c_{i,1}$ and $c_i$ for grouped job expectation data. ....	162
5.1	Smoothed estimates for motorcycle data showing time and head acceleration after a simulated impact. ....	175

5.2 Illustration of B-splines bases, one isolated and several overlapping ones for degree 1 and degree 2. ....	177
5.3 Smoothed probability of rainfall $\lambda = 4064$ . ....	199
5.4 Smoothed probability of rainfall $\lambda = 32$ . ....	199
5.5 Generalized cross-validation criterion, with logarithmic scale for $\lambda$ . ....	200
5.6 Probability of short-term unemployment estimated by the parametric logit model, P-splines, smoothing splines, and local likelihood with the smooth estimates determined by the AIC criterion. ....	203
5.7 Number of observations of unemployment data for given age groups. ....	203
5.8 Probability of short term unemployment with same degree of smoothing for smooth estimates. ....	204
5.9 Local fit of logit model for Tokyo rainfall data. ....	205
5.10 Cross-validation for local fit for Tokyo rainfall data. ....	205
5.11 Response surface for the nonoccurrence of vaso constriction based on the local fit of a logit model. ....	206
5.12 Response surface for the nonoccurrence of vaso constriction based on a fitted logit model. ....	206
5.13 Effects of concentration of dust and exposure time on the probability of chronic bronchitis for the generalized additive logit model. ....	211
5.14 Local linear fit of $f(dust, years)$ with nearest neighborhood to probability of chronic bronchitis. ....	212
5.15 Local likelihood estimates of intercept and slope for gender varying across years of education compared with estimates for the linear model within each gender group. ....	214
5.16 Fitted probabilities for women and men based on local likelihood and parametric model with pointwise error bands $1.96 * \text{standard error}$ . ....	215
5.17 Nonoccurrence of vaso constriction of the skin smoothed by an additive model with $\lambda_1 = \lambda_2 = 0.001$ . ....	222
5.18 Nonoccurrence of vaso constriction of the skin smoothed by an additive model with $\lambda_1 = \lambda_2 = 0.003$ . ....	222
5.19 Estimated effects of floor space and year of construction for the rent data. Shown are the posterior means within 80% credible regions. ....	232
5.20 Estimated effects of duration and amount of credit. ....	238
6.1 Monthly number of polio cases in the United States from 1970 to 1983. ....	254
6.2 Predicted polio incidence $\hat{\mu}_t$ based on a log-linear AR( $l = 5$ )-model fit. ....	255
6.3 Damage class distribution by time. ....	277

6.4	Estimated global odds ratios (lines) and empirically observed global odds ratios (points). Note that there is a different scale for the combination $l = 2, r = 1$ . . . . .	279
6.5	Estimated thresholds $\hat{f}_1(t)$ (left plot) and $\hat{f}_2(t)$ (right plot) with pointwise standard error bands (model based – dashed line, robust – boundary of shaded region). . . . .	280
6.6	Estimated effects of age and pH value with pointwise standard error bands . . . . .	280
6.7	Estimated probabilities $\text{pr}(Y_t = 1)$ , $\text{pr}(Y_t = 2)$ , and $\text{pr}(Y_t = 3)$ for age. From top to bottom: up to 50 years; between 50 and 120 years; above 120 years. . . . .	281
7.1	Relative frequency of the response “damage state” over the years. . . . .	324
7.2	Estimated nonparametric functions for the forest damage data. Shown are the posterior means with 80% credible regions. . . . .	324
8.1	Number of occurrences of rainfall in the Tokyo area for each calendar day during 1983–1984. . . . .	357
8.2	Smoothed probabilities $\hat{\pi}_t$ of daily rainfall, obtained by generalized Kalman and Gauss–Newton smoothing. . . . .	357
8.3	Smoothed trend and advertising effect. . . . .	359
8.4	Advertising data and fitted values. . . . .	359
8.5	Number of phone calls for half-hour intervals. . . . .	360
8.6	Observed and fitted values. . . . .	362
8.7	Tokyo rainfall data. . . . .	368
8.8	AHC data. Data and fitted probabilities (posterior median within 50%, 80%, and 95% credible regions). . . . .	370
8.9	Estimates of trends and thresholds. . . . .	376
8.10	Plot of the estimates of $b_{i1}$ against $b_{i2}$ for each unit. . . . .	377
8.11	Estimated time-changing covariate effects. Dashed vertical lines represent the month January of each year. . . . .	377
8.12	Estimated time-changing covariate effect of $G^+$ . Dashed vertical lines represent the month January of each year. . . . .	378
8.13	Posterior means of spatial effects. . . . .	381
9.1	Life table estimate for unemployment data. . . . .	400
9.2	Estimated survival function for unemployment data. . . . .	401
9.3	Life table estimate for duration of unemployment with causes full-time job (category 1) or part-time job (category 2). . . . .	420
9.4	Estimated nonparametric functions and seasonal effect. Shown are the posterior means within 80% credible regions. . . . .	426

xxiv List of Figures

9.5	Posterior mean and posterior probabilities of the district-specific effect. . . . .	427
9.6	Cubic-linear spline fit for head and neck cancer data and posterior mode smoother with $\pm$ standard deviation confidence bands. . . . .	431
9.7	Smoothed kernel estimate for head and neck cancer data. . . . .	431

# List of Tables

1.1	Data on infection from 251 births by Caesarian section . . . . .	3
1.2	Cellular differentiation data . . . . .	4
1.3	Grouped data for job expectations of psychology students in Regensburg . . . . .	5
1.4	Breathing results of Houston industrial workers . . . . .	6
1.5	Visual impairment data (reproduced from Liang, Zeger & Qaqish, 1992) . . . . .	8
1.6	Presence and absence of respiratory infection . . . . .	12
2.1	Simple exponential families with dispersion parameter . . . . .	21
2.2	Grouped data on infection . . . . .	31
2.3	Cellular differentiation data . . . . .	37
2.4	Logit model fit to Caesarian birth study data . . . . .	51
2.5	Logit model fit to credit-scoring data . . . . .	54
2.6	Log-linear model fits to cellular differentiation data based on Poisson-likelihoods . . . . .	54
2.7	Log-linear model fits to cellular differentiation data based on quasi-likelihoods . . . . .	60
3.1	Breathing results of Houston industrial workers . . . . .	81
3.2	Grouped data for job expectations of psychology students in Regensburg . . . . .	82
3.3	Estimates of cumulative models for breathing test data ( <i>p</i> -values in each second column) . . . . .	90
3.4	Cumulative model for job expectation data ( <i>p</i> -values are given in brackets) . . . . .	92
3.5	Tonsil size and Streptococcus pyogenes (Holmes & Williams, 1954) .	93
3.6	Fits for tonsil size data . . . . .	97
3.7	Sequential logit models for the breathing test data ( <i>p</i> -values in brackets) . . . . .	98
3.8	Clinical trial of a new agent and an active control . . . . .	100
3.9	Analysis of clinical trial data on rheumatoid arthritis . . . . .	102

3.10	Cross classification of gender, reported happiness, and years of schooling .....	115
3.11	Estimates for the cross classification of gender, reported happiness and years of schooling .....	117
3.12	Visual impairment data (from Liang et al., 1992) .....	127
3.13	Estimation results for visual impairment data .....	128
3.14	Parameter estimates and standard errors .....	133
4.1	Logit model fit to credit-scoring data .....	144
4.2	Vaso constriction data .....	149
4.3	Logit model fit to vaso constriction data.....	150
5.1	Estimates of selected covariate effects .....	231
5.2	Estimates of constant parameters for the credit-scoring data .....	237
6.1	Monthly number of poliomyelitis cases in the United States for 1970 to 1983 .....	253
6.2	Log-linear AR(5) model fit to polio data.....	254
6.3	Log-linear AR(1) model fits to polio data .....	256
6.4	Marginal model fit for polio data .....	260
6.5	Variables and questions of the IFO business test .....	266
6.6	Estimates of main effects .....	267
6.7	Presence and absence of respiratory infection.....	271
6.8	Marginal logit model fits for Ohio children data .....	272
6.9	Main effects model fits for Ohio children data .....	273
6.10	Effects of canopy density .....	278
7.1	Bitterness of wine data (Randall, 1989) .....	293
7.2	Random intercept logit model for Ohio children data (effect coding of smoking and age, standard deviations in parentheses) .....	318
7.3	ML estimates for Ohio children data based on finite mixtures .....	319
7.4	Estimation results of bitterness-of-wine data .....	320
7.5	Estimates of constant parameters for the forest damage data. ....	325
9.1	Duration of unemployment .....	409
9.2	Estimates of cause-specific logistic model for duration of unemployment data.....	419
9.3	Estimates of constant parameters in the unemployment data .....	429
9.4	Head and neck cancer (Efron, 1988).....	430

# 1

## Introduction

Classical statistical models for regression, time series, and longitudinal data analysis are generally useful in situations where data are approximately Gaussian and can be explained by some linear structure. These models are easy to interpret and the methods are theoretically well understood and investigated. However, the underlying assumptions may be too stringent and applications of the methods may be misleading in situations where data are clearly non-normal, such as categorical or counted data. Statistical modelling aims at providing more flexible model-based tools for data analysis.

Generalized linear models have been introduced by Nelder & Wedderburn (1972) as a unifying family of models for nonstandard cross-sectional regression analysis with non-normal responses. Their further development had a major influence on statistical modelling in a wider sense, and they have been extended in various ways and for more general situations. This book brings together and reviews a large part of recent advances in statistical modelling that are based on or related to generalized linear models. Throughout the text the focus is on discrete data. The topics include models for ordered multicategorical responses, multivariate correlated responses in cross-sectional or repeated measurements situations, nonparametric approaches, autoregressive-type and random effects extensions for non-normal time series and longitudinal data, state space and hidden Markov models for longitudinal and spatially correlated data, and discrete time survival models.

To make it accessible for a broader readership, the book is written at an intermediate mathematical level. The emphasis is on basic ideas and on explanations of methods from the viewpoint of an applied statistician. It is helpful to be familiar with linear regression models. Although some concepts like that of design matrices and coding of qualitative variables are introduced, they are not described at length. Of course, knowledge of matrix calculus and basic theory of estimation and testing is a necessary prerequisite. Extension of linear regression for continuous responses to nonparametric approaches, random coefficient models, and state space models

are reviewed in compact form at the beginning of corresponding chapters of this book. For a more detailed and deeper study of methods for continuous data, the reader should consult the literature cited there. Mathematical and technical details of the theory are kept at a comparably simple level. For detailed and rigorous proofs of some results, for example, asymptotics, the reader is referred to the literature. Moreover, some extensions and a number of mathematical and algorithmic details are deferred to starred sections. Knowledge of these sections is not necessary to understand the main body of this text. Real data examples and applications from various fields such as economics, social science, and medicine illustrate the text. They should encourage statisticians, students, and researchers working in these fields to use the methods for their own problems, and they should stimulate mathematically interested readers for a deeper study of theoretical foundations and properties. Many of the data sets can be downloaded from our data archive: [http://www.stat.uni-muenchen.de/welcome\\_e.html](http://www.stat.uni-muenchen.de/welcome_e.html).

## 1.1 Outline and Examples

The following survey of contents is illustrated by some of the examples that will be used in later chapters.

### **Modelling and Analysis of Univariate Cross-Sectional Data (Chapter 2)**

Chapter 2 gives a review of univariate generalized linear models and some basic extensions such as quasi-likelihood models. These models are useful for cross-sectional parametric regression analysis with non-normal response variables as in Examples 1.1, 1.2, and 1.3, where responses are binary or counts.

#### **Example 1.1: Caesarian birth study**

Table 1.1, kindly provided by Prof. R. E. Weissenbacher, Munich, Klinikum Großhadern, shows data on infection from births by Caesarian section. The response variable of interest is the occurrence or nonoccurrence of infection, with two types (I,II) of infection. Three dichotomous covariates, which may influence the risk of infection, were considered: Was the Caesarian section planned or not? Were risk-factors such as diabetes, excessive weight, and others present? Were antibiotics given as a prophylaxis? The aim is to analyze effects of the covariates on the risk of infection, especially whether antibiotics can decrease the risk of infection. If one ignores the two types of infection, the response variable is binary (infection yes/no); otherwise it is three-categorical. □

**Table 1.1.** Data on infection from 251 births by Caesarian section

	Caesarian planned			Not planned		
	Infection			Infection		
	I	II	non	I	II	non
<b>Antibiotics</b>						
Risk-factors	0	1	17	4	7	87
No risk-factors	0	0	2	0	0	0
<b>No antibiotics</b>						
Risk-factors	11	17	30	10	13	3
No risk-factors	4	4	32	0	0	9

**Example 1.2: Credit-scoring**

In the credit business, banks are interested in information about whether or not prospective consumers will pay back their credit as agreed to. The aim of credit-scoring is to model or predict the probability that a consumer with certain covariates (“risk factors”) is to be considered as a potential risk. The data set, which will be analyzed later, consists of 1000 consumers’ credits from a German bank. For each consumer the binary variable  $y$  “creditability” ( $y = 0$  for credit-worthy,  $y = 1$  for not credit-worthy) is available. In addition, 20 covariates that are assumed to influence creditability were collected, for example,

- running account, with categories no, medium, and good,
- duration of credit in months,
- amount of credit in “Deutsche Mark,”
- payment of previous credits, with categories good and bad,
- intended use, with categories private and professional.

The original data set is reproduced in Fahrmeir & Hamerle (1984, Appendix C) and can be downloaded from our data archive: [http://www.stat.uni-muenchen.de/welcome\\_e.html](http://www.stat.uni-muenchen.de/welcome_e.html) with a link to *data archive*. (Some information on the distribution of covariates is also given in Example 2.2).

We will analyze the effect of covariates on the variable “creditability” by a binary regression model. Other tools used for credit-scoring are discriminance analysis, classification trees, and neural networks. □

**Example 1.3: Cellular differentiation**

The effect of two agents of immuno-activating ability that may induce cell differentiation was investigated by Piegorsch, Weinberg & Margolin (1988). As response variable the number of cells that exhibited markers after exposure was recorded. It is of interest if the agents TNF (tumor necrosis factor) and IFN (interferon) stimulate cell differentiation independently or if there is a synergistic effect. The count data are given in Table 1.2.  $\square$

Generalized linear models extend classical linear regression for approximately normal responses to regression for non-normal responses, including binary responses or counts as in the preceding examples. In their original version, generalized linear models assume that the mean  $E(y|x)$  of the response given the covariates is related to a linear predictor  $\eta = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p = \beta_0 + x' \beta$  by a response or link function  $h$  in the form

$$E(y|x) = h(\beta_0 + x' \beta). \quad (1.1.1)$$

Due to the distributional assumptions, the variance function  $\text{var}(y|x)$  is then determined by choice of the specific distribution of  $y$  given  $x$ . Quasi-likelihood models and nonlinear or nonexponential family models allow us

**Table 1.2.** Cellular differentiation data

Number $y$ of cells differentiating	Dose of TNF(U/ml)	Dose of IFN(U/ml)
11	0	0
18	0	4
20	0	20
39	0	100
22	1	0
38	1	4
52	1	20
69	1	100
31	10	0
68	10	4
69	10	20
128	10	100
102	100	0
171	100	4
180	100	20
193	100	100

to weaken these assumptions. Most of the material of this chapter is fundamental for subsequent chapters.

### Models for Multicategorical Data (Chapter 3)

Model (1.1.1) is appropriate for unidimensional responses, like counts or binary 0-1-variables. However, in the Caesarian section study (Example 1.1) one may distinguish between different kinds of infection, getting a *multicategorical* response with categories “no infection,” “infection type I”, and “infection type II.” These response categories are not strictly ordered and therefore may be treated as unordered. In many cases the response of interest is given in ordered categories as in the following examples.

#### Example 1.4: Job expectation

In a study on the perspectives of students, psychology students at the University of Regensburg were asked if they expected to find adequate employment after getting their degree. The response categories were ordered with respect to their expectation. The responses were “don’t expect adequate employment” (category 1), “not sure” (category 2), and “immediately after the degree” (category 3). The data given in Table 1.3 show the response and the covariate age.  $\square$

**Table 1.3.** Grouped data for job expectations of psychology students in Regensburg

Observation number	Age in years	Response categories			$n_i$
		1	2	3	
1	19	1	2	0	3
2	20	5	18	2	25
3	21	6	19	2	27
4	22	1	6	3	10
5	23	2	7	3	12
6	24	1	7	5	13
7	25	0	0	3	3
8	26	0	1	0	1
9	27	0	2	1	3
10	29	1	0	0	1
11	30	0	0	2	2
12	31	0	1	0	1
13	34	0	1	0	1

**Example 1.5: Breathing test results**

Forthofer & Lehnen (1981) considered the effect of age and smoking on breathing test results for workers in industrial plants in Texas. The test results are given on an ordered scale with categories “normal,” “borderline”, and “abnormal.” It is of interest how age and smoking status are connected to breathing test results. Table 1.4 contains the data.  $\square$

Multicategorical responses cannot be considered as unidimensional. If the categories are labeled with numbers like  $1, \dots, k$ , the response looks like a univariate variable. But one has to keep in mind that the numbers are mere labels, especially if the categories are not ordered. Therefore, one has to consider a separate variable for each category in a multivariate modelling framework. Chapter 3 deals with multivariate approaches where the response variable is a vector  $y = (y_1, \dots, y_q)$ . Instead of one predictor  $\beta_0 + x'\beta$ , one has  $q$  predictors  $\beta_{01} + x'\beta_1, \dots, \beta_{0q} + x'\beta_q$ , that determine the response  $y_j$  in the form

$$E(y_j|x) = h_j(\beta_{01} + x'\beta_1, \dots, \beta_{0q} + x'\beta_q)$$

where  $h_j$  is a link function for the  $j$ th component of  $y$ . In analogy to the unidimensional model (1.1.1), one has

$$E(y|x) = h(\beta_{01} + x'\beta_1, \dots, \beta_{0q} + x'\beta_q)$$

with  $h = (h_1, \dots, h_q) : \mathbf{R}^q \rightarrow \mathbf{R}^q$  denoting the multidimensional link function.

This extension is particularly helpful in modelling multicategorical responses based on the multinomial distribution. Then  $y_j$  represents observations in category  $j$ . For unordered response categories the  $q$  predictors

**Table 1.4.** Breathing results of Houston industrial workers

Breathing test results					
Age	Smoking status	Normal	Borderline	Abnormal	
< 40	Never smoked	577	27	7	
	Former smoker	192	20	3	
	Current smoker	682	46	11	
40–59	Never smoked	164	4	0	
	Former smoker	145	15	7	
	Current smoker	245	47	27	

$\beta_{01} + x'\beta_1, \dots, \beta_{0q} + x'\beta_q$  usually are necessary. However, when the response is in ordered categories the use of the order information yields simpler models where the number of parameters may be reduced by assuming  $\beta_1 = \dots = \beta_q$ . A large part of this chapter considers ordinal response variables like the test results in breathing tests (Example 1.5) or expectations of students (Example 1.4).

In the preceding examples, we still considered the situation of *one* response variable, though multicategorical. The truly multivariate case arises if a vector of correlated responses is observed for each unit, as in the following example.

### Example 1.6: Visual impairment

In a visual impairment study (Liang, Zeger & Qaqish, 1992) binary responses  $y_1$  and  $y_2$  for both eyes of each individual of a sample population were recorded, indicating whether or not an eye was visually impaired. Covariates include age in years, race (white or black), and years of education. The main objective was to analyze the influence of race and age on visual impairment, controlling for education. The response variable “visual impairment” is bivariate, with two correlated binary components  $y_1, y_2$ , and an appropriate analysis has to take care of this.  $\square$

Similar situations occur, e.g., in twin and family studies or in dentistry. In other applications, the response vector consists of different variables, e.g., different questions in an interview. Section 3.5 surveys recent developments in regression analysis for correlated responses.

### Selecting and Checking Models (Chapter 4)

This chapter discusses some topics concerning the specification of models. The first topic is the selection of variables, which is important when one has 20 variables like in the credit-scoring example. Procedures and selection criteria that help to reduce the number of explanatory variables are considered. In Section 4.2 diagnostic tools are given within the framework of multivariate generalized linear models. These tools help to identify observations that are influential or determine the goodness-of-fit of the model in an extreme way. Residual analysis may show if the lack of fit is due to some specific observations (outliers) or if the model is inappropriate for most of the observations. In Section 4.3 an approach is outlined that does not investigate single observations but is based on tests that should reflect the appropriateness of a model. We do not consider tests for specific deviations, e.g., for the correct link function, but general specification tests that may be used in principle in any maximum likelihood type regression problem.

### Semi- and Nonparametric Approaches (Chapter 5)

Chapter 5 gives a brief introduction to semi- and nonparametric approaches. Nonparametric smoothing techniques are considered when one may keep faith with the data by low smoothing or one may produce a very smooth

**Table 1.5.** Visual impairment data (reproduced from Liang, Zeger & Qaqish, 1992)

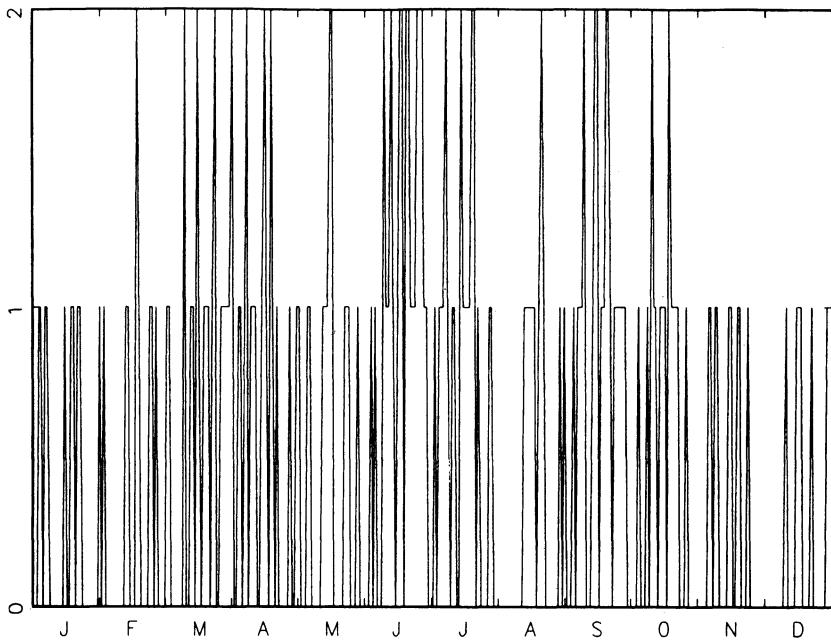
Visual impairment	White				Black				Total
	Age				Age				
	40–50	51–60	61–70	70+	40–50	51–60	61–70	70+	
Left eye									
Yes	15	24	42	139	29	38	50	85	422
No	617	557	789	673	750	574	473	344	4777
Right eye									
Yes	19	25	48	146	31	37	49	93	448
No	613	556	783	666	748	575	474	226	4751

estimate showing low fluctuations. The principle is not to use a parametric model but to let the data decide the functional form with a minimum of restrictions. Smoothing techniques are very flexible and allow one to look at data with varying degrees of smoothing.

### Example 1.7: Rainfall data

Figure 1.1 displays the number of occurrences of rainfall in the Tokyo area for each calendar day during the years 1983–1984 (Kitagawa, 1987). This is an example of a discrete time series. To compare it to similar data of other areas, or of other years, and to see some seasonal yearly pattern, it will be useful to estimate a smooth curve, representing the probability of rainfall on each calendar day.  $\square$

Simple regression smoothers like kernel smoothing are considered in Section 5.1 and extended to the case of categorical responses in Section 5.5. Alternative extensions of generalized linear models leading to additive models are based on the following approach: Instead of specifying the linear predictor  $x'\beta$ , covariates may enter the model by a smooth function  $f(x)$  yielding



**Figure 1.1.** Number of occurrences of rainfall in the Tokyo area for each calendar day 1983–1984.

$$E(y|x) = h(f(x))$$

with known response function. By splitting up the influence of covariates, one gets generalized additive models that have the form

$$E(y|x) = h(f_1(x_1) + \dots + f_p(x_p)),$$

where each component enters the model by a smooth function. Techniques of this type will give a picture for rainfall data that shows the pattern more clearly and is much more comfortable to look at.

### Fixed Parameter Models for Time Series and Longitudinal Data (Chapter 6)

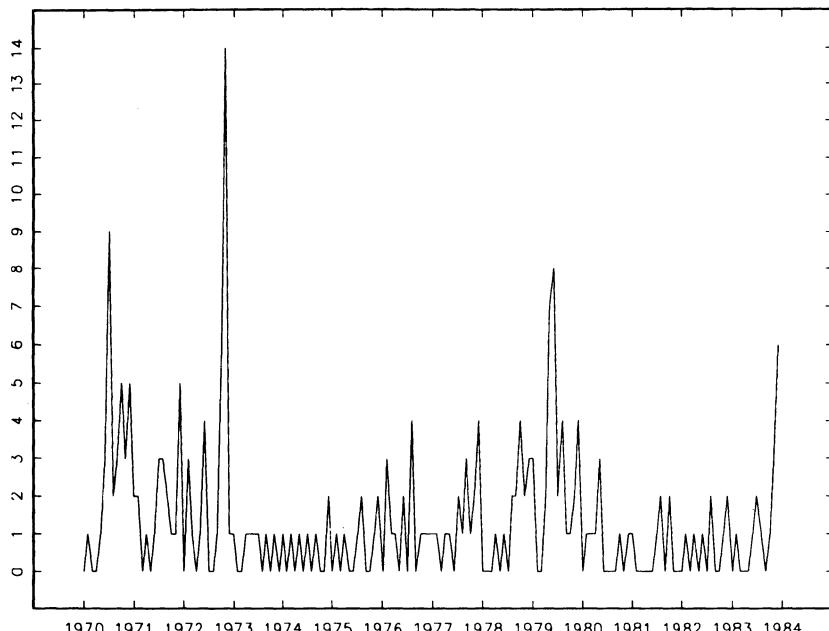
Chapter 6 extends the modelling approach of Chapters 2 and 3, which are mainly appropriate for (conditionally) independent observations  $(y_i, x_i), i = 1, \dots, n$  in a cross section, to time series data  $(y_t, x_t), t = 1, \dots, T$ , as in Example 1.8, and to longitudinal data, where many time series are observed simultaneously, as in Examples 1.9 and 1.10.

**Example 1.8: Polio incidence**

Figure 1.2 displays the monthly number of polio cases in the United States from 1970 to 1983 (Zeger, 1988a). This is a time series of count data, and analyzing it by traditional methods of time series analysis may give false conclusions. Looking at the plot in Figure 1.2, there seems to be some seasonality and, perhaps, a decrease in the rate of polio infection. Therefore, one might try to model the data with a log-linear Poisson model with trend and seasonal terms as covariates. However, in the time series situation one has to take into account dependence among observations. This can be accomplished by introducing past observations as additional covariates, as in common autoregressive models. Another possibility is to use a marginal modelling approach.  $\square$

**Example 1.9: IFO business test**

The IFO institute for economic research in Munich collects categorical monthly data of firms in various industrial branches. The monthly questionnaire contains questions on the tendency of realizations and expectations of variables like production, orders in hand, and demand. Answers are categorical, most of them trichotomous with categories like increase, decrease, or no change. Thus, for each firm the data form a categorical time series. Considering all firms within a certain branch, we have a categorical panel



**Figure 1.2.** Monthly number of polio cases in the United States from 1970 to 1983.

or longitudinal data. Based on such data, one may, for example, analyze the dependency of current production plans on demand and orders at hand.  $\square$

### **Example 1.10: Ohio children**

Within the Harvard Study of Air Pollution and Health, 537 children (Laird, Beck & Ware, 1984) were examined annually from age 7 to age 10 on the presence or absence of respiratory infection. Thus, we have four repeated measurements on one child or, in contrast to Example 1.9, many “short” binary time series. The only available covariate is the mother’s smoking status (regular smoker, nonregular smoker) at the beginning of the study. One of the primary goals was to analyze the influence of a mother’s smoking on children’s respiratory disease. Responses of one child, however, are likely to be correlated. This dependence should be taken into account in an adequate way. The data are given in Table 1.6, where “1” stands for infection and “0” for no infection.  $\square$

We distinguish between conditional or autoregressive-type models and marginal models. In conditional models, the independence assumption is dropped by including past observations  $y_{t-1}, y_{t-2}, \dots$  as additional covariates in generalized linear or quasi-likelihood models. This kind of modelling is quite analogous to autoregressive models for Gaussian time series, that means conditional densities  $f(y_t|y_{t-1}, y_{t-2}, \dots, x_t)$  are parametrically modelled. In certain applications, e.g., as in Example 1.10, the marginal effect of covariates  $x_t$  on  $y_t$  is of primary interest, whereas correlation of observations is regarded as a nuisance or only of secondary importance. Then it is more reasonable to base inference on marginal models for  $f(y_t|x_t)$  but taking into account correlation as in the work of Liang & Zeger (1986 and later).

### **Random Effects Models (Chapter 7)**

This chapter deals with random effects models for non-normal data. Such data appear in cross-sectional and longitudinal studies, when one has repeated measurements  $y_{it}, t = 1, \dots, T_i$ , from individual  $i$ . As an example consider the Ohio children data (Example 1.10). The presence or absence of respiratory infection is stated at four ages. Since children may respond differently to the smoking behavior of their mother, one may have so-called unobserved heterogeneity. Random effects models take into account unobserved heterogeneity, which arises from covariate effects varying from one individual to another or is due to omitted covariates. In analogy to linear random effects models for normal data, the models are defined as two-stage models or, equivalently, mixed models. In the first stage it is assumed that each of the repeated measurements  $y_{it}, t = 1, \dots, T_i$ , on individual  $i$  follows a (possibly autoregressive) generalized linear model with individual-specific unknown effects  $b_i$ . In the second stage, between-individual variation is introduced by assuming individual effects to be i.i.d. among individuals. Estimation of parameters and random effects, however, is more complicated

**Table 1.6.** Presence and absence of respiratory infection

Mother did not smoke					Mother smoked				
Age of child				Frequency	Age of child				Frequency
7	8	9	10		7	8	9	10	
0	0	0	0	237	0	0	0	0	118
0	0	0	1	10	0	0	0	1	6
0	0	1	0	15	0	0	1	0	8
0	0	1	1	4	0	0	1	1	2
0	1	0	0	16	0	1	0	0	11
0	1	0	1	2	0	1	0	1	1
0	1	1	0	7	0	1	1	0	6
0	1	1	1	3	0	1	1	1	4
1	0	0	0	24	1	0	0	0	7
1	0	0	1	3	1	0	0	1	3
1	0	1	0	3	1	0	1	0	3
1	0	1	1	2	1	0	1	1	1
1	1	0	0	6	1	1	0	0	4
1	1	0	1	2	1	1	0	1	2
1	1	1	0	5	1	1	1	0	4
1	1	1	1	11	1	1	1	1	7

than in linear random effects models. Several approaches based on posterior modes and means are discussed and applied.

### State Space and Hidden Markov Models (Chapter 8)

Chapter 8 surveys state space and hidden Markov modelling approaches for analyzing non-normal time series or longitudinal data, spatial data, and spatio-temporal data. State space and hidden Markov models have the common feature that they relate responses to unobserved “states” or “parameters” by an observation model. The states, which may represent, e.g., an unobserved temporal or spatial trend or time varying covariate effects, are assumed to follow a latent or “hidden” Markov model.

Traditionally, the terms state space and hidden Markov models are mostly used in the context of time series or longitudinal data  $\{y_t\}$ . Then the model consists of an observation model for  $y_t$  given the state  $\alpha_t$  and

a Markov chain model for the sequence  $\{\alpha_t\}$  of states. The term state space model is then mainly used for continuous states, and the term hidden Markov model for a finite state space. Given the observations  $y_1, \dots, y_t$  up to  $t$ , estimation of current, future, and past states (“filtering,” “prediction,” and “smoothing”) is a primary goal of inference. Sections 8.1 to 8.4 describe models and inference for time series and longitudinal data, with a focus on exponential family observation models, Gaussian state processes, and smoothing.

Section 8.5 considers some extensions to spatial and spatio-temporal data. For spatial data  $\{y_s\}$ , where  $s$  denotes the site or location in space of an observation, the latent or hidden model for the states  $\{\alpha_s\}$  now follows a Markov random field with continuous or discrete state space. For spatio-temporal data, where observations are available across time and space, hidden temporal and spatial Markov models are combined. We use the terms state space and hidden Markov models for all these cases.

### Survival Models (Chapter 9)

The objective here is to identify the factors that determine survival or transition. The case of continuous time is very briefly considered; only models that may be estimated in a way similar to generalized linear models are sketched. The focus in this chapter is on models for discrete time, which is a case often found in practice when only weeks or months of survival are recorded (see the following example). Some of the parametric approaches are strongly related to the models considered in ordinal modelling (Chapter 3). In addition to the parametric models, nonparametric smoothing techniques are briefly introduced.

### Example 1.11: Duration of unemployment

The data set comprises 1669 unemployed persons who are observed over several years in the socio-economic panel in Germany (Hanefeld, 1987). Time is measured in months. The focus of interest is how explanatory variables like gender, education level, or nationality influence the transition from unemployment to employment. Since the follow-up of persons over time is incomplete, one has to deal with the problem of censored observations. □

## 1.2 Remarks on Notation

When considering square roots of matrices, we use  $T$  as the transposed sign. Thus,  $A^{1/2}(A^{T/2})$  denotes the left (the corresponding right) square root of matrix  $A$  such that  $A^{1/2}A^{T/2} = A$  holds. The inverse matrices are denoted by  $A^{-1/2} = (A^{1/2})^{-1}$  and  $A^{-T/2} = (A^{T/2})^{-1}$ .

## 1.3 Notes and Further Reading

Several topics that are strongly related to generalized linear models but are treated extensively in books are only sketched here or simply omitted. McCullagh & Nelder (1989) is a standard source of information about generalized linear models. In particular, the univariate case is considered very extensively.

Since the focus here is on regression modelling, contingency table analysis is totally omitted. Agresti (1984, 1990) considers very thoroughly among other topics models for contingency tables with unordered or ordered categories. Sources for log-linear models are also Christensen (1997), Bishop, Fienberg & Holland (1975), and Tutz (2000). Whittaker (1990) and on a more formal level Lauritzen (1998) are books to consult when one is interested in graphical models. Santner & Duffy (1989) consider cross-classified data and univariate discrete data. A survey of exact inferences for discrete data is given in Agresti (1992).

Non- and semiparametric smoothing approaches are treated in more detail in Hastie & Tibshirani (1990), Green & Silverman (1994), Simonoff (1996), Loader (1999), and Härdle (1990a, 1990b). Interesting monographs that deal extensively with longitudinal data and repeated measurements are Jones (1993), Lindsey (1993), Longford (1993), and Diggle, Liang & Zeger (1994). Cressie (1993) is a comprehensive source for the analysis of spatial data.

A good survey on the rapidly growing field of Bayesian GLMs and extensions is provided by the chapters in Dey, Gosh & Mallick (1999).

Further sources of information in the area of generalized linear models are the Proceedings of GLIM and the International Workshop on Statistical Modelling (Gilchrist, Francis & Whittaker, 1985; Decarli, Francis, Gilchrist & Seeber, 1989; van der Heijden, Jansen, Francis & Seeber, 1992; Fahrmeir, Francis, Gilchrist & Tutz, 1992; Seeber, Francis, Hatzinger & Steckel-Berger, 1995; Forcina, Marchetti, Hatzinger & Galimacci, 1996; Minder & Friedl, 1997; Marx & Friedl, 1998; and Friedl, Berghold & Kauermann, 1999).

## 2

# Modelling and Analysis of Cross-Sectional Data: A Review of Univariate Generalized Linear Models

The material in this chapter provides an introduction to univariate generalized linear models and serves as a basis for the following chapters, which contain extensions, e.g., to multivariate, nonparametric, random effects, or dynamic models. It is not intended to replace a deeper study of detailed expositions like that in McCullagh & Nelder (1989) or, with focus on the GLIM package, that in Aitkin, Anderson, Francis & Hinde (1989). Shorter introductions are given, e.g., by Dobson (1989), Firth (1991), and Fahrmeir, Hamerle & Tutz (1996). Collett (1991) focuses on modelling of binary data and discusses practical aspects in more detail.

After an introductory section on the type of data that will be considered, Section 2.1 gives the general definition of generalized linear models and describes some important members. Binary and Poisson regression models are of particular relevance for the applications in this chapter as well as later. Binary regression models, such as the logit model, are appropriate for analyzing the relationship of a binary response with explanatory variables. In Example 1.1 the response is “infection” or “no infection”; in Example 1.2 the response is “creditworthy” or “not creditworthy.” Poisson regression models are useful if the response is a count like the number of cells differentiating, as in Example 1.3.

The common tools for estimation, testing, and simple goodness-of-fit criteria are contained in Section 2.2. Section 2.3 presents extensions to quasi-likelihood models for univariate responses, reviews Bayes approaches, and shortly considers nonlinear and nonexponential family models.

## 2.1 Univariate Generalized Linear Models

### 2.1.1 Data

Consider the common situation of cross-sectional regression analysis, where a univariate variable of primary interest, the response or dependent variable  $y$ , has to be explained by a vector  $x' = (x_1, \dots, x_m)$  of covariates or explanatory variables. The data

$$(y_i, x_i), \quad i = 1, \dots, n, \quad (2.1.1)$$

consist of observations on  $(y, x)$  for each unit or individual  $i$  of a cross section of size  $n$ . Responses can be continuous real variables as common in classical linear models, nonnegative, e.g., duration or income, counts as in Example 1.3, or binary as in Examples 1.1 and 1.2. Covariates may be quantitative (metrical), such as the duration of credit in months, the amount of credit in Example 1.2, the dose of TNF or IFN in Example 1.3, or qualitative (ordered or unordered categorical), such as the dichotomous covariates “Caesarian birth” (planned or not), “presence of risk factors” (yes/no), “antibiotics” (given or not) in Example 1.1, or “running account” (categories no, medium, good) in Example 1.2. Covariates may be deterministic, i.e., known values or experimental conditions, or they may be stochastic, i.e., observations of a random vector  $x$ .

#### Coding of Covariates

As in common linear regression, qualitative covariates have to be coded appropriately. A categorical covariate (or factor)  $x$  with  $k$  possible categories  $1, \dots, k$  will generally be coded by a “dummy vector” with  $q = k - 1$  components  $x^{(1)}, \dots, x^{(q)}$ . If 0 – 1 dummies are used, which is shortly referred to as “dummy coding”, then  $x^{(j)}$  is defined by

$$x^{(j)} = \begin{cases} 1 & \text{if category } j \text{ is observed,} \\ 0 & \text{otherwise,} \end{cases} \quad j = 1, \dots, q.$$

If the  $k$ th category, the “reference category,” is observed, then  $x$  is the zero vector.

An alternative coding scheme, which is referred to as “effect coding”, is defined by

$$x^{(j)} = \begin{cases} 1 & \text{if category } j \text{ is observed,} \\ -1 & \text{if category } k \text{ is observed,} \\ 0 & \text{otherwise,} \end{cases} \quad j = 1, \dots, q.$$

In the case of effect coding, the reference category  $k$  is given by the vector  $(-1, \dots, -1)$  instead of the zero vector. Other types of coding may sometimes be more useful for the purpose of interpretation, but will not be considered in this text.

### Grouped and Ungrouped Data

In (2.1.1), it was implicitly assumed that the data are *ungrouped*, i.e.,  $(y_i, x_i)$  is the original observation on unit  $i$  of the cross section. In this case, each covariate vector  $x'_i = (x_{i1}, \dots, x_{im})$  corresponds exactly to one unit  $i$  and to the  $i$ th row of the data matrix  $X$ :

$$\begin{array}{ll} \text{Unit 1} & \left[ \begin{array}{c} y_1 \\ \vdots \\ y_n \end{array} \right], \quad X = \left[ \begin{array}{ccc} x_{11} & \cdots & x_{1m} \\ \vdots & & \vdots \\ x_{i1} & \cdots & x_{im} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nm} \end{array} \right]. \\ \vdots & \\ \text{Unit } i, & \\ \vdots & \\ \text{Unit } n & \end{array}$$

The cellular differentiation data (Table 1.2) are an example of ungrouped data. If some of the covariate vectors or rows of  $X$  have identical covariate values  $(x_{i1}, \dots, x_{im})$ , the data may be *grouped*: After relabeling the index, only rows  $x_i$  with different combinations of covariate values appear in the data matrix  $X$ , together with the number  $n_i$  of repetitions and the arithmetic mean  $\bar{y}_i$  of the individual responses on the same vector  $x_i$  of covariates. Thus, grouped data are of the form

$$\begin{array}{ll} \text{Group 1} & \left[ \begin{array}{c} n_1 \\ \vdots \\ n_g \end{array} \right], \quad \bar{y} = \left[ \begin{array}{c} \bar{y}_1 \\ \vdots \\ \bar{y}_g \end{array} \right], \quad X = \left[ \begin{array}{ccc} x_{11} & \cdots & x_{1m} \\ \vdots & & \vdots \\ x_{i1} & \cdots & x_{im} \\ \vdots & & \vdots \\ x_{g1} & \cdots & x_{gm} \end{array} \right] \\ \vdots & \\ \text{Group } i, & \\ \vdots & \\ \text{Group } g & \end{array}$$

where  $g$  is the number of different covariates  $x_i$  in the data set,  $n_i$  denotes the number of units with equal covariate vector  $x_i$ , and  $\bar{y}_i$  denotes the arithmetic mean. Equivalently, instead of the arithmetic mean, the total sum of responses could be used. In particular if individual responses are binary, coded by 0–1 dummies, it is common to use absolute instead of relative frequencies and to display them in the form of contingency tables. It is seen that ungrouped data are a special case of grouped data for  $n_1 = \dots = n_g =$

1. Grouping of ungrouped raw data is generally possible and can lead to a considerable amount of data reduction if *all covariates are categorical*, as, e.g., in Example 1.1. In this example, individual responses in the raw data set were binary (infection yes or no), whereas Table 1.1 shows the grouped data. With *metrical covariates* it can often be impossible to group raw data, as, e.g., in Example 1.3, or there are only very few individuals with identical covariate values that can be grouped, as in Example 1.2.

For the following main reasons it is important to distinguish between grouped and ungrouped data:

- (i) Some statistical procedures are meaningful only for grouped data, in particular in connection with goodness-of-fit tests, residuals and influence analysis.
- (ii) Asymptotic results for grouped data can be obtained under  $n_i \rightarrow \infty$  for all  $i = 1, \dots, g$ , or under  $n \rightarrow \infty$ , without necessarily requiring  $n_i \rightarrow \infty$ . The latter type of asymptotics, which is more difficult to deal with, is appropriate for ungrouped data or grouped data with small  $n_i$ .
- (iii) From the computational point of view, considerable savings in computing times and memory requirements can be achieved by grouping the data as far as possible.

### 2.1.2 Definition of Univariate Generalized Linear Models

The *classical* linear model for ungrouped normal responses and deterministic covariates is defined by the relation

$$y_i = z'_i \beta + \epsilon_i, \quad i = 1, \dots, n,$$

where  $z_i$ , the design vector, is an appropriate function of the covariate vector  $x_i$  and  $\beta$  is a vector of unknown parameters. For a vector of metric variables the simplest form of  $z_i$  is  $z'_i = (1, x'_i)$ , for a vector of qualitative variables or a mixture of metric and qualitative variables dummy variables have to be included. The errors  $\epsilon_i$  are assumed to be normally distributed and independent,

$$\epsilon_i \sim N(0, \sigma^2).$$

We rewrite the model in a form that leads to generalized linear models in a natural way: The observations  $y_i$  are independent and normally distributed,

$$y_i \sim N(\mu_i, \sigma^2), \quad i = 1, \dots, n, \tag{2.1.2}$$

with  $\mu_i = E(y_i)$ . The mean  $\mu_i$  is given by the linear combination  $z'_i \beta$ ,

$$\mu_i = z'_i \beta, \quad i = 1, \dots, n. \tag{2.1.3}$$

If covariates are stochastic, we assume the pairs  $(y_i, x_i)$  to be independent and identically distributed. Then the model is to be understood conditionally, i.e., (2.1.2) is the conditional density of  $y_i$  given  $x_i$ , and the  $y_i$  are conditionally independent.

This remark applies also to the following definition of generalized linear models, where the preceding assumptions are relaxed in the following way:

1. *Distributional assumption:*

Given  $x_i$ , the  $y_i$  are (conditionally) independent, and the (conditional) distribution of  $y_i$  belongs to a simple exponential family with (conditional) expectation  $E(y_i | x_i) = \mu_i$  and, possibly, a common scale parameter  $\phi$ , not depending on  $i$ .

2. *Structural assumption:*

The expectation  $\mu_i$  is related to the linear predictor  $\eta_i = z'_i \beta$  by

$$\mu_i = h(\eta_i) = h(z'_i \beta) \quad \text{resp.,} \quad \eta_i = g(\mu_i),$$

where

$h$  is a known one-to-one, sufficiently smooth response function,

$g$  is the link function, i.e., the inverse of  $h$ ,

$\beta$  is a vector of unknown parameters of dimension  $p$ , and

$z_i$  is a design vector of dimension  $p$ , which is determined as an appropriate function  $z_i = z(x_i)$  of the covariates.

Thus, a specific generalized linear model is fully characterized by three components:

- *the type of the exponential family,*
- *the response or link function, and*
- *the design vector.*

Some important models, in particular for binary responses, are considered in more detail in the following subsections. Therefore, only general remarks on the three components are given here.

(i) Exponential families and some of its properties are described in more detail in Appendix A1. For univariate generalized linear models as considered in this chapter, the densities of responses  $y_i$  can always be written as

$$f(y_i | \theta_i, \phi, \omega_i) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi} \omega_i + c(y_i, \phi, \omega_i) \right\}, \quad (2.1.4)$$

where

$\theta_i$  is the so-called natural parameter,

$\phi$  is an additional scale or dispersion parameter,

$b(\cdot)$  and  $c(\cdot)$  are specific functions corresponding to the type of exponential family, and

$\omega_i$  is a weight with  $\omega_i = 1$  for ungrouped data ( $i = 1, \dots, n$ ) and  $\omega_i = n_i$  for grouped data ( $i = 1, \dots, g$ ) if the *average* is considered as response (or  $\omega_i = 1/n_i$  if the *sum* of individual responses is considered).

Important members are the normal, the binomial, the Poisson, the gamma, and the inverse Gaussian distributions. Their characteristics are expressed in exponential family terms in Table 2.1. The natural parameter  $\theta$  is a function of the mean  $\mu$ , i.e.,  $\theta_i = \theta(\mu_i)$ , which is uniquely determined by the specific exponential family through the relation  $\mu = b'(\theta) = \partial b(\theta)/\partial\theta$ . Moreover, the variance of  $y$  is of the form

$$\text{var}(y_i|x_i) = \sigma^2(\mu_i) = \phi v(\mu_i)/\omega_i \quad (2.1.5)$$

where the variance function  $v(\mu)$  is uniquely determined by the specific exponential family through the relation  $v(\mu) = b''(\theta) = \partial^2 b(\theta)/\partial\theta^2$ . Thus, specification of the mean structure by  $\mu = h(z'\beta)$  implies a certain variance structure. However, many results and procedures remain valid under appropriate modifications if the mean and the variance function are specified separately, thereby dropping the exponential family assumption and considering quasi-likelihood models (see Section 2.3.1).

(ii) The choice of appropriate response or link functions depends on the specific exponential family, i.e., on the type of response and on the particular application. For each exponential family there exists a so-called natural or canonical link function. Natural link functions relate the natural parameter directly to the linear predictor:

$$\theta = \theta(\mu) = \eta = z'\beta, \quad (2.1.6)$$

i.e.,  $g(\mu) \equiv \theta(\mu)$ . The natural link functions can thus be determined from Table 2.1, e.g.,

$$\begin{aligned} \eta &= \mu && \text{for the normal,} \\ \eta &= \log \mu && \text{for the Poisson,} \\ \eta &= \log \left[ \frac{\mu}{1-\mu} \right] && \text{for the Bernoulli} \end{aligned}$$

distribution. Natural link functions lead to models with convenient mathematical and statistical properties. However, this should not be the main reason for choosing them, and non-natural link functions may be more appropriate in a particular application.

(iii) Concerning the design vector, nothing new has to be said as compared to linear models: In most cases a constant, corresponding to the “grand mean,” is added so that  $z$  is of the form  $z = (1, w)$ . Metrical covariates can be incorporated directly or after appropriate transformations like  $\log(x), x^2, \dots$  etc. Categorical covariates, ordered or unordered, have to be

**Table 2.1.** Simple exponential families with dispersion parameter

$f(y \theta, \phi, \omega) = \exp \left\{ \frac{y\theta - b(\theta)}{\phi} \omega + c(y, \phi, \omega) \right\}$				
(a) Components of the exponential family				
Distribution	$\theta(\mu)$	$b(\theta)$	$\phi$	
Normal	$N(\mu, \sigma^2)$	$\mu$	$\theta^2/2$	$\sigma^2$
Bernoulli	$B(1, \pi)$	$\log(\pi/(1-\pi))$	$\log(1+\exp(\theta))$	1
Poisson	$P(\lambda)$	$\log \lambda$	$\exp(\theta)$	1
Gamma	$G(\mu, \nu)$	$-1/\mu$	$-\log(-\theta)$	$\nu^{-1}$
Inverse Gaussian	$IG(\mu, \sigma^2)$	$1/\mu^2$	$-(-2\theta)^{1/2}$	$\sigma^2$
(b) Expectation and variance				
Distribution	$E(y) = b'(\theta)$	var. fct. $b''(\theta)$	$\text{var}(y) = b''(\theta)\phi/\omega$	
Normal	$\mu = \theta$	1	$\sigma^2/\omega$	
Bernoulli	$\pi = \frac{\exp(\theta)}{1+\exp(\theta)}$	$\pi(1-\pi)$	$\pi(1-\pi)/\omega$	
Poisson	$\lambda = \exp(\theta)$	$\lambda$	$\lambda/\omega$	
Gamma	$\mu = -1/\theta$	$\mu^2$	$\mu^2\nu^{-1}/\omega$	
Inverse Gaussian	$\mu = (-2\theta)^{-1/2}$	$\mu^3$	$\mu^3\sigma^2/\omega$	

Derivatives are denoted by  $b'(\theta) = \partial b(\theta)/\partial\theta$ ,  $b''(\theta) = \partial^2 b(\theta)/\partial\theta^2$ . The weight  $\omega$  is equal to 1 for individual ungrouped observations. For grouped data,  $y$  denotes the average of individual responses, the densities are scaled, and the weight  $\omega$  equals the group size (i.e., the number of repeated observations in a group).

coded by a dummy vector as described in Section 2.1.1: For the unidimensional covariate  $x \in \{1, \dots, k\}$ , a meaningful linear predictor is

$$\eta = \beta_0 + x^{(1)}\beta_1 + \dots + x^{(q)}\beta_q$$

where  $x^{(1)}, \dots, x^{(q)}$  are dummy variables as given in Section 2.1.1.

If the original vector  $x$  comprises more than one qualitative component, the linear predictor will contain dummy variables for all the components and possibly products of dummy variables (interactions). In addition to “main effects,” the “interaction effects” that are of interest are obtained by adding the product of two or more covariates to the design vector. Metric covariates are multiplied directly; for categorical covariates, corresponding dummy vectors have to be multiplied. As in linear models, one has to be aware of the problem of multicollinearity or aliasing, to avoid problems of parameter identifiability.

(iv) A last remark concerns the relationship between models for grouped and ungrouped data. Suppose that ungrouped data  $(y_i, x_i)$ ,  $i = 1, \dots, n$ , are modelled by a generalized linear model with response and variance function

$$E(y_i|x_i) = \mu_i = h(z'_i\beta), \quad \text{var}(y_i|x_i) = \phi v(\mu_i).$$

If data are grouped as in Section 2.1.1, with  $\bar{y}_i$  denoting the arithmetic mean of  $n_i$  observations on the same covariate, resp. design value  $z_i = z(x_i)$ , then the distribution of  $\bar{y}_i$  is within the exponential family again, with the same mean structure

$$E(\bar{y}_i|x_i) = h(z'_i\beta)$$

as for ungrouped data, but with variance

$$\text{var}(\bar{y}_i|x_i) = \frac{\phi v(\mu_i)}{n_i};$$

see Appendix A.1. So the definition of generalized linear models applies to grouped data as well if the variance function is properly adjusted by  $1/n_i$ , or, equivalently, by defining the weights  $\omega_i$  in (2.1.4) as  $\omega_i = n_i$ . Note that in Table 2.1 the response  $y$  refers to the average over  $n_i$  observations if observations are grouped. Then  $\omega$  is the number of grouped observations instead of  $\omega = 1$  as for the ungrouped case.

### 2.1.3 Models for Continuous Responses

#### Normal Distribution

Assuming a normal distribution choice of the natural link function leads to the classical linear normal model

$$\mu = \eta = z'\beta.$$

Sometimes a nonlinear relationship

$$\mu = h(\eta)$$

will be more appropriate, e.g.,

$$h(\eta) = \eta^2, \quad h(\eta) = \log \eta, \quad \text{or} \quad h(\eta) = \exp \eta.$$

This type of nonlinear normal regression can easily be handled within the framework of generalized linear models.

## Gamma Distribution

The gamma distribution is useful for regression analysis of (nearly) continuous nonnegative variables, such as life spans, insurance claims, amount of rainfall, etc. In terms of its mean  $\mu > 0$  and the shape parameter  $\nu > 0$ , the density is given by

$$f(y|\mu, \nu) = \frac{1}{\Gamma(\nu)} \left( \frac{\nu}{\mu} \right)^\nu y^{\nu-1} \exp \left( -\frac{\nu}{\mu} y \right), \quad y \geq 0.$$

From Table 2.1 we have

$$\text{var}(y) = \sigma^2(\mu) = \phi\mu^2,$$

with  $\phi = 1/\nu$ . The shape parameter determines the form of the density. For  $0 < \nu < 1$ ,  $f(y)$  decreases monotonically, with  $f(y) \rightarrow \infty$  for  $y \rightarrow 0$  and  $f(y) \rightarrow 0$  for  $y \rightarrow \infty$ . For  $\nu = 1$ , the exponential distribution is obtained as a special case. For  $\nu > 1$ , the density is zero at  $y = 0$ , has a mode at  $y = \mu - \mu/\nu$ , and is positively skewed. Figure 2.1 displays the density for  $\nu = 0.5, 1.0, 2.0$ , and  $5.0$  ( $\mu = 1$ ). It can be seen that the densities are all positively skewed, but a normal limit is attained as  $\nu \rightarrow \infty$ .

The natural response function is the reciprocal

$$\mu = \eta^{-1}.$$

Since  $\mu > 0$ , the linear predictor is restricted to  $\eta = z'\beta > 0$ , implying restrictions on  $\beta$ . Two other important response functions are the identity

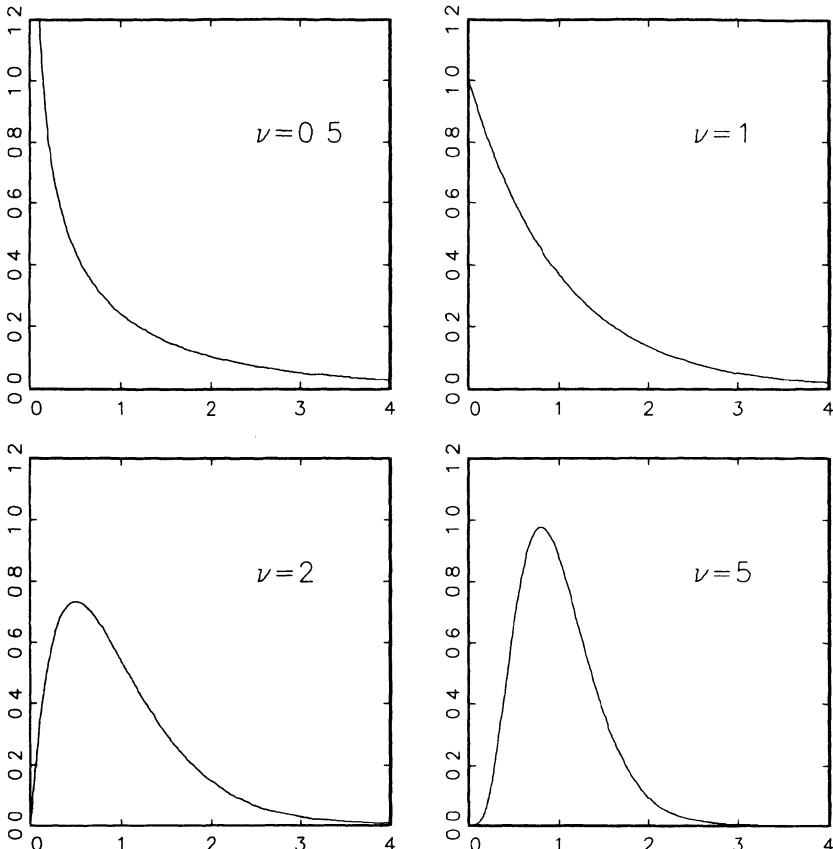
$$h(\eta) = \eta = \mu$$

and the exponential response function

$$h(\eta) = \exp(\eta) = \mu,$$

or, equivalently, the log-link

$$g(\mu) = \log(\mu) = \eta.$$



**Figure 2.1.** The gamma distribution:  $G(\mu = 1, \nu)$ .

### Inverse Gaussian Distribution

This distribution can be applied for nonsymmetric regression analysis and for lifetimes. Detailed expositions are given in Folks & Chhikara (1978) and Jorgensen (1982).

#### 2.1.4 Models for Binary and Binomial Responses

Consider first the case of ungrouped binary responses, coded by 0 and 1. Given the covariate  $x$ , a binary variable  $y$  is completely determined by its response probability

$$E(y|x) = P(y = 1|x) = \pi,$$

implying

$$\text{var}(y|x) = \pi(1 - \pi).$$

If binary data are grouped as in Section 2.1.1, we let  $\bar{y}$  denote the relative frequency of observed 1s for the, say  $m$ , independent binary observations with the same covariate vector  $x$ . The absolute frequencies  $m\bar{y}$  are binomially distributed with

$$E(m\bar{y}|x) = m\pi, \quad \text{var}(m\bar{y}|x) = m\pi(1 - \pi).$$

The relative frequencies  $\bar{y}$  are *scaled binomial*, i.e., they take the corresponding values  $0, 1/m, 2/m, \dots, 1$  with the same binomial probabilities as  $m\bar{y}$ , and

$$E(\bar{y}|x) = \pi, \quad \text{var}(\bar{y}|x) = \frac{\pi(1 - \pi)}{m}.$$

Thus, for grouped binary data, response functions are the same as for individual binary responses, whereas the variance function has to be divided by  $m$ , i.e.,  $\omega = m$ .

If the individual (ungrouped) response (given  $x$ ) is binomially distributed with  $y \sim B(m, \pi)$ , we will usually consider the scaled binomial or relative frequency  $y/m$  as response. Because  $y/m$  may be understood as an average of individual independent binary observations, this case may be treated as occurring from grouped observations. Then the variance

$$\text{var}(y/m) = \pi(1 - \pi)/m$$

is the same as for grouped binary observations when grouping is done over  $m$  observations. Models for binary and binomial responses are determined by relating the response probability  $\pi$  to the linear predictor  $\eta = z'\beta$  via some response function  $\pi = h(\eta)$ , resp. link function  $g(\pi) = \eta$ . The following models are the most common.

### Linear Probability Model

In analogy to linear normal models,  $\pi$  is related directly to  $\eta$  by the identity link

$$\pi = \eta = z'\beta.$$

Though easy to interpret, this model has a serious drawback: As  $\pi$  is a probability,  $z'\beta$  has to be in  $[0, 1]$  for all possible values of  $z$ . This implies severe restrictions on  $\beta$ .

This disadvantage is avoided by the following models. They relate  $\pi$  to the linear predictor  $\eta$  by

$$\pi = F(\eta), \tag{2.1.7}$$

where  $F$  is a strictly monotonous distribution function on the whole real axis, so that no restrictions on  $\eta$  and on  $\beta$  have to be imposed.

## Probit Model

The probit model is defined by

$$\pi = \Phi(\eta) = \Phi(z'\beta),$$

where  $\Phi$  is the standard normal distribution function. It imposes no restrictions on  $\eta$ ; however, it is computationally more demanding when one computes likelihoods.

## Logit Model

The logit model corresponds to the natural link function

$$g(\pi) = \log \left\{ \frac{\pi}{1 - \pi} \right\} = \eta,$$

with the logistic distribution function

$$\pi = h(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

as the resulting response function. The logistic distribution function also has support on the entire real axis and is symmetric, but it has somewhat heavier tails than the standard normal. Apart from  $\pi$  values near 0 or 1, which correspond to the tails, fits by probit or logit models are generally quite similar (see Figure 2.2). Since the logistic function is easier to compute than the standard normal, logit models are often preferred in practice.

## Complementary Log-Log Model

This model has the link function

$$g(\pi) = \log(-\log(1 - \pi)),$$

and the extreme minimal-value distribution function

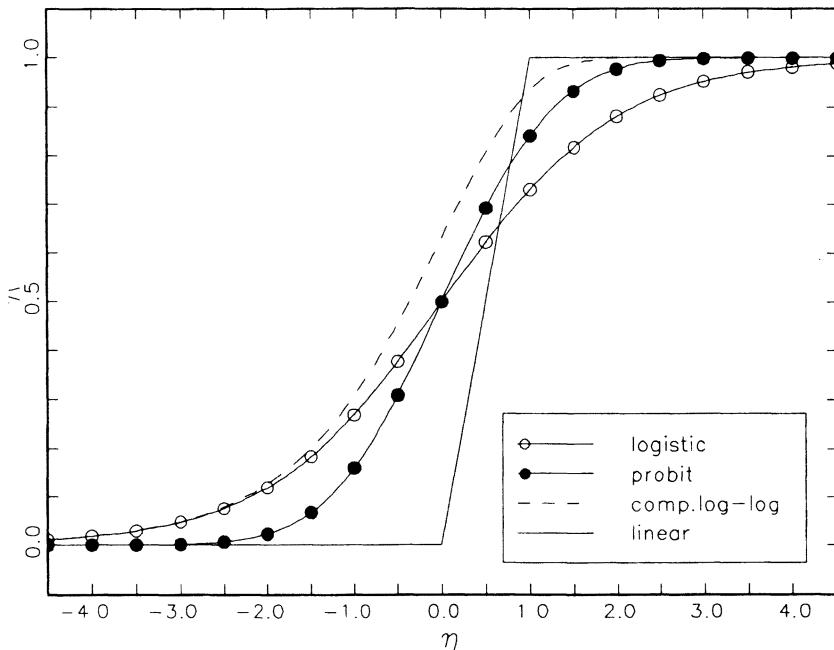
$$h(\eta) = 1 - \exp(-\exp(\eta))$$

as the corresponding response function. It is nonsymmetric, close to the logistic function for small  $\pi$ , but with a considerably less heavy right tail.

## Complementary Log-Model

The link function that gives this model its name is specified by

$$g(\pi) = -\log(1 - \pi).$$



**Figure 2.2.** Response functions for binary responses.

The corresponding response function is based on the exponential distribution with

$$h(\eta) = \begin{cases} 1 - \exp(-\eta), & \eta \geq 0 \\ 0 & \eta < 0 \end{cases}$$

In contrast to the other models, the response function of the complementary model may not be differentiated at the value  $\eta = 0$ . Consequently, estimation problems sometimes occur. For further motivation of the model and an overview of existing applications, see Piegorsch (1992).

The response functions corresponding to the linear probability, probit, logit, and complementary log-log models are displayed in Figure 2.2, where the values of  $\eta$  are plotted against  $\pi$ . Figure 2.2 suggests that the response functions are quite different. However, one should keep in mind that the predictor  $\eta$  is linear in the form  $\eta = z'\beta$ , for simplicity  $\eta = \beta_0 + x\beta_1$ . So if instead of the distribution function  $F$  the transformed distribution function  $\tilde{F}(u) = F(\frac{u-\mu}{\sigma})$  is used the models

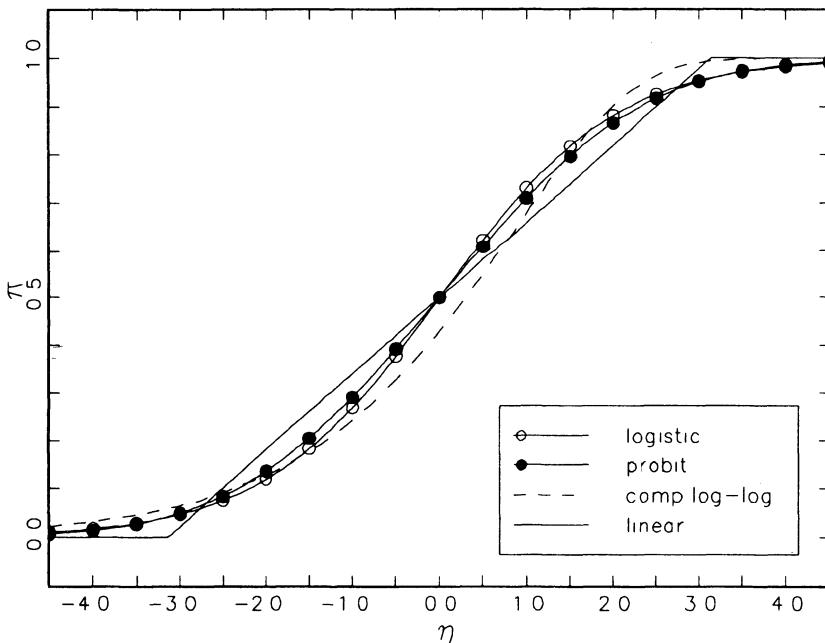
$$\pi = F(\beta_0 + x\beta_1) \quad \text{and} \quad \pi = \tilde{F}(\tilde{\beta}_0 + x\tilde{\beta}_1)$$

are equivalent by setting  $\tilde{\beta}_0 = \sigma\beta_0 + \mu$  and  $\tilde{\beta}_1 = \sigma\beta_1$ . Therefore, models should be compared for the appropriate scaling of the predictor  $\eta$ . This

may be done by transforming  $F$  so that the mean and variance according to different distribution functions are the same. For the functions used earlier mean and variance are given by

Response function $F$	Mean	Variance
linear	0.5	$1/12$
probit	0.0	1
logistic	0.0	$\pi^2/3$
compl. log-log	-0.5772	$\pi^2/6$

where  $\pi = 3.14159$ . Figure 2.3 displays the four response functions with  $\eta$  adjusted to have the logistic mean and variance for all four response functions.



**Figure 2.3.** Response functions for binary responses adjusted to the logistic function (that means linear transformation yielding mean zero and variance  $\pi^2/3$ ).

In contrast to Figure 2.2, the logistic and probit functions, which now have variance  $\pi^2/3$ , are nearly identical. Therefore, fits of probit and logit models are generally quite similar after the adjustment of  $\eta$ , which is most often implicitly done in estimation. The complementary log-log function,

however, is steeper than the logistic and probit functions even after the adjustment. Thus, for small values of  $\eta$  it approaches 0 more slowly, and as  $\eta$  approaches infinity it approaches 1 faster than the logistic and adjusted probit functions.

### Binary Models as Threshold Models of Latent Linear Models

As presented so far, binary response models seem to be ad hoc specifications having some useful properties. However, all these models can be derived as threshold models, where binary responses  $y$  are based on a latent continuous variable  $\tilde{y}$  that obeys a linear model

$$\tilde{y} = \alpha_0 + w' \alpha + \sigma \epsilon,$$

where  $\epsilon$  is distributed according to  $F(\cdot)$ , e.g., a logistic or standard normal distribution function, and  $\sigma$  is a scale parameter. The relation between  $y$  and  $\tilde{y}$  is given by

$$y = \begin{cases} 1, & \tilde{y} \leq \tau, \\ 0, & \tilde{y} > \tau, \end{cases}$$

with a threshold value  $\tau$ . From this assumption one gets

$$P(y = 1) = P(\alpha_0 + w' \alpha + \sigma \epsilon \leq \tau) = F\left(\frac{\tau - \alpha_0 - w' \alpha}{\sigma}\right).$$

Defining

$$\beta = \left(\frac{\tau - \alpha_0}{\sigma}, \frac{\alpha'}{\sigma}\right)', \quad z' = (1, -w'),$$

one obtains the general model (2.1.7). It should be noted that covariate effects  $\alpha$  of the underlying linear model can be identified only up to the common but generally unknown factor  $1/\sigma$ , and that the grand mean parameter  $\alpha_0$  cannot be identified at all if  $\tau$  is unknown.

### Parameter Interpretation

If we base the binary model on a latent linear model as above, covariate effects  $\beta = (\beta_1, \beta_2, \dots)$  may be interpreted with respect to this latent model. The preceding discussion shows that only relative values, e.g.,  $\beta_1/\beta_2$ , but not absolute values  $\beta_1, \beta_2$  are meaningful.

Parameter interpretation becomes more difficult for direct interpretation of covariate effects on the binary response  $y$ . For the logistic model, we have a linear model for the “log odds,” i.e., the logarithm of the odds  $\pi/(1 - \pi)$  of a response  $y = 1$ . So interpretation of covariate effects on the log odds is

the same as on the expectation  $\mu = E(y)$  in the linear model. In particular, in a medical context, the odds are often called “relative risk.” For a linear predictor of the form  $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ , say, we have

$$\frac{\pi}{1 - \pi} = \exp(\beta_0) \exp(\beta_1 x_1) \exp(\beta_2 x_2)$$

so that (exponentials of) covariate effects have a multiplicative effect on the relative risk. However, this kind of interpretation is not possible for other models. Generally it seems best to break up interpretation in two stages:

1. Interpret covariate effects on the linear predictor  $\eta = z'\beta$  in the same way as in linear models.
2. Transform this linear effect on  $\eta$  into a nonlinear effect on  $\pi$  with the aid of a graph of the response function

$$\pi = h(\eta),$$

as given in Figures 2.2 and 2.3.

### **Example 2.1: Caesarian birth study**

Recall Example 1.1 where the response variable of interest is the occurrence or nonoccurrence of infection following birth by Caesarian section. Ignoring the two different types of infection, we form a binary response variable  $y$  with  $y = 0$  if there is no infection and  $y = 1$  if there is infection of type I or II. The risk of infection will be modelled with respect to the following three covariates, which are all dichotomous:

- NOPLAN : whether or not the Caesarian was planned,  
 FACTOR : presence of one or more risk factors, such as diabetes,  
           excessive weight, early labor pains, and others,  
 ANTIB : whether antibiotics were given as prophylaxis.

Table 2.2 summarizes the data of 251 births and is obtained from Table 1.1 by combining the two types I and II of infection.

We model the probability of infection by a logit model. All covariates are 0–1-coded (NOPLAN = 1 for “the Caesarian was not planned,” ANTIB = 1 for “there were antibiotics given,” FACTOR = 1 for “there were risk factors present”). Including only three covariates (and no interactions) in the explanatory linear term, the model can be explicitly written in the form

$$\log \frac{P(\text{infection})}{P(\text{no infection})} = \beta_0 + \beta_1 \text{NOPLAN} + \beta_2 \text{FACTOR} + \beta_3 \text{ANTIB}.$$

The estimates of the covariate effects are given by

covariate	1	NOPLAN	FACTOR	ANTIB
effect	-1.89	1.07	2.03	-3.25

**Table 2.2.** Grouped data on infection

		Caesarian planned		Not planned	
		Infection		Infection	
		yes	no	yes	no
<b>Antibiotics</b>					
Risk factors		1	17	11	87
No risk factors		0	2	0	0
<b>No antibiotics</b>					
Risk factors		28	30	23	3
No risk factors		8	32	0	9

The estimates can be interpreted using the model formula above. Antibiotics lower the log odds of infection. Since the logarithm is strictly monotone, we can directly conclude that antibiotics lower the odds of infection itself. Similarly, risk-factors and the fact that a Caesarian was not planned do increase the odds of infection.

For a better interpretation of the parameter values, we can rewrite the model in the form

$$\frac{P(\text{infection})}{P(\text{no infection})} = \exp(\beta_0) \cdot \exp(\beta_1 \text{NOPLAN}) \cdot \exp(\beta_2 \text{FACTOR}) \cdot \exp(\beta_3 \text{ANTIB}).$$

It can now be easily seen that when the Caesarian was not planned the odds of infection increase by a factor of  $\exp(\hat{\beta}_1) = \exp(1.07) = 2.92$ . Additionally, with risk factors present the odds of infection is  $\exp(\hat{\beta}_2) = 7.6$  times the odds without risk factors present.

Note, however, that all these values are valid only if the model fitted is accurate. For comparison we also fit a probit model with the same covariates. Parameter estimates are now  $-1.09, 0.61, 1.20$ , and  $-1.90$ . Although absolute values are quite different, relative values or ratios of effects are nearly the same as for the logit model.  $\square$

### Example 2.2: Credit-scoring

If banks give a credit to a client, they are interested in estimating the risk that the client will not pay back the credit as agreed by contract. The aim of credit-scoring systems is to model or predict the probability that a client with certain covariates (“risk factors”) is to be considered as a potential risk. We will analyze the effect of covariates on the binary response “creditability” by a logit model. Other tools currently used in credit-scoring

are (linear) discriminance analysis, classification and regression trees, and neural networks.

The data set consists of 1000 consumers' credits from a southern German bank. The response variable of interest is "creditability," which is given in dichotomous form ( $y = 0$  for creditworthy,  $y = 1$  for not creditworthy). In addition, 20 covariates that are assumed to influence creditability were collected. The raw data are recorded in Fahrmeir & Hamerle (1984, see p. 334 ff and p. 751 ff.) and are available on electronic file. Fahrmeir & Hamerle (1984, p. 285–86) used a logit model to analyze a subset of these data containing only the following seven covariates, which are partly metrical and partly categorical:

- X1 running account, trichotomous with categories "no running account" (= 1), "good running account" (= 2), "medium running account" ("less than 200 DM" = 3 = reference category)
- X3 duration of credit in months, metrical
- X4 amount of credit in DM, metrical
- X5 payment of previous credits, dichotomous with categories "good," "bad" (= reference category)
- X6 intended use, dichotomous with categories "private" or "professional" (= reference category)
- X7 and X8 are dummies for gender and marital status with reference category "male" resp. "living alone."

Since there are only three clients with the same covariate values, it is not possible to present the data in grouped form. Individual data of clients look like this:

Client	Y	X1	X3	X4	X5	X6	X7	X8
No. 1	1	1	18	1049	0	1	0	0
No. 23	0	3	36	2348	1	1	0	0

It is important to note that the data come from a *stratified sample*: 300 clients were drawn from the stratum defined by  $y = 1$  (not creditworthy), while 700 clients are from the stratum  $y = 0$ . Within these strata, the empirical distributions of the covariates  $X1$ ,  $X3$ ,  $X4$ ,  $X5$  are represented below (relative frequencies in percents):

$X1$ : account	$y$	
	1	0
no account	45.0	19.9
good	15.3	49.7
medium	39.7	30.2

<i>X3: duration in months</i>		<i>y</i>	
	1		0
$\leq 6$	3.00	10.43	
$6 < \dots \leq 12$	22.33	30.00	
$12 < \dots \leq 18$	18.67	18.71	
$18 < \dots \leq 24$	22.00	22.57	
$24 < \dots \leq 30$	6.33	5.43	
$30 < \dots \leq 36$	12.67	6.84	
$36 < \dots \leq 42$	1.67	1.17	
$42 < \dots \leq 48$	10.67	3.14	
$48 < \dots \leq 54$	.33	.14	
$54 < \dots \leq 60$	2.33	1.00	

<i>X4: amount of credit in DM</i>		<i>y</i>	
	1		0
$\leq 500$	1.00	2.14	
$500 < \dots \leq 1000$	11.33	9.14	
$1000 < \dots \leq 1500$	17.00	19.86	
$1500 < \dots \leq 2500$	19.67	24.57	
$2500 < \dots \leq 5000$	25.00	28.57	
$5000 < \dots \leq 7500$	11.33	9.71	
$7500 < \dots \leq 10000$	6.67	3.71	
$10000 < \dots \leq 15000$	7.00	2.00	
$15000 < \dots \leq 20000$	1.00	.29	

<i>X5: previous credit</i>		<i>y</i>	
	1		0
good	82.33	94.85	
bad	17.66	5.15	

<i>X6: intended use</i>		<i>y</i>	
	1		0
private	57.53	69.29	
professional	42.47	30.71	

Looking at these empirical distributions, we see that variable  $X1$  is distinctly different for  $y = 1$  and 0, while variable  $X4$  does not differ very much for the two strata. Indeed it was found in a variable selection procedure (Example 4.1) that variable  $X4$  has no significant linear influence on  $Y$ ; the same is true for  $X7$ . Both variables have therefore been neglected in the following analysis.

The probability of being “not creditworthy” is assumed to follow the logit model

$$\pi = P(y = 1|x) = \frac{\exp(z'\beta)}{1 + \exp(z'\beta)},$$

with design vector  $z' = (1, X1[1], X1[2], X3, X5, X6, X8)$ . The dummy  $X1[1]$  stands for “no running account,”  $X1[2]$  for “good running account.” All qualitative covariates are (0–1)-coded with reference categories as described above. The maximum likelihood estimates of the covariate effects  $\beta$  are given as follows:

Covariate	1	$X1[1]$	$X1[2]$	$X3$	$X5$	$X6$	$X8$
Effect	0.026	0.617	-1.32	0.039	-0.988	-0.47	-0.533

Before interpreting the results, we remark the following: Although the data come from a sample stratified with respect to the binary response, covariate effects are estimated correctly; see Anderson (1972). However, instead of the intercept term  $\beta_0$ , one estimates  $\beta_0 + \log p(1)/N(1) - \log p(0)/N(0)$ , where  $N(1) = 300$ ,  $N(2) = 700$  are the (fixed) sample sizes in the stratified sample, while  $p(1)$ ,  $p(0)$  are the corresponding prior probabilities in the population. If the latter are known or can be estimated, the intercept term can be adjusted to obtain a consistent estimate of  $\beta_0$ .

First let us consider the effect of the metrical covariate “duration of credit” ( $X3$ ). Due to the positive effect of  $X3$ , an increasing credit duration increases the probability of being “not creditworthy.” The effect of the (0–1)-coded qualitative covariates on the creditability may be interpreted with respect to the chosen reference category. As an example consider the effect of “running account” ( $X1$ ). For persons with “no running account” ( $X1[1]$ ) the probability of being “not creditworthy” is higher than for those with “medium running account” (effect is 0 by definition) and for those with “good running account” ( $X1[2]$ ). Vice versa, for persons with “good running account” the probability of being “creditworthy” is higher than for those with “medium running account” and “no running account.” Furthermore, persons “living alone” ( $X8$ , effect is 0 by definition) are less “creditworthy” in probability than others, and persons who intend the “private use” ( $X6$ ) of the credit are also more “creditworthy” in probability than those who intend to use the credit in a professional way (effect is 0 by definition).  $\square$

## Overdispersion

A phenomenon often observed in applications is that the actual variance of binary data is larger than explained by the nominal variance of the (scaled) binomial model. This phenomenon is called “overdispersion” or “extra binomial variation.” There are a number of causes why overdispersion may be observed. Two main reasons are *unobserved heterogeneity* not taken into account by covariates in the linear predictor, and *positive correlation between individual binary responses*, e.g., when individual units belong to a cluster, as, e.g., members of a household. In the latter situation, the simple formula for the sum of independent binary variables leading to the binomial variance is no longer valid since positive correlations also contribute to the variance of the sum or the average. This type of overdispersion can only be caused if the local sample size  $n_i > 1$ . Unobserved heterogeneity, not modelled by the linear predictor, also leads to positive correlation (see Chapter 7), so that the same effect is observed. The simplest way to account for overdispersion is to introduce a multiplicative *overdispersion* parameter  $\phi > 1$  and to assume that

$$\text{var}(y_i|x_i) = \phi \frac{\pi_i(1 - \pi_i)}{n_i}.$$

Since only  $\pi_i = E(y_i|x_i)$  and  $\text{var}(y_i|x_i)$  are actually needed in the estimation of effects  $\beta$  and of  $\phi$ , statistical inference may be formally carried out as if  $\phi$  were the nuisance parameter occurring in an exponential family distribution, e.g., as in the gamma distribution; see the next section. However, the introduction of a multiplicative overdispersion parameter leads in fact already to the simplest form of a quasi-likelihood model (Section 2.3.1): Although there may exist distributions with variance  $\phi\pi_i(1 - \pi_i)/n_i$ , such as the beta-binomial model (see, e.g., Collett, 1991, Chapter 6), a genuine likelihood is not actually needed for the purpose of inference. For reasons of robustness it may be preferable to specify only the mean and the variance function. Note that simply allowing  $\phi \neq 1$  in the exponential family form of the binomial distribution in Table 2.1 will not give a normed density function.

More complex but explicit approaches to deal with overdispersion are contained in Section 3.5, where correlated binary responses are considered, and in Chapter 7, where unobserved heterogeneity is modelled by introducing random effects in the linear predictor. A nice discussion of overdispersion is given in Collett (1991, Chapter 6). Overviews and case studies are found in Hinde & Démétrio (1998), Poortema (1999), and Liang & McCullagh (1993); tests for overdispersion are described, e.g., in Dean (1992) and in Cameron & Trivedi (1998).

### 2.1.5 Models for Count Data

Count data appear in many applications, e.g., as the number of certain events within a fixed period of time (insurance claims, accidents, deaths, births, etc.) or as the frequencies in each cell of a contingency table. Under certain circumstances, such data may be approximately modelled by models for normal data, or, if only some small values  $0, 1, \dots, q$  are observed, by models for multicategorical data (Chapter 3). Generally the Poisson distribution or some modification should be the first choice.

#### Log-linear Poisson Model

This model relates  $\mu$  and  $\eta$  by the natural link

$$\log(\mu) = \eta = z'\beta, \quad \mu = \exp(\eta).$$

The dependence of  $\mu$  on the design vector is multiplicative, which is a sensible assumption in many applications. If all covariates are categorical and appropriate interaction effects are included in  $z$ , this leads to log-linear modelling of frequencies in higher-dimensional contingency tables. We do not discuss these models here, but refer the reader to Bishop, Fienberg & Holland (1975), McCullagh & Nelder (1989), Agresti (1990), and Chapter 10 in Fahrmeir, Hamerle & Tutz (1996).

#### Linear Poisson Model

The direct relation

$$\mu = z'\beta$$

may be useful if covariates are assumed to be additive. Since  $z'\beta$  has to be nonnegative for all  $z$ , restrictions on  $\beta$  are implied.

If  $y$  is exactly Poisson-distributed, its variance equals its expectation:

$$\text{var}(y|x) = \mu.$$

For similar reasons as for binomial data, overdispersion can be present in count data. For count data, as a rule, a nuisance parameter should be included so that

$$\text{var}(y|x) = \sigma(\mu) = \phi\mu.$$

More complex distributions for count data, which allow for various deviations like over- or underdispersion or “excess zeros” from the Poisson model, are available. The negative binomial and truncated Poisson distributions are members of the exponential family, but others like zero inflated Poisson or hurdle models are not. Early references are Breslow (1984) and

Cameron & Trivedi (1986). Two recent reference books with an emphasis on econometrics are Winkelmann (1997) and Cameron & Trivedi (1998).

### Example 2.3: Cellular differentiation

In a biomedical study of the immuno-activating ability of two agents, TNF (tumor necrosis factor) and IFN (interferon), to induce cell differentiation, the number of cells that exhibited markers of differentiation after exposure to TNF and/or IFN was recorded. At each of the 16 dose combinations of TNF/INF, 200 cells were examined. The number  $y$  of cells differentiating in one trial and the corresponding dose levels of the two factors are given in Table 2.3, which is reproduced from Piegorsch, Weinberg & Margolin (1988).

**Table 2.3.** Cellular differentiation data

Number $y$ of cells differentiating	Dose of TNF(U/ml)	Dose of IFN(U/ml)
11	0	0
18	0	4
20	0	20
39	0	100
22	1	0
38	1	4
52	1	20
69	1	100
31	10	0
68	10	4
69	10	20
128	10	100
102	100	0
171	100	4
180	100	20
193	100	100

An important scientific question is whether the two agents stimulate cell differentiation synergistically or independently. Example 2.6 treats this problem within a log-linear Poisson model of the form

$$\mu = E(y|TNF, IFN) = \exp(\beta_0 + \beta_1 TNF + \beta_2 IFN + \beta_3 TNF * IFN),$$

where  $E(y|TNF, IFN)$  denotes the expected number of cells differentiating after exposure to TNF and IFN. The synergistic effect between TNF and IFN is represented by the influence of the two-factor interaction  $TNF * IFN$ . The following estimates were obtained:

$$\hat{\beta}_0 = 3.436, \quad \hat{\beta}_1 = 0.016, \quad \hat{\beta}_2 = 0.009, \quad \hat{\beta}_3 = -0.001.$$

Therefore, it seems doubtful whether there is synergistic effect. A more refined analysis that relaxes the assumption of Poisson-distributed counts  $y$  is given in Example 2.7.  $\square$

## 2.2 Likelihood Inference

Regression analysis with generalized linear models is based on likelihoods. This section contains the basic inferential tools for parameter estimation, hypothesis testing, and goodness-of-fit tests, whereas more detailed material on model choice and model checking is deferred to Chapter 4. The methods rely on the genuine method of maximum likelihood, i.e., it is assumed that the model is completely and correctly specified in the sense of the definition in Section 2.1.2. In many applications, this assumption may be too idealistic. Quasi-likelihood models, where only the mean and the variance function are to be specified, are considered in Section 2.3.1.

### 2.2.1 Maximum Likelihood Estimation

Given the sample  $y_1, \dots, y_i, \dots$ , together with the covariates  $x_1, \dots, x_i, \dots$ , or design vectors  $z_1, \dots, z_i, \dots$ , a maximum likelihood estimator (MLE) of the unknown parameter vector  $\beta$  in the model  $E(y_i|x_i) = \mu_i = h(z'_i\beta)$  is obtained by maximizing the likelihood. To treat the cases of individual data ( $i = 1, \dots, n$ ) and of grouped data ( $i = 1, \dots, g$ ) simultaneously, we omit  $n$  or  $g$  as the upper limit in summation signs. Thus, sums may run over  $i$  from 1 to  $n$  or from 1 to  $g$ , and weights  $\omega_i$  have to be set equal to 1 for individual data and equal to  $n_i$  for grouped data.

We first assume that the *scale* parameter  $\phi$  is known. Since  $\phi$  appears as a factor in the likelihood, we may set  $\phi = 1$  in this case without loss of generality if we are only interested in a point estimate of  $\beta$ . Note, however, that  $\phi$  (or a consistent estimate) is needed for computing variances of the MLE. Consistent estimation of an unknown  $\phi$  by a method of moments, which is carried out in a subsequent step, is described at the end of this subsection. The parameter  $\phi$  may also be considered as an *overdispersion* parameter, and it may be treated *formally* in the same way as a scale parameter. (Note, however, that only the mean  $\mu_i = h(z'_i\beta)$  and the variance function are then

properly defined, so that one has to start with the expression for the score function  $s(\beta)$  instead of the log-likelihood  $l(\beta)$  in (2.2.1)).

To avoid additional complexities concerning parameter identifiability, it is assumed from now on that the “grand” design matrix

$$Z = (z_1, \dots, z_i, \dots)' \quad \text{has (full) rank } p,$$

or, equivalently,

$$\sum_i z_i z_i' = Z' Z \quad \text{has rank } p.$$

### Log-likelihood, Score Function and Information Matrix

According to (2.1.4), the log-likelihood contribution of observation  $y_i$  is, up to an additive constant,

$$l_i(\theta_i) = \log f(y_i | \theta_i, \phi, \omega_i) = \frac{y_i \theta_i - b(\theta_i)}{\phi} \omega_i.$$

The function  $c(y_i, \phi, \omega_i)$ , which does not contain  $\theta_i$ , has been omitted. After inserting the relation  $\theta_i = \theta(\mu_i)$  between the natural parameter and the mean, as given in Table 2.1 and Appendix A.1, this contribution becomes a function of  $\mu_i$ :

$$l_i(\mu_i) = \frac{y_i \theta(\mu_i) - b(\theta(\mu_i))}{\phi} \omega_i,$$

using  $l$  as a generic symbol for log-likelihoods. For example, in the case of binary responses ( $\mu_i = \pi_i$ ), one obtains the well-known form

$$l_i(\pi_i) = y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i).$$

For (scaled) binomial responses, i.e., relative frequencies  $\bar{y}_i$  and repetition number  $n_i$ , we have

$$l_i(\pi_i) = n_i (\bar{y}_i \log \pi_i + (1 - \bar{y}_i) \log(1 - \pi_i)).$$

In the case of Poisson responses ( $\mu_i = \lambda_i$ ), one has

$$l_i(\lambda_i) = y_i \log \lambda_i - \lambda_i.$$

Inserting the mean structure  $\mu_i = h(z_i' \beta)$  finally provides

$$l_i(\beta) = l_i(h(z_i' \beta))$$

as a function of  $\beta$ . Since  $y_1, \dots, y_i, \dots$  are observations of independent random variables, the log-likelihood of the sample is the sum of the individual contributions:

$$l(\beta) = \sum_i l_i(\beta). \quad (2.2.1)$$

Its first derivative is the  $p$ -dimensional score function

$$s(\beta) = \frac{\partial l}{\partial \beta} = \sum_i s_i(\beta).$$

The individual score function contributions are

$$s_i(\beta) = z_i D_i(\beta) \sigma_i^{-2}(\beta) [y_i - \mu_i(\beta)], \quad (2.2.2)$$

where

$$\mu_i(\beta) = h(z'_i \beta), \quad \sigma_i^2(\beta) = v(h(z'_i \beta)) \phi / \omega_i,$$

and

$$D_i(\beta) = \frac{\partial h(z'_i \beta)}{\partial \eta} \quad (2.2.3)$$

is the first derivative of the response function  $h(\eta)$  evaluated at  $\eta_i = z'_i \beta$ . The parameter  $\phi$  may be interpreted here as a scale parameter in the likelihood or as an overdispersion factor for the variance function.

The expected Fisher information matrix is

$$F(\beta) = \text{cov } s(\beta) = \sum_i F_i(\beta), \quad (2.2.4)$$

with

$$F_i(\beta) = z_i z'_i w_i(\beta)$$

and the weight functions

$$w_i(\beta) = D_i^2(\beta) \sigma_i^{-2}(\beta). \quad (2.2.5)$$

The observed information matrix is the matrix

$$F_{obs}(\beta) = -\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'}$$

of negative second derivatives. Its explicit form will not be needed in the sequel but is given in Appendix A.1. Observed and expected information matrices are related by

$$F(\beta) = E(F_{obs}(\beta)).$$

For natural link functions, score functions and Fisher information matrices simplify to

$$s(\beta) = \frac{1}{\phi} \sum_i \omega_i z_i [y_i - \mu_i(\beta)], \quad F(\beta) = \frac{1}{\phi} \sum_i \omega_i v(\mu_i(\beta)) z_i z_i'.$$

Moreover, expected and observed Fisher information are identical,

$$F(\beta) = F_{obs}(\beta).$$

For some purposes, matrix notation is convenient. For the more general grouped data case, one has

$$\begin{aligned} y &= \begin{bmatrix} y_1 \\ \vdots \\ y_g \end{bmatrix}, \quad \mu(\beta) = \begin{bmatrix} \mu_1(\beta) \\ \vdots \\ \mu_g(\beta) \end{bmatrix}, \quad \Sigma(\beta) = \begin{bmatrix} \sigma_1^2(\beta) & & 0 \\ & \ddots & \\ 0 & & \sigma_g^2(\beta) \end{bmatrix}, \\ D(\beta) &= \begin{bmatrix} D_1(\beta) & & 0 \\ & \ddots & \\ 0 & & D_g(\beta) \end{bmatrix}, \\ W(\beta) &= \begin{bmatrix} w_1(\beta) & & 0 \\ & \ddots & \\ 0 & & w_g(\beta) \end{bmatrix}, \end{aligned}$$

and one obtains

$$s(\beta) = Z' D(\beta) \Sigma^{-1}(\beta) [y - \mu(\beta)], \quad F(\beta) = Z' W(\beta) Z.$$

For canonical link functions one obtains

$$s(\beta) = \frac{1}{\phi} Z' \Omega [y - \mu(\beta)]; \quad F(\beta) = \frac{1}{\phi} Z' \Omega V(\beta) Z,$$

with  $\Omega = \text{diag}(\omega_i)$ ,  $V(\beta) = \text{diag}(v(\mu_i))$ .

### Numerical Computation of the MLE by Iterative Methods

Generally, MLEs  $\hat{\beta}$  are not computed as global maximizers of  $l(\beta)$ , but as solutions of the likelihood equations

$$s(\hat{\beta}) = 0, \tag{2.2.6}$$

which correspond to local maxima, i.e., with  $F_{obs}(\hat{\beta})$  positive definite. For many important models, however, the log-likelihood  $l(\beta)$  is concave so that local and global maxima coincide. For strictly concave log-likelihoods, the

MLE is even unique whenever it exists. Existence means that there is at least one  $\hat{\beta}$  within the admissible parameter set  $B$  such that  $l(\hat{\beta})$  is a global or local maximum. Some additional information on the important questions of existence and uniqueness is given later.

The likelihood equations are in general nonlinear and have to be solved iteratively. The most widely used iteration scheme is Fisher scoring or iteratively weighted least-squares. Starting with an initial estimate  $\hat{\beta}^{(0)}$ , Fisher scoring iterations are defined by

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + F^{-1}(\hat{\beta}^{(k)}) s(\hat{\beta}^{(k)}), \quad k = 0, 1, 2, \dots$$

Note that the dispersion parameter  $\phi$  cancels out in the term  $F^{-1}(\hat{\beta}^{(k)}) \cdot s(\hat{\beta}^{(k)})$ . As a simple initial estimate  $\hat{\beta}^{(0)}$ , one may compute the unweighted least squares estimate for the data set  $(g(y_i), z_i)$ ,  $i = 1, \dots, n$ , thereby slightly modifying observations  $y_i$  for which the link function is undefined (e.g.,  $g(y_i) = \log(y_i)$  for  $y_i = 0$ ). Iterations are stopped if some termination criterion is reached, e.g., if

$$\frac{\|\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}\|}{\|\hat{\beta}^{(k)}\|} \leq \varepsilon$$

for some prechosen small number  $\varepsilon > 0$ .

If one defines the “working observation vector”

$$\begin{aligned}\tilde{y}(\beta) &= (\tilde{y}_1(\beta), \dots, \tilde{y}_n(\beta))', \\ \tilde{y}_i(\beta) &= z'_i \beta + D_i^{-1}(\beta) [y_i - \mu_i(\beta)],\end{aligned}$$

then the Fisher scoring iterations may be expressed by

$$\hat{\beta}^{(k+1)} = (Z' W^{(k)} Z)^{-1} Z' W^{(k)} \tilde{y}^{(k)},$$

where  $W^{(k)}$ ,  $\tilde{y}^{(k)}$  means evaluation of  $W$  and  $\tilde{y}$  at  $\beta = \hat{\beta}^{(k)}$ . This form can be interpreted as iteratively weighted least squares, and it has the advantage that a number of results in linear and nonlinear least squares estimation can be used after appropriate modifications.

Of course, other iterative schemes may be applied to solve the likelihood equations. The Newton-Raphson scheme is obtained from Fisher scoring if expected information  $F(\beta)$  is replaced by observed information  $F_{obs}(\beta)$ . However,  $F(\beta)$  is easier to evaluate and always positive semidefinite. Quasi-Newton methods are often better alternatives than the simple Newton-Raphson scheme.

In defining the scoring iterations, we have tacitly assumed that  $F(\beta)$  is nonsingular, i.e., positive definite, for the sequence of iterates  $\hat{\beta}^{(k)}$ . Since full rank of  $Z'Z$  is assumed, this is the case if (most of) the weights  $w_i(\beta)$  are positive. In this case, iterations will usually stop after a few iterations

near a maximum. If they diverge, i.e., if successive differences  $\|\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}\|$  increase, this can indicate a bad initial estimate or, more often, nonexistence of an MLE within the admissible parameter set  $B$ . If  $B = \mathbb{R}^p$ , this means that at least one component of  $\hat{\beta}^{(k)}$  tends to infinity. In the following, some results for discrete response models are discussed more formally.

### Uniqueness and Existence of MLEs\*

The questions of whether MLEs exist, whether they lie in the interior of the parameter space, and whether they are unique have been treated by various authors. Results of Haberman (1974, log-linear and binomial models), Wedderburn (1976, normal, Poisson, gamma, and binomial models), Silvapulle (1981) and Kaufmann (1988, binomial and multicategorical models) are based on concavity of the log-likelihood. We restrict discussion to binomial and Poisson models; for other models we refer to Wedderburn (1976) and see also Fahrmeir & Kredler (1984).

Consider the general binary model (2.1.7), with distribution function  $F$  as a response function. If  $F$  is a continuous distribution function on the real line without constancy intervals, then, without further restrictions, the admissible parameter set is  $B = \mathbb{R}^p$ . In this case existence means that there is a finite  $\hat{\beta}$  where  $l(\beta)$  attains its maximum. Furthermore, let  $F$  be such that  $\log F$  and  $\log(1 - F)$  are strictly concave. This is fulfilled, e.g., for the logit, probit, and double exponential models. Then, for full rank of  $Z$ , existence and uniqueness are equivalent to the condition that the equality/inequality system

$$y_i z'_i \beta \geq 0, \quad (1 - y_i) z'_i \beta \leq 0, \quad \text{for all } i,$$

has only the trivial solution  $\beta = 0$ . Though easy to formulate, this condition can be difficult to check in practice. Note that according to our convention,  $y_i$  is the relative frequency of ones observed for the design value  $z_i$ . A sufficient but rather strong condition is that the matrix

$$\sum_i y_i (1 - y_i) z_i z'_i = Z' \text{ diag}(y_i (1 - y_i))$$

has full rank. This condition cannot be fulfilled for purely binary responses.

If  $F$  is a continuous distribution with  $0 < F(x) < 1$  only for a subinterval  $(a, b)$  of the real line, then the admissible set

$$B = \{\beta : a \leq z'_i \beta \leq b \text{ for all } i\}$$

is restricted but still convex. Existence now means that there is a  $\hat{\beta} \in B$  maximizing  $l(\beta)$ . If  $\log F$ , resp.  $\log(1 - F)$ , is strictly concave on  $(a, b]$ , resp.  $[a, b)$ , then a unique MLE always exists if  $Z$  has full rank. The most prominent model of this type is the linear probability one.

For the log-linear Poisson model, the admissible parameter space is  $B = \mathbb{R}^p$ . A finite and unique MLE exists if the matrix

$$\sum_i y_i z_i z_i' \quad \text{has full rank } p. \quad (2.2.7)$$

For the linear Poisson model, the admissible parameter space is given by

$$B = \{\beta : z_i' \beta \geq 0 \text{ for all } i\},$$

and a nonnegative MLE exists if  $Z$  has full rank. It is unique if and only if (2.2.7) holds. Let us conclude with some general remarks: If  $Z$  has full rank,  $F(\beta) = F_{obs}(\beta)$  is always positive definite for models with canonical link or response function. Conditions for existence are often difficult to check. In practice, it may be easier to start ML iterations to see whether divergence or convergence occurs. Nonexistence may be overcome by larger sample sizes, since the asymptotic theory guarantees asymptotic existence, or by Bayes estimation with a strictly concave informative prior for  $\beta$ ; see Section 2.3.2.

## Asymptotic Properties

Inferential methods for GLMs rely on asymptotic properties of ML estimators. Under “regularity assumptions,” discussed informally later, the following properties hold.

### *Asymptotic existence and uniqueness:*

The probability that  $\hat{\beta}$  exists and is (locally) unique tends to 1 for  $n \rightarrow \infty$ .

### *Consistency:*

If  $\beta$  denotes the “true” value, then for  $n \rightarrow \infty$  we have  $\hat{\beta} \rightarrow \beta$  in probability (weak consistency) or with probability 1 (strong consistency).

### *Asymptotic normality:*

The distribution of the (normed) MLE becomes normal for  $n \rightarrow \infty$ , or, somewhat more informally, for large  $n$

$$\hat{\beta} \stackrel{a}{\sim} N(\beta, F^{-1}(\beta)),$$

i.e.,  $\hat{\beta}$  is approximately normal with approximate (or “asymptotic”) covariance matrix

$$\text{cov}(\hat{\beta}) \stackrel{a}{=} A(\hat{\beta}) = F^{-1}(\hat{\beta}),$$

where  $A(\beta) := F^{-1}(\beta)$  is the inverse of the Fisher matrix. For an unknown scale parameter  $\phi$ , all results remain valid if it is replaced by a consistent estimate  $\hat{\phi}$ .

Furthermore, the MLE is asymptotically efficient compared to a wide class of other estimators.

Heuristic arguments for consistency and asymptotic normality of the MLE are as follows: Since  $Ey_i = \mu_i(\beta)$  for the “true”  $\beta$ , it follows from (2.2.2) that  $Es_i(\beta) = 0$  and

$$E s(\beta) = \sum_{i=1}^n E s_i(\beta) = 0.$$

Therefore, by some law of large numbers,  $s(\beta)/n \rightarrow 0$  in probability. Since  $s(\hat{\beta})/n = 0$  holds for the MLE  $\beta$ , one obtains  $\hat{\beta}_n \rightarrow \beta$  in probability, that is, (weak) consistency, by continuity arguments.

Using  $Es(\beta) = 0$ , we can easily derive

$$\text{cov } s(\beta) = E [s(\beta)s(\beta)'] = F(\beta),$$

with the expected Fisher information  $F(\beta)$  given as in (2.2.4). Applying a central limit theorem to the sum  $s(\beta) = \sum s_i(\beta)$ , we obtain approximate normality

$$s(\beta) \xrightarrow{a} N(0, F(\beta))$$

of the score function for large  $n$ .

A first-order Taylor expansion of  $s(\hat{\beta}) = 0$  about  $\beta$  gives

$$0 = s(\hat{\beta}) \xrightarrow{a} s(\beta) + H(\beta)(\hat{\beta} - \beta),$$

where  $H(\beta) = -\partial^2 l(\beta)/\partial\beta\partial\beta' = F_{obs}(\beta)$ . Replacing the observed information matrix by its expectation  $F(\beta)$ , we get

$$s(\beta) \xrightarrow{a} F(\beta)(\hat{\beta} - \beta)$$

and

$$\hat{\beta} - \beta \xrightarrow{a} F^{-1}(\beta)s(\beta).$$

Approximate normality

$$\hat{\beta} - \beta \xrightarrow{a} N(0, F^{-1}(\beta))$$

of the MLE follows from approximate normality of  $s(\beta)$ , with approximate covariance matrix  $A(\beta) = F^{-1}(\beta)F(\beta)F^{-1}(\beta) = F^{-1}(\beta)$ . Replacing  $\beta$  by its consistent estimate  $\hat{\beta}$  gives  $A(\hat{\beta}) = F^{-1}(\hat{\beta})$ .

Under appropriate regularity assumptions, these heuristic arguments are the basis for rigorous proof; see Appendix A.2.

### Discussion of Regularity Assumptions\*

With respect to the underlying regularity assumptions and the complexity of proofs, one can distinguish three types of asymptotics:

(i) Asymptotic theory for grouped data assumes a fixed number  $g$  of groups and  $n_i \rightarrow \infty$ ,  $i = 1, \dots, g$ , such that  $n_i/n \rightarrow \lambda_i$  for fixed “proportions”  $\lambda_i > 0$ ,  $i = 1, \dots, n$ . For applications with finite sample size, this means that there is a sufficient number of repeated observations for each design vector  $z_i$ . This assumption will normally be violated in the presence of continuous covariates.

(ii) Standard asymptotic theory for ungrouped data requires only that the total sample size  $n$  tends to infinity, but as a typical regularity assumption it is required that

$$\frac{F(\beta)}{n} \quad \text{has a positive definite limit,} \quad (2.2.8)$$

together with additional moment conditions. This type of asymptotic analysis is standard in the sense that it is rather near the case of i.i.d. observations. Therefore, results of this type are often stated without proof under “mild regularity conditions.” However, the convergence condition (2.2.8) induces convergence conditions on the covariates itself. It is commonly fulfilled for stochastic regressors in the population case, where  $(y_i, x_i)$  are i.i.d. drawings from the joint distribution of  $(y, x)$ . In planned experiments (2.2.8) can be fulfilled if regressors are similarly scattered as stochastic regressors. However, (2.2.8) is typically violated for trending or growing regressors.

(iii) General asymptotic theory requires only divergence of  $F(\beta)$ , i.e.,

$$A(\beta) = F^{-1}(\beta) \rightarrow 0 \quad (2.2.9)$$

together with additional continuity properties of  $F(\beta)$ , or conditions on the moments of responses, on the sequence of covariates, etc. Condition (2.2.9) seems to be an indispensable requirement. It guarantees that information in the data increases with the sample size. Such general results are given in Haberman (1977, for natural link functions) and in Fahrmeir & Kaufmann (1985). Proofs require additional effort, in particular reducing general conditions on  $F(\beta)$  to conditions on the sequence of covariates in certain models. As a “sample” of more specific conditions, some results of Fahrmeir & Kaufmann (1986) for the logit and probit models, as well as the linear and log-linear Poisson models are reviewed.

For bounded regressors, i.e.,  $\|z_n\| < c$  for all  $n$ , divergence of  $Z'Z$  or, equivalently,

$$(Z'Z)^{-1} \rightarrow 0 \quad \text{for } n \rightarrow \infty, \quad (2.2.10)$$

implies all the asymptotic properties. In the classical linear model, condition (2.2.10) alone is necessary and sufficient for (weak and strong) consistency

and for asymptotic normality. Although bounded regressors cover a large number of situations, there are applications where growing regressors will be of interest, e.g., to model certain trends in a longitudinal analysis or in dose response experiments. Under a slight sharpening of (2.2.9), it can be shown that

$$\|z_n\| = O(\log n) \quad \text{for the logit model and the log-linear Poisson model,}$$

$$\|z_n\| = O(\log n)^{\frac{1}{2}} \quad \text{for the probit model,}$$

$$\|z_n\| = O(n^\alpha) \quad \text{for the linear Poisson model, with some } \alpha > 0,$$

are sharp upper bounds for the admissible growth of regressors, ensuring asymptotic properties. Compared to linear models, where, e.g., exponential growth is admissible, growth of regressors is much more restricted, e.g., to sublogarithmic growth in the logit model.

### Additional Scale or Overdispersion Parameter

If the scale or an overdispersion parameter  $\phi$  is unknown, it can be consistently estimated by

$$\hat{\phi} = \frac{1}{g-p} \sum_{i=1}^g \frac{(\hat{y}_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)/n_i}, \quad (2.2.11)$$

with  $\hat{\mu}_i = h(z'_i \hat{\beta})$  and  $v(\hat{\mu}_i)/n_i$  as the estimated expectation and variance function of  $y_i$ , after grouping the data as far as possible. In all expressions where  $\phi$  occurs, e.g., in  $F(\hat{\beta})$ , it is replaced by its estimate  $\hat{\phi}$  to obtain correct standard errors.

Note that for linear Gaussian regression the moment estimate (2.2.11) reduces to the well-known estimate for  $\sigma^2$  from the sum of squared residuals.

### 2.2.2 Hypothesis Testing and Goodness-of-Fit Statistics

Most of the testing problems for  $\beta$  are linear hypotheses of the form

$$H_0 : C\beta = \xi \quad \text{against} \quad H_1 : C\beta \neq \xi, \quad (2.2.12)$$

where the matrix  $C$  has full row rank  $s \leq p$ . An important special case is

$$H_0 : \beta_r = 0 \quad \text{against} \quad H_1 : \beta_r \neq 0, \quad (2.2.13)$$

where  $\beta_r$  is a subvector of  $\beta$ . This corresponds to testing the submodel defined by  $\beta_r = 0$  against the full model. In the following it is assumed that

unknown scale or overdispersion parameters  $\phi$  are replaced by consistent estimates.

The *likelihood ratio statistic*

$$\lambda = -2\{l(\tilde{\beta}) - l(\hat{\beta})\}$$

compares the unrestricted maximum  $l(\hat{\beta})$  of the (log-)likelihood with the maximum  $l(\tilde{\beta})$  obtained for the restricted MLE  $\tilde{\beta}$ , computed under the restriction  $C\beta = \xi$  of  $H_0$ . If the unrestricted maximum  $l(\hat{\beta})$  is significantly larger than  $l(\tilde{\beta})$ , implying that  $\lambda$  is large,  $H_0$  will be rejected in favor of  $H_1$ . A likelihood ratio test of (2.2.13), i.e., testing of a submodel defined by  $\beta_r = 0$ , requires new scoring iterations for the submodel, whereas considerable more effort is required to estimate  $\tilde{\beta}$  under the general  $H_0$  of (2.2.12).

In case of an unknown scale parameter  $\phi$ , all results remain valid if it is replaced by a consistent estimate  $\hat{\phi}$ . Note that the likelihood ratio statistic is not properly defined for overdispersion models where the distributional assumptions are not fully given. As an approximation, however, it is common to work with the usual log-likelihoods for binomial or Poisson models, additionally divided by the estimate  $\hat{\phi}$  from (2.2.11), which one obtains from the larger model or, if several nested models are compared, from some maximal model containing both models under consideration.

The Wald test and the score test are computationally attractive quadratic approximations of the likelihood ratio statistic. The *Wald statistic*

$$w = (C\hat{\beta} - \xi)' [C F^{-1}(\hat{\beta}) C']^{-1} (C\hat{\beta} - \xi)$$

uses the weighted distance between the unrestricted estimate  $C\hat{\beta}$  of  $C\beta$  and its hypothetical value  $\xi$  under  $H_0$ . The weight is determined by the inverse of the asymptotic covariance matrix  $CF^{-1}(\hat{\beta})C'$  of  $C\hat{\beta}$ . The Wald statistic is useful if the unrestricted MLE has already been computed, as, e.g., in subset selection procedures. The *score statistic*

$$u = s'(\tilde{\beta}) F^{-1}(\tilde{\beta}) s(\tilde{\beta})$$

is based on the following idea: The score function  $s(\beta)$  for the unrestricted model is the zero vector if it is evaluated at the unrestricted MLE  $\hat{\beta}$ . If  $\hat{\beta}$  is replaced by the MLE  $\tilde{\beta}$  under  $H_0$ ,  $s(\tilde{\beta})$  will be significantly different from zero if  $H_0$  is not true. The distance between  $s(\tilde{\beta})$  and zero is measured by the score statistic  $u$ , with the inverse information matrix  $F^{-1}(\tilde{\beta})$  acting as a weight. The score test is of advantage if a restricted model has already been fitted and is to be tested against a more complex model, as, e.g., in a forward selection procedure. No new scoring iterations for computing the MLE  $\hat{\beta}$  of the larger model are required.

An advantage of the Wald and score statistics is that they are properly defined for models with overdispersion since only first and second moments are involved.

Let  $A(\beta) = F^{-1}(\beta)$  denote the inverse information matrix. For the special hypothesis (2.2.13), the statistics reduce to

$$w = \hat{\beta}'_r \hat{A}_r^{-1} \hat{\beta}_r \quad (2.2.14)$$

and

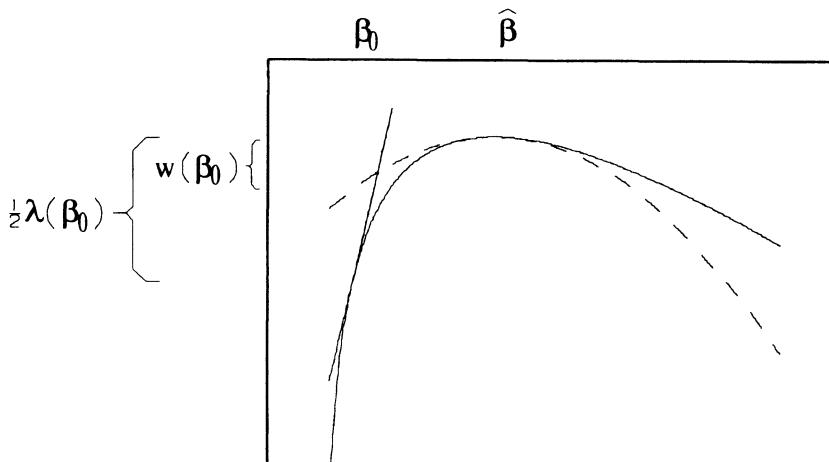
$$u = \tilde{s}'_r \tilde{A}_r \tilde{s}_r,$$

where  $A_r$  is the submatrix of  $A = F^{-1}$  corresponding to the elements of  $\beta_r$ ,  $s_r$  is the corresponding subvector of  $s$ , and “ $\hat{\phantom{x}}$ ” or “ $\tilde{\phantom{x}}$ ” means evaluation at  $\beta = \hat{\beta}$  or  $\beta = \tilde{\beta}$ .

For the special case where  $\beta_r$  consists of only a scalar component of  $\beta$ , the Wald statistic is the square of the “ $t$ -value”

$$t_r = \frac{\hat{\beta}_r}{\sqrt{\hat{a}_{rr}}},$$

the standardized estimate of  $\beta_r$ , with  $\hat{a}_{rr} = \text{var}(\hat{\beta}_r)$  the  $r$ th diagonal element of the (estimated) asymptotic covariance matrix  $A(\hat{\beta}) = F^{-1}(\hat{\beta})$  of  $\hat{\beta}$ .



**Figure 2.4.** Log-likelihood (—) and quadratic approximation (---) for Wald test and slope for score test.

Under  $H_0$  the three test statistics are asymptotically equivalent and have the same limiting  $\chi^2$ -distribution with  $s$  degrees of freedom,

$$\lambda, w, u \xrightarrow{a} \chi^2(s), \quad (2.2.15)$$

under similar general conditions, which ensure asymptotic properties of the MLE (Fahrmeir, 1987a). Critical values or  $p$ -values for the testing procedure

are determined according to this limiting  $\chi^2$ -distribution. Informally, the approximate  $\chi^2$ -distribution of the score statistic  $u$  and the Wald statistic  $w$  follows directly from approximate normality of the score function  $s$  and the MLE  $\beta$ . By a Taylor expansion of  $l(\beta)$  about  $\hat{\beta}$ , see Appendix (A.2),  $w$  and  $s$  are quadratic approximations of  $\lambda$ , so that  $\lambda$  is also approximately  $\chi^2$ -distributed. In particular  $p$ -values corresponding to the squared  $t$ -values  $t_r^2$  of effects  $\beta_r$ , are computed from the  $\chi^2(1)$  distribution.

For finite sample size  $n$ , the quality of approximation of the distribution of  $\lambda$ ,  $\omega$ ,  $u$  to the limiting  $\chi^2(s)$ -distribution depends on  $n$  and on the form of the log-likelihood function. As an example consider the linear Poisson model  $\mu = \beta$  for a single observation  $y$  that takes the value 3. The quality of approximation can be seen from Figure 2.4 where the special case of testing  $H_0 : \beta = \beta_0$  against  $H_1 : \beta \neq \beta_0$  is treated. In this case  $\tilde{\beta} = \beta_0$ . Furthermore,  $\hat{\beta} = 3$  denotes the MLE. The likelihood ratio statistic takes the vertical log-likelihood distance  $l(\hat{\beta}) - l(\beta_0)$  as a measure of evidence for or against  $H_1$ . The Wald statistic is based on the quadratic log-likelihood approximation  $\tilde{l}$ , which is obtained by a second-order Taylor series expansion around  $\hat{\beta}$ . It takes into account the distance  $\tilde{l}(\hat{\beta}) - \tilde{l}(\beta_0)$ . The larger the horizontal distance  $(\hat{\beta} - \beta_0)$  the larger is the discrepancy between the likelihood ratio and Wald statistic as long as the log-likelihood is nonquadratic. The score statistic looks at the (squared) slope of  $l(\beta)$  at  $\beta_0$ , weighted by the inverse of the curvature, which should be near to zero if  $\beta_0$  is near to  $\hat{\beta}$ . If the log-likelihood is a quadratic function of  $\beta$ , then all three test statistics coincide. For large sample size  $n$ , asymptotic theory shows that  $l(\beta)$  becomes approximately quadratic, so that  $\lambda$ ,  $w$  and  $u$  will tend to be close to each other. For medium or small  $n$ , however, differences can become more serious. A more detailed discussion is presented in Buse (1982).

The hypotheses considered so far form linear subspaces. In some applications, other hypotheses may be of interest, e.g., inequality restrictions. Already in the classical linear normal model the null distribution of the LR statistic is then a mixture of  $\chi^2$ -distributions (e.g., Gourieroux, Holly & Montfort, 1982; Wolak, 1987; Kaufmann, 1989). Following Wolak (1989), such results should carry over to generalized linear models under appropriate regularity assumptions.

## Goodness-of-Fit Statistics

Two summary statistics often used to assess the adequacy of a model are the *Pearson statistic*

$$\chi^2 = \sum_{i=1}^g \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}$$

and the *deviance*

$$D = -2\phi \sum_{i=1}^g \{l_i(\hat{\mu}_i) - l_i(y_i)\},$$

where  $\hat{\mu}_i$ ,  $v(\hat{\mu}_i)$  are the estimated mean and variance function, and  $l_i(y_i)$  is the individual log-likelihood where  $\mu_i$  is replaced by  $y_i$  (the maximum likelihood achievable). For both cases data should be grouped as far as possible. If grouped data asymptotics apply, i.e., if the number of observations is sufficiently large in all groups, both statistics are approximately  $\phi\chi^2(g-p)$ -distributed, where  $p$  is the number of estimated coefficients. Then the statistics may be used for formally testing the goodness-of-fit of a model. However, if  $n$  is large and  $n_i$  remains small, in particular  $n_i = 1$ , formal use of these test statistics becomes dangerous; see McCullagh & Nelder (1989, pp. 118–119). Large values of  $\chi^2$  or  $D$  cannot necessarily provide evidence for lack of fit.

Model checking has to be supported by additional formal and informal diagnostic tools; see Chapter 4.

**Example 2.4: Caesarian birth study** (Example 2.1, continued)

In Example 2.1, a logit model was fitted for the risk of infection. In addition to the point estimates  $\hat{\beta}$  of effects, already given and interpreted there, estimated standard deviations and  $t$ -values are given in Table 2.4.

**Table 2.4.** Logit model fit to Caesarian birth study data

	$\hat{\beta}$	$\sqrt{\text{var}(\hat{\beta})}$	$t$
1	-1.89	0.41	-4.61
NOPLAN	1.07	0.43	2.49
FACTOR	2.03	0.46	4.41
ANTIB	-3.25	0.48	-6.77

For the deviance, the value 10.997 was obtained, with 3 degrees of freedom. The model is rejected for  $\alpha = 5\%$ , indicating a bad fit. Let us check the model by comparing observed relative frequencies to fitted probabilities (in parentheses):

	Caesarian planned	Not planned
Antibiotics		
Risk factors	0.06 (0.04)	0.11 (0.11)
No risk factors	0.00 (0.01)	
No antibiotics		
Risk factors	0.48 (0.53)	0.88 (0.77)
No risk factors	0.20 (0.13)	0.00 (0.30)

This shows that the fit is extremely bad in the group defined by Caesarian not planned/no antibiotics/no risk. Let us try to improve the fit by including interaction terms. The last column of the following table shows the difference of deviance to the main effect model if the interaction term NOPLAN\*ANTIB is included.

	Deviance	df	Difference to main effect model
Main effect model	10.997	3	-
Main effects + NOPLAN*ANTIB	10.918	2	0.07839

This indicates that this interaction may be omitted. A problem occurs for the inclusion of the interactions FACTOR\*ANTIB and NOPLAN\*FACTOR. Some program packages yield estimates for this model, with very large values for the NOPLAN and NOPLAN\*FACTOR effects, but without any warning. Actually, in the first print of this book such misleading estimates were given (thanks to J.-L. Fouley for his hints). In fact, no finite ML estimate exists, i.e., parameter estimates for ANTIB and FACTOR\*ANTIB as well as for NOPLAN and NOPLAN\*FACTOR go to infinity. The reason is that data are too sparse with responses for the cell NOPLAN = 0, FACTOR = 0 and ANTIB = 1 all in the "yes" category, and for the cell NOPLAN = 1, FACTOR = 0 and ANTIB = 0 all in the "no" category. So this is a typical example where the log-likelihood is concave but has no finite global maximum; see the discussion on the uniqueness and existence of MLEs. Instead of using robustified estimates, a common and simple method to deal with such a situation is to add a fictitious datum to these cells, with 0.5 as response category and 0.5 in the other one. Doing this, the deviance for the model with inclusion of the interaction term FACTOR\*ANTIB is nonsignificant. In contrast, the interaction term NOPLAN\*FACTOR is significant. Estimated effects and standard deviations are given in the table below.

Covariate	1	NOPLAN	FACTOR	ANTIB	NOPLAN*FACTOR
$\hat{\beta}$	-1.39	-1.56	1.36	-3.83	3.41
$\sqrt{\text{var}(\hat{\beta})}$	0.40	1.50	0.47	0.60	1.61

The deviance for this model is 0.955, with 2 degrees of freedom, providing strong evidence for including the interaction NOPLAN\*FACTOR. Although some parameter estimates change considerably, one can easily see that there is only little change in the linear predictor for most of the covariate combinations. For example, the contribution of NOPLAN and FACTOR to the linear predictor in the main effect model was 3.10 for NOPLAN = 1 and FACTOR = 1, whereas it is 3.21 in the model including the interaction NOPLAN\*FACTOR. However, there is a distinct change for NOPLAN = 1 and FACTOR = 0, leading to improved fitted probabilities. Note that omission of NOPLAN, suggested by the standard deviation, would impair both fit and deviance.  $\square$

	Caesarian planned		Not planned	
Antibiotics				
Risk factors	0.06	(0.02)		0.11 (0.12)
No risk factors	0.00	(0.01)		
No antibiotics				
Risk factors	0.48	(0.49)		0.88 (0.86)
No risk factors	0.20	(0.20)		0.05 (0.05)

### Example 2.5: Credit-scoring (Example 2.2, continued)

Consider the logit model of Example 2.2, where the response variable was the consumers' creditability. Table 2.5 gives the MLEs  $\hat{\beta}_r$ , the standard errors  $\sqrt{\text{var}(\hat{\beta}_r)}$ , the standardized MLEs  $t_r = \hat{\beta}_r / \sqrt{\text{var}(\hat{\beta}_r)}$ , and their  $p$ -values for each component  $\beta_r$  of the parameter vector  $\beta$ . With a 0.05 significance level, the hypothesis  $H_0 : \beta_r = 0$  is rejected for all parameters with the exception of the grand mean "1."

In addition, the logit model yields deviance 1017.35 and  $\chi^2 = 1006.53$  with 991 degrees of freedom. Assuming an approximate  $\chi^2(991)$ -distribution, the model is not rejected, as can be seen from the  $p$ -values 0.387 (deviance) and 0.277 ( $\chi^2$ ). However, the assumption of a  $\chi^2(991)$ -distribution may be dangerous, even approximately, since the fit is based on binary observations that cannot be grouped because of the metrical covariate X3.  $\square$

**Table 2.5.** Logit model fit to credit-scoring data

	$\hat{\beta}$	$\sqrt{\text{var}(\hat{\beta})}$	$t$	$p\text{-value}$
1	0.026	0.316	0.082	0.933
X1[1]	0.617	0.176	3.513	0.0
X1[2]	-1.320	0.202	-6.527	0.0
X3	0.039	0.006	6.174	0.0
X5	-0.988	0.253	-3.910	0.0
X6	-0.470	0.160	-2.940	0.003
X8	-0.533	0.160	-3.347	0.001

**Example 2.6: Cellular differentiation** (Example 2.3, continued)

Recall the log-linear model of Example 2.3, where the logarithm of the expected number  $y$  of cells differentiating depends linearly on the main effects TNF, IFN, and on the interaction between these two factors. The first column of Table 2.6 contains the MLEs  $\hat{\beta}$ , which are based on a Poisson likelihood with nuisance-parameter  $\phi$  set equal to 1. Concerning the effect of the interaction between TNF and IFN, the  $p$ -value (in brackets) would suggest a high significance. However, deviance 142.4 and  $\chi^2 = 140.8$  at 12 degrees of freedom indicate a high level of extravariation or overdispersion that is not explained by the fitted Poisson model. Since the counts  $y$  are rather large, the asymptotic  $\chi^2(12)$ -distribution of deviance and  $\chi^2$  seem to be justified, so that the fitted Poisson model has to be rejected.

**Table 2.6.** Log-linear model fits to cellular differentiation data based on Poisson-likelihoods

	Poisson, $\phi = 1$	Poisson, $\hat{\phi} = 11.734$
1	3.436 (.0)	3.436 (.0)
TNF	.016 (.0)	.016 (.0)
IFN	.009 (.0)	.009 (.0)
TNF*IFN	-.001 (.0)	-.001 (.22)

To take into account overdispersion, the nuisance parameter  $\phi$  has to be estimated by (2.2.11). Since the likelihood equations (2.2.6) do not depend on  $\phi$  the estimated nuisance-parameter  $\hat{\phi} = 11.734$  does not affect the MLEs so that  $\hat{\beta}$  is the same for  $\phi$  set to 1 and  $\hat{\phi} = 11.734$ . The asymptotic variance-covariance matrix  $\text{cov}(\hat{\beta}) = F^{-1}(\hat{\beta})$ , however, depends on  $\phi$  and has to

be corrected with  $\hat{\phi} = 11.734$ . Due to this correction, the Wald statistics  $\hat{\beta}_r / \sqrt{\text{var}(\hat{\beta}_r)}$ ,  $r = 0, \dots, 3$ , as well as the  $p$ -values (given in brackets) change. In contrast to the fit where  $\phi$  was set to 1 (column 1), the interaction effect is no longer significant. This result is also obtained by using a quasi-likelihood approach (see Example 2.7).  $\square$

## 2.3 Some Extensions

The original class of generalized linear models and related techniques for statistical inference have been modified and extended in several ways, further enhancing its flexibility and potential in applications. In this section we first describe two approaches for univariate cross-sectional models that also play an important role in later chapters where responses are correlated (Section 3.5 and Chapters 6, 7, and 8): quasi-likelihood and Bayes models. A further generalization, nonlinear or nonexponential family models, is addressed briefly in Section 2.3.3.

### 2.3.1 Quasi-likelihood Models

One of the basic assumptions in the definition of generalized linear models is that the true density of the responses belongs to a specific exponential family, e.g., normal, binomial, Poisson, gamma, etc. Apart from the normal family, choice of the mean structure  $\mu = h(z'\beta)$  implies a certain variance structure  $v(\mu) = v(h(z'\beta))$ . For example, in a linear Poisson model  $\mu = z'\beta$  implies  $v(\mu) = \mu = z'\beta$ . If this is not consistent with the variation of the data, it was proposed to introduce an additional overdispersion parameter to account for extravariation, so that  $\text{var}(y) = \phi\mu = \phi z'\beta$ . The resulting score function, with typical contribution  $z(\phi z'\beta)^{-1}(y - z'\beta)$ , is no longer the first derivative from a Poisson likelihood but rather from some “quasi-likelihood.” Quasi-likelihood models drop the exponential family assumption and separate the mean and variance structure. No full distributional assumptions are necessary; only first and second moments have to be specified. Under appropriate conditions, parameters can be estimated consistently, and asymptotic inference is still possible under appropriate modifications.

#### Basic Models

Wedderburn (1974), McCullagh (1983), and McCullagh & Nelder (1989) assume that the mean and variance structure are correctly specified by

$$E(y|x) = \mu = h(z'\beta), \quad \text{var}(y|x) = \sigma^2(\mu) = \phi v(\mu), \quad (2.3.1)$$

where  $v(\mu)$  is a variance function, generally defined separately and without reference to some exponential family, and  $\phi$  is the dispersion parameter. Then a “quasi-likelihood”  $Q(\beta, \phi)$  or an extended version (Nelder & Pregibon, 1987; McCullagh & Nelder, 1989) can be constructed such that  $\partial Q/\partial\beta$  has the form (2.2.2) of a score function, with  $\sigma^2(\mu) = \phi v(\mu)$  given by (2.3.1). For the original definition of quasi-likelihood, an equivalent genuine likelihood exists if there is a simple exponential family with the same variance function (Morris, 1982). In a closely related approach, Gourieroux, Monfort & Trognon (1984) assume only a correctly specified mean structure. Estimation is based on a “pseudo”-exponential family model that need not contain the true distribution. Consequently, the true variance function will be different from the variance function appearing in the score function corresponding to the quasi-model.

In the following, both approaches are considered in a unifying way. We suppose that the *mean is correctly specified* by  $\mu = h(z'\beta)$  as in (2.3.1), but the *true variance*

$$\text{var}(y|x) = \sigma_0^2(x)$$

*may be different* from  $\sigma^2(\mu) = \phi v(\mu)$  in (2.3.1), which is used as a “*working*” variance function. The basic assumption of independent responses is maintained in this section. Estimation is based on the *quasi-score function* or *generalized estimating function* (GEE)

$$s(\beta) = \sum_i z_i D_i(\beta) \sigma_i^{-2}(\beta) [y_i - \mu_i(\beta)], \quad (2.3.2)$$

where  $\mu_i(\beta) = h(z'_i\beta)$  is the correctly specified mean, and  $D_i(\beta)$  is the first derivative of  $h$  evaluated at  $\eta_i = z'_i\beta$  as in Section 2.2.1. However,  $\sigma_i^2(\beta) = \phi v(\mu_i(\beta))$  is now a “*working*” variance: The variance function  $v(\mu)$  may in principle be specified freely, as in the approach of Wedderburn (1976) and others, or it is implied by a “*working*” pseudo-likelihood model  $l(\beta)$ , as in Gourieroux, Monfort & Trognon (1984). For reasons of efficiency, the working variance should be close to the true variance, which means in accordance with the variability of the data. However, consistent estimation of  $\beta$  is possible with any choice. Therefore, specification of  $v(\mu)$  will be a compromise between simplicity and loss of efficiency.

A global Q(uasi)MLE would be a global maximizer of an associated quasi-likelihood  $l(\beta)$ . As in the case of correctly specified generalized linear models, we consider only local QMLEs, i.e., roots of the quasi-score function where the matrix  $\partial s(\beta)/\partial\beta'$  is negative definite. Global and local QMLEs may be different, but for many models of interest they coincide.

Negative first derivatives  $-\partial s(\beta)/\partial\beta'$  have the same form as  $-\partial^2 l(\beta)/\partial\beta\partial\beta'$  in Section 2.2.1 resp. Appendix A.1. Moreover,

$$F(\beta) = E\left(-\frac{\partial s(\beta)}{\partial\beta'}\right) = \sum_i z_i z'_i w_i(\beta), \quad (2.3.3)$$

where the weights are given as in (2.2.5), with  $\sigma_i^2(\beta)$  as the working variances. In (2.3.3) and in the following,  $E$ , cov, var, etc., are to be understood with respect to the true but incompletely or incorrectly specified data-generating probability mechanism. Equation (2.3.3) follows from Appendix A.1, since  $E(y_i) - \mu_i(\beta) = 0$ . However, the expected quasi-information  $F(\beta)$  is generally different from  $\text{cov } s(\beta)$ , so that (2.2.4) will not hold. In fact,

$$V(\beta) = \text{cov } s(\beta) = \sum_i z_i z'_i D_i^2(\beta) \frac{\sigma_{0i}^2}{\sigma_i^4(\beta)}, \quad (2.3.4)$$

where  $\sigma_{0i}^2 = \sigma_0^2(x_i)$  is the true (conditional) variance  $\text{var}(y_i|x_i)$ . Equation (2.3.4) follows from (2.3.2) using  $Es(\beta) = 0$ ,  $E((y_i - \mu_i(\beta))^2|x) = \text{var}(y_i|x)$ , and the assumption of independent observations. Comparing (2.3.3) and (2.3.4), it is seen that generally  $F(\beta) \neq V(\beta)$ . However, if the variance structure is correctly specified, i.e., if

$$\sigma_0^2(x) = \phi v(\mu), \quad \mu = h(z'\beta),$$

as Wedderburn and others assume, then  $F(\beta) = \text{cov}(s(\beta)) = V(\beta)$  holds again.

Asymptotic properties of QMLEs can be obtained under appropriate regularity conditions, similar to that in Section 2.2.1. For standard asymptotic theory we refer to some of the work cited earlier; nonstandard results can be obtained along the lines of Fahrmeir (1990). Under appropriate assumptions, a (local) QMLE  $\hat{\beta}$  as a root of the estimating equation (2.3.2) exists asymptotically and is consistent and asymptotically normal,

$$\hat{\beta} \xrightarrow{a} N(\beta, \hat{F}^{-1} \hat{V} \hat{F}^{-1}),$$

with estimates  $\hat{F} = F(\hat{\beta})$  and

$$\hat{V} = \sum_i z_i z'_i D_i^2(\hat{\beta}) \frac{[y_i - h(z'_i \hat{\beta})]^2}{\sigma_i^4(\hat{\beta})}$$

for  $V(\beta)$  and  $F(\beta)$ .

Compared with the corresponding result for completely specified models, essentially only the asymptotic covariance matrix  $\hat{\text{cov}}(\hat{\beta})$  has to be corrected to the “sandwich”-matrix  $\hat{A} = \hat{F}^{-1} \hat{V} \hat{F}^{-1}$ . Thus, the quasi-likelihood approach allows consistent and asymptotically normal estimation of  $\beta$  under quite weak assumptions, with some loss of efficiency, however, due to the corrected asymptotic covariance matrix. To keep this loss small, the working variance structure should be as close as possible to the true variance structure.

Informally, approximate normality of the QMLE  $\hat{\beta}$  can be shown with only a slight modification compared to the arguments for the MLE. Since

now  $\text{cov } s(\beta) = V(\beta)$ , we have  $s(\beta) \stackrel{a}{\sim} N(0, V(\beta))$ . With the same Taylor expansion as for the MLE, we get  $\hat{\beta} - \beta \stackrel{a}{\sim} F^{-1}(\beta)s(\beta)$ . This implies

$$\hat{\beta} \stackrel{a}{\sim} N(\beta, A(\beta)),$$

with the sandwich matrix  $A(\beta) = F^{-1}(\beta)V(\beta)F^{-1}(\beta)$ . Replacing  $F$  and  $V$  by their estimates gives the result.

Tests of linear hypotheses of the form

$$H_0 : C\beta = \xi \quad \text{against} \quad H_1 : C\beta \neq \xi$$

are still possible by appropriately modified Wald and score statistics. For example, the *modified Wald statistic* is

$$w_m = (C\hat{\beta} - \xi)' [C\hat{A}C']^{-1} (C\hat{\beta} - \xi),$$

with the corrected covariance matrix  $\hat{A} = \hat{F}^{-1}\hat{V}\hat{F}^{-1}$  instead of  $\hat{F}$  as in the common Wald statistic  $w$ . It has a limiting  $\chi^2$ -distribution with  $r = \text{rank}(C)$  degrees of freedom. Results of this kind are contained in White (1982) for the i.i.d. setting and they carry over without major difficulties to the present framework. In contrast to testing in completely specified models, “quasi-likelihood ratio test statistics” are in general not asymptotically equivalent to modified Wald or score statistics and may not have a limiting  $\chi^2$ -distribution (see, e.g., Foutz & Srivastava, 1977, in the i.i.d. setting).

## Variance Functions with Unknown Parameters

Until now the variance function was assumed to be a known function  $v(\mu)$  of the mean. Several authors (e.g., Nelder & Pregibon, 1987) relaxed this requirement by allowing unknown parameters  $\theta$  in the variance function, so that

$$\text{var}(y|x) = \phi v(\mu; \theta).$$

For example, a useful parameterized variance function is obtained by considering powers of  $\mu$ :

$$v(\mu; \theta) = \mu^\theta.$$

The values  $\theta = 0, 1, 2, 3$  correspond to the variance functions of the normal, Poisson, gamma, and inverse Gaussian distributions, respectively.

For fixed  $\theta$ , a QMLE  $\hat{\beta}$  (and an estimate  $\hat{\phi}$ ) can be obtained as before. Estimation of  $\theta$ , given  $\beta$  and  $\phi$ , can, e.g., be carried out by some method of moments. Cycling between the two steps until convergence gives a joint estimation procedure. Asymptotic results for  $\hat{\beta}$  remain valid if  $\theta$  is replaced by a consistent estimate  $\hat{\theta}$ .

### Nonconstant Dispersion Parameter

A further extension is quasi-likelihood models where both the mean  $\mu$  and the dispersion parameter are modelled as functions of separate linear predictors (Pregibon, 1984; Nelder & Pregibon, 1987; Efron, 1986; McCullagh & Nelder, 1989, Ch. 10; Nelder, 1992):

$$\mu = h(z'\beta), \quad \phi = \phi(w'\theta), \quad \text{var}(y) = \phi v(\mu).$$

Here  $w$  is a vector of covariates affecting the dispersion parameter. A two-step estimating procedure is proposed by cycling between the generalized estimating equation (GEE) for  $\beta$ , holding the current iterate  $\hat{\phi} = \phi(w'\hat{\theta})$  fixed, and a second generalized estimating equation for  $\theta$ , holding  $\hat{\beta}$  and  $\hat{\mu} = h(z'\hat{\beta})$  fixed. This second GEE is obtained by differentiating Nelder and Pregibon's extended quasi-likelihood. Alternatively, joint estimation of  $\beta$  and  $\theta$  is, in principle, possible by the general techniques for fitting likelihood and quasi-likelihood models described by Gay & Welsch (1988); see also Section 2.3.3. Consistent estimation of  $\beta$  and  $\theta$  requires that not only the mean but also the dispersion parameter are correctly specified.

#### **Example 2.7: Cellular differentiation** (Examples 2.3, 2.6, continued)

In Example 2.3 a log-linear model was proposed to analyze the synergistic effect of TNF and IFN on  $\mu$ , the expected number  $y$  of cells differentiating after exposure to TNF and/or IFN. Example 2.6 gives the estimation results, which are based on Poisson-likelihood fits involving the variance structure  $\sigma^2(\mu) = \text{var}(y|\text{TNF}, \text{IFN}) = \phi\mu$ . However, a comparison of the sample mean and sample variance for each group of counts having the same dose level of TNF reveals that a variance structure proportional to  $\mu$  is less adequate:

	Dose of TNF			
	0	1	10	100
$\bar{x}$	22	45.25	74	161.5
$s^2$	107.5	300.7	1206.5	1241.25

The same holds for IFN. The sample moments lead one to suppose that the variance structure has the form  $\sigma^2(\mu) = \text{var}(y|\text{TNF}, \text{IFN}) = \phi\mu^2$  or  $\mu + \theta\mu^2$ , where the second one corresponds to the variance of the negative binomial distribution. QMLEs  $\hat{\beta}$ , which are based on the log-linear mean structure proposed in Example 2.3 and on three alternative variance structures  $\sigma^2(\mu)$ , are given together with their  $p$ -values (in brackets) and moment estimators for the dispersion parameter  $\phi$  or  $\theta$  in Table 2.7. Parameter estimates and  $p$ -values for the variance assumptions being quadratic in the mean  $\mu$  are nearly identical and do not differ very much from the results, which are

**Table 2.7.** Log-linear model fits to cellular differentiation data based on quasi-likelihoods

	$\sigma^2(\mu) = \phi\mu$	$\sigma^2(\mu) = \phi\mu^2$	$\sigma^2(\mu) = \mu + \theta\mu^2$
1	3.436 (0.0)	3.394 (0.0)	3.395 (0.0)
TNF	0.016 (0.0)	0.016 (0.0)	0.016 (0.0)
IFN	0.009 (0.0)	0.009 (0.003)	0.009 (0.003)
TFN*IFN	-0.001 (0.22)	-0.001 (0.099)	-0.001 (0.099)
$\hat{\phi}$	11.734	0.243	—
$\hat{\theta}$	—	—	0.215

based on a variance depending linearly on  $\mu$ . However, in contrast to the fit, which is based on a Poisson model with unknown nuisance-parameter (see the second column of Table 2.6), the  $p$ -value of the interaction TNF\*IFN only moderately supports the presence of a synergistic effect between TNF and IFN. Moreover, the Poisson assumption seems to be less appropriate due to the discrepancy between estimated means and variances.  $\square$

### 2.3.2 Bayesian Models

GLMs were initially introduced and, so far, have been described in this chapter from a classical viewpoint. Bayesian methods for analyzing GLMs are of more recent origin and have rapidly grown in the past decade. This is mainly due to a breakthrough in methods to compute the required posteriors such as simulation based techniques using Markov chain Monte Carlo (MCMC, see Appendix A.5), and due to the flexibility of Bayesian hierarchical modelling in extending GLMs to more complex problems, such as semiparametric GLMs (Ch. 5), generalized random effects and mixed models (Ch. 7), dynamic or state-space approaches to non-normal time series, longitudinal data and time-space processes (Ch. 8), and survival or event history data (Ch. 9). This section gives a short review of Bayesian estimation of GLMs.

Many recent results on Bayesian GLMs and their extensions are brought together in the reference book edited by Dey, Gosh & Mallick (1999), and the chapter by Gelfand & Ghosh (1999) extends and complements the following presentation.

Bayesian models assume that  $\beta$  is a random vector with prior density  $p(\beta)$ , and Bayes estimators are based on the posterior density  $p(\beta|Y = (y_1, \dots, y_n))$  of  $\beta$  given the data. Bayes' theorem relates prior and posterior densities by

$$p(\beta|Y) = \frac{L(\beta|Y) p(\beta)}{\int L(\beta|Y) p(\beta) d\beta}, \quad (2.3.5)$$

where  $L(\beta|Y) = p(Y|\beta)$  is the likelihood of the data. Marginal posterior densities for components of  $\beta$ , posterior means, and covariance matrices, etc., can be obtained from the posterior by integration (resp., summation in the discrete case). For example,

$$E(\beta|Y) = \int \beta p(\beta|Y) d\beta \quad (2.3.6)$$

is the *posterior mean*, which is an optimal estimator of  $\beta$  under quadratic loss, and

$$\text{cov}(\beta|Y) = \int (\beta - E(\beta|Y)) (\beta - E(\beta|Y))' p(\beta|Y) d\beta \quad (2.3.7)$$

is the associated *posterior covariance matrix*, which is a measure for the precision of the posterior mean estimate. So, at first glance, the Bayesian paradigm seems to be easily implemented. However, exact analytic solutions of the preceding integrations are available only for some special models, e.g., for the normal linear model. For most of the important models, e.g., binomial logit models with at least one covariate, no conjugate priors that would allow convenient analytic treatment exist. Therefore, direct implementation of the Bayesian estimation approach via (2.3.6) and (2.3.7) requires numerical or Monte Carlo integration. Since the integrals have the dimension of  $\beta$ , which may be high-dimensional, this is not a trivial task. A number of techniques have been proposed and discussed. We refer the reader to Naylor & Smith (1982), Zellner & Rossi (1984), Smith, Skene, Shaw, Naylor & Dransfield (1985), and West & Harrison (1989, Ch. 13) for methods such as Gauss-Hermite integration or Monte Carlo integration. However, application of these methods is limited to models with parameter vectors  $\beta$  of low dimension, often less than about 5. At least for higher dimensions and as building blocks in more complex models, MCMC simulation techniques described further below are now in common use.

*Posterior mode estimation* is an alternative to full posterior analysis or posterior mean estimation, which avoids numerical integrations or simulation methods. It has been proposed by a number of authors, e.g., Leonard (1972), Laird (1978), Stiratelli, Laird & Ware (1984), Zellner & Rossi (1984), Duffy & Santner (1989), and Santner & Duffy (1989). The posterior mode estimator  $\hat{\beta}_p$  maximizes the posterior density  $p$  or equivalently the log posterior likelihood

$$l_p(\beta|Y) = l(\beta) + \log p(\beta), \quad (2.3.8)$$

where  $l(\beta)$  is the log-likelihood of the generalized linear model under consideration. If a normal prior

$$\beta \sim N(\alpha, Q), \quad Q > 0,$$

is chosen, (2.3.8) specializes to

$$l_p(\beta|Y) = l(\beta) - \frac{1}{2}(\beta - \alpha)'Q^{-1}(\beta - \alpha), \quad (2.3.9)$$

dropping terms that are constant with respect to  $\beta$ . The criterion (2.3.9) is a penalized likelihood, with the penalty  $(\beta - \alpha)'Q^{-1}(\beta - \alpha)$  for deviations from the prior parameter  $\alpha$ . The addition of such a concave log prior to  $l(\beta)$  also helps to avoid problems of nonexistence and nonuniqueness for ML estimators. In the limiting case  $Q \rightarrow \infty$  or  $Q^{-1} = 0$  of a flat, noninformative prior for  $\beta$ , the penalty term vanishes, and the posterior mode estimator coincides with the MLE. For a normal prior with  $\alpha = 0$  and diagonal  $Q = \tau^2 I$ , the posterior mode estimator takes the form of a ridge estimator with shrinkage parameter  $\lambda = 1/2\tau^2$ . The penalty term reduces to  $\lambda(\beta_1^2 + \dots + \beta_p^2)$ , and the parameter  $\lambda$  or  $\tau^2$  thus regularizes shrinkage of the MLE  $\hat{\beta}$  toward zero. Generalized ridge regression is a remedy in the presence of ill-conditioned information, caused, for example, by near collinearity or high correlations among the covariates; see Marx, Eilers & Smith (1992) for a non-Bayesian treatment. Posterior mode estimation is also closely related to the Laplace approximation method suggested by Tierney & Kadane (1986); see Breslow & Clayton (1993, Section 2.1). For normal priors,  $s(\beta)$  is modified to

$$s_p(\beta) = \frac{\partial l_p(\beta|Y)}{\partial \beta} = s(\beta) - Q^{-1}(\beta - \alpha),$$

and  $F(\beta) = -E(\partial^2 l(\beta|Y)/\partial \beta \partial \beta')$  to

$$F_p(\beta) = -E\left(\frac{\partial^2 l_p(\beta|Y)}{\partial \beta \partial \beta'}\right) = F(\beta) + Q^{-1}.$$

Numerical computation of  $\hat{\beta}_p$  can be carried out by modifying the iteratively weighted least squares form of Fisher scoring in Section 2.2 to

$$\hat{\beta}_p^{(k+1)} = (Z'W^{(k)}Z + Q^{-1})^{-1}(Z'W^{(k)}\tilde{y}^{(k)} + Q^{-1}\alpha).$$

For large  $n$ ,  $\hat{\beta}_p$  becomes approximately normal,

$$\hat{\beta}_p \xrightarrow{a} N(\beta, F_p^{-1}(\hat{\beta}_p)),$$

under essentially the same conditions that ensure asymptotic normality of the MLE. Then the posterior mode and the (expected) curvature  $F_p^{-1}(\hat{\beta}_p)$  of  $l_p(\beta)$ , evaluated at the mode, are good approximations to the posterior mean (2.3.6) and covariance matrix (2.3.7).

Until now, we have tacitly assumed that the prior is completely specified, that is,  $\alpha$  and  $Q$  of the normal prior are known. Empirical Bayes analysis

considers  $\alpha$  and  $Q$  as unknown constants (“hyperparameters”) that have to be estimated from the data. ML estimation by direct maximization of the marginal likelihood and indirect maximization by application of the exact EM algorithm again require numerical integration. This can be avoided if an *EM-type algorithm* is used instead, where posterior means and covariances appearing in the E-step are replaced by posterior modes and curvatures; see, e.g., Santner & Duffy (1989, p. 249) and Duffy & Santner (1989). We will use such EM-type algorithms in Chapters 7 and 8.

*Fully Bayesian inference via MCMC* is based on samples drawn from the posterior and approximating posterior means, variances, etc., by their sampling analog. If we assume a normal prior  $\beta \sim N(\alpha, Q)$ , the posterior  $p(\beta|y)$  is given by

$$p(\beta|y) \propto \exp \left( l(\beta) - \frac{1}{2} (\beta - \alpha)' Q^{-1} (\beta - \alpha) \right).$$

Apart from special cases, there does not exist any closed-form expression for the norming constant. Dellaportas & Smith (1993) show how the Gibbs sampler with adaptive rejection sampling (Gilks & Wild, 1992) can be used to sample from log-concave posteriors. They consider the case of a diagonal  $Q$ , and, as usual with the basic form of Gibbs sampling, each component of  $\beta$  is sampled in turn from its full conditional given the data and the rest of the other parameters. Clayton (1996) describes an extension to non-diagonal  $Q$ .

Updating the entire vector of  $\beta$  can be done in principle by Metropolis-Hastings steps (Appendix A.5) with a random walk proposal  $q(\beta, \beta^*)$ , but a serious problem is tuning, i.e., specifying a suitable covariance matrix for the proposal that guarantees high acceptance rates and good mixing. Especially when the dimension of  $\beta$  is high, with significant correlations among components, tuning “by hand” is no longer feasible. A good alternative is the weighted least squares proposal suggested by Gamerman (1997a). Here a Gaussian proposal is used with mean  $m(\beta)$  and covariance matrix  $C(\beta)$ , where  $\beta$  is the current state of the chain. The mean  $m(\beta)$  is obtained by making one Fisher scoring step as in (2.3.11) to maximize the full conditional  $p(\beta|y)$  and  $C(\beta)$  is the inverse of the expected Fisher information, evaluated at the current state  $\beta$  of the chain. In this case the acceptance probability of a proposed new vector  $\beta^*$  is

$$\min \left\{ 1, \frac{p(\beta^*|\cdot)q(\beta^*, \beta)}{p(\beta|\cdot)q(\beta, \beta^*)} \right\}. \quad (2.3.10)$$

Note that  $q$  is not symmetric, because the covariance matrix  $C$  of  $q$  depends on  $\beta$ . Thus the fraction  $q(\beta^*, \beta)/q(\beta, \beta^*)$  cannot be omitted from the above equation. This algorithm can also be applied for a flat, non-informative prior with  $Q^{-1} = 0$ . Then the posterior equals the likelihood, so that Bayesian analysis is close to a likelihood analysis. As a consequence, of course, the algorithm will fail to converge in this case when the MLE, that is, the posterior mode, does not exist for finite  $\beta$ .

The Bayesian models considered so far are based on the parameters  $\beta$ . Analytically more tractable expressions for the implementation of the estimation approach can be deduced if the Bayes model is based on the means  $\mu_i$  of the data densities (2.1.4). Albert (1988), for example, considers the means  $\mu_i$  to be independent random variables with prior densities  $p(\mu_i)$ , where the prior means  $\nu_i$  of  $\mu_i$  are assumed to satisfy the GLM

$$E(\mu_i|x_i) = \nu_i = h(z'_i\beta). \quad (2.3.11)$$

Model (2.3.11) allows for uncertainty concerning the specification of the means  $\mu_i$  in the original GLM. The precision of the belief in the original GLM is reflected by the prior variance of  $\mu_i$ . If the variance of the prior  $p(\mu_i)$  approaches 0, the prior density becomes concentrated about the mean  $\nu_i$  and the Bayesian model (2.3.11) in this limiting case is equivalent to the original GLM. The larger the prior variance, the more uncertain is the original GLM. That means there are additional sources of variation that cannot be adequately explained by the original GLM.

Estimation of  $\beta$  and other unknown parameters in the prior  $p(\mu_i)$  can be carried out by empirical or hierarchical Bayes procedures. For simplicity we restrict discussion to the case that only  $\beta$  is unknown. In the empirical Bayes context,  $\beta$  is considered an unknown constant (hyperparameter). Estimation is based on the independent marginal densities

$$f(y_i|\beta) = \int f(y_i|\mu_i) p(\mu_i|\beta) d\mu_i, \quad (2.3.12)$$

where  $p(\mu_i|\beta)$  denotes the prior density of  $\mu_i$ , which depends on  $\beta$  via (2.3.11). Fortunately, the prior  $p(\mu_i)$  can be chosen from the conjugate family so that analytic solutions of the integrations (2.3.12) are available. Densities that are conjugate to simple exponential families are described by Cox & Hinkley (1974), among others. For example, the conjugate prior for the Poisson density is the gamma density, which yields a marginal density  $f(y_i|\beta)$  of the negative binomial or Poisson-gamma type. For such closed-form solutions of (2.3.12), estimation of  $\beta$  can be carried out by maximizing the marginal likelihood

$$L(\beta|Y) = \prod_{i=1}^n f(y_i|\beta) \quad (2.3.13)$$

with respect to  $\beta$  or by applying the maximum quasi-likelihood principle (Section 2.3.1), which is based on the first two moments of  $f(y_i|\beta)$ . The latter is of advantage if the marginal density  $f(y_i|\beta)$  does not belong to the simple exponential family class. See, e.g., Williams (1982) for dichotomous data and Breslow (1984) and Lawless (1987) for count data.

Hierarchical Bayes estimation of  $\beta$  has been considered by Leonard & Novick (1986) and Albert (1988); see also Gelfand & Ghosh (1999). In that context a two-stage prior is assigned to the means  $\mu_i$  of the data density. In addition to the prior  $p(\mu_i|\beta)$ , a prior density  $p(\beta)$  is assigned to  $\beta$ . This formulation is called hierarchical because there is a hierarchy of densities; one for the parameters in the data density and one for the parameters in the first-stage prior density. Estimation of  $\beta$  is based on the posterior density (2.3.5), where  $L(\beta|Y)$  is given by (2.3.13), which represents the product of independent mixture densities of the form (2.3.12). However, exact analytic solutions of the integrations in (2.3.5) are available only for some special cases. Albert (1988) suggests approximations as alternatives to direct numerical integrations. It seems, however, that MCMC methods that can deal directly with the parameters  $\beta$  of interest are preferable in the interim.

### 2.3.3 Nonlinear and Nonexponential Family Regression Models\*

In generalized linear (quasi-) models considered so far it was assumed that the predictors are linear in the parameters  $\beta$ . This assumption may be too restrictive and needs to be relaxed for some parts of the later chapters (e.g., in Section 3.5) by a nonlinear predictor, leading to *nonlinear exponential family regression models*. Suppose the mean structure is defined by

$$E(y|x) = h(x;\beta), \quad (2.3.14)$$

where the response function  $h$  has smoothness properties as in the original definition. Stressing dependence on  $\beta$ , we write  $\mu(\beta) := h(x;\beta)$ . Models with common response functions and nonlinear predictors  $\eta(x;\beta)$ ,

$$\mu(\beta) = h(\eta(x;\beta)),$$

are of course covered by the general nonlinear model (2.3.14). Models with composite link functions (Thompson & Baker, 1981) and parametric link functions with linear predictors

$$\mu(\beta;\theta) = h(z'\beta;\theta)$$

(Pregibon, 1980; Scallan, Gilchrist & Green, 1984; Czado, 1992) are also within the general framework.

Given the data  $(y_i, x_i)$ ,  $i = 1, 2, \dots$ , score functions and information matrices are now given by straightforward generalizations of (2.2.2) and (2.2.4):

$$s(\beta) = \sum_i M_i(\beta) \sigma_i^{-2}(\beta) [y_i - \mu_i(\beta)], \quad (2.3.15)$$

$$F(\beta) = \sum_i M_i(\beta) \sigma_i^{-2}(\beta) M'_i(\beta), \quad (2.3.16)$$

where  $\mu_i(\beta) = h(x_i; \beta)$  and  $M_i(\beta) = \partial \mu_i(\beta) / \partial \beta$ .

Defining  $M(\beta) = (M_1(\beta), \dots, M_n(\beta))'$  and  $y$ ,  $\mu(\beta)$ , and  $\Sigma(\beta) = \text{diag}(\sigma_i^2(\beta))$  as in Section 2.2.1, we have

$$s(\beta) = M'(\beta) \Sigma^{-1}(\beta) [y - \mu(\beta)], \quad F(\beta) = M'(\beta) \Sigma^{-1}(\beta) M(\beta)$$

in matrix notation. Fisher-scoring iterations for the MLE  $\hat{\beta}$  can again be formulated as iterative weighted least squares (Green, 1984, 1989). While generalization of asymptotic theory is comparably straightforward under appropriate regularity conditions, questions of finite sample existence and uniqueness are difficult to deal with and no general results are available.

One may even go a step further and drop the exponential family assumption to obtain rather *general parametric regression models* as, e.g., in Green (1984, 1989) or Jorgensen (1992). Estimation for a wide class of (quasi-) likelihood models for independent observations is discussed by Gay & Welsch (1988). They consider objective functions of the form

$$l(\beta, \theta) = \sum_i l_i(\eta(x_i; \beta); \theta),$$

where the (quasi-) likelihood contribution  $l_i$  of observation  $i$  is a function of a nonlinear predictor  $\eta(x_i; \beta)$  and a vector  $\theta$  of nuisance parameters. This framework includes linear and nonlinear exponential family models, robust regression models as described in Huber (1981) and Holland & Welsch (1977), and in the (extended) quasi-likelihood models of Section 2.3.1. Green (1989) and Seeber (1989) show that Fisher scoring can again be written in the form of iteratively weighted least squares and give a geometric interpretation.

A very comprehensive class of non-normal distributions is covered by the class of dispersion models; see Jorgensen (1997). Dispersion models include distributions for statistical modelling of a wide range of data beyond the original definition of generalized linear models, for example, count data with overdispersion, positive continuous responses, nonnegative responses where the outcome zero occurs with positive probability, and many others.

Most of the extensions of GLMs in the following chapters, e.g., to multivariate, nonparametric, or longitudinal data analysis, are written within the exponential family framework and, particularly, with a focus on discrete data. It should be noted, however, that much of the material can be extended to nonexponential family distributions, opening a broad field for current and future research.

## 2.4 Notes and Further Reading

Other topics extending the basic framework of generalized linear models are not described in this chapter or others in this book. Generalized linear models with errors in variables have deserved considerable attention. For a survey, see Carroll, Ruppert & Stefanski (1995) and Carroll, Küchenhoff, Lombard & Stefanski (1996). Principal component and ridge regression are considered by Marx & Smith (1990) and Marx, Eilers & Smith (1992). For bootstrapping in generalized linear models, see, e.g., Moulton & Zeger (1989, 1991), Gigli (1992), Hinde (1992), Shao & Tu (1995), Efron & Tibshirani (1993), and Davison & Hinkley (1997).

# 3

## Models for Multicategorical Responses: Multivariate Extensions of Generalized Linear Models

In this chapter the concept of generalized linear models is extended to the case of a vector-valued response variable. Consider Example 2.1, where we were interested in the effect of risk factors and antibiotics on infection following birth by caesarian section. In this example the response was binary, distinguishing only between occurrence and nonoccurrence of infection, and thereby ignoring that the data originally provided information on the type of infection (type I or II) as well. It is possible, however, to use this information by introducing a response variable with three categories (no infection, infection type I, infection type II). Naturally, these categories cannot be treated as a unidimensional response. We have to introduce a (dummy) variable for each category, thus obtaining a *multivariate* response variable. Therefore, link and response functions for the influence term will be vector-valued functions in this chapter. The focus is on multicategorical response variables and multinomial models. Variables of this type are often called polychotomous, the possible values are called categories. Extension to other multivariate exponential family densities is possible but not considered in this text.

After an introductory section we first consider the case of a nominal response variable (Section 3.2). If the response categories are ordered, the use of this ordering yields more parsimoniously parameterized models. In Section 3.3 several models of this type are discussed. Section 3.4 outlines statistical inference for the multicategorical case. In many applications, more than one response variable is observed, for example, when several measurements are made for each individual or unit or when measurements are

observed repeatedly. Approaches for this type of multivariate responses with correlated components are outlined in Section 3.5.

## 3.1 Multicategorical Response Models

### 3.1.1 Multinomial Distribution

For the categorical responses considered in this chapter, the basic distribution is the multinomial distribution. Let the response variable  $Y$  have  $k$  possible values, which for simplicity are labeled  $1, \dots, k$ . Sometimes consideration of  $Y \in \{1, \dots, k\}$  hides the fact that we actually have a *multivariate* response variable. This becomes obvious by considering the response vector of the dummy variables  $y' = (\tilde{y}_1, \dots, \tilde{y}_q)$ ,  $q = k - 1$ , with components

$$\tilde{y}_r = \begin{cases} 1 & \text{if } Y = r, \quad r = 1, \dots, q, \\ 0 & \text{otherwise.} \end{cases} \quad (3.1.1)$$

Then we have

$$Y = r \Leftrightarrow y = (0, \dots, 1, \dots, 0).$$

The probabilities are simply connected by

$$P(Y = r) = P(y_r = 1).$$

Given  $m$  independent repetitions  $y_1, \dots, y_m$  (or equivalently  $Y_1, \dots, Y_m$ ), it is useful to consider as a response variable the number of trials where we get outcome  $r$ . For the repetitions  $(y_1, \dots, y_m)$ , we get the sum of vectors

$$y = \sum_{i=1}^m \tilde{y}_i.$$

Then the vector  $y$  is multinomially distributed with distribution function

$$\begin{aligned} P(y = (m_1, \dots, m_q)) &= \frac{m!}{m_1! \cdots m_q! (m - m_1 - \dots - m_q)!} \\ &\cdot \pi_1^{m_1} \cdots \pi_q^{m_q} (1 - \pi_1 - \dots - \pi_q)^{m - m_1 - \dots - m_q}, \end{aligned} \quad (3.1.2)$$

where  $\pi_r = P(Y_i = r)$ ,  $i = 1, \dots, m$ . The multinomial distribution of  $y$  is abbreviated by

$$y \sim M(m, \pi), \quad \text{where } \pi' = (\pi_1, \dots, \pi_q).$$

Sometimes it is useful to consider the scaled multinomial distribution  $\bar{y} \sim M(m, \pi)/m$ , where  $\bar{y} = y/m$ , instead of the multinomial  $y \sim M(m, \pi)$ . The mean  $\bar{y}$  is an unbiased estimate of the underlying probability vector  $\pi$  with covariance matrix

$$\begin{aligned}\text{cov}(\bar{y}) &= \frac{1}{m}(\text{diag}(\pi) - \pi\pi') \\ &= \frac{1}{m} \begin{bmatrix} \pi_1(1 - \pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_q \\ -\pi_2\pi_1 & \ddots & & \\ \vdots & & \ddots & \\ -\pi_q\pi_1 & \cdots & \cdots & \pi_q(1 - \pi_q) \end{bmatrix}.\end{aligned}$$

### 3.1.2 Data

The data for regression type problems have the same form as in the case of univariate responses. Let

$$(y_i, x_i), \quad i = 1, \dots, n,$$

denote the observations of a cross section, where  $x'_i = (x_{i1}, \dots, x_{im})$  is the vector of covariates and  $y'_i = (y_{i1}, \dots, y_{iq})$  is the  $q$ -dimensional response vector, e.g., representing dummy variables for categories. Grouped and ungrouped data may be distinguished for the modelling of the conditional response  $y_i$  given  $x_i$ . In analogy to Section 2.1.1, the covariates of ungrouped data are given by a matrix  $X$ . However, observations of the dependent variable now form an  $(n \times q)$ -matrix:

Response variable	Explanatory variables
-------------------	-----------------------

$$\begin{array}{ll} \text{Unit 1} & \left[ \begin{array}{ccc} y_{11} & \cdots & y_{1q} \\ \vdots & & \vdots \end{array} \right] \quad \left[ \begin{array}{ccc} x_{11} & \cdots & x_{1m} \\ \vdots & & \vdots \end{array} \right] \\ \vdots & \\ \text{Unit } n & \left[ \begin{array}{ccc} y_{n1} & \cdots & y_{nq} \end{array} \right] \quad \left[ \begin{array}{ccc} x_{n1} & \cdots & x_{nm} \end{array} \right] \end{array}.$$

If some of the covariates  $x_i$  are identical, the data may be grouped. Let  $y_i^{(1)}, \dots, y_i^{(n_i)}$  denote the  $n_i$  observed responses for fixed covariate  $x_i$ . Then the arithmetic mean

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_i^{(j)}$$

or the sum of responses  $n_i \bar{y}_i$  may be considered a response given fixed covariates  $x_i$ .

The grouped data may be condensed in the form

	Response variable	Explanatory variables
Group 1 ( $n_1$ observations)	$\begin{bmatrix} y_{11} & \cdots & y_{1q} \\ \vdots & & \vdots \\ y_{g1} & \cdots & y_{gq} \end{bmatrix}$	$\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & & \vdots \\ x_{g1} & \cdots & x_{gm} \end{bmatrix},$

where  $x'_i = (x_{i1}, \dots, x_{im})$  now stands for the covariates in the  $i$ th group and  $y'_i = (y_{i1}, \dots, y_{iq})$  stands for the sum or the arithmetic mean of responses given fixed covariates  $x_i$ ,  $i = 1, \dots, g$ ,  $g \leq n$ , where  $g$  is the number of covariates with different values in the data set.

### 3.1.3 The Multivariate Model

Univariate models with response  $y_i$  given  $x_i$  as considered in Chapter 2 have the form

$$\mu_i = h(z'_i \beta).$$

For a dichotomous response variable  $y_i \in \{0, 1\}$ , for example, the logistic model is given by

$$\pi_i = \frac{\exp(z'_i \beta)}{1 + \exp(z'_i \beta)},$$

where  $\pi_i = P(y_i = 1|x_i)$ . In the multinomial case  $\pi_i = \mu_i = E(y_i|x_i)$  is a  $(q \times 1)$ -vector  $\pi'_i = (\pi_{i1}, \dots, \pi_{iq})$  rather than a scalar as earlier. Here the model has the form

$$\pi_i = h(Z_i \beta) \tag{3.1.3}$$

where  $h$  is a vector-valued response function,  $Z_i$  is a  $(q \times p)$ -design matrix composed from  $x_i$ , and  $\beta$  is a  $(p \times 1)$ -vector of unknown parameters. In analogy to the univariate case, the linear influence term will be abbreviated  $\eta_i = Z_i \beta$ .

As an example, consider the widely used multicategorical logit model that is treated more extensively in Section 3.2 and illustrated in Example 3.1. It is given by

$$P(Y_i = r) = \frac{\exp(\beta_{r0} + z'_i \beta_r)}{1 + \sum_{s=1}^q \exp(\beta_{s0} + z'_i \beta_s)}, \tag{3.1.4}$$

which may be written equivalently as

$$\log \frac{P(Y_i = r)}{P(Y_i = k)} = \beta_{r0} + z'_i \beta_r. \quad (3.1.5)$$

Here  $z_i$  is the vector of covariables determining the log odds for category  $r$  with respect to the reference category  $k$ . From the latter form of the model one immediately gets the response function  $h = (h_1, \dots, h_q)$  with

$$h_r(\eta_1, \dots, \eta_q) = \frac{\exp(\eta_r)}{1 + \sum_{s=1}^q \exp(\eta_s)}, \quad r = 1, \dots, q, \quad (3.1.6)$$

the design matrix

$$Z_i = \begin{bmatrix} 1 & z'_i \\ & 1 & z'_i \\ & & \ddots \\ & & & 1 & z'_i \end{bmatrix},$$

and the parameter vector  $\beta' = (\beta_{10}, \beta'_1, \dots, \beta_{q0}, \beta'_q)$ .

An alternative form of model (3.1.3) is given by using the link function  $g$ , which is the inverse of  $h$ , i.e.,  $g = h^{-1}$ . Then the model has the form

$$g(\pi_i) = Z_i \beta.$$

As is immediately seen from (3.1.5), the link function of the logit model is given by  $g = (g_1, \dots, g_q)$ , where

$$g_r(\pi_{i1}, \dots, \pi_{iq}) = \log \frac{\pi_{ir}}{1 - (\pi_{i1} + \dots + \pi_{iq})}.$$

This simple example shows that multivariate models (for multinomial responses) are characterized by two specifications:

- the response function  $h$  (or the link function  $g = h^{-1}$ );
- the design matrix that depends on the covariables and the model.

### Example 3.1: Caesarian birth study

Consider the data on infection following Caesarian birth given in Table 1.1 of Chapter 1 (p. 1). The effect of risk factors and antibiotics on infection has already been examined in Example 2.1 of Chapter 2, though the analysis was simplified there by only distinguishing between occurrence and nonoccurrence of infection. In contrast to Chapter 2, we now want to distinguish between the two different types of infection. Therefore, our response variable  $Y$  has three possible outcomes (infection type I, infection type II, no infection), which are labeled 1 to 3, respectively, thus having  $Y \in \{1, 2, 3\}$ . We introduce a multivariate response vector of dummy variables  $y'_i = (y_{i1}, y_{i2})$  to take into account the categorical character of  $Y$ . Let

$y_{i1} = 1$  if the  $i$ th Caesarian was followed by infection I, and  $y_{i2} = 1$  if it was followed by infection II. Assuming dummy coding, no infection leads to  $y'_i = (0, 0)$ . There are three binary covariates, namely, NOPLAN, FACTOR, and ANTIB, which we assume to be dummy coded with NOPLAN = 1 for “the Caesarian has not been planned,” FACTOR = 1 for “risk factors were present,” ANTIB = 1 for “antibiotics were given as a prophylaxis.” According to Section 3.1.2, the grouped data may be condensed in the form

	$y_{i1}$	$y_{i2}$	Response variable			Explanatory variables		
			NOPLAN	ANTIB	FACTOR	NOPLAN	ANTIB	FACTOR
Group 1	$n_1 = 40$	$\begin{bmatrix} 4 & 4 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$					
Group 2	$n_2 = 58$	$\begin{bmatrix} 11 & 17 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$					
Group 3	$n_3 = 2$	$\begin{bmatrix} 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$					
Group 4	$n_4 = 18$	$\begin{bmatrix} 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$					
Group 5	$n_5 = 9$	$\begin{bmatrix} 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$					
Group 6	$n_6 = 26$	$\begin{bmatrix} 10 & 13 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 1 \end{bmatrix}$					
Group 7	$n_7 = 98$	$\begin{bmatrix} 4 & 7 \end{bmatrix}$	$\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$					

We fit a multicategorical logit model to the data taking “no infection” as the reference category of  $Y$ . According to (3.1.5) the model can be written as

$$\begin{aligned} \log \frac{P(\text{infection type I})}{P(\text{no infection})} &= \beta_{10} + z_i' \beta_1, \\ \log \frac{P(\text{infection type II})}{P(\text{no infection})} &= \beta_{20} + z_i' \beta_2. \end{aligned}$$

Note that the link function  $g$  is now a *vector*-valued function of  $(\pi_{i1}, \pi_{i2})$ :

$$g(\pi_{i1}, \pi_{i2}) = \left( \log \frac{\pi_{i1}}{1 - \pi_{i1} - \pi_{i2}}, \log \frac{\pi_{i2}}{1 - \pi_{i1} - \pi_{i2}} \right),$$

As for the binary model it is useful to look at the contribution of explanatory variables to the odds of an infection. For infection type  $i$ , ( $i=\text{I,II}$ ), the model is given by

$$\log \frac{P(\text{infection type } i)}{P(\text{no infection})} = \beta_{i0} + \beta_{iN} \text{NOPLAN} + \beta_{iF} \text{FACTOR} + \beta_{iA} \text{ANTIB},$$

or equivalently

$$\frac{P(\text{infection type } i)}{P(\text{no infection})} = \exp(\beta_{i0}) \exp(\beta_{iN})^{\text{NOPLAN}} \exp(\beta_{iF})^{\text{FACTOR}} \\ \cdot \exp(\beta_{iA})^{\text{ANTIB}}.$$

Thus, the exponential of the parameter gives the factorial contribution to the odds if the corresponding explanatory variable takes value 1 instead of 0. Alternatively, the exponential of the parameter may be seen as odds ratio between odds for value 1 and odds for value 0, e.g., for NOPLAN one obtains

$$\exp(\beta_{iN}) = \frac{\frac{P(\text{infection type } i|\text{NOPLAN} = 1, F, A)}{P(\text{no infection}|\text{NOPLAN} = 1, F, A)}}{\frac{P(\text{infection type } i|\text{NOPLAN} = 0, F, A)}{P(\text{no infection}|\text{NOPLAN} = 0, F, A)}},$$

where the values of FACTOR and ANTIBIOTICS are not further specified because they may take any fixed value.

The estimates of the parameters and the exponential parameter are given in the following table.

	$\beta$	$\exp(\beta)$		$\beta$	$\exp(\beta)$
constant	-2.621	0.072	constant	-2560	0.077
NOPLAN[1]	1.174	3.235	NOPLAN[2]	0.996	2.707
ANTIB[1]	-3.520	0.030	ANTIB[2]	-3.087	0.046
FACTOR[1]	1.829	6.228	FACTOR[2]	2.195	8.980

According to these estimates, antibiotics given as a prophylaxis seem to decrease the odds for infection type I slightly more than that for type II. However, the presence of risk factors seems to increase the odds for infection type II more than that for type I. While for type I infection the odds increase by the factor 6.228, for type II infection the odds increase by the factor 8.980. We will examine the differences of parameter estimates for different types of infection in more detail in Section 3.4.  $\square$

### 3.1.4 Multivariate Generalized Linear Models

Multinomial response models like (3.1.4) may be considered special cases of multivariate generalized linear models. In analogy to the univariate case (Section 2.1.2), multivariate generalized linear models are based on a distributional assumption and a structural assumption. However, the response variable  $y_i$  is now a  $q$ -dimensional vector with expectation

$$\mu_i = E(y_i|x_i).$$

1. *Distributional assumption:*

Given  $x_i$ , the  $y_i$ s are (conditionally) independent and have a distribution that belongs to a simple exponential family, which has the form

$$f(y_i|\theta_i, \phi, \omega_i) = \exp \left\{ \frac{[y'_i \theta_i - b(\theta_i)]}{\phi} \omega_i + c(y_i, \phi, \omega_i) \right\}.$$

2. *Structural assumption:*

The expectation  $\mu_i$  is determined by a linear predictor

$$\eta_i = Z_i \beta$$

in the form

$$\mu_i = h(\eta_i) = h(Z_i \beta),$$

where

- the response function  $h : S \rightarrow M$  is defined on  $S \subset \mathbb{R}^q$ , taking values in the admissible set  $M \subset \mathbb{R}^q$ ,
- $Z_i$  is a  $(q \times p)$ -design matrix, and
- $\beta' = (\beta_1, \dots, \beta_p)$  is a vector of unknown parameters from the admissible set  $B \subset \mathbb{R}^p$ .

For the case of a mult categorial response one has to consider the multinomial distribution, which may be embedded into the framework of a simple (multivariate) exponential family. For  $y'_i = (y_{i1}, \dots, y_{iq}) \sim M(n_i, \pi_i)$ , the distribution of the arithmetic mean  $\bar{y}_i = y_i/n_i$  has the form

$$f(\bar{y}_i|\theta_i, \phi, \omega_i) = \exp \left\{ \frac{\bar{y}'_i \theta_i - b(\theta_i)}{\phi} \omega_i + c(y_i, \phi, \omega_i) \right\}, \quad (3.1.7)$$

where the natural parameter  $\theta_i$  is given by

$$\theta'_i = \left[ \log \left( \frac{\pi_{i1}}{1 - \pi_{i1} - \dots - \pi_{iq}} \right), \dots, \log \left( \frac{\pi_{iq}}{1 - \pi_{i1} - \dots - \pi_{iq}} \right) \right],$$

and

$$\begin{aligned} b(\theta_i) &= -\log(1 - \pi_{i1} - \dots - \pi_{iq}), \\ c(y_i, \phi, \omega_i) &= \log \left( \frac{n_i!}{y_{i1}! \cdots y_{iq}!(n - y_{i1} - \dots - y_{iq})!} \right), \\ \omega_i &= n_i. \end{aligned}$$

The parameter  $\phi$  may be treated as an additional dispersion parameter. In the following it is considered fixed with  $\phi = 1$ .

For the multinomial distribution the conditional expectation  $\mu_i$  is the vector of probabilities  $\pi'(x_i) = (\pi_{i1}, \dots, \pi_{iq})$ ,  $\sum_{r=1}^q \pi_{ir} < 1$ . Therefore, the admissible set of expectations  $M$  is given by

$$M = \left\{ (u_1, \dots, u_q) \mid 0 < u_i < 1, \sum_i u_i < 1 \right\}.$$

For *metrical* response a multivariate generalized linear model in wide use is the multivariate linear model assuming a normal distribution.

## 3.2 Models for Nominal Responses

Let the response variable  $Y$  have possible values  $1, \dots, k$  where the numbers are mere labels for the categories, i.e., neither ordering nor difference between category numbers is meaningful. As an example one may consider biometrical problems where the categories  $1, \dots, k$  stand for alternative types of infection (see Example 3.1). Nominal responses often occur in situations where an individual faces  $k$  choices. Then the categories are given by the alternatives at hand. For example, in the choice of transportation mode the alternatives may be bus, train, or automobile. In the following, models for nominal responses are motivated for the frequent situation where the response can be interpreted as resulting from an individual choice.

### 3.2.1 The Principle of Maximum Random Utility

Models for unordered categories may be motivated from the consideration of latent variables. In probabilistic choice theory it is often assumed that an unobserved utility  $U_r$  is associated with the  $r$ th response category. More generally, let  $U_r$  be a latent variable associated with the  $r$ th category. For the choice of transportation mode the underlying variable may be interpreted as the consumers' utility connected to the transportation mode.

Let  $U_r$  be given by

$$U_r = u_r + \epsilon_r,$$

where  $u_r$  is a fixed value associated with the  $r$ th response category and  $\epsilon_1, \dots, \epsilon_k$  are i.i.d. random variables with continuous distribution function  $F$ . Following the principle of maximum random utility, the observable response  $Y$  is determined by

$$Y = r \Leftrightarrow U_r = \max_{j=1, \dots, k} U_j. \quad (3.2.1)$$

This means that the response is category  $r$  if the latent variable  $U_r$  underlying this category is maximal. In choice situations the alternative that has maximal utility is chosen.

From (3.2.1) it follows that

$$\begin{aligned}
P(Y = r) &= P(U_r - U_1 \geq 0, \dots, U_r - U_k \geq 0) \\
&= P(\epsilon_1 \leq u_r - u_1 + \epsilon_r, \dots, \epsilon_k \leq u_r - u_k + \epsilon_r) \\
&= \int_{-\infty}^{\infty} \prod_{s \neq r} F(u_r - u_s + \epsilon) f(\epsilon) d\epsilon
\end{aligned} \tag{3.2.2}$$

where  $f = F'$  is the density function of  $\epsilon_r$  and  $F$  denotes the distribution function of  $\epsilon_r$ .

Depending on the distributional assumption for the noise variables  $\epsilon_r$ , equation (3.2.2) yields different models. If the  $\epsilon$ 's are independently normally distributed, one gets the independent probit model. The more general multivariate probit model also allows correlated noise variables. A simpler model is generated by assuming independent noise variables following the extreme-value distribution

$$F(x) = \exp(-\exp(-x)). \tag{3.2.3}$$

Then by simple integration one gets the multinomial logit model

$$P(Y = r) = \frac{\exp(u_r)}{\sum_{s=1}^k \exp(u_s)}. \tag{3.2.4}$$

Since only differences of utilities are identifiable, it is useful to consider the alternative form

$$P(Y = r) = \frac{\exp(u_r - u_k)}{1 + \sum_{s=1}^{q-1} \exp(u_s - u_k)} = \frac{\exp(\tilde{u}_r)}{1 + \sum_{s=1}^{q-1} \exp(\tilde{u}_s)}, \tag{3.2.5}$$

where  $\tilde{u}_r = u_r - u_k$  is the difference between the  $r$ th utility and the reference utility  $u_k$ . We can see that the principle of maximum random utility in combination with a specific distribution function  $F$  does determine the link or response function of the model.

The response function  $h : S \rightarrow M$ ,  $S \subset \mathbf{R}^q$ , determines how the expectation  $\mu$  is related to the linear predictor  $\eta = Z\beta$ . For the multinomial logit model the response function  $h = (h_1, \dots, h_q)$  is given by (3.1.6).

The connection between the extreme-value distribution and the logit model has been considered by Yellott (1977) and McFadden (1973). More general models based on stochastic utility maximization have been treated in the literature. McFadden (1981) considered the generalized extreme-value distribution, Hausman & Wise (1978), Daganzo (1980), Lerman & Manski (1981), and McFadden (1984) considered probit models that do not assume independent utilities (see also Small, 1987, Börsch-Supan, 1990).

### 3.2.2 Modelling of Explanatory Variables: Choice of Design Matrix

For multivariate models there are more possibilities of specifying the influence term than there are for the univariate models considered in Chapter 2. In particular several types of variables might influence the response. Let us consider the situation where an individual faces  $k$  choices and a set of variables characterizes the individual. Let the  $i$ th individual be characterized by the vector  $z_i = (z_{i1}, \dots, z_{im})$  containing variables like age, gender and income. Since we now have to distinguish between individuals (observations), the index  $i$  is added. Consequently,  $u_{ir}$  will denote the utility of the  $r$ th category for individual  $i$ , and  $Y_i$  will denote the categorical response variable. A simple linear model for the utility  $u_{ir}$  is given by

$$u_{ir} = \alpha_{r0} + z'_i \alpha_r,$$

where  $\alpha_r = (\alpha_{r1}, \dots, \alpha_{rm})$  is a parameter vector. That means the preference of the  $r$ th alternative for the  $i$ th individual is determined by  $z_i$  and a parameter  $\alpha_r$  that depends on the category. For example, in the choice of transportation mode, the individual income may be regarded as a covariate affecting the preference of different alternatives (automobile, bus, train). The corresponding parameters  $\alpha_r$  depend on the category, e.g., for increasing income, an increasing preference of automobile but a decreasing preference of bus may be expected. In the following a parameter that depends on the category will be called *category-specific*. Since the explanatory variables  $z_i$  do not depend on the response categories the variables  $z_i$  are called *global*.

For the differences between utilities  $\tilde{u}_{ir} = u_{ir} - u_{ik}$ , one gets

$$\tilde{u}_{ir} = \beta_{r0} + z'_i \beta_r,$$

where  $\beta_{r0} = \alpha_{r0} - \alpha_{k0}$ ,  $\beta_r = (\alpha_r - \alpha_k)$ . Assuming only global variables  $z_i$ , we see that the multinomial logit model (3.2.5) has the form

$$P(Y_i = r|z_i) = \frac{\exp(\beta_{r0} + z'_i \beta_r)}{1 + \sum_{s=1}^q \exp(\beta_{s0} + z'_i \beta_s)}.$$

For the generalized linear model  $\mu_i = h(Z_i \beta)$  with response function (3.1.6), one has the design matrix

$$Z_i = \begin{bmatrix} 1 & z'_i & & 0 \\ & & 1 & z'_i \\ & & & \ddots \\ 0 & & & 1 & z'_i \end{bmatrix}$$

and the parameter vector  $\beta' = (\beta_{10}, \beta'_1, \dots, \beta'_{q0}, \beta_q)$ ; see Example 3.1.

In addition to global variables let the alternatives  $1, \dots, k$  themselves be characterized by variables. For example, when buying a new automobile the consumer faces choices characterized by price, type, speed, etc. That means we have  $k$  sets of variables  $w_1, \dots, w_k$  that are connected to the  $k$  alternatives. Consequently, variables of this type are called *alternative-specific*. Now let the utility be determined by

$$u_{ir} = \alpha_{r0} + z'_i \alpha_r + w'_r \gamma,$$

where the additional term  $w'_r \gamma$  with parameter  $\gamma$  accounts for the influence of the characteristics of alternatives (e.g., price, speed of transportation modes). Since  $\gamma$  does not depend on the category, it is called a *global* parameter. The identifiable differences  $\tilde{u}_{ir} = u_{ir} - u_{ik}$  are given by

$$\tilde{u}_{ir} = \beta_{r0} + z'_i \beta_r + (w_r - w_k)' \gamma.$$

The design matrix now has the form

$$Z_i = \begin{bmatrix} 1 & z'_i & & w'_1 - w'_k \\ & 1 & z'_i & w'_2 - w'_k \\ & & \ddots & \\ & 1 & z'_i & w'_q - w'_k \end{bmatrix}$$

and the parameter vector is given by  $\beta' = (\beta_{10}, \beta'_1, \dots, \beta'_{q0}, \beta'_q, \gamma')$ .

As is seen from the design matrices, variables that are weighted by category-specific parameters induce a diagonal structure, whereas variables that are weighted by global parameters induce a column in the design matrix. This structure does not depend on whether a variable is global (specific for the choice maker) or alternative-specific (specific for the alternatives at hand).

In an even more general case the alternative-specific variables may also vary across individuals. For the choice of transportation mode the price often depends on the location of the residence. That means that price is a variable  $\omega_{ir}$  that depends on both the alternative  $r = 1, \dots, k$  and the individual  $i = 1, \dots, n$ . We can include this type of alternative-specific variables  $\omega_{ir}$  into the design matrix  $Z_i$  by adding an extra column the same way we did with  $\omega_r$ . Note, however, that this column will contain values that differ among individuals, whereas the column with variables  $\omega_r$  does not depend on the individual, i.e., it is the same for all  $Z_i$  (see above).

Discrete choice modelling is treated extensively in Ben-Akiva & Lerman (1985), Maddala (1983), Pudney (1989), and Ronning (1991).

### 3.3 Models for Ordinal Responses

Response variables that have more than two categories are often ordinal. That means the events described by the category numbers  $1, \dots, k$  can be considered as ordered. However, one should keep in mind that only the ordering of the category numbers is meaningful. A formal theory of scale levels is developed in measurement theory (see, e.g., Roberts, 1979 or Krantz, Luce, Suppes & Tversky, 1971), but it is not necessary here. A formal characterization of a regression model as nominal or ordinal based on invariance principles is given in Tutz (1993). In the following we consider a more informal model ordinal if the ordering of response categories is taken into account. In particular for categorical data where the sample size is often critical, it is necessary to make use of all the information available. Consequently, the ordering of the response categories has to be taken into account, allowing for simpler models.

#### **Example 3.2: Breathing test results**

Forthofer & Lehnen (1981, p. 21) investigated the effect of age and smoking on breathing test results for workers in industrial plants in Texas (see also Agresti, 1984, p. 96 ff). The test results have been classified in three categories, namely normal, borderline, and abnormal. Thus, the response variable “breathing results” may be considered an ordinal variable. Table 3.1 gives the data from Forthofer & Lehnen (1981).  $\square$

#### **Example 3.3: Job expectation**

In a study on the expectations of students at the University of Regensburg, psychology students were asked if they expected to find adequate employment within a reasonable time after getting their degree. The response categories were ordered with respect to their expectation. The categories 1 (don’t expect adequate employment), 2 (not sure), and 3 (immediately

**Table 3.1.** Breathing results of Houston industrial workers

		Breathing Test Results		
Age	Smoking status	Normal	Borderline	Abnormal
< 40	Never smoked	577	27	7
	Former smoker	192	20	3
	Current smoker	682	46	11
40–59	Never smoked	164	4	0
	Former smoker	145	15	7
	Current smoker	245	47	27

after getting the degree) reflect this ordering. Table 3.2 shows the data for different ages of the students.  $\square$

Ordinal variables may stem from quite different mechanisms. Anderson (1984) distinguishes between *grouped continuous* variables and *assessed ordered* categorical variables. The first type is a mere categorized version of a continuous variable, which in principle may be observed itself. For example, if the breathing test of Example 3.2 provides measurement on a physical scale, e.g., volume of breath, and the distinction normal, borderline, and abnormal corresponds to intervals of the physical scale, then the variable “breath result” could be considered as a grouped continuous variable. The second type of ordered variable arises when an assessor processes an unknown amount of information leading to the judgment of the grade of the ordered categorical scale. In the job expectation example the students weighed some information about their age, their grades at the university, the level of unemployment in the country, etc., in making their judgment on the category of their expectation. Consequently, the response may be considered an assessed ordered variable.

**Table 3.2.** Grouped data for job expectations of psychology students in Regensburg

Observation number	Age in years	Response categories			$n_i$
		1	2	3	
1	19	1	2	0	3
2	20	5	18	2	25
3	21	6	19	2	27
4	22	1	6	3	10
5	23	2	7	3	12
6	24	1	7	5	13
7	25	0	0	3	3
8	26	0	1	0	1
9	27	0	2	1	3
10	29	1	0	0	1
11	30	0	0	2	2
12	31	0	1	0	1
13	34	0	1	0	1

### 3.3.1 Cumulative Models: The Threshold Approach

The most widely used model in ordinal regression is based on the so-called category boundaries or threshold approach, which dates back at least to Edwards & Thurstone (1952). It is assumed that the observable variable  $Y$  is merely a categorized version of a latent continuous variable  $U$ . In the case of a grouped continuous response variable,  $U$  may be considered the unobserved underlying variable. In the case of an assessed ordered variable,  $U$  is the assessment on the underlying continuous scale. In both cases the latent variable is primarily used for the construction of this type of model. Although interpretation is simpler when the latent variable is taken into account, the model may also be interpreted without reference to the underlying continuous variable.

For a given vector  $x$  of explanatory variables the category boundary approach postulates that the observable variable  $Y \in \{1, \dots, k\}$  and the unobservable latent variable  $U$  are connected by

$$Y = r \Leftrightarrow \theta_{r-1} < U \leq \theta_r, \quad r = 1, \dots, k \quad (3.3.1)$$

where  $-\infty = \theta_0 < \theta_1 < \dots < \theta_k = \infty$ . That means that  $Y$  is a coarser (categorized) version of  $U$  determined by the thresholds  $\theta_1, \dots, \theta_{k-1}$ . Moreover, it is assumed that the latent variable  $U$  is determined by the explanatory variables by the linear form

$$U = -x'\gamma + \epsilon, \quad (3.3.2)$$

where  $\gamma' = (\gamma_1, \dots, \gamma_p)$  is a vector of coefficients and  $\epsilon$  is a random variable with distribution function  $F$ . The “ $-$ ” may be incorporated into the parameter vector, yielding  $U = x'(-\gamma) + \epsilon$ . It is used only to obtain a simpler form of the model.

From these assumptions it follows immediately that the observed variable  $Y$  is determined by the model

$$P(Y \leq r|x) = P(U \leq \theta_r) = F(\theta_r + x'\gamma). \quad (3.3.3)$$

Since the left side of the equation is the sum of probabilities  $P(Y = 1|x) + \dots + P(Y = r|x)$ , model (3.3.3) is called a cumulative model with distribution function  $F$ . Alternatively, the model is often called a threshold model. This name is due to the derivation based on the thresholds  $\theta_1, \dots, \theta_{k-1}$  of the latent variable. These thresholds are unknown parameters that represent some sort of intercepts for the multivariate response.

#### Cumulative Logistic Model or Proportional Odds Model

Specific choices of the distribution function lead to specific cumulative models. A common choice of the distribution function is the logistic distribution

function  $F(x) = 1/(1 + \exp(-x))$ . Consequently, the cumulative logistic model has the form

$$P(Y \leq r|x) = \frac{\exp(\theta_r + x'\gamma)}{1 + \exp(\theta_r + x'\gamma)}, \quad (3.3.4)$$

$r = 1, \dots, q = k - 1$ . It is simple to show that equivalent forms are given by

$$\log \left\{ \frac{P(Y \leq r|x)}{P(Y > r|x)} \right\} = \theta_r + x'\gamma \quad (3.3.5)$$

or

$$\frac{P(Y \leq r|x)}{P(Y > r|x)} = \exp(\theta_r + x'\gamma). \quad (3.3.6)$$

Another way of looking at model (3.3.3) is to consider the density  $f = F'$  of the underlying continuous response  $U$ . The response mechanism (3.3.1) cuts the density of  $U$  into slices, whereby the cutoff points are determined by the thresholds. The explanatory term  $-x'\gamma$  in (3.3.2) determines the shift of the response  $U$  on the latent scale. Depending on the strength and direction of the shift, the observable probabilities of response categories increase or decrease. In Figure 3.1 for three distribution functions the densities for two subpopulations corresponding to explanatory terms  $-x_1'\gamma$  and  $-x_2'\gamma$  are plotted.

The logistic cumulative model has also been called the proportional odds model (McCullagh, 1980). This name is due to a special property of model (3.3.4). Since one has

$$\frac{P(Y \leq r|x)}{P(Y > r|x)} = \frac{P(Y \leq r|x)}{1 - P(Y \leq r|x)},$$

the left side of equation (3.3.6) represents the odds that  $Y \leq r$  occurs instead of  $Y > r$ . This proportion may be considered as cumulative odds.

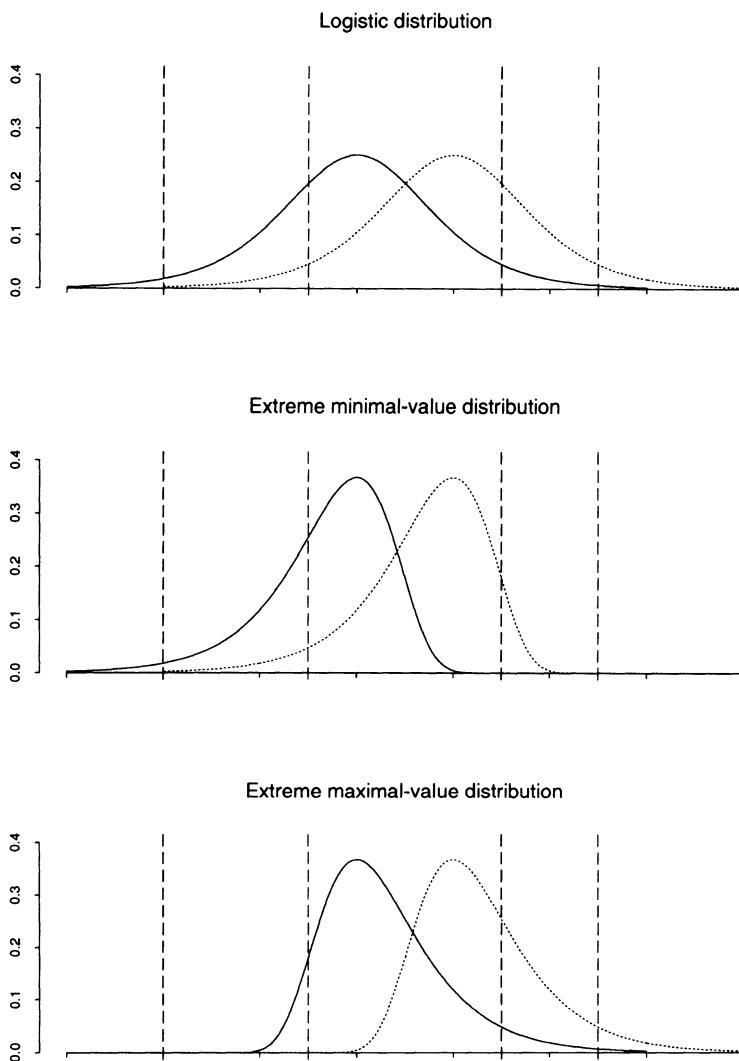
According to (3.3.5) the logarithms of the cumulative odds

$$\log \left\{ \frac{P(Y \leq r|x)}{P(Y > r|x)} \right\}, \quad r = 1, \dots, q,$$

are determined by a linear form of the explanatory variables. If two populations characterized by explanatory variables  $x_1$  and  $x_2$  are considered, the ratio of the cumulative odds for the two populations is given by

$$\frac{P(Y \leq r|x_1)/P(Y > r|x_1)}{P(Y \leq r|x_2)/P(Y > r|x_2)} = \exp\{(x_1 - x_2)'\gamma\} \quad (3.3.7)$$

and therefore does not depend on the category. That means the ratio of the cumulative odds for two populations is postulated to be the same for



**Figure 3.1.** Densities of the latent response for two subpopulations with different values of  $x$  (logistic, extreme minimal-value, extreme maximal-value distributions).

all the cumulative odds. This effect is essentially due to the derivation of the model from the latent regression model (3.3.2) where the parameter vector  $\gamma$  does not and cannot depend on the category. In extended versions of the cumulative model the parameter may also depend on the category (see Section 3.3.2). The property (3.3.7) yields a *strict stochastic ordering* of populations (see also Section 3.3.5).

### Grouped Cox Model or Proportional Hazards Model

Another choice of the distribution function  $F$  is the extreme minimal-value distribution  $F(x) = 1 - \exp(-\exp(x))$ . By inserting this distribution function in (3.3.3), one gets the model

$$P(Y \leq r|x) = 1 - \exp\{-\exp(\theta_r + x'\gamma)\}, \quad r = 1, \dots, q, \quad (3.3.8)$$

or, equivalently with complementary log-log links,

$$\log[-\log P(Y > r|x)] = \theta_r + x'\gamma, \quad r = 1, \dots, q.$$

The latter form shows that  $\log(-\log P(Y > r|x))$  is determined by a linear influence term composed of the explanatory variables. The model is called the *grouped Cox model* since it may be derived as a grouped version of the continuous Cox or proportional hazards model, which is well known in survival analysis (e.g., Kalbfleisch & Prentice, 1980).

As is seen in Figure 3.1, for small values the underlying density is rather similar to that of the cumulative logistic model. Consequently the models often yield very similar fits.

### Extreme Maximal-value Distribution Model

Instead of the extreme minimal-value distribution  $F(x) = 1 - \exp(-\exp(x))$ , the extreme maximal-value distribution  $F(x) = \exp(-\exp(-x))$  may be used. While the former is rather similar to the logistic distribution for small  $x$ , the latter is rather similar to the logistic distribution for large  $x$ . The model based on the extreme maximal-value distribution is given by

$$P(Y \leq r|x) = \exp\{-\exp\{-(\theta_r + x'\gamma)\}\} \quad (3.3.9)$$

or, equivalently with log-log links,

$$\log[-\log P(Y \leq r|x)] = -(\theta_r + x'\gamma). \quad (3.3.10)$$

Although model (3.3.8) is not equivalent to model (3.3.9), it may be estimated using the latter one. To do so, one has to construct first a variable  $\tilde{Y} = k + 1 - Y$  having support  $1, \dots, k$  but with inverse ordering of categories, and then estimate model (3.3.9) using  $\tilde{Y}$  instead of  $Y$ . In a second

step, the obtained estimates have to be multiplied by  $-1$  to yield the estimates for model (3.3.8), and the order of the threshold parameters  $\theta_r$  has to be reversed. More formally, let the ordering of categories be reversed by  $\tilde{Y} = k + 1 - Y$ . Then the extreme minimal-value model (3.3.8) for  $Y$  with parameters  $\theta_r, \beta$  is equivalent to the extreme maximal-value model (3.3.9) for  $\tilde{Y}$  with parameters  $\tilde{\theta}_r = -\theta_{k-\tilde{r}}, \tilde{\beta} = -\beta$ .

Several families of link functions have been proposed for the binary response model by Prentice (1976), Pregibon (1980), Aranda-Ordaz (1983), Morgan (1985), Stukel (1988), Friedl (1991), and Czado (1992). In principle, those families may also be used for ordinal response variables. Genter & Farewell (1985) consider a family of link functions within the cumulative model approach. Their family includes the probit model as well as the extreme minimal-value link and extreme maximal-value link for the modelling of ordinal responses.

### 3.3.2 Extended Versions of Cumulative Models

The simple cumulative model (3.3.3) is based on the assumption that the explanatory variable causes a shift on the latent scale but does not change the thresholds  $\theta_1, \dots, \theta_q$ . In a more general model the threshold may also depend on explanatory variables  $w' = (w_1, \dots, w_m)$  in the linear form

$$\theta_r = \beta_{r0} + w' \beta_r,$$

where  $\beta'_r = (\beta_{r1}, \dots, \beta_{rm})$  is a category-specific parameter vector (see Terza, 1985). The extended model follows directly from the mechanism (3.3.1) and the parameterization of the latent variable (3.3.2). It is given by

$$P(Y \leq r|x, w) = F(\beta_{r0} + w' \beta_r + x' \gamma). \quad (3.3.11)$$

Model (3.3.11) still assumes the category boundaries mechanism, but now only for explanatory variables  $x$ . The variables  $w$  only determine the thresholds that lie on the latent score. As is immediately seen, the assumption  $w_i = x_i, i = 1, \dots, m, m = p$ , makes  $\gamma$  an unidentifiable parameter. Thus, one has to distinguish strictly between threshold variables  $w_i$  and shift variables  $x_i$ . Threshold variables are always weighted by a category-specific parameter vector  $\beta_r$  and shift variables are always weighted by a global parameter vector  $\gamma$ . Of course, the  $x$ -variables may be weighted by category-specific parameters yielding

$$P(Y \leq r|x) = F(\beta_{r0} + x' \beta_r), \quad (3.3.12)$$

which generalizes model (3.3.3) by assuming category-specific parameters on  $x$ . Model (3.3.3) is the special case where  $\beta_1 = \dots = \beta_q = \gamma$ .

### 3.3.3 Link Functions and Design Matrices for Cumulative Models

The embedding into the framework of generalized linear models is done by specifying the link (or response) function and the design matrix. The link function  $g = (g_1, \dots, g_q)$  is immediately given from (3.3.3) or (3.3.11) by

$$g_r(\pi_1, \dots, \pi_q) = F^{-1}(\pi_1 + \dots + \pi_r),$$

$r = 1, \dots, q$ . For the grouped Cox model one has

$$g_r(\pi_1, \dots, \pi_q) = \log\{-\log(1 - \pi_1 - \dots - \pi_r)\},$$

$r = 1, \dots, q$ . For the design matrix one has to distinguish between the simple model (3.3.3) and the general model (3.3.11). For the simple model the linear term  $\eta_i = Z_i\beta$  is determined by

$$Z_i = \begin{bmatrix} 1 & & x'_i \\ & 1 & x'_i \\ & & \ddots & \vdots \\ & & & 1 & x'_i \end{bmatrix} \quad (3.3.13)$$

and the parameter vector  $\beta = (\theta_1, \dots, \theta_q, \gamma)$ . For the more general model with threshold variables  $w_i$  and shift variables  $x_i$ , the design matrix has the form

$$Z_i = \begin{bmatrix} 1 & w'_i & & x'_i \\ & 1 & w'_i & x'_i \\ & & \ddots & \vdots \\ & & & 1 & w'_i & x'_i \end{bmatrix}. \quad (3.3.14)$$

The parameter vector  $\beta$  is given by  $\beta' = (\beta_{10}, \beta'_1, \dots, \beta_{q0}, \beta'_q, \gamma')$ . Here, neither the threshold variables nor the shift variables contain a constant. The constant itself corresponds to the thresholds  $\beta_{r0}$  that are category-specific.

Sometimes, e.g., in random coefficient models (Chapter 7) or in dynamic modelling (Chapter 8), it is useful to consider an alternative form for the link function and design matrix. For cumulative models the parameters may not vary freely. The parameters for the simple model (3.3.3) are restricted by  $\theta_1 < \dots < \theta_q$ . For the general model, the restriction is determined by  $\beta_{10} + \beta'_1 w < \dots < \beta_{q0} + \beta'_q w$  for all possible values of the shift variables  $w$ . For link functions and design matrix as given earlier, these restrictions are not taken into account. Since the restriction must hold for each  $w$ , the severity

of the restriction depends on the range of the covariates  $w$ . If the constraint is not explicitly used in estimation, the iterative estimation procedure may fail by fitting inadmissible parameters. There will be no problems if, for the simple model (3.3.3), the thresholds are well separated. However, there might be numerical problems in the estimation procedure if some thresholds are very similar. For model (3.3.3) these problems may be simply avoided by using an alternative formulation. The model may be reparameterized by

$$\alpha_1 = \theta_1, \quad \alpha_r = \log(\theta_r - \theta_{r-1}), \quad r = 2, \dots, q,$$

or, respectively,

$$\theta_1 = \alpha_1, \quad \theta_r = \theta_1 + \sum_{i=2}^r \exp(\alpha_i), \quad r = 2, \dots, q.$$

Then the parameters  $\alpha_1, \dots, \alpha_q$  are not restricted and we have  $(\alpha_1, \dots, \alpha_q) \in \mathbb{R}^q$ . The linear structure of the model becomes obvious in the form

$$F^{-1}(P(Y = 1|x)) = \alpha_1 + x'\gamma, \\ \log \left[ F^{-1}\{P(Y \leq r|x)\} - F^{-1}\{P(Y \leq r-1|x)\} \right] = \alpha_r, \quad r = 2, \dots, q.$$

From this the link function is determined.

For the special case of the logistic cumulative model, one gets the link function

$$g_1(\pi_1, \dots, \pi_q) = \log \left( \frac{\pi_1}{1-\pi_1} \right), \\ g_r(\pi_1, \dots, \pi_q) = \log \left[ \log \left\{ \frac{\pi_1 + \dots + \pi_r}{1-\pi_1 - \dots - \pi_r} \right\} \right. \\ \left. - \log \left\{ \frac{\pi_1 + \dots + \pi_{r-1}}{1-\pi_1 - \dots - \pi_{r-1}} \right\} \right],$$

$r = 1, \dots, q$ . Of course, when one uses this alternative link function for the logit-type model, the design matrix has to be adapted. The design matrix for observation  $(y_i, x_i)$  now has the form

$$Z_i = \begin{bmatrix} 1 & & x'_i \\ & 1 & 0 \\ & \ddots & \vdots \\ & & 1 & 0 \end{bmatrix} \quad (3.3.15)$$

and the parameter vector is given by  $\beta = (\alpha_1, \dots, \alpha_q, \gamma)$ .

**Example 3.4: Breathing test results** (Example 3.2, continued)

The data given in Table 3.1 are used to investigate the effect of age and smoking on breathing test results. The analysis is based on the cumulative model in three variants, namely, the cumulative logistic model, the proportional hazards model, and the extreme maximal-value model. If one takes the deviance (see Section 3.4) as a measure of distance between the data and the fitted values, the proportional hazards model shows the best fit. The extreme maximal-value model shows the worst fit. Thus, it is neglected in the following.

Table 3.3 gives the estimates for the three models of the simple type (3.3.3). All variables are given in effect coding, with  $-1$  for the last category. A positive sign signals a shift on the latent scale to the left end, yielding higher probabilities for normal and borderline categories (i.e., categories 1 and 2). As is to be expected, low age and low smoking categories produce better breathing test results.

Parameter estimates must be different for the models because of the differing variance of the logistic and the extreme-value distributions. As seen in Figure 2.2 (Chapter 2, p. 15), the extreme minimal-value distribution function, underlying the proportional hazards model, is steeper than the logistic distribution function. Therefore, to achieve the same amount of shifting on the latent scale a larger effect (measured by the parameter) is necessary for the latter distribution function. Consequently, the parameters

**Table 3.3.** Estimates of cumulative models for breathing test data (*p*-values in each second column)

	Cumulative logistic		Proportional hazards		Extreme maximal-value	
Threshold 1	2.370	0.0	0.872	0.0	2.429	0.0
Threshold 2	3.844	0.0	1.377	0.0	3.843	0.0
AGE[1]	0.114	0.29	0.068	0.04	0.095	0.37
SMOKE[1]	0.905	0.0	0.318	0.0	0.866	0.19
SMOKE[2]	-0.364	0.01	-0.110	0.02	-0.359	0.14
AGE[1]*SMOKE[1]	-0.557	0.0	-0.211	0.00	-0.529	0.19
AGE[1]*SMOKE[2]	0.015	0.91	0.004	0.92	0.021	0.14
Deviance	8.146		3.127		9.514	

for the logistic model are larger for all the variables. However, the tendency of the parameters is about the same.

In this data set the strong interaction effect of age and smoking is interesting. It is much more impressive than the main effect of age. The interaction  $\text{AGE}[1]*\text{SMOKE}[1] = -0.211$  (proportional hazards model) shows that the positive tendency given by the strong influence  $\text{SMOKE}[1] = 0.318$  is not so impressive when the person is still young. Since effect coding was used, the interaction effects not given in Table 3.3 are easily computed via the restriction that they sum to zero. This leads to the following table of interaction effects.

	SMOKE[1]	SMOKE[2]	SMOKE[3]
AGE[1]	-0.211	0.004	0.207
AGE[2]	0.211	-0.004	-0.207

From this table of interactions it can be easily seen that the smoking history becomes quite influential for higher ages. Note that the interaction  $\text{AGE}[2]*\text{SMOKE}[3] = -0.207$  has to be added to the negative effects  $\text{AGE}[2] = -0.068$  and  $\text{SMOKE}[3] = -0.208$ . The same conclusions are drawn from the logistic model, although the fit is inferior in comparison to the proportional hazards model.  $\square$

### Example 3.5: Job expectation (Example 3.3, continued)

For the job expectation data given in Table 3.2,  $\log \text{AGE}$  is considered in the influence term. Two models have been fitted:

the simple cumulative logistic model

$$P(Y \leq r | \text{AGE}) = F(\theta_r + \gamma \log \text{AGE}),$$

and the extended version

$$P(Y \leq r | \text{AGE}) = F(\beta_{r0} + \beta_r \log \text{AGE}),$$

where  $\log \text{AGE}$  is a threshold variable.

Estimates are given in Table 3.4. In particular, the simple model where  $\log \text{AGE}$  is a shift variable shows a rather bad fit in Pearson's  $\chi^2$ . The fit of the extended version is not overwhelming but not so bad. The strong difference between deviance and Pearson's  $\chi^2$  may be a hint that the assumptions for asymptotics may be violated (see increasing cells asymptotic in Section 3.4.3). The extended version yields estimates for the parameters and an obvious change in the threshold values. The negative values of  $\hat{\beta}_r$  (and  $\hat{\gamma}$ ) signal that increasing age yields lower probabilities for low categories. Students seem to be more optimistic when getting closer to the examinations. However, the effect on the cumulative odds

**Table 3.4.** Cumulative model for job expectation data (*p*-values are given in brackets)

	Log AGE as global shift variable	Log AGE as threshold variable
Threshold 1	14.987 (0.010)	9.467 (0.304)
Threshold 2	18.149 (0.002)	20.385 (0.002)
Slope $\gamma$	-5.402 (0.004)	
Category-specific slope $\beta_1$		-3.597 (0.230)
Category-specific slope $\beta_2$		-6.113 (0.004)
<hr/>		
Pearson's $\chi^2$	42.696 (0.007)	33.503 (0.055)
Deviance	26.733 (0.267)	26.063 (0.248)

$$\frac{P(Y \leq r|x_1)/P(Y > r|x_1)}{P(Y \leq r|x_2)/P(Y > r|x_2)} = \exp(\beta_{r0} + \beta_r(x_1 - x_2))$$

depends on the category  $r$ . Consider age groups  $x_1 > x_2$ . Then the cumulative odds for  $r = 1$  measures the tendency toward strongly negative expectation (category 1: don't expect adequate employment); for  $r = 2$  the cumulative odds measure the tendency toward strongly negative *or* uncertain expectation (category 2: not sure) in comparison to the positive statement of category 3. Since  $\hat{\beta}_2 < \hat{\beta}_1$ , the effect of age on the latter cumulative odds is stronger than for the former cumulative odds. Looking at the *p*-value 0.230 of  $\beta_1$  shows that the real negative expectation (category 1 versus categories 2 and 3) may even be influenced by age. However, even the extended model shows a very high value of Pearson's  $\chi^2$ , which is a goodness-of-fit measure given in Section 3.4.2. The bad fit suggests further investigation of the residuals, which is given in Chapter 4.  $\square$

### 3.3.4 Sequential Models

In many applications the ordering of the response categories is due to a sequential mechanism. The categories are ordered since they can be reached

only successively. A response variable of this type is children's tonsil size, as considered in the following example.

**Example 3.6: Tonsil size**

A data set that several authors have considered (e.g., Holmes & Williams, 1954; McCullagh, 1980), focuses on the tonsil size of children. Children have been classified according to their relative tonsil size and whether or not they are carriers of Streptococcus pyogenes. The data are given in Table 3.5.

**Table 3.5.** Tonsil size and Streptococcus pyogenes (Holmes & Williams, 1954)

	Present but not enlarged	Enlarged	Greatly enlarged
Carriers	19	29	24
Noncarriers	497	560	269

It may be assumed that tonsil size always starts in the normal state "present but not enlarged" (category 1). If the tonsils grow abnormally, they may become "enlarged" (category 2); if the process does not stop, tonsils may become "greatly enlarged" (category 3). But in order to get greatly enlarged tonsils, they first have to be enlarged for the duration of the intermediate state "enlarged."  $\square$

For data of this type, models based on a sequential mechanism will often be more appropriate. In a similar way as for the cumulative model, sequential models may be motivated from latent variables.

Let latent variables  $U_r$ ,  $r = 1, \dots, k - 1$ , have the linear form  $U_r = -x'\gamma + \epsilon_r$ , where  $\epsilon_r$  is a random variable with distribution function  $F$ . The response mechanism starts in category 1 and the first step is determined by

$$Y = 1 \quad \Leftrightarrow \quad U_1 \leq \theta_1,$$

where  $\theta_1$  is a threshold parameter. If  $U_1 \leq \theta_1$ , the process stops. For the tonsil size data  $U_1$  may represent the latent tendency of growth in the initial state of normal tonsil size. If  $U_1$  is below threshold  $\theta_1$ , the tonsil size remains normal ( $Y = 1$ ); if not, at least enlarged tonsils result ( $Y \geq 2$ ). That means that if  $U_1 > \theta_1$ , the process is continuing in the form

$$Y = 2 \quad \text{given} \quad Y \geq 2 \quad \Leftrightarrow \quad U_2 \leq \theta_2$$

and so on. The latent variable  $U_2$  may represent the unobservable tendency of growth when the tonsils are already enlarged. Generally, the complete mechanism is specified by

$$Y = r \quad \text{given} \quad Y \geq r \quad \Leftrightarrow \quad U_r \leq \theta_r$$

or, equivalently,

$$Y > r \quad \text{given} \quad Y \geq r \quad \Leftrightarrow \quad U_r > \theta_r, \quad (3.3.16)$$

$r = 1, \dots, k - 1$ . The sequential mechanism (3.3.16) models the transition from category  $r$  to category  $r + 1$  given that category  $r$  is reached. A transition takes place only if the latent variable determining the transition is above a threshold that is characteristic for the category under consideration.

The main difference to the category boundaries approach is the conditional modelling of transitions. The sequential mechanism assumes a binary decision in each step. Given that category  $r$  is reached, it must be decided whether the process stops (thus getting  $r$  as the final category) or whether it continues with a resulting higher category. Only the final resulting category is observable.

The sequential response mechanism (3.3.16) combined with the linear form of the latent variables  $U_r = -x'\gamma + \epsilon_r$  immediately leads to the sequential model with distribution function  $F$

$$P(Y = r|Y \geq r, x) = F(\theta_r + x'\gamma), \quad (3.3.17)$$

$r = 1, \dots, k$ , where  $\theta_k = \infty$ . The probabilities of the model are given by

$$P(Y = r|x) = F(\theta_r + x'\gamma) \prod_{i=1}^{r-1} \{1 - F(\theta_i + x'\gamma)\},$$

$r = 1, \dots, k$ , where  $\prod_{i=1}^0 \{\cdot\} = 1$ . Model (3.3.17) is also called the continuation ratio model (e.g., Agresti, 1984). Note that no ordering restriction is needed for the parameters  $\theta_1, \dots, \theta_q$ , as was the case for the cumulative-type model.

So far we have considered only the general form of the sequential model. There are several sequential models depending on the choice of the distribution function  $F$ . If  $F$  is chosen as the logistic distribution  $F(x) = 1/(1 + \exp(-x))$ , we get the sequential logit model

$$P(Y = r|Y \geq r, x) = \frac{\exp(\theta_r + x'\gamma)}{1 + \exp(\theta_r + x'\gamma)}$$

or, equivalently,

$$\log \left\{ \frac{P(Y = r|x)}{P(Y > r|x)} \right\} = \theta_r + x'\gamma.$$

For the extreme-value-distribution  $F(x) = 1 - \exp(-\exp(x))$ , the model has the form

$$P(Y = r|Y \geq r, x) = 1 - \exp(-\exp(\theta_r + x'\gamma)) \quad (3.3.18)$$

or, equivalently,

$$\log \left[ -\log \left\{ \frac{P(Y > r|x)}{P(Y \geq r|x)} \right\} \right] = \theta_r + x' \gamma. \quad (3.3.19)$$

It is noteworthy that this sequential model is equivalent to the *cumulative* model with distribution function  $F(x) = 1 - \exp(-\exp(x))$ . That means model (3.3.19) is a special parametric form of the grouped Cox model. The equivalence is easily seen by the reparameterization

$$\theta_r = \log \{ \exp(\tilde{\theta}_r) - \exp(\tilde{\theta}_{r-1}) \}, \quad r = 1, \dots, k-1,$$

or

$$\tilde{\theta}_r = \log \left( \sum_{i=1}^r \exp(\theta_i) \right).$$

If  $\theta_r$  is inserted in (3.3.19), one gets the grouped Cox model

$$P(Y \leq r|x) = 1 - \exp(-\exp(\tilde{\theta}_r + x' \gamma))$$

(see also Lääärä & Matthews, 1985, Tutz, 1991c).

Another member of the sequential family is the exponential sequential model, which is based on the exponential distribution  $F(x) = 1 - \exp(-x)$ . It is given by

$$P(Y = r|Y \geq r, x) = 1 - \exp(-(\theta_r + x' \gamma))$$

or, equivalently,

$$-\log \left( \frac{P(Y > r|x)}{P(Y \geq r|x)} \right) = \theta_r + x' \gamma.$$

Comparisons between cumulative models and sequential models are found in Armstrong & Sloan (1989) and Greenland (1994); see also the overview in Barnhart & Sampson (1994) and the investigation of sequential models in Cox (1988).

### Generalized Sequential Models

In the same way as for the cumulative model, the thresholds  $\theta_r$  of the sequential model may be determined by covariates  $z' = (z_1, \dots, z_m)$ . From the linear form

$$\theta_r = \delta_{ro} + z' \delta_r,$$

where  $\delta'_r = (\delta_{r1}, \dots, \delta_{rm})$  is a category-specific parameter vector, one gets the generalized sequential model

$$P(Y = r|Y \geq r, x) = F(\delta_{r0} + z'\delta_r + x'\gamma). \quad (3.3.20)$$

Alternatively, model (3.3.20) can be derived directly from the sequential mechanism. One has to assume that the latent variables  $U_r$  have the form  $U_r = -x'\gamma - z'\delta_r + \epsilon_r$  and that the response mechanism

$$Y > r \text{ given } Y \geq r \text{ if } U_r > \delta_{r0}$$

is given for the fixed thresholds  $\delta_{r0}$ .

It is implicitly assumed that the influence of variables  $z$  on the transition from category  $r$  to  $r+1$  depends on the category. The effect of the variables  $z$  is nonhomogeneous over the categories, whereas the effect of variables  $x$  is homogeneous. Given that category  $r$  is reached, the transition to a higher category is always determined by  $x'\gamma$ . Since the shift of the underlying score is determined by  $x'\gamma$  and is constant over categories, in analogy to the general cumulative model, variables  $x$  are called shift variables. Although the assumption of linearly determined thresholds  $\theta_r = \delta_{r0} + z'\delta_r$  is optional, variables  $z$  with a category-specific weight  $\delta_r$  are called threshold variables. Thus, the distinction between shift and threshold variables corresponding to global or category-specific weighting is the same as for cumulative models.

### **Example 3.7: Tonsil size** (Example 3.6, continued)

Table 3.6 shows the estimates for the tonsil size data for the cumulative logit model and for the sequential logit model. Pearson statistic and deviance suggest a better fit of the sequential model. However, the extremely small values of the statistics for the sequential model may be a hint for underdispersion (see Section 3.4). Parameters have to be interpreted with reference to the type of model used. The estimated parameter  $\gamma = -0.301$  for the cumulative model means that the odds  $P(Y \leq r|x)/P(Y > r|x)$  of having normal-sized tonsils ( $r = 1$ ) as well as the odds of having normal-sized or enlarged tonsils ( $r = 2$ ) is  $\exp((x_1 - x_2)'\gamma) = \exp(-0.301(-1 - 1)) \approx 1.8$  times as large for noncarriers ( $x = -1$ ) as for carriers ( $x = 1$ ;  $x$  was effect-coded). Within the sequential model the parameter  $\gamma$  gives the strength with which the transition from category 1 to 2, and from 2 to 3, is determined. The estimated value  $\gamma = -0.264$  means that the odds  $P(Y = r|x)/P(Y > r|x)$  of having normal-sized tonsils ( $r = 1$ ) is  $\exp((x_1 - x_2)'\gamma) = \exp(-0.264(-1 - 1)) \approx 1.7$  times as large for noncarriers as for carriers. The same proportion holds for the odds  $P(Y = 2|x)/P(Y > 2|x)$  of having merely enlarged tonsils given that the tonsils are not normal. The sequential model as a stepwise model assumes that the process of enlargement as far as the comparison of carriers and noncarriers is concerned does not stop in the “enlarged” category but goes on and leads to greatly enlarged tonsils, where  $(x_1 - x_2)'\gamma = 2\gamma$  determines the comparison between populations.  $\square$

### **Example 3.8: Breathing test results** (Examples 3.2 and 3.4, continued)

The effect of age and smoking history on breathing test results has already been investigated by use of the cumulative model. However, the categories

**Table 3.6.** Fits for tonsil size data

	Cumulative logit	Sequential logit
$\theta_1$	-0.809 (0.013)	-0.775 (0.011)
$\theta_2$	1.061 (0.014)	0.468 (0.012)
$\gamma$	-0.301 (0.013)	-0.264 (0.010)
Pearson	0.301	0.005
Deviance	0.302	0.006
DF	1	1

The variable is given in effect coding,  $p$ -values are given in parentheses.

“normal,” “borderline,” and “abnormal” may be seen as arising from a sequential mechanism starting with “normal test results.” Consequently, the sequential model may be used for this data set. Table 3.7 gives the estimates for two sequential logit models. The first is of the simple type (3.3.17) with AGE, SMOKE, and the interaction AGE\*SMOKE. The second one is of the generalized type (3.3.20) where AGE is a shift variable (with global weight) and SMOKE and the interaction AGE\*SMOKE are threshold variables (with category-specific weight). The deviance for the first model is 4.310 on 5 d.f., which shows that the sequential model fits the data better than the cumulative model (see Table 3.3). The estimates of the generalized-type model are given for comparison. It is seen that for variable SMOKE the category-specific effects are not very different for the two thresholds. The first has significant effects; for the second threshold the  $p$ -values are rather high, an effect due to the low number of observations in category 3. The thresholds for the interaction AGE\*SMOKE[2] have quite different effects but both of them turn out to be not significant.  $\square$

**Table 3.7.** Sequential logit models for the breathing test data (*p*-values in brackets)

Simple sequential logit model			Sequential logit model with threshold variables	
Threshold 1	2.379	(0.0)	2.379	(0.0)
Threshold 2	1.516	(0.0)	1.510	(0.0)
AGE	0.094	(0.368)	0.092	(0.385)
SMOKE[1]	0.882	(0.0)	0.915	(0.0)
			0.675	(0.108)
SMOKE[2]	-0.356	(0.008)	-0.375	(0.008)
			-0.163	(0.609)
AGE*SMOKE[1]	-0.601	(0.001)	-0.561	(0.003)
			-0.894	(0.047)
AGE*SMOKE[2]	0.092	(0.492)	0.015	(0.912)
			0.532	(0.161)

### Link Functions of Sequential Models

Response and link functions may be derived directly from model (3.3.17). The link function  $g = (g_1, \dots, g_q)$  is given by

$$g_r(\pi_1, \dots, \pi_q) = F^{-1}(\pi_r / (1 - \pi_1 - \dots - \pi_{r-1})), \quad r = 1, \dots, q,$$

and the response function  $h = (h_1, \dots, h_q)$  has the form

$$h_r(\eta_1, \dots, \eta_q) = F(\eta_r) \prod_{i=1}^{r-1} \left(1 - F(\eta_i)\right), \quad r = 1, \dots, q.$$

For the sequential logit model we have

$$g_r(\pi_1, \dots, \pi_q) = \log \left( \frac{\pi_r}{1 - \pi_1 - \dots - \pi_r} \right), \quad r = 1, \dots, q,$$

and

$$h_r(\eta_1, \dots, \eta_q) = \exp(\eta_r) \prod_{i=1}^{r-1} \left(1 + \exp(\eta_i)\right)^{-1}, \quad r = 1, \dots, q.$$

For the other models the functions may be easily derived.

Naturally, the design matrices depend on the specification of shift and threshold variables. But as far as design matrices are concerned, there is no difference between sequential and cumulative models. Thus, the design matrices from Section 3.3.3 apply.

### 3.3.5 Strict Stochastic Ordering\*

In Section 3.3.1 it is shown that the ratio of the cumulative odds for the subpopulations does not depend on the category. This property, called strict stochastic ordering (McCullagh, 1980), is shared by all simple cumulative models of the type (3.3.3). Let two subpopulations be represented by covariates  $x_1, x_2$ . Then for the cumulative model (3.3.3) the difference

$$\Delta_c(x_1, x_2) = F^{-1}\left\{P(Y \leq r|x_1)\right\} - F^{-1}\left\{P(Y \leq r|x_2)\right\}$$

is given by  $\gamma'(x_1 - x_2)$ . This means that, e.g., for the logistic cumulative model, where  $F^{-1}\{P(Y \leq r|x)\} = \log\{P(Y \leq r|x)/P(Y > r|x)\} = l_r(x)$ , the difference between “cumulative” log-odds

$$\Delta_c(x_1, x_2) = l_r(x_1) - l_r(x_2) = \gamma'(x_1 - x_2)$$

does not depend on the category. Thus if one of the cumulative log-odds  $l_r$  is larger for population  $x_1$  than for population  $x_2$ , then this holds for the cumulative log-odd of any category. For the grouped Cox model one gets

$$F^{-1}\{P(Y \leq r|x)\} = \log(-\log P(Y > r|x)).$$

Thus the difference of log-log-transformed cumulative probabilities does not depend on  $r$ .

It should be noted that for the *general* cumulative model (3.3.11) the strict stochastic ordering of subpopulations no longer holds. Omitting the term  $x'\gamma$  and using  $w = x$  yields a special generalization of the simple model (3.3.3). For this model a test of  $\beta_1 = \dots = \beta_q$  may be considered a test of strict stochastic ordering. Armstrong & Sloan (1989) used this approach and considered the relative efficiency of using dichotomized outcomes compared with ordinal models. For tests of the type  $\beta_1 = \dots = \beta_q$ , see Section 3.4.2.

Strict stochastic ordering is not restricted to the cumulative models. As is seen from the sequential model (3.3.17), for two subpopulations characterized by covariates  $x_1, x_2$ , we get

$$\begin{aligned}\Delta_s(x_1, x_2) &= F^{-1}\{P(Y = r|Y \geq r, x_1)\} - F^{-1}\{P(Y = r|Y \geq r, x_2)\} \\ &= \gamma'(x_1 - x_2).\end{aligned}$$

That means the difference  $\Delta_s(x_1, x_2)$  does not depend on the category. If the hazard  $P(Y = r|Y \geq r)$  for subpopulation  $x_2$  is larger than the hazard for subpopulation  $x_1$ , this relation holds for any category  $r$  (see also Tutz, 1991c).

### 3.3.6 Two-Step Models

Both types of models, cumulative and sequential, only make use of the ordering of the response categories  $1, \dots, k$ . However, often the response categories quite naturally may be divided into sets of categories with very homogeneous responses where the sets are heterogeneous.

#### Example 3.9: Rheumatoid arthritis

Mehta, Patel & Tsatsis (1984) analyzed data of patients with acute rheumatoid arthritis. A new agent was compared with an active control, and each patient was evaluated on a five-point assessment scale. The data are given in Table 3.8.

**Table 3.8.** Clinical trial of a new agent and an active control

Drug	Global assessment				
	Much improved	Improved	No change	Worse	Much worse
New agent	24	37	21	19	6
Active control	11	51	22	21	7

The global assessment in this example may be subdivided into the coarse response “improvement,” “no change,” and “worse.” On a higher level “improvement” is split up into “much improved” and “improved” and the “worse” category is split into “worse” and “much worse.” Thus, there is a split on the response scale after category 2 and a split after category 3, yielding three sets of homogeneous responses.  $\square$

For data of this type it is useful to model in a first step the coarse response, and in a second step the response within the sets of homogeneous categories. More generally, let the categories  $1, \dots, k$  be subdivided into  $t$  basic sets  $S_1, \dots, S_t$ , where  $S_j = \{m_{j-1} + 1, \dots, m_j\}$ ,  $m_0 = 0$ ,  $m_t = k$ .

Let the response in the *first* step in one of the sets be determined by a cumulative model based on an underlying latent variable  $U_o = -x'\gamma_o + \epsilon$ , where the random variable  $\epsilon$  has distribution function  $F$ . The response mechanism is given by

$$Y \in S_j \Leftrightarrow \theta_{j-1} < U_o \leq \theta_j.$$

In the *second* step let the conditional mechanism conditioned on  $S_j$  also be determined by a cumulative model based on a latent variable  $U_j = -x'\gamma_j + \epsilon_j$ , with  $\epsilon_j$  also having distribution function  $F$ . We assume

$$Y = r | Y \in S_j \Leftrightarrow \theta_{j,r-1} < U_j \leq \theta_{jr}.$$

With the assumption of independent noise variables  $\epsilon_j$ , the resulting model is given by

$$\begin{aligned} P(Y \in T_j | x) &= F(\theta_j + x' \gamma_o), \\ P(Y \leq r | Y \in S_j, x) &= F(\theta_{jr} + x' \gamma_j), \end{aligned} \quad (3.3.21)$$

where  $T_j = S_1 \cup \dots \cup S_j$ ,  $\theta_1 < \dots < \theta_{t-1}$ ,  $\theta_t = \infty$ ,

$$\theta_{j,m_{j-1}+1} < \dots < \theta_{j,m_j-1}, \quad \theta_{j,m_j} = \infty, \quad j = 1, \dots, t.$$

The underlying process in both steps is based on cumulative mechanisms. Thus, the model is called a *two-step cumulative model* or a *compound cumulative model*.

An advantage of the two-step model is that different parameters are involved in different steps. The choice between the basic sets is determined by the parameter  $\gamma_o$ , and the choice on the higher level is determined by the parameters  $\gamma_i$ . Thus, in the model the influence of the explanatory variables on the dependent variable may vary. The choice between the basic sets, e.g., the alternatives “improvement” or “no improvement,” may be influenced by different strength or even by different variables than the choice within the “improvement” set and the “no improvement” set.

### **Example 3.10: Rheumatoid arthritis** (Example 3.9, continued)

The arthritis data may serve as an illustration example. Here the basic sets are given by  $S_1 = \{1, 2\}$ ,  $S_2 = \{3\}$ ,  $S_3 = \{4, 5\}$ . Table 3.9 gives the results for the fitting of the compound logit model. The standardized estimates  $\hat{\beta}/\sqrt{\text{var}(\hat{\beta})}$  are given in brackets. The simple cumulative model yields deviance 5.86 with 3 d.f. Thus, it will not be rejected, but it does not fit very well. The cumulative model assumes that the distribution of the underlying variable for the active control treatment group is shifted with respect to the new agent treatment group. Thus, the response probabilities of all categories are changed simultaneously, resulting in increasing probability for low response categories and decreasing probability for high response categories.

The compound cumulative model in the first step models whether according to the assessment health is getting better or worse. With  $\hat{\gamma}_0 = 0.04$  the effect is negligible. On a higher level it is investigated whether there is a “conditional shift” between treatment groups within the basic sets. The effect within the basic sets seems not to be negligible for the improvement categories. From the results in Table 3.9 it follows that for the compound model  $\gamma_0$  and  $\gamma_2$  may be omitted. If, in addition, we consider the model with  $\gamma_1 = 0$  the deviance is rather large. Thus, the only effect that may not be neglected is  $\gamma_1$ . The new drug seems to have an effect only if there is an improvement. The cumulative compound model with  $\gamma_0 = \gamma_2 = 0$  has the same number of parameters as the simple cumulative model but with deviance 0.094 a much better fit. This is because the cumulative model is unable to detect that the only essential transition is from category 1 to category 2.

Model (3.3.21) is only one representative of two-step models. We get alternative models, e.g., the cumulative-sequential model if the first step is based on a cumulative model and the second step is based on a sequential model. A model of this type is appropriate if we have three basic sets where  $S_2$  is a sort of starting point and  $S_1, S_3$  are stages of change in differing directions. The arthritis example is of this type. However, for this example  $S_1, S_3$  have only two categories, and there is therefore no difference between a cumulative and a sequential model in the second step. For alternative models and examples, see Tutz (1989), and Morawitz & Tutz (1990).  $\square$

### Link Function and Design Matrix for Two-Step Models

Two-step models may be written in the form of a generalized linear model. For the cumulative (-cumulative) model the link function is given by

**Table 3.9.** Analysis of clinical trial data on rheumatoid arthritis

	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	Log-likelihood	Deviance	DF
Two-step cumulative model	0.04 (0.28)	0.55 (2.60)	-0.03 (-0.08)	-315.502	0.008	1
Two-step cumulative model with $\gamma_0 = 0$	—	0.55	-0.02	-315.541	0.087	2
Two-step cumulative model with $\gamma_0 = \gamma_2 = 0$	—	0.55	—	-315.545	0.094	3
Two-step cumulative model with $\gamma_0 = \gamma_1 =$ $= \gamma_2 = 0$	—	—	—	-319.132	7.269	4

$$\begin{aligned} g_j(\pi_1, \dots, \pi_q) &= F^{-1}(\pi_1 + \pi_2 + \dots + \pi_{m_j}), \quad j = 1, \dots, t, \\ g_{jr}(\pi_1, \dots, \pi_q) &= F^{-1} \left\{ \frac{\pi_{m_{j-1}+1} + \dots + \pi_{m_{j-1}+1+r}}{\pi_{m_{j-1}+1} + \dots + \pi_{m_j}} \right\}, \\ j &= 1, \dots, t, \quad r = 1, \dots, m_j - 1. \end{aligned}$$

The design matrix is somewhat more difficult than for the simple cumulative and sequential model. It has the form

$$\left[ \begin{array}{ccc} 1 & & x'_i \\ \ddots & \vdots & \\ 1 & x'_i & \\ & 1 & x'_i \\ & \ddots & \vdots \\ & 1 & x'_i & \\ & & 1 & x'_i \\ & & \ddots & \vdots \\ & & 1 & x'_i & \\ & & & 1 & x'_i \\ & & & \ddots & \vdots \\ & & & 1 & x'_i \end{array} \right]$$

with parameter vector

$$\beta' = (\theta_1, \dots, \theta_t, \gamma'_0, \theta_{11}, \dots, \theta_{1,m_1-1}, \gamma'_1, \dots, \theta_{t1}, \dots, \theta_{t,m_t-1}, \gamma'_t).$$

### 3.3.7 Alternative Approaches

Alternative approaches in ordinal regression have been considered and are shortly sketched in the following. Anderson (1984) introduced the so-called stereotype regression model, which in the simple one-dimensional form is given by

$$P(Y = r|x) = \frac{\exp(\beta_{r0} - \phi_r \beta' x)}{1 + \sum_{i=1}^q \exp(\beta_{i0} - \phi_i \beta' x)},$$

$r = 1, \dots, q$ . In order to get an ordered regression model the parameters  $\phi_1, \dots, \phi_k$  must fulfill the constraints

$$1 = \phi_1 > \cdots > \phi_k = 0.$$

Most often the model is estimated without imposing the constraints a priori. However, if the estimated values  $\hat{\phi}_i$  are allowed to yield the ordering of categories the order is a result of the model and not a trait of the variable considered. Then the model is not an ordinal regression model because it makes no use of the information provided by the ordering. Anderson (1984) also considered the concept of indistinguishability, meaning that response categories are indistinguishable if  $x$  is not predictive between these categories. For a similar concept in cumulative and sequential models, see Tutz (1991a). A comparison of the proportional odds model and the stereotype model is given by Holtbrügge & Schuhmacher (1991); see also Greenland (1994).

Another type of model assumes given scores for the categories of the response  $Y$ . Williams & Grizzle (1972) consider the model

$$\sum_{r=1}^k s_r P(Y = r|x) = x' \beta,$$

where  $s_1, \dots, s_k$  are given scores. Instead of using the support  $\{1, \dots, k\}$ , one may consider  $Y \in \{s_1, \dots, s_k\}$  and write the model by

$$\sum_{r=1}^k s_r P(Y = s_r|x) = x' \beta.$$

Obviously models of this type are not suited for responses that are measured on ordinal scale level. By introducing scores, a higher-scale level is assumed for the discrete response.

A third type of model that is only mentioned here is adjacent categories logits (e.g., Agresti, 1984). The model

$$\log [P(Y = r|x)/P(Y = r - 1|x)] = x' \beta_r$$

is based on the consideration of the adjacent categories  $\{r - 1, r\}$ . Logits are built locally for these adjacent categories. Another form of the model is

$$P(Y = r|Y \in \{r, r + 1\}, x) = F(x' \beta_r),$$

where  $F$  is the logistic distribution function. The latter form shows that the logistic distribution function may be substituted for any strictly monotone increasing distribution function. Moreover, it shows that it may be considered as a dichotomous response model given  $Y \in \{r, r + 1\}$ . Very similar models are used in item response theory (Masters, 1982) and are often misinterpreted as sequential process models (see Tutz, 1990, 1997). The model may also be considered the corresponding regression model arising from the row-column (RC-) association model considered by Goodman (1979, 1981a, b).

## 3.4 Statistical Inference

Since maximum likelihood estimation plays a central role, it is considered separately in Section 3.4.1. The testing of linear hypotheses, which has been considered already in the univariate case, is only mentioned in the following section. In Section 3.4.2 we consider more general goodness-of-fit statistics than for the univariate case. The family of power-divergence statistics due to Cressie & Read (1984) and the associated minimum power-divergence estimation principle together with asymptotics for the classical and the “increasing-cells” cases are given in Section 3.4.3.

### 3.4.1 Maximum Likelihood Estimation

In the following, estimation is described in the general form of multivariate exponential families (compare to Appendix A.1). The special case of the multinomial distribution is given by (3.1.7). Based on the exponential family

$$f(y_i|\theta_i, \phi, \omega_i) = \exp\left\{\frac{y'_i \theta_i - b(\theta_i)}{\phi}\right\} \omega_i + c(y_i, \phi, \omega_i)$$

for observation vectors  $y_1, \dots, y_n$ , maximum likelihood estimation may be derived in analogy to the one-dimensional case outlined in Section 2.2. The log-likelihood kernel for observation  $y_i$  is given by

$$l_i(\mu_i) = \frac{y'_i \theta_i - b(\theta_i)}{\phi} \omega_i, \quad \theta_i = \theta(\mu_i), \quad (3.4.1)$$

and the log-likelihood for the sample has the form

$$l(\beta) = \sum_{i=1}^n l_i(\mu_i).$$

Note that in the grouped case  $n$  stands for the number of groups, whereas in the ungrouped case  $n$  stands for the number of units. Using the link  $\mu_i = h(Z_i \beta)$ , the score function  $s(\beta) = \partial l / \partial \beta = \sum_{i=1}^n s_i(\beta)$  has components

$$s_i(\beta) = Z'_i D_i(\beta) \Sigma_i^{-1}(\beta) [y_i - \mu_i(\beta)], \quad (3.4.2)$$

where

$$D_i(\beta) = \frac{\partial h(\eta_i)}{\partial \eta}$$

is the derivative of  $h(\eta)$  evaluated at  $\eta_i = Z_i \beta$  and

$$\Sigma_i(\beta) = \text{cov}(y_i)$$

denotes the covariance matrix of observation  $y_i$  given parameter vector  $\beta$ . Equation (3.4.2) is a direct generalization of (2.2.2), from Chapter 2. The alternative form

$$s_i(\beta) = Z'_i W_i(\beta) \frac{\partial g(\mu_i)}{\partial \mu'} [y_i - \mu_i(\beta)]$$

makes use of the weight matrix

$$W_i(\beta) = D_i(\beta) \Sigma_i^{-1}(\beta) D'_i(\beta) = \left\{ \frac{\partial g(\mu_i)}{\partial \mu'} \Sigma_i(\beta) \frac{\partial g(\mu_i)}{\partial \mu} \right\}^{-1}, \quad (3.4.3)$$

which may be considered an approximation of the inverse of the covariance matrix of the “transformed” observation  $g(y_i)$  in cases where  $g(y_i)$  exists. The expected Fisher information is given by

$$F(\beta) = \text{cov}(s(\beta)) = \sum_{i=1}^n Z'_i W_i(\beta) Z_i,$$

which is a direct generalization of the form in Chapter 2.

In matrix notation score function and Fisher matrix have the same form as in Chapter 2, namely,

$$s(\beta) = Z' D(\beta) \Sigma^{-1}(\beta) [y - \mu(\beta)], \quad F(\beta) = Z' W(\beta) Z,$$

where  $y$  and  $\mu(\beta)$  are given by

$$y' = (y'_1, \dots, y'_n), \quad \mu(\beta)' = (\mu_1(\beta)', \dots, \mu_n(\beta)').$$

The matrices have block diagonal form

$$\Sigma(\beta) = \text{diag}(\Sigma_i(\beta)), \quad W(\beta) = \text{diag}(W_i(\beta)), \quad D(\beta) = \text{diag}(D_i(\beta)),$$

and the total design matrix is given by

$$Z = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix}.$$

These formulas are given for individual observations  $y_1, \dots, y_n$ . For grouped observations, which are more convenient for computational purposes, the formulas are the same. The only difference is that the summation is over the grouped observations  $y_1, \dots, y_g$ , where  $y_i$  is the mean over  $n_i$  observations, and  $\Sigma_i(\beta)$  is replaced by  $\Sigma_i(\beta)/n_i$ .

Under regularity assumptions (comparable to the assumptions in Chapter 2) one gets asymptotic normality of the estimate:

$$\hat{\beta} \xrightarrow{a} N(\beta, F^{-1}(\hat{\beta})).$$

That means  $\hat{\beta}$  is approximately normal with covariance matrix  $\text{cov}(\hat{\beta}) = F^{-1}(\hat{\beta})$ .

## Numerical Computation

Numerical computation of maximum likelihood estimates has already been given for univariate models in Chapter 2. In the multivariate case one merely has to substitute vectors and matrices by the multivariate versions. The working or pseudo observation vector now is given by  $\tilde{y}(\beta) = (\tilde{y}_1(\beta), \dots, \tilde{y}_n(\beta))'$ , where

$$\tilde{y}_i(\beta) = Z_i\beta + (D_i^{-1}(\beta))' [y_i - \mu_i(\beta)]$$

is an approximation for  $g(\mu_i(\beta))$ . Given the  $k$ th estimate  $\hat{\beta}^{(k)}$ , the weighted least-squares estimate

$$\hat{\beta}^{(k+1)} = \left( Z' W(\hat{\beta}^{(k)}) Z \right)^{-1} Z' W(\hat{\beta}^{(k)}) \tilde{y}(\hat{\beta}^{(k)})$$

is equivalent to the Fisher scoring iteration

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \left( Z' W(\hat{\beta}^{(k)}) Z \right)^{-1} s(\hat{\beta}^{(k)}).$$

Thus, the Fisher scoring can be viewed as an iteratively weighted least-squares estimate.

### 3.4.2 Testing and Goodness-of-Fit

#### Testing of Linear Hypotheses

The testing of linear hypotheses of the form  $H_0 : C\beta = \xi$  against  $H_1 : C\beta \neq \xi$  has already been considered in Section 2.2.2. The statistics given there may be used in the multivariate case, too. One has only to replace score functions and Fisher matrices by their multivariate versions.

#### Goodness-of-Fit Statistics

The goodness-of-fit of the models may again be checked by the Pearson statistic and the deviance. Since we are considering multinomial data, the expectation  $\mu_i$  is equivalent to the probability vector  $\pi'_i = (\pi_{i1}, \dots, \pi_{iq})$ ,  $\pi_{ik} = 1 - \pi_{i1} - \dots - \pi_{iq}$ , where  $\pi_{ir} = P(Y = r|x_i)$ . The estimate of  $\pi_i$  based on the model is denoted by  $\hat{\mu}_i = \hat{\pi}_i = (\hat{\pi}_{i1}, \dots, \hat{\pi}_{iq})'$ . For the following, data should be grouped as far as possible so that the observation vector  $y'_i = (y_{i1}, \dots, y_{iq})$  consists of relative frequencies.

The *Pearson statistic* in general is given by

$$\chi^2 = \sum_{i=1}^g (y_i - \hat{\mu}_i)' \Sigma_i^{-1}(\hat{\beta}) (y_i - \hat{\mu}_i).$$

In the case of a mult categorial response variable with multinomial distribution  $n_i y_i \sim M(n_i, \pi_i)$ ,  $\chi^2$  may be written in the more familiar form

$$\chi^2 = \sum_{i=1}^g \chi_P^2(y_i, \hat{\pi}_i),$$

where

$$\chi_P^2(y_i, \hat{\pi}_i) = n_i \sum_{j=1}^k \frac{(y_{ij} - \hat{\pi}_{ij})^2}{\hat{\pi}_{ij}}$$

is the Pearson residual for the  $i$ th (grouped) observation with  $y_{ik} = 1 - y_{i1} - \dots - y_{iq}$ ,  $\hat{\pi}_{ik} = 1 - \hat{\pi}_{i1} - \dots - \hat{\pi}_{iq}$ .

The *deviance* or likelihood ratio statistic is given by

$$D = -2 \sum_{i=1}^g \{l_i(\hat{\pi}_i) - l_i(y_i)\}.$$

For multinomial data the more familiar form is given by

$$D = 2 \sum_{i=1}^g \chi_D^2(y_i, \hat{\pi}_i),$$

where

$$\chi_D^2(y_i, \hat{\pi}_i) = n_i \sum_{j=1}^k y_{ij} \log \left( \frac{y_{ij}}{\hat{\pi}_{ij}} \right)$$

is the deviance residual. If  $y_{ij} = 0$ , the term  $y_{ij} \log(y_{ij}/\hat{\pi}_{ij})$  is set to zero.

Under “regularity conditions” including in particular increasing sample sizes  $n_i \rightarrow \infty$ ,  $i = 1, \dots, g$ , such that  $n_i/n \rightarrow \lambda_i > 0$ , one gets approximately  $\chi^2$ -distributed goodness-of-fit statistics

$$\chi^2, D \xrightarrow{a} \chi^2(g(k-1) - p)$$

where  $g$  denotes the number of groups,  $k$  is the number of response categories, and  $p$  is the number of estimated parameters. For sparse data with small  $n_i$  alternative asymptotics is appropriate (see end of next section).

### **Example 3.11: Caesarian birth study** (Example 3.1, continued)

In Example 3.1 a mult categorial logit model was fitted to the data on infection following birth by Caesarian section. The response  $Y$  had three levels (infection type I (1)/infection type II (2)/no infection (3)), and three binary, dummy-coded covariates were included in the model, namely, NOPLAN (= 1 for the caesarian was not planned), FACTOR (= 1 for risk factors were

present), and ANTIB ( $= 1$  for antibiotics were given as a prophylaxis). Before testing differences between parameters, we recall the model fitted:

$$\begin{aligned}\log \frac{P(Y = 1)}{P(Y = 3)} &= \beta_{10} + \beta_{1N} \cdot \text{NOPLAN} + \beta_{1F} \cdot \text{FACTOR} \\ &\quad + \beta_{1A} \cdot \text{ANTIB}, \\ \log \frac{P(Y = 2)}{P(Y = 3)} &= \beta_{20} + \beta_{2N} \cdot \text{NOPLAN} + \beta_{2F} \cdot \text{FACTOR} \\ &\quad + \beta_{2A} \cdot \text{ANTIB}.\end{aligned}$$

The parameter estimates from Example 3.1 were as follows:

$\hat{\beta}_{10}$	-2.621	$\hat{\beta}_{20}$	-2.560
$\hat{\beta}_{1N}$	1.174	$\hat{\beta}_{2N}$	0.996
$\hat{\beta}_{1F}$	1.829	$\hat{\beta}_{2F}$	2.195
$\hat{\beta}_{1A}$	-3.520	$\hat{\beta}_{2A}$	-3.087

Looking at the parameter estimates we might assume that ANTIB has different effects on the log-odds of infection types I and II. On the contrary, the effects of NOPLAN and FACTOR seem to be about the same for both types of infection. We want to test the latter assumption, i.e.,  $H_0 : (\beta_{1N} = \beta_{2N} \text{ and } \beta_{1F} = \beta_{2F})$  against  $H_1 : (\beta_{1N} \neq \beta_{2N} \text{ or } \beta_{1F} \neq \beta_{2F})$ . The deviance for the unrestricted model under  $H_1$  is 11.830 with  $14 - 8 = 6$  degrees of freedom. The restricted model under  $H_0$  has two fewer parameters and thus 8 degrees of freedom; the deviance is 12.162. We can calculate the likelihood ratio statistic  $\lambda$  (see Section 2.2.2) as the difference of the deviances  $\lambda = 12.162 - 11.830 = 0.8467$ . Assuming  $\lambda$  to be  $\chi^2$ -distributed with 2 *df*, we obtain a *p*-value of 0.847 for the test, which is far from being significant.

Starting off with a new model where  $\beta_{1N} = \beta_{2N}$  and  $\beta_{1F} = \beta_{2F}$ , we might now test  $H_0 : \beta_{1A} = \beta_{2A}$  against  $H_1 : \beta_{1A} \neq \beta_{2A}$ . The deviance of the former model has been 12.162 at 8 degrees of freedom. The deviance of the  $H_0$ -model is 12.506 at 9 degrees of freedom. Thus,  $\lambda = 0.3437$  and is far from being significant (the *p*-value resulting from the  $\chi^2(1)$ -distribution is 0.557).  $\square$

### 3.4.3 Power-Divergence Family\*

For categorical data a more general single-parameter family of goodness-of-fit statistics is the power-divergence family, which was introduced by Cressie & Read (1984). The power-divergence statistic with parameter  $\lambda \in \mathbb{R}$  is given by

$$S_\lambda = \sum_{i=1}^g SD_\lambda(y_i, \hat{\pi}_i),$$

where the sum of deviations over observations at fixed point  $y_i$  is given by

$$SD_\lambda(y_i, \hat{\pi}_i) = \frac{2n_i}{\lambda(\lambda+1)} \sum_{j=1}^k y_{ij} \left[ \left( \frac{y_{ij}}{\hat{\pi}_{ij}} \right)^\lambda - 1 \right], \quad (3.4.4)$$

where  $-\infty < \lambda < \infty$  is a fixed value. The term “power divergence” is derived from the fact that the divergence of the empirical distribution  $y_i$  (the vector of relative frequencies) from the hypothesized distribution  $\hat{\pi}_i$  is measured through a sum of powers of the term  $y_{ij}/\hat{\pi}_{ij}$ . The cases  $\lambda = -1$  and  $\lambda = 0$  are defined by the continuous limits  $\lambda \rightarrow -1$ ,  $\lambda \rightarrow 0$ . Many statistics that are in common use in discrete data analysis turn out to be special cases of the power-divergence family. In particular, for  $\lambda = 1$  one obtains the Pearson statistic

$$S_1 = \sum_{i=1}^g n_i \sum_{j=1}^k \frac{(y_{ij} - \hat{\pi}_{ij})^2}{\hat{\pi}_{ij}},$$

and for  $\lambda \rightarrow 0$  one obtains the likelihood ratio statistic

$$S_0 = 2 \sum_{i=1}^g n_i \sum_{j=1}^k y_{ij} \log \left( \frac{y_{ij}}{\hat{\pi}_{ij}} \right).$$

Further interesting cases are Kullback's minimum discrimination information statistic

$$S_{-1} = 2 \sum_{i=1}^g n_i \sum_{j=1}^k \hat{\pi}_{ij} \log \left( \frac{\hat{\pi}_{ij}}{y_{ij}} \right)$$

and Neyman's minimum modified  $\chi^2$ -statistic

$$S_{-2} = \sum_{i=1}^g n_i \sum_{j=1}^k \frac{(y_{ij} - \hat{\pi}_{ij})^2}{y_{ij}},$$

where in comparison to the Pearson statistic only the denominator  $\hat{\pi}_{ij}$  is substituted for  $y_{ij}$ . Moreover, the Freeman-Tukey statistic

$$S_{-1/2} = 4 \sum_{i=1}^g \sum_{j=1}^k \left( \sqrt{n_i y_{ij}} - \sqrt{n_i \hat{\pi}_{ij}} \right)^2$$

is a special case (see Read & Cressie, 1988 and Bhapkar, 1980).

The power-divergence family provides a number of goodness-of-fit statistics that are determined by the single parameter  $\lambda$ . Read & Cressie (1988) recommend  $\lambda$ -values from the interval  $[-1, 2]$ . For  $\lambda$  outside this interval the values of the statistics tend to be too small and the statistic becomes too sensitive to single-cell departures.

The maximum likelihood estimator of  $\pi_i$  is given as the probability vector  $\hat{\pi}_i$ , which maximizes the likelihood or equivalently minimizes the deviance given the model is true, i.e.,  $g(\pi_i) = Z_i\beta$  holds. Thus, maximum likelihood estimation may be considered a minimum distance estimation procedure.

Alternative minimum distance estimators like the minimum discrimination information estimator (Kullback, 1959, 1985) or Neyman's minimum  $\chi^2$ -estimator (Neyman, 1949) have been considered in the literature. These estimation procedures follow quite naturally as special cases of minimum power-divergence estimation. Let the estimates  $\hat{\pi}_i = h(Z_i\hat{\beta})$  be determined by minimization of the power-divergence statistic for fixed parameter

$$S_\lambda = \sum_{i=1}^g SD_\lambda(y_i, \hat{\pi}_i) \rightarrow \min.$$

Then special cases of minimum power-divergence estimates are the maximum likelihood estimate ( $\lambda = 0$ ), the minimum  $\chi^2$ -estimate ( $\lambda = 1$ ), Kullback's minimum discrimination information estimate ( $\lambda = -1$ ), and Neyman's minimum modified estimate ( $\lambda = -2$ ). For further distance measures, see Read & Cressie (1988) and Parr (1981).

### Asymptotic Properties under Classical “Fixed Cells” Assumptions

Classical assumptions in asymptotic theory for grouped data imply

- a fixed number of groups,
- increasing sample sizes  $n_i \rightarrow \infty$ ,  $i = 1, \dots, g$ , such that  $n_i/n \rightarrow \lambda_i$  for fixed proportions  $\lambda_i > 0$ ,  $i = 1, \dots, n$ ,
- a fixed “number of cells”  $k$  in each group,
- a fixed number of parameters.

If Birch's (1963) regularity conditions hold, minimum power-divergence estimates will be best asymptotically normally (BAN-) distributed estimates. If BAN estimates  $\hat{\pi}_i$  are used under the assumption that the model holds, the power-divergence statistic for any  $\lambda$  is asymptotically  $\chi^2$ -distributed with

$$S_\lambda \sim \chi^2(g(k-1) - p),$$

where  $g$  is the number of groups,  $k$  denotes the number of possible outcomes, and  $p$  is the number of estimated parameters (see Read & Cressie, 1988,

Appendix A6). Therefore, in this setting the statistics are asymptotically equivalent for all  $\lambda \in \mathbb{R}$ . Moreover, the asymptotic equivalence still holds under local alternatives where the limit distribution is a noncentral  $\chi^2$ -distribution.

### Sparseness and “Increasing-Cells” Asymptotics

If several explanatory variables are considered, the number of observations for a fixed explanatory variable  $x$  is often small and the usual asymptotic machinery will fail. Alternatively, under such sparseness conditions it may be assumed that with increasing sample size  $n \rightarrow \infty$  the number of groups (values of the explanatory variables) is also increasing with  $g \rightarrow \infty$ . Read & Cressie (1988, Sections 4.3 and 8.1) give a review of “increasing-cells” asymptotics for parametric models. For the product-multinomial sampling scheme, Dale (1986) investigated the asymptotic distribution of Pearson’s  $\chi^2$  and the deviance. Extensions for the power-divergence family have been given by Rojek (1989) and Osius & Rojek (1992).

The main result in “increasing-cells” asymptotics is that  $S_\lambda$  is no longer  $\chi^2$  distributed but has an asymptotic normal distribution under the null-hypothesis that the model holds. Essential conditions beside smoothness and differentiability as considered by Osius & Rojek (1992) are the increase of information, i.e., there is a positive definite limit of  $F(\beta)/n$  and the probabilities  $P(Y = r|x)$  fulfill side conditions, e.g., are all bounded away from 0 and 1. Moreover, one needs consistency of the estimator  $\hat{\beta}$ . Further conditions assume that  $g/\text{var}(S_\lambda)$  is bounded and that group sizes are increasing such that  $\sum_i n_i/g \rightarrow \infty$  or  $(\sum_i n_i^2/ng)^{1/2} \rightarrow \infty$ . Then for the power-divergence statistic we have asymptotically

$$S_\lambda \sim N(\mu_\lambda, \sigma_\lambda^2), \quad \text{resp.,} \quad T_\lambda = \frac{(S_\lambda - \mu_\lambda)}{\sigma_\lambda} \rightarrow N(0, 1),$$

where  $\lambda > 1$  is assumed. As the notation already suggests, the parameters of the limiting distribution of  $S_\lambda$  depend on  $\lambda$ . In contrast to the classical asymptotics, the asymptotic distributions for different  $\lambda$ s are not equivalent under sparseness assumptions. Although  $\mu_\lambda$  and  $\sigma_\lambda^2$  may be computed, the computational effort becomes high for  $k > 2$ . Simple expressions are available for Pearson’s statistic  $\lambda = 1$  and several special cases of increasing sample sizes (see Osius & Rojek, 1992).

## 3.5 Multivariate Models for Correlated Responses

Until now models with only *one*, though possibly mult categorial, response variable have been considered. In many applications, however, one is confronted with the truly multivariate case: A vector of correlated or clustered

response variables is observed, together with covariates, for each unit in the sample. The response vector may include repeated measurements of units on the *same* variable, as in longitudinal studies or in subsampling primary units. Examples for the latter situation are common in genetic studies where the family is the cluster but responses are given by the members of the family, in ophthalmology where two eyes form a cluster with observations taken for each eye, or in studies on forest damage, where neighborhood interactions exist between damage states of trees standing close together. In other situations the response vector consists of *different* variables, e.g., different questions in an interview. The important case of *repeated measurements* or, in other words, *longitudinal data* with many short time series, is treated in Section 6.2, with a considerable overlap of methodological development.

For approximately Gaussian variables multivariate linear models have been extensively studied and applied for a long time. Due to the lack of analytically and computationally convenient multivariate distributions for discrete or mixed discrete/continuous variables, multivariate regression analysis for non-Gaussian data becomes more difficult, and research is more recent and still in progress. One can distinguish three main approaches: *Conditional models* (Section 3.5.1), often called data-driven models, specify the conditional distribution of each component of the response vector given covariates and the remaining components. If primary interest is in effects of covariates on the responses, *marginal models* (Section 3.5.2), which specify marginal distributions, are more appropriate. The term “marginal” emphasizes that the mean response modelled is conditional only on covariates and not on previous responses or random effects. In contrast to this population-averaged approach, *random effects models* allow for cluster- or subject-specific effects. The distinction of these types of modelling is studied, e.g., by Neuhaus, Hauck & Kalbfleisch (1991), Agresti (1993b), and Diggle, Liang & Zeger (1994). Cluster-specific approaches are considered in the context of random effects models (Chapter 7), dynamic mixed models (Chapter 8), and frailty models (Chapter 9).

This section provides a short survey on work in this area. An early bibliography of methods for correlated categorical data is given in Ashby et al. (1992). More recent references are in the book by Diggle, Liang & Zeger (1994), the review articles by Agresti (1999) and Molenberghs & Lesaffre (1999), with an emphasis on longitudinal data, and in Pendergast, Gange, Newton & Lindstrom (1996). Our focus is on binary and categorical responses, but the basic ideas extend to other types of responses. In the following we do not consider models based on a multivariate structure of latent variables to which the interpretation necessarily refers. Thus, we will not consider factor analysis for categorical variables (Bartholomew, 1980) or structural equation models with explanatory variables (e.g., Muthén, 1984; Arminger & Küsters, 1985; Küsters, 1987; Arminger & Sobel, 1990). A further important class of alternative models for the investigation of the asso-

ciation structure in a given set of variables is graphical chain models (see Lauritzen & Wermuth, 1989; Wermuth & Lauritzen, 1990; Whittaker, 1990; Cox & Wermuth, 1996; Lauritzen, 1998). These models, which also allow for sets of both discrete and continuous variables, are beyond the scope of this book.

### 3.5.1 Conditional Models

#### Asymmetric Models

In many applications the components of the response vector are ordered in a way that some components are prior to other components, e.g., if they refer to events that take place earlier. Let us consider the simplest case of a response vector  $Y = (Y_1, Y_2)$  with categorical components  $Y_1 \in \{1, \dots, k_1\}$ ,  $Y_2 \in \{1, \dots, k_2\}$ . Let  $Y_2$  refer to events that may be considered conditional on  $Y_1$ . It is quite natural to model first the dependence of  $Y_1$  on the explanatory variables  $x$  and then the conditional response  $Y_2|Y_1$ . In both steps the models of Sections 3.2 and 3.3 may be used according to the scale level of  $Y_1$  and  $Y_2$ . More generally, consider  $m$  categorical responses  $Y_1, \dots, Y_m$  where  $Y_j$  depends on  $Y_1, \dots, Y_{j-1}$  but not on  $Y_{j+1}, \dots, Y_m$ . A simple case when this assumption is appropriate is repeated measurements on the same variable. However, the components of  $Y = (Y_1, \dots, Y_m)$  may also stand for different variables when the causal structure is known, e.g., when  $Y_j$  stands for events that are previous to the events coded by  $Y_{j+1}$ . In Example 3.12 we will consider the responses years in school ( $Y_1$ ) and reported happiness ( $Y_2$ ). Since the years in school are previous to the statement about happiness,  $Y_2$  may be influenced by  $Y_1$ , but not vice versa. In the following, categorical variables with  $Y_j \in \{1, \dots, k_j\}$  are considered. For the special case of binary responses  $Y_j \in \{1, 2\}$ , we use 0–1 dummies  $y_j = 2 - Y_j$ .

Models that make use of this dependence structure are based on the decomposition

$$P(Y_1, \dots, Y_m|x) = P(Y_1|x) \cdot P(Y_2|Y_1, x) \cdots P(Y_m|Y_1, \dots, Y_{m-1}, x). \quad (3.5.1)$$

Simple models arise if each component of the decomposition is specified by a generalized linear model

$$P(Y_j = r|Y_1, \dots, Y_{j-1}, x) = h_j(Z_j \beta), \quad (3.5.2)$$

where  $Z_j = Z(Y_1, \dots, Y_{j-1}, x)$  is a function of previous outcomes  $Y_1, \dots, Y_{j-1}$  and the vector of explanatory variables  $x$ . Conditional models of this type are sometimes called *data-driven*, since the response is determined by previous outcomes.

*Markov-type transition models* follow from the additional assumption  $P(Y_j = r|Y_1, \dots, Y_{j-1}, x) = P(Y_j|Y_{j-1}, x)$ . For binary outcomes a simple model assumes

$$\begin{aligned}\log(P(y_1 = 1|x)/P(y_1 = 0|x)) &= \beta_{01} + z'_j \beta_1, \\ \log\left(\frac{P(y_j = 1|y_1, \dots, y_{j-1}, x)}{P(y_j = 0|y_1, \dots, y_{j-1}, x)}\right) &= \beta_{0j} + z'_j \beta_j + y_{j-1} \gamma_j.\end{aligned}$$

In the multicategorical case the binary model may be substituted by a multinomial logit model. Repeated measurements of Markov-type models (of first or higher order) may be interpreted as transition models since the transition from previous states to the actual state is modelled. Models of this type are further discussed in the framework of categorical time series and longitudinal data (Sections 6.1.1, 6.2.1). Similar models are also considered in discrete survival analysis (Section 9.2).

*Regressive logistic models* as considered by Bonney (1987) have the form

$$\log\left(\frac{P(y_j = 1|y_1, \dots, y_{j-1}, x)}{P(y_j = 0|y_1, \dots, y_{j-1}, x)}\right) = \beta_0 + z_j \beta + \gamma_1 y_1 + \dots + \gamma_{j-1} y_{j-1}.$$

In this model the number of included previous outcomes depends on the component  $y_j$ . Thus, no Markov-type assumption is implied. If variables are multicategorical (with possibly varying numbers of categories), regressive models may be based on the modelling approaches from previous sections as building blocks.

### Example 3.12: Reported happiness

Clogg (1982) investigated the association between gender ( $x$ ), years in school ( $Y_1$ ), and reported happiness ( $Y_2$ ). The data are given in Table 3.10. Since gender and years in school are prior to the statement about happiness, the latter variable is modelled conditionally on  $Y_1$  and  $x$ . Since the dependent

**Table 3.10.** Cross classification of gender, reported happiness, and years of schooling

Gender	Reported happiness	Years of school completed			
		<12	12	13–16	≥17
Males	Not too happy	40	21	14	3
	Pretty happy	131	116	112	27
	Very happy	82	61	55	27
Females	Not too happy	62	26	12	3
	Pretty happy	155	156	95	15
	Very happy	87	127	76	15

variables are ordinal, the analysis is based on a simple regressive cumulative model. For  $Y_1 \in \{1, \dots, k_1\}$ ,  $Y_2 \in \{1, \dots, k_2\}$ , the considered model is the regressive cumulative model given by

$$\begin{aligned} P(Y_1 \leq r|x) &= F(\theta_r + x'\beta_r^{(1)}), \quad r = 1, 2, 3 \\ P(Y_2 \leq s|Y_1 = r, x) &= F(\theta_{rs} + x'\beta_s^{(2)}), \quad r = 1, \dots, 4; s = 1, 2. \end{aligned} \quad (3.5.3)$$

Here the marginal distribution of  $Y_1$  (years of school) is given by a cumulative logit model. The conditional distribution of  $Y_2$  (reported happiness) is again specified by a cumulative logit model where the previous outcome influences the thresholds  $\theta_{rs}$  of the conditional responses. Of course, the marginal distribution of  $Y_2$  does not necessarily follow a cumulative model.

Deviance and Pearson's  $\chi^2$  for the model are given by 1.55 on 6 degrees of freedom. It is natural to look if the model can be simplified. Both parts of the model assume that gender  $x$  is a category-specific variable that varies with the response category. If  $x$  is considered as a global variable for the marginal distribution of  $Y_1$ , i.e.,  $\beta_1^{(1)} = \beta_2^{(1)} = \beta_3^{(1)}$ , the fit is bad with deviance 19.54 on 8 degrees of freedom. Thus, within a logit model gender may not be seen as producing a simple shift on the response years in school. The difference between deviances for model (3.5.3) and the model restricted by  $\beta_1^{(1)} = \beta_2^{(1)} = \beta^{(1)}$  may be considered as a test for the proportionality assumption (3.3.7). With a value of 17.99 on 2 degrees of freedom, proportionality is rejected. On the other hand, the conditional modelling of  $Y_2|x$  may be simplified. The model assuming  $\beta_s^{(2)} = 0$  yields deviance 13.27 and Pearson's  $\chi^2$  is 12.20 on 8 degrees of freedom, thus showing a satisfying fit. That means given the years of school ( $Y_1$ ), gender no longer influences the reported happiness. The estimates of the latter model are given in Table 3.11.  $\square$

Models of the type (3.5.2) may be embedded into the framework of generalized linear models. Link function and design matrix follow from (3.5.1) and (3.5.2). In the general case link function and design matrix might be very complex and are not available in standard programs. However, if  $k_1 = \dots = k_m$ , the local models (3.5.2) may often be chosen to have a link function  $h$  that does not depend on  $j$ . By choosing an appropriate design matrix (filled up with zeros) the decomposition (3.5.1) allows us to use standard software, where  $Y_1, Y_2|Y_1, \dots, Y_m|Y_1, \dots, Y_{m-1}$  (always given  $x$ ) are treated as separate observations (see, e.g., Bonney, 1987).

## Symmetric Models

If there is no natural ordering of the components of the response vector, or if one does not want to use this ordering, models that treat response components in a symmetric way are more sensible. An example from ophthalmology is visual impairment data from the Baltimore Eye Survey (Tielsch, Sommer,

**Table 3.11.** Estimates for the cross classification of gender, reported happiness and years of schooling. The estimates in Table 3.11 refer to model (3.5.3), where  $\beta_j^{(2)}=0$ .

	Estimate	Standard deviation	p-value
$\theta_1$	-0.545	0.053	0.000
$\theta_2$	0.841	0.056	0.000
$\theta_3$	2.794	0.112	0.000
$\beta_1^{(1)}$	0.001	0.053	0.984
$\beta_2^{(1)}$	-0.201	0.056	0.000
$\beta_3^{(1)}$	-0.388	0.112	0.000
$\theta_{11}$	-1.495	0.109	0.000
$\theta_{12}$	0.831	0.092	0.000
$\theta_{21}$	-2.281	0.153	0.000
$\theta_{22}$	0.528	0.091	0.000
$\theta_{31}$	-2.564	0.203	0.000
$\theta_{32}$	0.575	0.109	0.000
$\theta_{41}$	-2.639	0.422	0.000
$\theta_{42}$	0.133	0.211	0.527

Katz & Ezrene, 1989), analyzed in Liang, Zeger & Qaqish (1992); see Example 1.6 in Chapter 1 and Example 3.13 at the end of Section 3.5.2. Binary observations of visual impairment for both ( $m = 2$ ) eyes are taken for more than 5000 persons, together with demographic covariates such as age, race, gender, etc. Conditional models are useful if conditional distributions or moments of one response component given the others are of interest, e.g., for the purpose of prediction. If the main scientific objective is to analyze effects of covariates on responses, marginal models (Section 3.5.2) rather than conditional ones are useful.

For the following, attention is restricted to binary response components  $y_1, \dots, y_m$ , but ideas generalize to multicategorical and other types of variables. Symmetric *conditional* models can be developed by specification of the conditional distributions

$$P(y_j = 1|y_k, k \neq j; x_j) \quad (3.5.4)$$

of one response given the others. For binary responses such conditional distributions uniquely determine the joint distribution. Logistic regression models are a natural choice for the conditional distributions in (3.5.4): Starting from a log-linear model including all interactions between  $y_1, \dots,$

$\dots, y_m$ , one arrives at logit models where the linear predictor includes covariates  $x_j$ , main effects  $y_k$ ,  $k \neq j$ , and second- (and higher-) order interaction effects such as  $y_k y_l$ ,  $k, l \neq j$ , in additive form; see Zeger & Liang (1989). In developing methods for spatial statistics, Besag (1974) introduced the auto-logistic model having this form, although his proposal did not include covariates. A parsimonious class of conditional logistic models, treating the individual binary variables symmetrically, is considered by Qu, Williams, Beck & Goormastic (1987):

$$\pi_j = P(y_j = 1 | y_k, k \neq j; x_j) = h(\alpha(w_j; \theta) + x'_j \beta_j), \quad (3.5.5)$$

where  $h$  is the logistic function and  $\alpha$  is an arbitrary function of a parameter  $\theta$  and the sum  $w_j = \sum_{k \neq j} y_k$  of the conditioning  $k \neq j$ . In model (3.5.5), the “location parameter”  $\alpha$  depends on the conditioning  $y$ ’s, whereas covariate effects are kept constant. For the case of two components  $(y_1, y_2)$ , the sums  $w_1, w_2$  reduce to  $y_2, y_1$ , respectively. Then the simplest choice is a logistic model including the conditioning response as a further covariate:

$$\pi_j = P(y_j = 1 | y_k, k \neq j; x_j) = h(\theta_0 + \theta_1 y_k + x'_j \beta_j), \quad j, k = 1, 2. \quad (3.5.6)$$

Other choices for  $\alpha(w; \theta)$  are discussed in Qu, Williams, Beck & Goormastic (1987) and Conolly & Liang (1988). The joint density  $P(y_1, \dots, y_m; x_1, \dots, [1]x_m)$  can be derived from (3.5.5); however, it involves a normalizing constant, which is a complicated function of the unknown parameters  $\theta$  and  $\beta$ ; see Prentice (1988) and Rosner (1984). Full likelihood estimation may therefore become computationally cumbersome. Conolly & Liang (1988) propose a quasi-likelihood approach, with an “independence working” quasi-likelihood and quasi-score function for each cluster  $i = 1, \dots, n$ :

$$\begin{aligned} L_i(\beta, \theta) &= \prod_{j=1}^m \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}}, \\ s_i(\beta, \theta) &= \sum_{j=1}^m \frac{\partial \pi_{ij}}{\partial(\beta, \theta)} \sigma_{ij}^{-2} (y_{ij} - \pi_{ij}(\beta, \theta)), \end{aligned}$$

where  $y_i = (y_{i1}, \dots, y_{ij}, \dots, y_{im})$  are the responses in cluster  $i$ ,  $\pi_{ij}(\beta, \theta) = P(y_{ij} = 1 | \cdot)$  is defined by (3.5.5), and  $\sigma_{ij}^2 = \pi_{ij}(\beta, \theta)(1 - \pi_{ij}(\beta, \theta))$ . If  $\alpha$  is a linear function of  $\theta$ , as, e.g., in (3.5.6), this is the common form of the score function for  $m$  independent binary responses. Setting

$$\begin{aligned} M_i &= \left( \frac{\partial \pi_{i1}}{\partial(\beta, \theta)}, \dots, \frac{\partial \pi_{im}}{\partial(\beta, \theta)} \right), & \Sigma_i &= \text{diag}(\sigma_{i1}^2, \dots, \sigma_{im}^2), \\ && \pi_i &= (\pi_{i1}, \dots, \pi_{im})', \end{aligned}$$

we have

$$s_i(\beta, \theta) = M_i \Sigma_i^{-1} (y_i - \pi_i),$$

which is a multivariate extension of the (quasi-) score function in Section 2.3.1. The roots  $(\hat{\beta}, \hat{\theta})$  of the resulting generalized estimation equation

$$s(\beta, \theta) = \sum_{i=1}^n s_i(\beta, \theta) = 0$$

are, under regularity assumptions, consistent and asymptotically normal:

$$(\hat{\beta}, \hat{\theta}) \stackrel{a}{\sim} N((\beta, \theta), \hat{F}^{-1} \hat{V} \hat{F}^{-1}),$$

with

$$\hat{F} = \sum_{i=1}^n \hat{M}_i \hat{\Sigma}_i^{-1} \hat{M}_i, \quad \hat{V} = \sum_{i=1}^n \hat{s}_i \hat{s}_i',$$

where “ $\hat{\cdot}$ ” means evaluation at  $(\beta, \theta)$ .

Two drawbacks of conditional models should be kept in mind: First, by construction, they measure the effect of covariates on a binary component  $y_i$  having already accounted for the effect of other responses  $y_k$ ,  $k \neq j$ . If regression of covariates on the response is the scientific focus, this will often condition away covariate effects. A second drawback is that interpretation of effects depends on the dimension of  $y$ : Excluding components of  $y$  leads to different kinds of “marginal” distributions for the remaining components; the models are not “reproducible.” Therefore, conditional models are only meaningful if the dimension of  $y$  is the same for all observations. Both disadvantages can be avoided by the following marginal modelling approach.

### 3.5.2 Marginal Models

In many situations the primary scientific objective is to analyze the *marginal mean* of the responses given the covariates. The association between responses is often of secondary interest. In the Baltimore Eye Survey study on visual impairment (Example 1.6 and Example 3.13), the primary goal is to identify the influence of demographic variables, such as age, race, and education, on visual impairment. Data are available on both eyes, and the association between both responses has to be taken into account in a correct analysis. However, one is not primarily interested in this association or its relation with demographic variables. In a second application we consider a study on forest damage in Bavaria (Example 3.14). The damage state of trees is determined from aerial pictures and is recorded in ordered categories. Pictures are taken at predetermined points of the terrain, and the trees standing close to these points form clusters. Again, the scientific question of interest is to analyze the influence of covariates, which characterize the

stand, on the state of trees, whereas spatial correlation caused by neighborhood interactions is regarded as a nuisance. Similar situations occur in other applications, e.g., twin studies and dentistry.

Marginal models were first proposed by Liang & Zeger (1986), Zeger & Liang (1986) in the closely related context of longitudinal data with many short time series (see also Section 6.2.2). Their modelling and estimation approach is based on generalized estimating equations (GEE) and has subsequently been modified and generalized in various ways including suggestions for alternative modelling and fitting of association structures, extensions to correlated categorical responses and likelihood-based approaches. Reviews are given in Liang, Zeger & Qaqish (1992), Fitzmaurice, Laird & Rotnitzky (1993), Diggle, Liang & Zeger (1994), Agresti (1999), and Molenberghs & Lesaffre (1999), mainly from the viewpoint of longitudinal data analysis.

The focus in this section is on modelling and estimating marginal means of correlated binary and categorical data by first-order generalized estimating equations (GEE1). Extensions to other types of responses and likelihood-based approaches are briefly outlined. We first consider the simpler case of marginal models for binary and other types of univariate responses.

### Marginal Models for Correlated Univariate Responses

In marginal models, the effect of covariates on responses and the association between responses is modelled separately. This is in contrast to conditional models, which are completely defined by the conditional mean structure. Let  $y'_i = (y_{i1}, \dots, y_{im_i})$  be the vector of responses and  $x'_i = (x'_{i1}, \dots, x'_{im_i})$  the vector of covariates from cluster  $i$ ,  $i = 1, \dots, n$ . Here the term “cluster” is used instead of observation to emphasize that in many situations the components of  $y_i$  are correlated observations on the same type of variable. Generally, “cluster size” or the size  $m_i$  of the vector  $y'_i = (y_{i1}, \dots, y_{im_i})$  may depend on the cluster. This is in contrast to conditional models: Marginal models are “reproducible” and parameter interpretation does not depend on the dimension of  $y$ . In the Baltimore Eye Survey example (Example 1.6), when measuring the performance of two eyes one always has  $m_i = m = 2$ . However, if measurements are made in families of various sizes,  $m_i$  may stand for the family size: Each cluster (family) involves a possibly differing number of correlated responses. The covariate components may vary across units within a cluster or may be constant within clusters, i.e.,  $x_{i1} = \dots = x_{im_i}$ . Specification of marginal models is as follows.

- (i) The *marginal means* of  $y_{ij}$ ,  $j = 1, \dots, m_i$ , are assumed correctly specified by common univariate response models

$$\mu_{ij}(\beta) = E(y_{ij}|x_{ij}) = h(z'_{ij}\beta), \quad (3.5.7)$$

where  $h$  is a response function, e.g., a logit function, and  $z_{ij}$  is an appropriate design vector.

- (ii) The *marginal variance* of  $y_{ij}$  is specified as a function of  $\mu_{ij}$ ,

$$\sigma_{ij}^2 = \text{var}(y_{ij}|x_{ij}) = v(\mu_{ij})\phi, \quad (3.5.8)$$

where  $v$  is a known variance function.

- (iii) The *correlation* between  $y_{ij}$  and  $y_{ik}$  is a function of the marginal means  $\mu_{ij} = \mu_{ij}(\beta)$ ,  $\mu_{ik} = \mu_{ik}(\beta)$ , and perhaps of additional association parameters  $\alpha$ ,

$$\text{corr}(y_{ij}, y_{ik}) = c(\mu_{ij}, \mu_{ik}; \alpha), \quad (3.5.9)$$

with a known function  $c$ .

Responses from different clusters are assumed to be independent. Note that parameters  $\beta$  and  $\alpha$  are the same for each cluster. Therefore, marginal models are appropriate for analyzing “population-averaged” effects. Marginal regression coefficients describe the effect of covariates on the marginal means or average responses and have the same interpretation as in Section 2.1. This is in contrast to random effects models (Chapter 7), where effects vary from cluster to cluster.

An important feature of marginal models is the following: Marginal effects  $\beta$  can be consistently estimated even if the correlation function is incorrectly specified. This corresponds to quasi-likelihood models in Section 2.3.1, where the variance function can also be incorrectly specified, while the parameters  $\beta$  can still be estimated consistently, however, with some loss of efficiency, as long as the mean function is correctly specified.

Since the primary scientific objective is often the regression relationship, it is natural to spend more time in correct specification of the marginal mean structure (3.5.7) than in the covariance structure. It is therefore assumed that (3.5.7) is correctly specified, while  $c(\mu_{ij}, \mu_{ik}; \alpha)$  is a *working correlation* for the association between  $y_{ij}$  and  $y_{ik}$ . Together with the variance function (3.5.8), one obtains a *working covariance matrix*

$$\text{cov}(y_i) = \Sigma_i(\beta, \alpha),$$

which depends on  $\beta$  and  $\alpha$ . It may additionally depend on  $\phi$  through (3.5.8); however, we suppress  $\phi$  notationally. Generally this working covariance matrix will be different from the *true* covariance matrix.

Two main approaches for specifying working correlations or covariances have been considered in the literature. Liang & Zeger (1986) and Prentice (1988) use correlations as a measure for associations: The variance structure (3.5.8) is supplemented by a *working correlation matrix*  $R_i(\alpha)$ , so that the working covariance matrix is of the form

$$\Sigma_i(\beta, \alpha) = C_i^{1/2}(\beta) R_i(\alpha) C_i^{1/2}(\beta),$$

where

$$C_i(\beta) = \text{diag}[\text{var}(y_{ij}|x_{ij})] = \text{diag}[\sigma_{i1}^2, \dots, \sigma_{im_i}^2].$$

Various choices for  $R_i(\alpha)$  are suggested by Liang & Zeger (1986) within a longitudinal data context; see also Chapter 6. The simplest choice is a *working independence model*, i.e.,

$$R_i(\alpha) = I,$$

where  $I$  is the identity matrix. Another convenient choice is the *equicorrelation model* with

$$\text{corr}(y_{ij}, y_{ik}) = \alpha$$

for all  $j \neq k$ . If enough data are available one may leave  $R(\alpha)$  *completely unspecified*, i.e.,  $\alpha$  has the elements  $\alpha_{jk} = \text{corr}(y_{ij}, y_{ik})$ ,  $j < k$ .

For *binary responses* the odds ratio is an alternative measure of association that is easier to interpret and has some desirable properties. The odds ratio parameterization is suggested by Lipsitz, Laird & Harrington (1991). The odds ratio for components  $y_{ij}, y_{ik}$ ,  $j, k = 1, \dots, m$ ,  $j \neq k$ , is defined by

$$\gamma_{ijk} = \frac{P(y_{ij} = 1, y_{ik} = 1)}{P(y_{ij} = 1, y_{ik} = 0)} \frac{P(y_{ij} = 0, y_{ik} = 0)}{P(y_{ij} = 0, y_{ik} = 1)}.$$

From the relation (e.g., Lipsitz, Laird & Harrington, 1991; Liang, Zeger & Qaqish, 1992)

$$P(y_{ij} = y_{ik} = 1) = E(y_{ij}y_{ik}) \\ = \begin{cases} \frac{1 - (\pi_{ij} + \pi_{ik})(1 - \gamma_{ijk}) - s(\pi_{ij}, \pi_{ik}, \gamma_{ijk})}{2(\gamma_{ijk} - 1)}, & \gamma_{ijk} \neq 1 \\ \pi_{ij}\pi_{ik}, & \gamma_{ijk} = 1, \end{cases} \quad (3.5.10)$$

with  $s(\pi_{ij}, \pi_{ik}, \gamma_{ijk}) = [\{1 - (\pi_{ij} + \pi_{ik})(1 - \gamma_{ijk})\}^2 - 4(\gamma_{ijk} - 1)\gamma_{ijk}\pi_{ij}\pi_{ik}]^{1/2}$ , it is seen that the covariance matrix of  $y_i = (y_1, \dots, y_m)'$  can be expressed as a function of marginal probabilities and odds ratios.

One advantage of the odds ratio is that it is not constrained by the means. In contrast, correlations are constrained by means in the form:

$$\text{corr}(y_{ij}, y_{ik}) = \frac{P(y_{ij} = y_{ik} = 1) - \pi_{ij}\pi_{ik}}{[\pi_{ij}(1 - \pi_{ij})\pi_{ik}(1 - \pi_{ik})]^{1/2}},$$

where  $P(y_{ij} = y_{ik} = 1)$  is constrained by  $\max(0, \pi_{ij} + \pi_{ik} - 1) \leq P(y_{ij} = y_{ik} = 1) \leq \min(\pi_{ij}, \pi_{ik})$ . This may considerably narrow the range for admissible correlations. However, if correlation between pairs is weak, the correlation parameterization will do as well.

To reduce the number of parameters, odds ratios are parameterized by appropriate working assumptions. The simplest model is  $\gamma_{ijk} = \gamma$  for all

$i, j, k$ , analogous to an equicorrelation assumption. Alternatively, odds ratio might be functions of covariates. A common parameterization is a log-linear model  $\log \gamma_{ijk} = \alpha' w_{ijk}$  as in Example 3.13. Generally we write  $\gamma_{ijk} = \gamma_{ijk}(\alpha)$ , where  $\alpha$  is a vector of parameters to be estimated. Using relation (3.5.10), the covariance matrix of  $y_i$  can be expressed as a function  $\Sigma_i(\beta, \alpha)$  of  $\beta$  and  $\alpha$ .

The choice of the working covariance should be a compromise between simplicity and loss of efficiency due to incorrect specification. If association is only a nuisance or of secondary interest, simple working models will suffice. The study of McDonald (1993) favors the independence model whenever association is a nuisance. But Fitzmaurice (1995) showed that efficiency of the estimate depends on both, the correlation and the covariate design. He demonstrated that the assumption of independence can lead to a considerable loss of efficiency when the responses are strongly correlated and the design includes a within-cluster covariate. The coefficient associated with that covariate may be estimated rather inefficiently. If association is also of interest, a good specification of odds ratios will be desirable. If in doubt, one will try several models.

The following are some examples of marginal models.

*Continuous responses:*

- (i)  $\mu_{ij}(\beta) = E(y_{ij}|x_{ij}) = z'_{ij}\beta$ ,
- (ii)  $\text{var}(y_{ij}|x_{ij}) = \phi = \sigma^2$ ,
- (iii)  $\text{corr}(y_{ij}, y_{ik}) = \alpha_{jk}$ .

*Binary responses:*

- (i)  $\mu_{ij}(\beta) = \pi_{ij}(\beta) = P(y_{ij} = 1|x_{ij})$ , logit  $\pi_{ij}(\beta) = z'_{ij}\beta$ ,
- (ii)  $\text{var}(y_{ij}|x_{ij}) = \pi_{ij}(\beta)(1 - \pi_{ij}(\beta))$ ,
- (iii)  $\text{corr}(y_{ij}, y_{ik}) = 0$  (independence working assumption) or  $\gamma_{ijk} = \alpha$  (equal odds ratios).

*Count data:*

- (i)  $\log(\mu_{ij}(\beta)) = \log(E(y_{ij}|x_{ij})) = z'_{ij}\beta$ ,
- (ii)  $\text{var}(y_{ij}|x_{ij}) = \mu_{ij}(\beta)\phi$ ,
- (iii)  $\text{corr}(y_{ij}, y_{ik}) = \alpha$  (equicorrelation).

## The Generalized Estimating Approach for Statistical Inference

Let  $y'_i = (y_{i1}, \dots, y_{im_i})$ ,  $x'_i = (x'_{i1}, \dots, x'_{im_i})$  be the observations for cluster  $i$ ,  $i = 1, \dots, n$ , and  $\mu_i(\beta)' = (\mu_{i1}(\beta_1), \dots, \mu_{im_i}(\beta_{m_i}))$ ,  $\Sigma_i(\beta, \alpha)$  be the corresponding marginal mean vectors and working covariance matrices for  $y_i$  as described above. In linear models for Gaussian responses, ML estimation is no problem, since specification of means and covariances determines the likelihood. This is not the case with discrete data, where specification

of the likelihood requires additional assumptions about higher order moments. Even when additional assumptions are made, ML estimation can become computationally cumbersome or even intractable; see the discussion on likelihood-based methods at the end of this section. Therefore, a generalized estimation approach is proposed. Keeping the association parameters  $\alpha$  as well as  $\phi$ , if present, fixed for the moment, the generalized estimating equation (GEE) for effect  $\beta$  is

$$s_\beta(\beta, \alpha) = \sum_{i=1}^n Z_i' D_i(\beta) \Sigma_i^{-1}(\beta, \alpha) (y_i - \mu_i(\beta)) = 0, \quad (3.5.11)$$

with design matrices

$$Z_i' = (z_{i1}, \dots, z_{im_i})$$

and diagonal matrices  $D_i(\beta) = \text{diag}(D_{ij}(\beta))$ ,  $D_{ij}(\beta) = \partial h / \partial \eta_{ij}$  evaluated at  $\eta_{ij} = z_{ij}' \beta$ .

Equation (3.5.11) is a multivariate version of the GEE (2.3.2) in Section 2.3.1, with a correctly specified mean  $E(y_i | x_i) = \pi_i(\beta)$  and a possibly misspecified covariance matrix  $\text{cov}(y_i | x_i) = \Sigma_i(\beta, \alpha)$ . Note that  $\Sigma_i(\beta, \alpha)$ , and therefore  $s_\beta(\beta, \alpha)$ , may also depend on an unknown scale or overdispersion parameter  $\phi$ , though it has been suppressed notationally. A GEE estimate  $\hat{\beta}$  is computed by iterating between a modified Fisher scoring algorithm for  $\hat{\beta}$  and estimation of  $\alpha$  (and  $\phi$ ). Given current estimates  $\hat{\alpha}$  (and  $\hat{\phi}$ ), the GEE (3.5.11) for  $\hat{\beta}$  is solved by the iterations

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + (\hat{F}^{(k)})^{-1} \hat{s}_\beta^{(k)},$$

with  $\hat{s}_\beta^{(k)}$  as the (quasi-) score function evaluated at current parameter values and

$$\hat{F}^{(k)} = \sum_{i=1}^n Z_i' D_i(\hat{\beta}^{(k)}) \Sigma_i^{-1}(\hat{\beta}^{(k)}, \hat{\alpha}) D_i(\hat{\beta}^{(k)}) Z_i.$$

For the case of an independence working model,  $R_i(\alpha) = I$ , the working covariance is  $\Sigma_i(\beta) = \text{diag}(\sigma_{i1}^2, \dots, \sigma_{im_i}^2)$ . The GEE has then the usual form of ML scoring equations, as if observations were independent, and no association parameter  $\alpha$  has to be estimated jointly with  $\beta$ . An additional scale or overdispersion parameter  $\phi$  can be estimated by the method of moments as in Section 2.2. Note that the iteration does not depend on  $\phi$ . Therefore,  $\phi$  can be estimated after the last iteration.

If unknown association parameters  $\alpha$  are present, several ways of estimating them have been suggested. Liang & Zeger (1986) use a method of moments based on Pearson residuals

$$\hat{r}_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{(v(\hat{\mu}_{ij}))^{\frac{1}{2}}}.$$

The dispersion parameter is estimated consistently by

$$\hat{\phi} = \frac{1}{N-p} \sum_{i=1}^n \sum_{j=1}^{m_i} \hat{r}_{ij}^2, \quad N = \sum_{i=1}^n m_i.$$

Estimation of  $\alpha$  depends on the choice of  $R_i(\alpha)$ . For exchangeable correlation,

$$\hat{\alpha} = \frac{1}{\hat{\phi}\{\sum_{i=1}^n \frac{1}{2}m_i(m_i - 1) - p\}} \sum_{i=1}^n \sum_{k>j} \hat{r}_{ik} \hat{r}_{ij}.$$

An unspecified  $R = R(\alpha)$  can be estimated by

$$\frac{1}{n\hat{\phi}} \sum_{i=1}^n C_i^{-\frac{1}{2}} (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' C_i^{-\frac{1}{2}}$$

if cluster sizes are all equal to  $m$  and small compared to  $n$ . Cycling between Fisher scoring steps for  $\beta$  and estimation of  $\alpha$  and  $\phi$  leads to a consistent estimation of  $\beta$ . (Note that  $\hat{\alpha}$  need only be consistent for *some*  $\alpha$ , not a *true*  $\alpha^0$ , which does not exist for working correlations.)

Alternatively,  $\alpha$  (and possibly  $\phi$ ) can be estimated by simultaneously solving an additional estimating equation, as first suggested by Prentice (1988). For each cluster define the vector  $w_i = (w_{i12}, w_{i13}, \dots, w_{im_{i-1}m_i})'$  of products  $w_{ijk} = (y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik})$  and the vector  $v_i = E(w_i)$  of corresponding expectations. The second GEE for  $\alpha$  is then

$$s_\alpha(\beta, \alpha) = \sum_{i=1}^n \frac{\partial v_i}{\partial \alpha} B_i^{-1} (w_i - v_i) = 0. \quad (3.5.12)$$

The matrix  $B_i$  is a further working covariance matrix for  $w_i$ . “Optimal” specification of  $B_i$  requires additional assumptions about higher order moments. Simpler choices are  $B_i = I$  and  $B_i = \text{diag}(\text{var}(w_{i12}), \text{var}(w_{i13}), \dots)$ . With binary responses, we obtain

$$\nu_{ijk} = E(w_{ijk}) = E(y_{ij}y_{ik}) - \pi_{ij}\pi_{ik},$$

with  $E(y_{ij}y_{ik})$  given by (3.5.10) and

$$\text{var}(w_{ijk}) = \pi_{ij}(1 - \pi_{ij}) - \nu_{ijk}^2 + \nu_{ijk}(1 - 2\pi_{ik})(1 - 2\pi_{ij}).$$

Note that Prentice (1988) suggests working with residuals  $r_{ijk}$  instead of  $w_{ijk}$ . GEE estimates  $\hat{\beta}$  and  $\hat{\alpha}$  are obtained by cycling between Fisher scoring equations for  $s_\beta$  and  $s_\alpha$ .

Under regularity assumptions,  $\hat{\beta}$  is consistent and asymptotically normal,

$$\hat{\beta} \xrightarrow{a} N(\beta, F^{-1} V F^{-1}),$$

with

$$F = \sum_{i=1}^n Z_i' D_i \Sigma_i^{-1} D_i Z_i, \quad V = \sum_{i=1}^n Z_i' D_i \Sigma_i^{-1} S_i \Sigma_i^{-1} D_i Z_i,$$

where  $S_i$  is the *true* covariance matrix of  $y_i$ .  $F^{-1}$  and  $V$  can be estimated by replacing  $\beta$ ,  $\alpha$ , and  $\phi$  by their estimates and  $S_i$  by  $(y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)'$ , i.e.,  $\text{cov}(\hat{\beta})$  is approximated by the “sandwich matrix”

$$\hat{A} \approx \hat{F}^{-1} \left\{ \sum_{i=1}^n Z_i' \hat{D}_i \Sigma_i^{-1} (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' \hat{\Sigma}_i^{-1} \hat{D}_i Z_i \right\} \hat{F}^{-1}.$$

The asymptotic normality result for  $\hat{\beta}$  is a multivariate extension of the corresponding result for quasi-likelihood estimation in Section 2.3.1. Keeping  $\alpha$  (and  $\phi$ ) fixed, it follows from the form of the quasi-score function  $s_\beta(\beta, \alpha)$  in (3.5.11) that

$$\text{cov } s_\beta(\beta, \alpha) = V = \sum_{i=1}^n Z_i' D_i \Sigma_i^{-1} S_i \Sigma_i^{-1} D_i Z_i.$$

Approximate normality  $s_\beta(\beta, \alpha) \xrightarrow{a} N(0, V)$  follows from a central limit theorem. The same Taylor expansion as for the MLE (Section 2.2) and the QLME (Section 2.3.1) gives

$$s_\beta(\beta, \alpha) \xrightarrow{a} H(\beta)(\hat{\beta} - \beta),$$

with  $H(\beta) = -\partial s_\beta(\beta, \alpha)/\partial\beta'$ . Replacing  $H(\beta)$  by its expectation  $F(\beta)$  and solving for  $\hat{\beta} - \beta$  gives

$$\hat{\beta} - \beta \xrightarrow{a} F^{-1}(\beta) s_\beta(\beta, \alpha),$$

and the approximate normality result follows from approximate normality of  $s_\beta(\beta, \alpha)$ .

It is important again to note that  $\hat{\beta}$  is consistent and asymptotically normal even if the covariance matrix  $y_i$  has been misspecified. Only correct specification of the marginal means is needed.

Compared to the nominal covariance matrix  $\hat{F}$  of an independence model, the adjusted sandwich covariance matrix  $\hat{F}^{-1} \hat{V} \hat{F}^{-1}$  protects against assessing false standard deviations for estimated effects  $\hat{\beta}$ . For the case of positively correlated observations within clusters, standard deviations of effects for covariates that are constant within clusters, as in the visual impairment example, are generally too small if they are based on  $\hat{F}$  instead of  $\hat{F}^{-1} \hat{V} \hat{F}^{-1}$ . On the other side, standard deviations are too large for effects

of subject-specific covariates. If observations within clusters are negatively correlated, “too large” and “too small” have to be exchanged.

### Example 3.13: Visual impairment

For 5199 individuals bivariate binary responses were observed, indicating whether or not an eye was visually impaired (VI). Covariates include age in years centered at 60 (A), race (R: 0–white, 1–black), and education in years centered at 9 (E). Table 3.12 gives the data of VI of right and left eyes for race  $\times$  age combinations. The age-specific risks for whites increase

**Table 3.12.** Visual impairment data (from Liang et al., 1992)

Visual impairment	White				Black				Total	
	Age				Age					
	40–50	51–60	61–70	70+	40–50	51–60	61–70	70+		
Left eye										
Yes	15	24	42	139	29	38	50	85	422	
No	617	557	789	673	750	574	473	344	4777	
Right eye										
Yes	19	25	48	146	31	37	49	93	448	
No	613	556	783	666	748	575	474	226	4751	

clearly with age. Blacks are more affected than whites, and the discrepancy increases with age. Risks are quite similar for both eyes, as one might expect.

The main objective is to analyze the influence of age and race on visual impairment controlling for education, which serves as a surrogate for socioeconomic status. Therefore, Liang, Zeger & Qaqish (1992) fitted a marginal logistic model

$$\log \left( \frac{\pi_1}{1 - \pi_1} \right) = \log \left( \frac{\pi_2}{1 - \pi_2} \right) = \beta_0 + \beta_1 A + \beta_2 A^2 + \beta_3 R + \beta_4 (A \times R) + \beta_5 (A^2 \times R) + \beta_6 E$$

that an eye was visually impaired. For the odds ratio, a log-linear regression model

$$\log \gamma_{12} = \alpha_0 + \alpha_1 R$$

was used. For comparison this is contrasted with a corresponding conditional model, given the response of one eye, and an ordinary logistic model assuming all binary observations are independent.

**Table 3.13.** Estimation results for visual impairment data

Covariate	Marginal m.	Conditional m.	Logistic m.
Intercept	-2.83 (-37)	-3.16 (-43)	-2.82
Age in years (centered at 60)	0.049 (7.1)	0.038 (6.6)	0.049
$(\text{Age} - 60)^2$	0.0018 (5.3)	0.0012 (4.5)	0.0018
Race (0=white; 1=black)	0.33 (3.2)	0.099 (0.9)	0.33
Age $\times$ Race interaction	0.00066 (0.07)	-0.0039 (-0.5)	0.0011
$(\text{Age} - 60)^2$ by race interaction	-0.0010 (-2.1)	-0.00075 (-2.0)	-0.0011
Education in years (centered at 9)	-0.060 (-0.35)	-0.045 (-3.8)	-0.059
Log odds ratio intercept	2.3 (8.7)	2.3 (13)	—
Log odds ratio race	0.54 (1.3)	0.55 (2.2)	—

Table entries are: parameter (parameter/standard error).

For the conditional model the linear predictor included, in addition to that of the marginal model above, the response for the other eye and its interaction with race.

Table 3.12 reproduces parameter estimates for the marginal model (with  $\alpha_0, \alpha_1$  estimated by a second GEE), the conditional model, and an ordinary logistic model. Estimates of the latter and the marginal model are quite close, while effects in the conditional model tend to be attenuated compared to the marginal effects because of explicit control for the other eye. For example, the marginal effect 0.33 and the *t*-value 3.2 for race indicate sig-

nificant influence of race on visual impairment, while the conditional effect 0.099 and the  $t$ -value 0.9 do not. In this epidemiological context, marginal modelling is more sensible.  $\square$

### Marginal Models for Correlated Categorical Responses

Consider now the situation where categorical responses  $Y_{ij}$ ,  $j = 1, \dots, m_i$ , are observed in each cluster  $i$ . For simplicity we assume the same number  $k$  of categories for each response. Coding responses  $Y_{ij}$  by a vector  $y_{ij} = (y_{ij1}, \dots, y_{ijq})'$  of  $q = k - 1$  dummy variables, with  $k$  as reference category, data are given in the same notation as before by responses  $y'_i = (y'_{i1}, \dots, y'_{im_i})$  and covariates  $x'_i = (x'_{i1}, \dots, x'_{im_i})$  in each cluster  $i$ ,  $i = 1, \dots, n$ . Marginal categorical response models can be defined and fitted with the same basic principles as for binary responses: Specify marginal means or response probabilities for  $Y_{ij}$  by one of the multinomial response models of Section 3.2 or 3.3, and supplement it by some working association model. Depending on the type of response – nominal or ordinal, the association model and the method of estimation, a variety of approaches is conceivable. Like most authors, we focus on ordinal responses, as in our application to forest damage (Example 3.14). Marginal GEE models for ordinal responses are described by Miller, Davis & Landis (1993), using pairwise correlations as measures of association, and by Williamson, Kim & Lipsitz (1995), Fahrmeir & Pfitscher (1996) and Heagerty & Zeger (1996), based on global cross-ratio parameterizations. They are defined as follows:

- (i) The vector of marginal means or responses probabilities of  $Y_{ij}$  is assumed correctly specified by an ordinal response model

$$\pi_{ij}(\beta) = (\pi_{ij1}(\beta), \dots, \pi_{ijq}(\beta))' = h(Z_{ij}\beta),$$

with  $\pi_{ijr} = P(Y_{ij} = r | x_{ij}) = P(y_{ijr} = 1 | x_{ij})$ , and the response function  $h$  and the design matrix  $Z_{ij}$  chosen to represent one of the ordinal models in Section 3.2, e.g., a cumulative logit model.

- (ii) The marginal covariance function of  $y_{ij}$  is given by

$$\Sigma_{ij} = \text{cov}(y_{ij} | x_{ij}) = \text{diag}(\pi_{ij}) - \pi_{ij}\pi'_{ij},$$

i.e., the covariance matrix of a multinomial random variable.

- (iii) Association between  $Y_{ij}$  and  $Y_{ik}$  is modelled by a working correlation matrix  $R_i$  or by global cross-ratios.

The choice  $R_i = I$  corresponds to independent working assumptions. The working matrix for exchangeable correlations is

$$R_i(\alpha) = \begin{bmatrix} I & Q & \dots & Q \\ Q' & I & \dots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ Q' & \dots & \dots & I \end{bmatrix},$$

where the  $(q \times q)$ -matrix  $Q$  generally contains unknown parameters  $\alpha$  that can be estimated by a method of moments, involving Pearson residuals as in the binary case.

Global cross-ratios (GCR), for each pair of categories  $l$  and  $m$  of  $Y_{ij}$  and  $Y_{ik}$ , are defined as

$$\gamma_{ijk}(l, m) = \frac{P(Y_{ij} \leq l, Y_{ik} \leq m) P(Y_{ij} > l, Y_{ik} > m)}{P(Y_{ij} > l, Y_{ik} \leq m) P(Y_{ij} \leq l, Y_{ik} > m)}.$$

Global cross-ratios are modelled log-linearly as in the binary case, e.g., by

$$\log(\gamma_{ijk}(l, m)) = \alpha_{lm},$$

or by a regression model including covariate effects. For similar reasons as in the binary case, we prefer modelling associations by GCRs.

Marginal models for nominal multicategorical response can be specified along similar lines, but with some modifications: Obviously, marginal response probabilities have to obey one of the nominal response models in Section 3.1. Second, since GCRs are reasonably defined only for ordinal responses, other types of odds ratios like local odds ratios (see Agresti, 1990, Ch. 2) have to be used as measures of association.

Regression and association parameters are estimated by a multivariate extension of the GEE approach. Setting  $\mu_i(\beta) \equiv \pi_i(\beta)$ , the GEE  $s_\beta(\beta, \alpha) = 0$  for  $\beta$  is again given by (3.5.11), with obvious multivariate modifications in the definitions of  $Z_i$ ,  $D_i(\beta)$ , and  $\Sigma_i(\beta, \alpha)$ . For example, if a working correlation matrix  $R_i(\alpha)$  is specified, then

$$\Sigma_i(\beta, \alpha) = C_i^{1/2}(\beta) R_i(\alpha) C_i^{1/2}(\beta),$$

where  $C_i(\beta) = \text{diag}(\Sigma_{i1}, \dots, \Sigma_{im_i})$  is now block-diagonal. If GCRs are used, off-diagonal elements  $\text{cov}(y_{ijl}, y_{ikm}) = E(y_{ijl}y_{ikm} - \pi_{ijl}\pi_{ikm})$  of  $\Sigma_i(\beta, \alpha)$  have to be computed. This can be done by using the relationship (3.5.10) for cumulative indicator variables  $\tilde{y}_{ijl} = I(Y_{ij} \leq l)$  and associated cumulative probabilities  $\pi_{ijl} = P(\tilde{y}_{ijl} = 1)$ ; see Dale (1986), Fahrmeir & Pritscher (1996), or Gieger (1998) for details. To estimate association parameters  $\alpha$  jointly with  $\beta$ , a multivariate version of the second GEE  $s_\alpha(\beta, \alpha) = 0$ , (3.5.12) is introduced. Now the vector  $w_i$  contains all products  $w_{ijk}(l, m) = (y_{ijl} - \pi_{ijl})(y_{ikm} - \pi_{ikm})$  and  $\nu_i = \nu_i(\beta, \alpha)$  is the vector of corresponding expectations  $\nu_{ijk}(l, m) = E(y_{ijl}y_{ikm}) - \pi_{ijl}\pi_{ikm}$ .

The simplest choice for the weight matrix is again  $B_i = I$ . Other choices are  $B_i = \text{diag}[\text{var}(w_{ijk}(l, m))]$  or a block-diagonal  $B_i$ , proposed by Heagerty & Zeger (1996) and Heumann (1997, Section 8.2.5). Cycling between Fisher scoring iterations for solving the two GEEs until convergence gives estimates  $\hat{\beta}$  and  $\hat{\alpha}$ , and  $\hat{\beta}$  is consistent and approximately normal again, with the multivariate version of the sandwich matrix as an approximation of  $\text{cov}(\hat{\beta})$ .

In our experience a good strategy is to start with an independent working assumption and to compare standard errors obtained from “naive” ML estimation and from the “robust” sandwich matrix. If differences are significant, a more complex GEE approach based on GCRs as association measure is worthwhile.

### **Example 3.14: Forest damage**

We illustrate the GEE approach for correlated ordinal responses with an analysis of forest damage data from a survey conducted in a forest district in the northeastern part of Bavaria. A primary goal was to get information about damage state of spruce, the dominating tree species in this area, and about the influence of other variates on it. To determine damage state, infra-red colored aerial pictures of the area were taken by helicopters, and the degree of defoliation of tree tops was used as an indicator of damage state. On infrared-colored pictures healthy green tops without defoliation have an intensive red color, while strongly damaged trees appear without any red coloring. In this way damage state was classified into five categories, indicating the degree of defoliation.

A sample of spruce trees was obtained in the following way: Aerial pictures were related to a digital terrain model with a  $250\text{ m} \times 250\text{ m}$  grid. Taking grid points as primary units, the damage state of the eight spruce trees next to a grid point was determined. The whole data set of the survey consists of clusters with 8 trees for each of 771 grid points, giving a total sample of 6168 trees.

For regression analysis the response variable damage D was further condensed into three ordered categories: strong, distinct, and light. Figure 3.2 shows corresponding relative frequencies in the sample. All covariates are categorical and due to the survey plan constant within clusters, i.e., trees, belonging to the same cluster, have the same covariate values.

In preliminary exploratory analysis (see Fahrmeir & Pritscher, 1996, for more details), the following covariates were found to be most influential.

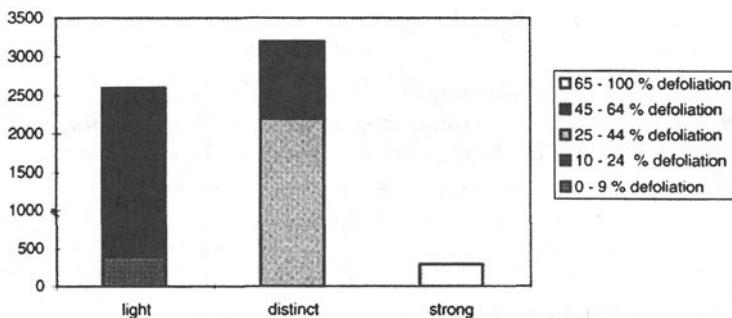
C canopy density: very low (1), low (2), medium (3), high (4).

M mixture of stand: coniferous (1) or mixed (2).

U utilization method: second commercial thinning (1), first commercial thinning (2), precommercial thinning (3). This variable can be considered as a surrogate for age of the stand, since utilization methods change with age.

- S site: bedrock or nonfertile granite weathering (1), fertile granite weathering (2), gneiss weathering (3), soil with water surplus (4).  
 A altitude: 500–600 m (1), 601–650 m (2), 651–700m (3), 701–750 m (4), 751–900 m (5).

In addition to main effects of the covariates, it is reasonable to consider possible interaction effects, for example, between utilization method and canopy density or site and altitude.



**Figure 3.2.** Frequency distribution of damage classes.

To analyze the data we assume a cumulative logit model with  $q = 2$  relevant categories for the marginal probabilities of light or distinct damage D. In addition to the thresholds the model includes five main effects (canopy density C, mixture stand M, utilization method U, site S, and altitude A given in effect-coding and an interaction term  $U \times C$  for interactions between U and C. For each covariate, the last category is taken as the reference category. In the working correlation matrix for exchangeable association, Q is a  $2 \times 2$  matrix with four unknown elements  $q_{11}, q_{12}, q_{21}, q_{22}$ . Correspondingly, the cross-ratios are parameterized by  $\log \gamma_{ijk}(l, m) = \alpha_{lm}$ ,  $l, m = 1, 2$ . No covariate effects are included, since covariates give no information on association in our application. Table 3.14 gives parameter estimates and standard errors for the independence model. Additionally, naive standard errors, corresponding to unmodified maximum likelihood estimation for the independence model, are given in the second column. Estimates for other interaction effects have been omitted since they turned out to be nonsignificant. Comparing naive standard errors obtained from unmodified ML estimation to robust standard errors, we see that naive standard errors are distinctly smaller. Therefore, naive ML estimation will lead to overinterpretation of results. In particular, interactions effects, mixture of stands \* altitude and site \* altitude, would be falsely considered as significant. Therefore, they are omitted in Table 3.14.

**Table 3.14.** Parameter estimates and standard errors

Covariates	Working association model				Exchangeable		Gl. cross-ratio	
	Independence		Correlation		Estim.	SE (robust)	Estim.	
	Estim.	SE (naive)	SE	Estim.			SE	Estim.
<b>Threshold parameters</b>								
	-0.635	0.067	0.117	-0.566	0.078	-0.576	0.097	
	2.765	0.082	0.134	2.047	0.081	2.513	0.109	
<b>Canopy density</b>								
very low	-0.685	0.157	0.313	-0.648	0.199	-0.644	0.252	
low	0.182	0.082	0.137	0.159	0.097	0.160	0.117	
medium	0.373	0.070	0.122	0.353	0.082	0.357	0.101	
high	0.129	0.104	0.166	0.136	0.106	0.128	0.136	
<b>Mixture of stand</b>								
coniferous	0.104	0.029	0.037	0.094	0.029	0.101	0.035	
mixed	-0.104	0.029	0.037	-0.094	0.029	-0.101	0.035	
<b>Utilization Method</b>								
2. commercial thinning	-0.544	0.078	0.134	-0.438	0.085	-0.514	0.109	
1. commercial thinning	0.262	0.078	0.124	0.181	0.083	0.211	0.104	
precommercial thinning	0.281	0.106	0.216	0.256	0.140	0.303	0.175	
<b>Site</b>								
bedrock	-0.082	0.049	0.067	-0.047	0.052	-0.061	0.062	
granite weathering	-0.134	0.060	0.080	-0.114	0.057	-0.122	0.072	
gneiss weathering	-0.021	0.068	0.084	-0.029	0.064	-0.042	0.079	
soil with water surplus	0.238	0.059	0.073	0.191	0.058	0.224	0.068	
<b>Altitude</b>								
500–600 m	0.342	0.086	0.106	0.296	0.081	0.318	0.100	
601–650 m	-0.106	0.059	0.078	-0.078	0.058	-0.093	0.071	
651–700 m	0.013	0.054	0.072	-0.015	0.056	-0.001	0.066	
701–750 m	-0.034	0.068	0.088	0.000	0.068	-0.009	0.080	
750–900 m	-0.215	0.079	0.107	-0.204	0.079	-0.215	0.097	
<b>Utilization Method * Canopy density</b>								
2.c.th./very low	0.600	0.170	0.328	0.599	0.208	0.556	0.265	
1.c.th./very low	0.465	0.198	0.330	0.330	0.218	0.360	0.272	
p.c.th./very low	-1.066	0.283	0.615	-0.930	0.389	-0.916	0.492	
2.c.th./low	0.044	0.099	0.158	0.001	0.108	0.025	0.133	
1.c.th./low	0.019	0.097	0.146	-0.007	0.105	0.011	0.126	
p.c.th./low	-0.064	0.148	0.260	0.006	0.185	-0.036	0.221	
2.c.th./medium	-0.305	0.091	0.147	-0.254	0.096	-0.257	0.121	
1.c.th./medium	-0.091	0.087	0.136	-0.031	0.095	-0.048	0.116	
p.c.th./medium	0.396	0.117	0.226	0.286	0.152	0.305	0.186	
2.c.th./high	-0.339	0.164	0.255	-0.346	0.158	-0.324	0.206	
1.c.th./high	-0.394	0.135	0.198	-0.291	0.130	-0.323	0.165	
p.c.th./high	0.733	0.142	0.251	0.637	0.165	0.647	0.206	

Parameter estimates obtained from the GEE model with exchangeable correlations are smaller in absolute value. Standard errors are smaller than for the independence model, but still mostly larger than naive standard errors, leading to similar conclusions for significance or nonsignificance of effects. For the GCR model, parameter estimates are often quite near to those

for the exchangeable correlation model, while standard errors are higher but still smaller than for the independence model due to improved efficiency.

In this application, all three models lead to very similar conclusions and the following interpretations. Due to effect coding, parameter estimates for the categories of each covariate sum up to zero. High (positive) values indicate a positive influence on minor damage, while low (negative) values show a positive influence on damage.

*Mixture of stand:* In the particular survey area, the probability for low damage is significantly higher for coniferous stands compared to mixed stands. For example, the estimated effect 0.101 in the cross-ratio model leads to an odds ratio increase of  $1.1 = \exp(.101)$  for low damage in coniferous stands.

*Site:* As to be expected, soil with water surplus is significantly beneficial for low damage of spruce. There is no significant difference between the influence of the remaining categories of site.

*Altitude:* Altitudes below 600 m are most favorable; on the other hand, probability of high damage increases significantly above 750 m. Other altitudes have no significant influence.

*Utilization method:* The main effect of utilization shows that stands with second commercial thinning corresponding to higher age have a clearly higher portion of damaged spruce. Although the difference between first and precommercial thinning is not significant, there is evidence of increasing damage with an increase in age.

*Canopy density:* Stands with very low density have distinctly increased probability for high damage, while medium canopy density is clearly beneficial.

*Utilization \* Canopy density:* Since there are significant interactions between categories of utilization method and canopy density, main and interaction effects should be interpreted together by adding them up. For example, stands with precommercial thinning with high or medium canopy density have a clearly lower probability of damage than indicated by the main effect alone. On the other side, the rather rare combination of precommercial thinning and very low canopy density is very unfavorable. In contrast, stands of higher age (first and second commercial thinning) are less affected by very low canopy density. Going through the other interactions, one may summarize as follows: Lower canopy density becomes more favorable with increasing age.  $\square$

### Likelihood-based Inference for Marginal Models

The GEE approach described above is not likelihood-based and does not require specification of the joint distribution of the multivariate response vectors  $y_i$ . The difficulties with likelihood-based or full ML inference is due to the lack of convenient and easily tractable joint distributions for multivariate non-Gaussian, in particular discrete responses. To specify joint distributions and likelihoods, additional assumptions about higher order moments are required. Even when additional assumptions are made, the likelihood can become computationally inconvenient or intractable for higher-dimensional responses. To illustrate some of the issues, we discuss the case of binary responses, following the presentation in Fitzmaurice, Laird & Rotnitzky (1993).

To simplify notation, we assume that cluster sizes are all equal to  $m$ . Then the joint distribution of response  $y_i = (y_{i1}, \dots, y_{im})'$  is multinomial with a  $2^m$  probability vector  $\pi_i = \{\pi_{ij_1, \dots, j_m} = P(y_{i1} = j_1, \dots, y_{im} = j_m)\}$ , where  $j_1, \dots, j_m$  are 0 or 1. The marginal response probabilities  $\pi_{ij}$  for  $y_{ij}$  are obtained from these joint probabilities by summation. The fully parameterized distribution has  $2^m - 1$  parameters, since probabilities are constrained to sum up to 1. For marginal models,  $m$  of these parameters are the marginal means or probabilities  $\pi_{ij}$ ,  $j = 1, \dots, m$ . Obviously, the remaining  $2^m - m - 1$  parameters can be specified in many possible ways. One particular parameterization is Bahadur's representation, describing the joint distribution in terms of marginal means and correlations. A drawback, however, is that the marginal correlations must satisfy constraints determined by the marginal probabilities.

An alternative is to start from the log-linear specification

$$f(y_i, \psi_i, \omega_i) = \exp\{\psi_i'y_i + \omega_i'w_i - A(\psi_i, \omega_i)\},$$

where  $w_i = (y_{i1}y_{i2}, \dots, y_{im-1}y_{im}, \dots, y_{i1}y_{i2}, \dots, y_{im})'$  is a  $2^m - m - 1$  vector of two- and higher-way cross-products of responses  $\psi_i = (\psi_{i1}, \dots, \psi_{im})'$  and  $\omega_i = (\omega_{i12}, \dots, \omega_{im-1,m}, \dots, \omega_{i12\dots m})'$  are vectors of canonical parameters, and  $A(\psi_i, \omega_i)$  is a normalizing constant. The parameters in  $\psi_i$  have interpretations as conditional probabilities,  $\psi_{ij} = \text{logit}\{P(y_{ij} = 1 | y_{ik} = 0, k \neq j)\}$ , while the parameters of  $\omega_i$  can be interpreted in terms of log-conditional odds ratios.

The log-linear specification is a flexible general form, from which various models can be obtained by constraining or transforming parts of the “canonical” parameters  $\psi_i$  and  $\omega_i$ . Zhao & Prentice (1990) suggest basing inference on a “quadratic exponential family” model, setting three- and higher-way association parameters to zero. They make a one-to-one transformation of the remaining conditional parameters to first and second moment parameters  $\mu_i$  and  $v_i$  (as defined above) to obtain joint likelihood equations in a form called GEE2 by Liang, Zeger & Qaqish (1992). A serious drawback of this method is that in contrast to the GEE approach, consistency of  $\beta$

requires correct specification for the means *and* the pairwise marginal correlations. Fitzmaurice & Laird (1993) avoid this problem by a one-to-one transformation from  $(\psi_i, \omega_i)$  to marginal means  $\mu_i$  and conditional association parameters. Their approach can be viewed as a special case of partly exponential models introduced by Zhao, Prentice & Self (1992). Generally, any partition of canonical parameters into subsets of lower and higher order can be transformed into corresponding subsets of marginal lower and canonical higher order association parameters; see Heumann (1997) and Gieger (1998) for a detailed discussion in the broader context of multicategorical response. Glonek (1996), Glonek & McCullagh (1996), and Molenberghs & Lesaffre (1992, 1994, 1999) employ likelihood-based inference through a full marginal parameterization. However, correct specification of lower *and* higher order parameters is needed with these approaches for consistent estimation.

An important class of simultaneous models has been introduced by Lang & Agresti (1994). These models have the generalized log-linear matrix form  $C \log A\mu = X\beta$ , where  $\mu$  is the vector of expected frequencies,  $A$  and  $C$  are transformation matrices which may define marginal distributions by containing ones and zeros, and  $X$ ,  $\beta$  have the usual meaning of design matrix and unknown parameters. This general form allows simultaneous modelling of joint and marginal distributions. Thus, for example, in longitudinal studies the association between adjacent time points may be modelled by a first- or second-order Markov model while simultaneously the marginal distribution for each variable is of interest. The models may be fitted by maximum likelihood methods subject to constraints (see also Lang, 1996).

An important issue is computational feasibility. Full ML estimation requires recovery of a  $2^m$  joint probability vector. Computations grow exponentially with  $m$  and quickly become impractical. This problem is even more serious for multicategorical responses, where already a GEE2 approach can become computationally infeasible.

## 3.6 Notes and Further Reading

In recent years the construction of multivariate distributions by means of copulas has gained much interest. Song (2000) presents multivariate dispersion models, generated from Gaussian copula, which are marginally closed. Multivariate binary or Poisson models are included as a special case. Moreover, a close relationship between likelihood regression analysis based on these models and the GEE approach is established.

### Bayesian Inference

The Bayesian approach requires that a statistical model specifies a genuine likelihood for the data given the parameters. Therefore, in principle,

Bayesian analysis is possible for all multivariate response models of this chapter with the exception of GEE methods for marginal models. Whenever the likelihood of the response  $Y$  can be calculated directly, as for the nominal multinomial logistic model or the cumulative logistic model, inference can be carried out essentially with the same techniques as described in Section 2.3.2. An attractive alternative is to base Bayesian analysis on underlying linear models for latent variables or utilities  $U$  as in Sections 3.2 and 3.3. This is particularly attractive for latent Gaussian models, leading to various versions of probit models. Albert & Chib (1993) suggested a fully Bayesian approach for the case of independent ordered and unordered categorical response. They propose a Gibbs sampling scheme, where sampling from full conditionals for  $U$  given  $Y$  requires draws from truncated normal distributions, whereas sampling from full conditionals for parameters given  $U$  can be carried out as in Gaussian models. This idea has been generalized to more complex data situations, including correlated data as in Sections 3.5; see Chib (1999) and Chen & Dey (1999) for recent surveys. The book by Johnson & Albert (1999) is also a valuable reference for Bayesian approaches.

# 4

# Selecting and Checking Models

Fitting data by a certain generalized linear model means choosing appropriate forms for the predictor, the link function, and the exponential family or variance function. In the previous chapters Pearson's  $\chi^2$ , the deviance and, in the multinomial case, the power-divergence family were introduced as general goodness-of-fit statistics. This chapter considers more specific tools to select and check models. Section 4.1 deals with variable selection, i.e., which variables should be included in the linear predictor. Diagnostic methods based on the hat matrix and on residuals are described in Section 4.2, and Section 4.3 covers general misspecification tests, such as Hausman-type tests and tests for nonnested models. We do not treat tests for specific directions, such as testing the correct form of the link function by embedding it in a broader parametric class of link functions. A survey of tests of this type is contained in Chapter 11.4 of McCullagh & Nelder (1989). In addition to the methods of this chapter, nonparametric approaches, as in Chapter 5, may also be used to check the adequacy of certain parametric forms.

## 4.1 Variable Selection

Regression analysis is often used in situations where there is a catalog of many potentially important covariates. Variable selection methods aim at determining submodels with a moderate number of parameters that still fit the data adequately. For normal linear models, various selection methods ranging from traditional stepwise approaches to “all-subsets” methods (e.g., Furnival & Wilson, 1974) are available. The relative merits and drawbacks of stepwise procedures, lower computational costs versus suboptimality, have been mainly discussed within the linear regression context (e.g., Hocking, 1976; Seeber, 1977; Miller, 1984, 1989). For non-normal regression models, Lawless & Singhal (1978, 1987) developed efficient screening and all-subsets

procedures based on various selection criteria. Yet stepwise variable selection is a useful data analysis tool within these more complex models: First, maximum likelihood estimates have to be computed iteratively, so that all subset selection by likelihood-based criteria like AIC can become computationally critical in situations with large covariate sets. Second, the problem of nonexistence of estimates, which is usually negligible for classical linear models, becomes much more serious for some non-normal models involving a large number of parameters, in particular for models with multicategorical variables. In this case, a complete model search, or even a stepwise backward selection, may break down from the beginning since the full model cannot be fitted, whereas stepwise forward-backward selection (e.g., Fahrmeir & Frost, 1992) is still applicable. Stepwise selection can also be useful for choosing a good initial model in refined selection procedures (Edwards & Havranek, 1987) or in combination with multiple test procedures.

Though we have generalized linear models in mind, the methods apply to any parametric regression models with log-likelihoods of the general form

$$l(\beta; \theta) = \sum_{i=1}^n l_i(\eta_i; \theta),$$

where  $l_i(\eta_i; \theta)$  is the log-likelihood contribution of  $(y_i, x_i)$ . The predictor

$$\eta_i = Z_i \beta$$

is linear in the parameters  $\beta$ , with design matrices  $Z_i = Z_i(x_i)$  as functions of the covariate vector, in particular  $Z_i = (1, x_i)$  for univariate models. The additional parameter vector  $\theta$  may, e.g., contain an unknown dispersion parameter  $\phi$ , but it may also be void. Variable selection is then understood in the sense of selecting associated sets of subvectors of  $\beta$ . If  $\theta$  is nonvoid, it is always included in the model. It may be estimated by a method of moments, as, e.g., the dispersion parameter  $\phi$  in generalized linear models, or by simultaneous ML estimation together with  $\beta$ .

### 4.1.1 Selection Criteria

A submodel corresponds to a subvector, say  $\beta_1$ , of  $\beta$ . Without loss of generality, let  $\beta$  be partitioned as  $(\beta_1, \beta_2)$ . The adequacy of a certain submodel can be formally tested by

$$H_0 : \beta_2 = 0, \quad \beta_1, \theta \text{ unrestricted}$$

against

$$H_1 : \beta_1, \beta_2, \theta \text{ unrestricted},$$

where  $H_1$  stands for the “full” model, or, as in stepwise procedures, for some supermodel. Let the score function  $s(\beta)$  for  $\beta$ , the expected (or observed) information matrix  $F(\beta)$ , and its inverse  $A(\beta) = F^{-1}(\beta)$  be partitioned in conformity with the partitioning of  $\beta$ :

$$s = \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}, \quad F = \begin{bmatrix} F_{11} & F_{12} \\ F'_{12} & F_{22} \end{bmatrix}, \quad A = \begin{bmatrix} A_{11} & A_{12} \\ A'_{12} & A_{22} \end{bmatrix}.$$

In the following  $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$  denotes the unrestricted MLE under  $H_1$ , whereas  $\tilde{\beta} = (\beta_1, 0)$  is the MLE under the restriction of  $H_0$ . Correspondingly,  $\hat{\theta}$  and  $\tilde{\theta}$  are consistent estimators of  $\theta$  under  $H_0$  and  $H_1$ . The three common test statistics (compare with Section 2.2.2) are the *likelihood ratio statistic*

$$\lambda = -2\{l(\tilde{\beta}_1, 0, \tilde{\theta}) - l(\hat{\beta}_1, \hat{\beta}_2, \hat{\theta})\},$$

the *Wald statistic*

$$w = \hat{\beta}'_2 \hat{A}_{22}^{-1} \hat{\beta}_2, \tag{4.1.1}$$

and the *score statistic*

$$u = \tilde{s}'_2 \tilde{A}_{22} \tilde{s}_2, \tag{4.1.2}$$

where “ $\hat{\cdot}$ ” (resp., “ $\tilde{\cdot}$ ”) means evaluation at  $(\hat{\beta}, \hat{\theta})$ , resp.,  $(\tilde{\beta}_1, 0, \tilde{\theta})$ . As a fourth test statistic one may consider a *modified likelihood ratio statistic*

$$\lambda_m = -2\{l(\bar{\beta}_1, 0, \bar{\theta}) - l(\hat{\beta}_1, \hat{\beta}_2, \hat{\theta})\},$$

where  $\bar{\beta}_1$  is a first-order approximation to  $\tilde{\beta}$ , given by

$$\bar{\beta}_1 = \hat{\beta}_1 - \hat{A}'_{12} \hat{A}_{22}^{-1} \hat{\beta}_2. \tag{4.1.3}$$

Under  $H_0$ , all test statistics are asymptotically  $\chi^2(r)$ ,  $r = \dim(\beta_2)$ , provided that appropriate regularity conditions hold. Generally, the likelihood ratio statistic  $\lambda$  is preferred for a comparably small number of covariates and moderate sample size. For larger sample sizes, the test statistics tend to

agree closely, and it is reasonable to use the statistics  $w$ ,  $u$ , and  $\lambda_m$  because they are easier and faster to compute. This can be of considerable importance in selection procedures for larger models.

The Wald statistic  $w$  is of advantage if the unrestricted MLE has already been computed, as in all-subsets procedures or stepwise backward procedures. The statistic  $\lambda_m$  can provide a better approximation to  $\lambda$ , but requires more computation. The score statistic  $u$  is quite useful if a restricted model has been fitted and shall be tested against a more complex model, e.g., as in a stepwise forward selection procedure.

To compare models with different numbers of parameters in an all-subset search, Akaike's information criterion (AIC)

$$AIC = -2l(\tilde{\beta}_1, 0, \tilde{\theta}) + 2(r + s), \quad (4.1.4)$$

$r = \dim(\tilde{\beta}_1)$ , and  $s = \dim(\tilde{\theta})$ , penalizes models with many parameters. Alternatively, one may consider Schwarz' criterion (SC) (Schwarz, 1978). Since the overall maximum log-likelihood  $\hat{l}$  is a constant, one can replace  $-2l(\tilde{\beta}_1, 0, \tilde{\theta})$  in (4.1.4) by  $w$ ,  $u$ , or  $\lambda_m$  and use the resulting criteria for model comparison.

### 4.1.2 Selection Procedures

#### All-Subsets Selection

Lawless & Singhal (1987) adopt and extend Furnival and Wilson's algorithm to generalized linear models, making efficient use of symmetric "sweeps" on (inverse) information matrices. Sweeps are the necessary algebraic manipulations to compute the relevant parts of (inverse) information matrices  $F$  and  $A$  for submodels defined by  $\beta_2 = 0$ . For example, sweeping on  $A_{22}$  in the augmented tableau

$$\begin{pmatrix} A_{11} & A_{12} & \beta_1 \\ A'_{12} & A_{22} & \beta_2 \\ \beta'_1 & \beta'_2 & 0 \end{pmatrix}$$

produces  $A_{22}^{-1}$ , and it transforms 0 to the Wald statistic  $w$  in (4.1.1). As a by-product one obtains the one-step approximation  $\bar{\beta}_1$  to  $\tilde{\beta}_1$  in (4.1.3). Submodels are tested against the full model, and they are labeled as good if the chosen test statistic ( $\lambda$ ,  $w$ , or  $\lambda_m$ ) has a small value. They propose two options for screening models: (i) The "best"  $m$  models for fixed  $r$  (dimension of  $\beta_1$ ) are determined by the models having the  $m$  smallest values of the specified test statistics. (ii) The "best"  $m$  models overall (i.e., of any dimension) are those with the  $m$  smallest modified AIC-values.

## Stepwise Backward and Forward Selection

For large models and large data sets, stepwise procedures are a useful additional tool for the reasons mentioned at the beginning of this section. Stepwise selection based on Wald and score tests is described in Fahrmeir & Frost (1992).

Backward steps use Wald tests, starting from a maximal model  $M$ , e.g., the “full” model containing all covariates. It is tested against a set  $U$  of admissible submodels (e.g., all submodels of  $M$  obtained by removing parameters associated with certain variables or all hierarchical submodels). Performing sweeps as in the tableau above, one may compute Wald statistics and  $p$ -values  $\alpha_L$  of the models  $L \in U$ . Then the submodel  $L_0$  with

$$\alpha_{L_0} = \max_{L \in U} \alpha_L \quad \text{and} \quad \alpha_{L_0} > \alpha_{out},$$

$\alpha_{out}$  a prechosen exclusion level, is selected. Then the MLE  $\hat{\beta}$  for  $L_0$  and its covariance matrix are computed, using the first approximation (4.1.3) as a starting value for the iterations. Since  $\bar{\beta}$  is already a good approximation to  $\hat{\beta}$ , only one or two iterations are required, or they may even be omitted for some of the selection steps. Then  $L_0$  is redefined as  $M$ , and the backward procedure is applied iteratively to admissible submodels. It terminates if there is no  $p$ -value  $\alpha_L > \alpha_{out}$ .

Forward steps use score tests starting from a minimal model  $L$ , e.g., the model without covariates. It is tested against a list  $V$  of admissible supermodels, obtained, e.g., by adding covariates not contained in  $L$ . Score statistics and  $p$ -values  $\alpha_M$  for all  $M \in V$  are computed by efficient sweeps; see Fahrmeir & Frost (1992) for details. The supermodel  $M_0 \in V$  with

$$\alpha_{M_0} = \min_{M \in V} \alpha_M \quad \text{and} \quad \alpha_{M_0} < \alpha_{in},$$

$\alpha_{in}$  a prechosen inclusion level, is selected. Then  $M_0$  is redefined as  $L$ , and forward steps are iterated until there is no  $p$ -value smaller than  $\alpha_{in}$ .

Combining forward steps with backward steps in the usual way as in linear regression or discriminant analysis yields forward/backward model selection algorithms. Pure forward or forward/backward stepwise selection is quite useful in situations with a large number of variables (parameters) and with comparably sparse data. Typical examples are categorical response models or log-linear models for high-dimensional contingency tables. If many categorical covariates together with higher-order interaction terms are contained in the “full” model  $M$  and if data are sparse in some categories of the response variable or in some of the cells, a finite ML estimate will probably not exist, i.e., ML iterations will not converge and no “full” model can be fitted. Thus, all-subset selection or even stepwise pure backward selection, which both need the fitted full model, break down.

Selection procedures based on score and Wald tests can also easily be adapted to quasi-likelihood models, replacing  $w$  and  $u$  by their modified counterparts  $w_m$  and  $u_m$  defined in Chapter 2, Section 2.3.1, whereas the likelihood ratio statistic may not even be properly defined.

**Example 4.1: Credit-scoring** (Examples 2.2, 2.5, continued)

This data set, which is contained in Fahrmeir & Hamerle (1984, see p. 334 ff. and p. 751 ff.), has been analyzed by Kredler (1984) and is reanalyzed for the purpose of comparison. The sample consists of 1000 consumers' credits, the binary response variable "creditability" and 20 covariates, which are assumed to influence creditability. Based on a binary logit model with  $y = \text{creditworthy}$  as the reference category and with main effects of the covariates only, pure backward selection (exclusion level 0.05) and pure forward selection (inclusion level 0.05) eliminate the same nine variables as in Fahrmeir & Hamerle (1984, p. 284).

A full subset selection among  $2^{20} = 1048576$  possible models becomes computationally infeasible. To compare all-subset selection with stepwise selection, we confined analysis to a logit model with the seven covariates X1, X2, X4, X5, X6, X7 and X8, which are given in Example 2.2. The estimation results for the full model containing all main effects are given in Table 4.1. With exclusion/inclusion levels of 0.05, backward and forward selection eliminate the variables X4 and X7. All-subset selections with the Wald-, log-likelihood-, Akaike's AIC-, and Schwarz' SC-criterion lead to the same best model (see Frost, 1991, for more details).  $\square$

**Table 4.1.** Logit model fit to credit-scoring data

Variable	ML estimate	p-value
Grand mean	- 0.188121	0.614225
X1[1]	+ 0.634647	0.000321
X1[2]	- 1.317027	0.000000
X3	+ 0.035027	0.000008
X4	+ 0.000032	0.330905
X5[1]	- 0.988369	0.000093
X6[1]	- 0.474398	0.003113
X7[1]	+ 0.223511	0.311471
X8[1]	- 0.385423	0.078926

## 4.2 Diagnostics

Diagnostic tools for assessing the fit of a classical linear regression model are in common use. They are designed to detect discrepancies between the data and the fitted values as well as discrepancies between a few data and the rest. Most of these techniques are based on graphical presentations of residuals, the hat matrix, and case deletion measures (see, e.g., Belsley, Kuh & Welsch, 1980, and Cook & Weisberg, 1982). As an example consider a plot of residuals against fitted values. If the pattern of residuals increases or decreases with the fitted values, some model departure, e.g., a wrong variance function and/or omitted covariates, is to be suspected. On the other hand, if only a few residuals are far from the rest, the corresponding data points represent outliers and should be checked further. More precisely, one must examine whether fit or parameter estimates greatly change by the omission of an outlying observation, and if so, whether the outlier is caused by an extreme response and/or extreme covariate values. Beside displays of residuals, plots of the hat matrix and case deletion measures provide information about this question. Note, however, that a model departure indicated by some diagnostic plot may be caused by a faulty choice of the variance function. But it may also arise because one or more covariates are missing, or because of some outliers. That means the source of deviation cannot be determined exactly. Therefore, one has to be careful in the interpretation of such plots. Nevertheless, diagnostic methods are an important tool revealing the influence of single observations or observation sets on the fit.

For non-normal regression models, diagnostics are based on extended versions of residuals, the hat matrix, and case deletion measures. In this section we report on such developments within the context of generalized linear models. The methods can be easily extended to quasi-likelihood and nonlinear, nonexponential family models.

We do not report on diagnostics for detecting misscaled covariates or covariates that should be included in the linear predictor. In classical linear regression such methods are known as *partial residual* and *added variable plots* (see, e.g., Cook & Weisberg, 1982). Landwehr, Pregibon & Shoemaker (1984) extended the partial residual plot to logistic regression. An improved partial residual plot, the so-called *constructed variable plot*, for the generalized linear model was derived by Wang (1987). It is based on an added variable plot for generalized linear models that Wang (1985) introduced. Interesting ways of combining robustness and a forward search through the data with diagnostic and graphical tools were more recently proposed by Atkinson & Riani (2000).

### 4.2.1 Diagnostic Tools for the Classical Linear Model

Consider the classical linear model

$$y = X\beta + \varepsilon,$$

where  $\varepsilon \sim N(0, \sigma^2 I)$  is normally distributed,  $X$  is an  $(n \times p)$ -matrix of covariates, and  $\beta$  is a  $(p \times 1)$ -vector of coefficients.

Based on the least-squares estimate  $\hat{\beta} = (X'X)^{-1}X'y$ , the *residual vector* is given by

$$r = y - \hat{y},$$

where  $\hat{y} = X\hat{\beta}$  is the vector of fitted values. The residual vector shows which points are ill-fitting.

Another quantity that determines how much influence or leverage the data have on the fitted values is the *hat matrix*

$$H = X(X'X)^{-1}X'.$$

It is called the hat matrix because one has  $\hat{y} = Hy$ . Thus, the matrix  $H$  maps  $y$  into  $\hat{y}$ . Consider the  $p$ -dimensional subspace  $\{X\beta, \beta \in \mathbb{R}^p\}$ . Then  $\hat{y}$  is the perpendicular projection of  $y$  into this subspace. The matrix  $H$  acts as a projection matrix, which means  $H$  is symmetric and idempotent, i.e.,  $H^2 = H$  holds.

From  $\hat{y} = Hy$  it is seen that the element  $h_{ij}$  of the hat matrix  $H = (h_{ij})$  shows the amount of leverage or influence exerted on  $\hat{y}_i$  by  $y_j$ . Since  $H$  depends only on  $X$ , this influence is due to the “design” not to the dependent variable. The most interesting influence is that of  $y_i$  on the fitted value  $\hat{y}_i$ , which is reflected by the diagonal element  $h_{ii}$ . For the projection matrix  $H$  one has

$$\text{rank}(H) = \sum_{i=1}^n h_{ii} = p$$

and  $0 \leq h_{ii} \leq 1$ . Therefore,  $p/n$  is the average size of a diagonal element. As a rule of thumb an  $x$ -point for which  $h_{ii} > 2p/n$  holds is considered a high-leverage point (e.g., Hoaglin & Welsch, 1978).

In addition to the basic residual vector  $r$ , several standardized versions that have a constant variance are in common use. If  $r' = (r_1, \dots, r_n)$  denotes the basic residual vector, the standardization

$$r_i^* = r_i / \sqrt{1 - h_{ii}}$$

has variance  $\sigma^2$ , which can be estimated consistently by  $\hat{\sigma}^2 = r'r/(n-p)$ . The standardized residual  $r_i^*$  is scaled to variance 1 by dividing  $r_i^*$  by  $\hat{\sigma}$ .

The role of residual vector and hat matrix may also be seen by looking at the effect of omitting single observations. If the  $i$ th observation is omitted, the change in LS estimates is given by

$$\Delta_i \hat{\beta} = \hat{\beta} - \hat{\beta}_{(i)} = (X'X)^{-1}x_ir_i/(1-h_{ii}),$$

where  $\hat{\beta}_{(i)}$  is the LS estimate when the  $i$ th observation is omitted and  $x_i$  is the design vector from the  $i$ th component, i.e.,  $y_i = x_i'\beta + \varepsilon_i$ . The change  $\Delta_i \hat{\beta}$  increases with increasing residual  $r_i$  and increasing diagonal element  $h_{ii}$ .

### 4.2.2 Generalized Hat Matrix

Let us now consider a (univariate or multivariate) generalized linear model which for the  $i$ th  $q$ -dimensional observation has mean of the form

$$\mu_i = h(Z_i\beta).$$

For univariate GLMs one has  $q = 1$ ; for multicategorical responses as considered in Chapter 3, one has  $q = k - 1$  where  $k$  is the number of categories. As is common for simple goodness-of-fit statistics (Sections 2.2.2, 3.4.2), data should be grouped as far as possible. Therefore, the number  $g$  of grouped observations will be used in the following. In this section we will consider a generalized form of the hat matrix based on ML estimation. The hat matrix yields a measure for the leverage of data and is a building block of regression diagnostics in the sense of Pregibon (1981). It is useful to remember that  $A^{1/2}(A^{T/2})$  denotes a left (the corresponding right) square root of matrix  $A$  such that  $A^{1/2}A^{T/2} = A$  holds. The inverse matrices are denoted by  $A^{-1/2} = (A^{1/2})^{-1}$  and  $A^{-T/2} = (A^{T/2})^{-1}$ .

The iterative procedure for maximum likelihood estimates is given in Sections 2.2.1 and 3.4.1. At convergence the estimate has the form

$$\hat{\beta} = (Z'W(\hat{\beta})Z)^{-1}Z'W(\hat{\beta})\tilde{y}(\hat{\beta}),$$

where  $\tilde{y}(\hat{\beta}) = Z\hat{\beta} + (D^{-1}(\hat{\beta}))'(y - \mu(\hat{\beta}))$ . The estimate  $\hat{\beta}$  is a weighted least-squares solution of the linear problem  $\tilde{y}(\hat{\beta}) = Z\beta + \varepsilon$  or an unweighted least-squares solution of the linear problem

$$\tilde{y}_0(\hat{\beta}) = Z_0\beta + \tilde{\varepsilon},$$

where  $\tilde{y}_0(\hat{\beta}) = W^{T/2}(\hat{\beta})\tilde{y}(\hat{\beta})$ ,  $Z_0 = W^{T/2}(\hat{\beta})Z$ . The hat matrix corresponding to this model has the form

$$\begin{aligned} H &= Z_0(Z'_0 Z_0)^{-1} Z'_0 = W^{T/2}(\hat{\beta})Z(Z'W(\hat{\beta})Z)^{-1}Z'W^{1/2}(\hat{\beta}) \\ &= W^{T/2}(\hat{\beta})ZF^{-1}(\hat{\beta})Z'W^{1/2}(\hat{\beta}). \end{aligned}$$

From  $Z_0\hat{\beta} = H\tilde{y}_0(\hat{\beta})$  it is seen that  $H$  maps the observation  $\tilde{y}_0(\hat{\beta}) = W^{T/2}(\hat{\beta})\tilde{y}(\hat{\beta})$  into the “fitted” value  $W^{T/2}(\hat{\beta})Z\hat{\beta}$ .

The  $(qq \times qq)$ -matrix  $H$  is idempotent and symmetric and therefore may be viewed as a projection matrix for which  $\text{tr}(H) = \text{rank}(H)$  holds. For univariate response ( $q = 1$ ) the diagonal elements  $h_{ii}$  of the matrix  $H = (h_{ij})$  are determined by  $0 \leq h_{ii} \leq 1$ , and high values of  $h_{ii}$  (close to 1) correspond to extreme points in the design space. However, in contrast to the classical linear model, the hat matrix does not depend only on the design matrix but also on the fit. Therefore, extreme points in the design space do not necessarily have a high value of  $h_{ii}$ . For multivariate responses it is useful to consider the blocks  $H_{ij}$  of the matrix  $H = (H_{ij})$ ,  $i, j = 1, \dots, g$ , where  $H_{ij}$  is a  $(q \times q)$ -matrix. The  $(q \times q)$ -submatrix  $H_{ii}$  corresponds to the  $i$ th observation, and  $\det(H_{ii})$  or  $\text{tr}(H_{ii})$  may be used as an indicator for the leverage of  $y_i$ .

### Example 4.2: Vaso constriction

The data taken from Finney (1947) were obtained in a carefully controlled study in human physiology where a reflex “vaso constriction” may occur in the skin of the digits after taking a single deep breath. The response  $y$  is the occurrence ( $y = 1$ ) or nonoccurrence ( $y = 0$ ) of vaso constriction in the skin of the digits of one subject after he or she inhaled a certain volume of air at a certain rate. The responses of three subjects are available. The first contributed 9 responses, the second contributed 8 responses, and the third contributed 22 responses. The  $i = 1, \dots, 39$  observations are listed in Table 4.2.

Although the data represent repeated measurements, an analysis that assumes independent observations may be applied, as claimed by Pregibon (1981). Therefore, we use the binary logit model

$$\log(P(y = 1)/P(y = 0)) = \beta_0 + \beta_1 \log(\text{volume}) + \beta_2 \log(\text{rate})$$

to analyze the effect of volume of air and inspiration rate on the occurrence of vaso constriction. The deviance of the fit is 29.23 on 36 degrees of freedom. Pearson’s  $\chi^2$  takes value 34.23. Both statistics are less than the expectation 36 of an asymptotic  $\chi^2(36)$ -distribution, so that the fitted logit model seems to be adequate. The MLEs for  $\beta$ , their estimated standard errors, and the  $p$ -values for testing  $H_0 : \beta_r = 0$ ,  $r = 0, 1, 2$ , are given in Table 4.3. The MLEs

**Table 4.2.** Vaso constriction data

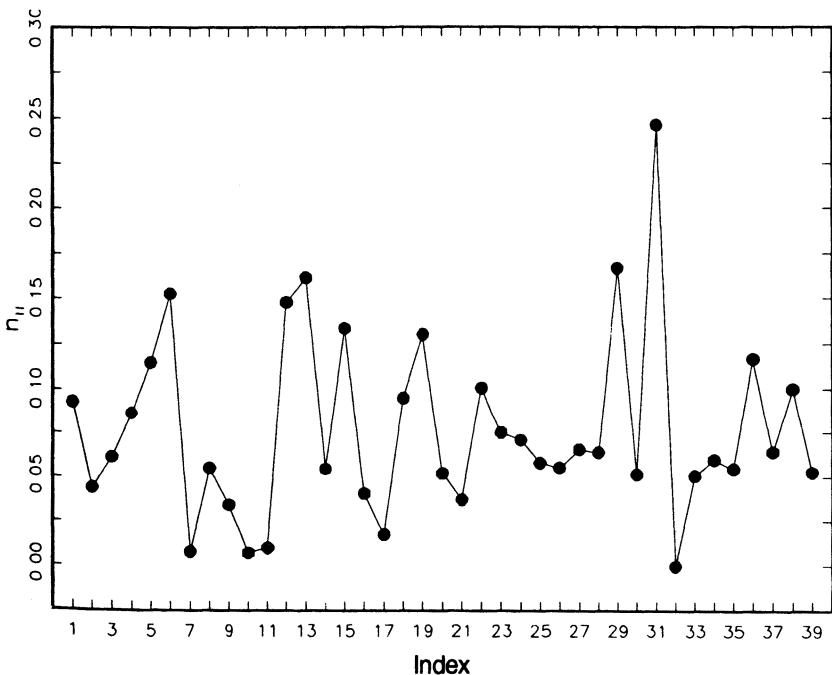
Index	Volume	Rate	Y	Index	Volume	Rate	Y
1	3.70	0.825	1	20	1.80	1.800	1
2	3.50	1.090	1	21	0.40	2.000	0
3	1.25	2.500	1	22	0.95	1.360	0
4	0.75	1.500	1	23	1.35	1.350	0
5	0.80	3.200	1	24	1.50	1.360	0
6	0.70	3.500	1	25	1.60	1.780	1
7	0.60	0.750	0	26	0.60	1.500	0
8	1.10	1.700	0	27	1.80	1.500	1
9	0.90	0.750	0	28	0.95	1.900	0
10	0.90	0.450	0	29	1.90	0.950	1
11	0.80	0.570	0	30	1.60	0.400	0
12	0.55	2.750	0	31	2.70	0.750	1
13	0.60	3.000	0	32	2.35	0.030	0
14	1.40	2.330	1	33	1.10	1.830	0
15	0.75	3.750	1	34	1.10	2.200	1
16	2.30	1.640	1	35	1.20	2.000	1
17	3.20	1.600	1	36	0.80	3.330	1
18	0.85	1.415	1	37	0.95	1.900	0
19	1.70	1.060	0	38	0.75	1.900	0
				39	1.30	1.625	1

for  $\beta_1$  and  $\beta_2$  show that with increasing volume as well as with increasing rate the probability for occurrence of vaso constriction becomes higher. In addition, all three parameters are highly significant, as can be seen from the  $p$ -values.

The leverage of the observation  $y_i, i = 1, \dots, 39$ , can be examined by an index plot of the diagonal elements  $h_{ii}$  of the hat matrix  $H$ . Figure 4.1 gives such a plot. Observation 31 has the highest value of  $h_{ii}$ . Since its value is higher than  $2p/n = 0.1533$ , one would suspect that observation 31 corresponds to a high-leverage point. However, as can be seen from Table 4.2, observation 31 is not extreme concerning the values of “volume” and “rate,” and its effect on the fit is minor in comparison to other observations identified by residual analytic tools (see Example 4.4).  $\square$

**Table 4.3.** Logit model fit to vaso constriction data

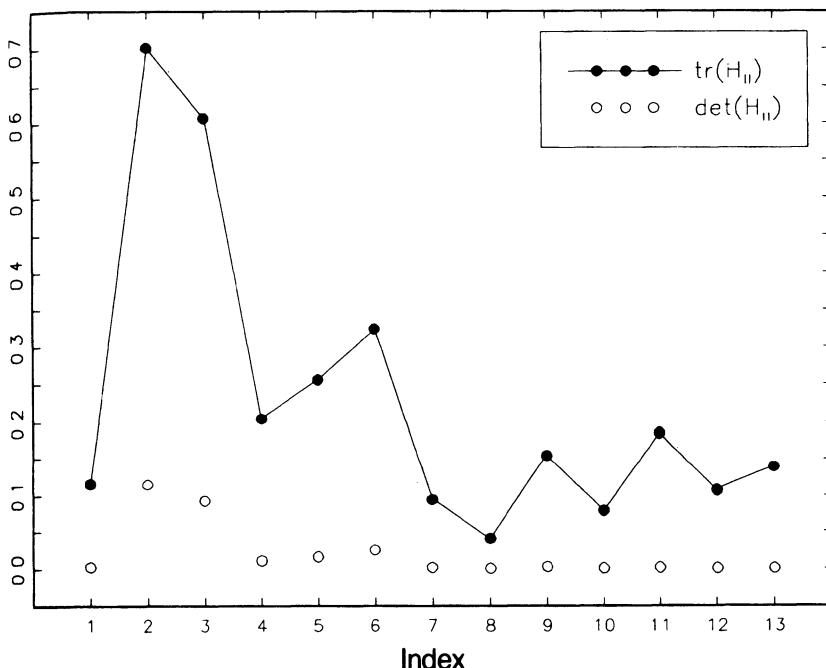
	MLE	Standard error	<i>p</i> -value
$\beta_0$	-2.875	1.319	0.029
$\beta_1$	5.179	1.862	0.005
$\beta_2$	4.562	1.835	0.013

**Figure 4.1.** Index plot of  $h_{ii}$  for vaso constriction data.**Example 4.3: Job expectation** (Examples 3.3, 3.5, continued)

Recall the grouped data for job expectation of psychology students given in Table 3.2. In Example 3.5 the relationship between the trichotomous response “job expectation” ( $Y$ ) and “age” was analyzed by the cumulative logistic model

$$P(Y \leq r)|\text{age}) = F(\theta_r + \gamma \log(\text{age})), \quad r = 1, 2.$$

Results of the fit are given in Table 3.4. Now we are interested in the leverage of the  $i$ th observation group,  $i = 1, \dots, 13$ . Note that these observation groups differ in the only available covariate “age,” and in the local sample size  $n_i$  as can be seen from Table 3.2 (p. 82). Since the response variable is three-categorical, we use  $\det(H_{ii})$  [resp.,  $\text{tr}(H_{ii})$ ] as a measure for leverage. Figure 4.2 gives both measures plotted against the index  $i$ . In contrast to the values of  $\text{tr}(H_{ii})$ , which take into account only the diagonal elements of  $H_{ii}$ , the values of  $\det(H_{ii})$  are based on all elements of  $H_{ii}$  and are much smaller. Both measures, however, identify the observation groups 2 and 3 as those having the highest leverage values. These values, however, are primarily caused by the relatively large local sample sizes  $n_i$  and not by an extremeness in the design space.  $\square$



**Figure 4.2.** Index plot of  $\text{tr}(H_{ii})$  and  $\det(H_{ii})$  for grouped job expectation data.

### 4.2.3 Residuals and Goodness-of-Fit Statistics

The quantities considered in residual analysis help to identify poorly fitting observations that are not well explained by the model. Moreover, they should reveal the impact of observations on the goodness-of-fit. Systematic

departures from the model may be checked by residual plots. The methods considered in the following are based on ML estimates for generalized linear models with  $q$ -dimensional observations  $y_i$ .

Generalized residuals should have properties similar to standardized residuals for the normal linear model. In particular, the following forms of generalized residuals are widely used.

The Pearson residual for observation  $y_i$  is based on the *Pearson goodness-of-fit statistic*

$$\chi^2 = \sum_{i=1}^q \chi_P^2(y_i, \hat{\mu}_i)$$

with the  $i$ th component

$$\chi_P^2(y_i, \hat{\mu}_i) = (y_i - \hat{\mu}_i)' \Sigma_i^{-1}(\hat{\beta})(y_i - \hat{\mu}_i),$$

where  $\Sigma_i(\hat{\beta})$  denotes the estimated covariance matrix of  $y_i$  and  $\hat{\mu}_i = \bar{\mu}_i(\hat{\beta})$ .

For the standardized form of the residual, one needs a standardization of  $y_i - \mu_i(\hat{\beta})$ . The *Pearson residual* is given by

$$r_i^P = \Sigma_i^{-1/2}(\hat{\beta})(y_i - \hat{\mu}_i),$$

which is a vector of length  $q$ . The Pearson residual may be considered a square root of  $\chi_P^2(y_i, \hat{\mu}_i)$  since  $\chi_P^2(y_i, \hat{\mu}_i) = (r_i^P)' r_i^P$  holds. For increasing local sample size  $n_i \rightarrow \infty$  as well as  $n \rightarrow \infty$ , the covariance of  $y_i - \hat{\mu}_i$  may be approximated by

$$\text{cov}(y_i - \hat{\mu}_i) = \Sigma_i^{1/2}(\hat{\beta})(I - H_{ii})\Sigma_i^{T/2}(\hat{\beta}),$$

where  $H_{ii}$  denotes the  $i$ th  $(q \times q)$ -matrix in the diagonal of the hat matrix  $H$  and  $I$  denotes the identity matrix. Therefore, the asymptotic covariance of  $r_i^P$  is  $\text{cov}(r_i^P) = I - H_{ii}$  and a *studentized version* of the *Pearson residual*  $r_i^P$  is given by

$$r_{i,s}^P = (I - H_{ii})^{-1/2} r_i^P = (I - H_{ii})^{-1/2} \Sigma_i^{-1/2}(\hat{\beta})(y_i - \hat{\mu}_i),$$

which for large  $n_i$  should be approximately multinormal.

For univariate models the Pearson residual takes the simple form

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\text{var}}(y_i)}}.$$

Since these residuals are skewed and thus for small  $n_i$  cannot be considered approximately normally distributed, transformed values  $t(y_i)$  and  $E(t(y_i))$  may be used instead of  $y_i$  and  $\mu_i(\beta) = E(y_i)$ . McCullagh & Nelder (1989, Section 2.4.2) give transformed *Anscombe residuals* for important cases. Alternative forms of transformations aim at the vanishing of the first-order asymptotic skewness when the local sample size increases are given by Pierce & Schafer (1986). For the Poisson distribution they give the Anscombe residual

$$r_i^A = \frac{3}{2} \frac{y_i^{2/3} - (\hat{\mu}_i^{2/3} - \hat{\mu}_i^{-1/3}/9)}{\hat{\mu}_i^{1/6}},$$

and for the binomial distribution  $y_i \sim B(n_i, \pi_i)$  they give the Anscombe residual

$$r_i^A = \sqrt{n_i} \frac{t(y_i) - [t(\hat{\pi}_i) + (\hat{\pi}_i(1 - \hat{\pi}_i))^{-1/3}(2\hat{\pi}_i - 1)/6n_i]}{\hat{\pi}_i(1 - \hat{\pi}_i)^{1/6}},$$

where  $t(u) = \int_0^u s^{-1/3}(1 - s)^{-1/3} ds$ . For alternative variance-stabilizing residuals, see Pierce & Schafer (1986).

Another type of residual is based on the *deviance*

$$D = \sum_{i=1}^g \chi_D^2(y_i, \hat{\mu}_i)$$

with components

$$\chi_D^2(y_i, \hat{\mu}_i) = 2(l_i(y_i) - l_i(\hat{\mu}_i)),$$

where

$$l_i(\mu_i) = [y'_i \theta_i - b(\theta_i)]\omega_i/\phi + c(y_i, \phi_i)$$

is the likelihood of  $\mu_i$  based on observation  $y_i$  and  $\hat{\mu}_i = \mu_i(\hat{\beta})$ . If  $\theta_i$  is degenerate,  $l_i(\mu_i)$  is defined by  $l_i(\mu_i) = 0$ . For univariate response models the deviance residual is given by

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i) \sqrt{\chi_D^2(y_i, \hat{\mu}_i)}.$$

Explicit forms for the various distributions are easily derived from Table 2.1 (p. 21) in Chapter 2. For example, for the Poisson distribution one gets the deviance residual

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i) \{2[y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)]\}^{1/2}.$$

*Adjusted deviances*, which show better approximation to the normal distribution, are given by

$$r_i^{AD} = r_i^D + \rho(\theta_i(\hat{\mu}_i))/6,$$

where  $\rho(\theta) = E_\theta\{[(y - \mu)/s(y)]^3\}$  with  $s(y) = \sqrt{\text{var}(y)}$ . Pierce & Schafer (1986) give the explicit forms for several distributions:

$$\begin{aligned} y \sim B(n, \pi) : \rho(\theta) &= (1 - 2\pi)/\{n\pi(1 - \pi)\}^{1/2}, \\ y \sim P(\lambda) : \rho(\theta) &= 1/\sqrt{\lambda}, \\ y \sim \Gamma(\mu, \nu) : \rho(\theta) &= 2/\sqrt{\nu}. \end{aligned}$$

For these cases they compare the true tail probabilities and normal approximations based on several types of residuals. Anscombe residuals and adjusted deviance residuals, which are very nearly the same, yield rather good approximations with a slight preference for the adjusted deviances. For the case of integer-valued  $y$ , Pierce & Schafer (1986) suggest making a continuity correction by replacing  $y$  in the formulas by  $y \pm 1/2$  toward the center of distribution.

For a nonparametric approach to assessing the influence of observations on the goodness-of-fit, see Simonoff & Tsai (1991).

#### Example 4.4: Vaso constriction (Example 4.2, continued)

Figure 4.3 gives an index plot of the Pearson residuals  $r_i^P$ , the standardized forms  $r_{i,s}^P$ , and the deviance residuals  $r_i^D$ . All residuals are based on the fit of the logistic model given in Example 4.2. Although the response is strictly binary, the three residuals behave quite similarly; observations 4 and 18 show the largest values for  $r_i^P$ ,  $r_{i,s}^P$  as well as for  $r_i^D$ . Obviously, the standardization of  $r_i^P$  by  $(1 - h_{ii})^{1/2}$ , which yields  $r_{i,s}^P$ , can be neglected since the leverage values  $h_{ii}$  are rather balanced as can be seen in Figure 4.1.

Observations 4 and 18 are also identified as outliers in a normal probability plot of the standardized Pearson residuals,  $r_{i,s}^P$  that can be seen in Figure 4.4. In Figure 4.4 the ordered values of  $r_{i,s}^P$ ,  $i = 1, \dots, 39$ , are plotted against the order statistic of an  $N(0, 1)$ -sample. In the case of approximately  $N(0, 1)$ -distributed residuals, such a plot shows approximately a straight line as long as model departures and/or outliers are absent. Clearly, since the response is strictly binary, the residuals  $r_{i,s}^P$  are not  $N(0, 1)$ -distributed. However, the plot can be used to detect badly fitted observations that are far away from the rest. Although observations 4 and 18 are badly fitted they are essential for the fitting of the logit model. The data set is very close to the case of perfect separation where the data are divided into non-overlapping groups with responses 0 in one group and 1 in the other. If observations 4, 18, and 24 are detected, perfect separation occurs, and parameter estimates

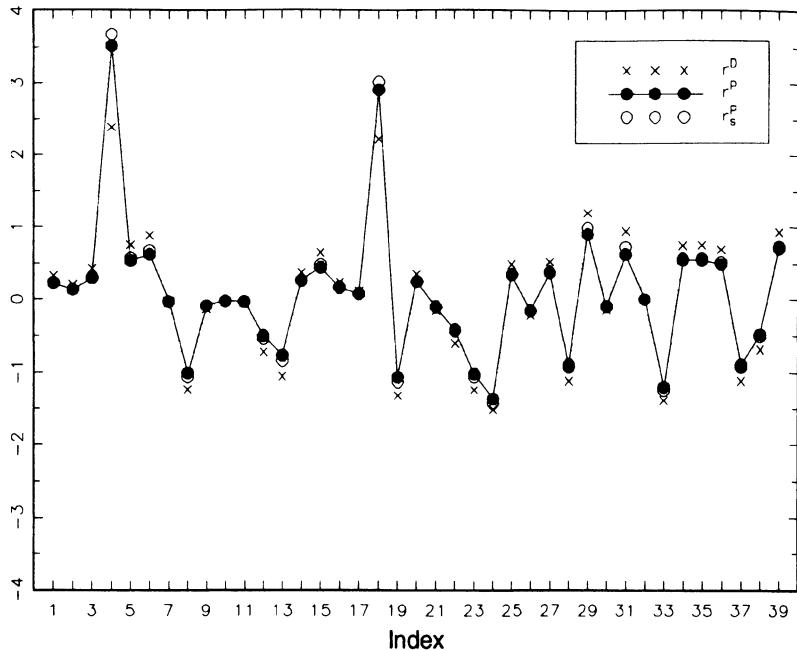


Figure 4.3. Index plot of  $r_i^P$ ,  $r_{i,s}^P$ , and  $r_i^D$  for vaso constriction data.

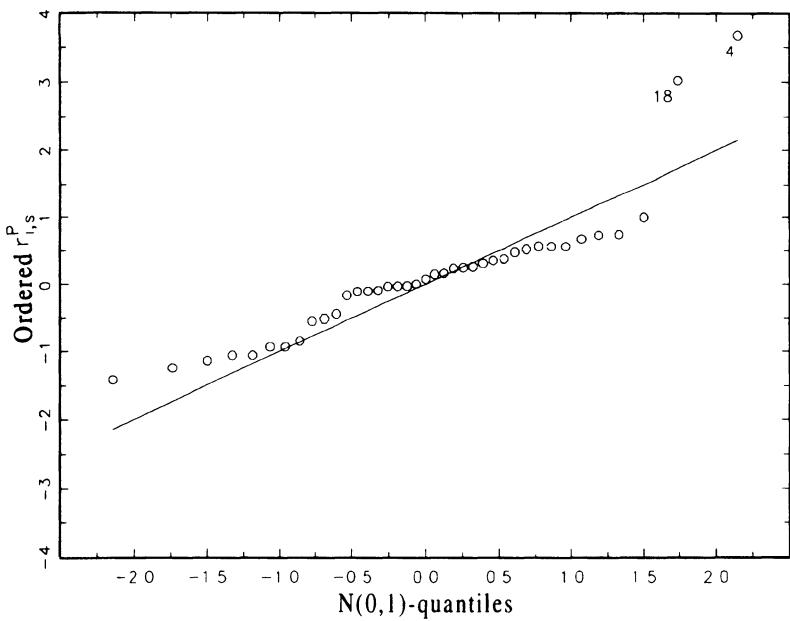


Figure 4.4.  $N(0, 1)$ -probability plot of  $r_{i,s}^P$  for vaso constriction data.

go to infinity. Further analysis including alternative link functions is found in Atkinson & Riani (2000).  $\square$

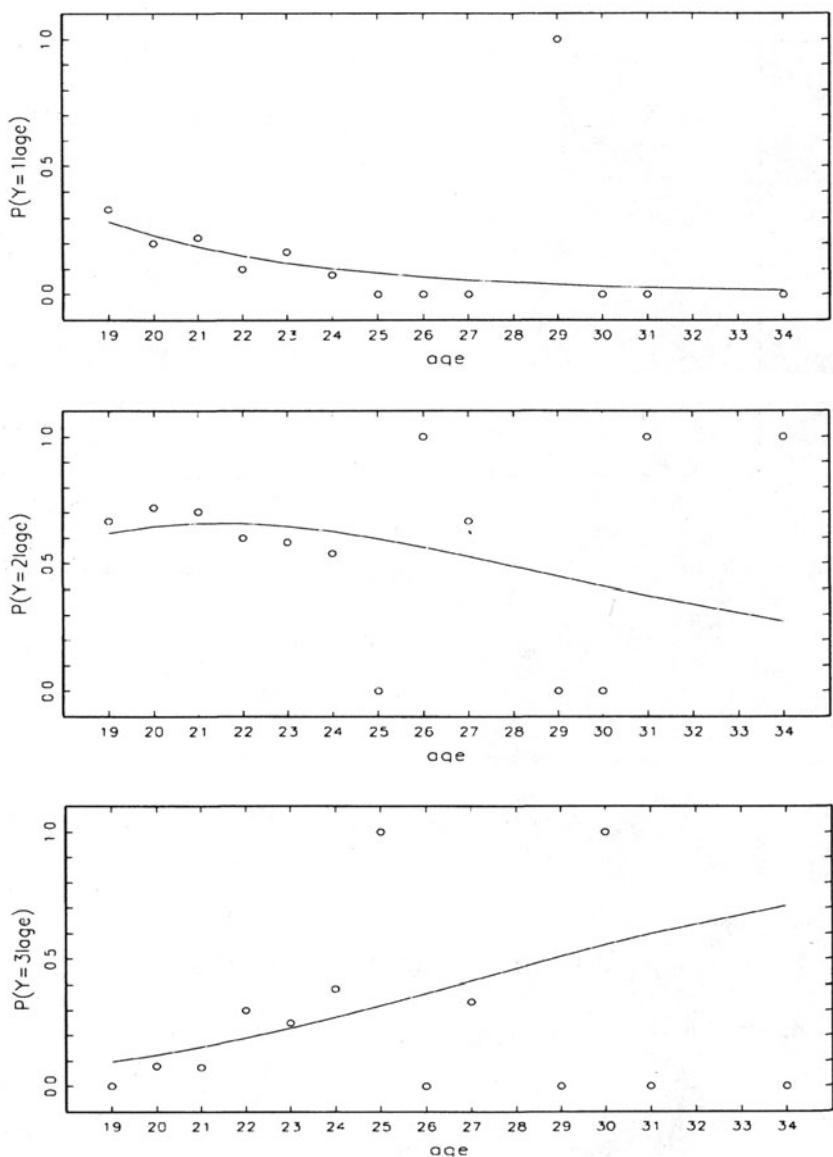
**Example 4.5: Job expectation** (Examples 3.3, 3.5, 4.3, continued)

For illustration of residuals in the case of a multicategorical response, we refer to the job expectation data. Figure 4.6 shows an index plot of the squared standardized Pearson residuals  $(r_{i,s}^P)'(r_{i,s}^P)$  that are based on the cumulative logit model given in Example 4.3. Note that the index  $i, i = 1, \dots, 13$ , stands for the  $i$ th observation group as given in Table 3.2 (p. 82). The plot in particular identifies observation group 10 as an outlying point. From Table 3.2 it can be seen that this observation group contains only one observation that is 29 years old and does not expect to find an adequate job after getting the degree in psychology. Such a response-covariate combination, however, does not fit into the estimated cumulative logit model, which suggests that the probability of not finding an adequate job decreases with increasing age. This is demonstrated by Figure 4.5, which shows the relative frequencies and probability for the fitted model. A similar argument holds for observation group 7 (25 years of age). A look at Figure 4.5 shows that observations 12 (31 years of age) and 13 (34 years of age) are also far from the expected value. However, the local sample size for both observations is only 1.

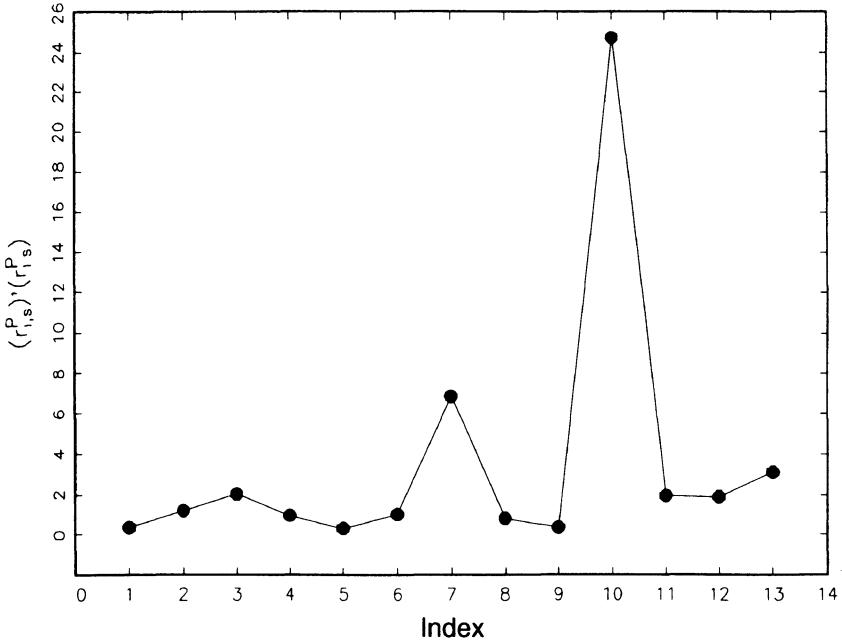
Since the local sample sizes  $n_i$  of most observation groups are larger than 1, one may compare the squared standardized Pearson residuals  $(r_{i,s}^P)'(r_{i,s}^P)$  with a  $\chi^2$ -distribution. More specifically, a  $\chi^2(2)$ -probability plot is carried out by plotting the ordered values  $(r_{i,s}^P)'(r_{i,s}^P)$  against the order statistics of a  $\chi^2(2)$ -sample. As can be seen in Figure 4.7, the plot approximates a straight line rather well, with the exception of observation 10. This leads one to suppose that the “bad” fit given in Table 3.4 is mainly caused by the outlying observation 10 (25 years).  $\square$

#### 4.2.4 Case Deletion

An indicator for the influence of the  $i$ th observation  $(y_i, Z_i)$  on the vector  $\beta$  can be calculated by the difference  $(\hat{\beta} - \hat{\beta}_{(i)})$ , where  $\hat{\beta}_{(i)}$  is the MLE obtained from the sample without observation  $(y_i, Z_i)$  and  $\hat{\beta}$  is the MLE from all observations. If  $\hat{\beta}_{(i)}$  is substantially different from  $\hat{\beta}$ , observation  $(y_i, Z_i)$  may be considered influential. Measures of this type have been given by Cook (1977) for linear regression models. Since the estimation of unknown parameters requires an iterative procedure, it is computationally expensive to subsequently delete each observation and refit the model. A one-step approximation of  $\hat{\beta}_{(i)}$  is obtained by performing just one step of the iterative process when  $\hat{\beta}$  is the starting value. The one-step estimate is given by



**Figure 4.5.** Relative frequencies and response curves of the fitted cumulative logistic model (responses correspond to “don’t expect adequate employment,” “not sure,” “expect employment immediately after getting the degree” from top to bottom).



**Figure 4.6.** Index plot of  $(r_{i,s}^P)'(r_{i,s}^P)$  for grouped job expectation data.

$$\hat{\beta}_{(i),1} = F_{(i)}^{-1}(\hat{\beta}) Z'_{(i)} W_{(i)}(\hat{\beta}) \tilde{y}(\hat{\beta}),$$

where the reduced Fisher matrix is given by

$$F_{(i)}(\hat{\beta}) = F(\hat{\beta}) - Z'_i W_i(\hat{\beta}) Z_i$$

and

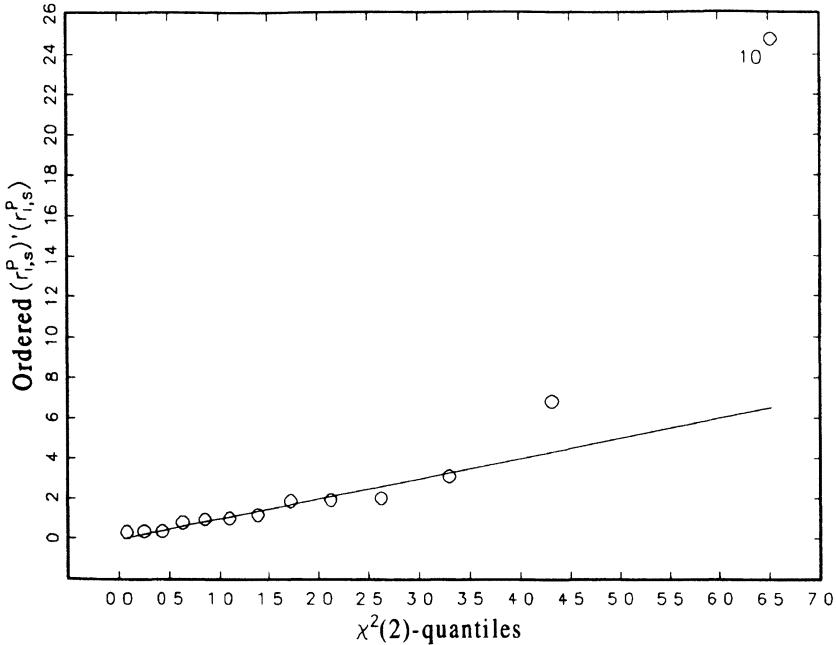
$$Z'_{(i)} W_{(i)}(\hat{\beta}) \tilde{y}(\hat{\beta}) = Z' W(\hat{\beta}) \tilde{y}(\hat{\beta}) - Z'_i W_i(\hat{\beta}) \tilde{y}_i(\hat{\beta}).$$

A simpler form of  $\hat{\beta}_{(i),1}$  given by Hennevogl & Kranert (1988) is

$$\hat{\beta}_{(i),1} = \hat{\beta} - F^{-1}(\hat{\beta}) Z'_i W_i^{1/2}(\hat{\beta}) (I - H_{ii})^{-1} \Sigma_i^{-1/2}(\hat{\beta}) (y_i - \mu_i(\hat{\beta})), \quad (4.2.1)$$

where  $H_{ii}$  is the  $i$ th block diagonal of the hat matrix evaluated at  $\hat{\beta}$ . The difference between the one-step estimate  $\hat{\beta}_{(i),1}$  and the original estimate  $\hat{\beta}$  may be used as an indicator for the impact of observation  $(y_i, Z_i)$  on the estimated parameter.

To determine the influence of observations on the estimate  $\hat{\beta}$ , one has to consider all the components of  $\hat{\beta}$ . Therefore, it is often useful to have an



**Figure 4.7.**  $\chi^2(2)$ -probability plot of  $(r_{i,s}^P)'(r_{i,s}^P)$  for grouped job expectation data.

overall measure as considered by Cook (1977). An asymptotic confidence region for  $\beta$  is given by the log-likelihood distance

$$-2\{l(\beta) - l(\hat{\beta})\} = c,$$

which is based on its asymptotic  $\chi^2(p)$ -distribution. Approximation of  $l(\beta)$  by a second-order Taylor expansion yields an approximate confidence region given by

$$(\beta - \hat{\beta})' \text{cov}(\hat{\beta})^{-1} (\beta - \hat{\beta}) \approx c. \quad (4.2.2)$$

If  $\beta$  is replaced by the one-step estimate  $\hat{\beta}_{(i),1}$ , one gets

$$\begin{aligned} c_{i,1} &= (\hat{\beta}_{(i),1} - \hat{\beta})' \text{cov}(\hat{\beta})^{-1} (\hat{\beta}_{(i),1} - \hat{\beta}) \\ &= (y_i - \mu_i(\hat{\beta}))' \Sigma_i^{-T/2} (\hat{\beta}) (I - H_{ii})^{-1} H_{ii} \\ &\quad \cdot (I - H_{ii})^{-1} \Sigma_i^{-1/2} (\hat{\beta}) (y_i - \mu_i(\hat{\beta})) \\ &= (r_i^P)' (I - H_{ii})^{-1} H_{ii} (I - H_{ii})^{-1} r_i^P. \end{aligned} \quad (4.2.3)$$

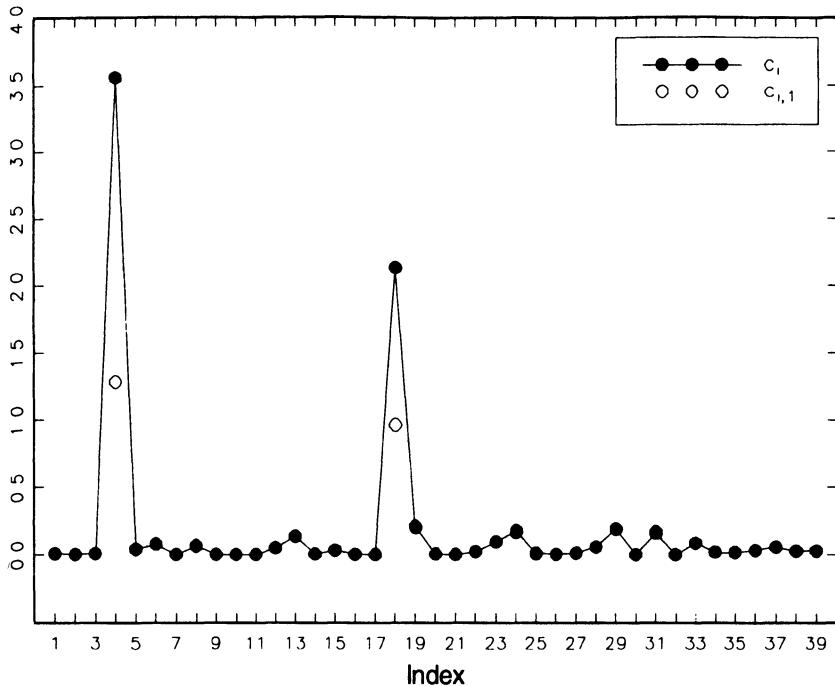


Figure 4.8. Index plot of  $c_{i,1}$  and  $c_i$  for vaso constriction data.

Pregibon (1981) refers to  $c_{i,1}$  as a confidence interval displacement diagnostic. The quantity  $c_{i,1}$  measures the displacement of  $\hat{\beta}$  by omitting observation  $(y_i, Z_i)$ . As seen in (4.2.3),  $c_{i,1}$  is composed from previously defined diagnostic elements. Again the generalized hat matrix  $H_{ii}$  for the  $i$ th observation plays an important role. Large values in  $H_{ii}$  will produce large values  $c_{i,1}$ . The same holds for the Pearson residual  $r_i^P$ . The original Cook measure as proposed by Cook (1977) for the normal linear model makes use of the leaving-one-out estimate  $\hat{\beta}_{(i)}$  instead of the one-step approximation  $\hat{\beta}_{(i),1}$ . This corresponds to replacing  $\beta$  by  $\hat{\beta}_{(i)}$  in (4.2.2), yielding  $c_i = (\hat{\beta}_{(i)} - \hat{\beta})\text{cov}(\hat{\beta})^{-1}(\hat{\beta}_{(i)} - \hat{\beta})$ . Applications for the generalized linear model show that the use of  $\hat{\beta}_{(i),1}$  often tends to underestimate the original Cook measure. The quantities  $c_i$  and  $c_{i,1}$  assess the effect of omitting observation  $(y_i, Z_i)$  on the estimate of the whole parameter vector  $\beta$ . Other quantities that measure the change in various other aspects of the fit are given by Pregibon (1981), Williams (1987), Lee (1988), and Lesaffre & Albert (1989). For example, Williams (1987) measures the change in likelihood ratio statistics, and Lee (1988) derives influence quantities that measure the effect of deleting observation  $(y_i, Z_i)$  on the estimates of a subset or single components of the parameter vector  $\beta$ .

**Example 4.6: Vaso constriction** (Examples 4.2, 4.4, continued)

Figure 4.8 gives an index plot of the approximate Cook distances  $c_{i,1}$  as well as of the exact measures  $c_i$  for the logistic model given in Example 4.2. The approximate values  $c_{i,1}$  are quite similar to the exact values  $c_i$ . Only for observations having an undue influence on the MLE  $\hat{\beta}$ , i.e., observations 4 and 18, does the approximative measure  $c_{i,1}$  underestimate the exact value  $c_i$ . The large influence of observations 4 and 18, given by  $c_{4,1}$  and  $c_{18,1}$ , is primarily caused by large residuals, as can be seen in Figure 4.3, and not by large leverage values, as can be seen in Figure 4.1.  $\square$

**Example 4.7: Job expectation** (Examples 3.3, 3.5, 4.3, 4.5, continued)

Approximate Cook distances  $c_i$  for the grouped job expectation data of Table 3.2 are given in Figure 4.9. The values  $c_i$  were obtained by fitting the cumulative logit model of Example 4.3 to the reduced data that do not contain observation group  $i$ . In terms of  $c_i$  observations 3, 7, 10, and 13 have the strongest influence. Observation 10 and (somewhat weaker) observations 7 and 13 show high residuals (Figure 4.6), whereas observation 3 shows a large value for  $\text{tr}(H_{ii})$ . The one-step approximation does not work so well. In particular, observations 2 and 10 show a quite different value. Since  $c_{i,1}$  is a direct function of residuals and hat matrix, the high value of  $\text{tr}(H_{ii})$  of observation 2 may cause this effect.  $\square$

## 4.3 General Tests for Misspecification\*

Throughout this section we maintain the basic assumption of (conditionally) independent responses. This assumption will generally be violated in the time series and longitudinal data situation. Consequences of its violation can be studied along the lines of White (1984). The tests for misspecification described in the sequel should be of potential usefulness for testing whether a model is misspecified for one of the following reasons:

- (i) The linear predictor does not reflect the influence of covariates correctly, e.g., due to omitted covariates, a wrong design matrix in mult categorial models, or nonlinearities not covered by the chosen predictor.
- (ii) Link or response function violation occurs, e.g., when a logit model is used instead of another binary model.
- (iii) The exponential family or the variance function is incorrectly specified.

A number of tests of deviations in one of these particular directions have been developed in the literature. Most of them include further parameters

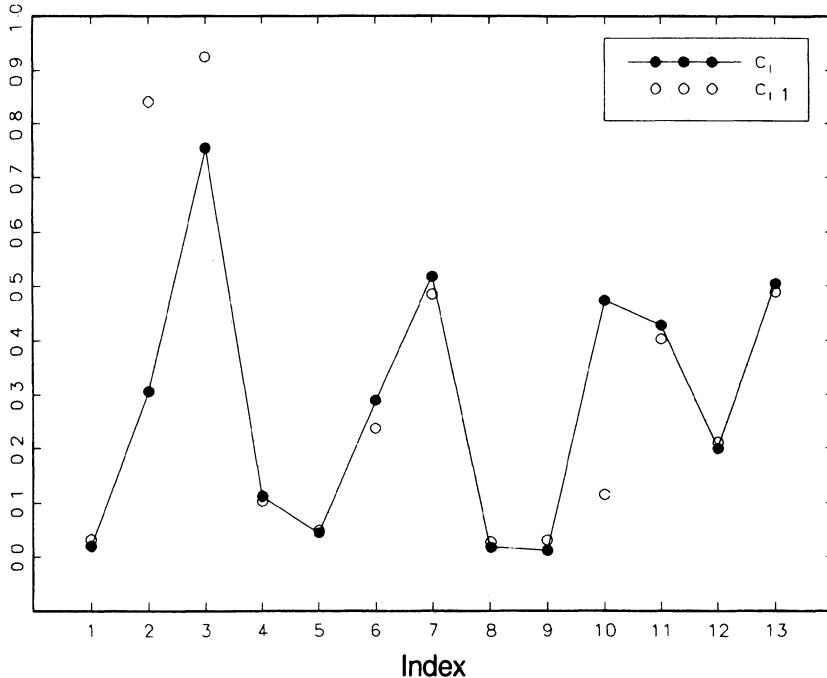


Figure 4.9. Index plot of  $c_{i,1}$  and  $c_i$  for grouped job expectation data.

in the model, defining, e.g., a generalized family of link functions as in Pregibon (1980), Stukel (1988), and Czado (1992), and test whether or not this inclusion significantly improves the fit. For a survey of tests of this type the reader is referred to Chapter 11.4 of McCullagh & Nelder (1989). A nonparametric approach for checking the adequacy of the link function is proposed in Azzalini, Bowman & Härdle (1989) and Horowitz & Härdle (1994).

In Sections 4.3.2 and 4.3.3 we will take a look at some alternative general tests for misspecification that have mainly been discussed in the econometric literature. Most of these tests were developed, at least in the beginning, with a view toward linear and nonlinear regression models for metric responses, but they are applicable, in principle, to any maximum likelihood-type regression problem. We start with a short discussion of the consequences of model assumption violations, adapting White's (1982, 1984) work to generalized linear models. A more detailed presentation can be found in Fahrmeir (1987a, 1990).

### 4.3.1 Estimation under Model Misspecification

In the following, terms like  $E$ ,  $\text{cov}$ , etc. refer to "true" expectations, covariances, etc., corresponding to the "true" probabilistic process  $P$  generating

the data  $(y_1, x_1), \dots, (y_n, x_n)$ . After one chooses a certain GLM or quasi-likelihood model, estimation is based on the *quasi-score function*

$$s(\beta) = \sum_{i=1}^n M_i(\beta) \Sigma_i^{-1}(\beta) (y_i - \mu_i(\beta)), \quad (4.3.1)$$

where  $\mu_i(\beta)$  and  $\Sigma_i(\beta)$  are the mean and variance structure implied by the model specification, and  $M_i(\beta) = \partial \mu_i(\beta) / \partial \beta = Z'_i D_i(\beta)$ . Compared to the quasi-likelihood modelling considered in Section 2.3.1, not only the variance function but also the mean may be misspecified, i.e., there is no “true”  $\beta_0$  in the admissible set  $B$  of parameters such that  $E(y_i|x_i) = \mu_i(\beta_0)$ ,  $i = 1, \dots, n$ , for the *true mean*  $m_i = E(y_i|x_i)$ . A quasi-MLE  $\hat{\beta}$  is computed by the usual algorithms as a root of  $s(\beta)$ . What can we say about such a QMLE? An investigation of its properties is based on the heuristic idea that a local QMLE should be near to a root  $\beta^*$  of the *expected quasi-score function*

$$s^*(\beta) = Es(\beta) = \sum_{i=1}^n M_i(\beta) \Sigma_i^{-1}(\beta) (m_i - \mu_i(\beta)). \quad (4.3.2)$$

Such a root  $\beta^*$  of (4.3.2) may be termed a quasi-true parameter, as it plays a similar role as the “true” parameter  $\beta_0$  in correctly specified models. Comparing it with (4.3.1), one sees that  $Es(\beta)$  is obtained from  $s(\beta)$  simply by replacing responses  $y_i$  by their true means  $m_i = E(y_i|x_i)$ . In the case where  $s(\beta)$  is obtained as the derivative of a proper log-likelihood model,  $l(\beta)$ ,  $\beta^*$  can be interpreted as a minimizer of the Kullback-Leibler distance between  $l(\beta)$  and the true log-likelihood.

Under certain regularity assumptions (White, 1982; Gourieroux, Monfort & Trognon, 1984; Fahrmeir, 1990), which are similar but somewhat sharper than in the case of correctly specified models, the QMLE  $\hat{\beta}$  is asymptotically normal,

$$\hat{\beta} \stackrel{a}{\sim} N(\beta^*, H^*(V^*)^{-1} H^*),$$

where  $H(\beta) = -\partial s(\beta) / \partial \beta'$  is the quasi-information matrix,

$$V(\beta) = \text{cov } s(\beta) = \sum_{i=1}^n M_i(\beta) \Sigma_i^{-1}(\beta) S_i \Sigma_i^{-1}(\beta) M'_i(\beta),$$

with  $S_i = \text{cov}(y_i|x_i)$  as the true covariance matrix of  $y_i$ , and “\*” means evaluation at  $\beta = \beta^*$ .

Remarks:

- (i) If the model is correctly specified, then  $\beta^* = \beta_0$  (the “true” parameter) and  $S_i = \Sigma_i(\beta_0)$ , so that  $V(\beta)$  reduces to the common expected

information  $F(\beta)$ . Moreover,  $H(\beta)$  is asymptotically equivalent to  $F(\beta)$ , so that asymptotics boils down to the usual asymptotic normality result for GLMs.

(ii) If only the mean is correctly specified, i.e.,

$$m_i = h(Z'_i \beta_0), \quad i = 1, 2, \dots \quad \text{for some } \beta_0,$$

then we still have  $\beta^* = \beta_0$ , since  $\beta_0$  is a root of (4.3.2). Estimating  $S_i$  by  $(y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)'$  and  $H^*$  by  $\hat{F}$  essentially reduces to the asymptotic normality result for quasi-likelihood estimation in Section 2.3.1.

(iii) A special but important case of mean misspecification occurs if the linear predictor is still correctly specified by

$$\eta = \alpha + x' \gamma$$

say (only univariate models are considered), but the response function  $h(\cdot)$  is wrong. Extending previous work on estimation under link violation, Li & Duan (1989) show the following: In the population case where  $(y_i, x_i)$  are i.i.d. replicates of  $(y, x)$ , the slope vector  $\gamma$  can still be estimated consistently up to a scale factor, i.e.,

$$\hat{\gamma} \rightarrow \gamma^0 = c\gamma,$$

provided that  $h$  is a response function that leads to a concave log-likelihood, and that the regressor vector  $x$  is sampled randomly from a continuous distribution with the property that the conditional expectation  $E(x'a|x'\gamma)$  is linear in  $x'\gamma$  for any linear combination  $x'a$ . The latter condition is fulfilled for elliptically symmetric distributions, including the multivariate normal. This result indicates that estimates  $\hat{\beta}$  can still be meaningful even if the response function is grossly misspecified but the linear predictor is correctly specified: We still can estimate the ratios  $\gamma_i/\gamma_k$  of the slope vector consistently, and these are the key quantities for interpretation of the effects of covariates  $x_i$  and  $x_k$ . In addition,  $\hat{\gamma}$  is asymptotically normal (Li & Duan, 1989, p. 1030).

(iv) For applications, an estimator of the asymptotic covariance matrix is needed. Whereas  $H^*$  can be estimated by  $\hat{H}$  or  $\hat{F}$ , where “ $\hat{\phantom{x}}$ ” means evaluation at  $\hat{\beta}$ , no general solution seems to be available for  $V^*$  yet, since it contains the unknown covariance matrices  $S_i$ . If the mean is not too grossly misspecified in the sense that  $\hat{\mu}_i$  fits  $y_i$  not too badly, it seems reasonable to use the estimator

$$\hat{V} = \sum_{i=1}^n \hat{M}_i \hat{\Sigma}_i^{-1} (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' \hat{\Sigma}_i^{-1} \hat{M}_i'.$$

### 4.3.2 Hausman-type Tests

#### Hausman Tests

The original proposal of Hausman (1978) was made in the context of linear regression, with primary interest in developing a test for misspecification that detects correlation between covariates and errors caused, e.g., by omitted variables. However, the test can also be sensitive to other alternatives, and the basic approach can be extended beyond the frame of linear models. The general idea is to construct two alternative estimators  $\hat{\beta}$  and  $\tilde{\beta}$  that are consistent for the “true” parameter  $\beta_0$  under the null hypothesis  $H_0$  of correct specification and that do not converge to the same limit if the model is misspecified. (In the latter case both estimators may be inconsistent.) Furthermore both estimators have to be asymptotically normal. A Hausman test statistic of the Wald type is then defined by (see also Hausman & Taylor, 1981; Holly, 1982)

$$w_h = (\hat{\beta} - \tilde{\beta})' \hat{C}(\hat{\beta}, \tilde{\beta})^{-1} (\hat{\beta} - \tilde{\beta}), \quad (4.3.3)$$

where  $C(\hat{\beta}, \tilde{\beta})$  is the asymptotic covariance matrix of  $\hat{\beta} - \tilde{\beta}$  under  $H_0$ , and  $\hat{C}(\hat{\beta}, \tilde{\beta})$  is a consistent estimate of  $C$ . If  $\hat{\beta}$  is chosen as the MLE or as any other asymptotically efficient estimator, then

$$\hat{C}(\hat{\beta}, \tilde{\beta}) = \hat{V}(\hat{\beta}) - \tilde{V}(\tilde{\beta}),$$

where  $\hat{V}$  and  $\tilde{V}$  are the asymptotic covariance matrices of  $\hat{\beta}$  and  $\tilde{\beta}$ . Generally, the asymptotic covariance matrix  $C$  can be obtained from the joint asymptotic distribution of  $(\hat{\beta}, \tilde{\beta})$ ; see, e.g., Arminger & Schoenberg (1989).

Given certain additional regularity conditions, the Wald-type Hausman statistic  $w_h$  has an asymptotic  $\chi^2$ -distribution with  $r = \text{rank}(C)$  degrees of freedom under  $H_0$ . Large values of  $w_h$  should therefore indicate model misspecification. In its general formulation the Hausman test is more a principle than a specific test. Some important questions – in particular, applications of the basic idea – are still left open.

A main problem is the choice of alternative estimators  $\hat{\beta}$  and  $\tilde{\beta}$ . Taking  $\hat{\beta}$  as the MLE or any other efficient estimator, one has to find an inefficient alternative estimator  $\tilde{\beta}$  such that  $w_h$  has enough power to detect various sources of misspecification. For testing correlation among regressors and errors in linear models, estimators  $\tilde{\beta}$  based on instrumental variables or weighted least-squares have been proposed. For generalized linear models there are more possible candidates, e.g., unweighted least squares estimators or QMLEs  $\tilde{\beta}$  maximizing weighted quasi-log-likelihoods

$$l_w(\beta) = \sum_{i=1}^n w_i l_i(\beta).$$

In the context of nonlinear regression, White (1981) suggested choosing weights  $w_i$  in such a way that the region of the regressor space where the  $y_i$ s are fitted poorly is heavily weighted compared to regions with a good fit. Arminger & Schoenberg (1989) adapt this proposal to quasi-likelihood models.

A further problem is: How well does the  $\chi^2$ -approximation work for finite sample sizes? Two questions arise in this context. First, can the shape of the finite sample distribution be adequately approximated by a  $\chi^2$ -distribution, in particular in the tails? Monte Carlo studies would provide some evidence, but there seems to be a lack of such investigations in the published literature. Second, as first observed by Krämer (1986) for linear regression models, the determination of  $\text{rank}(C)$  by the rank of the estimate  $\hat{C}$  may be misleading: Despite the singularity of  $C$ , the estimate  $\hat{C}$  may be nonsingular or have higher rank than  $C$ . Then wrong inferences will be drawn due to wrong degrees of freedom. Moreover, even the asymptotic  $\chi^2$ -distribution can be in question: A sufficient and necessary condition that  $\hat{C}^-$  is a consistent estimate of  $C^-$  is  $\text{rank}(\hat{C}) \rightarrow \text{rank}(C)$  in probability as  $n \rightarrow \infty$ ; see also Andrews (1987). There are counterexamples where the latter condition is violated, implying that  $\hat{C}^-$  is not a consistent estimate of  $C^-$  although  $\hat{C}$  is a consistent estimate of  $C$ . Since consistency of  $\hat{C}^-$  is crucial in conventional proofs, even the asymptotic  $\chi^2$ -distribution can be in question, a fact that has often been overlooked in earlier work.

In a simulation study Frost (1991) investigated several versions of Hausman tests, including unweighted LSEs and weighted QMLEs  $\hat{\beta}$ , for categorical regression models. The results were disappointing: Various versions of the test often failed to hold its asymptotic significance level and had insufficient power.

### Information Matrix Test

White (1982) proposed comparing estimates  $\hat{F}$  and  $\hat{V}$  of the (expected) quasi-information  $F(\beta) = EH(\beta)$  and of the covariance matrix  $V(\beta)$  given in the previous section. Under correct specification the usual regularity assumptions of ML estimation imply  $F = V, \hat{F} - \hat{V} \xrightarrow{P} 0$ , while  $F \neq V$  under misspecification. The information matrix test based on this idea can be interpreted as a Hausman-type test: The difference of alternative estimators is replaced by the vectorized difference

$$\hat{d} = \text{vec}(\hat{V} - \hat{F}),$$

and the Hausman-type statistic for the information matrix test is

$$i = \tilde{d}' \hat{C}^{-1} \tilde{d},$$

where  $\tilde{d}$  is a subvector of  $\hat{d}$  selected such that its asymptotic covariance matrix  $C$  is nonsingular. Under  $H_0$  it should be asymptotically  $\chi^2$  with  $r = \text{rank}(C)$  degrees of freedom. The main problems connected with its implementation are selecting an appropriate subvector  $\tilde{d}$  of  $\hat{d}$  and finding consistent and numerically stable estimators  $\hat{C}$  for  $C$ .

In a number of simulation studies we found that the information matrix test behaved unsatisfactorily. In particular, the probability for falsely rejecting a correctly specified model was often too high. Similar observations have also been made by Andrews (1988).

To summarize: Although Hausman-type tests are implemented in some program packages, in our experience one should be careful when using current versions of Hausman-type tests as a general tool for detecting misspecification in GLMs. It seems that further investigation is necessary here.

### 4.3.3 Tests for Nonnested Hypotheses

Hausman-type tests are pure significance tests in the sense that there is no particular alternative  $H_1$  to the model  $H_0$ . The tests considered in this section require a specific alternative model  $H_1$ . In contrast to classical tests null and alternative hypotheses need not be nested. For binomial responses typical nonnested hypotheses are, e.g.,

$$\begin{aligned} H_0: & \text{logit model with predictor } \eta = z'\beta, \\ H_1: & \text{linear model with the same predictor,} \end{aligned}$$

or

$$\begin{aligned} H_0: & \text{logit model with predictor } \eta_0 = z'_0\beta_0, \\ H_1: & \text{logit model with different predictor } \eta_1 = z'_1\beta_1, \end{aligned}$$

or a combination of both. Generally two models  $H_0$  and  $H_1$  are said to be nonnested if  $H_0$  is not nested within  $H_1$ , and  $H_1$  is not nested within  $H_0$ .

Much of the theoretical literature on nonnested hypotheses testing has its roots in the basic work of Cox (1961, 1962), who proposed a modified likelihood ratio statistic. However, “Cox-type tests” are often difficult to implement, and a number of authors proposed simplified versions; see McKinnon (1983) for a survey. A major class of models that is more convenient to deal with is based on the idea of embedding  $H_0$  and  $H_1$  in a supermodel.

### Tests Based on Artificial Nesting

For linear models Davidson & McKinnon (1980) formulated an artificial supermodel

$$y = (1 - \lambda)z'_0\beta_0 + \lambda z'_1\beta_1 + \varepsilon$$

in order to test  $H_0 : \lambda = 0$  against  $H_1 : \lambda \neq 0$ . For identifiability reasons,  $\beta_1$  has to be replaced by some estimator  $\hat{\beta}_1$ , e.g., the OLSE of the alternative model  $y = z'_1\beta_1 + \varepsilon_1$ . Davidson & McKinnon (1980) suggested a common  $t$ -test (i.e., a Wald test) for testing  $H_0 : \lambda = 0$ , which requires joint estimation of  $b = (1 - \lambda)\beta_0$  and  $\lambda$ . Their “J-test” can be carried out easily with standard software for linear models.

For generalized linear models, Gourieroux (1985) derived a score statistic for testing  $\lambda = 0$ . Assuming that both models belong to the same exponential family, as in the examples for binomial responses earlier, the test is based on artificial models of the form

$$Ey = \mu = (1 - \lambda)h_0(x'_0\beta_0) + \lambda h_1(x'_1\beta_1),$$

where  $h_0(x'_0\beta_0)$  and  $h_1(x'_1\beta_1)$  stand for the null and alternative models, respectively. For the resulting score statistic we refer the reader to Gourieroux (1985).

### Generalized Wald and Score Tests

A different approach, which is closer to the idea of Hausman’s test, has been taken by Gourieroux, Monfort & Trognon (1983b). Let

$$s_0(\beta_0) = \sum_{i=1}^n M_{i0}\Sigma_{i0}^{-1}(y_i - h_0(Z'_{i0}\beta_0))$$

and

$$s_1(\beta_1) = \sum_{i=1}^n M_{i1}^{-1}\Sigma_{i1}^{-1}(y_i - h_1(Z'_{i1}\beta_1))$$

denote the (quasi-) score functions under  $H_0$  and  $H_1$ . Based on the theory of estimation under misspecification (Section 4.3.1), the idea is to compare the (quasi-) MLE  $\hat{\beta}_1$ , i.e., a root of  $s_1(\beta_1)$ , with the corresponding root  $\hat{\beta}_{10}$  of

$$\hat{s}_1(\beta_1) = \sum_{i=1}^n M_{i1} \Sigma_{i1}^{-1} (\hat{h}_{i0} - h_1(Z'_{i1} \beta_1)),$$

where  $\hat{h}_{i0} = h_0(Z'_{i0} \hat{\beta}_0)$  and  $\hat{\beta}_0$  is a root of  $s_0(\beta_0)$ . Compared to the expected (under  $H_0$ ) score function  $E_0 s_1(\beta_1)$ , see (4.3.2), the expectation  $m_{i0} = E_0 y_i$  is replaced by its estimate  $\hat{h}_{i0}$ . Therefore, it is plausible that under  $H_0$  both  $\hat{\beta}_1$  and  $\hat{\beta}_{10}$  should tend to the quasi-true value  $\beta_1^*$ . As a consequence of the asymptotic results in Section 4.3.1, it can be conjectured that the difference  $\hat{\beta}_1 - \hat{\beta}_{10}$  is asymptotically normal. Gourieroux, Monfort & Trognon (1983a) gave a formal proof of this conjecture. A Wald statistic of the form

$$w = (\hat{\beta}_1 - \hat{\beta}_{10})' \hat{C}_w^- (\hat{\beta}_1 - \hat{\beta}_{10}),$$

where  $\hat{C}_w$  is an estimate of the possibly singular asymptotic covariance matrix of  $\hat{\beta}_1 - \hat{\beta}_{10}$ , can be used to test  $H_0$  versus  $H_1$ . Large values of  $\hat{\beta}_1 - \hat{\beta}_{10}$  and  $w$  are in favor of  $H_1$ . Alternatively, the following asymptotically equivalent score statistic can be derived:

$$s = \hat{s}'_{10} \hat{C}_s^- \hat{s}_{10}, \quad (4.3.4)$$

with  $\hat{s}_{10} = s_1(\hat{\beta}_{10})$  and the estimated asymptotic covariance matrix

$$\begin{aligned} \hat{C}_s &= \hat{C}_{11} - \hat{C}_{10} \hat{C}_{00}^{-1} \hat{C}_{01}, \\ \hat{C}_{00} &= \sum_{i=1}^n \hat{M}_{i0} \hat{\Sigma}_{i0}^{-1} \hat{M}'_{i0}, \\ \hat{C}_{01} &= \sum_{i=1}^n \hat{M}_{i0} \hat{\Sigma}_{i1}^{-1} \hat{M}'_{i1}, \hat{C}_{10} = \sum_{i=1}^n M_{i1} \hat{\Sigma}_{i0}^{-1} \hat{M}'_{i0} = \hat{C}'_{01}, \\ \hat{C}_{11} &= \sum_{i=1}^n \hat{M}_{i1} \hat{\Sigma}_{i1}^{-1} \hat{\Sigma}_{i0} \hat{\Sigma}_{i1}^{-1} \hat{M}'_{i1}, \end{aligned}$$

where “ $\hat{\cdot}$ ” means evaluation at  $\beta_0 = \hat{\beta}_0$  or  $\beta_1 = \hat{\beta}_{10}$ , respectively. Compared to the Wald statistic,  $\hat{C}_s$  is easier to implement and numerically more stable than  $\hat{C}_w$ . If the  $H_0$  model is true,  $s$  is asymptotically  $\chi^2$ -distributed with rank ( $C_s$ ) degrees of freedom.

In a number of simulation experiments Frost (1991) compared generalized score and Wald tests with the Gourieroux test based on artificial nesting. All in all, the results are in favor of the generalized score test.

To decide between two rival models in applications, one has to carry out the test twice by changing the role of  $H_0$  and  $H_1$ . Correspondingly, there are four possible results:

- (i)  $H_0$  not rejected,  $H_1$  rejected: choice of  $H_0$ ,

- (ii)  $H_0$  rejected,  $H_1$  not rejected: choice of  $H_1$ ,
- (iii)  $H_0$  and  $H_1$  rejected: both models rejected,
- (iv)  $H_0$  nor  $H_1$  rejected: no conclusion possible.

**Example 4.8: Credit-scoring** (Examples 2.2, 2.5, 4.1, continued)

As an illustration we use the credit-scoring data already analyzed previously. We will compare nonnested rival logit models characterized by different covariate vectors using the generalized score statistic (4.3.4). With the same abbreviations as in Example 2.2, we first test

$$\begin{aligned} H_0 &: \text{model including } X1, X3, X5, \\ H_1 &: \text{model including } X4, X6, X7, X8. \end{aligned}$$

The resulting score statistic is  $s = 20.30$ , with  $\text{df} = 4$  degrees of freedom. The  $p$ -value  $\alpha = 0.00043$  leads to rejection of  $H_1$ . Interchanging  $H_0$  and  $H_1$ , one obtains  $s = 150.49$ ,  $\text{df} = 4$ ,  $\alpha = 0.00$ , so that the model including  $X1, X3, X5$  is rejected as well. This suggests that important covariates are omitted in both models.

Enlarging the covariate vector of the models to  $X1, X3, X5, X6$  (resp.,  $X4$ ),  $X5, X6, X7, X8$  and testing again leads to rejection of both models. Finally, let us test

$$\begin{aligned} H_0 &: \text{model including } X1, X3, X5, X6, X8 \\ H_1 &: \text{model including } X3, X4, X5, X6, X7, X8. \end{aligned}$$

One obtains  $s = 2.66$ ,  $\text{df} = 6$ ,  $\alpha = 0.85$ , so that  $H_0$  is not rejected. Interchanging  $H_0$  and  $H_1$  leads to rejection of the model including  $X3, X4, X5, X6, X7, X8$  in favor of the model including  $X1, X3, X5, X6, X8$ . This conclusion is consistent with the results obtained by variable selection in Example 4.1.  $\square$

## 4.4 Notes and Further Reading

### Bayesian Model Determination

There is a variety of procedures and suggestions in the literature for Bayesian model checking with diagnostic tools and model choice based on tests or selection criteria. Gelfand (1996), Raftery (1996), Gelman & Meng (1996), and several sections in Dey, Gosh & Mallick (1999) give reviews of existing methods, but there is still much work in progress.

Several strategies exist for *diagnostic checking*. Gelfand, Dey & Chang (1992) propose various diagnostics based on the cross-validation predictive distribution  $f(y_i|Y_{-i})$ , where  $Y_{-i}$  is data without observation  $y_i$ ; see also Gelfand (1996). Gelman, Meng & Stern (1995) suggest posterior predictive model checking; see also the survey of Gelman & Meng (1996). The idea is as follows: Assume  $\beta_1, \dots, \beta_s$  are samples from the posterior distribution  $p(\beta|Y)$ , obtained, for example, via MCMC. Then hypothetical replications  $y_i^{\text{rep}}$ ,  $i = 1, \dots, s$ , are generated, with  $y_i^{\text{rep}}$  drawn from the sampling distribution of  $Y$  given parameter  $\beta_i$ . To check the model, these replications are compared to the observed data via appropriate discrepancy variables. Another approach uses mixtures to embed the model under consideration into a wider class of models. Albert & Chib (1997) use this type of model choice procedures in the expanded model. Müller & Parmigiani (1995) combine model expansion with posterior-prior comparison. Based on the latent variable or utility approach for probit models (see Chapter 3, Sections 3.2, 3.3, and 3.6), Albert & Chib (1995) and Chen & Dey (1999) use latent residuals, obtained from the Gibbs outputs, for model diagnostics. Recently, Dey et al. (1998) suggested a general simulation-based method requiring only model specification and posterior simulation techniques like MCMC. It seems that there is a need for systematically evaluating the relative merits of these alternative diagnostic tools.

The classical Bayesian approach for *model choice* and *variable selection* considers a finite number of plausible models  $m = 1, \dots, M$ , assuming one of them is the “true” model. Specifying prior probabilities  $p(m)$ , the posterior probabilities are

$$p(m|Y) = \frac{f(Y|m) p(m)}{\sum_{m=1}^M f(Y|m) p(m)},$$

where  $f(Y|m) = \int f(Y|\beta_m)p(\beta_m|m) d\beta_m$  is the marginal likelihood. The model that maximizes  $p(m|Y)$ , or equivalently  $f(Y|m)p(m)$ , is selected. In particular, when comparing two models  $m_1$  and  $m_2$ , the ratio  $p(m_1|Y)/p(m_2|Y)$  is called the Bayes factor. A Bayes factor  $> 1$  supports model  $m_1$ ; a value  $< 1$  supports model  $m_2$ . Note that the models need not be nested.

Obviously, calculation of posterior probabilities for a collection of possibly complex models is a problem. Current approaches based on MCMC simulations comprise methods to estimate marginal likelihoods directly (Chib, 1995), to sample over a product space of model indicators and parameters (Carlin & Chib, 1995; Green, 1995), or combining simulations and asymptotic approximations (Di Ciccio, Kass, Raftery & Wasserman, 1997). Dellaportas, Forster & Ntzoufras (1999) discuss variable selection procedures using Bayes factors. A comparative review of MCMC methods for computing Bayes factors is given by Han & Carlin (2000), including further references on recent work.

Another problem is that Bayes factors are not well defined in the case of partially diffuse or improper priors. Such priors are quite common in complex hierarchical models (Section 5.4) or state space models (Chapter 8) with diffuse initial priors. Therefore, other criteria for model selection, in the spirit of AIC or BIC criteria, have been suggested more recently. Gelfand & Ghosh (1998) suggest a utility maximization approach, based on posterior prediction. Another criterion, using a “Bayesian deviance,” is proposed by Spiegelhalter, Best & Carlin (1998). Their deviance information criterion DIC can be viewed as a generalization of AIC, adjusting for the effective number of parameters. Both criteria are easily computed from posterior samples. However, as with diagnostic tools, we feel that additional experience is necessary for routine use.

## Robust Estimates

Pregibon (1982) considered robustified estimates, which are supposed to be less sensitive to outliers. Instead of minimization of the deviance  $D = 2 \sum_i (l(p_i) - l(\hat{p}_i))$ , Pregibon considered minimization of a transformation  $\sum_i \lambda(l_i(p_i) - l_i(\hat{p}_i))$ , where  $\lambda$  is a monotone transformation, e.g., Huber’s function  $\lambda(x) = x$  if  $x \leq c$ ,  $\lambda(x) = 2\sqrt{xc} - c$  if  $x > c$ . Alternatively, Copas (1988) considered resistant fitting by allowing that in a binary regression model contamination of the response, i.e., a transposition between 0 and 1, happens with a small probability. The misclassification maximum likelihood estimate for this model for contaminated data turns out to be more robust than the estimate. Carroll & Pederson (1993) consider more general versions of the misclassification estimate.

## Model Tests Against Smooth Alternatives

In recent years the use of nonparametric regression (see Chapter 5) to check the fit of a parametric regression model has been extensively investigated; see Azzalini, Bowman & Härdle (1989), Staniswalis & Severini (1991), Firth, Glosup & Hinkley (1991), le Cessie & van Houwelingen (1991), Eubank & Hart (1993), Azzalini & Bowman (1993), Stute (1997), Stute, González-Monteiga & Presedo-Quindimil (1998), Hart (1997), and Kauermann & Tutz (1999, 2000b).

# 5

## Semi- and Nonparametric Approaches to Regression Analysis

In this chapter we give developments that lead beyond the framework of parametric models. Instead of assuming a functional form that specifies how explanatory variables determine dependent variables, the functional form is mostly assumed to be in some way smooth, and the data are allowed to determine the appropriate functional form under weak restrictions.

In the first section the case of a continuous metric response variable is considered. The concepts are developed first for the familiar case of continuous responses because they are basic for extensions to the non-Gaussian case. The first concept is the expansion in basis functions including regression splines. The second one is based on penalization techniques that yield smoothing splines. The last one refers to local polynomial fitting. In Section 5.2 the concepts are extended to smoothing in generalized linear models. Multifunctional modelling with multiple covariates and Bayesian approaches are considered in Sections 5.3 and 5.4, respectively.

More extensive treatments of special approaches may be found in the literature. Härdle (1990a) considers nonparametric smoothing for continuous responses, Eubank (1988) gives a thorough treatment of splines, and Hastie & Tibshirani (1990) show the possibilities of generalized additive modelling. Green & Silverman (1994) give a very thorough account of the roughness penalty approach to smoothing, Simonoff (1996) introduces to smoothing techniques highlighting the smoothing of sparse categorical data, and Fan & Gijbels (1996) demonstrate the usefulness of local polynomial modelling. More recently, Loader (1999) extensively treats local regression techniques.

## 5.1 Smoothing Techniques for Continuous Responses

This section gives a short survey of smoothing techniques for the continuous case. Observations are bivariate data  $(y_i, x_i), i = 1, \dots, n$ , where the response  $y_i$  and the explanatory variable  $x_i$  are measured on an interval scale level. It is assumed that the dependence of  $y_i$  on  $x_i$  is given by

$$y_i = f(x_i) + \varepsilon_i,$$

where  $f$  is an unspecified smooth function and  $\varepsilon_i$  is a noise variable with  $E(\varepsilon_i) = 0$  that is at least approximately Gaussian. A scatterplot smoother  $\hat{f}$  is a smooth estimate of  $f$  based on observations  $(y_i, x_i)$ ,  $i = 1, \dots, n$ . Most of the smoothers considered in the following are linear. That means if one is interested in the fit at an observed value  $x_i$ , the estimate has the linear form

$$\hat{f}(x_i) = \sum_{j=1}^n s_{ij} y_j.$$

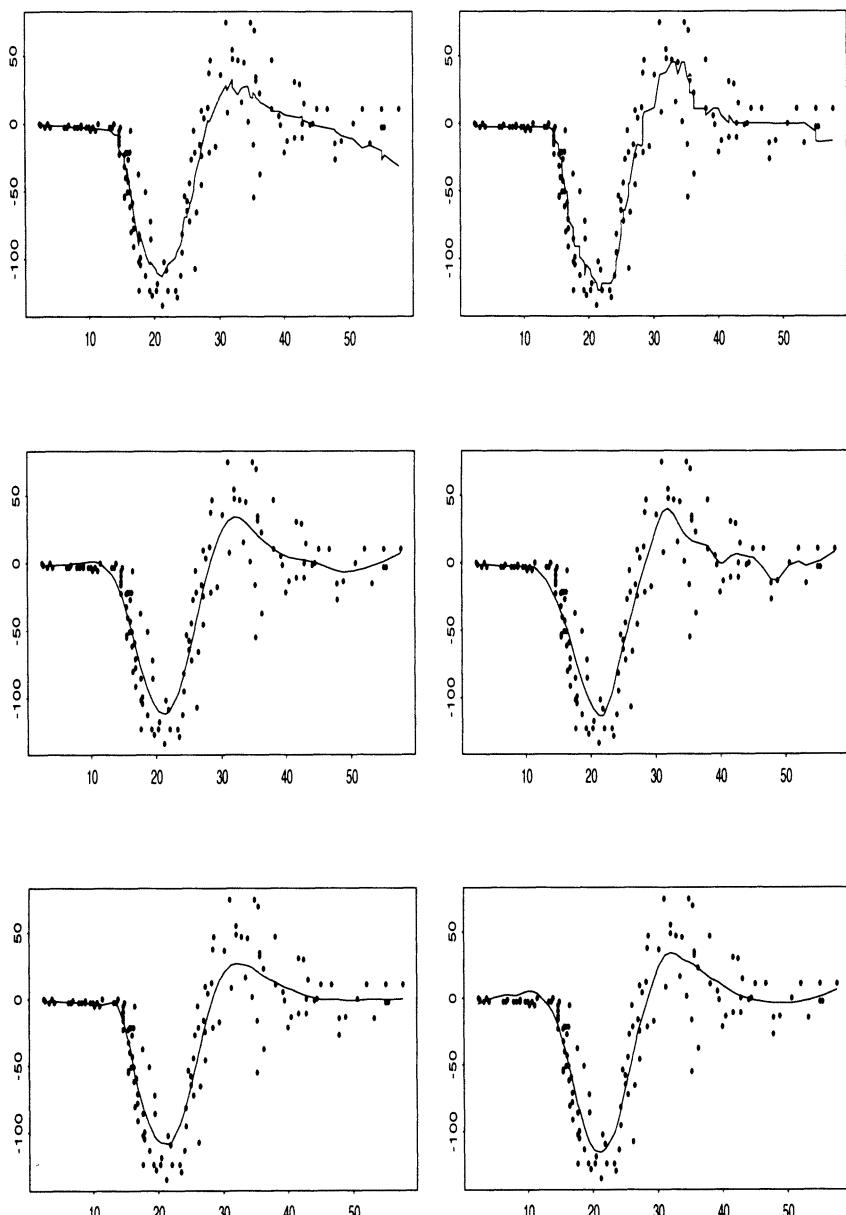
The weights  $s_{ij} = s(x_i, x_j)$  depend on the target point  $x_i$  where the response is to be estimated and on the value  $x_j$  where the response  $y_j$  is observed. To give an impression of the resulting estimates, we consider an example that uses smoothing techniques, which will be outlined in the following.

### Example 5.1: Motorcycle data

For an illustration, let us consider an example with continuous response. In a data set from Härdle (1990a, Table 1) the time dependence (in milliseconds) of head acceleration after a simulated impact with motorcycles is investigated. Figure 5.1 shows the data as dots and smoothed estimates. The top left shows the running line smoother, the top right, the running median smoother, and the second row shows the kernel smoother based on the normal kernel and the estimate resulting from cubic splines. The last row shows local linear and local quadratic fits. The amount of smoothing has been chosen so that the residual sum of the smoothers is about the same. It is seen that kernel and spline methods yield much smoother curves. In particular, the running median smoother gives a quite jagged curve.  $\square$

### 5.1.1 Regression Splines and Other Basis Functions

The simple linear regression model assumes that the function  $f(x)$  is given by  $f(x) = \beta_0 + x\beta$  with unknown coefficients  $\beta_0, \beta$ . A first step in the direction of a more flexible model is to include power functions of  $x$ , i.e.,  $x^2$ ,



**Figure 5.1.** Smoothed estimates for motorcycle data showing time ( $x$ -axis) and head acceleration ( $y$ -axis) after a simulated impact. Smoothers are running lines (top left), running medians (top right), normal kernel (middle left), cubic splines (middle right), local linear (bottom left), and local quadratic (bottom right).

$x^3$ , and so on. By including a finite number of power functions, one obtains the familiar case of polynomial regression. Although the method has been widely used, it has severe drawbacks. For example, individual observations may have a strong influence on remote areas of the curve, and increasing the degree of the polynomial may yield highly fluctuating curves. These drawbacks are no longer valid if a polynomial of low degree is fitted only within small ranges of the covariate domain. This is essentially the concept behind regression splines.

## Regression Splines

One chooses a sequence of breakpoints or knots  $\xi_1 < \dots < \xi_s$  from  $(x_1, x_n)$ , where the data are distinct and given in the form  $x_1 < \dots < x_n$ . The basic idea is to fit a polynomial piecewise in the range  $[\xi_i, \xi_{i+1})$  where  $\xi_0$  and  $\xi_{s+1}$  are additional boundary knots. In addition, the polynomials are supposed to join smoothly at the knots. For the case of cubic splines, polynomials are constrained to have continuous first and second derivatives at the knots.

For a given set of knots one representation of cubic splines is given by the so-called *truncated power series basis*

$$f(x) = \delta_0 + \delta_1 x + \delta_2 x^2 + \delta_3 x^3 + \sum_{i=1}^s \delta_{3+i} (x - \xi_i)_+^3, \quad (5.1.1)$$

where  $(x - \xi_i)_+ = \max\{0, x - \xi_i\}$ . It is easily seen that  $f(x)$  is a cubic polynomial in  $[\xi_i, \xi_{i+1})$  with continuous first and second derivatives. The smooth function (5.1.1) is a weighted sum of the  $s + 4$  functions  $P_0(x) = 1, P_1(x) = x, \dots, P_{s+3}(x) = (x - \xi_s)_+^3$ .

For a cubic spline on the interval  $[\xi_0, \xi_{s+1}]$ , the so-called natural boundary conditions specify that its second and third derivatives are zero at  $\xi_0$  and  $\xi_{s+1}$ . Then the spline is said to be a *natural cubic spline* with  $f$  linear on the two extreme intervals  $[\xi_0, \xi_1]$  and  $[\xi_s, \xi_{s+1}]$ .

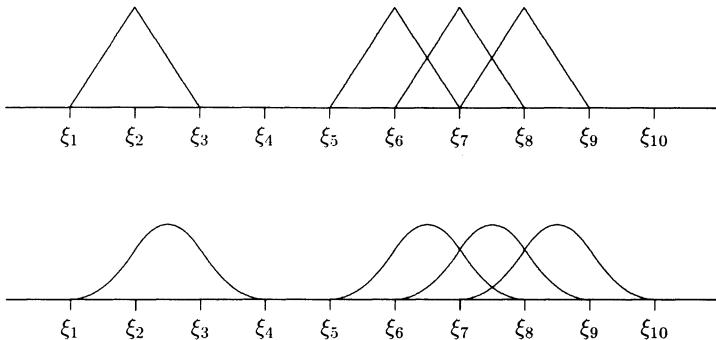
Expression (5.1.1) is linear in the  $s + 4$  parameters  $\delta_0, \dots, \delta_{3+s}$  and may be fitted by usual linear regression. However, numerically superior alternatives are available by using alternative basis functions called B-splines. Basic references for B-splines are De Boor (1978) and Schumaker (1993). Here, we follow the presentation of Eilers & Marx (1996), who present B-splines in a nutshell.

A polynomial B-spline consists of polynomial pieces between knots; the form of the spline function itself depends on the degree. In Figure 5.2 linear (degree = 1) and quadratic (degree = 2) spline functions are given for the case of equidistant knots. The simple linear spline consists of two linear pieces, one piece between knots  $\xi_i$  and  $\xi_{i+1}$ , the other from  $\xi_{i+1}$  to  $\xi_{i+2}$ . In the left part of Figure 5.2 (top) just one spline function is given, while in the right part several functions are shown. The B-spline of degree 2 (Figure 5.2, bottom) consists of three quadratic pieces, joined at two knots. It is

not equal to zero only between the knots  $\xi_i, \xi_{i+1}, \xi_{i+2}, \xi_{i+3}$ , where the first derivatives (not the second derivatives) are all equal.

These examples just illustrate the properties of B-splines. The general properties of a B-spline of degree  $q$  are as follows:

- it consists of  $q + 1$  polynomial pieces, each of degree  $q$ ,
- the polynomial pieces join at  $q$  inner knots; at the joining knots, derivatives up to order  $q - 1$  are continuous,
- the spline function is positive on a domain spanned by  $q + 2$  knots; elsewhere it is zero,
- at a given  $x$ ,  $q + 1$  B-splines are nonzero,
- except at boundaries, the spline function overlaps with  $2q$  polynomial pieces of its neighbors.



**Figure 5.2.** Illustration of B-splines bases, one isolated and several overlapping ones for degree 1 (top) and degree 2 (bottom).

A recursive algorithm to compute B-splines from B-splines of lower degree is given in De Boor (1978). Since  $q$  determines the degree of the polynomials for natural B-splines, some boundary modifications are necessary.

For a basis of B-spline functions (of fixed degree), the fitted function has the form

$$f(x) = \sum_j \delta_j B_j(x), \quad (5.1.2)$$

where the  $\delta_j$  stand for weights and  $B_j(x)$  is the  $j$ th B-spline. Thus, the function  $f$  is approximated by a weighted sum of known functions with compact local support.

Cubic splines are a common choice. A more challenging task is the selection of knots. The number and position of knots strongly determine the

degree of smoothing. The position of knots may be chosen uniformly over the data (cardinal splines), at appropriate quantiles or by more complex data-driven schemes. For a detailed discussion of these issues, see Friedman & Silverman (1989).

## Other Basis Functions

The general linear basis function approach is given by the approximation

$$f(x) = \sum_{j=1}^m \delta_j B_j(x), \quad (5.1.3)$$

where  $B_j(x)$  are known basis functions. The basis functions can be power functions  $B_j(x) = x^j$  or, as in the case of regression splines, polynomials with compact support.

An alternative choice of a set of basis functions is given by the Fourier basis  $B(x) = 1$ ,  $B_{2k}(x) = \sqrt{2} \cos(2\pi kx)$ ,  $B_{2k+1}(x) = \sqrt{2} \sin(2\pi kx)$ . The Fourier basis is an orthonormal basis, fulfilling  $\int B_i(x)B_j(x)dx = \delta_{ij}$  with  $\delta_{ij}$  being the Kronecker delta function. It is useful, in particular, for modelling seasonal effects if the covariate is time. Basis functions often used in the neural network community are the so-called radial basis functions, which have the form of a localized density, e.g., the Gaussian basis function is given by

$$B_j(x) = \exp\left(-\frac{|x - \mu_j|^2}{2h_j^2}\right), \quad (5.1.4)$$

where  $\mu_j$  denotes the center of the basis function and  $h_j$  determines the spread. It should be noted that in (5.1.3) basis functions are not necessarily smooth. If jumps are possible this may be modelled by appropriate basis functions. A flexible tool that often allows an economic expansion in a few basis functions are wavelet bases. They are orthonormal and, in contrast to bases like the Fourier functions or Gaussian radial basis functions, have local support (see, e.g., Strang, 1993; Mallat, 1989; Nason & Silverman, 2000). For metric response the theory of wavelets is well developed and outlined in the books of Ogden (1997) and Vidakovic (1999).

The linear basis function approach given by (5.1.3) looks like simple parametric modelling. For fixed basis functions one just has to estimate the parameters  $\delta_j$ ,  $j = 1, \dots$ , e.g., by minimizing the sum of squared residuals  $\sum_{i=1}^n (y_i - f(x_i))^2$ .

What makes (5.1.3) a flexible nonparametric approach is the choice of the basis functions, which itself may depend on parameters that are chosen adaptively from the data. For the radial basis functions (5.1.4) the center  $\mu_s$  and the width  $h_s$  are adaptive parameters of this type. For regression splines flexibility and smoothness are obtained by making the functions specific to

certain intervals of the  $x$ -range with intervals (knots) being determined by the data themselves.

However, even if the set of basis functions is specified, the questions of which basis functions and how many should be taken remain. In particular, the number of basis functions controls the smoothness of the estimated function. An approach that suggests itself is to include basis functions and control for the residuals. This is done, e.g., in extended linear modelling (see Section 5.3.1). Another way to control smoothness is based on regularization, with penalization of smoothness being one possibility.

## Penalization

One approach to optimizing the fitting procedure in a nonparametric setting is to alter the usual fitting procedure by adding a penalty term. Thus, instead of fitting by simple least squares, one may minimize the penalized (weighted) sum of squares

$$\sum_{i=1}^n w_i(y_i - f(x_i))^2 + \lambda J(f), \quad (5.1.5)$$

where  $f(x)$  has the functional form  $f(x) = \sum_j \delta_j B_j(x)$  and  $J(f)$  is a roughness functional. An often used roughness functional is the integrated squared second derivative  $J(f) = \int (f''(u))^2 du$  as a global measure for curvature or roughness of the function  $f$ . Thus, (5.1.5) is composed of two parts, the residual of squares, which reflects the distance between data and estimates, and a term that penalizes the roughness of the estimates. The parameter  $\lambda \geq 0$  is a smoothing parameter that controls the trade-off between the smoothness of the curve and the faith with the data. Large values of the smoothing parameter  $\lambda$  give large weight to the penalty term, therefore enforcing smooth functions with small variance but possibly high bias. For rather small  $\lambda$ , the function  $f$  will be closer to the data. For distinct data  $x_i$  the weights  $w_i$  in (5.1.5) may be chosen by  $w_i = 1$ . They are introduced here because they will be necessary for extensions to the non-Gaussian case.

If  $J(f) = \int (f''(x))^2 dx$ , one obtains for B-splines of degree 3 a penalty function given by

$$J(f) = \delta' K_B \delta,$$

with  $K_B = (k_{ij})$  a banded matrix with entries  $k_{ij} = \int B_i''(x) B_j''(x) dx$  and  $\delta' = (\delta_1, \dots)$  the vector of parameters. Then, with  $y' = (y_1, \dots, y_n)$ , the criterion (5.1.5) is given by

$$(y - B\delta)' W(y - B\delta) + \lambda \delta' K_B \delta,$$

where  $B = (b_{ij})$  is a design matrix with entries  $b_{ij} = B_j(x_i)$  and  $W = \text{diag}(w_1, \dots, w_n)$  is the diagonal matrix of weights. Minimization yields

$$\hat{\delta} = (B'WB + \lambda K_B)^{-1} B'W\delta, \quad (5.1.6)$$

which may be computed efficiently by exploiting that the matrices  $B$  and  $K_B$  are banded.

Eilers & Marx (1996) use a penalty of the form

$$J_d = \sum_{j=d+1}^m (\Delta^d \delta_j)^2 = \delta' K_d \delta,$$

where  $\Delta$  is the difference operator operating on adjacent B-spline coefficients, i.e.,  $\Delta \delta_j = \delta_j - \delta_{j-1}$ ,  $\Delta^2 \delta_j = \Delta(\delta_j - \delta_{j-1}) = \delta_j - 2\delta_{j-1} + \delta_{j-2}$ . The matrix  $K_d$  has a banded structure and follows from representing the differences in matrix form. The vector of first differences is given by  $D_1\delta$ , with

$$D_1 = \begin{pmatrix} -1 & 1 & 0 & 0 & \dots \\ 0 & -1 & 1 & 0 & \dots \\ 0 & & -1 & 1 & \dots \\ \vdots & & & & \end{pmatrix}.$$

The differences of order  $d$  are given as  $d$ th-order polynomial contrasts by  $D_d\delta = D_1 D_{d-1}\delta$ . Then the sum of squared differences  $J_d$  takes the simple form  $\delta' K_d \delta$  with  $K_d = D'_d D_d$ . Eilers & Marx (1996) proposed using a large number of equally spaced knots but prevented overfitting by attaching the difference penalty  $J_d$  on adjacent B-spline coefficients, thereby ensuring smoothness of the estimated function. The combination of B-splines and penalization yields so-called *P-splines*. For a thorough treatment of the properties, see Eilers & Marx (1996, with discussion).

Alternative penalties in common use have the form

$$\sum_j |\delta_j|^q, \quad q > 0,$$

which for  $q = 2$  is equivalent to  $J_d$  with  $d = 0$ . The case  $q = 2$  is also familiar from ridge regression (Hoerl & Kennard, 1970). The choice  $q = 1$  is known as soft thresholding (Bickel, 1983). In connection with wavelet bases soft thresholding has attractive minimax properties (Donoho & Johnstone, 1995; Donoho et al., 1995; Klinger, 1998).

In the next section the concept of penalization will be considered again but without the restriction that  $f$  is a sum of basis functions. The result are so-called smoothing splines, which of course are strongly connected to regression splines.

### 5.1.2 Smoothing Splines

Smoothed estimators may be considered compromises between faith with the data and reduced roughness caused by the noise in the data. This view is made explicit in the construction of smoothing splines. The starting point is the following minimization problem: Find the function  $f$  that minimizes

$$\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int_{-\infty}^{\infty} (f''(u))^2 du \quad (5.1.7)$$

where  $f$  has continuous first and second derivatives  $f'$  and  $f''$ , and  $f''$  is quadratically integrable. The first term in (5.1.7) is the residual sum of squares, and the second term penalizes roughness in the way it is done in (5.1.5).

The solution  $\hat{f}_\lambda$  of the penalized least-squares estimation problem (5.1.7) is a *natural cubic smoothing spline* with knots at each distinct  $x_i$  (Reinsch, 1967). This means  $\hat{f}_\lambda$  is a cubic polynomial between successive  $x$ -values and first and second derivatives are continuous at the observation points. At the boundary points the second derivative is zero. Since these piecewise cubic polynomials are defined by a finite number of parameters, optimization of (5.1.7) with respect to a set of functions actually reduces to a finite-dimensional optimization problem. It can be shown (Reinsch, 1967; De Boor, 1978, Ch. 14) that minimization of (5.1.7) is equivalent to minimizing the *penalized least-squares (PLS) criterion*

$$PLS(f) = (y - f)'(y - f) + \lambda f' K f, \quad (5.1.8)$$

where  $y' = (y_1, \dots, y_n)$  are the data and  $f' = (f(x_1), \dots, f(x_n))$  for  $x_1 < \dots < x_n$  is the vector of evaluations of the function  $f(x)$ . The penalty matrix  $K$  has a special structure and is given by

$$K = D' C^{-1} D, \quad (5.1.9)$$

where  $D = (d_{ij})$  is an  $(n - 2, n)$  upper-tridiagonal matrix,

$$D = \begin{pmatrix} \frac{1}{h_1} & -(\frac{1}{h_1} + \frac{1}{h_2}) & & \frac{1}{h_2} & & \\ & \frac{1}{h_2} & -(\frac{1}{h_2} + \frac{1}{h_3}) & & \frac{1}{h_3} & \\ & & \ddots & \ddots & & \ddots \\ & & & \frac{1}{h_{n-2}} & -(\frac{1}{h_{n-2}} + \frac{1}{h_{n-1}}) & \frac{1}{h_{n-1}} \end{pmatrix},$$

$h_i := x_{i+1} - x_i$ , and  $C = (c_{ij})$  is an  $(n - 2, n - 2)$  tridiagonal symmetric matrix,

$$C = \frac{1}{6} \begin{pmatrix} 2(h_1 + h_2) & & h_2 & & & \\ & h_2 & 2(h_2 + h_3) & h_3 & & \\ & & \ddots & & & \\ & & & \ddots & & \\ & & & & h_{n-2} & \\ & & & & & 2(h_{n-2} + h_{n-1}) \end{pmatrix}.$$

This specific form of the penalty matrix  $K$  can be derived by evaluating  $\int(f''(x))^2 dx$  and from the fact that  $f(x)$  is a natural cubic spline; see, e.g., Green & Silverman (1994). The minimizing function  $\hat{f}_\lambda$  can be obtained by equating the vector  $-2(y - f) + 2\lambda K f$  of first derivates of  $PLS(f)$  to zero. This yields the *linear smoother*

$$\hat{f}_\lambda = (I + \lambda K)^{-1} y, \quad (5.1.10)$$

with smoothing matrix  $S_\lambda = (I + \lambda K)^{-1}$ ,  $I$  denoting the identity matrix.

For computational reasons,  $\hat{f}_\lambda$  and the smoothing matrix  $S_\lambda$  are generally not computed directly by inversion of  $I + \lambda K$  (note that  $S_\lambda$  is an  $(n \times n)$ -matrix). Instead,  $\hat{f}_\lambda$  is computed indirectly in two steps, making efficient use of the band matrices  $D$  and  $C$ ; see Reinsch (1967) and De Boor (1978, Ch. 14) for details. Based on B-splines and knots at each distinct  $x_i$ ,  $\hat{f}_\lambda$  may also be estimated by the use of (5.1.6), which may be considered as an alternative to the Reinsch algorithm.

In (5.1.8) the distance between data and estimator is measured by a simple quadratic function. More generally, a weighted quadratic distance may be used. For given diagonal weight matrix  $W$ , a weighted penalized least-squares criterion is given by

$$(y - f)' W (y - f) + \lambda f' K f. \quad (5.1.11)$$

The solution is again a cubic smoothing spline, with the vector  $\hat{f}_\lambda$  of fitted values now given by

$$\hat{f}_\lambda = (W + \lambda K)^{-1} W y. \quad (5.1.12)$$

Thus, only the identity matrix  $I$  is replaced by the diagonal weight matrix  $W$ . Again the solution  $\hat{f}_\lambda$  of (5.1.11) is computed indirectly. The weighted spline version (5.1.11), (5.1.12) forms a basic tool for simple spline smoothing in generalized linear models (Section 5.2.2). It is also useful if several responses are observed at point  $x_i$ . When the mean is used as response, it has to be weighted by the number of responses used to compute the mean value.

### 5.1.3 Local Estimators

#### Simple Neighborhood Smoothers

A simple device for the estimation of  $f(x_i)$  is to use the average of response values in the neighborhood of  $x_i$ . These *local average estimates* have the form

$$\hat{f}(x_i) = \text{Ave}_{j \in N(x_i)}(y_j),$$

where Ave is an averaging operator and  $N(x_i)$  is a neighborhood of  $x_i$ . The extension of the neighborhood is determined by the *span* or *window size*  $w$ , which denotes the proportion of total points in a neighborhood.

Let the window size  $w$  range over  $(0,1)$  and let the integer part of  $wn$ , denoted by  $[wn]$ , be odd. Then a symmetric neighborhood may be constructed by

$$N(x_i) = \left\{ \max\left\{i - \frac{[wn] - 1}{2}, 1\right\}, \dots, i - 1, i, i + 1, \dots, \min\left\{i + \frac{[wn] - 1}{2}, n\right\} \right\}.$$

$N(x_i)$  gives the indices of the ordered data  $x_1 < \dots < x_n$ . It contains  $x_i$  and  $([wn] - 1)/2$  points on either side of  $x_i$  if  $x_i$  is from the middle of the data. The neighborhood is smaller near the endpoints, e.g., at  $x_1, x_n$ , which leads to quite biased estimates at the boundary. Of course,  $w$  determines the smoothness of the estimate. If  $w$  is small, the estimate is very rough; for large  $w$  the estimate is a smooth function.

A special case of a local average estimate is the *running mean*, where Ave stands for arithmetic mean. Alternatively, the mean may be replaced by the *median*, yielding an estimate that is more resistant to outliers. A drawback of the median is that the resulting smoother is nonlinear.

Another simple smoother is the *running-line smoother*. Instead of computing the mean, a linear term

$$\hat{f}(x_i) = \hat{\alpha}_i + \hat{\beta}_i x_i$$

is fitted where  $\hat{\alpha}_i, \hat{\beta}_i$  are the least-squares estimates for the data points in the neighborhood  $N(x_i)$ . In comparison to the running mean, the running-line smoother reduces bias near the endpoints. However, by using simple neighborhood smoothing, both smoothers, running mean and running line may produce curves that are quite jagged.

If the target point  $x$  is not from the sample  $x_1, \dots, x_n$ , one may interpolate linearly between the fit of the two  $\hat{y}_i$  values, which are observed at predictor values from the sample adjacent to  $x$ . Alternatively, one may use varying nonsymmetric neighborhoods. For  $k \in \mathbb{N}$  the *k-nearest neighborhood* (*k*-NN) estimate is a weighted average based on the varying neighborhood

$$N(x) = \{i | x_i \text{ is one of the } k\text{-nearest observations to } x\},$$

where “near” is defined by a distance measure  $d(x, x_i)$ . Then the degree of smoothing is determined by  $k$ . The proportion of points in each neighborhood is given by  $w = k/n$ . For example, the linear k-NN estimate has the form

$$\hat{f}(x) = \sum_{i=1}^n s(x, x_i) y_i$$

with weights

$$s(x, x_i) = \begin{cases} 1/k & \text{if } x_i \in N(x), \\ 0 & \text{otherwise.} \end{cases}$$

The weights fulfill the condition  $\sum_i s(x, x_i) = 1$ . The estimate is not restricted to unidimensional  $x$ -values. By appropriate choice of a distance measure, e.g., the Euclidean distance  $d(x, x_i) = \|x_i - x\|^2$ , the method may be applied to vector-valued  $x$ .

## Local Regression

A smoothing concept of increasing importance is the local fitting of a model from a parametric family. Estimation of the expected response  $f(x)$  at a target value  $x_0$  is obtained by fitting the model locally in a neighborhood around  $x_0$ . Although  $f(x)$  may not be a member of the parametric family, locally the parametric model is often a good approximation. In the case of *local polynomial regression* the procedure may be motivated by Taylor expansion. If a function is sufficiently smooth, i.e., the derivatives  $f^{(s)}(x) = d^s f(x)/dx^s$  exist, it may be approximated at  $x$  by a polynomial

$$\begin{aligned} f(x_i) &\approx \sum_{s=0}^m (f^{(s)}(x)/s!)(x_i - x)^s \\ &= f(x) + f'(x)(x_i - x) + (f''(x)/2)(x_i - x)^2 + \dots \\ &= \beta_0 + \sum_{s=1}^m \beta_s (x_i - x)^s. \end{aligned}$$

Thus, locally a polynomial fit is justified. One minimizes for fixed value  $x$

$$\sum_{i=1}^n \left( y_i - \beta_0 - \sum_{s=1}^m \beta_s (x_i - x)^s \right)^2 w_\lambda(x, x_i), \quad (5.1.13)$$

where  $w_\lambda(x, x_i)$  is a weighting function depending on the target value  $x$  and the measurement point  $x_i$  and, in addition, contains a smoothing parameter

$\lambda$ . Thus, locally around  $x$  a polynomial of degree  $m$  is fitted by using the familiar technique of least-squares fitting.

If  $\hat{\beta}_i$ ,  $i = 0, \dots, m$ , minimizes (5.1.13), the estimate  $\hat{f}(x)$  results from computing the polynomial  $\hat{\beta}_0 + \sum_s \hat{\beta}_s(x_i - x)^s$  at the value  $x_i = x$ , yielding

$$\hat{f}(x) = \hat{\beta}_0.$$

As is seen from the Taylor expansion, the derivatives of  $f$  may be estimated by

$$\hat{f}_\lambda^{(s)} = s! \hat{\beta}_s.$$

In order to get an estimate for the function  $f(x)$ , one has to minimize (5.1.13) for a grid of target values  $x$ . For each target value one gets specific parameter estimates  $\hat{\beta}_i$ . Thus, the coefficients  $\hat{\beta}_i$  actually depend on  $x$ , which is suppressed in the notation.

The most often used weights in (5.1.13) are kernel weights. If these weights are used, the strong connection between local polynomial regression and so-called kernel smoothing becomes obvious.

A kernel or density function  $K$  is a continuous symmetric function with

$$\int K(u)du = 1. \quad (5.1.14)$$

Widely used kernel functions are the Gaussian kernel, where  $K$  is the standard Gaussian density, the Epanechnikov kernel

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2) & \text{if } |u| \leq 1, \\ 0 & \text{otherwise} \end{cases} \quad (5.1.15)$$

(Epanechnikov, 1969), or the minimum variance kernel

$$K(u) = \begin{cases} \frac{3}{8}(3 - 5u^2) & \text{if } |u| \leq 1, \\ 0 & \text{otherwise.} \end{cases} \quad (5.1.16)$$

The weights  $w_\lambda(x, x_i)$  in (5.1.13) are chosen to be proportional to

$$K\left(\frac{x - x_i}{\lambda}\right).$$

This means that for reasonable kernel functions  $K$  (e.g., a Gaussian density) the weights will decrease with increasing distance  $|x - x_i|$ . The window-width

or bandwidth  $\lambda$  determines how fast the weights decrease. For small  $\lambda$ , only values in the immediate neighborhood of  $x$  will be influential; for large  $\lambda$ , values more distant from  $x$  may also influence the estimate.

The simplest polynomial that may be fitted locally is a constant (for polynomial of degree zero), resulting in locally constant fitting. The minimization of

$$\sum_i (y_i - \beta_0)^2 K\left(\frac{x - x_i}{\lambda}\right)$$

yields the linear estimate

$$\hat{f}(x) = \sum_{i=1}^n s(x, x_i) y_i, \quad (5.1.17)$$

where the weights on the observed responses are given by

$$s(x, x_i) = \frac{K\left(\frac{x - x_i}{\lambda}\right) / n\lambda}{\sum_{j=1}^n K\left(\frac{x - x_j}{\lambda}\right) / n\lambda}. \quad (5.1.18)$$

The kernel regression estimate (5.1.17) with weights (5.1.18) has been proposed by Nadaraya (1964) and Watson (1964) and is called *Nadaraya-Watson kernel estimator*. It may be derived in a quite different way by considering  $f(x) = E(y|x) = \int y g(x, y)/g(x) dy$ , where  $g(x, y)$  and  $g(x)$  are the densities of  $(x, y)$  and  $x$ , respectively. If  $g(x, y)$  is replaced by the density estimate

$$\hat{g}(x, y) = \frac{1}{n\lambda} \sum_{i=1}^n K\left(\frac{x - x_i}{\lambda}\right) K_y\left(\frac{y - y_i}{\lambda}\right),$$

and  $g(x)$  is replaced by the corresponding density estimate

$$\hat{g}(x) = \frac{1}{n\lambda} \sum_{i=1}^n K\left(\frac{x - x_i}{\lambda}\right),$$

the Nadaraya-Watson estimate follows from  $\int K_y(u) du = 1$  and  $\int u K_y(u) du = 0$ . The denominator in (5.1.18) is just the density estimator of  $g(x)$ . It yields the normalization of the weights that fulfill  $\sum_{i=1}^n s(x, x_i) = 1$ .

In the general case of local polynomial regression, the minimization of

$$\sum_{i=1}^n \left( y_i - \beta_0 - \sum_{s=1}^m \beta_s (x_i - x) \right)^2 K\left(\frac{x_i - x}{\lambda}\right)$$

is equivalent to solving a weighted least-squares problem. With

$$Z_x = \begin{pmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^m \\ \vdots & \vdots & & \vdots \\ 1 & x_n - x & & (x_n - x)^m \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

and weight matrix

$$W_x = \text{diag} \left( K \left( \frac{x_i - x}{\lambda} \right) \right),$$

one obtains for  $\beta' = (\beta_0, \dots, \beta_m)$  the estimate

$$\hat{\beta} = (Z_x' W_x Z_x)^{-1} Z_x' W_x y. \quad (5.1.19)$$

The estimate  $\hat{f}(x) = \hat{\beta}_0$  results from multiplication with the unit vector  $e'_1 = (1, 0, \dots, 0)$  and is given by

$$\hat{f}(x) = e'_1 (Z_x' W_x Z_x)^{-1} Z_x' W_x y = s_{xy}$$

with smoothing weights  $s_{xy} = e'_1 (Z_x' W_x Z_x)^{-1} Z_x W_x$ . It is obvious that  $\hat{f}(x)$  is a linear smoother, i.e., linear in the observed response  $y$ .

As already noted, the local constant fit yields the Nadaraya-Watson estimate  $f_\lambda(x) = \sum_i s(x, x_i) y_i$  with weights  $s(x, x_i)$  given by (5.1.18). An alternative weight function  $s(x, x_i)$  has been considered in a series of papers by Gasser & Müller (1979, 1984), Müller (1984), and Gasser, Müller & Mammitzsch (1985). It is based on ordered values  $0 = x_1 < \cdots < x_n = 1$ , with  $s_i \in [x_i, x_{i+1}]$ ,  $s_0 = 0$ ,  $s_n = 1$ , chosen between adjacent  $x$ -values. The *Gasser-Müller weights* given by

$$s(x_0, x_i) = \frac{1}{\lambda} \int_{s_{i-1}}^{s_i} K\left(\frac{x_0 - u}{\lambda}\right) du \quad (5.1.20)$$

again sum to 1. In comparison to the Nadaraya-Watson weights, the Gasser-Müller weights are an improvement because one has less bias, but this advantage is obtained at the expense of increasing the variability. Other weight functions related to Gasser-Müller weights have been given by Priestley & Chao (1972) and Benedetti (1977).

### Bias-Variance Trade-off

Smoothing always means a compromise between bias and variance. For local modelling when information from the neighborhood of a target value is borrowed, this becomes obvious by looking at the size of the neighborhood. If

the window-width  $\lambda$  in local regression becomes large, many observations are used, so bias should be high and variance low. If, in contrast,  $\lambda \rightarrow 0$ , the estimate becomes increasingly local, only few observations get substantial weights, with the consequence of low bias but high variance. For a point  $x$  away from the boundary region one obtains the bias, conditional on observations  $x_1, \dots, x_n$ , asymptotically ( $\lambda \rightarrow 0, n\lambda \rightarrow \infty$ ) as

$$E(\hat{f}_\lambda(x) - f(x)) = \frac{\lambda^{m+1} f^{(m+1)}(x)}{(m+1)!} \mu_{m+1}(K) + o_p(\lambda^{m+1})$$

if  $m$  is odd, and

$$\begin{aligned} E(\hat{f}_\lambda(x) - f(x)) &= \left[ \frac{\lambda^{m+2} f^{(m+1)}(x) g'(x)}{(m+1)! g(x)} + \frac{\lambda^{m+2} f^{(m+2)}(x)}{(m+2)!} \right] \\ &\quad \cdot \mu_{m+2}(K) + o_p(\lambda^{m+2}) \end{aligned}$$

if  $m$  is even. Here  $f^{(q)}$  denotes the  $q$ th derivative and  $\mu_q(K) = \int u^q K(u) du$  depends only on the kernel. For  $m$  odd,  $K$  has to be a kernel of order  $m+1$ ; for  $m$  even, it has to be of order  $m+2$ . A kernel of order  $s$  is defined by  $\mu_0(K) = 1, \mu_r(K) = 0, r = 0, \dots, s-1, \mu_s(K) \neq 0$ .

The corresponding asymptotic variance is given by

$$\text{var}(\hat{f}_\lambda(x)) = \frac{\sigma^2(x)}{n\lambda g(x)} \int K^2(u) du + o_p((n\lambda)^{-1}),$$

where  $\sigma^2(x)$  is the variance of  $y$  at target value  $x$ . For more details and additional assumptions, see e.g. Simonoff (1996, Ch. 5) and Fan & Gijbels (1996, Ch. 3).

The principal trade-off between the bias and the variance of the estimate is reflected by the mean squared error

$$E(\hat{f}_\lambda(x) - f(x))^2 = \text{var} \hat{f}_\lambda(x) + [E\hat{f}_\lambda(x) - f(x)]^2. \quad (5.1.21)$$

Asymptotic bias and variance show several interesting features. As was to be expected, in the leading term of the bias the smoothing parameter is found in the numerator and for variance it is found in the denominator. Thus, e.g., for  $\lambda \rightarrow 0$  the variance becomes large whereas the bias becomes low.

There is a difference between  $m$  odd and  $m$  even, leading to the same order of the bias for  $m = 0$  (constant) and  $m = 1$  (local linear), as well as for  $m = 2$  and  $m = 3$ , and so on. For example, for  $m = 0$  as well as for  $m = 1$ , the leading term of the bias contains  $\lambda^2$ , whereas for  $m = 2$  and  $m = 3$  one obtains  $\lambda^4$ . As is seen from the formulas, for  $m$  odd the bias does not depend on the density  $g(x)$ ; in this sense the estimate is *design adaptive*. This does not hold for  $m$  even. The Nadaraya-Watson estimate that corresponds to degree zero fitting ( $m = 0$ ) as compared to local linear

fit ( $m = 1$ ) has an additional term containing  $f^{(1)}(x) g'(x)/g(x)$  in the bias approximation. This term contains the density  $g(x)$  in the denominator, meaning that bias is lower if the density  $g(x)$  is high. Moreover, it contains the slope  $f^{(1)}(x)$  in the numerator, which is negligible when local linear fitting is applied. In addition, local polynomial fitting adapts to both random and fixed design. The expressions for bias and variance remain valid for fixed designs. In contrast, for the Gasser-Müller kernel, although having low bias, the unconditional variance is higher by a factor of 1.5 for random designs.

A further advantage of local polynomial fitting is the so-called boundary carpentry, meaning that local fitting automatically adapts to the special data situation occurring at the boundary of the design. In kernel regression the increase of bias near the boundary has led to the development of specific boundary kernels that modify the weighting scheme at the boundary. For local polynomial fitting no boundary modifications are necessary.

Although there are many advantages of fitting a polynomial of degree  $m \geq 1$  over simple kernel regression, there are also some problems in the practical application of the technique. The matrix  $Z_x' W_x Z_x$  in (5.1.19) has to be invertible, which might be a problem for small bandwidth  $\lambda$  in connection with the random design. Inversion may also be numerically unstable (for remedies, see Seifert & Gasser, 1996). Moreover, the use of kernels with finite support leads to estimators with infinite unconditional variance. Although the bias in boundary regions is corrected, the use of polynomials of higher orders tends to increase the variance in boundary regions. In practice local linear estimation is used most often. It has odd order and corrects the bias in the boundary region automatically. Local linear and quadratic estimates with locally varying bandwidth based on nearest neighbors are sometimes called *loess* or *lowess* estimators.

Local polynomial regression has a long tradition. Cleveland & Loader (1996) give a nice historical review dating the origins back to the 19th century. Modern work that reflects the properties of estimates within a model framework starts with Stone (1977) and Cleveland (1979). Hastie & Loader (1993) give a good introduction, highlighting the advantages over classical kernel regression. A comprehensive and thorough presentation of local polynomial regression has been given by Fan & Gijbels (1996). Loader (1999) gives an extensive treatment of local regression and likelihood.

## Relation to Other Smoothers

Smoothing procedures, like nearest neighborhood estimates and spline smoothing, are strongly related to kernel smoothing. Let the  $x$ -values be equidistant with  $x_i = i/n, i = 1, \dots, n$ . Then in the middle of the data the nearest neighborhood estimate is equivalent to a kernel estimate with weights (5.1.18) based on the uniform kernel  $K(u) = 1$  for  $u \in [-0.5, 0.5]$ . The smoothing parameter has to be chosen by  $\lambda = k/n$ , where  $k$  is the

number of neighborhood values. A combination of kernels and neighborhood estimates is given by using, e.g., the Nadaraya-Watson weights with smoothing parameter

$$\lambda_x = d(x, x^{(k)}),$$

where  $d(x, x^{(k)})$  is the distance between the target point  $x$  and its  $k$ th-nearest neighbor. Then the bandwidth is locally chosen with small values if the data are dense and strong smoothing if the data are sparse (see Stone, 1977; Mack, 1981). Alternative approaches to locally varying bandwidths are considered by Müller & Stadtmüller (1987) and Staniswalis (1989). Instead of varying smoothing parameters, the distance may be defined alternatively. The distance  $(x - x_i)$  in the kernel may be substituted by  $\hat{F}(x) - \hat{F}(x_i)$  where  $\hat{F}$  denotes the empirical distribution function (see Yang, 1981; Stute, 1984). Although it is not obvious, (cubic) spline smoothing yields linear estimates  $\hat{f}_\lambda(x) = \sum_i s(x, x_i)y_i$  with “spline weights”  $s$ . For  $x = x_i$  this is already seen from (5.1.10). Silverman (1984) showed that under regularity conditions the weights may be approximated by kernel weights based on a symmetric kernel  $K_s$  with locally varying bandwidths. For plots of the kernel  $K_s$  and effective spline weight functions, see Silverman (1984) and Engle, Granger, Rice & Weiss (1986).

### 5.1.4 Selection of Smoothing Parameters

The mean squared error (5.1.21) is a local criterion considered for a fixed point  $x$ . Global distance measures are the *average squared error*

$$ASE(\lambda) = \frac{1}{n} \sum_{i=1}^n \{\hat{f}_\lambda(x_i) - f(x_i)\}^2 \quad (5.1.22)$$

and the *mean average squared error*

$$MASE(\lambda) = \frac{1}{n} \sum_{i=1}^n E\{\hat{f}_\lambda(x_i) - f(x_i)\}^2. \quad (5.1.23)$$

The latter may be considered a discrete approximation to the *integrated squared error*

$$ISE(\lambda) = \int (\hat{f}_\lambda(x) - f(x))^2 dx. \quad (5.1.24)$$

A prediction-oriented measure is the *average predictive squared error*

$$PSE(\lambda) = \frac{1}{n} \sum_{i=1}^n E\{\hat{f}_\lambda(x_i) - y_i^*\}^2, \quad (5.1.25)$$

where  $y_i^*$  is a new independent observation at  $x_i$ . For the model  $y_i = f(x_i) + \varepsilon_i$  with  $\sigma^2 = \text{var}(\varepsilon_i)$ , the connection to  $MASE$  is given by  $PSE = MASE + \sigma^2$ .

Minimization of  $ASE$  yields a smoothing parameter that is oriented on the data at hand whereas minimization of  $MASE$  and  $PSE$  aims at the average over all possible data sets. Härdle, Hall & Marron (1988) show for kernel smoothing based on Priestley-Chao weights that for  $n \rightarrow \infty$  and equally spaced design points  $x_i$   $ASE$  and  $MASE$  are very close.

A naive measure that may actually be minimized for the data at hand is the *average squared residual*

$$ASR(\lambda) = \frac{1}{n} \sum_i \{y_i - \hat{f}_\lambda(x_i)\}^2. \quad (5.1.26)$$

However, minimization of the  $ASR$  does not yield a good approximation of global theoretical measures because the data are used to construct the estimate and simultaneously to assess the properties.

A more appropriate measure is the *cross-validation criterion*

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n \{y_i - \hat{f}_\lambda^{-i}(x_i)\}^2, \quad (5.1.27)$$

where  $\hat{f}_\lambda^{-i}(x_i)$  is a version of  $\hat{f}_\lambda(x_i)$  computed by leaving out the  $i$ th data point. A simple version of the leaving-one-out estimate is obtained by correcting the weights computed for the full set of  $n$  data points. For linear smoothers  $\hat{f}(x_i) = \sum_j s_{ij} y_j$ , one may choose

$$\hat{f}_\lambda^{-i}(x_i) = \sum_{j \neq i} \frac{s_{ij}}{1 - s_{ii}} y_j, \quad (5.1.28)$$

where the modified weights  $s_{ij}/(1 - s_{ii})$  now sum to 1. Thus, one gets the simple form

$$\hat{f}_\lambda^{-i}(x_i) = \frac{1}{1 - s_{ii}} \hat{f}_\lambda(x_i) - \frac{s_{ii}}{1 - s_{ii}} y_i.$$

Then the essential term  $y_i - \hat{f}_\lambda^{-i}(x_i)$  in CV is given by

$$y_i - \hat{f}_\lambda^{-i}(x_i) = \frac{y_i - \hat{f}_\lambda(x_i)}{1 - s_{ii}}$$

and may be computed from the regular fit  $\hat{f}_\lambda(x_i)$  based on  $n$  observations and the weight  $s_{ii}$ . For spline smoothing and kernel smoothing with Nadaraya-Watson weights, (5.1.28) is equivalent to the estimate based on  $n-1$  observations. This does not hold for all smoothers because, for example, a smoother designed for an equidistant grid changes when one observation is left out. By using (5.1.28) one gets the criterion

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1 - s_{ii}} \right\}^2.$$

Generalized cross-validation as introduced by Craven & Wahba (1979) replaces  $s_{ii}$  by the average  $\sum_i s_{ii}/n$ . The resulting criterion

$$GCV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1 - \frac{1}{n} \sum_j s_{jj}} \right\}^2 = ASR(\lambda)(1 - \frac{1}{n} \sum_i s_{ii})^{-2}$$

is easier to compute as it is the simple averaged squared error corrected by a factor. For weighted cubic spline smoothing as in (5.1.11), the squared “studentized” residuals in  $CV(\lambda)$  and  $GCV(\lambda)$  have to be multiplied by the corresponding weights  $w_i$  of the weight matrix  $W = \text{diag}(w_1, \dots, w_n)$ . For example,  $GCV(\lambda)$  is modified to

$$GCV(\lambda) = \frac{1}{n} \sum_{i=1}^n w_i \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1 - \text{tr}(S_\lambda)/n} \right\}^2,$$

where  $S_\lambda$  is the smoother matrix of the linear smoother  $\hat{f}_\lambda = S_\lambda y$ . Generalized cross-validation is a special case of minimizing

$$\log(\hat{\sigma}^2) + \psi(S_\lambda),$$

where  $\psi(\cdot)$  is a penalty function that decreases with increasing smoothness of  $\hat{f}_\lambda$ , and  $\hat{\sigma}^2$  is the average squared residual given in (5.1.6), i.e.,  $\hat{\sigma}^2 = (1/n) \sum_i (y_i - \hat{f}_\lambda(x_i))^2$ . The choice  $\psi(S_\lambda) = -2 \log\{1 - \text{tr}(S_\lambda)/n\}$  yields the generalized cross-validation criterion, while  $\psi(S_\lambda) = 2 \text{tr}(S_\lambda/n)$  yields the AIC criterion

$$\log(\hat{\sigma}^2) + 2 \text{tr}(S_\lambda/n). \quad (5.1.29)$$

The usual form of the AIC criterion is given by  $AIC = -2\{\log(L) - p\}$ , where  $\log(L)$  is the maximal log-likelihood and  $p$  stands for the number of parameters. Under the assumption of normally distributed response  $y_i \sim N(\mu_i, \sigma^2)$ , one obtains apart from additive constants  $AIC = n\{\log(\hat{\sigma}^2) +$

$\frac{2}{n}p\}$ . In (5.1.29) the trace  $\text{tr}(S_\lambda)$  plays the role of the effective numbers of parameters used in the smoothing fit (see also Hastie & Tibshirani, 1990). Thus, replacing  $p$  by  $\text{tr}(S_\lambda)$  yields the criterion (5.1.29). If  $\psi(S_\lambda) = -\log\{1 - 2 \text{tr}(S_\lambda)/n\}$  is chosen, one obtains the criterion suggested by Rice (1984).

Since the choice of the smoothing parameter is of crucial importance to the performance of the estimator, this has been a topic of extensive research. Classical selection procedures like cross-validation, the AIC criterion, and risk estimation are considered in Härdle, Hall & Marron (1988) and Rice (1984). Other approaches are plug-in methods where the estimate  $\hat{f}$  is written as an approximative function of the unknown  $f$  and an estimate is plugged in to obtain an estimate of measures of fit. In a second step the “optimal” bandwidth is chosen by considering the measure of fit. Thus, the actual selection is based on a pilot estimate. Moreover, some assumptions that are often doubtful are necessary to obtain an optimal parameter. The approach has been strongly criticized by Loader (1995). A nice overview and an alternative asymptotically efficient selection procedure is found in Hall & Johnstone (1992). For new developments in cross-validation techniques, see Hart & Yi (1996) and Hurvich, Simonoff & Tsai (1998).

## 5.2 Smoothing for Non-Gaussian Data

In the previous section the response variable  $y$  was assumed to be approximately Gaussian and measured on a metric scale. If the response is binary or given by counts, the methods considered there are no longer appropriate, because binary or count data have an expectation-variance structure that is different from continuous, normally distributed responses. In the following, the concepts from Section 5.1 are extended to the case of more general responses, but basically the same main ideas, namely expansion in basis functions, penalization, or localization, are used. The role of generalized models is that of a background model which, for example, is fitted locally. Thus, fitting procedures that are familiar from generalized linear models are needed, but, of course, the modelling itself is smooth and no longer parametric.

### 5.2.1 Basis Function Approach

Let the function  $f(x)$  be specified by a basis function in the form

$$f(x) = \sum_j \delta_j B_j(x) \quad (5.2.1)$$

with  $B_j(x)$  denoting the basis functions, e.g., B-splines of a fixed degree. The basic modification in comparison with Section 5.1.2 is that the least-squares

fit criterion in (5.1.5), which is equivalent to the log-likelihood for normal data, is replaced by log-likelihoods for non-normal data. In the context of generalized linear models, the linear predictor  $\eta_i = \beta_0 + x'_i\beta$  is replaced by the more general predictor  $\eta_i = f(x_i) = \sum_j \delta_j B_j(x_i)$ , so that

$$E(y_i|x_i) = h(\sum_j \delta_j B_j(x_i))$$

with known response function  $h$ .

The log-likelihood criterion to be maximized is given by

$$\sum_{i=1}^n l_i(y_i; f(x_i)), \quad (5.2.2)$$

where  $l_i$  is the log-likelihood contribution of observation  $y_i$  when, for the linear predictor,  $\eta_i = f(x_i)$  is assumed.

As for metric responses, instead of (5.2.2) a regularized version with a penalty term may be considered for maximization. One considers the penalized likelihood

$$PL(\delta) = \sum_{i=1}^n l_i(y_i; f(x_i)) - \frac{\lambda}{2} J(f),$$

where  $J(f)$  is again a penalty function. Penalty functions of the type  $\sum \delta_j^2$ , which correspond to ridge regression, have been considered by Nyquist (1990) and Segerstedt (1992); functions of the type  $\sum |\delta_j|$  which correspond to soft-thresholding, have been treated by Klinger (1998) and Donoho & Johnstone (1995). Within the P-spline framework (Eilers & Marx, 1996) the penalty is given by  $J_d = \sum_j (\Delta^d \delta_j)^2$ . The matrix representation of the difference operator uses matrices  $D_d = D_{d-1}D_1$  with  $D_1$  a banded matrix that contains only the contrasts  $-1, 1$  (compare Section 5.1.1). The resulting penalized likelihood has the form

$$PL(\delta) = \sum_{i=1}^n l_i(y_i; f(x_i)) - \frac{\lambda}{2} \delta' K_d \delta,$$

where  $K_d = D'_d D_d$ . As outlined in the following, maximization may be performed by Fisher scoring. Penalization for smoothing splines that is of the type  $J(f) = \int (f''(x))^2 dx$  is considered in Section 5.2.2.

### Fisher Scoring for Penalized Likelihood\*

Derivatives of the penalized likelihood  $PL(\delta)$  yield

$$\frac{\partial PL(\delta)}{\partial \delta} = s(\delta) - \lambda K_d \delta, \quad E \left( -\frac{\partial PL(\delta)}{\partial \delta \partial \delta'} \right) = B' W B + \lambda K_d,$$

where  $s(\delta)$  is the vector of first derivatives of the log-likelihood given in the form (as in Section 2.2) by

$$s(\delta) = B'D(\delta)\Sigma^{-1}(y - \mu(\delta))$$

with

$$y' = (y_1, \dots, y_n), \quad \mu(\delta)' = (h(\eta_1), \dots, h(\eta_n)), \quad \eta_i = \Sigma_\gamma \delta_\gamma B_\gamma(x_i),$$

and matrices

$$\begin{aligned} D(\delta) &= \text{diag}(\partial h(\eta_1)/\partial \delta, \dots, \partial h(\eta_n)/\partial \delta), \\ \Sigma(\delta) &= \text{diag}(\sigma_1^2(\delta), \dots, \sigma_n^2(\delta)), \\ B &= (b_{ij}), \quad b_{ij} = B_j(x_i). \end{aligned}$$

The weight matrix  $W$  is given in the usual form

$$W = \text{diag}(w_1, \dots, w_n) \quad \text{with } w_i = (\partial h(\eta_i)/\partial \delta)^2 / \sigma_i^2.$$

The Fisher scoring step to go from the current estimate  $\delta^{(k)}$  to  $\delta^{(k+1)}$ ,  $k = 0, 1, 2, \dots$ , is given by

$$\delta^{(k+1)} = \delta^{(k)} + (B'W^{(k)}B + \lambda K_d)^{-1}(s(\delta^{(k)}) - \lambda K_d \delta^{(k)}).$$

A simple derivation shows that iterations may be expressed in the iteratively weighted least-squares form

$$\delta^{(k+1)} = (B'W^{(k)}B + \lambda K_d)^{-1}B'W^{(k)}\tilde{y}^{(k)}$$

with “working observations”

$$\tilde{y}^{(k)} = B\delta^{(k)} + D(\delta^{(k)})^{-1}(y - \mu(\delta^{(k)})).$$

### 5.2.2 Penalization and Spline Smoothing

Throughout this section we consider only models for univariate responses, but extensions to the multivariate case are possible. Moreover, we focus on the special but important case of cubic spline smoothing. More general forms can be dealt with along similar lines by penalized maximum likelihood estimation as in Green & Yandell (1985) and Green (1987).

The basic idea is to consider the penalized log-likelihood criterion

$$\sum_{i=1}^n l_i(y_i; f(x_i)) - \frac{1}{2}\lambda \int (f''(u))^2 du \rightarrow \max, \quad (5.2.3)$$

where  $l_i$  is the log-likelihood contribution of observation  $y_i$  and  $f$  has continuous first and second derivatives  $f'$  and  $f''$ , where  $f''$  is quadratically integrable. Compared to the penalized least-squares criterion (5.1.7),

the squared Euclidean distance  $(y_i - f(x_i))^2$  is replaced by a Kullback-Leibler distance. For normally distributed observations  $y_i$ , the penalized log-likelihood criterion (5.2.3) reduces to the penalized least-squares criterion (5.1.7). As in Section 5.1.2, the penalized log-likelihood criterion explicitly formulates the compromise between faith to the data measured by the first term and roughness of the function expressed by the penalty term, with the smoothing parameter  $\lambda$  controlling this compromise. The solution is again a cubic spline. Parameterizing by the evaluations  $f_i = f(x_i)$  of the cubic splines at the observed points  $x_1 < \dots < x_n$ , one reduces (5.2.3) to

$$PL(f) = \sum_{i=1}^n l_i(y_i; f_i) - \frac{1}{2} \lambda f' K f \rightarrow \max_f, \quad (5.2.4)$$

with  $f' = (f_1, \dots, f_n) = (f(x_1), \dots, f(x_n))$  and  $K$  defined as in (5.1.9).

Except for normal likelihoods, this is no longer a quadratic optimization problem with an explicit solution  $\hat{f}_\lambda$  as in (5.1.10). The solution can be obtained by Fisher-scoring iterations. In each step, the next iterate,  $f^{(k+1)}$  say, is computed as a (weighted) spline smoother of the type (5.1.12) applied to a “working observation” vector  $\tilde{y}^{(k)}$ , computed from the current iterate  $f^{(k)}$ . This is in complete analogy to computing maximum likelihood estimates in generalized linear models by iteratively weighted least-squares; compare with Section 2.2.1. Details of the Fisher scoring algorithm are described in the following.

### Fisher Scoring for Generalized Spline Smoothing\*

The first and expected negative second derivatives of  $PL(f)$  are

$$\frac{\partial PL(f)}{\partial f} = s - \lambda K f, \quad E\left(-\frac{\partial PL(f)}{\partial f \partial f'}\right) = W + \lambda K, \quad (5.2.5)$$

where  $s$  is the vector of first derivatives of the log-likelihood term and  $W$  is the diagonal matrix of expected negative second derivatives.

The score function  $s$  and the diagonal weight matrix  $W$  have the same structure as in common parametric GLMs (compare with Chapter 2):

$$s' = (s_1, \dots, s_n), \quad s_i = \frac{D_i}{\sigma_i^2} (y_i - h(f_i)), \quad (5.2.6)$$

with  $D_i = \partial h / \partial f_i$  as the first derivative of the response function and  $\sigma_i^2$  the variance evaluated at  $\mu_i = h(f_i)$ ,

$$W = \text{diag}(w_i), \quad w_i = D_i^2 / \sigma_i^2. \quad (5.2.7)$$

The Fisher scoring step to go from the current iterate  $f^{(k)}$  to  $f^{(k+1)}$ ,  $k = 0, 1, 2, \dots$ , is thus given by

$$(W^{(k)} + \lambda K)(f^{(k+1)} - f^{(k)}) = s^{(k)} - \lambda K f^{(k)}. \quad (5.2.8)$$

In the same way as for GLMs with linear predictors, the iteration (5.2.8) can be rewritten in iteratively weighted least-squares form:

$$f^{(k+1)} = (W^{(k)} + \lambda K)^{-1} W^{(k)} \tilde{y}^{(k)} = S^{(k)} \tilde{y}^{(k)}, \quad (5.2.9)$$

with the “working” observation vector

$$\begin{aligned} \tilde{y}^{(k)} &= (W^{(k)})^{-1} s^{(k)} + f^{(k)} = (\tilde{y}_1^{(k)}, \dots, \tilde{y}_n^{(k)})', \\ \tilde{y}_i^{(k)} &= \frac{y_i - h(f_i^{(k)})}{D_i^{(k)}} + f_i^{(k)}, \end{aligned}$$

and the smoother matrix

$$S^{(k)} = (W^{(k)} + \lambda K)^{-1} W^{(k)}.$$

Comparing with cubic spline smoothing in Section 5.1, it is seen that (5.2.9) is the weighted cubic spline smoother (5.1.12) applied to the working observation vector  $\tilde{y}^{(k)}$ . Thus, in each Fisher scoring step computationally efficient weighted spline smoothing algorithms can be applied to  $\tilde{y}^{(k)}$  to obtain the next iterate  $f^{(k+1)}$  in  $O(n)$  operations. Iterations are stopped according to a termination criterion, e.g., if

$$\|f^{(k+1)} - f^{(k)}\| / \|f^{(k)}\| < \varepsilon$$

is fulfilled.

### Choice of Smoothing Parameter

Cross-validation or generalized cross-validation may be adapted to the present situation for the data-driven choice of the smoothing parameter  $\lambda$ . O’Sullivan, Yandell & Raynor (1986) replace the usual squared residuals by squared Pearson residuals and define the generalized cross-validation (*GCV*) criterion

$$GCV_P(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{(y_i - \hat{\mu}_i)/\hat{\sigma}_i}{1 - \text{tr}(S_\lambda)/n} \right\}^2,$$

where  $\hat{\mu}_i = h(\hat{f}_i)$  and the variances  $\hat{\sigma}_i^2$  are evaluated at the last iterate of the scoring algorithm. Another possibility would be to replace Pearson residuals by deviance residuals; compare to Hastie & Tibshirani (1990, Ch. 6.9). Other approaches for estimating  $\lambda$  would be possible, e.g., an empirical Bayes approach in combination with an EM-type algorithm as in Chapter 8.

### Example 5.2: Rainfall data

As an illustration, let us consider smoothing the rainfall data already presented in Chapter 1, Example 1.7. The data, given by the number of occurrences of rainfall in Tokyo for each day during the years 1983–1984, are reproduced in Figures 5.3 and 5.4 as relative frequencies with values 0, 0.5, and 1. To obtain a smooth curve for the (marginal) probability  $\pi_t$  of rainfall at calendar day  $t$ , a logistic cubic spline model

$$\pi(t) = \frac{\exp(f(t))}{1 + \exp(f(t))}$$

with  $x(t) = t$  as “covariate” was fitted. The smoothing parameter was estimated by generalized cross-validation, resulting in two local minima at  $\lambda = 32$  and  $\lambda = 4064$ ; compare with Figure 5.5. The smooth curve in Figure 5.3 ( $\lambda = 4064$ ) shows a clear seasonal pattern, while in Figure 5.4 ( $\lambda = 32$ ) the curve is rougher but closer to the data. The pattern in Figure 5.3, with peaks for wet seasons, nicely reflects the climate in Tokyo. It would be fairly difficult to see this by only looking at the raw data. A further analysis of these data is in Chapter 8 with closely related categorical state-space methods.  $\square$

### 5.2.3 Localizing Generalized Linear Models

The basic idea is to fit a generalized linear model locally. The choice of the GLM to be fitted depends primarily on the type of the response variable. While a logit model is appropriate for binary data, a cumulative model makes a good choice for ordinal categorical data. For a (multivariate) generalized linear model  $\mu_i = h(Z_i\beta)$ , the log-likelihood contribution is given by

$$l_i = (y'_i\theta_i - b(\theta_i))/\phi,$$

where the natural parameter  $\theta_i = \theta(\mu_i)$  is a function of the expectation  $\mu_i$ . *Local likelihood estimation* at target value  $x$  means maximization of the local likelihood

$$l_x(\beta) = \sum_{i=1}^n [(y'_i\theta_i - b(\theta_i))/\phi] w_\lambda(x, x_i), \quad (5.2.10)$$

where  $w_\lambda(x, x_i)$  is again a weight function depending on a smoothing parameter  $\lambda$  (e.g., kernel weights of the type  $K((x - x_i)/\lambda)$ ) and  $\theta_i = \theta(\mu_i) = \theta(Z_i\beta)$  is the natural parameter of the underlying model  $\mu_i = h(Z_i\beta)$ . With  $\hat{\beta}$  the resulting estimate, one obtains  $\hat{\mu}(x) = h(Z(x)\hat{\beta})$ , where the design matrix is built from the target value  $x$ .

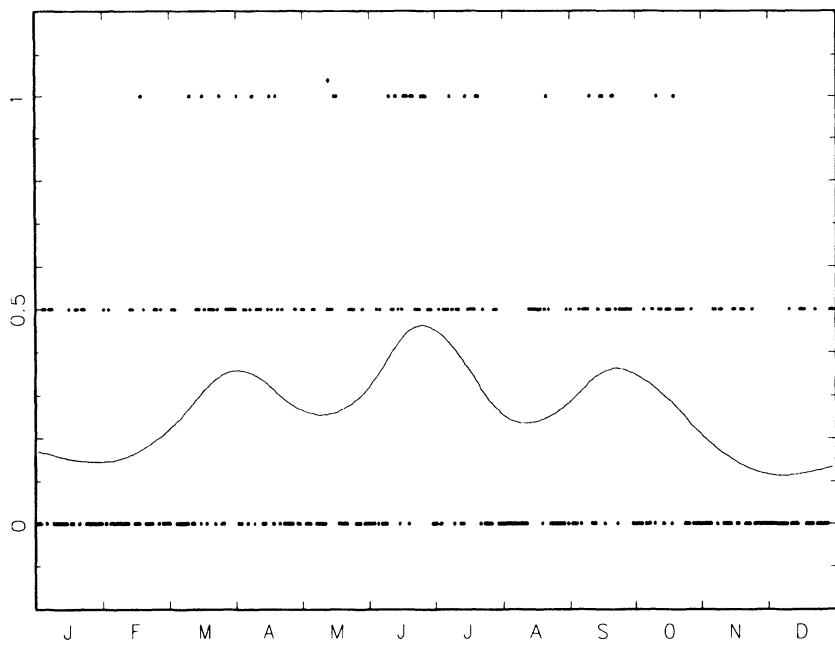


Figure 5.3. Smoothed probability of rainfall  $\lambda = 4064$ .

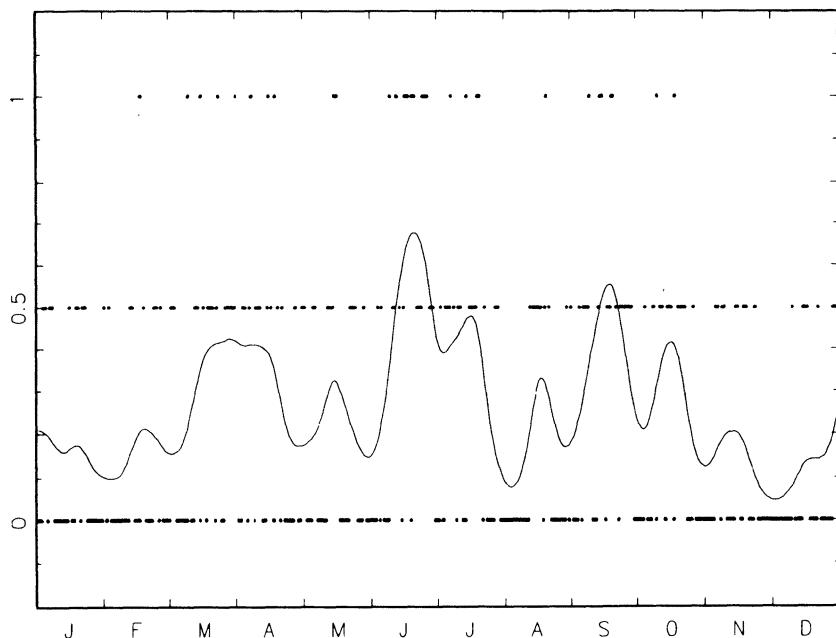
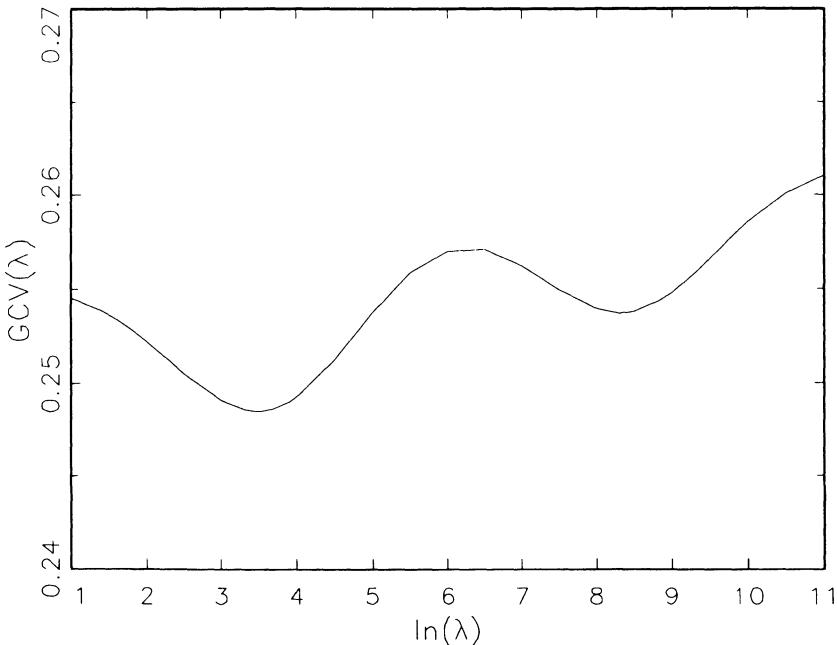


Figure 5.4. Smoothed probability of rainfall  $\lambda = 32$ .



**Figure 5.5.** Generalized cross-validation criterion, with logarithmic scale for  $\lambda$ .

As for normally distributed responses *local polynomial regression* for univariate responses may be motivated by Taylor expansion. Let  $l_i(y_i, \eta_i)$  denote the log-likelihood where  $\mu_i = h(\eta_i)$  is assumed. Then local polynomial regression at the target value  $x$  means that the unknown function  $\mu(x) = h(\eta(x))$  is approximated by

$$\begin{aligned} \eta_i = \eta(x_i) &\approx \eta(x) + \eta'(x)(x_i - x) + (\eta''(x)/2)(x_i - x)^2 + \dots \\ &= \beta_0 + \beta_1(x_i - x) + \beta_2(x_i - x)^2 + \dots \end{aligned}$$

The local model to be fitted is based on the design vector  $z(x_i)' = (1, (x_i - x), (x_i - x)^2, \dots)$ , i.e.,  $\eta_i = z(x_i)' \hat{\beta}$ . For the case of normally distributed observations  $y_i$  and design vector  $z_i$ , the local likelihood (5.2.10) is equivalent to (5.1.13), the form used in local polynomial regression for continuous response. In local polynomial fitting of generalized linear models the choice of the link function is not crucial because the model is fitted only locally in order to get a smooth, nonparametric response function. Thus, for computational convenience the canonical link function may be used. As in polynomial regression, the estimate of  $\mu(x)$  itself is determined by computing the polynomial at target value  $x$ , which corresponds to using the design vector  $z(x)' = (1, 0, \dots, 0)$ . One obtains  $\hat{\mu}(x) = h(z(x)' \hat{\beta}) = h(\hat{\beta}_0)$ , where  $\hat{\beta}$  is the parameter vector estimated locally at target value  $x$ .

As an example let us consider the case of a binary response. For a binary response variable  $y_i \in \{0, 1\}$ , the simplest model that may be fitted is a polynomial of degree zero, i.e.,  $\pi = \beta_0$ , yielding the categorical analog of the Nadaraya-Watson estimate

$$\hat{\pi}(x) = \sum_{i=1}^n w(x, x_i) y_i$$

with  $w(x, x_i) = K((x - x_i)/\lambda) / \sum_i K((x - x_i)/\lambda)$ . As for continuous responses, the local fit of a polynomial with degree  $m \geq 1$  reduces the bias. In order to avoid problems with the restriction  $0 \leq \pi \leq 1$ , it is preferable not to fit a polynomial for  $\pi$  itself but a polynomial for the transformed probability. A natural candidate is the logit model  $\log(\pi/(1-\pi)) = z'\beta$ , which as a model with canonical link is computationally simpler than models based on other link functions.

The localizing approach allows us also to make use of the ordinal information in the response variable by fitting an ordinal model like the cumulative or the sequential model considered in Section 3.3. Then smoothing across response categories is done by the fitted model. Alternatively, explicit smoothing across the categorical response categories may be based on ordinal categorical kernels (see, e.g., Aitken, 1983; Titterington & Bowman, 1985; or in combination with explanatory variables, Tutz, 1991a and b, and 1995b). However, local likelihood with ordinal models is a more elegant and computationally simpler approach.

The concept of local likelihood estimation has been formulated by Tibshirani & Hastie (1987); see also Hastie & Tibshirani (1990). A detailed analysis of local polynomial fitting for univariate GLMs is given in Fan & Gijbels (1996, Ch. 5); some asymptotics for the multivariate case are given in Kauermann & Tutz (2000a). Ahead we show that computation is easily done within the usual framework of generalized linear models.

## Local Fitting by Weighted Scoring

Derivation of the local likelihood (5.2.10) yields the local score function

$$s_x(\beta) = \sum_{i=1}^n w_\lambda(x, x_i) Z'_i D_i(\beta) \Sigma_i^{-1}(\beta) (y_i - \mu_i(\beta)),$$

which, apart from the inclusion of weights, has the usual form. The local estimation equation  $s_\lambda(\beta) = 0$  may be solved by iterative Fisher scoring of the form

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + F_\lambda(\beta^{(k)})^{-1} s_x(\beta^{(k)}),$$

where

$$F_\lambda(\hat{\beta}) = \sum_{i=1}^n w_\lambda(x, x_i) Z_i' D_i(\beta) \Sigma_i(\beta) D_i(\beta)' Z_i$$

is a weighted Fisher matrix that differs from the usual Fisher matrix only by the weights  $w_\lambda(x, x_i)$ . Thus, the usual estimation procedures may be used; the only modification is to include weights. Under regularity conditions the Fisher matrix may also be used as an approximation for the variance in the form  $\text{cov}(\hat{\beta}) = F_\lambda(\hat{\beta})^{-1}$  (e.g., Kauermann & Tutz, 2000b).

### **Example 5.3: Short-term unemployment**

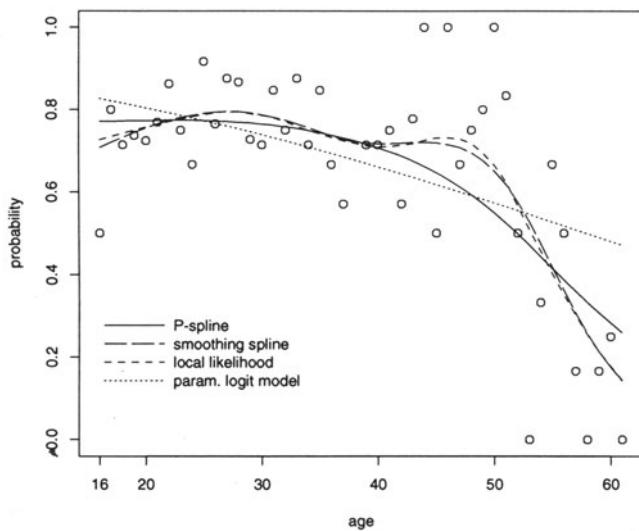
The objective is to investigate the effect of age on the probability of short-term unemployment, meaning unemployment of up to six months. The unemployed persons considered are men with a German passport. Figure 5.6 shows the data and several estimates. The data given as dots show the relative frequency of short-term unemployment for fixed age. Points have to be taken more seriously if they are based on many observations. Thus, in Figure 5.7 the number of underlying data is given. The number of observations is rather high around 25 years of age but rather low for higher ages and some intermediate values. The estimates considered are P-splines (40 knots, second difference penalty, cubic spline,  $\lambda = 1015$ ), smoothing splines, local likelihood with neighborhood smoothing (loess), and as a reference the parametric linear logit model. The estimates optimized with respect to the AIC criterion are given in Figure 5.6. It is seen that the parametric model yields a rather flat function that particularly for higher ages is far from the data. The P-spline estimate yields a strictly concave curve, whereas smoothing splines and local likelihood show more variation. This stronger variation is mainly the effect of the chosen smoothing parameter. If the smoothing parameter for smoothing splines and local likelihood estimates are chosen to match the smoothness of the P-spline approach, one obtains very similar estimates (see Fig. 5.8).  $\square$

### **Example 5.4: Rainfall data** (Example 5.2, continued)

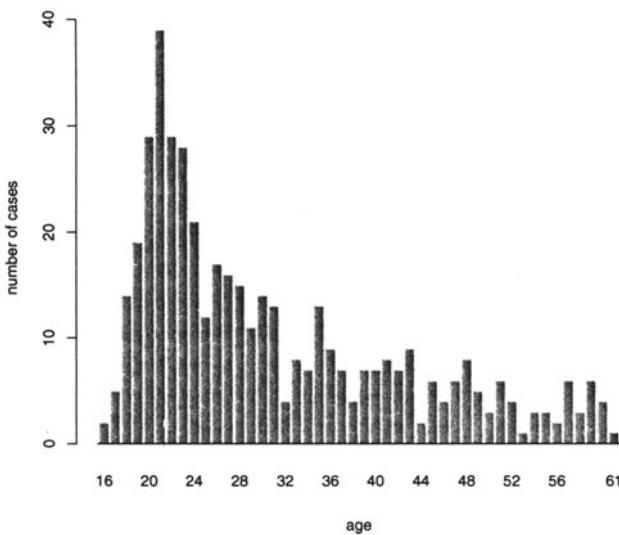
In Figures 5.3 and 5.4 (p. 199) the rainfall in Tokyo for each day during the years 1983–1984 has been smoothed by the use of a logistic cubic spline model. In Figure 5.9 a logit model has been fitted locally based on a normal kernel with smoothing parameter  $\lambda = 17$  that results from cross-validation, as given in Figure 5.10. The resulting curve is quite similar to the one resulting from cubic splines (Figure 5.3, p. 199).  $\square$

## **5.3 Modelling with Multiple Covariates**

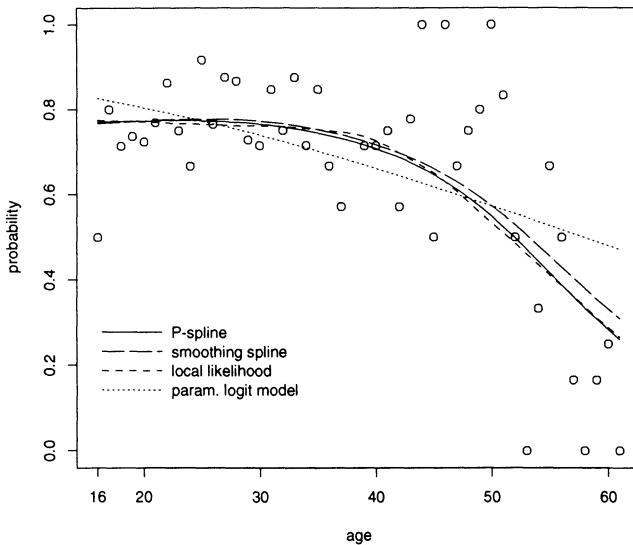
One possibility to extend nonparametric regression techniques to the case of more than one covariate is to assume that the predictor is a smooth function



**Figure 5.6.** Probability of short-term unemployment estimated by the parametric logit model, P-splines, smoothing splines, and local likelihood with the smooth estimates determined by the AIC criterion.



**Figure 5.7.** Number of observations of unemployment data for given age groups.



**Figure 5.8.** Probability of short term unemployment with same degree of smoothing for smooth estimates.

$\eta = \eta(x_1, x_2, \dots)$  of the covariates and try to estimate it by some surface smoother. For example, within the localizing approach it is straightforward to allow for vector-valued explanatory variables by using a corresponding weight function. With  $x'_i = (x_{i1}, \dots, x_{ip})$  and  $x' = (x_{01}, \dots, x_{0p})$  denoting the observation and target value, respectively, one may use a product kernel

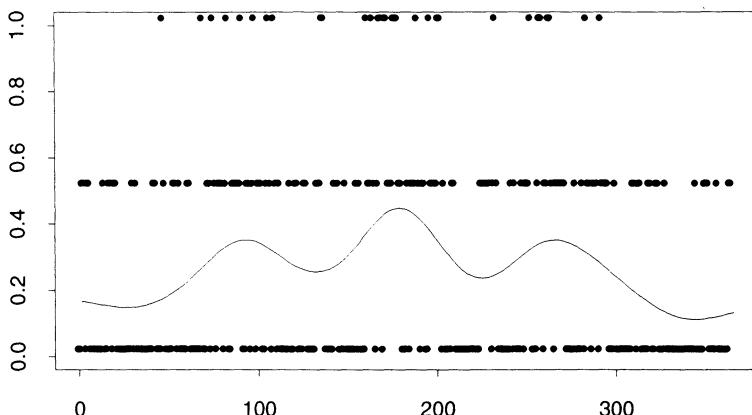
$$w(x, x_i) = \prod_{j=1}^p K\left(\frac{x_{0j} - x_{ij}}{\lambda_j}\right)$$

or a distance based kernel

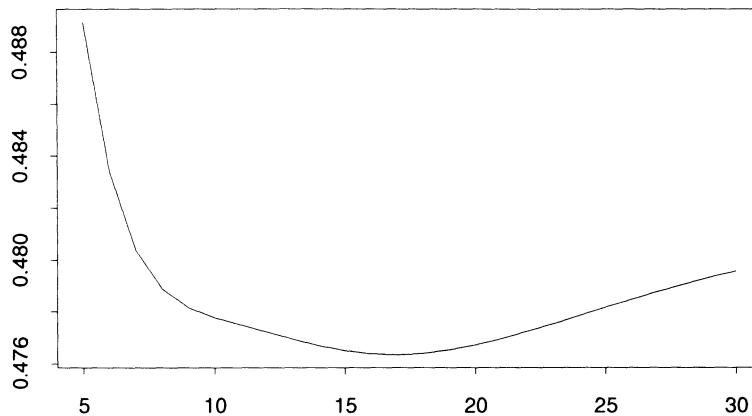
$$w(x, x_i) = K\left(\frac{d(x, x_i)}{\lambda}\right),$$

where  $d(x, x_i)$  is some  $p$ -dimensional distance function. The product kernel allows us to use smoothing parameters  $\lambda_j$ , which may vary across the components of  $x$ . The following example uses the product kernel with  $\lambda_j = \lambda$ .

**Example 5.5: Vaso constriction data** (Examples 4.2, 4.4, continued)  
The objective in this data set is to investigate the effect of rate and volume of air inspired on a transient vaso constriction in the skin of the digits. In Figure 5.11 the local constant fit of a logit model with weights from a

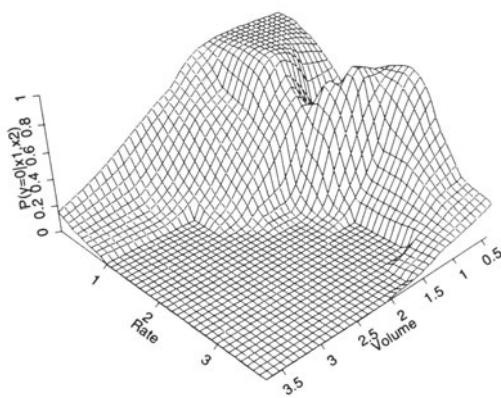


**Figure 5.9.** Local fit of logit model for Tokyo rainfall data.

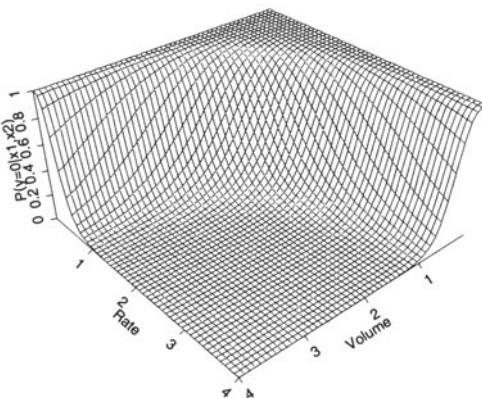


**Figure 5.10.** Cross-validation for local fit for Tokyo rainfall data.

normal product kernel ( $\lambda = 0.24$ ) is given. The plot suggests that a logistic model may be appropriate. For comparison, in Figure 5.12 the surface is plotted for a logit model based on log rate and log volume. The estimated coefficients based on Pregibon's (1982) resistant fit are 7.62 for log rate and 8.59 for log volume. Observations 4 and 18, which are not well fitted by the model (see Examples 4.2 and 4.4, and in particular, Figure 4.6), show in the smoothed fit. Both observations have volume about 0.8 and rate about 1.5. In Figure 5.11 they appear as a gap near the point (0.8, 1.5), making the surface unsteady. Comparison with Figure 5.12 shows that parametric and nonparametric estimates are roughly comparable. From the nonparametric fit, however, it is seen where the data produce deviations from the parametric model.  $\square$



**Figure 5.11.** Response surface for the nonoccurrence of vaso constriction based on the local fit of a logit model.



**Figure 5.12.** Response surface for the nonoccurrence of vaso constriction based on a fitted logit model.

Extensions of cubic spline smoothing to two or more dimensions have been considered by O'Sullivan, Yandell & Raynor (1986) and Gu (1990). Such approaches are computationally rather demanding and, more seriously, their performance may suffer from sparse data in dimensions higher than

two or three: the curse of dimensionality. The same holds for the localizing techniques as considered in Section 5.2.3.

In the following, first an overview of alternative modelling approaches is given, and then some estimation procedures are considered.

### 5.3.1 Modelling Approaches

The models to be considered in this section are either fully nonparametric or semiparametric in the sense that they contain parametric as well as nonparametric components. The basic idea is to attack the curse of dimensionality by implying some structure in the predictor space. Throughout the section, models are considered within the framework of generalized linear models, and the objective is to weaken the strict assumption of a linear predictor  $\eta_i = \mathbf{x}'_i \boldsymbol{\beta}$  by replacing the linear predictor by a smooth function depending on covariates. In the first model, which is the widely used generalized additive model, this is done by assuming additive predictors.

#### Generalized Additive Models

Generalized additive models retain an important feature of GLMs: Covariate effects are additive. However, these effects are generally nonlinear, so that the predictor is assumed to have the form

$$\eta = \alpha + \sum_{j=1}^p f_{(j)}(x_j), \quad (5.3.1)$$

where  $f_{(1)}, \dots, f_{(p)}$  are unspecified unknown functions and the covariate vector  $\mathbf{x}' = (x_1, \dots, x_p)$  can be a function of original covariates, including, for example, interactions.

While the distributional assumptions remain the same as for the definition of univariate GLMs in Chapter 2, the mean of  $y$  given  $x$  is now assumed to have the form

$$\mu = E(y|x) = h(\eta), \quad \eta = \alpha + \sum_{j=1}^p f_{(j)}(x_j). \quad (5.3.2)$$

Additive models replace the problem of estimating a function of a  $p$ -dimensional covariate vector  $x$  by one of estimating  $p$  separate one-dimensional functions. Of course, by doing so, the range of multivariate surfaces that can be generated by the model is restricted. More flexibility is obtained by adding interaction terms of the form  $f_{(ij)}(x_i, x_j)$ . However, for a larger set of covariates the number of combinations grows rapidly. Moreover, some of the advantages of the model, namely simplicity and easy interpretation of effects, get lost by including interactions.

## Partially Linear Models

A special case of generalized additive models follows when one covariate function, say  $f_{(1)}$ , is modelled nonparametrically while the remaining covariates still enter as a linear combination, say  $x_2\beta_2 + \dots + x_p\beta_p$ . More generally, in this type of model the explanatory variables are split into two groups of variables that are treated in different ways. Let  $(x, z)$  denote the covariates. The linear predictor is assumed to be given by

$$\eta = f(x) + z'\beta, \quad (5.3.3)$$

where  $\beta$  represents weights on the  $z$ -vector and  $f(x)$  is a smooth function in  $x$ . If  $x$  is vector-valued, one may also postulate some more structure for the nonparametric term. Assuming an additive structure yields

$$\eta = \sum_{j=1}^q f_{(j)}(x_j) + z'\beta,$$

where  $x' = (x_1, \dots, x_q)$ . The model itself has the form  $\mu = h(\eta)$  with fixed response function  $h$ . While there is no restriction on the type of variables in  $z$ ,  $x$  should be composed from metrically scaled, preferably continuous variables. Models of this type are useful if, for a specific variable, linearity is doubtful whereas for the rest of the variables, for example, categorical ones, linearity seems adequate. Models of this type have also been called partially linear models because parts of the model have a strictly linear form.

## Varying-Coefficient Models

An approach that formally comprises partially linear and generalized additive models is basically semiparametric. As in partially linear models there are two groups of variables, the predictor  $z_1, \dots, z_p$  and the so-called effect modifiers  $x_1, \dots, x_p$ . The predictor has the form

$$\eta = \beta_0 + z_1\beta_1(x_1) + \dots + z_p\beta_p(x_p).$$

Thus, the effect modifiers  $x_1, \dots, x_p$  modify the effect of the predictors  $z_1, \dots, z_p$ . Although the model is given as a linear combination, the functions  $\beta_i(u_i)$  are not parametrically specified and have to be estimated nonparametrically, e.g., by smoothing techniques. The variation of the parameters across the effect modifiers may be seen as a nonparametric form of interaction of between  $z_i$  and  $x_i$  with the advantage that the variation of the parameter  $\beta(x_i)$  has simple interpretation and is easily seen from plotting  $\beta(x_i)$ .

Although estimation procedures may differ, formally the model comprises many special cases. For constant  $x$ , i.e.,  $x_i \equiv 1$ ,  $i = 1, \dots, p$ , one obtains the usual parametric model

$$\eta = \beta_0 + z_1\beta_1 + \cdots + z_p\beta_p,$$

while for  $z_1 \equiv 1, x_2 = \cdots = x_p = 1$ , one has the partial linear model

$$\eta = \beta_0 + \beta_1(u) + z_2\beta_2 + \cdots + z_p\beta_p,$$

and for  $z_1 \equiv \cdots \equiv z_p \equiv 1$  one obtains the generalized additive model with predictor

$$\eta = \beta_0 + \beta_1(x_1) + \cdots + \beta_p(x_p).$$

An interesting special case is that where the effect modifiers are the same,  $x_1 = \cdots = x_p$ . If the modifier corresponds to time, one obtains with  $t = x_1 = \cdots = x_p$  the dynamic model

$$\eta = \beta_0 + z_1\beta(t) + \cdots + z_p\beta(t)$$

with coefficients varying across time. Models of this type are also called state space models and are considered extensively in Chapter 8.

The concept of varying coefficients has been introduced by Hastie & Tibshirani (1993) with estimation based on penalized likelihood functions. Kauermann & Tutz (2000a) and Tutz & Kauermann (1997) consider local likelihood estimation that allows the derivation of asymptotic results. Alternative estimates within the P-spline framework have been given by Marx & Eilers (1998). For Bayesian approaches, see Section 5.4.

## Projection Pursuit Regression

Projection pursuit (Friedman & Stützle, 1981) handles the problem of dimensionality by assuming a predictor of the form

$$\eta = \sum_{j=1}^m v_j(x'\beta_j),$$

where  $x'\beta_j$  is a one-dimensional projection of the vector  $x$  and  $v_j$  denotes an arbitrary univariate function that is unknown. The case  $m = 1$  where one considers only one projection is also called a *single-index model* (see also Horowitz, 1998). A generalization that is also a partially linear model of type (5.3.3) is the model with predictor

$$\eta = v(x'\alpha) + z'\beta, \quad (5.3.4)$$

where again two sorts of variables are considered. The first,  $x$ , is projected and underlies an unknown transformation  $v$ ; the second,  $z$ , is specified in the usual linear form. With  $x$  being a one-dimensional continuous variable, one has a partially linear model. The class of models specified by (5.3.4) has been called *generalized partially single-index models* and has been investigated by Carroll, Fan, Gijbels & Wand (1997).

## Basis Function Approach

The underlying idea is to expand the predictor in a linear form with basis functions in the form

$$\eta(x) = \sum_{j=1}^m \delta_j B_j(x),$$

where  $B_j(x)$  are basis functions such as B-splines or polynomials. The approach obtains its flexibility from the number of possible choices of number and type of basis functions. If it seems adequate to the problem, also trigonometric basis functions or jump functions may be used. Moreover, it is easy, for example, to impose an additive structure by using the linear combination

$$\eta(x) = \sum_{j=1}^p \sum_{s=1}^m \delta_{js} B_{js}(x_j),$$

where  $x' = (x_1, \dots, x_p)$  and the basis functions  $B_{js}(x_j)$  contain only the  $j$ th component of  $x$ .

A specific approach of this type are multivariate adaptive regression splines (MARS), which were introduced by Friedman (1991). MARS with linear splines use basis functions of the form

$$B_j(x) = (x_1 - t_{1j})_+ (x_2 - t_{2j})_+ \dots (x_r - t_{rj})_+,$$

with  $t_{js}$  denoting knots in the range of the variable  $x_j$ .

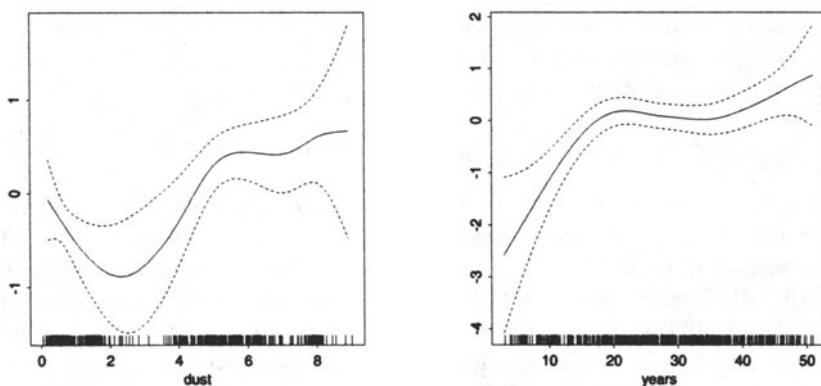
A characteristic of the MARS approach is the successive addition of the basis functions. In the MARS program (Friedman, 1991) initially the constant model  $f(x) = \delta_1$  is fitted. Then, models with a fixed number of basis functions are successively extended by considering the addition of new basis functions  $B_m(x)(x_i - t)_+$  and  $B_m(x)(t - x_i)_+$ , where  $x_i$  is one of the components of  $x$ ,  $t$  is a new knot in the range of  $x_i$ , and  $B_m(x)$  is a basis function already in the model. For given set of basis functions the unknown parameters may be estimated by ML estimation.

The successive addition of basis functions makes it similar to classification and regression trees (CART), which are treated extensively in Breiman, Friedman, Olshen & Stone (1984). CART uses (0–1) jump functions, i.e., the knots are interpretable as splits. The concept of growing a tree and subsequent pruning, which is the usual procedure for CARTs, may also be used in the MARS approach. Stone, Hansen, Kooperberg & Truong (1997) consider a modification of MARS called POLYMARS, which allows for a multidimensional response variable and defines an allowable space by listing its basis functions. Stone, Hansen, Kooperberg & Truong (1997) give a more general treatment of polynomial splines and tensor products within an approach called extended linear modelling.

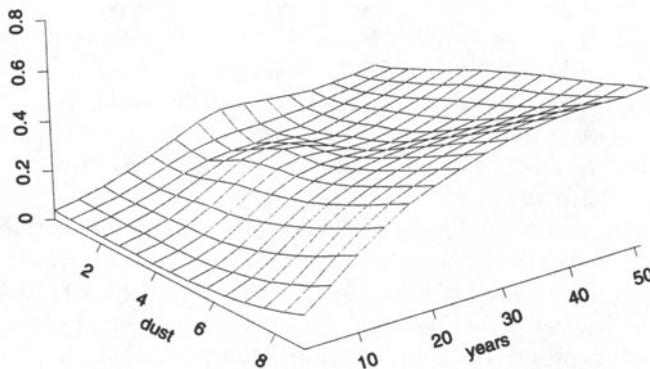
### Example 5.6: Chronic bronchitis

An occupational epidemiology multicenter study focused on the dependence of the development of chronic bronchitis on the average concentration of dust at the working place and the time of exposure. A detailed description is given in Küchenhoff & Ulm (1997). The analysis is based on 920 workers, all of whom were smokers. A generalized additive logit model based on spline smoothing has been fitted with  $y = 1$  as the occurrence of bronchitis. Figure 5.13 shows the additive components together with error bands provided by S-PLUS. The components signal a generally increasing probability of bronchitis with increasing duration of the exposure. The effect of dust concentration is strictly increasing above  $2.5 \text{ mg/m}^3$  dust. The increase only for higher concentrations may suggest that there is a threshold-limiting value of dust concentration, where damage in the form of increased probability of bronchitis occurs only if the concentration is above this threshold. Küchenhoff & Ulm (1997) give various approaches to the modelling of threshold-limiting values.

Figure 5.14 shows the local linear two-dimensional fit of  $f(\text{dust}, \text{years})$  based on nearest neighborhood with the smoothing parameter chosen such that it uses about 50% of the data. It is seen that the gap of observations between  $1.5$  and  $4 \text{ mg/m}^3$  is smoothed by the nearest neighborhood fit, with the dip around  $2.5 \text{ mg/m}^3$  being much less prominent. There is no strong indication for interaction between dust and duration, which is also supported by the fact that an interaction term in the simple linear logit model is not significant (compare with Simonoff & Tutz, 2000).  $\square$



**Figure 5.13.** Effects of concentration of dust and exposure time on the probability of chronic bronchitis for the generalized additive logit model.



**Figure 5.14.** Local linear fit of  $f(\text{dust}, \text{years})$  with nearest neighborhood to probability of chronic bronchitis.

### Example 5.7: Women's role in society

The data set is part of two general social surveys carried out by the National Opinion Research Center, University of Chicago, and analyzed before by Haberman (1978) and Collett (1991). Respondents have been asked if he or she agree ( $y = 1$ ) or disagree ( $y = 0$ ) with the statement, "Women should take care of running their homes and leave running the country up to men." Analysis may be based on the dichotomous logit model with the covariates gender and years of education. A simple parametric model is given by

$$\text{logit}(Y = 1|i, \text{years}) = \beta_{0i} + \text{years} \beta_{Y,i}, \quad i = 1, 2, \quad (5.3.5)$$

where  $i = 1$  stands for women and  $i = 2$  stands for men. Thus, differing intercepts and slopes are allowed for men and women. Constraining the slopes to be equal yields an increase in deviance which is highly significant. On the other hand, as Collett (1991) states, the resulting deviance of model (5.3.5) is 57.103 on 37 d.f. and thus "is uncomfortably high." In particular, for the female subpopulation quadratic terms seem to be necessary. But then comparison of slopes no longer makes sense. Thus, alternatively the varying-coefficient model

$$\text{logit}(Y = 1|x_G, \text{years}) = \beta_0(\text{years}) + x_G \beta_G(\text{years}) \quad (5.3.6)$$

may be fitted with  $x_G$  as a dummy variable for gender (1: female, 0: male), and the global intercept as well as  $\beta_G$  may vary across years of education.

Figure 5.15 shows the local likelihood estimates of  $\beta_0$  and  $\beta_G$  based on the normal kernel with smoothing parameter  $\gamma = 2.0$  together with approximated pointwise 0.95 confidence intervals. It is seen that the effect of years of education apart from the gender effect is rather stable below 6 years and decreases above 6 years. The linear effect of gender is modified by years with a high positive effect at 2 to 3 years of education and very low values at 17 to 18 years of education. One may compare the results with the parametric model (5.3.5). If model (5.3.5) holds,  $\beta_0$  and  $\beta_G$  both depend linearly on years in the form

$$\begin{aligned}\beta_0(\text{years}) &= \beta_{02} + \text{years } \beta_{Y,2}, \\ \beta_G(\text{years}) &= (\beta_{01} - \beta_{02}) + \text{years}(\beta_{Y,1} - \beta_{Y,2}).\end{aligned}$$

If, in addition, slopes are equal for men and women,  $\beta_G(\text{years}) = \beta_{01} - \beta_{02}$  does not depend on years and  $\beta_G(\text{years})$  should be a horizontal line. The fits of the linear model is also shown in Fig. 5.15. The main effect of years decreases across years, but in particular for low values, it is rather constant over years. The gender specific slopes, which are assumed in (5.3.5), yield a linear effect that lies within the confidence bends. However, the simplification  $\beta_{Y,1} = \beta_{Y,2}$  seems to hold only in the medium range of years, but for low and high values of years of education, the estimates deviate from a horizontal line. In model (5.3.6) the modification of the effect of gender across years of education is not specified parametrically and may vary freely. It becomes obvious in which areas the parametric model is not appropriate.

In Figure 5.16 the fitted probabilities are shown for the smooth approach as well as for the parametric approach (5.3.5). For males the parametric models are quite close to the estimates for the varying-coefficient model. This is different for females. For least-squares estimates of this example compare with Tutz & Kauermann (1998).  $\square$

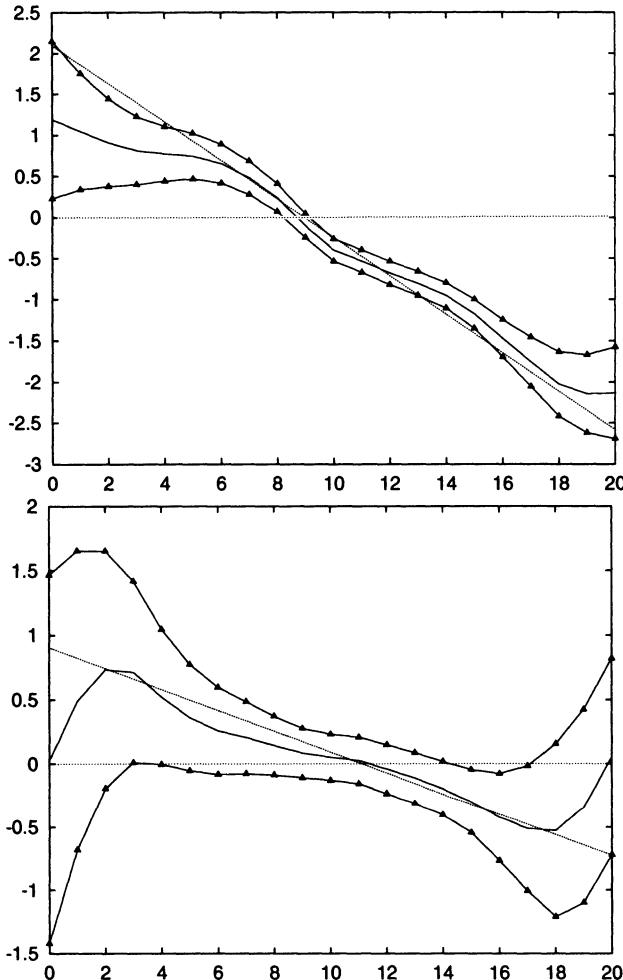
### 5.3.2 Estimation Concepts

Now we consider estimation of models with multiple covariates. Since some of the concepts have been mentioned in the previous section, we restrict the presentation to some basic concepts.

#### Backfitting Algorithm for Generalized Additive Models

A widely used algorithm for fitting additive models is the backfitting algorithm (Friedman & Stützle, 1981). For simplicity we start with the simple case of an additive model where the response is usually a metrically scaled variable and one assumes the additive structure

$$\mu_i = E(y_i|x_i) = \alpha + \sum_{j=1}^p f_{(j)}(x_{ij})$$

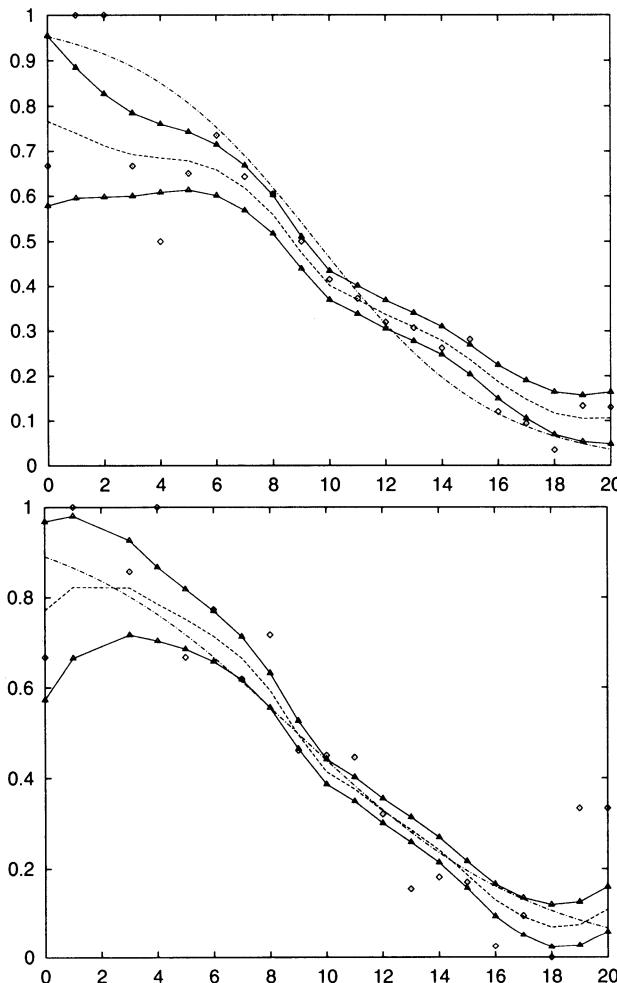


**Figure 5.15.** Local likelihood estimates of intercept (above) and slope (below) for gender varying across years of education compared with estimates for the linear model within each gender group. For the local estimates pointwise error bands are given by  $1.96 * \text{standard error}$ .

for data  $y_i, x'_i = (x_{i1}, \dots, x_{iq}), i = 1, \dots, n$ .

The basic principle is to start from initial estimates and iteratively fit the smooth components  $f_{(j)}$  by smoothing partial residuals that are computed without  $f_{(j)}$ . Since  $\alpha$  may be incorporated in any of the functions  $f_{(1)}, \dots, f_{(p)}$ , identification problems arise for the model in the given form. Therefore,  $\alpha$  is fixed as the mean response, i.e.,  $\hat{\alpha} = \bar{y}$ .

Let  $y' = (y_1, \dots, y_n)$  denote the vector of observations and  $f'_j = (f_{(j)}(x_{1j}), \dots, f_{(j)}(x_{nj}))$  denote the vector of evaluations.



**Figure 5.16.** Fitted probabilities for women (above) and men (below) based on local likelihood and parametric model with pointwise error bands  $1.96 * \text{standard error}$ .

*Backfitting algorithm:*

- (1) Initialize  $\hat{f}_j = f_j^{(0)}$ ,  $j = 1, \dots, p$ ,  $\hat{f}'_0 = (\hat{\alpha}, \dots, \hat{\alpha})$ .
- (2) For  $j = 1, \dots, p$ , compute updates  $\hat{f}_j^{(0)} \rightarrow \hat{f}_j^{(1)}$  of the form

$$\hat{f}_j^{(1)} = S_j \left( y - \hat{f}_0 - \sum_{s < j} \hat{f}_s^{(1)} - \sum_{s > j} \hat{f}_s^{(0)} \right),$$

where  $S_j$  is a smoother matrix and  $y - \hat{f}_0 - \sum_{s < j} \hat{f}_s^{(1)} - \sum_{s > j} \hat{f}_s^{(0)}$  is the partial residual resulting from previously fitted components without  $\hat{f}_j$ .

- (3) Continue (2) until the residual sum of squares  $(y - \hat{f}_0 - \sum_{j=1}^p \hat{f}_j)'(y - \hat{f}_0 - \sum_{j=1}^p \hat{f}_j)$  fails to decrease.

In step (2) a linear smoother is used for each component. Of course, the corresponding smoothing matrices  $S_1, \dots, S_p$  may use different amounts of smoothing.

In the more general case of a generalized additive model,

$$\mu_i = h(\alpha + \sum_{j=1}^p f_{(j)}(x_{ij}))$$

Fisher scoring is used locally to obtain the fitted components. Fisher scoring iterations as considered in Section 5.2.1 may be expressed as weighted least-squares fittings of working observations that have the form

$$\tilde{y}_i = \hat{\eta}_i + \frac{y_i - h(\hat{\eta}_i)}{\hat{D}_i},$$

where  $\hat{D}_i = \partial h(\hat{\eta}_i)/\partial \eta$ . The weights on observation  $i$  are determined by  $w_i = \hat{D}_i^2/\hat{\sigma}^2$ .

In the local scoring algorithm, for each Fisher scoring step there is a inner loop that fits the additive structure of the linear predictor. In this inner loop a weighted version of the backfitting algorithm from above is performed.

*Backfitting with Fisher scoring:*

- (1) *Initialization:*  $\alpha^{(0)} = g(\sum_{i=1}^n y_i/n)$ ,  $f_1^{(0)} = \dots = f_p^{(0)} = 0$ , where  $g = h^{-1}$  is the link function.  
(2) *Scoring steps* for  $k = 0, 1, 2, \dots$ : Compute the current working observations

$$\tilde{y}_i^{(k)} = \hat{\eta}_i^{(k)} + \frac{y_i - h(\hat{\eta}_i^{(k)})}{\hat{D}_i^{(k)}}$$

and the weights  $w_i^{(k)} = (\hat{D}_i^{(k)}/\hat{\sigma}_i^{(k)})^2$  with  $\hat{\eta}_i^{(k)} = \hat{\alpha}^{(k)} + \sum_{j=1}^p f_{(j)}^{(k)}(x_{ij})$ ,  $i = 1, \dots, n$ .

*Inner backfitting loop:*

- (i) Initialize  $\hat{\alpha} = \frac{1}{n} \sum_{i=1}^n \tilde{y}_i^{(k)}$ ,  $f_j^0 := f_j^{(k)}$ ,  $j = 1, \dots, p$ ,  
(ii) Compute updates  $f_j^0 \rightarrow f_j^1$ ,  $j = 1, \dots, p$ , in each backfitting iteration from the working observations  $\tilde{y}' = (\tilde{y}_1^{(k)}, \dots, \tilde{y}_n^{(k)})$  by

$$f_j^1 = S_j(\tilde{y} - \hat{\alpha} - \sum_{s < j} f_s^1 - \sum_{s > j} f_s^0)$$

by application of weighted cubic spline smoothing to the “working residual”  $\tilde{y} - \hat{\alpha} - \sum f_s^1 - \sum f_s^0$ . Set  $f_j^0 := f_j^1$ ,  $j = 1, \dots, p$ , after such iteration. Stop backfitting iterations until  $\|f_j^0 - f_j^1\|$ ,  $j = 1, \dots, p$  is nearly zero. Set  $f_j^{(k+1)} := f_j^1$  for the final iterate.

- (3) *Stop* if some termination criterion, for example,

$$\frac{\sum_{j=1}^p \|f_j^{(k+1)} - f_j^{(k)}\|}{\sum_{j=1}^p \|f_j^{(k)}\|} \leq \varepsilon,$$

is reached.

The next section explicitly shows how the backfitting algorithm may be derived from Fisher scoring. Convergence properties of the backfitting algorithm are derived in Buja, Hastie & Tibshirani (1989), Opsomer & Ruppert (1997), and Opsomer (2000). Alternative estimation procedures for additive models have been proposed e.g., by Wahba (1990), who suggests to estimate the functions simultaneously, Linton & Härdle (1996), and Marx & Eilers (1998).

### Backfitting with Spline Functions\*

For generalized additive models with predictor  $\eta_i = \sum_i f_{(j)}(x_{ij})$ , one approach in estimation is based on penalization. In order to estimate the functions  $f_{(1)}, \dots, f_{(p)}$  by cubic splines, criterion (5.2.3) is generalized to

$$\sum_{i=1}^n l_i(y_i; \eta_i) - \frac{1}{2} \sum_{j=1}^p \lambda_j \int (f_{(j)}''(u))^2 du \rightarrow \max, \quad (5.3.7)$$

with separate penalty terms and smoothing parameters for each function. The maximizing functions are again cubic splines. Parameterizing by the vectors of evaluations

$$f_j = (f_{(j)}(x_{1j}), \dots, f_{(j)}(x_{nj}))', \quad j = 1, \dots, p,$$

(5.3.7) can be rewritten as the penalized log-likelihood criterion

$$PL(f_1, \dots, f_p) = \sum_{i=1}^n l_i(y_i; \eta_i) - \frac{1}{2} \sum_{j=1}^p \lambda_j f_j' K_j f_j, \quad (5.3.8)$$

where penalty matrices  $K_j$  for each predictor  $f_j$  are defined analogously to  $K$  in (5.1.9). The predictor in vector form is given by  $\eta = f_1 + \dots + f_p$ . The constant term  $\alpha$  is omitted here and in the following derivation of the Fisher-scoring iterations. It is used in the backfitting algorithm only in order to guarantee uniqueness of the smoothers by centering the working observations; compare with Hastie & Tibshirani (1990, Sections 5.2, 5.3).

Partial derivatives of  $\partial PL/\partial f_j, j = 1, \dots, p$ , are

$$\begin{aligned}\frac{\partial PL}{\partial f_j} &= \left( \frac{D_1}{\sigma_1^2}(y_1 - \mu_1), \dots, \frac{D_n}{\sigma_n^2}(y_n - \mu_n) \right)' - \lambda_j K_j f_j \\ &= \text{diag}(D_i/\sigma_i^2)(y - \mu) - \lambda_j K_j f_j,\end{aligned}$$

with  $D_i = \partial h(\eta_i)/\partial \eta$  as the first derivative of the response function and  $\sigma_i^2$  the variance function evaluated at  $\mu_i = h(\eta_i)$ . One obtains the equations

$$s = \lambda_1 K_1 f_1, \dots, s = \lambda_p K_p f_p, \quad (5.3.9)$$

where the derivative  $s' = (s_1, \dots, s_n)$  of the log-likelihood is given by  $s_i = (D_i/\sigma_i^2)(y_i - \mu_i)$ .

The second derivative of  $PL$  is given by

$$\frac{\partial^2 PL}{\partial f_j \partial f'_j} = (y - \mu) \frac{\partial}{\partial f'_j} \text{diag}(D_i/\sigma_i^2) + \text{diag}(D_i/\sigma_i^2) \frac{\partial}{\partial f'_j} (y - \mu) - \lambda_j K_j,$$

yielding the expected value

$$E \left( -\frac{\partial^2 PL}{\partial f_j \partial f'_j} \right) = \text{diag}(D_i/\sigma_i^2) D_i + \lambda_j K_j = \text{diag}(D_i^2/\sigma_i^2) + \lambda_j K_j.$$

In complete analogy one obtains for  $s \neq j$   $E(-\partial^2 PL/\partial f_j \partial f'_j) = \text{diag}(D_i^2/\sigma_i^2)$ . In vector form one obtains for  $f' = (f'_1, \dots, f'_p)$  the Fisher scoring algorithm by

$$f^{(k+1)} = f^{(k)} + \left\{ E \left( -\frac{\partial^2 PL}{\partial f \partial f'} \right) \right\}^{-1} \frac{\partial PL(\hat{f}_1^{(k)}, \dots, \hat{f}_p^{(k)})}{\partial f}.$$

Using  $W = \text{diag}(D_i^2/\sigma_i^2)$  yields Fisher scoring iterations in the form

$$\begin{bmatrix} W^{(k)} + \lambda_1 K_1 & W^{(k)} & \dots & W^{(k)} \\ W^{(k)} & W^{(k)} + \lambda_2 K_2 & \dots & W^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ W^{(k)} & W^{(k)} & \dots & W^{(k)} + \lambda_p K_p \end{bmatrix} \begin{bmatrix} f_1^{(k+1)} - f_1^{(k)} \\ f_2^{(k+1)} - f_2^{(k)} \\ \vdots \\ f_p^{(k+1)} - f_p^{(k)} \end{bmatrix}$$

$$= \begin{bmatrix} s^{(k)} - \lambda_1 K_1 f_1^{(k)} \\ s^{(k)} - \lambda_2 K_2 f_2^{(k)} \\ \vdots \\ s^{(k)} - \lambda_p K_p f_p^{(k)} \end{bmatrix},$$

where  $W^{(k)}$  and  $s^{(k)}$  are  $W$  and  $s$  evaluated at  $\eta^{(k)} = \eta(f_1^{(k)}, \dots, f_p^{(k)})$ . If one defines the working observation vector

$$\tilde{y}^{(k)} = \eta^{(k)} + (W^{(k)})^{-1} s^{(k)}$$

and the smoother matrices

$$S_j^{(k)} = (W^{(k)} + \lambda_j K_j)^{-1} W^{(k)},$$

the iterations can be written as

$$\begin{bmatrix} I & S_1^{(k)} & \dots & S_1^{(k)} \\ S_2^{(k)} & I & \dots & S_2^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ S_p^{(k)} & S_p^{(k)} & \dots & I \end{bmatrix} \begin{bmatrix} f_1^{(k+1)} \\ f_2^{(k+1)} \\ \vdots \\ f_p^{(k+1)} \end{bmatrix} = \begin{bmatrix} S_1^{(k)} \tilde{y}^{(k)} \\ S_2^{(k)} \tilde{y}^{(k)} \\ \vdots \\ S_p^{(k)} \tilde{y}^{(k)} \end{bmatrix}. \quad (5.3.10)$$

A direct solution of the  $np$ -dimensional system (5.3.10) to obtain the next iterate  $f^{(k+1)} = (f_1^{(k+1)}, \dots, f_p^{(k+1)})$  will be computationally feasible only in special cases. Instead (5.3.10) is rewritten as

$$\begin{bmatrix} f_1^{(k+1)} \\ f_2^{(k+1)} \\ \vdots \\ f_p^{(k+1)} \end{bmatrix} = \begin{bmatrix} S_1^{(k)}(\tilde{y}^{(k)} - \sum_{j \neq 1} f_j^{(k+1)}) \\ S_2^{(k)}(\tilde{y}^{(k)} - \sum_{j \neq 2} f_j^{(k+1)}) \\ \vdots \\ S_p^{(k)}(\tilde{y}^{(k)} - \sum_{j \neq p} f_j^{(k+1)}) \end{bmatrix} \quad (5.3.11)$$

and solved iteratively by the “backfitting” or Gauss-Seidel algorithm in the inner loop, with the special choice of smoothing matrix given above.

Smoothing in generalized additive models is not restricted to cubic spline smoothing. The approach works for other symmetric linear smoothers based on smoother matrices  $S_1, \dots, S_p$  as well: One only has to define the penalty matrices  $K_1, \dots, K_p$  in the penalized log-likelihood criterion by

$$K_j = S_j^- - I,$$

where  $S_j^-$  is any generalized inverse of  $S_j$ . Weighted cubic spline smoothing in the inner loop (ii) of the backfitting algorithm is then substituted by a corresponding symmetric linear smoother, for example, a running line smoother.

### Choice of Smoothing Parameter

A data-driven choice of the smoothing parameters  $\lambda' = (\lambda_1, \dots, \lambda_p)$  by generalized cross-validation is possible in principle. For additive models the criterion is

$$GCV_P(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{(y_i - \hat{\mu}_i)/\hat{\sigma}_i}{1 - \text{tr}(R_\lambda)/n} \right\}^2,$$

where  $R_\lambda$  is the smoother matrix that generates the additive predictor  $\hat{y} = R_\lambda \tilde{y}$  in the last iteration step. However, optimization of the criterion would require efficient computation of  $\text{tr}(R_\lambda)$  in each step of a multidimensional search algorithm. Moreover, it may be very difficult to find a global minimum in the multidimensional case. It seems that additional research is necessary for automatic smoothing parameter selection; compare with Hastie & Tibshirani (1990, Section 6.9).

### Partially Linear Models

For partially linear models with predictor  $\eta_i = f(u_i) + x'_i \beta$ , there are several possibilities of estimating  $\beta$  and the smooth function  $f$ . The penalizing approach is based on the maximization of

$$\sum_{i=1}^n l_i(y_i; \eta_i) - \frac{1}{2} \lambda \int (f''(u))^2 du,$$

where the linear predictor is given by (5.3.3). Maximization may again be performed by Fisher scoring (see Green & Silverman, 1994, Ch. 5.3). For the case of normally distributed response with identity link, one may use the iterative backfitting procedure considered in the previous section for generalized additive models or a direct method where the estimate of  $\beta$  results from solving an equation that does not depend on  $\hat{f}(u_i)$  (see Green & Silverman, 1994, Ch. 4.3). Alternative estimation procedures have been considered by Speckman (1988), Hunsberger (1994), and Severini & Staniswalis (1994). The latter consider the concept of generalized profile likelihood according to which, first, the parametric component of the model is held fixed and an estimate of the smooth component (using some smoothing method) is obtained. Then a profile likelihood for the parametric component is constructed which uses the estimate of the smooth component. They obtain the useful result of a  $\sqrt{n}$  rate convergence of the parametric component.

**Example 5.8: Vaso constriction data** (Example 5.5, continued)

Figures 5.17 and 5.18 show the nonoccurrence of vaso constriction estimated with an additive logit model. The additive structure resulting from the sum of the component functions is clearly seen.

The effects of the two explanatory variables may be considered separately. The effect of rate on the occurrence of constriction is almost monotone: If rate decreases, the probability of nonoccurrence increases. The effect of volume is different: There is a decrease in probability if volume decreases between 1.6 and 0.8. For stronger smoothing (Figure 5.18) the decrease is almost a plateau. A comparison with the kernel-based smoother in Figure 5.12, p. 206, shows the difference between approaches: Data that indicate nonmonotone behavior influence the behavior of the smoothing components in the additive model, whereas they yield more local deviation concerning both components in the case of kernel smoothing.  $\square$

## 5.4 Semiparametric Bayesian Inference for Generalized Regression

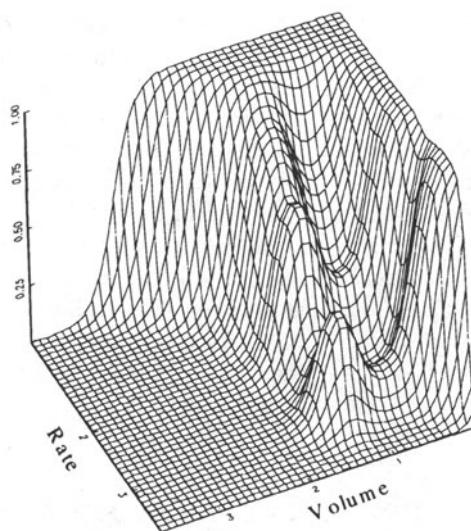
Mainly due to breakthroughs in computer-intensive inference via MCMC simulation, Bayesian approaches for non- and semiparametric regression have become useful tools for practical data analysis. Two main directions for semiparametric extensions for GLMs are, first, to keep the link function fixed and allow for more flexible forms of the predictor as in GAMs, or, vice versa, retain the linear predictor and model the link function nonparametrically. The second option includes various probabilistic specifications such as discrete mixture and Dirichlet processes; see, e.g., Escobar & West (1995), Newton, Czado & Chappell (1996), Mukhopadhyay & Gelfand (1997), and Gelfand (1998) for a general review. This section provides a brief survey of recent advances in Bayesian non- and semiparametric modelling of the predictor. We first consider the case of Gaussian models and then move on to non-Gaussian models, in particular those with discrete responses. The emphasis is on smoothness prior approaches, which can be regarded as a stochastic generalization of penalization approaches, but methods based on basis functions will also be briefly outlined.

### 5.4.1 Gaussian Responses

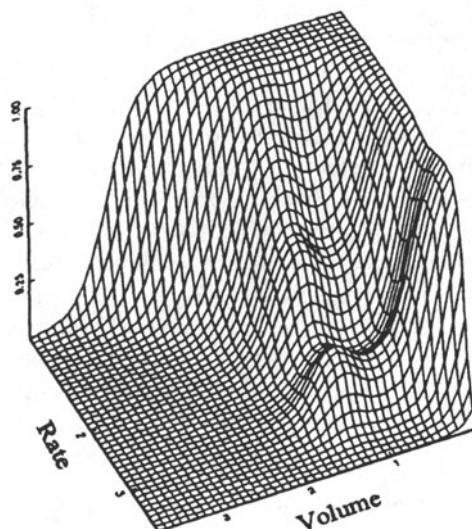
#### Smoothness Priors Approaches

The smoothness priors approach for Bayesian function estimation in the classical case of a nonparametric Gaussian regression model

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, n, \tag{5.4.1}$$



**Figure 5.17.** Nonoccurrence of vaso constriction of the skin smoothed by an additive model with  $\lambda_1 = \lambda_2 = 0.001$ .



**Figure 5.18.** Nonoccurrence of vaso constriction of the skin smoothed by an additive model with  $\lambda_1 = \lambda_2 = 0.003$ .

with observations  $(y_i, x_i)$ ,  $i = 1, \dots, n$ , and i.i.d. Gaussian errors  $\varepsilon_i \sim N(0, \sigma^2)$  can be based on the close connection between penalized least-squares and state space models; compare to Chapter 8 and the review in Fahrmeir & Knorr-Held (2000). For simplicity we first consider the case of equidistant covariate values or design points  $x_1 < \dots < x_n$ .

For nonparametric estimation using smoothness priors, the observation model (5.4.1) is supplemented by assigning an appropriate prior to the vector  $f = (f(x_1), \dots, f(x_n))'$  of function evaluations. This can be done by assuming *local smoothness priors* in the form of random walks of first or second order,

$$f(x_i) = f(x_{i-1}) + u_i \quad \text{or} \quad f(x_i) = 2f(x_{i-1}) - f(x_{i-2}) + u_i, \quad (5.4.2)$$

with i.i.d. Gaussian errors  $u_i \sim N(0, \tau^2)$ . Initial values are specified by  $f(x_1) \sim N(0, c_1)$ ,  $f(x_2) \sim N(0, c_2)$ . Diffuse initial priors are defined as the limiting case  $c_1, c_2 \rightarrow \infty$ .

The random walk model (5.4.2) locally penalizes deviations from constant functions or straight lines, or, equivalently, deviations of first differences  $\Delta^1 f(x_i) = f(x_i) - f(x_{i-1})$  or second differences  $\Delta^2 f(x_i) = f(x_i) - 2f(x_{i-1}) + f(x_{i-2})$  from zero. The variance  $\tau^2$  controls the amount of smoothness: The penalty increases as the variance  $\tau^2$  becomes smaller, since larger differences  $\Delta^1 f(x_i)$  or  $\Delta^2 f(x_i)$  become more unlikely, thus enforcing smoothness.

The close connection to penalized least-squares estimation and smoothing splines can be seen as follows: For given variances or “hyperparameters”  $\sigma^2$  and  $\tau^2$ , the observation model (5.4.1) implies the Gaussian distribution

$$p(y|f) \propto \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - f(x_i))^2 \right\} \quad (5.4.3)$$

for the data  $y$  conditional upon  $f$ . If one assumes diffuse initial priors for  $f(x_1)$  and  $f(x_2)$ , the local smoothness priors (5.4.2) define a multivariate Gaussian distribution

$$p(f) \propto \exp \left\{ -\frac{1}{2\tau^2} \sum_{i=d+1}^n (\Delta^d f(x_i))^2 \right\} \quad (5.4.4)$$

for the vector  $f$  of function evaluations. Thus, the posterior  $p(f|y) \propto p(y|f)p(f)$  is also Gaussian and characterized by the posterior mean  $\hat{f} = E(f|y)$  and covariance matrix  $\text{cov}(f|y)$ . Due to normality,  $\hat{f}$  coincides with the posterior mode of  $p(f|y)$  and can be obtained by maximizing  $p(y|f)p(f)$ . Taking logarithms and assuming diffuse initial priors, it is easily seen that  $\hat{f}$  is obtained by minimizing the penalized least-squares criterion

$$PLS(f) = \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \sum_{i=d+1}^n (\Delta^d f(x_i))^2 \longrightarrow \min_f \quad (5.4.5)$$

with  $\Delta^d$ ,  $d = 1, 2$ , as the first- or second-order difference operator and smoothness parameter  $\lambda = \sigma^2/\tau^2$ . Estimation of  $f$  via (5.4.5) is the classical “method of graduation” of Whittaker (1923) and has a non-Bayesian interpretation, with the first term as a measure of goodness-of-fit, the second as a roughness penalty, and  $\lambda$  controlling the bias-variance trade off. The penalty terms are discretized versions of corresponding penalty terms

$$\int (f'(x))^2 dx, \quad \text{resp.,} \quad \int (f''(x))^2 dx,$$

leading to quadratic (resp., cubic) spline smoothing as in Section 5.1.2. Already for a moderate number of observations, smoothing splines and discretized versions  $\hat{f}$  are virtually indistinguishable.

In matrix notation (5.4.5) can be written as

$$PLS(f) = (y - f)'(y - f) + \lambda f' K f \quad (5.4.6)$$

with appropriately defined penalty matrix  $K$ . For example, one gets

$$K = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}$$

for a first-order random walk. Criterion (5.4.6) has the same form as the corresponding penalized least-squares criterion (5.1.8) leading to smoothing splines, but with a different penalty matrix  $K$ .

The posterior mean or mode estimate  $\hat{f}$  is given by

$$\hat{f} = E(f|y) = (I + \lambda K)^{-1} y,$$

with smoother matrix  $S = (I + \lambda K)^{-1}$ ; compare with (5.1.10) for the case of cubic smoothing splines.

For non-equally spaced observations, local smoothness priors have to be modified appropriately to account for non-equal distances  $h_i = x_i - x_{i-1}$  between observations. Now  $x_1 < \dots < x_m$ ,  $m \leq n$ , are the *strictly ordered, different* observations of the covariate  $x$ , and  $f = (f(x_1), \dots, f(x_m))'$  is the vector of unknown function evaluations. First-order random walk models are now specified by

$$f(x_i) = f(x_{i-1}) + u_i, \quad u_i \sim N(0; h_i \sigma^2),$$

i.e., by adjusting error variances from  $\sigma^2$  to  $h_i \sigma^2$ . Second-order random walk models generalize to

$$f(x_i) = \left(1 + \frac{h_i}{h_{i-1}}\right) f(x_{i-1}) - \frac{h_i}{h_{i-1}} f(x_{i-2}) + u_i, \quad u_i \sim N(0; g_i \sigma^2),$$

where  $g_i$  is an appropriate weight. The simplest weight is again  $g_i = h_i$ , but more complex forms may also be used. Again, the posterior mean  $\hat{f}$  can be obtained by minimizing the PLS criterion (5.4.6), with a modified penalty matrix  $K$ . For example, it is easy to see that for a random walk of first-order the penalty matrix is

$$K = \begin{pmatrix} h_2^{-1} & -h_2^{-1} & & & \\ -h_2^{-1} & h_2^{-1} + h_3^{-1} & -h_3^{-1} & & \\ & -h_3^{-1} & h_3^{-1} + h_4^{-1} & -h_4^{-1} & \\ & & & \ddots & \\ & -h_{m-2}^{-1} & h_{m-2}^{-1} + h_{m-1}^{-1} & -h_{m-1}^{-1} & \\ & & -h_{m-1}^{-1} & h_{m-1}^{-1} + h_m^{-1} & -h_m^{-1} \\ & & & -h_m^{-1} & h_m^{-1} \end{pmatrix}.$$

From a Bayesian viewpoint, we are interested not only in the posterior mean but in the entire posterior distribution. In matrix form the Gaussian observation density (5.4.3) is

$$p(y|f) \propto \exp\left(-\frac{1}{2\sigma^2}(y-f)'(y-f)\right),$$

and the Gaussian smoothness prior (5.4.4) is

$$p(f) \propto \exp\left(-\frac{1}{2\tau^2}f'Kf\right). \quad (5.4.7)$$

(Note that this is a partially improper prior, since  $\text{rank}(K) = n-d$  for first- ( $d=1$ ) or second- ( $d=2$ ) order random walk priors.)

Inserting into  $p(f|y) \propto p(y|f)p(f)$  and rearranging yields

$$p(f|y) \propto \exp\left(-\frac{1}{2}(f-Sy)'(S\sigma^2)^{-1}(f-Sy)\right),$$

i.e., the multivariate Gaussian posterior

$$f|y \sim N(Sy, S\sigma^2). \quad (5.4.8)$$

Result (5.4.8) remains valid for any choice of the penalty matrix as long as the smoother matrix  $S = (I + \lambda K)^{-1}$  is well defined and nonsingular. For example, with the penalty matrix (5.1.14) for cubic smoothing splines, we obtain a Bayesian version for spline fitting. Assuming a stochastic process prior in form of an integrated Wiener process, Wahba (1978) derived Bayesian spline models with posterior (5.4.8).

In practice it often occurs that the same  $x$ -values are observed for different observations. Some modifications are necessary then. Now let  $x_{(1)} < x_{(2)} < \dots < x_{(m)}$ ,  $m \leq n$ , denote the  $m$  *distinct* and *ordered*  $x$ -values observed in the sample  $(x_i, y_i)$ ,  $i = 1, \dots, n$ , and

$$f = (f(x_{(1)}), \dots, f(x_{(m)}))'$$

be the corresponding vector of function evaluations. Local or global smoothness priors remain the same; only  $x_i$  and  $n$  are replaced by  $x_{(i)}$  and  $m$  everywhere. The smoother matrix generalizes to

$$S = (X'X + \lambda K)^{-1}, \quad (5.4.9)$$

and the posterior mean is

$$\hat{f} = E(f|y) = SX'y. \quad (5.4.10)$$

Here  $X = (x_{ij})$  is an  $n \times m$  incidence matrix with  $x_{ij} = 1$  if observation  $i$  has covariate value  $x_i = x_{(j)}$ , and  $x_{ij} = 0$  otherwise. It is easy to see that  $X'X$  is diagonal, where the  $j$ th diagonal element is the number of observations with covariate value  $x_{(j)}$ , and  $X'y$  is the sum of response values observed at  $x_{(j)}$ .

Instead of first defining local smoothness priors, implying certain forms for  $K$ , Hastie & Tibshirani (2000) start directly from a *global smoothness prior* (5.4.7) with an appropriately chosen penalty matrix  $K$ . With  $K$  as in (5.1.9), they get a Bayesian cubic spline smoother with posterior (5.4.8).

Until now variances  $\sigma^2$  and  $\tau^2$  were considered fixed or known. In practice they have to be determined as well. In a fully Bayesian approach, priors are put on the variances and their posteriors are estimated along with the function  $f$ . Since  $\tau^2$  acts as a smoothness parameter, the Bayesian approach therefore allows simultaneous estimation of smooth functions and the smoothness parameter. The standard choices are independent highly dispersed inverse gamma (IG) priors

$$p(\sigma^2) \sim IG(a_0, b_0), \quad p(\tau^2) \sim IG(a_1, b_1).$$

The posterior is then

$$p(f, \sigma^2, \tau^2 | y) \propto p(y|f, \sigma^2)p(f|\tau^2)p(\sigma^2)p(\tau^2).$$

The posterior  $p(f|y)$  is no longer Gaussian, but intractable analytically. Therefore, Bayesian inference is carried out by Gibbs sampling (Appendix A.5). Posterior samples are drawn sequentially from the full conditionals

$$\begin{aligned} p(f|y, \sigma^2, \tau^2) &\sim N(Sy, S\sigma^2), \\ p(\sigma^2|y, f, \tau^2) &= p(\sigma^2|y, f), \\ p(\tau^2|y, f, \sigma^2) &= p(\tau^2|f). \end{aligned}$$

Efficient sampling from the high-dimensional Gaussian posterior for  $f$  avoids explicit computation of  $S$ . This can be done either by sampling schemes of Frühwirth-Schnatter (1994) or Carter & Kohn (1994a), developed in the related context of state space models and Kalman filtering, or by making clever use of the sparse structure of penalty matrices as Hastie & Tibshirani (2000) suggest for Bayesian smoothing splines and Rue (2000) in a more general context. An alternative, which also works for non-Gaussian responses, is the MCMC technique with conditional prior proposals developed by Knorr-Held (1999), again in the context of state space models.

The posteriors for  $\sigma^2$  and  $\tau^2$  are inverse gamma distributions

$$p(\sigma^2|\cdot) \propto IG(a'_0, b'_0), \quad p(\tau^2|\cdot) \propto IG(a'_1, b'_1)$$

with updated parameters

$$\begin{aligned} a'_0 &= a_0 + n/2, & b'_0 &= b_0 + (y - f)'(y - f)/2 \\ \text{and} \quad a'_1 &= a_1 + \text{rank}(K)/2, & b'_1 &= b_1 + \frac{1}{2}f'Kf. \end{aligned} \quad (5.4.11)$$

Thus, posterior samples for  $\tau^2$  and  $\sigma^2$  can be drawn directly, and estimation of variances or smoothing parameters is automatically included in Bayesian inference via Gibbs samling.

A related approach is to work with stochastic differential smoothness priors for smoothing splines and state space models for non-equally spaced observations derived from them; see Carter & Kohn (1994a) and Wong & Kohn (1996).

## Basis Function Approaches

A second-option is Bayesian basis function approaches where  $f$  is modelled as a linear combination

$$f(x) = \sum_{j=1}^m \delta_j B_j(x)$$

of linearly independent basis functions, like spline functions, piecewise polynomials, and wavelets. Generally, it is difficult to determine which basis functions should be included. If the basis has too many functions, estimates for  $\delta_i$  will have high variability or may even produce interpolation in the extreme case. Conversely, if too few functions are in the basis, the resulting estimator will be severely biased.

Smith & Kohn (1996) propose approximating  $f(x)$  by cubic regression splines using the truncated power-series basis. A Bayesian variable selection approach is used to determine the significant knots and to estimate the parameters  $\delta$ . Denison, Mallick & Smith (1998) use a wider class of piecewise polynomials as basis functions that incorporate smoothing splines as

a subclass. However, it is more flexible because unsmooth functions with rapidly varying first and second derivatives or even discontinuities can also be modelled adequately. The number and location of knots are considered to be unknowns, and sampling from posteriors for both the number and the location of knots is addressed using the reversible jump MCMC simulation technique of Green (1995). The method is some hybrid technique between an empirical and a fully Bayesian approach in that the  $\beta$ 's are estimated by least squares conditional upon the number and location of knots as well as  $\sigma^2 = \text{var}(\varepsilon_i)$ . A fully Bayesian approach using a B-spline basis and reversible jump MCMC for automatic knot selection and estimation of basis coefficients is developed in Biller (2000a).

A Bayesian version of P-splines described in Section 5.1.1 is obtained if the basis functions  $B_j(x)$  are B-splines and the coefficients  $\delta_1, \dots, \delta_j, \dots, \delta_m$  follow a random walk of first- or second-order. Replacing the vector  $f = (f(x_{(1)}), \dots, f(x_{(m)}))'$  by the coefficient vector  $\delta = (\delta_1, \dots, \delta_m)'$  and defining the matrix  $X$  in (5.4.7), (5.4.8) by  $x_{ij} = B_j(x_i)$ , the same arguments as above lead to the full conditionals

$$p(\delta|y, \sigma^2, \tau^2) \sim N(SX'y, S\sigma^2).$$

Together with inverse gamma full conditionals for  $\sigma^2$  and  $\tau^2$ , we obtain a Gibbs sampler for Bayesian P-splines, including estimation of the smoothing parameter  $\tau^2$ . Details are provided in Lang & Brezger (2000).

## Models with Multiple Covariates

Univariate function sampling serves as a building block for Bayesian additive models

$$y_i = \alpha + \sum_{j=1}^p f_{(j)}(x_{ij}) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

For identifiability reasons, the unknown functions are centered about zero.

The Gibbs sampler is tailor-made for drawing posteriors from Bayesian additive models and can be viewed as “Bayesian backfitting” (Hastie & Tibshirani, 2000). Suppressing dependence on hyperparameters like variances, etc., the generic form of the Gibbs sampler for the vectors  $f_j = (f_{(j)}(x_{ij}), \dots, f_{(j)}(x_{nj}))'$  of function evaluations is as follows:

Step 1 Initialize  $f_0 = \hat{\alpha}$ , e.g.,  $\hat{\alpha} = \bar{y}$ , and  $f_j = f_j^0$ , e.g.,  $f_j^0 = 0$  or a penalized least-squares estimate.

Step 2 Draw repeated posterior samples from the Gaussian full conditionals

$$p(f_j|f_l, l \neq j, y).$$

In the “Bayesian backfitting” iterations of Step 2, one of the univariate function samplers described above can be employed. Smoothing parameters or hyperparameters can be included by assigning hyperpriors, for example, inverse gamma priors to the variances  $\sigma^2$  and  $\tau_j^2$ ,  $j = 1, \dots, p$ , as in the univariate case.

Extensions to more general forms of additive predictors are conceptually straightforward. For example, in semiparametric models

$$y_i = \sum_{j=1}^p f_{(j)}(x_{ij}) + w_i' \beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

a diffuse or informative Gaussian prior is assigned to the vector  $\beta$  of covariate effects, and an additional updating step is included to sample from the posterior  $p(\beta|f_j, j = 1, \dots, p, y)$ .

For the smoothness priors approach we assign independent priors

$$p(f_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2} f_j' K_j f_j\right), \quad j = 1, \dots, p, \quad (5.4.12)$$

to the vectors  $f_j$  of function evaluations, where  $K_j$  are penalty matrices, inverse Gamma priors  $IG(a_j, b_j)$ ,  $j = 0, 1, \dots, p$ , to the variances  $\sigma^2, \tau_1^2, \dots, \tau_p^2$  and a Gaussian or diffuse prior  $p(\beta)$  for  $\beta$ . Bayesian model specification is completed by assuming conditionally independent observations, given covariates and parameters, and independent priors for parameters. Then the posterior factorizes into

$$\begin{aligned} & p(\beta, \sigma^2, f_1, \dots, f_p, \tau_1^2, \dots, \tau_p^2 | y) \\ & \propto \prod_{i=1}^n L_i(y_i; \eta_i) \times \prod_{j=1}^p \{p(f_j|\tau_j^2)\} p(\tau_j^2) \pi(\sigma^2) \pi(\beta), \end{aligned} \quad (5.4.13)$$

where  $L_i(y_i; \eta_i)$  is the likelihood contribution of observation  $y_i$ , and  $\eta_i$  is the additive or semiparametric predictor. Bayesian inference via Gibbs sampling can be carried out by iterative draws from full conditionals. With a diffuse prior  $p(\beta) \propto const$  for  $\beta$ , full conditionals can be derived as

- (i)  $p(f_j|\cdot) \sim N(S_j X_j'(y - \tilde{\eta}_j), S_j \sigma^2)$ ,  $j = 1, \dots, p$ ,
- (ii)  $p(\beta|\cdot) \sim N((W'W)^{-1} W'(y - \tilde{\eta}_0), (W'W)^{-1} \sigma^2)$ ,
- (iii)  $p(\sigma^2|\cdot) \sim IG(a'_0, b'_0)$ ,
- (iv)  $p(\tau_j^2|\cdot) \sim IG(a'_j, b'_j)$ ,  $j = 1, \dots, p$ .

Here “.” denotes the remaining parameters and the data;  $W$  is the design matrix for fixed effects  $\beta$ ; and  $\tilde{\eta}_j$ ,  $j = 0, \dots, p$ , is the part of the predictor associated with the remaining effects in the model. The updated inverse Gamma parameters are

$$\begin{aligned} a'_0 &= a_0 + \frac{n}{2}, & b'_0 &= b_0 + \frac{1}{2} \sum_{i=1}^n (y_i - \eta_i)^2, \\ a'_j &= a_j + \frac{\text{rank}(K_j)}{2}, & b'_j &= b_j + \frac{1}{2} f'_j K_j f_j. \end{aligned} \quad (5.4.14)$$

The Gibbs sampler draws iteratively from (i) to (iv) until convergence and enough samples are obtained. As for univariate function estimation, draws from the high-dimensional Gaussian posteriors  $p(f_j|\cdot)$  have to be implemented efficiently, using banded or sparse matrix operations.

### Example 5.9: Rental rates

According to German rental law, owners of apartments or flats can base an increase in the amount they charge for rent on “average rents” for flats comparable in type, size, equipment, quality, and location in a community. To provide information about these “average rents,” most larger cities publish rental guides, which can be based on regression analysis with rent as the dependent variable.

We use data for a random sample of more than 3000 flats from the city of Munich in 1998. The response variable is

$Y$  monthly net rent per square meter in German marks (DM), that is, the monthly rent minus calculated or estimated utility costs.

Covariates characterizing the flat were constructed from almost 200 variables out of a questionnaire answered by tenants. City experts assessed the flat’s location in the city in three categories (average, good, top). In the following reanalysis, we include 27 covariates. Here is a selection of typical covariates:

- $F$  floor space in square meters
- $A$  year of construction
- $L_1$  0|1 indicator for good location in the city (0 = average)
- $L_2$  0|1 indicator for top location in the city (0 = average)
- $H$  0|1 no central heating indicator
- $B$  0|1 no bathroom indicator
- $E$  0|1 indicator of bathroom equipment above average
- $K$  0|1 indicator of kitchen equipment above average
- $W$  0|1 indicator of no central warm-water system
- $S$  0|1 indicator of large balcony facing the south or west
- $O$  0|1 indicator of simple, old building constructed before 1949
- $R$  0|1 indicator of old building, renovated in good state
- $N$  0|1 indicator of new, modern building built after 1978.

In this section we analyze the data with a Gaussian semiparametric model with predictor

$$\eta = \alpha + f_{(1)}(F) + f_{(2)}(A) + w'\beta + \beta_1 L_1 + \beta_2 L_2,$$

where  $w$  is a vector of 25 binary indicators, including those mentioned above. Previous data analysis showed that the assumption of a Gaussian error structure and the form of the predictor provide a reasonably realistic model. Priors for  $f_{(1)}, f_{(2)}$  are second-order random walk models. The prior for  $\beta$  is diffuse.

Figure 5.19 shows the influence of floor space on rents: Small flats and apartments are more expensive than larger ones, but this nonlinear effect becomes smaller with increasing floor space. The effect of year of construction on rents is more or less constant until the 1950s. It then distinctly increases until about 1990, and it stabilizes on a high level in the 1990s.

Estimates for fixed effects of selected binary indicators are given in Table 5.1, showing positive or negative effects as to be expected. By construction, the estimated effect is the additional positive or negative amount of net rent per square meter caused by the presence (= 1) of the indicator for a flat.

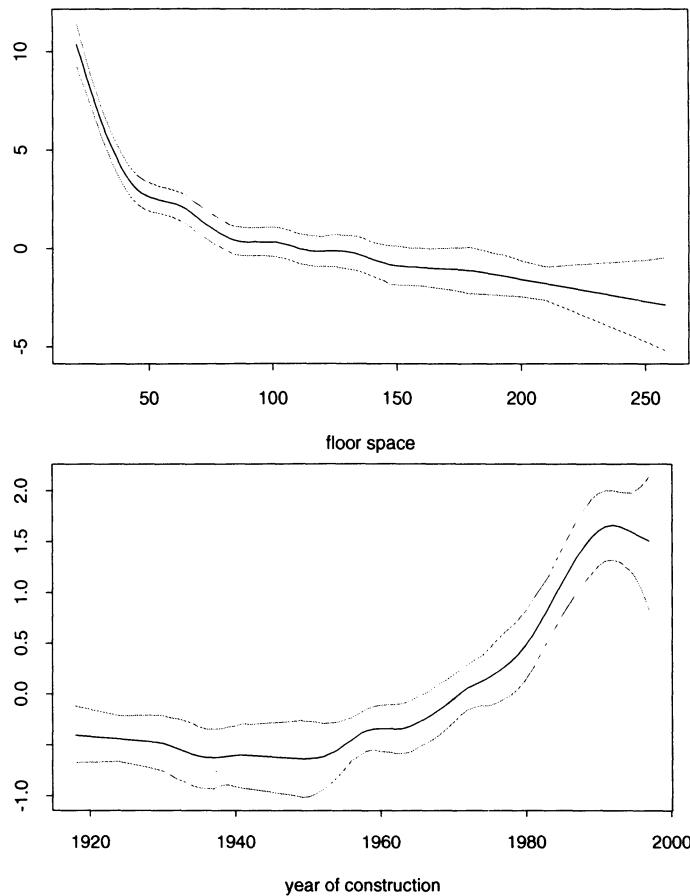
Covariate	Mean	10% quantile	90% quantile
<i>const</i>	13.32	12.57	14.04
<i>H</i>	-1.37	-1.75	-1.00
<i>B</i>	-1.33	-1.78	-0.89
<i>E</i>	0.50	0.15	0.83
<i>K</i>	0.88	0.48	1.30
<i>W</i>	-2.42	-2.88	-1.97
<i>S</i>	0.74	0.32	1.18
<i>O</i>	-1.91	-2.23	-1.59
<i>R</i>	1.48	0.79	2.15
<i>N</i>	0.85	0.47	1.25
<i>L</i> <sub>1</sub>	1.09	0.90	1.29
<i>L</i> <sub>2</sub>	2.35	1.82	2.88

**Table 5.1.** Estimates of selected covariate effects

In Section 8.5 we will carry out a spatial analysis of the effect of location, without using the experts' assessment. Such a data-driven nonparametric estimate can then be used to validate the experts' opinion.  $\square$

### 5.4.2 Non-Gaussian Responses

For non-Gaussian responses, the observation model (5.4.1) is replaced by the usual exponential family assumption for responses  $y_i$ , conditional on



**Figure 5.19.** Estimated effects of floor space and year of construction for the rent data. Shown are the posterior means within 80% credible regions.

covariates and unknown functions and parameters, and conditional means  $\mu_i$  are related to additive or semiparametric predictors

$$\eta_i = \sum_{j=1}^p f_{(j)}(x_{ij}) + w'_i \beta$$

by a link  $\mu_i = h(\eta_i)$ . For Bayesian inference, smoothness priors and basis function approaches are again suitable, in principle. However, a main feature of Bayesian inference for generalized additive models is that the posteriors for unknown vectors  $f_j$  of function evaluations are no longer multivariate Gaussian. Hence direct Gibbs sampling is no longer feasible, and more general Metropolis-Hastings algorithms are needed.

Hastie & Tibshirani (2000) make a corresponding suggestion for extending their “Bayesian backfitting” algorithm to generalized additive models.

Fahrmeir & Lang (1999) retain the local smoothness priors approach by assuming independent random walk models (5.4.2), resp., priors (5.4.12), for unknown functions  $f_j$ ,  $j = 1 \dots, p$ , diffuse or Gaussian priors for  $\beta$ , and inverse gamma priors  $IG(a_j, b_j)$  for the error variances  $\tau_j^2$ ,  $j = 1, \dots, p$  of the random walk models. To obtain posterior samples for function evaluations  $f_j$ , they adopt and extend the MH algorithm with conditional prior proposals developed by Knorr-Held (1999) for dynamic generalized linear models, compare Section 8.3.2. Computational efficiency combined with good convergence and mixing behavior of the generated Markov chains is achieved by partitioning each vector  $f_j$  of function evaluations into  $v_j$  subvectors or “blocks”

$$f'_j = (f'_j[1], \dots, f'_j[r], \dots, f'_j[v_j])$$

and updating blocks instead of the entire vector  $f_j$ .

Updating steps for  $\beta$  are carried out with Gamerman's (1997a) MH algorithm described in Section 2.3.2, while drawings for variance parameters are directly available from inverse gamma posteriors with updated parameters.

Taken together, the following hybrid MCMC scheme is obtained:

- (i) Partition each vector  $f_j$  of function evaluations into blocks

$$(f'_j[1], \dots, f'_j[r], \dots, f'_j[v_j])'$$

- (ii) For each  $j$  and each block, draw samples from the full conditionals, given all other function evaluations,

$$p(f_j[1]|\cdot), \dots, p(f_j[r]|\cdot), \dots, p(f_j[v_j]),$$

by MH steps with conditional prior proposals (as described in Section 8.3.2).

- (iii) Draw samples from  $p(\beta|\cdot)$  by Gamerman's weighted least-squares proposal (2.3.10), described in Section 2.3.2.
- (iv) Draw samples from inverse Gamma posteriors

$$p(\tau_j^2|\cdot) \sim IG(a'_j, b'_j), \quad j = 1, \dots, p,$$

with parameters  $a'_j, b'_j$  as in (5.4.14).

- (v) Iterate steps (i) to (iv) until enough samples are obtained.

A B-spline *basis function* approach with adaptive knot selection for generalized additive models with semiparametric predictor has been developed by Biller (2000c). Several examples indicate good convergence properties. Mallick, Denison & Smith (1999) propose Bayesian multivariate adaptive regression splines by an extension of the curve-fitting method for Gaussian responses of Denison, Mallick & Smith (1998). For the extension to GAMs

they use a simple Metropolis-Hastings proposal, but the sampler seems to have slow convergence. As for Gaussian responses, Bayesian P-spline inference can be developed by assigning a smoothness prior to the vector  $\delta_j = (\delta_{1j}, \dots, \delta_{mj})'$  of B-spline coefficients for  $f_j(s_j) = \sum_j \delta_j B_j(x_j)$ . The resulting hybrid MCMC scheme resembles the sampling scheme above, replacing  $f_j$  by  $\delta_j$ ; see Lang & Brezger (2000) for details.

### Latent Variable Models for Categorical Responses

For categorical responses  $Y_i$ , Bayesian inference can be based on semiparametric regression models for latent continuous variables or utilities  $U_i$ . These latent models generate semiparametric categorical response models via a threshold of utility mechanisms; see Sections 3.2 and 3.3. This is particularly attractive for latent Gaussian models, defining probit models for observed responses. Sampling schemes for Gaussian responses of Section 5.4.1 can then be used as major building blocks.

For the case of ordinal responses  $Y_i$  with ordered categories  $1, \dots, k$ , the simplest mechanism is a threshold mechanism

$$Y_i = r \iff \theta_{r-1} < U_i \leq \theta_r, \quad r = 1, \dots, k,$$

with ordered thresholds  $-\infty = \theta_0 < \theta_1 < \dots < \theta_q < \theta_k = +\infty$ , leading to cumulative models. If  $U_i$  obeys a semiparametric Gaussian model

$$U_i = \eta_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1),$$

with predictor

$$\eta_i = \sum_{j=1}^p f_{(j)}(x_{ij}) + w_i' \beta, \quad (5.4.15)$$

then  $Y_i$  follows a corresponding semiparametric probit model

$$P(Y_i \leq r) = \Phi(\theta_r - \eta_i).$$

For identifiability reasons, the linear term must not contain an intercept  $\beta_0$ . Otherwise, one of the thresholds has to be fixed, e.g., by setting

$$\theta_1 = 0.$$

Latent variable mechanisms for more complex models (e.g., sequential models) are described in Section 3.3.

For nominal responses  $Y_i$  let  $U_i$  be latent variables associated with category  $r$  and assume a latent semiparametric Gaussian regression model

$$U_{ir} = \eta_{ir} + \varepsilon_{ir}$$

with category-specific predictors

$$\eta_{ir} = \alpha_r + \sum_{j=1}^p f_{(j)}^r(x_{ij}) + w_i' \beta_r, \quad r = 1, \dots, q. \quad (5.4.16)$$

For the reference category  $k$  we set  $\eta_{ik} = 0$  to ensure identifiability. The principle of random utility postulates that

$$Y_{ir} \iff U_{ir} = \max_{l=1, \dots, k} U_l.$$

If the  $\varepsilon$ 's are i.i.d. standard normal, one gets the independent probit model. Correlated errors lead to multivariate probit models.

The following smoothness priors approach has been developed in Fahrmeir & Lang (2000). With diffuse priors  $p(\beta) \propto \text{const}$ ,  $p(\theta) \propto \text{const}$  for fixed parameters and thresholds, posterior analysis is now based on

$$p(f, \tau^2, \beta, \theta, U|Y) \propto \prod_{i=1}^n \{p(Y_i|U_i)L_i(U_i; \eta_i)\} p(f|\tau^2)p(\tau^2),$$

where  $f$  and  $\tau^2$  are vectors containing all function evaluations and variances,  $U = (U_1, \dots, U_n)'$ , with  $U_i = (U_{i1}, \dots, U_{ik})'$  for nominal response. Threshold parameters  $\theta$  have to be included for ordinal models only. The “likelihood” contributions  $L_i(U_i; \eta_i)$  are defined by the latent Gaussian models (5.4.15) and (5.4.16), with the same priors for  $f$  and  $\tau^2$  as for Gaussian observations in (5.4.13). The conditional “likelihoods”  $p(Y_i|U_i)$  are determined by the latent variable mechanisms.

For cumulative ordinal models, we get

$$p(Y_i|U_i) = \sum_{l=1}^k I(\theta_{l-1} < U_l \leq \theta_l) I(Y_i = l),$$

with the 0 – 1 indicator function  $I(\cdot)$ . Note that  $p(Y_i|U_i)$  is 1 if  $U_i$  obeys the restriction imposed by the observed value of  $Y_i$  and is 0 otherwise.

For a nominal response, we have

$$p(Y_i|U_{i1}, \dots, U_{ik}) = \sum_{l=1}^k I(\max(U_{i1}, \dots, U_{ik}) = U_{ir}) I(Y_i = r).$$

For a Gibbs sampling procedure, additional draws from full conditionals for the latent variables are necessary. For the following we assume (independent) Gaussian errors, that is, we consider semiparametric probit models. Then these full conditionals are truncated normals, subject to the restraints imposed by the formulas for  $p(Y_i|U_i)$  and  $p(Y_i|U_{i1}, \dots, U_{ik})$ .

For cumulative models, the latent variables  $U_i$ ,  $i = 1, \dots, n$ , are sampled as follows. If  $Y_i = r$ , then  $U_i$  is generated from a normal distribution with

mean  $\eta_i$  (evaluated at the current sampling values of functions and parameters) and variance 1, subject to the truncation constraint  $\theta_{r-1} < U_i \leq \theta_r$ . The full conditionals for threshold parameters  $\theta_r$  are uniform on the intervals

$$[\max\{\max\{U_i : Y_i = r\}, \theta_{r-1}\}, \min\{\min\{U_i : Y_i = r+1\}, \theta_{r+1}\}];$$

see Albert & Chib (1993). As Chen & Dey (1999) point out, posterior samples may be slowly converging. Therefore, improvements have been suggested; see Chen & Dey (1999) and references given therein.

For nominal responses, we choose  $k$  as the reference category. As only differences of utilities can be identified (see Section 3.2), we may either set the predictor  $\eta_{ik}$  to zero or the latent variable  $U_{ik}$ . Setting  $U_{ik} \equiv 0$ , latent variables  $U_{ir}$ ,  $r = 1, \dots, k-1$ , are generated as follows for each observation  $Y_i$ ,  $i = 1, \dots, n$ . If  $Y_i = r$ ,  $r \neq k$ , then  $U_{ir}$  is generated first from a normal distribution with mean  $\eta_{ir}$  and variance 1, subject to the constraints  $U_{ir} > U_{il}$ ,  $l \neq k$ , and  $U_{ir} > 0$  ( $\equiv U_{ik}$ ). Next we generate  $U_{il}$  for  $l \neq r$  from a normal distribution with mean  $\eta_{il}$  and variance 1, subject to the constraint that  $U_{il}$  is less than the  $U_{ir}$  generated just before.

If  $Y_i = k$  (the reference category), then we generate  $U_{il}$ ,  $l = 1, \dots, k$ , from a normal with mean  $\eta_{il}$  and variance 1, subject to the constraint  $U_{il} < 0$ .

Since observations  $Y$  are completely determined by latent variables  $U$ , full conditionals  $p(f_j|U, Y, \cdot)$ ,  $f(\tau_j^2|U, Y, \cdot)$ , and  $p(\beta|U, Y, \cdot)$  for functions and parameters reduce to Gaussian full conditionals  $p(f_j|U, \cdot)$  and  $p(\beta|U, \cdot)$ , and inverse gamma full conditionals  $f(\tau_j^2|U)$ . Therefore, posterior samples for these full conditionals can be generated as in Section 5.4.1 for Gaussian responses. With the smoothness priors approach, posterior samples for  $f_j$ ,  $\tau_j^2$ ,  $j = 1, \dots, p$ , and  $\beta$  can be generated just as in Section 5.4.1, replacing  $Y$  by  $U$ ; see Fahrmeir & Lang (2000). Basis function approaches are another possibility, see Yau, Kohn & Wood (2000) for recent work.

We conclude with a final remark: In this section we have considered Bayesian inference for non- and semiparametric regression models as an alternative to the non-Bayesian methods described in the foregoing sections of this chapter. One of the major advantages of Bayesian approaches is that they can be extended to more complex situations in a unified but flexible framework. In Chapters 7, 8, and 9 we consider extensions to problems where correlation caused by individual, serial, or spatial heterogeneity is taken into account by incorporating appropriate types of random effects within a joint model.

### **Example 5.10: Credit scoring revisited**

Analysis is based on the same data set as in Examples 2.3, 2.7, and 4.1. A parametric logit model for the probability  $P(y = 1|x)$  of being not credit-worthy led to the conclusion that the covariate “amount of credit” has no

significant influence on the risk. Here, we reanalyze the data with a semiparametric logit model

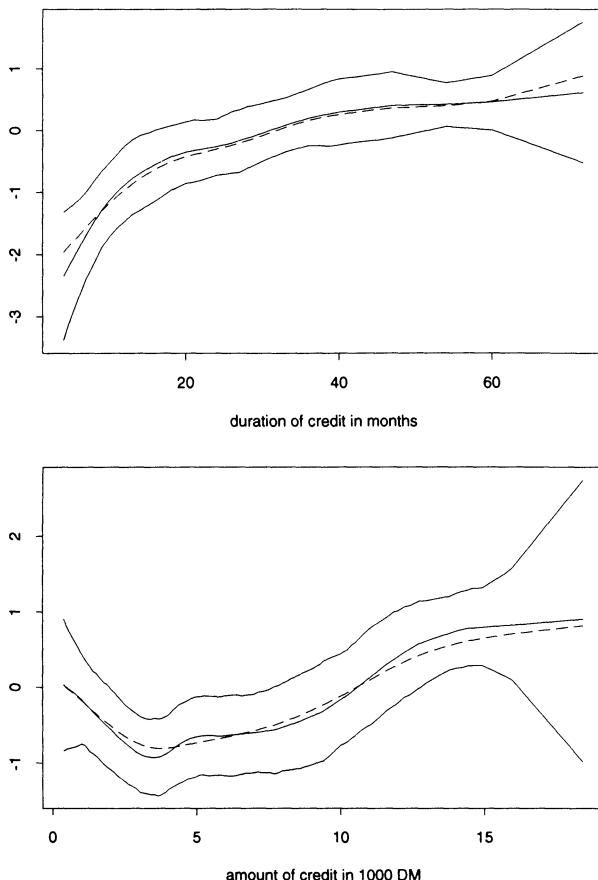
$$\log \frac{\text{pr}(y = 1|x)}{1 - \text{pr}(y = 1|x)} = \beta_0 + \beta_1 x_1^1 + \beta_2 x_1^2 + f_3(x_3) + f_4(x_4) \\ + \beta_5 x_5 + \beta_6 x_6 + \beta_8 x_8,$$

where  $x_1^1$  and  $x_1^2$  are dummies for the categories “good” and “medium” running accounts. The predictor has a semiparametric additive form: The smooth functions  $f_3(x_3), f_4(x_4)$  of the metrical covariates “duration of credit” and “amount of credit,” are estimated nonparametrically using second-order random walk models for non-equally spaced observations. The constant  $\beta_0$  and the effects  $\beta_1, \beta_2, \beta_5, \beta_6, \beta_8$  of the remaining categorical covariates are considered as fixed and estimated jointly with the curves. In contrast to previous analysis, effect coding is used for the categorical covariate. Figure 5.20 shows estimates for the curves  $f_3$  and  $f_4$ . For comparison, cubic smoothing splines are included in addition to posterior mean estimates. Although cubic splines are posterior mode estimators and the penalty terms are not exactly the same, both estimates are close. While the effect of the variable “duration of credit” is almost linear, the effect of “amount of credit” is clearly nonlinear. The curve has a bathtub shape and indicates that not only high credits but also low credits increase the risk, compared to “medium” credits between 3000–6000 DM. Apparently, if the influence is misspecified by assuming a linear function  $\beta_4 x_4$  instead of  $f_4(x_4)$ , the estimated effect  $\hat{\beta}_4$  will be near zero, corresponding to an almost horizontal line  $\hat{\beta}_4 x_4$  near zero, and falsely considered as nonsignificant.

Covariate	Mean	10% quantile	90% quantile	ML estimator
$x_1^1$	0.85	0.62	1.08	0.86
$x_1^2$	-1.09	-1.34	-0.82	-1.09
$x_5$	-0.50	-0.76	-0.24	-0.50
$x_6$	-0.22	-0.38	-0.05	-0.22
$x_8$	-0.27	-0.44	-0.10	-0.26

**Table 5.2.** Estimates of constant parameters for the credit-scoring data

Table 5.2 gives the posterior means together with 80% credible intervals and, for comparison, maximum likelihood estimates of the remaining effects. Both estimates have similar values. Transforming between effect and dummy coding also shows that estimates are quite close to the estimates in Examples 2.3 and 2.7, so that interpretation remains qualitatively the same for these constant effects.



**Figure 5.20.** Estimated effects of duration and amount of credit. Shown are the posterior means within 80% credible regions and, for comparison, cubic smoothing splines (dotted lines).

Closeness of fixed effects is in agreement with asymptotic normality results for penalized likelihood estimation in semiparametric GLMs by Mammen & van de Geer (1997). Closeness of posterior means and methods of nonparametric components is in agreement with empirical evidence and our own experience in the related context of non-Gaussian state space models. A heuristic justification is given in Fahrmeir & Wagenpfeil (1997), but rigorous asymptotic results seem to be missing.

A main interest with credit-scoring is prediction, that is, to predict the probability of failure  $\pi_0$  for a new client with characteristics  $x_0$ . Within a Bayesian framework this can easily be done just by drawing samples from the predictive distribution. The advantage compared to other semi-

parametric approaches is that we get not only point estimates but also exact credible intervals to assess the uncertainty of the prediction. Suppose, for example, that we have a new client with covariate vector  $x_0 = (-1, -1, 16, 1.4, 1, -1, 1)$ . Then sampling from the predictive distribution results in a predicted probability of  $\pi_0 = 0.35$  (posterior mean) and a credible interval of (0.2, 0.54) based on the posterior 10% and 90% quantiles. It should be noted that the credit scoring data come from a stratified sample with 300 “bad” and 700 “good” credits. This explains the high value  $\pi_0$  of risk. To obtain realistic values, a stratification correction is necessary; compare to Anderson (1972).  $\square$

The example illustrates the following features of Bayesian generalized additive modelling: MCMC simulation provides samples from all posteriors of interest and permits estimation of posterior means, medians, quantiles, confidence bands, and predictive distributions. No approximations based on (conjectures of) asymptotic normality have to be made, and no “plug in” procedures are needed. A data driven choice of smoothing parameters is automatically included.

On the other side, convergence of MCMC is always an issue. Therefore, if possible, comparison with penalized likelihood estimates, or, from a Bayesian viewpoint, posterior mode estimates is recommended.

## 5.5 Notes and Further Reading

There is a growing body of literature that uses nonparametric fitting techniques as a diagnostic tool for linear or generalized linear models; see, e.g., Azzalini, Bowman & Härdle (1989), Staniswalis & Severini (1991), Firth, Glosup & Hinkley (1991), le Cessie & van Houwelingen (1991), Eubank & Hart (1993), Azzalini & Bowman (1993), and Kauermann & Tutz (1999, 2000b). A comprehensive overview of smooth tests can be found in Hart (1997). Surveys of categorical data smoothing are found in Simonoff (1996) and Simonoff & Tutz (2000).

Nonparametric and semiparametric approaches in this chapter are based on the assumption that responses  $y_i$  given the covariates  $x_i$  are (conditionally) independent. This assumption is mainly appropriate for cross-sectional data. However, the basic concepts, namely basis function approaches, localization, and penalization, can be transferred to more complex data situations such as longitudinal, spatial, survival, and event history data. Some extensions of the non- and semiparametric approaches of this chapter are discussed in the following chapters.

Our short presentation of Bayesian approaches focuses on approaches with basis functions and (Gaussian) smoothness priors for semiparametric

additive predictors. Alternative approaches are based on non-Gaussian priors or on stochastic process priors, e.g., (mixtures) of Dirichlet process priors (Müller, Erkanli & West, 1996), Lèvy processes, etc. The contributions in Dey, Müller & Sinha (1998) and the references therein provide a good survey on these methods as well as other semiparametric Bayesian statistical modelling (e.g. for neural networks and wavelets).

Gaussian process priors are also used for modelling error process in the design and analysis of computer experiments; see Koehler & Owen (1996).

# 6

## Fixed Parameter Models for Time Series and Longitudinal Data

The methods of the preceding chapters are mainly appropriate for modelling and analyzing a broad class of non-normal cross-sectional data. Extensions to time-dependent data are possible in a variety of ways. Time series are repeated observations  $(y_t, x_t)$  on a response variable  $y$  of primary interest and on a vector of covariates taken at times  $t = 1, \dots, T$ . Discrete time longitudinal or panel data are repeated observations  $(y_{it}, x_{it})$  taken for units  $i = 1, \dots, n$  at times  $t = 1, \dots, T_i$ . The restriction to integral times is made to simplify notation but is not necessary for most of the approaches. Longitudinal data may be viewed as a cross section of individual time series, reducing to a single time series for  $n = 1$ , or as a sequence of cross-sectional observations where units are identifiable over time. If a comparably small number of longer time series is observed, models and methods will be similar to those for single time series. If, however, many short time series have been observed, models, and often the scientific objective, can be different.

In this Chapter, we consider extensions of generalized linear models for time series (Section 6.1) and longitudinal data (Section 6.2) where parameters or covariate effects are fixed. In each section we consider conditional (“observation-driven”) and marginal models. This corresponds to the distinction between conditional and marginal approaches in Section 3.5. For longitudinal data, Chapters 8 and 10 in Diggle, Liang & Zeger (1994) provide a detailed introduction to both approaches. Models with parameters varying across units according to a mixing distribution or across time according to a stochastic process are treated in Chapters 7 and 8 of this book.

## 6.1 Time Series

While time series analysis for approximately Gaussian data has a long tradition, models for non-Gaussian data have received attention only more recently. Typical examples are binary or multicategorical time series, e.g., daily rainfall with categories no/yes or no/low/high (Example 1.7), and time series of counts, e.g., monthly number of cases of poliomyelitis (Example 1.8), daily number of purchases of some good, etc. Categorical and discrete valued time series were often analyzed as time homogeneous Markov chains, i.e., Markov chains with stationary transition probabilities. However, without further constraints the number of parameters increases exponentially with the order of the Markov chain, and in many applications non-homogeneous Markov chains are more appropriate since exogenous variables possibly give rise to nonstationary transition probabilities. For binary time series, Cox (1970) proposed an autoregressive logistic model, where covariates and a finite number of past outcomes are part of the linear predictor. Autoregressive generalized linear models of this kind and extensions are considered in Section 6.1.1, which covers conditional models for  $y_t$  given  $y_{t-1}, \dots, y_1$  and  $x_t$ . In certain applications, the marginal effect of covariates  $x_t$  on the response  $y_t$  is of primary interest, whereas the dependence of observations is regarded as a nuisance. Then it is more reasonable to base inference on the marginal distributions of  $y_t$  given  $x_t$  only, since in conditional models the influence of past observations may condition away covariate effects (Section 6.1.3). Brumback et al. (2000) discuss these different approaches in terms of a unifying framework.

### 6.1.1 Conditional Models

A broad class of non-normal time series models can be obtained by the following simple dynamization of generalized linear models or quasi-likelihood models for independent observations: As in autoregressive models for normal responses, conditional distributions or moments of  $y_t$  given the past can be defined by including past values  $y_{t-1}, \dots, y_{t-l}, \dots$  together with covariates.

#### Generalized Autoregressive Models

Consider first the case of univariate responses  $y_t$  and let

$$H_t = \{y_{t-1}, y_{t-2}, \dots, y_1, x_t, x_{t-1}, \dots, x_1\}$$

be the “history” of past observations and of present and past covariates. Generalized autoregressive models are characterized by the following structure:

- (i) The conditional densities  $f(y_t|H_t), t = 1, 2, \dots$  are of the exponential family type.
- (ii) The conditional expectation  $\mu_t = E(y_t|H_t)$  is of the form

$$\mu_t = h(z'_t \beta), \quad (6.1.1)$$

where  $h$  is a response function as in Chapter 2 and the  $p$ -dimensional design vector  $z_t$  is a function of  $H_t$ , i.e.,  $z_t = z_t(H_t)$ .

Assumptions (i) and (ii) imply that the conditional variance  $\sigma_t^2 = \text{var}(y_t|H_t)$  is given by

$$\sigma_t^2 = v(\mu_t)\phi,$$

where  $v(\cdot)$  is the variance function corresponding to the specific exponential family and  $\phi$  is the scale parameter. Thus, the definition is formally identical to that of GLMs for independent observations in Chapter 2; however, *conditional* rather than *marginal* distributions and moments are modelled.

In general, the design vector  $z_t$  is constructed from any lagged values of the response process and (or) any covariates that evolve in time simultaneously with the response process. Also, lagged values of covariates and any interactions are allowed.

Often, however, only a finite number of past observations,  $y_{t-1}, \dots, y_{t-l}$  say, will be included in the design vector. We will call this a *generalized autoregressive model* or *Markov model of order l*.

Let us consider some examples.

(i) *Binary and binomial time series.*

For a binary time series  $\{y_t\}, y_t \in \{0, 1\}$ , the conditional distribution of  $y_t$  given  $H_t$  is determined by

$$\pi_t = P(y_t = 1|H_t).$$

If no covariates are observed, then

$$\pi_t = h(\beta_0 + \beta_1 y_{t-1} + \dots + \beta_l y_{t-l}) = h(z'_t \beta), \quad t > l,$$

with  $z'_t = (1, y_{t-1}, \dots, y_{t-l}), \beta = (\beta_0, \dots, \beta_l)'$ , defines a purely autoregressive model of order  $l$ , with variance  $\sigma_t^2 = \pi_t(1 - \pi_t)$  and  $\phi = 1$ . Using the logistic response function, this is the Markov chain of order  $l$  suggested by Cox (1970). Additional interaction terms such as  $y_{t-1}y_{t-2}$ , etc., may be included. Incorporation of all possible interactions up to order  $l$  yields a saturated model, containing the same number of parameters as the general homogeneous Markov chain.

Inclusion of covariates, e.g., by

$$\begin{aligned}\pi_t &= h(\beta_0 + \beta_1 y_{t-1} + \dots + \beta_l y_{t-l} + x_t' \gamma), \quad t > l, \\ z_t' &= (1, y_{t-1}, \dots, y_{t-l}, x_t'), \quad \beta' = (\beta_0, \dots, \beta_l, \gamma'),\end{aligned}$$

allows the modelling of *nonstationary Markov chains*. Interaction terms of past observations and covariates, e.g., as in

$$\pi_t = h(\beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + \beta_3 y_{t-1} x_t), \quad (6.1.2)$$

can be useful or even necessary. Model (6.1.2) is equivalent to the following parameterization of nonhomogeneous transition probabilities:

$$\begin{aligned}\pi_{i1} &= P(y_t = 1 | y_{t-1} = i, x_t) = h(\alpha_{0i} + \alpha_{1i} x_t), \quad i = 0, 1, \\ \pi_{i0} &= P(y_t = 0 | y_{t-1} = i, x_t) = 1 - \pi_{i1}.\end{aligned}$$

Setting  $y_{t-1} = 0$  or  $1$  in (6.1.2) allows us to easily see that

$$\alpha_{00} = \beta_0, \alpha_{01} = \beta_0 + \beta_1, \alpha_{10} = \beta_2, \alpha_{11} = \beta_2 + \beta_3.$$

Thus, *nonhomogeneous models for transition probabilities*, which some authors have suggested, see, e.g., Garber (1989), are contained in the general model  $\pi_t = h(z_t' \beta)$  by appropriate inclusion of interaction terms.

Finally, the Markovian property may be dropped completely, as for  $z_t = (1, \bar{y}_{t-1})'$ , with the arithmetic or some weighted mean  $\bar{y}_{t-1}$  of past observations.

Extension to the case where  $\{y_t\}$  is a time series of relative frequencies assumed to be (scaled) binomial with repetition number  $n_t$  and conditional expectation  $\mu_t = E(y_t | H_t)$  is straightforward: Now  $\mu_t = h(z_t' \beta)$  with analogous models as earlier, but  $\sigma_t^2 = \mu_t(1 - \mu_t)/n_t$ .

## (ii) Count data

For counts, log-linear Poisson models for  $y_t | H_t$  are reasonable specifications. In complete analogy to the preceding binary models, one may assume

$$\lambda_t = E(y_t | H_t) = \exp(\beta_0 + \beta_1 y_{t-1} + \dots + \beta_l y_{t-l} + x_t' \gamma).$$

For  $\gamma = 0$ , such purely autoregressive models have been considered by Wong (1986). Although this model seems sensible, certain limitations are implied. Consider the case  $l = 1$  for simplicity.  $\lambda_t$  grows exponentially for  $\beta_1 > 0$ , while  $\beta_1 < 0$  corresponds to a stationary process for  $\gamma = 0$ . Therefore,

other specifications may sometimes be more appropriate. For example,  $\lambda_t = \exp(\beta_0 + \beta_1 \log y_{t-1})$  is equivalent to

$$\lambda_t = \lambda(y_{t-1})^{\beta_1}, \quad \lambda = \exp(\beta_0).$$

For  $\beta_1 = 0$ , the rate is constant,  $\lambda_t = \lambda$ . For  $\beta_1 > 0$ ,  $\lambda_t$  is increased by the previous outcome, while for  $\beta_1 < 0$  it is decreased. For  $y_{t-1} = 0$ ,  $\lambda_t = 0$ , i.e.,  $y_{t-1} = 0$  is absorbing. Models of this type, but more general, have been considered by Zeger & Qaqish (1988); see also below.

Also, for cross-sectional data, other distributions for  $y_t|H_t$  allow for deviations from the Poisson model. Fokianos (2000) suggests (doubly) truncated Poisson models. They are still within the exponential family framework, but other models like zero inflated Poisson models might be considered as well.

### (iii) Categorical time series

Generalized autoregressive models are easily extended to multivariate responses. For multicategorical time series  $\{y_t\}$ , where  $y_t$  is observed in  $k$  categories, let the response be coded by  $y_t = (y_{t1}, \dots, y_{tq})'$ ,  $q = k - 1$ ,

$$y_{tj} = \begin{cases} 1 & \text{if category } j \text{ has been observed,} \\ 0 & \text{otherwise,} \end{cases} \quad j = 1, \dots, q.$$

Correspondingly,  $\pi_t = (\pi_{t1}, \dots, \pi_{tq})$  is the vector of conditional probabilities

$$\pi_{tj} = P(y_{tj} = 1|H_t), \quad j = 1, \dots, q.$$

The general *autoregressive model for categorical time series* (e.g., Fahrmeir & Kaufmann, 1987; Kaufmann, 1987) now becomes

$$\pi_t = h(Z_t \beta), \quad t > l,$$

where  $h$  is a  $q$ -dimensional response function as in Chapter 3, but the design matrix  $Z_t = Z_t(H_t)$  is now a function of past observations and past and present covariates. In this way all response models for unordered and ordered categories are easily adapted to the time series situation, admitting a flexible and parsimonious treatment of higher-order dependence and, to some extent, of nonstationarity. With the various possibilities for  $h$  and  $Z_t$ , a wide variety of models can be covered. Choosing, e.g., a multinomial logit model with  $Z_t = \text{diag}(z'_t)$  and

$$z'_t = (1, y_{t-1,1}, \dots, y_{t-1,q}, \dots, y_{t-l,1}, \dots, y_{t-l,q}, x'_t),$$

one obtains a nonhomogeneous Markov chain of order  $l$ , with  $k$  unordered states. Interactions may be included, as in the binary case.

## Quasi-Likelihood Models and Generalized Autoregression Moving Average Models

Generalized autoregressive models are genuine likelihood models in the sense that the (conditional) distribution of  $y_1, \dots, y_t$ , given the covariates  $x_1, \dots, x_t$ , is completely determined by specification of the exponential family type and the mean structure (6.1.1), thereby implying a certain variance structure. As in the case of independent observations, one may wish to separate the mean and variance structures. Also, it may not always be sensible to treat covariate effects and effects of past observations symmetrically, as in the simple mean structure model (6.1.1). Zeger & Qaqish (1988) consider the following class of quasi-likelihood Markov models: The conditional mean structure is assumed to be

$$\mu_t = h(x'_t \gamma + \sum_{i=1}^l \beta_i f_i(H_t)), \quad (6.1.3)$$

where  $h$  is a response function, and the functions  $f_i$  are functions of past responses  $y_{t-i}$  and, possibly, past linear combinations  $x'_{t-i} \gamma$ . The conditional variance is assumed to be

$$\text{var}(y_t | H_t) = v(\mu_t) \phi, \quad (6.1.4)$$

where  $v$  is a variance function and  $\phi$  is an unknown dispersion parameter.

Some examples, not contained in the more restricted framework of generalized autoregressive models, are as follows.

### (iv) Autoregressive conditionally heteroscedastic (ARCH) models for Gaussian responses

Linear autoregressive models with covariates are defined by

$$\mu_t = \beta y_{t-1} + x'_t \gamma, \quad v(\mu_t) = 1,$$

setting  $l = 1$  for simplicity. Alternatively, for

$$\mu_t = \beta(y_{t-1} - x'_{t-1} \gamma), \quad v(\mu_t) = \mu_t^2 / \beta = \beta(y_{t-1} - x'_{t-1} \gamma)^2,$$

we get an ARCH model (e.g., Engle, 1982). Such models can account for time series with periods of increased variability.

### (v) Counts

As an alternative to Example (ii), consider a log-linear Poisson model with

$$\lambda_t = \exp\{x'_t \gamma + \beta[\log(y^*_{t-1}) - x'_{t-1} \gamma]\}$$

and  $\text{var}(y_t|y_{t-1}, x_t) = \lambda_t \phi$ , where  $y_{t-1}^* = \max(y_{t-1}, c)$ ,  $0 < c < 1$  (Zeger & Qaqish, 1988). This is equivalent to

$$\lambda_t = \exp(x'_t \gamma) \left[ \frac{y_{t-1}^*}{\exp(x'_{t-1} \gamma)} \right]^\beta. \quad (6.1.5)$$

Compared to the common log-linear model, the rate at time  $t$  is modified by the ratio of the past response and of  $\exp(x'_{t-1} \gamma)$ . Positive (negative) values of  $\beta$  correspond to positive (negative) autocorrelation. The parameter  $c$  determines the probability that  $y_t > 0$  in the case of  $y_{t-1} = 0$ , so that  $y_{t-1} = 0$  is no longer an absorbing state.

(vi) *Conditional gamma models*

In analogy to (iv) and (v), Zeger & Qaqish (1988) propose a model with canonical link

$$\mu_t^{-1} = x'_t \gamma + \sum_{i=1}^l \beta_i \left( \frac{1}{y_{t-i}} - x'_{t-i} \gamma \right)$$

and  $\sigma_t^2 = \mu_t^2 \phi$ . Thus, the linear predictor  $x'_t \gamma$  at time  $t$  is modified by a weighted sum of past “residuals”  $1/y_{t-i} - x'_{t-i} \gamma$ . They illustrate a simplified version of this model, where  $x'_t \gamma = \gamma$ , with an analysis of interspike times collected from neurons in the motor cortex of a monkey.

(vii) *Generalized autoregressive moving average (GARMA) models*

Benjamin, Rigby & Stasinopoulos (2000) extend the well-known Gaussian ARMA process to exponential family observations  $y_t|H_t$ . They postulate a mean  $E(y_t|H_t) = \mu_t = h(\eta_t)$  with predictor

$$\eta_t = x'_t \gamma + \sum_{j=1}^p \phi_j \{g(y_{t-j}) - x'_{t-j} \gamma\} + \sum_{j=1}^q \theta_j \{g(y_{t-j}) - \eta_{t-j}\},$$

where  $g = h^{-1}$  is the link function. For certain functions  $g$  it may again be necessary to replace  $y_{t-j}$  with some modification  $y_{t-j}^*$  as in (v) above. The first sum represents the autoregressive terms and the second sum the moving average terms.

For Gaussian  $y_t|H_t$  and the identity link  $\mu_t = \eta_t$ , the general GARMA model reduces to the Gaussian ARMA( $p, q$ ) model. A log-linear Poisson GARMA model is

$$\log(\mu_t) = x'_t \gamma + \sum_{j=1}^p \phi_j \{\log(y_{t-j}^* - x'_{t-j} \gamma)\} + \sum_{j=1}^q \theta_j \{\log(y_{t-j}^*/\mu_{t-j})\},$$

where  $y_{t-j}^* = \max(y_{t-j}, c)$  and  $\theta < c < 1$ . For  $\theta_j = 0$   $j = 1, \dots, q$ , this reduces to the autoregressive model (6.1.5). Similarly, the conditional gamma model in (vi) is extended to a gamma GARMA process and (iv) to a GARMA-GARCH process.

The additional flexibility of models of this type does not come for free. Generally, they cannot be reformulated in terms of the mean structure (6.1.1), which allows application of the usual fitting procedures of generalized linear models; see Section 6.1.2. However, keeping either  $\beta$  or  $\gamma$  fixed, examples (v) and (vi) fit into the usual framework. Therefore, a second level of iteration will be necessary for simultaneously estimating  $\beta$  and  $\gamma$ ; see Section 6.1.2. For GARMA models, Benjamin, Rigby & Stasinopoulos (2000) developed an iteratively dependent variable regression.

In the models considered so far, it was assumed that the conditional densities  $f(y_t|H_t)$ , or at least the conditional first and second moments, are correctly specified. In the case of quasi-likelihood models for independent observations (Section 2.3.1), we have seen that consistent parameter estimation, however, with some loss of efficiency, is still possible if the mean is correctly specified, whereas the true variance function is replaced by some “working” variance. One may ask if similar results hold for the time series situation. This is indeed the case. For generalized autoregressive models, conditioning in  $f(y_t|H_t)$ ,  $E(y_t|H_t)$ , etc., is on the complete “history”  $H_t$ . If one works with some Markov model of order  $l$ , it has to be *assumed* that

$$f(y_t|H_t) = f(y_t|y_{t-1}, \dots, y_{t-l}, x_t).$$

If this Markov property holds, then estimation can be based on genuine likelihoods. If this is not the case, or if the correct  $f(y_t|H_t)$  cannot be determined, we may *deliberately* condition only on the  $l$  most recent observations  $y_{t-1}, \dots, y_{t-l}$ . If  $f(y_t|y_{t-1}, \dots, y_{t-l}, x_t)$  is correctly specified, e.g., by one of the autoregressive models of order  $l$  discussed earlier, then consistent quasi-likelihood estimation is still possible. Compared to the case of independent observations, however, the correction for the asymptotic covariance matrix of the estimators becomes more difficult, and asymptotic theory requires more stringent assumptions; see Section 6.1.2.

One may even go a step further and assume only a correctly specified conditional mean  $E(y_t|y_{t-1}, \dots, y_{t-l}, x_t)$  together with some appropriate variance and autocorrelation structure. The extreme case, where conditioning is only on  $x_t$  and not on past responses, leads to marginal models in Section 6.1.3.

Some final remarks concern extensions to nonlinear and nonexponential family time series models in analogy to corresponding models for independent observations mentioned in Section 2.3.3. Nonlinear and robust models for

metrical outcomes have been considered and discussed by various authors. Several approaches to discrete outcomes, not covered by the conditional models here, may be cast into the more general nonlinear-nonexponential framework. In their development of D(iscrete) ARMA processes, Jacobs & Lewis (1983) were guided by the autocorrelation structure of ARMA processes for continuous variables. Guan & Yuan (1991) and Ronning & Jung (1992) discuss estimation of integer-valued AR models with Poisson marginals. In other approaches, binary time series  $\{y_t\}$  are assumed to be generated by truncation of a latent series  $\{\tilde{y}_t\}$  of continuous outcomes (e.g., Gourieroux, Monfort & Trognon, 1983b; Grether & Maddala, 1982). Pruscha (1993) proposes a categorical time series model that combines the autoregressive model (iii) with a recursion for the probability vector  $\pi_t$ . In the most general model (Heckman, 1981, for longitudinal data), the latent variable  $\tilde{y}_t$  is a linear combination of covariates, past values of  $\{y_t\}$  and  $\{\tilde{y}_t\}$ , and an error term. However, fitting such models requires considerable computational effort and no asymptotic theory is available for this general case.

### 6.1.2 Statistical Inference for Conditional Models

Estimation and testing for generalized autoregressive models can be based on genuine likelihood models. In the case of deterministic covariates, the joint density of the observations  $y_1, \dots, y_T$  factorizes into a product of conditional densities,

$$f(y_1, \dots, y_T | \beta) = \prod_{t=1}^T f(y_t | y_{t-1}, \dots, y_1; \beta),$$

and the conditional densities are determined by assuming some specific generalized autoregressive model. If covariates are stochastic, the joint density factorizes into

$$f(y_1, \dots, y_T, x_1, \dots, x_T | \beta) = \prod_{t=1}^T f(y_t | H_t; \beta) \prod_{t=1}^T f(x_t | C_t),$$

where  $H_t = (y_1, \dots, y_{t-1}, x_1, \dots, x_t)$  and  $C_t = (x_1, \dots, x_{t-1})$ . Assuming that the second product does not depend on  $\beta$ , estimation can be based on the (partial) likelihood defined by the first product. In any case, the log-likelihood of  $y_1, \dots, y_T$  is given by

$$l(\beta) = \sum_{t=1}^T l_t(\beta), \quad l_t(\beta) = \log f(y_t | H_t; \beta),$$

where the conditional densities are given by the definition of generalized autoregressive models. For Markov models of order  $l$ , this is, strictly speaking, a conditional log-likelihood, given the starting values  $y_0, \dots, y_{-l+1}$ . The first derivative of  $l(\beta)$ , the *score function*, is

$$s(\beta) = \sum_{t=1}^T Z_t' D_t(\beta) \Sigma_t^{-1}(\beta) (y_t - \mu_t(\beta)), \quad (6.1.6)$$

where  $\mu_t(\beta) = h(Z_t(\beta))$  is the *conditional* expectation,  $\Sigma_t(\beta) = \text{cov}(y_t | H_t)$  the *conditional* covariance matrix, and  $D_t(\beta) = \partial h / \partial \eta$  evaluated at  $\eta_t = Z_t \beta$ . Note that (6.1.6) is written in multivariate notation to include categorical time series. For univariate responses,  $Z_t$  reduces to  $z'_t$ , and  $\Sigma_t(\beta)$  to the conditional variance  $\sigma_t^2$ . Comparing these with corresponding expressions for independent observations in Chapters 2 and 3, we see that they are formally identical if past responses are treated like additional covariates. Of course, observed information matrices  $F_{obs}(\beta) = -\partial^2 l(\beta) / \partial \beta \partial \beta'$  are also of the same form. However, it is generally not possible to write down the *unconditional* expected information matrix  $F(\beta) = EF_{obs}(\beta)$  in explicit form. Instead, for dependent observations, the *conditional information matrix*

$$G(\beta) = \sum_{t=1}^T \text{cov}(s_t(\beta) | H_t),$$

where  $s_t(\beta) = \partial \log f(y_t | H_t; \beta) / \partial \beta = Z_t' D_t(\beta) \Sigma_t^{-1}(\beta) (y_t - \mu_t(\beta))$  is the individual score function contribution, plays an important role in computations and asymptotic considerations. In our context it is given by

$$G(\beta) = \sum_{t=1}^T Z_t' D_t(\beta_t) \Sigma_t^{-1} D_t'(\beta) Z_t \quad (6.1.7)$$

and has the same form as the expected information matrix for independent observations. Integrating out the observations  $\{y_t\}$ , we get, in principle, the unconditional expected information  $F(\beta) = \text{cov } s(\beta)$ .

To compute an MLE  $\hat{\beta}$  as a root of  $s(\beta)$  corresponding to a (local) maximum of  $l(\beta)$ , we can use the same iterative algorithms as for independent observations, treating past responses just like additional covariates. For the Fisher-scoring algorithm or its equivalent iterative weighted least-squares procedure, this means that conditional information matrices (6.1.7) are used instead of unconditional ones.

Under appropriate “regularity assumptions,” the MLE  $\hat{\beta}$  is consistent and asymptotically normal,

$$\hat{\beta} \stackrel{a}{\sim} N(\beta, G^{-1}(\hat{\beta})),$$

with the inverse of the conditional information matrix the asymptotic covariance matrix. Since dependent observations contain less information than independent observations, such regularity assumptions are generally somewhat more restrictive. Often stationarity or ergodicity conditions are assumed, allowing the application of limit theorems for stationary sequences of random variables. Certain forms of nonstationarity can be admitted; however, additional mathematical effort has to be invested. Kaufmann (1987) provides rigorous asymptotic estimation theory for Markov models with categorical responses; see also Fahrmeir & Kaufmann (1987). (As a key tool, martingale limit theory (e.g., Hall & Heyde, 1980) can be applied to the sequence  $s_t(\beta) = \partial l_t(\beta)/\partial\beta$  of individual conditional score function contributions, since they form a martingale sequence.) To illustrate what kind of nonstationarity can be admitted, consider an autoregressive logit or probit model of the form (6.1.1) and design vector  $z'_t = (1, y_{t-1}, \dots, y_{t-l}, x'_t)$  without interaction terms between past responses and covariates. Then the following two conditions are sufficient for consistency and asymptotic normality of the MLE (corollary 1 in Fahrmeir & Kaufmann, 1987):

- (i) the covariate sequence  $\{x_t\}$  is bounded;
- (ii) the empirical covariance matrix

$$S_t = \sum_{s=1}^t (x_s - \bar{x})(x_s - \bar{x})'$$

of the regressors diverges, i.e.,  $\lambda_{\min} S_t \rightarrow \infty$  or, equivalently,  $S_t^{-1} \rightarrow 0$ .

This condition corresponds exactly to (2.2.10) of Chapter 2 for independent observations, after removal of the constant 1. No convergence assumptions on  $S_t$  such as  $S_t/t \rightarrow S$  are required. Although no rigorous proof has been given, we conjecture that growing regressors are admissible with similar growth rates as for the case of independent observations.

Testing linear hypotheses by likelihood ratio, Wald statistics, and score statistics is possible in complete analogy to Section 2.2.2 if unconditional information matrices are replaced by conditional ones. The common asymptotic properties of test statistics remain valid under essentially the same general conditions required for consistency and asymptotic normality of the MLE (e.g., Fahrmeir, 1987a, 1988). All other tools for statistical inference that rely on the MLE and test statistics such as goodness-of-fit statistics and variable selection methods may be used in the same way as in Chapters 2, 3, and 4.

For categorical time series, Fokianos & Kedem (1998) extend these large sample results by considering stochastic time-dependent covariates and by

dropping any Markovian assumption. They use the concept of partial likelihood which simplifies conditional inference and obviates the Markov assumptions.

For the more general quasi-likelihood models defined by (6.1.3) and (6.1.4), the predictor

$$\eta_t = x'_t \gamma + \sum_{i=1}^l \beta_i f_i(H_t)$$

is no longer linear in  $\gamma$  and  $\beta$  if  $f_i(H_t)$  depends on  $\gamma$ . The MLE for  $(\beta, \gamma)$  may in principle be computed by the method of Fisher scoring; however, score functions and conditional information matrices are not of the simple form (6.1.6), (6.1.7) and have to be redefined by evaluating  $\partial \mu_t / \partial \gamma$  and  $\partial \mu_t / \partial \beta$ . A similar fitting procedure is suggested by Benjamin, Rigby & Stasinopoulos (2000) for ML estimation of GARMA models. If  $f_i(H_t)$  is linear in  $\gamma$  as in Examples (v) and (vi), a two-step iteration procedure alternating between estimates for  $\gamma$  and  $\beta$  may be easier to implement; see Zeger & Qaqish (1988, p. 1024). General asymptotic theory can be applied and the (quasi-) MLE will possess the usual asymptotic properties under appropriate regularity conditions.

Finally, let us briefly discuss the situation considered at the end of Section 6.1.1, where modelling and estimation are based on a Markov model of order  $l$  and the resulting quasi-log-likelihood

$$l(\beta) = \sum_{t=1}^T l_t(\beta) = \sum_{t=1}^T \log f(y_t | y_{t-1}, \dots, y_{t-l}, x_t; \beta).$$

Since  $f(y_t | H_t) \neq f(y_t | y_{t-1}, \dots, y_1, x_t; \beta)$ , this is not the true log-likelihood of the data. Yet  $\beta$  can be estimated consistently and asymptotically normal, but the asymptotic covariance matrix has to be corrected. Due to the dependence of the observations, this correction is not as simple as in the case of independent observations. For stationary processes and under suitable mixing assumptions, Azzalini (1983), Levine (1983), and White (1984) give relevant results. However, it seems that one cannot move too far away from stationarity.

### **Example 6.1: Polio incidence in the United States**

Table 6.1 lists a time series of the monthly number  $y_t$  of cases of poliomyelitis reported by the U.S. Centers for Disease Control for the years 1970 to 1983,  $t = 0, \dots, 167$ , taken from Zeger (1988b).

Let us analyze whether polio incidence has been decreasing since 1970. A plot of the time series given in Figure 6.1 reveals some seasonality but does not provide clear evidence for a long-term decrease in the rate of U.S.

polio infection. Thus, we will regress  $y_t$  on a linear trend, as well as sine, cosine pairs at the annual and semiannual frequencies. Since the counts  $y_t$  and  $y_{t+\tau}$ ,  $\tau > 0$ , are not independent, we also take into account the effect of past polio counts by conditioning on these. As a basis we use a conditional log-linear Poisson model of the form

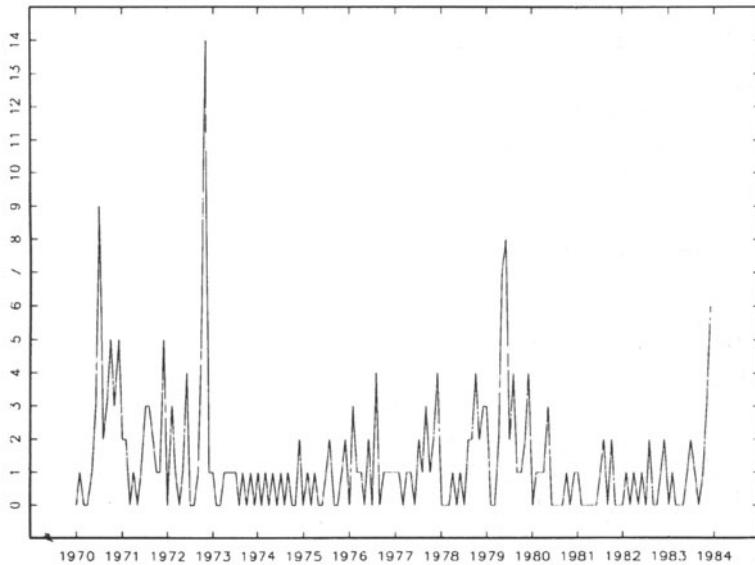
$$\lambda_t = E(y_t | y_{t-1}, \dots, y_{t-l}) = \exp(\alpha + \beta t \times 10^{-3} + z'_t \delta + \sum_{j=1}^l \gamma_j y_{t-j}),$$

where the term  $z'_t$  including  $\cos(2\pi t/12)$ ,  $\sin(2\pi t/12)$ ,  $\cos(2\pi t/6)$ ,  $\sin(2\pi t/6)$  represents the seasonal part. For the autoregressive model of order  $l = 5$  the fitted incidence rate  $\hat{\mu}_t$  can be seen in Figure 6.2. MLEs on which the fitted incidence rate  $\hat{\mu}_t$  is based are given in Table 6.2 together with  $p$ -values that already take into account the estimated nuisance parameter  $\hat{\phi} = 1.761$ . A long-term decrease in polio incidence is indicated by the fitted curve in Figure 6.2 as well as by the negative sign of the MLE for the trend component. This is also confirmed by the Wald test of the hypothesis  $H_0: \beta = 0$ , which cannot clearly be rejected due to the  $p$ -value 0.095.

Concerning time dependence, the autoregressive order  $l = 5$  should not be reduced further since the  $p$ -value of  $y_{t-5}$  indicates a strong time dependence. The effect of successive dropping of autoregressive terms on the

**Table 6.1.** Monthly number of poliomyelitis cases in the United States for 1970 to 1983

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1970	0	1	0	0	1	3	9	2	3	5	3	5
1971	2	2	0	1	0	1	3	3	2	1	1	5
1972	0	3	1	0	1	4	0	0	1	6	14	1
1973	1	0	0	1	1	1	1	0	1	0	1	0
1974	1	0	1	0	1	0	1	0	1	0	0	2
1975	0	1	0	1	0	0	1	2	0	0	1	2
1976	0	3	1	1	0	2	0	4	0	1	1	1
1977	1	1	0	1	1	0	2	1	3	1	2	4
1978	0	0	0	1	0	1	0	2	2	4	2	3
1979	3	0	0	3	7	8	2	4	1	1	2	4
1980	0	1	1	1	3	0	0	0	0	1	0	1
1981	1	0	0	0	0	0	1	2	0	2	0	0
1982	0	1	0	1	0	1	0	2	0	0	1	2
1983	0	1	0	0	0	1	2	1	0	1	3	6

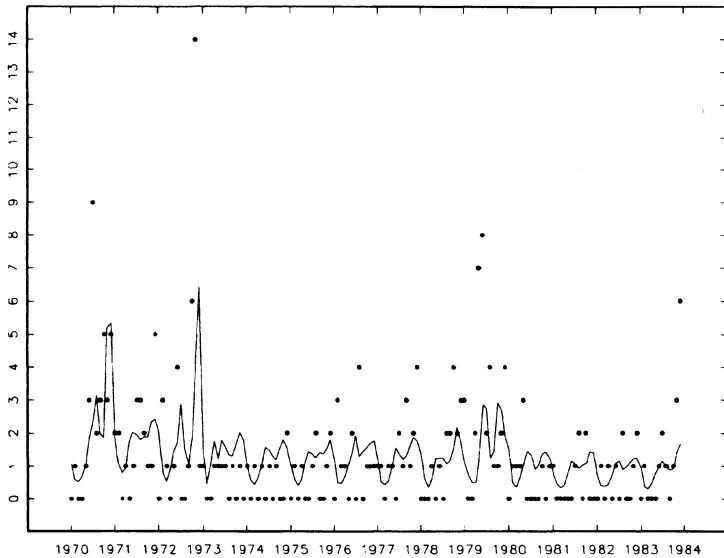


**Figure 6.1.** Monthly number of polio cases in the United States from 1970 to 1983.

**Table 6.2.** Log-linear AR(5) model fit to polio data

	MLE	p-value
1	0.160	0.523
$t \times 10^{-3}$	-3.332	0.095
$\cos(2\pi t/12)$	-0.217	0.116
$\sin(2\pi t/12)$	-0.462	0.002
$\cos(2\pi t/6)$	0.128	0.354
$\sin(2\pi t/6)$	-0.372	0.008
$y_{t-1}$	0.085	0.022
$y_{t-2}$	0.040	0.445
$y_{t-3}$	-0.040	0.501
$y_{t-4}$	0.029	0.532
$y_{t-5}$	0.081	0.059

intercept  $\alpha$  and trend component  $\beta$  is documented in Table 6.3. Obviously *different* autoregressive models yield *different* trend component estimates. Such a phenomenon is typical for conditional models, which usually are not robust against different forms of time-dependence specification. Moreover, interpretation of the trend components changes from one autoregressive



**Figure 6.2.** Predicted polio incidence  $\hat{\mu}_t$  based on a log-linear AR( $l = 5$ )-model fit.

model to the other, since the conditioning on past observations changes. Effects of seasonal components behave quite similarly. Therefore, marginal modelling as in the next section may be more sensible if one is interested in the marginal effect of trend and season.

Benjamin, Rigby & Stasinopoulos (2000) fit negative binomial and Poisson GARMA models to the data, reporting a significantly better fit for the negative binomial model and concluding that there is no significant evidence of a decreasing trend in polio infection. Fokianos (2000) improves the fit by applying a truncated Poisson model. Again effects change with different models.  $\square$

### 6.1.3 Marginal Models

In a number of applications, the main scientific objective is not to fit a conditional or predictive model for  $y_t$  but to express  $y_t$  or its expectation as a simple function of  $x_t$ . More formally, this means one is rather interested in the marginal distribution  $f(y_t|x_t)$  or the marginal expectation  $E(y_t|x_t)$  instead of some conditional distribution  $f(y_t|y_{t-1}, \dots, x_t)$  or conditional expectation. This is in complete analogy to Section 3.5.2, where conditional and marginal models for correlated responses have been contrasted within a cross-sectional context.

For illustration, let us first consider two simple Gaussian models. The conditional model

**Table 6.3.** Log-linear AR(1) model fits to polio data

AR-order	1	$t \times 10^{-3}$
$l = 1$	0.352 (0.193)	-3.934 (1.914)
$l = 2$	0.297 (0.218)	-3.715 (1.955)
$l = 3$	0.335 (0.225)	-3.882 (1.983)
$l = 4$	0.278 (0.242)	-3.718 (2.008)
$l = 5$	0.160 (0.251)	-3.332 (1.997)

MLEs and standard errors (in brackets) of intercept and trend component

$$y_t = x_t\beta + \gamma y_{t-1} + u_t, \quad u_t \sim N(0, \sigma^2),$$

has conditional expectation  $x_t\beta + \gamma y_{t-1}$ , whereas its marginal expectation is (for  $|\gamma| < 1$ )

$$E y_t = \left( \sum_{s=0}^{t-1} \gamma^s x_{t-s} \right) \beta,$$

which is not easy to interpret. As an alternative, consider the model

$$\tilde{y}_t = x_t\beta + \varepsilon_t, \quad \varepsilon_t = \gamma\varepsilon_{t-1} + u_t. \quad (6.1.8)$$

Now the marginal expectation is

$$E \tilde{y}_t = x_t\beta,$$

which is quite easily interpreted.

Generally, the basic idea is to model marginal and not conditional expectations and to supplement the marginal mean structure by a variance-covariance structure. This leads to the following specification of marginal models for univariate generalized time series models (Zeger, Diggle & Yasui, 1990), which is analogous to the definition in Section 3.5. Assume

- (i) a *marginal mean structure*

$$E(y_t|x_t) = \mu_t = h(z_t'\beta), \quad z_t = z_t(x_t),$$

as for independent observations, together with

(ii) a *variance function*

$$\text{var}(y_t|x_t) = v(\mu_t)\phi,$$

(iii) and an (*auto-*) *covariance function*

$$\text{cov}(y_t, y_{t+r}) = c(\mu_t, \mu_{t+r}; \alpha),$$

where  $\alpha$  is a further parameter.

The most crucial point is the specification of the covariance function. However, consistent estimation by a generalized estimating approach is still possible under regularity assumptions if  $c(\mu_t, \mu_{t+r}; \alpha)$  is only a “working,” i.e., possibly incorrect, covariance.

As an example we first consider a regression model for time series of counts (Zeger, 1988b). It is a log-linear model specified in analogy to the linear Gaussian model (6.1.8). Conditional on an unobserved latent process  $\varepsilon_t$ , the counts are assumed to be independent with conditional mean and variance

$$E(y_t|\varepsilon_t) = \text{var}(y_t|\varepsilon_t) = \exp(x'_t\beta)\varepsilon_t. \quad (6.1.9)$$

Furthermore, suppose that  $\{\varepsilon_t\}$  is a stationary process with  $E(\varepsilon_t) = 1$ , for simplicity, and  $\text{cov}(\varepsilon_t, \varepsilon_{t+\tau}) = \sigma^2 \rho_\varepsilon(\tau)$ . Then the marginal moments of  $y_t$  are

$$\mu_t = E y_t = \exp(x'_t\beta), \quad \text{var}(y_t) = \mu_t + \sigma^2 \mu_t^2, \quad (6.1.10)$$

$$\text{cov}(y_t, y_{t+\tau}) = \mu_t \mu_{t+\tau} \sigma^2 \rho_\varepsilon(\tau). \quad (6.1.11)$$

If the autocorrelation function  $\rho_\varepsilon$  is fully specified by a further set of parameters  $\alpha$ , as, e.g.,  $\rho_\varepsilon(\tau) = \rho^\tau$  for a first-order autoregressive process  $\varepsilon_t$ , then the covariance function is in the form (iii) with  $\alpha = (\sigma^2, \rho)$ .

For binary time series, a marginal logit model

$$\pi_t = E(y_t|x_t) = x'_t\beta, \quad \sigma_t^2 = \text{var}(y_t|x_t) = \pi_t(1 - \pi_t),$$

and a specification of the odds ratio (OR)

$$OR(y_t, y_{t+\tau}) = \frac{P(y_t = y_{t+\tau} = 1)P(y_t = y_{t+\tau} = 0)}{P(y_t = 1, y_{t+\tau} = 0)P(y_t = 0, y_{t+\tau} = 1)}$$

are suggested. For a more parsimonious parameterization, “stationary” and “truncated” (log) odds ratios

$$\log OR(y_t, y_{t+\tau}) = \begin{cases} \gamma(\tau) & \text{if } \tau \leq h, \\ 0 & \text{if } \tau > h, \end{cases}$$

are useful as a “working” autocovariance structure.

### Estimation of Marginal Models

If covariance parameters  $\alpha$  or consistent estimates are known, regression parameters  $\beta$  can be estimated by adapting the generalized estimating equation approach in Sections 2.3.1 and 3.5.2 to the time series case. Setting

$$y = (y_1, \dots, y_T)', \quad Z = (z_1, \dots, z_T)', \quad \mu = (\mu_1, \dots, \mu_T)',$$

$$D(\beta) = \text{diag}(D_t(\beta)), \quad M(\beta) = Z'D(\beta), \quad \Sigma(\beta) = \text{cov}(y),$$

the quasi-score function (generalized estimating function) can be written in matrix form as

$$s(\beta) = Z'D(\beta)\Sigma^{-1}(\beta)(y - \mu).$$

With independent observations,  $\Sigma(\beta)$  is diagonal. With time series data,  $\Sigma(\beta)$  is nondiagonal with elements

$$(\Sigma(\beta))_{st} = \text{cov}(y_s, y_t) = c(\mu_s, \mu_t; \alpha).$$

A quasi-MLE  $\hat{\beta}$  is obtained as a root of the generalized estimating equation  $s(\beta) = 0$  and can, in principle, be computed by the method of Fisher scoring. However, this requires inversion of the  $(T \times T)$ -matrix  $\Sigma(\beta)$  in each iteration step. To simplify computations, it will be useful to approximate the actual covariance matrix by a simpler “working” covariance matrix  $\Sigma_w$ , e.g., a band diagonal matrix.

Under regularity assumptions that admit the application of limit theorems for dependent data, the quasi-MLE  $\hat{\beta}$  is consistent and asymptotically normal with asymptotic covariance matrix

$$\hat{A} = (Z'D(\hat{\beta})\Sigma^{-1}(\hat{\beta})D'(\hat{\beta})Z)^{-1}$$

if  $\Sigma(\beta)$  is *correctly specified*. If a working covariance matrix  $\Sigma_w(\beta)$  is used to simplify computations, then – in complete analogy to Section 3.5.2 – the covariance matrix has to be modified to the sandwich matrix

$$\hat{A} = (\hat{M} \hat{\Sigma}_w^{-1} \hat{M}')^{-1} \hat{M} \hat{\Sigma}_w^{-1} \hat{\Sigma} \hat{\Sigma}_w^{-1} \hat{M}' (\hat{M} \hat{\Sigma}_w^{-1} \hat{M}')^{-1},$$

where “ $\hat{\cdot}$ ” means evaluation at  $\hat{\beta}$  and  $\hat{M} = Z'D(\hat{\beta})$ . Of course, direct application of this result requires that the autocovariance structure  $\Sigma(\beta)$  is known. The asymptotic normality result remains true if the “nuisance” parameter  $\alpha$  is replaced by a consistent estimate  $\hat{\alpha}$ . Such an estimate can be obtained, e.g., by a method of moments. For example, in the log-linear model for count data,  $\text{var}(y_t) = \mu_t + \sigma^2 \mu_t^2$ . Hence,  $\sigma^2$  can be estimated by

$$\hat{\sigma}^2 = \sum_t \{(y_t - \hat{\mu}_t)^2 - \hat{\mu}_t\} / \sum_t \hat{\mu}_t^2;$$

similarly, the autocorrelation function can be estimated by

$$\hat{\rho}_\varepsilon(\tau) = \hat{\sigma}^{-2} \sum_{t=\tau+1}^T \{(y_t - \hat{\mu}_t)(y_{t-\tau} - \hat{\mu}_{t-\tau})\} / \sum_{t=\tau+1}^T \hat{\mu}_t \hat{\mu}_{t-\tau}.$$

If  $\rho_\varepsilon$  is itself parameterized by  $\rho_\varepsilon(\tau; \alpha)$ , then  $\hat{\rho}_\varepsilon(\tau, \alpha)$  is solved for  $\alpha$ . For example, if  $\varepsilon_t$  is assumed to be a stationary autoregressive process, then  $\hat{\alpha}$  can be obtained by solving the Yule-Walker equations.

As stated in Section 3.5.2, in some cases simple working covariances matrices will yield good estimates. However, the assumption of independence can also lead to a considerable loss of efficiency when the responses are strongly correlated and the design includes a within-cluster covariate (see Fitzmaurice, 1995).

**Example 6.2: Polio incidence in the United States** (Example 6.1, continued)

Zeger (1988b) applied a marginal log-linear Poisson model (6.1.9) – (6.1.11) to the polio incidence data of Table 6.1, including trend and seasonal components in the linear predictor as in Example 6.1 but omitting autoregressive terms. Estimates  $\hat{\sigma}^2, \hat{\rho}_\varepsilon$  were obtained as above and a tridiagonal “working” covariance matrix  $\Sigma_w$  was used in the Fisher-scoring steps. The results of the trend and seasonal components are given in Table 6.4.

Compared to conditional modelling in Example 6.1, signs of the parameter estimates remain mostly the same, but absolute values of marginal effects differ significantly from corresponding effects in the conditional autoregressive model of order 5 (Table 6.2). In particular, the trend term in Table 6.4 indicates a long-term decrease in the rate of polio infection more clearly. In contrast, this effect is attenuated with conditional models.  $\square$

**Table 6.4.** Marginal model fit for polio data

	Estimate	Std. error
$t \times 10^{-3}$	-4.35	2.68
$\cos(2\pi t/12)$	-0.11	0.16
$\sin(2\pi t/12)$	-0.48	0.17
$\cos(2\pi t/6)$	-0.20	0.14
$\sin(2\pi t/6)$	-0.41	0.14
$\hat{\sigma}^2$	0.77	
$\hat{\rho}(1)$	0.25	

## 6.2 Longitudinal Data

In this section we consider longitudinal data in discrete time, where time series data or repeated observations  $(y_{it}, x_{it}), t = 1, \dots, T_i$ , are available for each individual or unit  $i = 1, \dots, n$  of a population. Linear models for Gaussian outcomes  $y_{it}$  have been treated extensively for a long time; see, e.g., Hsiao (1986) and Dielman (1989). Corresponding models for discrete and non-normal outcomes have received attention only more recently, but the gap is rapidly closing. The text of Diggle, Liang & Zeger (1994) is an important contribution with a focus on biostatistical applications.

Longitudinal data may be viewed as data with correlated responses within clusters (Section 3.5): The cluster  $i$  contains repeated responses  $y_{it}$ ,  $t = 1, \dots, T_i$  on the same subject  $i$ ,  $i = 1, \dots, n$ , and it is very likely that these repeated responses are correlated. Therefore, methods for analyzing correlated responses will reappear in this section in modified form. For example, due to the longitudinal data situation, special covariance structures will be of interest. As in Section 3.5 and as for the pure time series situation in Section 6.1, it is important to distinguish between conditional and marginal models. For example, consider an epidemiologic study as in Example 1.10, where  $y_{it}$  is the health or illness state, measured in categories, of individual  $i$  at time  $t$  and  $x_{it}$  is a vector of risk factors or individual characteristics, possibly changing over time. If one is interested in analyzing conditional probabilities for a certain state or transitions to a state given the individual's history and the effect of covariates on these probabilities, then conditional models are needed. If the main scientific objective is the effect of covariates on the state of health, marginal models are appropriate. The distinction is also important from the methodological point of view: Most models for non-normal outcomes are nonlinear, so that conditional and marginal models will generally not be of the same form. For example, marginal probabilities calculated from a conditional logistic model will not be of the logistic form.

### 6.2.1 Conditional Models

Notation is simplified if observations at time  $t$  are collected in “panel waves”

$$(y_t, x_t) = (y_{1t}, \dots, y_{nt}; x_{1t}, \dots, x_{nt}), \quad t = 1, \dots, T,$$

where  $T = \max(T_i)$  is the maximal length of individual time series and  $y_{it}, x_{it}$  are omitted if  $t > T_i$ . The “history” of covariates and past responses at time  $t$  is then denoted by

$$H_t = \{x_t, \dots, x_1, y_{t-1}, \dots, y_1\}. \quad (6.2.1)$$

Initial values  $y_0, y_{-1}, \dots$ , etc., which are needed in autoregressive models, are assumed to be part of the covariates. Conditional models for time series (Section 6.1.1) are easily extended to panel data if we assume that individual responses  $y_{it}$  given  $H_t$  are conditionally independent:

$$f(y_t|H_t) = \prod_{i \in R_t} f(y_{it}|H_t), \quad t = 1, \dots, T, \quad (6.2.2)$$

where  $R_t$  is the set of units still observed at time  $t$ . This condition is clearly fulfilled if the individual  $n$  time series  $\{y_{it}, t = 1, \dots, T\}$  are totally independent. However, (6.2.2) is weaker since interaction among units may be introduced via the common history. Condition (6.2.2) may be weakened further by conditioning  $y_{it}$  additionally on  $y_{jt}, j \neq i$ , and similarly for conditional symmetric models for multivariate cross-sectional observations in Section 3.5.1; see Zeger & Liang (1989). However, estimation becomes more difficult.

#### Generalized Autoregressive Models, Quasi-Likelihood Models

The simplest models of generalized autoregressive form are obtained if it is assumed that parameters are constant across time and units and that conditional densities of  $y_{it}$  given  $H_t$  belong to a uni- or multivariate simple exponential family. Conditional means are supposed to be given by

$$\mu_{it} = E(y_{it}|H_t) = h(Z_{it}\beta), \quad (6.2.3)$$

where the design matrix is a function of  $H_t$ , i.e.,  $Z_{it} = Z_{it}(H_t)$ . As for cross-sectional and time series data, various specifications of the response function and the design matrix are possible, in particular for multicategorical responses. As an important subclass, Markov models are obtained if only a finite number of past responses is included in  $H_t$ .

Together with a specific exponential family the mean structure (6.2.3) completely specifies the variance function. Separation of the mean and variance structures is possible by adopting the extended models defined by (6.1.3) and (6.1.4) to the panel data situation.

The assumption that parameters  $\beta$  are homogeneous with respect to time and the population can be relaxed, in principle, in a number of ways. Whether this is sensible will also depend on the data available. For example, if we allow for individual-specific parameters  $\beta_i$ , (6.2.3) is replaced by

$$\mu_{it} = h(Z_{it}\beta_i), i = 1, \dots, n.$$

If individual time series are long enough, the  $\beta_i$ 's can be consistently estimated separately. However, if individual time series are short, this will lead to serious bias or even the nonexistence of estimates. To avoid such problems, other conceivable working hypotheses are homogeneity of parameters in subpopulations or homogeneity with respect to a part of the parameters as, e.g., in models where only the intercept term varies. An attractive alternative is random effects models (Chapter 7). Analogous remarks apply to parameters  $\beta_t$  varying over time. If the population is large enough, it may be possible to fit a model separately for each panel wave  $y_t$  as in Stram, Wei & Ware (1988). For smaller cross sections one will run into trouble due to bias or nonexistence of estimates, and some kind of smoothing as in Chapter 8 will be useful.

Though the class of conditional models considered so far seems to be quite large, it does not contain all approaches proposed previously in the literature. Heckman (1981) derives dynamic models for panel data with binary responses  $y_{it}$  by relating them to latent continuous responses  $\tilde{y}_{it}$  via a threshold mechanism:  $y_{it} = 1 \Leftrightarrow \tilde{y}_{it} \geq 0; y_{it} = 0$  otherwise. He assumes a linear autoregressive model  $\tilde{y}_{it} = \eta_{it} + \varepsilon_{it}$ , where  $\eta_{it}$  is a linear combination of covariates  $x_{it}$ , past responses, and past latent variables. The inclusion of  $\tilde{y}_{i,t-1}, \tilde{y}_{i,t-2}, \dots$  formalizes the idea that former latent propensities to choose a state influence the probability for the current choice. The conditional probability for  $y_{it}$  given the past cannot be expressed in the simple mean structure model above.

## Statistical Inference

Under the conditional independence assumption (6.2.2), the log-likelihood given all data factorizes into individual contributions. Therefore, log-likelihoods, score functions, and conditional and observed information matrices can be written as a sum of the individual contributions. For generalized autoregressive models with parameters  $\beta$  constant across time and units, we obtain the log-likelihood

$$l(\beta) = \sum_{i,t} l_{it}(\beta), \quad l_{it}(\beta) = \log f(y_{it}|H_t, \beta), \quad (6.2.4)$$

the score function

$$s(\beta) = \sum_{i,t} s_{it}(\beta) = \sum_{i,t} Z'_{it} D_{it}(\beta) \Sigma_{it}^{-1} (y_{it} - \mu_{it}(\beta)),$$

and the conditional information matrix

$$G(\beta) = \sum_{i,t} G_{it}(\beta) = \sum_{i,t} Z'_{it} D_{it}(\beta) \Sigma_{it}^{-1} D'_{it}(\beta) Z_{it} \quad (6.2.5)$$

with the usual notation for individual conditional means  $\mu_{it}$ , variances  $\Sigma_{it}$ , etc. The modifications necessary for parameters varying over time or units are obvious. MLEs are obtained by the usual iterative methods from  $s(\hat{\beta}) = 0$ . For quasi-likelihood models one starts directly from the appropriately modified quasi-score function.

For asymptotics, three main situations can be considered:

$$\begin{aligned} n \rightarrow \infty, & \quad t \text{ finite, } \beta \text{ constant across units;} \\ n \text{ finite,} & \quad t \rightarrow \infty, \beta \text{ constant across time;} \\ n \rightarrow \infty, & \quad t \rightarrow \infty. \end{aligned}$$

These possibilities correspond to longitudinal data with many short time series, a moderate number of longer time series, and many longer time series, respectively. The second situation can be treated similarly as for single time series. For  $n \rightarrow \infty$  and  $t \rightarrow \infty$ , it may be possible to admit some parameters varying over units and others varying over time, e.g., as for an additive intercept term  $\beta_{0it} = \beta_{0i} + \beta_{it}$ . Under appropriate regularity assumptions the (quasi-) MLE will be consistent and asymptotically normal; however, to our knowledge rigorous proofs are not available. Yet it seems reasonable to assume that

$$\hat{\beta} \xrightarrow{a} N(\beta, G^{-1}(\hat{\beta}))$$

and to base test statistics, goodness-of-fit tests, etc., on this working assumption. Anyway, finite sample behavior requires further investigation.

## Transition Models

For discrete outcomes  $y_{it}$  a main objective often is the modelling of transition probabilities between pairs of categories or “states” at successive occasions  $t - 1$  and  $t$ . If a first-order Markov assumption holds, transitions may be analyzed with the models and methods above by inclusion of  $y_{i,t-1}$  as an additional covariate. It can be preferable, however, to model transition probabilities

$$p_{jk}(x_{it}) = P(y_{it} = k | y_{i,t-1} = j, x_{it}; \beta), \quad t = 2, \dots, T,$$

separately, e.g., by a logistic model for all pairs  $(j, k)$  and to express the likelihood as the product of transition probabilities of all observed transitions. This direct approach, which has been used, e.g., by Garber (1989) and Hopper & Young (1989), is more convenient and flexible if one has to distinguish among different types of categories, such as transient, recurrent, or absorbing states.

If the first-order Markov property holds and transition probabilities are correctly specified, usual likelihood theory applies. One may ask, however, whether transition probabilities can be consistently estimated when the Markov property is violated. This is no problem if the population is large enough to estimate separate response models for each of the  $t - 1$  occasions; see, e.g., Ware, Lipsitz & Speizer (1988). Otherwise the question is analogous to the estimation of marginal models from time series (Section 6.1.3) or longitudinal data (Section 6.2.2). If transition probabilities are correctly specified, then results of White (1984, Theorem 3.1) show that consistent parameter estimation is possible for  $T \rightarrow \infty$  under regularity assumptions. Again little is known about finite sample behavior, in particular concerning covariance matrix estimation.

## Subject-specific Approaches and Conditional Likelihood

A binary logistic model that includes a subject-specific parameter is given by

$$P(y_{it} = 1 | x_{it}) = \frac{\exp(\alpha_i + x_{it}'\beta)}{1 + \exp(\alpha_i + x_{it}'\beta)}.$$

The model accounts for heterogeneity by introducing the parameter  $\alpha_i$ . Models of this type have been studied in the psychometrics literature in the context of item response theory. The simplest version is the Rasch model, where  $\alpha_i$  stands for the subject's ability and  $x_{it}'\beta$  simplifies to a parameter  $\gamma_t$  that stands for the item difficulty (see Rasch, 1961; Andersen, 1980).

The basic idea proposed by Andersen (1973) is to consider the conditional likelihood given  $y_{i \cdot} = \sum_{t=1}^T y_{it}$ , where it is assumed that the number of observations does not depend on  $i$ . From the form

$$P(y_{i1}, \dots, y_{iT} | y_{i \cdot}) = \frac{\exp\left(\sum_{t=1}^T y_{it} x'_{it} \beta\right)}{\sum_{\{\tilde{y}_{i1}, \dots, \tilde{y}_{iT} : \sum_j \tilde{y}_{ij} = y_{i \cdot}\}} \exp\left(\sum_{t=1}^T \tilde{y}_{it} x'_{it} \beta\right)},$$

it is seen that the conditional probability given  $y_{i \cdot}$  does not depend on the heterogeneity parameter  $\alpha_i$ , as  $y_{i \cdot}$  is a sufficient statistic for  $\alpha_i$ . Therefore, conditioning on  $y_{1 \cdot}, \dots, y_{n \cdot}$  allows conditional inference upon the parameter  $\beta$ . Originally developed for item response models, this conditioned likelihood approach has been considered more recently for categorical repeated measurement and panel data (Conaway, 1989; Hamerle & Ronning, 1992). An alternative approach to subject-specific models based on the connection of the Rasch model and quasi-symmetry has been investigated in a series of papers by Agresti (1993a, 1993b) and Agresti & Lang (1993).

### Example 6.3: IFO business test

The IFO Institute in Munich collects micro-data of firms each month for its “business tests.” The monthly questionnaire contains questions on the tendency of successive change of realizations, plans, and expectations of variables like production, orders in hand, demand, etc. Answers are categorical, most of them trichotomous with categories like “increase” (+), “decrease” (-), and “no change” (=). Currently several thousand firms from various industry branches participate in this survey on a voluntary basis. Table 6.5 contains some typical variables and questions.

One of the main objectives of previous research has been to study the dependency of certain variables, such as production plans and prices, on other economic variables and to check whether the results are in agreement with economic theory; see, e.g., König, Nerlove & Oudiz (1981) and Nerlove (1983). Based on this work, we will analyze the dependency of the variable  $P_t$  at month  $t$  on the explanatory variables  $D_t$  and  $O_t$  and on the production plans  $P_{t-1}$  in the previous month. In this section, we consider 317 firms of the machine-building industry from January 1980 to December 1990. Previous work of Morawitz & Tutz (1990) showed that simple cumulative models (Section 3.3.1) are not always appropriate for analyzing these data, while extended versions with explanatory variables as threshold variables (Section 3.3.2) provide reasonable fits. For work on business tendency surveys, see also Ronning (1980, 1987).

For the following each of the trichotomous ( $k = 3$ ) variables  $P$ ,  $D$  and  $O$  is described by two ( $q = 2$ ) dummy variables, with “decrease” (-) as the

**Table 6.5.** Variables and questions of the IFO business test

Variable	Questions
Production plan $P$	Our production with respect to product X is planned for the next three months (corrected for seasonal variations) to be increased, the same, decreased.
Expected business $D$	Our business conditions for product X are expected to be in the next six months (corrected for seasonal variation) improved, about the same, deteriorated.
Orders in hand $O$	Our orders in hand (domestic and foreign) for X are, in relation to the preceding month, higher, the same, lower.

(Since the variable “expected business condition” is considered as a substitute for “expected demand,” we have chosen the mnemonic  $D$ ).

reference category. The relevant dummy variables for “increase” (+) and “no change” (=) are shortened by  $P^+$ ,  $P^-$ ,  $D^+$ ,  $D^-$ ,  $O^+$ , and  $O^-$ . Then an extended cumulative model of the form (3.3.11), from Chapter 3, where all covariates are threshold variables, is specified by

$$\begin{aligned} P(P_t = "+") &= F(\beta_{10} + w'\beta_1), \\ P(P_t = "+" \text{ or } "=") &= F(\beta_{20} + w'\beta_2), \end{aligned}$$

where  $P(P_t = "+")$  and  $P(P_t = "+" \text{ or } P_t = "=")$  stand for the probability of increasing and nondecreasing production plans, and  $F$  is the logistic distribution function. The parameters  $(\beta_{10}, \beta_1)$  and  $(\beta_{20}, \beta_2)$  are category-specific, and the threshold design vector  $w$  contains all six main effects and interaction effects if necessary:

$$w = (P_{t-1}^+, P_{t-1}^-, D_t^+, D_t^-, O_t^+, O_t^-, P_{t-1}^+ * D_t^+, \dots).$$

For the data at hand it turned out that the model without three-factor interactions but including all two-factor interactions fits the data reasonably well, while all smaller models are rejected by the Pearson or deviance statistic. Table 6.6 contains estimation results for the main effects of the fitted model. Effects of two-factor interactions are generally quite smaller and are omitted for a simplified presentation. The significant influence of previous production plans on current production plans provides a clear indication for the continuity of production planning. Also, the strong influence of an improved expected business condition indicates high dependency of economic growth on positive expectation, as described in several economic theories, and the effect of “orders in hand” allows a similar interpretation.  $\square$

**Table 6.6.** Estimates of main effects

Category	Parameter	Estimator	<i>p</i> -value
+	threshold	-6.51	0.000
=	threshold	-1.45	0.000
+	$P_{t-1}^+$	4.34	0.000
=	$P_{t-1}^+$	2.35	0.000
+	$P_{t-1}^=$	1.11	0.003
=	$P_{t-1}^=$	1.82	0.000
+	$D_t^+$	3.78	0.000
=	$D_t^+$	1.61	0.000
+	$D_t^=$	2.13	0.000
=	$D_t^=$	1.68	0.000
+	$O_t^+$	3.54	0.000
=	$O_t^+$	1.50	0.000
+	$O_t^+$	1.59	0.000
=	$O_t^=$	0.98	0.000
Pearson's $\chi^2$ :		24.17, 16 df,	<i>p</i> -value 0.09
Deviance:		24.05, 16 df,	<i>p</i> -value 0.09

### 6.2.2 Marginal Models

In many longitudinal studies, data consist of a small or moderate number of repeated observations for many subjects, and the main objective is to analyze the effect of covariates on a response variable, without conditioning on previous responses. Thus, one has essentially the same situation as in Section 3.5 but within a longitudinal context of many usually short time series. In this setting it is more natural to view the data as a cross section of individual time series

$$\mathbf{y}_i = (y'_{i1}, \dots, y'_{iT_i})', \quad \mathbf{x}_i = (x'_{i1}, \dots, x'_{iT_i})', \quad i = 1, \dots, n,$$

and to assume that, given the covariates, individual time series are mutually independent. Marginal models for univariate responses  $y_{it}$  (non-normal longitudinal data) have been introduced by Liang & Zeger (1986) and Zeger & Liang (1986).

Their GEE approach has been extended in various ways, in particular to multicategorical responses and to likelihood-based models; see the book by Diggle, Liang & Zeger (1994), the bibliography in Agresti (1999) and Ziegler, Kastner & Blettner (1998), the review of Molenberghs & Lesaffre

(1999), and other work, already cited in Section 3.5.2. To avoid unnecessary repetition and too much overlap with the presentation given there, we give here only a brief summary in a notation covering univariate and multivariate responses. As usual, a multivariate response  $Y_{it}$  with  $k$  categories is coded by a vector  $y'_{it} = (y_{it1}, \dots, y_{itq})$  of  $q = k - 1$  dummy variables, and we assume, for simplicity, that the number of categories is the same for all observations.

Marginal GEE models for longitudinal data are then defined as follows:

- (i) The *marginal mean* of the component  $y_{it}$  of  $y_i$  is correctly specified by a response model

$$\mu_{it}(\beta) = E(y_{it}|x_{it}) = h(\eta_{it}), \quad \eta_{it} = Z_{it}\beta,$$

where  $h$  is a known response function and  $Z_{it}$  a design matrix constructed from covariates  $x_{it}$ . For univariate response,  $Z_{it}$  reduces to a design vector  $z'_{it}$  and the predictor  $\eta_{it} = z'_{it}\beta$  is scalar. For multivariate response, the response function and the design matrix are chosen to represent a multivariate regression model as in Sections 3.2 and 3.3, and the predictor  $\eta_{it}$  is  $q$ -dimensional with components  $\eta_{itr} = z'_{itr}\beta$ , where  $z'_{itr}$  is the  $r$ th row of  $Z_{it}$ .

- (ii) The association structure for  $y_i = (y'_{i1}, \dots, y'_{iT_i})'$  is specified by a *working* covariance matrix  $\Sigma_i(\beta, \alpha)$ , depending on regression parameters  $\beta$ , on association parameters  $\alpha$ , and, possibly but notationally suppressed, on an additional nuisance parameter  $\phi$ . This working covariance matrix can be determined as in Section 3.5.2 by specification of a variance function for  $y_{it}$  together with a working correlation matrix  $R_i(\alpha)$  or some odds ratio model  $\gamma_i(\alpha)$ .

Compared to Section 3.5.2, the choice of the association structure can be guided by models borrowed from time series analysis. For example, for univariate responses the choice

$$(R_i(\alpha))_{st} = \alpha(|t - s|) \tag{6.2.6}$$

corresponds to stationary correlations between  $y_{is}$  and  $y_{it}$ .

The special form  $\alpha(|t - s|) = \alpha^{|t-s|}$  mimics the autocorrelation function of a Gaussian AR(1) process. Further examples can be found in Liang & Zeger (1986) or obtained from an underlying random effects model as, e.g., in Hamerle & Nagl (1988) and Thall & Vail (1990).

## Statistical Inference

Estimation of  $\beta$  using GEEs is carried out in complete analogy to Section 3.5.2. According to the model assumptions, the observation vectors  $y_i = (y'_{i1}, \dots, y'_{iT_i})'$ ,  $i = 1, \dots, n$ , are independent and have means  $E(y_i|x_i) = \mu_i(\beta) = (\mu'_{i1}, \dots, \mu'_{iT_i})'$  and working covariance matrices  $\Sigma_i(\beta, \alpha)$ . Defining

design matrices  $Z'_i = (Z_{i1}, \dots, Z_{iT_i})$  and (block-) diagonal matrices  $D_i(\beta) = \text{diag}(D_{it}(\beta))$ ,  $D_{it}(\beta) = \partial h/\partial\eta_{it}$  evaluated at  $\eta_{it} = Z_{it}\beta$ , the generalized estimating equation for  $\beta$  is

$$s_\beta(\beta, \alpha) = \sum_{i=1}^n Z'_i D_i(\beta) \Sigma_i^{-1}(\beta, \alpha) (y_i - \mu_i(\beta)) = 0. \quad (6.2.7)$$

In the special case of the working independence assumption the working covariance matrix  $\Sigma_i(\beta, \alpha)$  is (block-)diagonal and (6.2.7) reduces to the usual form of the score function for independent observations. Generally,  $\Sigma_i(\beta, \alpha)$  is a  $(qT_i \times qT_i)$ -matrix that depends on the covariate effects  $\beta$ , the correlation parameter  $\alpha$ , and the dispersion parameter. To compute  $\hat{\beta}$ , one cycles between a modified Fisher scoring for  $\beta$  and estimation of  $\alpha$  and  $\phi$ , e.g., by some method of moments or by a second GEE, in complete analogy to Section 3.5.2. Given current estimates  $\hat{\alpha}$  and  $\hat{\phi}$ , the GEE (6.2.7) for  $\hat{\beta}$  is solved by the iterations

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + (\hat{F}^{(k)})^{-1} \hat{s}^{(k)},$$

with the “working” Fisher matrix

$$\hat{F}^{(k)} = \sum_{i=1}^n Z_i D_i(\hat{\beta}^{(k)}) \Sigma_i^{-1}(\hat{\beta}^{(k)}, \hat{\alpha}) D'_i(\hat{\beta}^{(k)}) Z'_i,$$

where “ $\hat{\cdot}$ ” means evaluation at  $\beta = \hat{\beta}^{(k)}$ , and the corresponding quasi-score function  $\hat{s}^{(k)} = s(\hat{\beta}^{(k)}, \hat{\alpha}, \hat{\phi})$ . Given a current estimate of  $\beta$ , the parameters  $\alpha$  and  $\phi$  can be estimated from Pearson residuals

$$\hat{r}_{it} = \frac{y_{it} - \hat{\mu}_{it}}{(v(\hat{\mu}_{it}))^{1/2}}$$

by the method of moments detailed in Section 3.5.2, with some minor change in notation. The dispersion parameter is estimated consistently by

$$\hat{\phi} = \frac{1}{N-p} \sum_{i=1}^n \sum_{t_1}^{T_i} \hat{r}_{it}^2, \quad N = \sum_{i=1}^n T_i.$$

Estimation of  $\alpha$  depends on the choice of  $R_i(\alpha)$ . For the exchangeable correlation model  $\alpha$  can be estimated by

$$\hat{\alpha} = \frac{1}{\hat{\phi} \left\{ \sum_{i=1}^n \frac{1}{2} T_i(T_i - 1) - p \right\}} \sum_{i=1}^n \sum_{t>s} \hat{r}_{it} \hat{r}_{is}.$$

The parameters of the stationary model (6.2.6) can be estimated analogously by averages of the corresponding residuals  $\hat{r}_{it}, \hat{r}_{is}$ .

A totally unspecified  $R = R(\alpha)$  can be estimated as in Section 3.5.2 if  $T$  is small compared to  $n$ .

In particular for binary and categorical observations, estimation via a second GEE

$$\sum_{i=1}^n \frac{\partial v_i}{\partial \alpha} B_i^{-1} (w_i - \nu_i) = 0 \quad (6.2.8)$$

for  $\alpha$  is often preferable. In (6.2.8), the vector  $w_i = (\dots, w_{ist}(l, m), \dots)'$  contains the centered products  $w_{ist}(l, m) = (y_{isl} - \pi_{isl})(y_{itm} - \pi_{itm})$ ,  $l, m = 1 \dots, q, s < t = 1, \dots, T_i$ , and  $\nu_i = (\dots, \nu_{ist}(l, m), \dots)'$  is the vector of covariances  $\nu_{i,st}(l, m) = E(y_{isl}y_{itm}) - \pi_{isl}\pi_{itm}$ . The matrix  $B_i$  is a further working covariance matrix for  $w_i$ , for example,  $B_i = I$ ; see Section 3.5.2. Cycling between Fisher scoring for (6.2.7) and (6.2.8) until convergence gives GEE estimates  $\hat{\beta}$  and  $\hat{\alpha}$ .

Consistency and asymptotic normality results for  $\hat{\beta}$  can be obtained on the lines of asymptotic theory of quasi-MLEs for independent observations if  $T_i, i = 1, \dots, n$  is fixed and  $n \rightarrow \infty$ ; see Section 3.5.2 and Appendix A.2. Standard asymptotic theory imposes “mild” regularity conditions leading to  $n^{1/2}$ -asymptotics as in Theorem 2 of Liang & Zeger (1986): If  $\hat{\alpha}$  is  $n^{1/2}$ -consistent given  $\beta, \phi$ , and  $\hat{\phi}$  is  $n^{1/2}$ -consistent given  $\beta$ , then  $\hat{\beta}$  is  $n^{1/2}$ -consistent and asymptotically multivariate Gaussian; briefly,

$$\hat{\beta} \xrightarrow{a} N(\beta, F^{-1}VF^{-1}),$$

with

$$F = \sum_{i=1}^n Z_i' D_i \Sigma_i^{-1} D_i Z_i, \quad V = \sum_{i=1}^n Z_i' D_i \Sigma_i^{-1} \text{cov}(y_i) \Sigma_i^{-1} D_i Z_i.$$

The asymptotic covariance matrix  $A = F^{-1}VF^{-1}$  can be consistently estimated by replacing  $\beta, \alpha$ , and  $\phi$  by their estimates and  $\text{cov}(y_i)$  by  $(y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)'$ , i.e., by the sandwich matrix

$$\hat{A} = \text{cov}(\hat{\beta}) \xrightarrow{a} \hat{F}^{-1} \left\{ \sum_{i=1}^n Z_i' \hat{D}_i \hat{\Sigma}_i^{-1} (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' \hat{\Sigma}_i^{-1} \hat{D}_i Z_i \right\} \hat{F}^{-1}. \quad (6.2.9)$$

Based on the preceding asymptotic result, confidence intervals for  $\beta$  and Wald and score tests may be constructed; compare with Section 2.3.1. If the covariance structure  $\Sigma_i$  is correctly specified or consistently estimated as in the totally unspecified case, then estimation is asymptotically efficient. If  $R_i(\alpha)$  is just a working assumption, then some loss of efficiency will occur.

For univariate response, Liang & Zeger (1986) and Hamerle & Nagl (1988) provide some evidence that this loss can be quite small for simple choices of  $R_i(\alpha)$ . The study of McDonald (1993) recommends the *independence* working model whenever correlation is merely regarded as a nuisance. In our experience, loss of efficiency matters more with mult categorical response.

As an alternative to the GEE approach, likelihood-based methods have been developed; see Section 3.5.2. For computational reasons, however, its applicability is restricted to only a few repeated observations, particularly with mult categorical response.

#### **Example 6.4: Ohio children**

Zeger, Liang & Albert (1988) analyzed a subset of data from the Harvard Study of Air Pollution and Health, reported by Laird, Beck & Ware (1984) and reproduced in Table 6.7. The data consist in reports for 537

**Table 6.7.** Presence and absence of respiratory infection

Mother Did Not Smoke					Mother Smoked				
Age of child				Frequency	Age of child				Frequency
7	8	9	10		7	8	9	10	
0	0	0	0	237	0	0	0	0	118
0	0	0	1	10	0	0	0	1	6
0	0	1	0	15	0	0	1	0	8
0	0	1	1	4	0	0	1	1	2
0	1	0	0	16	0	1	0	0	11
0	1	0	1	2	0	1	0	1	1
0	1	1	0	7	0	1	1	0	6
0	1	1	1	3	0	1	1	1	4
1	0	0	0	24	1	0	0	0	7
1	0	0	1	3	1	0	0	1	3
1	0	1	0	3	1	0	1	0	3
1	0	1	1	2	1	0	1	1	1
1	1	0	0	6	1	1	0	0	4
1	1	0	1	2	1	1	0	1	2
1	1	1	0	5	1	1	1	0	4
1	1	1	1	11	1	1	1	1	7

children from Ohio, examined annually from ages 7 to 10. Responses  $y_{it}$ ,  $i = 1, \dots, 537$ ,  $t = 1, \dots, 4$  are binary, with  $y_{it} = 1$  for the presence and  $y_{it} = 0$  for the absence of respiratory infection. To analyze the influence of mother's smoking status and of age on children's respiratory disease, we assume that the marginal probability of infection follows a logit model

$$\log \frac{P(\text{infection})}{P(\text{no infection})} = \beta_0 + \beta_S x_S + \beta_{A1} x_{A1} + \beta_{A2} x_{A2} + \beta_{A3} x_{A3} + \\ + \beta_{SA1} x_S x_{A1} + \beta_{SA2} x_S x_{A2} + \beta_{SA3} x_S x_{A3},$$

where smoking status is coded by  $x_S = 1$  for smoking,  $x_S = -1$  for non-smoking;  $x_{A1}$ ,  $x_{A2}$ ,  $x_{A3}$  represent age in effect-coding, with  $x_{A4}$  as reference category, and  $x_S x_{A1}$ ,  $x_S x_{A2}$ ,  $x_S x_{A3}$  are corresponding interaction terms. Table 6.8 shows estimates based on three correlation assumptions:  $R = I$  (independence assumption),  $R_{st} = \alpha$ ,  $s \neq t$  (exchangeable correlation), and unspecified correlation  $R$ . For all three working correlations, estimates are identical for the first relevant digits, so only one column is given for point estimates and robust standard deviations, based on the sandwich matrix (6.2.9). For comparison, naive standard deviations, obtained from the usual ML method for independent observations, are also displayed.

Looking at the parameter estimate  $\hat{\beta}_S$  alone, the results seem to indicate a positive effect of mother's smoking on the probability of infection. The naive standard deviation also slightly supports this finding. However, the correct standard deviation, which is larger since smoking status is a time-independent covariate, shows that the effect of smoking is overinterpreted if analysis is falsely based on common ML estimation for independent observations. The effects  $\hat{\beta}_{A1}$  and  $\hat{\beta}_{A3}$  of ages 7 and 9 are slightly positive but nonsignificant. The positive effect  $\hat{\beta}_{A2}$  of age 8 is slightly more significant. Comparing standard deviations, it is seen that robust

**Table 6.8.** Marginal logit model fits for Ohio children data

Parameter	Standard Deviation		
	Effect	Robust	Naive
$\hat{\beta}_0$	-1.696	0.090	0.062
$\hat{\beta}_S$	0.136	0.090	0.062
$\hat{\beta}_{A1}$	0.059	0.088	0.107
$\hat{\beta}_{A2}$	0.156	0.081	0.104
$\hat{\beta}_{A3}$	0.066	0.082	0.106
$\hat{\beta}_{SA1}$	-0.115	0.088	0.107
$\hat{\beta}_{SA2}$	0.069	0.081	0.104
$\hat{\beta}_{SA3}$	0.025	0.082	0.106

**Table 6.9.** Main effects model fits for Ohio children data

Parameter	Effect		Standard Deviation	
	Independent	Exchangeable/Unspecified	Robust	Naive
$\hat{\beta}_0$	-1.695	-1.696	0.090	0.062
$\hat{\beta}_S$	0.136	0.130	0.089	0.062
$\hat{\beta}_{A1}$	0.087	0.087	0.086	0.103
$\hat{\beta}_{A2}$	0.141	0.141	0.079	0.102
$\hat{\beta}_{A3}$	0.060	0.060	0.080	0.103

estimates are smaller since age is a time-dependent covariate. However,  $\hat{\beta}_{A4} = -\hat{\beta}_{A1} - \hat{\beta}_{A2} - \hat{\beta}_{A3} = -0.28$  of age 10 is negative and highly significant (stand. dev. = 0.094). This means that the probability of infection is significantly lower at age 10. It seems that childrens' constitution is more stable at this age. Interaction effects are nonsignificant on the basis of the estimated standard deviations. Therefore, their contribution on the probability of infection should only be interpreted with great caution. This is also supported by the estimates in Table 6.9, obtained from fitting marginal models without interaction effects.  $\square$

A closely related class of models for ordered categorical outcomes has been proposed by Stram, Wei & Ware (1988) and Stram & Wei (1988). Admitting time-dependent parameters, they propose fitting a marginal cumulative response model

$$\pi_{it} = h(Z_{it}\beta_t)$$

with time-dependent parameter  $\beta_t$  separately for each occasion. For  $n \rightarrow \infty$ , the combined estimate  $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_T)$  becomes asymptotically normal, and its asymptotic covariance matrix can be estimated empirically. Zeger (1988a) shows that  $\hat{\beta}$  can be viewed as the solution of a GEE with working correlation matrix  $R(\alpha) = I$ , and that the covariance matrix is identical to the one obtained from the GEE approach. A similar approach has been considered by Moulton & Zeger (1989). They combine the estimated coefficients at each time point by using bootstrap methods or weighted least-squares methods. For applications of these approaches to clinical trials, see Davis (1991).

In principle, the estimate  $\hat{\beta}$  could be used to examine trends in the sequence  $\beta_t$ , to combine estimates, etc. However, enough data for each occasion are required to obtain reasonable estimates. If this is not the case, models with time-varying parameters according to a stochastic process or a "state space" model (Chapter 8) are preferable.

Additional aspects that arise in longitudinal studies are time-varying covariates and incomplete data due to dropouts. Problems with time-varying covariates are discussed in Fitzmaurice & Laird (1993), Pepe & Anderson (1994), and Robin, Rotnitzky & Zhao (1995). Effects of dropouts or missing data are studied by Fitzmaurice, Lipsitz & Molenberghs (1995), Lesaffre, Molenberghs & Dewulf (1996), and Robin, Rotnitzky & Zhao (1995). More recently, nonparametric approaches for flexible modelling and exploring mean and association structures have been suggested; see the following section.

### 6.2.3 Generalized Additive Models for Longitudinal Data

An important feature of conditional and marginal models in the previous sections is that they still rely on the structural assumption of linear parametric predictors. Extensions to more flexible predictors for classes of generalized additive models defined in Chapter 5 is straightforward for *conditional models*, since past responses can be formally treated like covariates. Any of the basic principles for nonparametric modelling and fitting, that is, penalization, localization, and basis function methods, can be applied again. More recently, *generalized additive marginal models* for longitudinal data have been developed. Wild & Yee (1996) have introduced an additive extension to GEE-based models for univariate responses using penalized GEEs. Gieger (1998, 1999) extends the penalized GEE approach to multicategorical responses and to additive models for the mean *and* association structure. Semi- and nonparametric modelling of predictors in GEEs based on local regression techniques have recently been considered by Carroll, Ruppert & Welsh (1998) and more specifically for longitudinal data with ordinal responses by Kauermann (2000). Heagerty & Zeger (1998) proposed a nonparametric model for association if scientific interest is focused on the dependence structure. We briefly outline the penalized GEE approach.

Additive and semiparametric marginal models assume that marginal responses  $y_{it}$  have exponential family density with means specified by  $\mu_{it} = E(y_{it}|x_{it}) = h(\eta_{it})$ , with predictors  $\eta_{it}$  in generic additive form

$$\eta_{it} = \alpha + \sum_{j=1}^p f_{(j)}(x_{itj}) \quad (6.2.10)$$

or one of the other semiparametric forms of Section 5.3.1.

Association is modelled either parametrically by specifying working correlations or odds ratios as in Sections 3.5.2 and 6.2.2, or nonparametrically, assuming an additive predictor for association parameters, e.g., a nonparametric model for odds ratios.

Estimation is based on penalized GEEs. The basic idea derives from penalized quasi-likelihood. It generalizes penalized likelihood approaches by replacing the log-likelihood  $l(\eta; y)$  of the data by a quasi log-likelihood  $Q(\eta; y)$ ,

say. Denoting vectors of unknown function evaluations by  $f_j$ ,  $j = 1, 2, \dots$ , penalized quasi-likelihood estimators  $\hat{f}_j$  maximize

$$Q(\eta; y) - \frac{1}{2} \sum_j \lambda_j f'_j K_j f_j,$$

with penalty matrices  $K_j$  and smoothing parameters  $\lambda_j$  as in Section 5.3. Under the GEE framework, a quasi-likelihood may not be well defined. However, it still makes sense to start directly from a penalized quasi-score function or GEE. Gathering all function evaluations in a big vector  $f = (\dots, f'_j, \dots)'$ , smoothing parameters in  $\Lambda = \text{diag}(\dots, \lambda_j, \dots)'$ , and penalty matrices in  $K = \text{blockdiag}(\dots, K_j, \dots)$ , the penalized GEE is obtained by adding the first derivative of the penalty term above to the GEE (6.2.7). This defines the PGEE

$$s(f) = \sum_{i=1}^n D_i \Sigma_i^{-1} (y_i - \mu_i) - \Lambda K f = 0,$$

with  $D_i = \partial h / \partial \eta$ ,  $\Sigma_i$  and  $\mu_i$  evaluated at  $\eta_i$  defined by (6.2.10). Berhane & Tibshirani (1998) model association by a working correlation matrix  $R_i(\alpha)$  as in Section 3.5.2. They estimate unknown functions by cycling between fitting an additive model to the PGEE via backfitting, given a current estimate for  $\alpha$ , and estimating  $\alpha$  by the method of moments. Gieger (1998) uses the Demmler-Reinsch basis for splines (see Eubank, 1988), thus avoiding backfitting, and defines a second (penalized) GEE for estimating working association structures parametrically or nonparametrically by cycling between modified Fisher scoring algorithms. We illustrate his approach with the following example.

### **Example 6.5: A longitudinal study on forest damage**

These longitudinal data are collected in a yearly visual forest damage inventory carried out in a forest district in the northern part of Bavaria since 1983. There are 80 observation points with occurrence of beech trees spread over the whole area. In this damage study we analyze the influence of calendar time and of covariates, e.g., age of the trees, pH value of soil, and canopy density at the stand, on the defoliation of beeches at the stand. A detailed survey and data description can be found in Göttlein & Pruscha (1996). As in Example 3.14 the degree of defoliation is an indicator for the damage state of trees. Due to the survey design, responses must be assumed to be serially correlated. The ordinal response variable  $Y_t$ , “damage state” at time  $t$ , is measured in three categories: none ( $Y_t = 1$ ), light ( $Y_t = 2$ ), and distinct/strong ( $Y_t = 3$  = reference) defoliation. Figure 6.3 shows the relative frequencies of the damage categories in the sample for the years 1983 to 1994.

A cumulative logistic model is chosen to relate the marginal probabilities of “damage state” to the following covariates:

- A age of beeches at the beginning of the study with categories: below 50 years (= 1), between 50 and 120 years (= 2), and above 120 years (= reference);
- pH pH value of the soil in 0–2 cm depth, measured as a metrical covariate ranging from a minimum of 3.3 to a maximum of 6.1;
- CD canopy density at the stand with categories: low(= 1), medium (= 2), and high (= reference).

The covariates pH value and canopy density vary over time for each stand, while the variable age is constant over time by construction. Based on preliminary exploratory data analysis, Gieger (1998, 1999) assumes a marginal cumulative logit model

$$\text{logit}(\text{pr}(Y_t \leq r)) = f_r(t) + f_3(t)A^{(1)} + f_4(t)A^{(2)} + f_5(pH_t) + \alpha_6 CD_t^{(1)} + \alpha_7 CD_t^{(2)},$$

for the probabilities of no damage ( $r = 1$ ) and light damage ( $r = 2$ ), with  $A^{(1)}, A^{(2)}, CD_t^{(1)}, CD_t^{(2)}$  as dummy variables for the categorical covariates  $A$  and  $CD$ . The threshold functions  $f_1$  and  $f_2$  and the effects  $f_4$  and  $f_5$  of the time constant variable age are assumed to vary smoothly with time  $t$ . Due to a lack of information about the form of the influence, it is reasonable to model the effect of pH value nonparametrically by an unspecified smooth function  $f_5$  as well. The effects  $\alpha_6$  and  $\alpha_7$  of canopy density are assumed to be fixed.

Pairwise association is modelled by global cross ratios. Supported by preliminary descriptive analysis, the association structure is parameterized by a log global cross ratio model

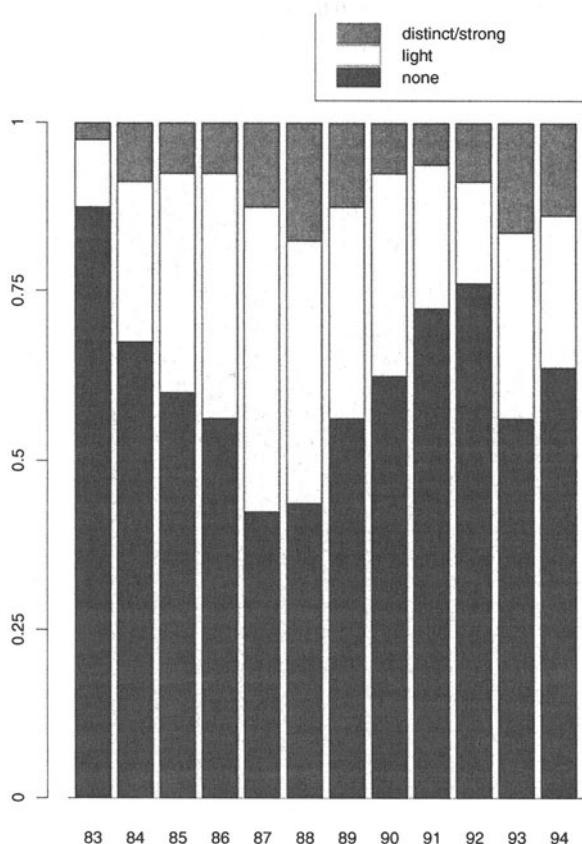
$$\log \gamma_{ist}(l, m) = f_{lm}(|t - s|), \quad l, m = 1, 2.$$

Thus, no specific parameter form is assumed for the dependence on the time lag  $|t - s|$ . The estimates for the association functions (Figure 6.4) are quite similar in their global shape, but the levels are different. We also recognize a distinct temporal structure.

Figure 6.5 shows the estimated threshold functions  $\hat{f}_1(t)$  (left plot) and  $\hat{f}_2(t)$  (right plot). Both curves decrease up to the year 1988 with a more pronounced decrease of the first threshold  $\hat{f}_1(t)$ . This indicates a shift to higher probabilities for the categories light and distinct/strong damage up to this year. After an improvement, i.e., a shift to the no damage category, up to 1992 there is another increase in damage until 1994. This result is true for beeches older than 120 years, i.e., for the reference category of the age. For the other two categories of age, the effects of the corresponding dummy variables  $A^{(1)}$  and  $A^{(2)}$  have to be added. Both effects are positive over the 12 years (Figure 6.6, left plot), indicating that younger trees are more robust. The positive effect (upper curve) is greater for trees younger than 50

years, and the increase of this effect after 1988 corrects the change to the worse after 1992, which is indicated by the threshold functions. These interpretations are further illustrated by Figure 6.7, where fitted probabilities stratified by age are plotted against time.

The estimated function  $f_5$  for the influence of pH value is more or less linear over the range of observed pH values (Figure 6.6, right plot). Stands with low pH values have a negative influence on damage state compared to stands with less acidic soils, i.e., low pH values aggravate the condition of the trees. However, the influence of pH value does not seem to be very strong. Table 6.10 contains parameter estimates for the effects of canopy density together with model based and robust standard errors.



**Figure 6.3.** Damage class distribution by time.

Covariate	Estimate	SE (model)	SE (robust)	<i>p</i> -value (model)	<i>p</i> -value (robust)
$CD_t^{(1)}$	-1.2822	0.3587	0.3104	0.0003	0.0000
$CD_t^{(2)}$	-0.5318	0.2481	0.2196	0.0320	0.0153

**Table 6.10.** Effects of canopy density

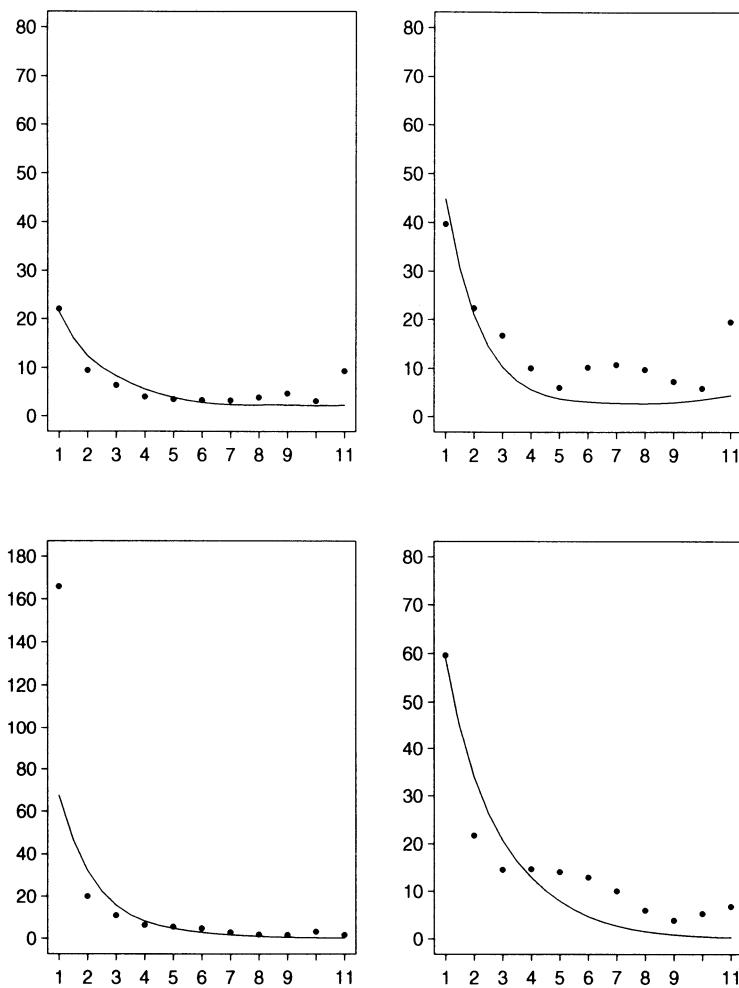
Both estimates are significantly negative. This means that stands with low ( $CD_t^{(1)}$ ) or medium ( $CD_t^{(2)}$ ) canopy density have an increased probability for high damage compared to stands with a high canopy density. The reason could be that lower canopy densities result in rougher conditions for the tree population, connected with stronger physiological, aerodynamic, and physical stress.

Model based and robust standard errors for curves and parameters for the marginal model are close together. This indicates that the association structure is modelled and fitted quite well.  $\square$

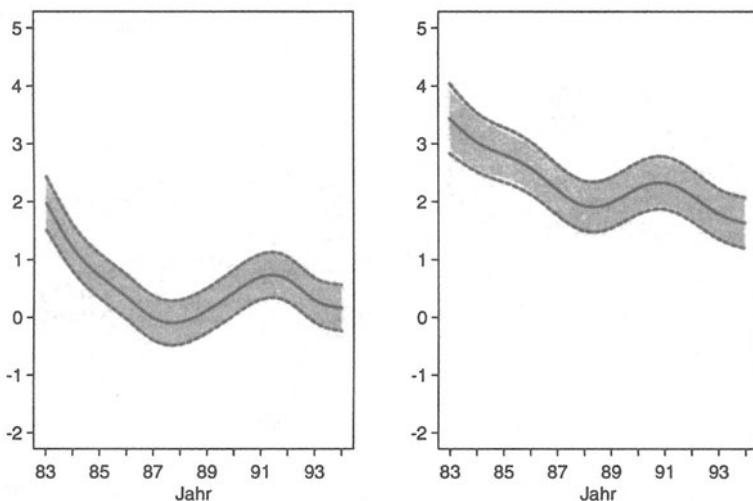
## 6.3 Notes and Further Reading

Whereas Bayesian inference for parametric conditional models for time series or longitudinal data can be carried out with the same methods as for the cross-sectional regression model in Chapters 2 and 3, this is not possible for the GEE approach for marginal models. As with cross-sectional multivariate categorical responses in Section 3.5, an attractive alternative is to base Bayesian analysis on latent Gaussian models for time series or longitudinal data. Chib (1999) and Chen & Dey (1999) and the references therein provide a good survey for correlated binary or ordinal responses.

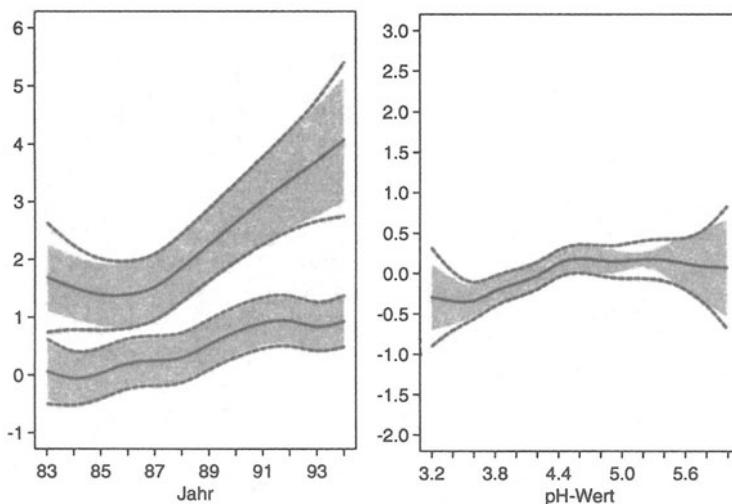
Artes & Jorgensen (2000) extend the GEE method to the class of dispersion models (Jorgensen, 1997). As shown by Song (2000), there exists a close relationship between the GEE approach and likelihood regression based on multivariate dispersion models generated from Gaussian copulas.



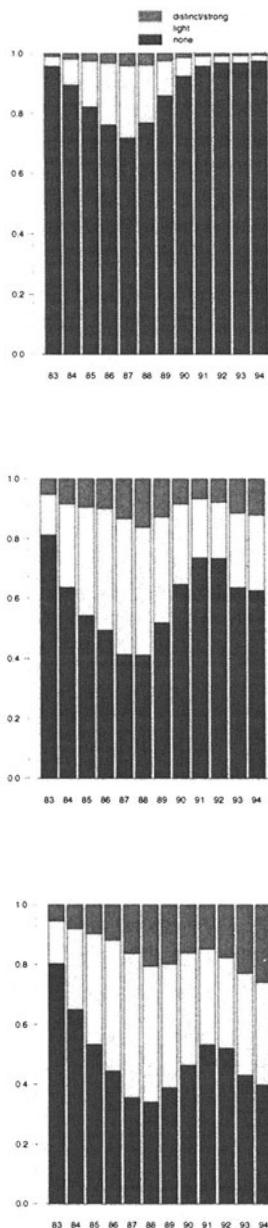
**Figure 6.4.** Estimated global odds ratios (lines) and empirically observed global odds ratios (points). Note that there is a different scale for the combination  $l = 2, r = 1$ .



**Figure 6.5.** Estimated thresholds  $\hat{f}_1(t)$  (left plot) and  $\hat{f}_2(t)$  (right plot) with pointwise standard error bands (model based – dashed line, robust – boundary of shaded region).



**Figure 6.6.** Estimated effects of age (left) with  $A^{(1)}$  (upper curve),  $A^{(2)}$  (lower curve), and pH value (right) with pointwise standard error bands (model based – dashed line, robust – boundary of shaded region).



**Figure 6.7.** Estimated probabilities  $\text{pr}(Y_t = 1)$ ,  $\text{pr}(Y_t = 2)$ , and  $\text{pr}(Y_t = 3)$  for age. From top to bottom: up to 50 years; between 50 and 120 years; above 120 years.

# Random Effects Models

This chapter is concerned with random effects models for analyzing non-normal data that are assumed to be clustered or correlated. The clustering may be due to repeated measurements over time, as in longitudinal studies, or to subsampling the primary sampling units, as in cross-sectional studies. In each of these cases the data consist of **repeated observations**  $(y_{it}, x_{it})$ ,  $t = 1, \dots, T_i$ , for each individual or unit  $i = 1, \dots, n$ , where  $y$  denotes a **response variable** of primary interest and  $x$  a **vector of covariates**. Typical examples include panel data, where the cluster-specific data

$$(y_i, x_i) = (y_{i1}, \dots, y_{iT_i}, x_{i1}, \dots, x_{iT_i})$$

correspond to a time series of length  $T_i$ , or large-scale health studies, where  $(y_i, x_i)$  represents the data of a primary sampling unit, say a hospital or a geographical region.

Let us consider the Ohio children data (Example 6.4), which are further investigated in this chapter. In this data set the objective is to investigate the effect of mother's smoking status on the presence/absence of respiratory infection. Since children's response is considered at ages 7, 8, 9, and 10, we have repeated measurements. In Example 6.4 a marginal fixed effect approach was used to analyze these data. However, susceptibility to respiratory infection differs in highly individual ways that cannot be traced back to only one covariate and time effects. If the focus is on *individual* risk probabilities, unobservable heterogeneity should be taken into account. Therefore, one may consider a logit model with covariates "mother's smoking status" and "age" and a random intercept that is allowed to vary randomly across children.

Most of the models considered in previous sections specify effects to be constant across clusters. However, when one is interested in individual risks, a better assumption is that the parameters, e.g., intercept and/or covariate effects, may vary across clusters. Models of the first type are called *population-averaged*, whereas the latter approach is *cluster- or subject-specific*. Parameters in population-averaged models may be interpreted with

respect to the marginal or population-averaged distribution. In subject-specific models the parameters refer to the influence of covariates for individuals. The distinction between these approaches is irrelevant for Gaussian outcomes, but it becomes important for categorical data since the mixture of subject-specific logistic models in general is not a logistic model for the population. For the comparison of these alternative model types, see also Zeger, Liang & Albert (1988), Neuhaus, Kalbfleisch & Hauck (1991), and Agresti (1993b).

The cluster- or subject-specific approach considered in this chapter is based on random effects: Cluster-specific effects are assumed to be independent and identically distributed according to a mixing distribution. In principle, cluster-specific effects can also be treated as fixed. In the fixed effects approach cluster-specific dummy variables have to be introduced such that each of the cluster-specific effects is treated as an unknown fixed parameter. However, the substantial number of parameters to be estimated often gives rise to serious problems, in particular when the cluster sizes  $T_i$  are small or moderate. Random effects approaches are much more parsimonious in the parameters, and estimation can be carried out even when cluster sizes are small.

Throughout the entire chapter we retain the basic assumption that random effects are independent and identically distributed. However, this assumption is not always appropriate. With longitudinal or spatial data, heterogeneity is often caused by serially or correlated random effects. Chapter 8 treats serially correlated random effects in detail and outlines extensions to spatial and temporal-spatial random effects.

For normal data, linear random effects models are commonly used in theory and practice. Estimation of unknown parameters and of cluster-specific random effects is easier since the mixing distribution can be chosen to be conjugate to the data density so that the posterior distribution is available in closed form. Therefore, random effects models for Gaussian data are only sketched in Section 7.1. In Section 7.2 random effects generalized linear models are introduced. The statistical analysis of random effects models for non-normal data becomes more difficult due to the lack of analytically and computationally tractable mixing distributions. Therefore, estimation methods based on full marginal and posterior distributions or on posterior means require repeated approximations of multidimensional integrals. In this chapter we report on work in this area that still is in progress. In Section 7.3 an estimation procedure based on posterior modes is given that has been used by Stiratelli, Laird & Ware (1984) and Harville & Mee (1984). It is closely related to approximate inference using penalized (quasi-) likelihood, suggested in Breslow & Clayton (1993) for generalized linear mixed models with independent as well as correlated random effects. In Section 7.4 integration-based two-step procedures are considered. First the fixed effects are estimated by maximizing the marginal likelihood based on Gauss-

Hermite quadrature or Monte Carlo techniques. Then an empirical Bayes approach is used to estimate the random effects. Techniques of this type have been used by Hinde (1982), Anderson & Aitkin (1985), Anderson & Hinde (1988), and Jansen (1990). In Section 7.7 a marginal estimation approach due to Zeger, Liang & Albert (1988) is sketched. Fully Bayesian methods using MCMC simulation are outlined in Section 7.6.

Several books extensively treat the case of normally distributed data, see, e.g., Goldstein (1987), Searle, Casella & McCulloch (1992). Readable books, which also treat univariate generalized linear models with random efficient coefficients, are Longford (1993) and Diggle, Liang & Zeger (1994).

## 7.1 Linear Random Effects Models for Normal Data

For clustered Gaussian data linear random effects models provide an efficient tool for analyzing cluster-specific intercepts and/or covariate effects. Although there is some development of nonlinear models (e.g., Lindstrom & Bates, 1990) we consider only linear models. This section gives a short review on modelling and estimation techniques. More detailed expositions may be found in Hsiao (1986), Rao & Kleffe (1988), Dielman (1989), Jones (1993), and Lindsey (1993).

### 7.1.1 Two-stage Random Effects Models

Linear random effects models as considered by Laird & Ware (1982) extend the classical linear model for Gaussian response variables. We first define the general model and describe familiar special cases further later.

At the *first stage* the normal responses  $y_{it}$  are assumed to depend linearly on unknown population-specific effects  $\beta$  and on unknown cluster-specific effects  $b_i$ ,

$$y_{it} = z'_{it}\beta + w'_{it}b_i + \varepsilon_{it}, \quad (7.1.1)$$

where  $z_{it}$  and  $w_{it}$  represent design vectors,  $w_{it}$  often being a subvector of  $z_{it}$ , and the disturbances  $\varepsilon_{it}$  being uncorrelated normal random variables with  $E(\varepsilon_{it}) = 0$  and  $\text{var}(\varepsilon_{it}) = \sigma^2$ . The design vector  $z_{it}$  and thus  $w_{it}$  may depend on deterministic or stochastic covariates and on past responses, as in longitudinal studies, such that

$$z_{it} = z_{it}(x_{it}, y_{i,t-1}^*), \quad \text{with } y_{i,t-1}^* = (y_{i,t-1}, \dots, y_{i1}).$$

Furthermore, it is assumed that  $z_{it}$  contains the intercept term “1.”

At the *second stage* the effects  $b_i$  are assumed to vary independently from one cluster to another according to a mixing distribution with mean  $E(b_i) = 0$ . Since the disturbances are Gaussian a normal mixing density with unknown covariance matrix  $\text{cov}(b_i) = Q$  is commonly chosen,

$$b_i \sim N(0, Q), \quad Q > 0, \quad (7.1.2)$$

and the sequences  $\{\epsilon_{it}\}$  and  $\{b_i\}$  are assumed to be mutually uncorrelated. Although it is possible to use a more general form of covariance structure for  $\{\epsilon_{it}\}$ , e.g., first-order autoregression (see Jones, 1993), we only consider the uncorrelated case. In matrix notation model (7.1.1) takes the form

$$y_i = Z_i \beta + W_i b_i + \varepsilon_i, \quad (7.1.3)$$

where  $y'_i = (y_{i1}, \dots, y_{iT_i})'$  is the response vector,  $Z'_i = (z_{i1}, \dots, z_{iT_i})'$ ,  $W'_i = (w_{i1}, \dots, w_{iT_i})'$  are design matrices, and  $\varepsilon'_i = (\varepsilon_{i1}, \dots, \varepsilon_{iT_i})'$  is the vector of within cluster errors,  $\varepsilon_i \sim N(0, \sigma_\varepsilon^2 I)$ .

The assumption of Gaussian errors allows us to rewrite the model as a multivariate heteroscedastic linear regression model

$$y_i = Z_i \beta + \varepsilon_i^*, \quad (7.1.4)$$

where the multivariate disturbances  $\varepsilon_i^* = (\varepsilon_{i1}^*, \dots, \varepsilon_{iT_i}^*)'$ , with components  $\varepsilon_{it}^* = w'_{it} b_i + \varepsilon_i$ , are independent and normally distributed,

$$\varepsilon_i^* \sim N(0, V_i), \quad \text{with } V_i = I\sigma_\varepsilon^2 + W_i Q W_i'. \quad (7.1.5)$$

Equation (7.1.4), together with the covariance structure (7.1.5), represents the *marginal version* of the linear random effects model, where the conditioning on the cluster-specific effects  $b_i$  is dropped. Marginally, responses are correlated within clusters, as can be seen from (7.1.5).

In the following we discuss some special versions of linear random effects models, which are applied to analyze varying intercepts, (partially) varying slopes or covariate effects, or cluster-specific effects being hierarchically nested.

## Random Intercepts

In many empirical studies cluster-specific characteristics that possibly determine the response variable in addition to the observed covariates  $x_{it}$  have not been collected due to technical or economical circumstances. To take account of such cluster-specific effects a linear model with cluster-specific intercepts  $\tau_i$  is appropriate,

$$y_{it} = \tau_i + x'_{it} \gamma + \varepsilon_{it}, \quad (7.1.6)$$

where the slope coefficients  $\gamma$  are constant and the intercepts  $\tau_i$  are assumed to be i.i.d. with unknown parameters  $E(\tau_i) = \tau$  and  $\text{var}(\tau_i) = \sigma^2$ . The

unobservable deviations between the population mean  $\tau$  and the cluster-specific realizations  $\tau_i$  may be interpreted as effects of omitted covariates. Assuming normality one immediately gets a linear random effects model of the form (7.1.1), (7.1.2) with

$$\beta' = (\tau, \gamma'), \quad z'_{it} = (1, x'_{it}), \quad w_{it} = 1, \quad b_i = (\tau_i - \tau) \sim N(0, \sigma^2).$$

Note that the random-intercept model also takes into account intracluster correlation of the Gaussian outcomes. From (7.1.5) it is easily seen that the correlation between  $y_{it}$  and  $y_{is}$  is given by

$$\rho_{ts} = \rho = \frac{\sigma^2}{\sigma_\varepsilon^2 + \sigma^2}, \quad t \neq s.$$

Random intercept models are also called *error components* or *variance components models* (see, e.g., Hsiao, 1986). The primary objective is to analyze the variance components  $\sigma_\varepsilon^2$  and  $\sigma^2$ , which stand for variability within (resp., between) the clusters.

### Random Slopes

Random intercept models do not alleviate the restrictive assumption that the slope coefficients are equal for each observation. Varying slope coefficients arise in particular in longitudinal studies, where intercept and slope coefficients are specific to each time series. To take into account such parameter heterogeneity, the random intercept model (7.1.6) can be extended, treating not only the intercept but all coefficients as random. The corresponding model has the form

$$y_{it} = \tau_i + x'_{it}\gamma_i + \varepsilon_{it}.$$

Note that in the longitudinal setting the  $i$ th cluster corresponds to a time series. Then effects of past responses are also allowed to vary from time series to time series.

Suppose now that the regression coefficients  $\beta'_i = (\tau_i, \gamma'_i)$  vary independently across clusters according to a normal density with mean  $E(\beta_i) = \beta$  and covariance matrix  $\text{cov}(\beta_i) = Q$  being positive definite,

$$\beta_i \sim N(\beta, Q).$$

The mean  $\beta$  can be interpreted as the population-averaged effect of the regressors  $z_{it}$ . The covariance matrix  $Q$  contains variance components indicating the parameter heterogeneity of the population as well as covariance components representing the correlation of single components of  $\beta_i$ . Rewriting the coefficients  $\beta_i$  as

$$\beta_i = \beta + b_i, \quad b_i \sim N(0, Q),$$

and assuming mutual independence of the sequences  $\{b_i\}$  and  $\{\epsilon_{it}\}$ , one obtains a linear random effects model with

$$z'_{it} = w'_{it} = (1, x'_{it}).$$

Models where all coefficients are assumed to vary randomly over clusters are also called *random coefficient regression models*; see, e.g., Hsiao (1986). Sometimes, however, assuming that some coefficients are cluster-specific is less realistic than assuming that some coefficients are constant across clusters. If  $\beta_{i1}$  denotes the cluster-specific coefficients and  $\beta_{i2}$  the remaining coefficients are constant across clusters, the parameter vector  $\beta_i$  can be partitioned into  $\beta'_i = (\beta'_{i1}, \beta'_{i2})$  with  $\beta_{i2} = \beta_2$  for all  $i$ . The design vector  $z'_{it} = (1, x'_{it})$  also has to be rearranged according to the structure  $z'_{it} = (z'_{it1}, z'_{it2})$ . Then the probability model for  $\beta_i$  can be expressed by a multivariate normal density with singular covariance matrix,

$$\beta_i = \begin{bmatrix} \beta_{i1} \\ \beta_{i2} \end{bmatrix} \sim N \left( \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix} \right), \quad (7.1.7)$$

where the submatrix  $Q$  is assumed to be positive definite. Rewriting the coefficients  $\beta_i$  as

$$\beta_i = \beta + a_i, \quad \text{with } a'_i = (b'_i, 0'), \quad b_i \sim N(0, Q),$$

one gets again a random effects model with

$$z'_{it} = (z'_{it1}, z'_{it2}), \quad w_{it} = z_{it1}. \quad (7.1.8)$$

Due to the mixing of “fixed” and “random” coefficients, models of this type are also called linear mixed models.

## Multilevel Models

The linear random effects models considered so far are based on a single level structure. In some applications, however, clustering occurs on more than one level and the clusters are hierarchically nested. For example, in a standard application in educational research there are classrooms with varying numbers of students, and each student has a pair of scores, one at the beginning and the other at the conclusion of a specific course. In this case we have a two-level structure: At the first level the clusters correspond to  $j = 1, \dots, n$  classrooms, each having  $i = 1, \dots, I_j$  students, and at the second level the clusters are formed by students, each having  $t = 1, \dots, T_i$  observations. Models that take into account unobservable classroom-specific effects and student-specific effects have hierarchically nested clusters. Models of this type have been considered by Aitkin & Longford (1986), among others.

### 7.1.2 Statistical Inference

The primary objective in analyzing linear random effects models is to estimate the parameters  $\beta$ ,  $\sigma_\varepsilon^2$ , and  $Q$ , and the random effects  $b_i$ . Estimation is often based on a frequentist approach, where  $\beta$ ,  $\sigma_\varepsilon^2$ , and  $Q$  are treated as “fixed” parameters. As an alternative, Bayesian estimation methods can be applied, where  $\beta$ ,  $\sigma_\varepsilon^2$ , and  $Q$  are treated as random variables with some prior distribution. In the following we start with the frequentist approach including “empirical Bayes” methods. Fully Bayesian methods will be treated in Section 7.6. For a better understanding we first treat the more hypothetical situation, where the variance-covariance components  $\sigma_\varepsilon^2$  and  $Q$  are known. For simplicity only single-level models are considered. Extensions to multi-level models can be derived; see, for example, Goldstein (1986), Longford (1987), and Goldstein (1989).

#### Known Variance-Covariance Components

First let the covariance matrix  $V_i$  (i.e.,  $\sigma_\varepsilon^2$  and  $Q$ ) be known. Then estimation of the parameter  $\beta$  is usually based on the marginal model defined by (7.1.4) and (7.1.5). For this model the maximum likelihood estimator (MLE) is equal to the weighted least-squares solution

$$\hat{\beta} = \left( \sum_{i=1}^n Z_i' V_i^{-1} Z_i \right)^{-1} \sum_{i=1}^n Z_i' V_i^{-1} y_i. \quad (7.1.9)$$

The MLE can also be shown to be a best linear unbiased estimator (BLUE); see, e.g., Harville (1977) and Rao & Kleffe (1988).

Estimation of the unobservable random effects  $b_i$  is based on the posterior density of  $b_i$ , given the data  $Y = (y_1, \dots, y_n)$ . Due to the normality and linearity assumptions the posterior of  $b_i$  is also normal. Moreover, the posterior depends only on  $y_i$  since the stochastic terms  $\epsilon_{it}$  and  $b_i$ ,  $t = 1, \dots, T_i$ ,  $i = 1, \dots, n$ , are assumed to be completely independent. Following Bayesian arguments the optimal point estimator is the posterior mean

$$\hat{b}_i = E(b_i | y_i) = Q W_i' V_i^{-1} (y_i - Z_i \hat{\beta}), \quad (7.1.10)$$

which also can be shown to be a BLUE in the present setting (see, e.g., Harville, 1976; Rao & Kleffe, 1988). Since the posterior is normal, posterior mean and posterior mode coincide. Therefore, the estimators  $\hat{b}_i$  are also obtained by maximizing the log-posterior density with respect to  $b_1, \dots, b_n$ .

#### Unknown Variance-Covariance Components

The estimating equations (7.1.9) and (7.1.10) are based on known variance-covariance components. Let  $\theta = (\sigma_\varepsilon^2, Q)$  denote the vector of variance and

covariance parameters. If  $\theta$  is unknown it has to be replaced by a consistent estimate  $\hat{\theta}$ . Harville (1977), among others, distinguishes maximum likelihood estimation (MLE) and restricted maximum likelihood estimation (RMLE). Some authors also suggest minimum norm quadratic unbiased estimation (MINQUE) and minimum variance quadratic unbiased estimation (MIVQUE), which, however, are less commonly used since such estimates may be negative (see, e.g., Hsiao, 1986; Rao & Kleffe, 1988).

The MLEs for  $\theta$  are obtained together with those for  $\beta$  by maximizing the marginal log-likelihood based on the marginal model defined by (7.1.4) and (7.1.5) with respect to  $\beta$  and  $\theta$ . This is equivalent to the minimization of

$$l(\beta, \theta) = \sum_{i=1}^n [\log |V_i| + (y_i - Z_i\beta)' V_i^{-1} (y_i - Z_i\beta)].$$

The criticism of MLE for  $\theta$  is that this estimator does not take into account the loss in degrees of freedom resulting from the estimation of  $\beta$ . The larger the dimension of  $\beta$ , the larger is the bias of the MLE for  $\theta$ . RMLEs generally yield a smaller bias. The idea of an RMLE is to construct a likelihood that depends only on  $\theta$ . Such a likelihood can be derived using a Bayesian formulation of the linear random effects model, where the parameters  $\beta$  are considered random variables having a vague or totally flat prior distribution, for example,

$$\beta \sim N(\beta^*, \Gamma), \quad \text{with} \quad \Gamma \rightarrow \infty \quad \text{or} \quad \Gamma^{-1} \rightarrow 0,$$

so that the prior density of  $\beta$  is just a constant. The choice of  $\beta^*$  is immaterial since the covariance matrix  $\Gamma$  becomes infinite. Maximizing the limiting (as  $\Gamma^{-1} \rightarrow 0$ ) marginal log-likelihood yields the RMLE for  $\theta$  (see Harville, 1976). The numerical calculation of MLEs and RMLEs for  $\theta$  is much more complicated than the estimation of  $\beta$  since the likelihoods depend nonlinearly on  $\theta$ . Thus, iterative methods are required. Harville (1977) considers Newton-Raphson and scoring algorithms, and Laird & Ware (1982) recommend several versions of the EM algorithm. Since the EM algorithm will also be used for generalized linear models with random coefficients, a version of this algorithm used for RMLE is sketched in the following.

Some details on its derivation are given below. The resulting EM algorithm jointly estimates  $\delta = (\beta, b_1, \dots, b_n)$  and  $\theta = (\sigma_\epsilon^2, Q)$  as follows:

1. Choose starting values  $\theta^{(0)} = (\sigma_\epsilon^{2(0)}, Q^{(0)})$ .

For  $p = 0, 1, 2, \dots$

2. Compute  $\hat{\delta}^{(p)} = (\hat{\beta}^{(p)}, \hat{b}_1^{(p)}, \dots, \hat{b}_n^{(p)})$  from (7.1.9) and (7.1.10) with variance-covariance components replaced by their current estimates  $\theta^{(p)} = (\sigma_\epsilon^{2(p)}, Q^{(p)})$ , together with current residuals  $e_i^{(p)} = y_i - Z_i\hat{\beta}^{(p)} - W_i\hat{b}_i^{(p)}$ ,  $i = 1, \dots, n$ , and posterior covariance matrices  $\text{cov}(b_i|y_i; \theta^{(p)})$ ,  $\text{cov}(\varepsilon_i|y_i; \theta^{(p)})$ .

3. EM-step: Compute  $\theta^{(p+1)} = (\sigma_\varepsilon^{2(p+1)}, Q^{(p+1)})$  by

$$\hat{\sigma}_\varepsilon^{2(p+1)} = \frac{1}{T_1 + \dots + T_n} \sum_{i=1}^n [(e_i^{(p)})' e_i^{(p)} + \text{tr}(\text{cov}(\varepsilon_i | y_i; \theta^{(p)}))], \quad (7.1.11)$$

$$Q^{(p+1)} = \frac{1}{n} \sum_{i=1}^n [b_i^{(p)} (b_i^{(p)})' + \text{cov}(b_i | y_i; \theta^{(p)})]. \quad (7.1.12)$$

The posterior covariance matrices can be obtained from the joint normal density defined by the model. For details and alternative versions, see Laird & Ware (1982) and Jones (1993).

However, RMLEs should not be used blindly, because RMLEs may have larger mean square errors than MLEs, as Corbeil & Searle (1976) have pointed out. Further tools of statistical inference on  $\beta$  and  $\theta$  are less well developed. An approximation for the distribution of  $\hat{\beta}$  that takes into account the additional variability due to the estimation of  $\theta$  has been proposed by Giesbrecht & Burns (1985). Approximative distributions of MLEs and RMLEs for  $\theta$  are considered by Rao & Kleffe (1988).

### Derivation of the EM algorithm\*

Indirect maximization of the marginal density by an EM algorithm starts from the joint log-density of observable data  $Y = (y_1, \dots, y_n)$  and unobservable effects  $\delta = (\beta, b_1, \dots, b_n)$ ; see Appendix A.3. Since a flat prior is assumed for  $\beta$ , the joint log-likelihood is

$$\log f(Y, \delta; \theta) = \log f(Y | \delta; \sigma_\varepsilon^2) + \log f(b_1, \dots, b_n; Q),$$

where the first term is determined by the first stage (7.1.1), resp., (7.1.3), of the model, and the second term by the second stage (7.1.2), the prior for  $b_1, \dots, b_n$ . From the model assumptions one obtains, up to constants,

$$\begin{aligned} S_1(\sigma_\varepsilon^2) &= -\frac{1}{2} \sum_{i=1}^n T_i \log \sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^n \varepsilon_i' \varepsilon_i, \\ S_2(Q) &= -\frac{n}{2} \log \det(Q) - \frac{1}{2} \sum_{i=1}^n b_i' Q^{-1} b_i \\ &= -\frac{n}{2} \log \det(Q) - \frac{1}{2} \sum_{i=1}^n \text{tr}(Q^{-1} b_i b_i') \end{aligned}$$

for the first and second terms. The E-step yields

$$M(\theta | \theta^{(p)}) = E\{S_1(\sigma_\varepsilon^2) | y; \theta^{(p)}\} + E\{S_2(Q) | y; \theta^{(p)}\}$$

with

$$\begin{aligned} E\{S_1(\sigma_\varepsilon^2)|y; \theta^{(p)}\} &= -\frac{1}{2} \sum_{i=1}^n T_i \log \sigma_\varepsilon^2 \\ &\quad - \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^n [(e_i^{(p)})' e_i^{(p)} + \text{tr cov}(\varepsilon_i|y_i; \theta^{(p)})], \end{aligned} \quad (7.1.13)$$

$$\begin{aligned} E\{S_2(Q)|y; \theta^{(p)}\} &= -\frac{n}{2} \log \det(Q) \\ &\quad - \frac{1}{2} \sum_{i=1}^n \text{tr}[Q^{-1} b_i^{(p)} (b_i^{(p)})'] + \text{cov}(b_i|y_i; \theta^{(p)}). \end{aligned} \quad (7.1.14)$$

Differentiation with respect to  $\sigma_\varepsilon^2$  and  $Q$  using matrix calculus setting derivatives to zero and solving for  $\sigma_\varepsilon^2$  and  $Q$  yields (7.1.11) and (7.1.12).

## 7.2 Random Effects in Generalized Linear Models

Let us consider two simple examples with categorical responses.

**Example 7.1: Ohio children data** (Example 6.4, continued)

As already mentioned at the beginning of this section, in the study of Ohio children one has the dichotomous response presence/absence of respiratory infection. The children are measured repeatedly (ages 7, 8, 9, 10) and the covariate of interest is the smoking behavior of mothers.  $\square$

**Example 7.2: Bitterness of white wines**

In a study on the bitterness of white wine (Randall, 1989) it is of interest whether treatments that can be controlled during pressing the grapes influence the bitterness of wines. The two factors considered are the temperature and the admission of contact with skin when pressing the grapes. Both factors are given in dichotomous form. For each factor combination two bottles of white wine were chosen randomly and the bitterness of each of the  $t = 1, \dots, 8$  bottles was classified on a 5-categorical ordinal scale ( $1 = \text{nonbitter}, \dots, 5 = \text{very bitter}$ ) by  $i = 1, \dots, 9$  professional judges. The 5-categorical responses  $y_{it}$  are given in Table 7.1. Since judges cannot be expected to have the same sensitivity to bitterness, an effect of judges should be incorporated. This may be done by allowing the judges to have shifted thresholds in a cumulative model.  $\square$

For non-normal data we cannot expect to get a closed form like  $y_i = Z_i\beta + \varepsilon_i^*$  as we had in (7.1.4) for normal data. For the introduction of random effects models in this case, it is useful to reconsider the linear random effects model as a two-stage model:

**Table 7.1.** Bitterness of wine data (Randall, 1989)

Judge	Low Temperature				High Temperature			
	No Contact		Contact		No Contact		Contact	
	Bottle	Bottle	Bottle	Bottle	Bottle	Bottle	Bottle	Bottle
1	2	3	3	4	4	4	5	5
2	1	2	1	3	2	3	5	4
3	2	3	3	2	5	5	4	4
4	3	2	3	2	3	2	5	3
5	2	3	4	3	3	3	3	3
6	3	2	3	2	2	4	5	4
7	1	1	2	2	2	3	2	3
8	2	2	2	3	3	3	3	4
9	1	2	3	2	3	2	4	4

At the first stage the observations  $y_{it}$  are treated as conditionally independent and normally distributed random variables, given the effects  $b_i$ ,

$$y_{it}|b_i \sim N(\mu_{it}, \sigma_\varepsilon^2), \quad (7.2.1)$$

where the *conditional* mean  $\mu_{it} = E(y_{it}|b_i)$  is given by  $\mu_{it} = z'_{it}\beta + w'_{it}b_i$ .

At the second stage the cluster-specific effects  $b_i$  are assumed to be i.i.d. with  $b_i \sim N(0, Q)$ . If the covariates  $x_{it}$  are stochastic or the responses  $y_{it}$  depend on past observations  $y_{it}^* = (y_{i,t-1}, \dots, y_{i1})$ , as in longitudinal studies, the model is to be understood conditionally, i.e., (7.2.1) is the conditional density of  $y_{it}$ , given  $x_{it}, y_{it}^*, b_i$ . However, to avoid unnecessary inflation of notation, the possible dependence on  $x_{it}$  or  $y_{it}^*$  is suppressed in the following.

### Generalized Linear Models with Random Effects

For non-normal responses  $y_{it}$  like the binary response infection (Example 7.1) or the multicategorical response bitterness (Example 7.2), model (7.2.1) can be extended in the following way:

At the *first stage* it is assumed that the conditional density  $f(y_{it}|b_i)$  is of the simple uni- or multivariate exponential family type with conditional mean

$$\mu_{it} = E(y_{it}|b_i) = h(\eta_{it}), \quad \text{with} \quad \eta_{it} = Z_{it}\beta + W_{it}b_i, \quad (7.2.2)$$

where  $h$  is one of the response functions in Sections 2.1, 3.2, or 3.3, and  $\eta_{it}$  is the linear predictor. The design matrix  $Z_{it}$  is a function of the covariates

$x_{it}$ , and possibly past responses  $y_{it}^*$  as considered in the previous chapters. Intercepts, covariates, and/or past responses whose effects are assumed to vary across the  $i = 1, \dots, n$  clusters are collected in the design matrix  $W_{it}$ . Though it is not necessary,  $W_{it}$  often is a submatrix of  $Z_{it}$ .

The *second stage* of linear random effects models is retained: Cluster-specific effects  $b_i$  are assumed to be independent and (if  $b_i$  in (7.2.2) is not restricted) normally distributed with mean  $E(b_i) = 0$  and unknown covariance matrix  $\text{cov}(b_i) = Q$ , where  $Q$  has to be positive definite. In principle, each parametric density  $f$  having mean  $E(b_i) = 0$  and unknown parameters  $Q$  is admissible so that the second stage of the generalized linear random effects model is specified more generally by independent densities

$$p(b_i; Q), \quad i = 1, \dots, n. \quad (7.2.3)$$

As an additional assumption, conditional independence of observations within and between clusters is required, i.e.,

$$f(Y|B; \beta) = \prod_{i=1}^n f(y_i|b_i; \beta), \quad \text{with} \quad f(y_i|b_i; \beta) = \prod_{t=1}^{T_i} f(y_{it}|b_i; \beta), \quad (7.2.4)$$

where  $Y = (y_1, \dots, y_n)$  and  $B = (b_1, \dots, b_n)$  represent the whole set of responses and random effects. If covariates are stochastic or responses depend on past observations, densities are to be understood conditional on these quantities. Marginally, where the conditioning on  $b_i$  is dropped, observations *within* clusters are dependent. Observations on *different* clusters are conditionally and marginally independent.

## Examples

### Example 1 (binary logistic model)

For univariate responses  $y_{it}$  the design matrices  $Z_{it}$  and  $W_{it}$  reduce to design vectors  $z_{it}$  and  $w_{it}$  that can be constructed along the lines of Section 7.1.1. Therefore, univariate generalized linear models imposing no restrictions on the predictor  $\eta$  are easily extended to include random effects. Let us consider a binary logistic model with intercepts varying according to a normal distribution,

$$\pi_{it} = P(y_{it} = 1|b_i) = \exp(\eta_{it})/(1 + \exp(\eta_{it})), \quad (7.2.5)$$

with

$$\eta_{it} = z'_{it}\beta + b_i, \quad b_i \sim N(0, \sigma^2),$$

where the design vector  $z_{it}$  is assumed to contain the intercept term “1.”

For Example 7.1 with random threshold the design for smoking mother and a child of seven years is given by

$$\eta_{it} = [1, 1, 1, 0, 0]\beta + [1]b_i$$

where  $\beta$  contains the fixed effects “constant,” “smoking status,” “age 7 years,” “age 8 years,” “age 9 years.”  $\square$

The mixed-effects logistic model (7.2.5) is cluster- or subject-specific in contrast to population-averaged models where the predictor has the simple form  $\eta_{it} = z'_{it}\beta$ . If the subject-specific model holds the covariate effects measured by the population-averaged model with  $\eta_{it} = z'_{it}\beta$  are closer to zero than the covariate effects of the underlying subject-specific model (see Neuhaus, Kalbfleisch & Hauck, 1991). The binary model (7.2.5) has been considered by Anderson & Aitkin (1985) in analyzing interviewer variability. More complex binary logistic models involving multinormally distributed random effects have been studied by Stiratelli, Laird & Ware (1984) for longitudinal data. Extensions to multilevel structures can be found in Preisler (1989) and Wong & Mason (1985). Log-linear Poisson models with a normally distributed intercept are treated by Hinde (1982) and Brillinger & Preisler (1983).

However, generalized linear random effects approaches are also suited for analyzing clustered *multicategorical* data. Let  $y'_{it} = (y_{it1}, \dots, y_{itk})$  denote the  $k$ -categorical response of the  $t$ th observation in cluster  $i$ , where  $y_{itj} = 1$  if category  $j$  is observed and  $y_{itj} = 0$  otherwise,  $j = 1, \dots, k$ . Then multinomial random effects models for nominal or ordered responses are completely specified by conditional probabilities  $\pi'_{it} = (\pi_{it1}, \dots, \pi_{itq}), q = k - 1$ , which depend on population-specific effects and cluster-specific effects,

$$\pi_{it} = h(\eta_{it}), \quad \eta_{it} = Z_{it}\beta + W_{it}b_i.$$

The structure of the design matrix  $W_{it}$  depends on the assumed parameter heterogeneity.

#### *Example 2 (multinomial logistic model)*

A multinomial logistic random effects model that takes into account the effect of unobservable characteristics being cluster- and category-specific is given by

$$\pi_{itj} = \exp(\eta_{itj}) / (1 + \sum_{m=1}^q \exp(\eta_{itm})),$$

with

$$\eta_{itj} = \alpha_{ij} + x'_{it}\gamma_j, \quad j = 1, \dots, q.$$

Assuming that the effects  $\alpha'_i = (\alpha_{i1}, \dots, \alpha_{iq})$  vary independently from one cluster to another according to a normal distribution with mean  $E(\alpha_i) = \tilde{\alpha}$  and covariance matrix  $\text{cov}(\alpha_i) = Q$ ,

$$\alpha_i \sim N(\tilde{\alpha}, Q),$$

the linear predictors can be rewritten as

$$\eta_{itj} = \tilde{\alpha}_j + b_{ij} + x'_{it} \gamma_j, \quad \text{where} \quad b_{ij} = \alpha_{ij} - \tilde{\alpha}_j.$$

Defining  $\beta' = (\tilde{\alpha}_1, \gamma'_1, \dots, \tilde{\alpha}_q, \gamma'_q)$  and  $b'_i = (b_{i1}, \dots, b_{iq}) \sim N(0, Q)$ , one obtains a conditional multinomial model of the form (7.2.2) with

$$Z_{it} = \begin{bmatrix} 1 & x'_{it} & & & 0 \\ & & 1 & x'_{it} & \\ & & & \ddots & \\ 0 & & & & 1 & x'_{it} \end{bmatrix}, \quad W_{it} = \begin{bmatrix} 1 & & & 0 \\ & \ddots & & \\ 0 & & & 1 \end{bmatrix}.$$

This model can also be developed by applying the random utility principle proposed in Section 3.3. Then the latent utility variable has to be based on a linear random effects model that is similar to (7.1.5). Inclusion of cluster-specific slopes and nested cluster-specific effects can be performed along the lines of Section 7.1.1.  $\square$

### *Example 3 (ordered response)*

In the bitterness-of-wine example judges responded on a five-point scale that may be considered ordinal. Random effects versions of *ordinal* response models (see Sections 3.3 and 3.4) can be derived in the same way as before. However, one has to be cautious with cumulative models. Such models are based on thresholds being constant for all observations. To allow for variation over clusters the thresholds may be parameterized by linear combinations of the observed covariates, as stated in Section 3.3.2. Sometimes, however, the heterogeneity of thresholds cannot be explained adequately by the observed covariates since there exist some unobservable cluster-specific characteristics. To take into account unobservable threshold heterogeneity cumulative approaches with randomly varying thresholds may be appropriate. If  $F$  denotes a known distribution function, e.g., the logistic, conditional response probabilities are given by

$$\pi_{it1} = F(\theta_{i1} + x'_{it}\gamma), \quad \pi_{itr} = F(\theta_{ir} + x'_{it}\gamma) - \pi_{it,r-1}, \quad r = 2, \dots, q,$$

with cluster-specific thresholds being ordered,

$$-\infty = \theta_{i0} < \theta_{i1} < \dots < \theta_{iq} < \theta_{im} = \infty, \quad (7.2.6)$$

and covariate effects  $\gamma$  being constant over clusters. The simplest random effects model is given by cluster-specific shifting of thresholds where

$$\theta_{ir} = \theta_r + b_i, \quad b_i \sim N(0, \sigma^2).$$

Then the linear predictor has the form

$$\eta_{itr} = \theta_r + b_i + x'_{it}\gamma.$$

The matrices in the linear form  $\eta_{it} = Z_{it}\beta + W_{it}b_i$  are given by

$$Z_{it} = \begin{pmatrix} 1 & & x'_{it} \\ & \ddots & \vdots \\ & & 1 & x'_{it} \end{pmatrix}, \quad W_{it} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix},$$

where  $\beta' = (\theta_1, \dots, \theta_q, \gamma')$ .

An extended version assumes all the thresholds to be cluster-specific. The simplest assumption is the normal distribution  $\theta'_i = (\theta_{i1}, \dots, \theta_{iq}) \sim N(\theta_i, Q)$ . However, if the threshold means  $\theta_r$  are not well separated and the variances are not small enough, the ordering (7.2.6) may be violated and numerical problems will occur in the estimation procedure. To overcome this problem the thresholds may be reparameterized as already stated in Section 3.3.3 using

$$\alpha_{i1} = \theta_{i1}, \quad \alpha_{ij} = \log(\theta_{ij} - \theta_{i,j-1}), \quad j = 2, \dots, q.$$

The reparameterized thresholds  $\alpha'_i = (\alpha_{i1}, \dots, \alpha_{iq})$  may vary unrestrictedly in  $\mathbb{R}^q$  according to a normal distribution with mean  $E(\alpha_i) = \tilde{\alpha}$  and covariance matrix  $\text{cov}(\alpha_i) = Q$ . The conditional response probabilities are now

$$\begin{aligned} \pi_{it1} &= F(\alpha_{i1} + x'_{it}\gamma), \\ \pi_{itj} &= F(\alpha_{i1} + \sum_{m=2}^j \exp(\alpha_{im}) + x'_{it}\gamma) - \pi_{it,j-1}, \quad j = 2, \dots, q. \end{aligned}$$

Rewriting the reparameterized thresholds as

$$\alpha_i = \tilde{\alpha} + b_i, \quad \text{with} \quad b_i \sim N(0, Q),$$

and defining  $\beta = (\tilde{\alpha}, \gamma)$  yields a cumulative logistic random effects model, where the design matrices are given by

$$Z_{it} = \begin{bmatrix} 1 & & 0 & x'_{it} \\ & \ddots & & 0 \\ & & \ddots & \vdots \\ 0 & & & 1 & 0 \end{bmatrix}, \quad W_{it} = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}.$$

It should be noted that a simple shifting of thresholds  $\theta_{ir} = \theta_r + b_i$  after reparameterization corresponds to the adding of a random effect to  $\alpha_{i1}$  yielding  $\alpha_{i1} = \alpha_1 + b_i$ ,  $b_i \sim N(0, \sigma^2)$ .

Cumulative random effects models allowing for varying slopes are obtained in the usual way, e.g., by using a linear random effects model as latent variable model (see, e.g., Harville & Mee, 1984; Jansen, 1990).  $\square$

## 7.3 Estimation Based on Posterior Modes

In this section we consider a procedure for the simultaneous estimation of the parameters  $\beta$ ,  $Q$  and the  $n$  random effects  $b_i$  which is based on posterior modes, thus avoiding integration. The procedure corresponds to an EM-type algorithm, where posterior modes and curvatures are used instead of posterior means and covariances. Such an approach was already sketched in Section 2.3.2 for analyzing Bayes models. For binary logistic models with Gaussian random effects it has been used by Stiratelli, Laird & Ware (1984). See also Harville & Mee (1984) for the application to cumulative random effects models and Wong & Mason (1985), who adopted the approach to logistic regression models with nested random effects structures. The EM-type algorithm may be derived from an EM algorithm that maximizes the marginal likelihood involving  $Q$  only. In linear random effects models the resulting estimators  $\hat{Q}$  correspond to RMLEs (see Section 7.1.2). Estimation of the parameters  $\beta$  and the random effects  $b_i$  is embedded in the E-step of the EM-type algorithm by maximizing the joint posterior density with  $Q$  fixed at  $\hat{Q}$ .

For fixed or known  $Q$ , posterior mode estimation of  $\beta$  and  $b_1, \dots, b_n$  is equivalent to the penalized quasi-likelihood approach of Breslow & Clayton (1993), derived from Laplace's method for integral approximation. However, their approximate profile quasi-likelihood method for estimating variance components and further suggestions in Breslow & Lin (1995) are different from the EM-type algorithm.

Since the complete estimation procedure is based on a marginal likelihood, which only depends on the variance-covariance components  $Q$ , a vague or flat prior density with covariance matrix  $\Gamma \rightarrow \infty$  is assigned to the parameters  $\beta$  (see Section 7.1.2). Collecting the population and cluster-specific effects into

$$\delta' = (\beta', b'_1, \dots, b'_n),$$

the limiting (as  $\Gamma \rightarrow \infty$ ) prior density of  $\delta$  satisfies the proportionality

$$p(\delta; Q) \propto \prod_{i=1}^n p(b_i; Q).$$

### 7.3.1 Known Variance-Covariance Components

For simplicity we first consider the estimation of  $\delta$  for known variance-covariance components  $Q$ . Estimation of  $\delta$  is based on its posterior density, given the data  $Y = (y_1, \dots, y_n)$ . Applying Bayes theorem and using the independence assumption (7.2.4) the posterior is given by

$$f(\delta|Y; Q) = \frac{\prod_{i=1}^n f(y_i|b_i, \beta) \prod_{i=1}^n p(b_i; Q)}{\int \prod_{i=1}^n f(y_i|b_i, \beta) p(b_i; Q) db_1 \cdots db_n d\beta}. \quad (7.3.1)$$

An optimal point estimator for  $\delta$  is the posterior mean that was used to estimate random effects for Gaussian data; see equation (7.1.10). However, due to the lack of analytically and computationally tractable random effects posterior densities, numerical integration or Monte Carlo methods are required in general. In the present setting, these methods imply an enormous numerical effort since the integral structure in (7.3.1) is nested. Therefore, *posterior mode estimation* is used. Posterior modes and posterior curvatures are obtained by maximizing the logarithm of the posterior (7.3.1). Because of the proportionality

$$f(\delta|Y; Q) \propto \prod_{i=1}^n f(y_i|b_i, \beta) \prod_{i=1}^n p(b_i; Q),$$

maximization of (7.3.1) is equivalent to maximizing the log-posterior

$$\sum_{i=1}^n \log f(y_i|b_i, \beta) + \sum_{i=1}^n \log p(b_i; Q) \quad (7.3.2)$$

with respect to  $\delta$ . In general, computation has to be carried out iteratively, e.g., by Fisher scoring. As an example, we consider a normal random effects density  $p(b_i; Q)$  with positive definite covariance matrix  $Q$  so that (7.3.2) corresponds to

$$l(\delta) = \sum_{i=1}^n \log f(y_i|b_i, \beta) - \frac{1}{2} \sum_{i=1}^n b_i' Q^{-1} b_i \quad (7.3.3)$$

after dropping terms that are constant with respect to  $b_i$ . Criterion (7.3.3) can be interpreted as a penalized log-likelihood for random effects deviations from the population mean zero. It can also be derived from the penalized quasi-likelihood criterion in Breslow & Clayton (1993, Section 2.1) for the special case of independent and identically distributed random effects. Similarly, as for spline smoothing in generalized linear models (Chapter 5), maximization of the penalized log-likelihood  $l(\delta)$  is carried out by Fisher scoring. In each scoring step, efficient use may be made of the block structure of (expected negative) second derivatives of  $l(\delta)$ . Details are given in Section 7.3.3.

### 7.3.2 Unknown Variance-Covariance Components

Generally the variance-covariance components  $Q$  are unknown. Estimation of  $Q$  can be based on an EM-type algorithm, which can be deduced from an EM algorithm that maximizes the marginal log-likelihood

$$l(Q) = \log \int \prod_{i=1}^n f(y_i|\beta, b_i) p(b_i; Q) db_1 \cdot \dots \cdot db_n d\beta \quad (7.3.4)$$

indirectly. In general, direct maximization of (7.3.4) is cumbersome due to the nested high-dimensional integral structure, which can be solved analytically only for special situations, e.g., linear random effects models. Indirect maximization of (7.3.4) by an EM algorithm starts from the joint log-density

$$\log f(Y, \delta; Q) = \log f(Y|\delta) + \log p(\delta; Q)$$

of the observable data  $Y$  and the unobservable effects  $\delta$ . For details of the EM algorithm, see Appendix A.3. In the  $(p+1)$ th cycle of the algorithm the E-step consists of computing

$$M(Q|Q^{(p)}) = E\{\log f(Y, \delta; Q)|Y; Q^{(p)}\},$$

which denotes the conditional expectation of the complete data log-density, given the observable data  $Y$  and the estimate  $Q^{(p)}$  from the previous cycle. Because now  $Q^{(p)}$  from the previous cycle is known, we have essentially the case of known variance-covariance components considered in Section 7.3.1. To avoid integrations necessary for computing conditional expectations of random effects (or quadratic forms of them) appearing in  $\log p(\delta; Q)$ , conditional expectations are replaced by conditional modes obtained from maximizing (7.3.3) for  $Q = Q^{(p)}$ . The resulting EM-type algorithm is described in detail in the following subsection.

Alternative estimation methods have been considered by Breslow & Clayton (1993, Section 2.4) and Breslow & Lin (1995), Lin & Breslow (1996), Schall (1991), and Wolfinger (1994).

### 7.3.3 Algorithmic Details\*

#### Fisher Scoring for Given Variance-Covariance Components

Let us first consider Fisher-scoring maximization of the penalized log-likelihood  $l(\delta)$  given by (7.3.3), with  $Q$  known or given. The components of the score function  $s(\delta) = \partial l / \partial \delta = (s_\beta, s_1, \dots, s_n)'$  are then given by

$$\begin{aligned} s_\beta &= \frac{\partial l(\delta)}{\partial \beta} = \sum_{i=1}^n \sum_{t=1}^{T_i} Z'_{it} D_{it}(\delta) \Sigma_{it}^{-1}(\delta) (y_{it} - \mu_{it}(\delta)), \\ s_i &= \frac{\partial l(\delta)}{\partial b_i} = \sum_{t=1}^{T_i} W'_{it} D_{it}(\delta) \Sigma_{it}^{-1}(\delta) (y_{it} - \mu_{it}(\delta)) - Q^{-1} b_i, \quad i = 1, \dots, n, \end{aligned}$$

with  $D_{it}(\delta) = \partial h(\eta_{it}) / \partial \eta$ ,  $\Sigma_{it}(\delta) = \text{cov}(y_{it}|\beta, b_i)$ , and  $\mu_{it}(\delta) = h(\eta_{it})$ .

The expected conditional (Fisher) information matrix  $F(\delta) = \text{cov } s(\delta)$  is partitioned into

$$F(\delta) = \begin{bmatrix} F_{\beta\beta} & F_{\beta 1} & F_{\beta 2} & \cdots & F_{\beta n} \\ F_{1\beta} & F_{11} & & & 0 \\ F_{2\beta} & & F_{22} & & \\ \vdots & & & \ddots & \\ F_{n\beta} & 0 & & & F_{nn} \end{bmatrix}$$

with

$$\begin{aligned} F_{\beta\beta} &= -E\left(\frac{\partial^2 l(\delta)}{\partial \beta \partial \beta'}\right) = \sum_{i=1}^n \sum_{t=1}^{T_i} Z'_{it} D_{it}(\delta) \Sigma_{it}^{-1}(\delta) D'_{it}(\delta) Z_{it}, \\ F_{\beta i} &= F'_{i\beta} = -E\left(\frac{\partial^2 l(\delta)}{\partial \beta \partial b'_i}\right) = \sum_{t=1}^{T_i} Z'_{it} D_{it}(\delta) \Sigma_{it}^{-1}(\delta) D'_{it}(\delta) W_{it}, \\ F_{ii} &= -E\left(\frac{\partial^2 l(\delta)}{\partial b_i \partial b'_i}\right) = \sum_{t=1}^{T_i} W'_{it} D_{it}(\delta) \Sigma_{it}^{-1}(\delta) D'_{it}(\delta) W_{it} + Q^{-1}. \end{aligned}$$

The posterior mode estimator  $\hat{\delta}$  that satisfies the equation  $s(\delta) = 0$  can be calculated by the Fisher-scoring algorithm

$$\delta^{(k+1)} = \delta^{(k)} + F^{-1}(\delta^{(k)}) s(\delta^{(k)}), \quad (7.3.5)$$

where  $k$  denotes an iteration index. However, the dimensionality of  $F(\delta)$  may cause some problems. These can be avoided due to the partitioned structure of  $F(\delta)$ . Since the lower right part of  $F(\delta)$  is block diagonal, the algorithm (7.3.5) can be reexpressed more simply as

$$\begin{aligned} F_{\beta\beta}^{(k)} \Delta \beta^{(k)} + \sum_{i=1}^n F_{\beta i}^{(k)} \Delta b_i^{(k)} &= s_{\beta}^{(k)}, \\ F_{i\beta}^{(k)} \Delta \beta^{(k)} + F_{ii}^{(k)} \Delta b_i^{(k)} &= s_i^{(k)}, \quad i = 1, \dots, n, \end{aligned}$$

where  $\Delta \beta^{(k)} = \beta^{(k+1)} - \beta^{(k)}$  and  $\Delta b_i^{(k)} = b_i^{(k+1)} - b_i^{(k)}$ . After some transformations the following algorithm is obtained, where each iteration step implies working off the data twice to obtain first the corrections

$$\Delta \beta^{(k)} = \{F_{\beta\beta}^{(k)} - \sum_{i=1}^n F_{\beta i}^{(k)} (F_{ii}^{(k)})^{-1} F_{i\beta}^{(k)}\}^{-1} \{s_{\beta}^{(k)} - \sum_{i=1}^n F_{\beta i}^{(k)} (F_{ii}^{(k)})^{-1} s_i^{(k)}\}$$

and then

$$\Delta b_i^{(k)} = (F_{ii}^{(k)})^{-1} \{s_i^{(k)} - F_{i\beta}^{(k)} \Delta \beta^{(k)}\}, \quad i = 1, \dots, n.$$

If the cluster sizes  $T_i$  are large enough, the resulting estimates  $\hat{\delta} = (\hat{\beta}, \hat{b}_1, \dots, \hat{b}_n)$  become approximately normal,

$$\hat{\delta} \stackrel{a}{\sim} N(\delta, F^{-1}(\delta)),$$

under essentially the same conditions, which ensure asymptotic normality of the MLE in GLMs. Then the posterior mode and the (expected) curvature  $F^{-1}(\hat{\delta})$  of  $l(\delta)$ , evaluated at the mode, are good approximations to the posterior mean and covariance matrix.  $F^{-1}(\delta)$  is obtained using standard formulas for inverting partitioned matrices (see, e.g., Magnus & Neudecker, 1988). The result is summarized as follows:

$$F^{-1}(\delta) = \begin{bmatrix} V_{\beta\beta} & V_{\beta 1} & V_{\beta 2} & \dots & V_{\beta n} \\ V_{1\beta} & V_{11} & V_{12} & \dots & V_{1n} \\ V_{2\beta} & V_{21} & V_{22} & \dots & V_{2n} \\ \vdots & \vdots & & & \vdots \\ V_{n\beta} & V_{n1} & \dots & \dots & V_{nn} \end{bmatrix},$$

with

$$V_{\beta\beta} = (F_{\beta\beta} - \sum_{i=1}^n F_{\beta i} F_{ii}^{-1} F_{i\beta})^{-1}, \quad V_{\beta i} = V'_{i\beta} = -V_{\beta\beta} F_{\beta i} F_{ii}^{-1},$$

$$V_{ii} = F_{ii}^{-1} + F_{ii}^{-1} F_{i\beta} V_{\beta\beta} F_{\beta i} F_{ii}^{-1}, \quad V_{ij} = V'_{ji} = F_{ii}^{-1} F_{i\beta} V_{\beta\beta} F_{\beta j} F_{jj}^{-1}, \quad i \neq j.$$

### EM Type Algorithm

In the M-step  $M(Q|Q^{(p)})$  has to be maximized with respect to  $Q$ . Since the conditional density of the observable data,  $f(Y|\delta)$ , is independent of  $Q$ , the M-step reduces to maximizing

$$M(Q|Q^{(p)}) = E\{\log p(\delta; Q)|Y; Q^{(p)}\} = \int \log p(\delta; Q) f(\delta|Y; Q^{(p)}) d\delta,$$

where  $f(\delta|Y; Q^{(p)})$  denotes the posterior (7.3.1), evaluated at  $Q^{(p)}$ . As an example consider again a normal random effects density, where the M-step simplifies to the update

$$Q^{(p+1)} = \frac{1}{n} \sum_{i=1}^n (\text{cov}(b_i|y_i; Q^{(p)}) + E(b_i|y_i; Q^{(p)})E(b_i|y_i; Q^{(p)})').$$

This step is the same as for Gaussian observations (see equation (7.1.12)). Although the algorithm is simple, it is difficult to carry out the update exactly, since the posterior mean and covariance are only obtained by numerical or Monte Carlo integration in general (see Section 7.4.2). Therefore, the posterior means are approximated by the posterior modes  $\hat{b}_i$  and the posterior covariances by the posterior curvatures  $\hat{V}_{ii}$ , both of which are calculated by the Fisher-scoring algorithm (7.3.5).

The resulting EM-type algorithm with the Fisher-scoring algorithm embedded in each E-step jointly estimates  $\delta$  and  $Q$  as follows:

1. Choose a starting value  $Q^{(0)}$ .

For  $p = 0, 1, 2, \dots$

2. Compute posterior mode estimates  $\hat{\delta}^{(p)}$  and posterior curvatures  $\hat{V}_i^{(p)}$  by the Fisher-scoring algorithm (7.3.5), with variance-covariance components replaced by their current estimates  $Q^{(p)}$ .
3. EM-step: Compute  $Q^{(p+1)}$  by

$$Q^{(p+1)} = \frac{1}{n} \sum_{i=1}^n (\hat{V}_i^{(p)} + \hat{b}_i^{(p)}(\hat{b}_i^{(p)})').$$

Note that the EM-type algorithm may be viewed as an approximate EM algorithm, where the posterior of  $b_i$  is approximated by a normal distribution. In the case of linear random effects models, the EM-type algorithm corresponds to an exact EM algorithm since the posterior of  $b_i$  is normal, and so posterior mode and mean coincide, as do posterior covariance and curvature.

## 7.4 Estimation by Integration Techniques

In this section we consider maximum likelihood estimation of the “fixed” parameters  $\beta$  and  $Q$  and posterior mean estimation of the  $b_1, \dots, b_n$  in the generalized linear random effects model. Together, both methods represent a two-step estimation scheme, where first the “fixed” parameters are estimated. Then random effects are predicted on the basis of estimated “fixed” parameters. If the dimension of the random effects is high, the simultaneous scheme from Section 7.3 is of computational advantage since numerical integration is avoided in contrast to the sequential scheme. For simplicity in the following we assume that the nuisance parameter  $\phi$  is known.

### 7.4.1 Maximum Likelihood Estimation of Fixed Parameters

As in linear random effects models the unknown fixed parameters  $\beta$  and  $Q$  can be estimated by maximizing the marginal log-likelihood

$$l(\beta, Q) = \sum_{i=1}^n \log L_i(\beta, Q), \quad \text{with} \quad L_i(\beta, Q) = \int f(y_i | b_i; \beta) p(b_i; Q) db_i, \tag{7.4.1}$$

that is obtained after integrating out the random effects  $b_i$  from the conditional exponential family densities, where  $p(b_i; Q)$  is the density of random effects with  $\text{cov}(b_i) = Q$ . With  $f(y_i)$  from (7.2.4) the marginal log-likelihood to be maximized has the form

$$l(\beta, Q) = \sum_{i=1}^n \log \int \prod_{t=1}^{T_i} f(y_{it}|b_i, \beta) p(b_i, Q) db_i. \quad (7.4.2)$$

Analytical solutions of the possibly high-dimensional integrals are only available for special cases, e.g., for linear random effects models of the form (7.1.2). Therefore, numerical or Monte Carlo integration techniques are required for a wide range of commonly used random effects models, like binomial logit or log-linear Poisson models with normally distributed random effects.

If the random effects density  $p(b_i; Q)$  is symmetric around the mean, application of such techniques is straightforward after reparameterizing the random effects. In the simplest case where  $b_i \sim N(0, \sigma^2)$ ,  $b_i$  may be represented by

$$b_i = \sigma a_i, \quad a_i \sim N(0, 1).$$

Then the parameters to be estimated from maximization of the marginal likelihood are  $\beta$  and  $\sigma$  or  $\beta$  and  $\sigma^2$ , respectively. If  $b_i$  is vector-valued with  $\text{cov}(\beta_i) = Q$ , reparameterization may have the form

$$b_i = Q^{1/2} a_i,$$

where  $Q^{1/2}$  denotes the left Cholesky factor, which is a lower triangular matrix, so that  $Q = Q^{1/2} Q^{T/2}$ , where  $T$  denotes the transpose, and  $a_i$  is a standardized random vector having mean zero and the identity matrix  $I$  as covariance matrix. To obtain a linear predictor in  $\beta$  as well as in the unknown variance-covariance components of  $Q^{1/2}$ , we apply some matrix algebra (see, e.g., Magnus & Neudecker, 1988, p. 30) so that

$$\eta_{it} = Z_{it}\beta + W_{it}Q^{1/2}a_i$$

can be rewritten in the usual linear form

$$\eta_{it} = [Z_{it}, a'_i \otimes W_{it}] \begin{bmatrix} \beta \\ \theta \end{bmatrix}, \quad (7.4.3)$$

where the operator  $\otimes$  denotes the Kronecker product and the vector  $\theta$  corresponds to the vectorization of  $Q^{1/2}$ ,  $\theta = \text{vec}(Q^{1/2})$ . In the case of a scalar random effect  $a_i$  the Kronecker product simplifies to  $a_i W_{it}$ .

For illustration consider a simple binary response model with two-dimensional random effect  $(b_{i1}, b_{i2})$ . Let  $Q^{1/2} = (q_{ij})$  denote the root of  $Q$  and  $\text{vec}(Q^{1/2}) = (q_{11}, q_{21}, q_{12}, q_{22})$  denote the corresponding vectorized form. The linear predictor is given by

$$\eta_{it} = z'_{it}\beta + (w_{it1}, w_{it2}) Q^{1/2} \begin{pmatrix} a_{i1} \\ a_{i2} \end{pmatrix},$$

where  $(a_{i1}, a_{i2}) \sim N(0, I)$ . For the second term simple computation shows

$$\begin{aligned} (w_{it1}, w_{it2}) Q^{1/2} \begin{pmatrix} a_{i1} \\ a_{i2} \end{pmatrix} &= (a_{i1}, a_{i2}) \otimes (w_{it1}, w_{it2}) \text{ vec}(Q^{1/2}) \\ &= a_{i1} w_{it1} q_{11} + a_{i1} w_{it2} q_{21} + a_{i2} w_{it1} q_{12} + a_{i2} w_{it2} q_{22}. \end{aligned}$$

Since for the Cholesky factor  $Q^{1/2}$  we have  $q_{12} = 0$ , we may omit  $q_{12}$  in  $\text{vec}(Q^{1/2})$  and the corresponding column in  $(a_{i1}, a_{i2}) \otimes (w_{it1}, w_{it2})$ .

Defining the new parameter vector

$$\alpha' = (\beta', \theta'),$$

the reparameterized log-likelihood

$$l(\alpha) = \sum_{i=1}^n \log L_i(\alpha), \quad \text{with } L_i(\alpha) = \int \prod_{t=1}^{T_i} f(y_{it}|a_i; \alpha) g(a_i) da_i, \quad (7.4.4)$$

is obtained, where  $g$  denotes a density with zero mean and the identity matrix  $I$  as covariance matrix. Specifying  $g$  and maximizing  $l(\alpha)$  with respect to  $\alpha$  yields MLEs for  $\beta$  and  $Q^{1/2}$ . Note that we do not assume  $Q^{1/2}$  to be positive definite. As a consequence  $Q^{1/2}$  is not unique. Uniqueness could be achieved by requiring the diagonal elements of  $Q^{1/2}$  to be strictly positive. However, to avoid maximization under such constraints we do not restrict ourselves to positive definite roots. Note also that  $Q^{1/2}$  contains elements that are zero by definition. To obtain derivatives of  $l(\alpha)$  we first differentiate with respect to the entire parameter vector  $\theta$  and then delete entries that correspond to elements that are zero by definition in  $\theta$ .

Two approaches for maximizing (7.4.1) or (7.4.4) are in common use: The *direct* approach applies Gauss-Hermite or Monte Carlo integration techniques directly to obtain numerical approximations of the marginal likelihood or, more exactly, of its score function. Iterative procedures are then used for obtaining the maximum likelihood estimates. The *indirect* approach for maximizing the marginal likelihood is an EM algorithm, where conditional expectations in the E-step are computed by Gauss-Hermite or Monte Carlo integrations, and maximization in the M-step is carried out by Fisher scoring. The indirect maximization method seems to be numerically more stable and simpler to implement. However, it takes more computation time. The direct approach, which is based on the fitting of an appropriate GLM, is outlined in the following; details of the indirect approach are given in Section 7.4.3.

### Direct Maximization Using Fitting Techniques for GLMs

Let us first consider estimation of  $\alpha$  based on the reparameterized marginal likelihood (7.4.4). Direct maximization is based on solving the ML equations

$$s(\alpha) = \frac{\partial l(\alpha)}{\partial \alpha} = \sum_{i=1}^n \frac{\partial L_i(\alpha)/\partial \alpha}{L_i(\alpha)} = 0, \quad (7.4.5)$$

where  $s(\alpha)$  denotes the score function. The problem is that the marginal likelihood contribution  $L_i(\alpha)$  contains a possibly high-dimensional integral that cannot be solved analytically in general.

If the mixing density  $g$  is normal, evaluation of the integral in (7.4.4) can be accomplished by *Gauss-Hermite quadrature*, (e.g., Bock & Aitkin, 1981; Hinde, 1982; Anderson & Aitkin, 1985). This is feasible in practice for low-dimensional random effects. For simplicity we confine ourselves to scalar random effects so that the integral is one-dimensional and  $\theta$  in (7.4.3) represents a single variance component. For multivariate random effects Monte Carlo integration will be considered.

In the case of a scalar random effect the *Gauss-Hermite approximation* (see Appendix A.4) of the likelihood contribution  $L_i(\alpha)$  is given by

$$L_i(\alpha) \approx L_i^{GH}(\alpha) = \sum_{j=1}^m v_j f(y_i|d_j; \alpha), \quad (7.4.6)$$

where  $d_j$  denotes one of the  $j = 1, \dots, m$  quadrature points, and  $v_j$  represents the weight associated with  $d_j$ . Using the identity

$$\frac{\partial f(y_i|d_j; \alpha)}{\partial \alpha} = f(y_i|d_j; \alpha) \frac{\partial \log f(y_i|d_j; \alpha)}{\partial \alpha},$$

the score approximation

$$s(\alpha) \approx s^{GH}(\alpha) = \sum_{i=1}^n \frac{\partial \log L_i^{GH}(\alpha)}{\partial \alpha} = \sum_{i=1}^n \sum_{j=1}^m c_{ij}^{GH}(\alpha) \frac{\partial \log f(y_i|d_j; \alpha)}{\partial \alpha} \quad (7.4.7)$$

is obtained, where

$$c_{ij}^{GH}(\alpha) = \frac{v_j f(y_i|d_j; \alpha)}{\sum_{k=1}^m v_k f(y_i|d_k; \alpha)}, \quad \text{with} \quad \sum_{j=1}^m c_{ij}^{GH}(\alpha) = 1,$$

denote weight factors that depend on the parameters  $\alpha$  that are to be estimated. The derivative

$$\partial \log f(y_i|d_j, \alpha) / \partial \alpha' = (\partial \log f(y_i|d_j, \alpha) / \partial \beta', \partial \log f(y_i|d_j, \alpha) / \partial \theta')$$

corresponds to the score function of the GLM

$$E(\tilde{y}_{itj}) = h(\eta_{itj}), \quad \eta_{itj} = Z_{it}\beta + d_j \cdot W_{it}\theta, \quad (7.4.8)$$

for observations  $\tilde{y}_{itj}, t = 1, \dots, T_i, j = 1, \dots, m$ , where  $\tilde{y}_{itj} = y_{it}$ . That means the original  $T_i$  observations for cluster  $i$  become  $T_i m$  observations in

the corresponding GLM. The essential point in (7.4.8) is that the quadrature point  $d_j$  becomes another observed variable in the regression. Therefore, the components of  $\partial \log f / \partial \alpha$  have the usual GLM form

$$\frac{\partial \log f(y_i|d_j; \alpha)}{\partial \beta} = \sum_{t=1}^{T_i} Z'_{it} D_{it}(\alpha, d_j) \Sigma_{it}^{-1}(\alpha, d_j) (y_{it} - \mu_{it}(\alpha, d_j)), \quad (7.4.9)$$

$$\frac{\partial \log f(y_i|d_j; \alpha)}{\partial \theta} = \sum_{t=1}^{T_i} d_j W'_{it} D_{it}(\alpha, d_j) \Sigma_{it}^{-1}(\alpha, d_j) (y_{it} - \mu_{it}(\alpha, d_j)), \quad (7.4.10)$$

with  $D_{it}(\alpha, d_j) = \partial h(\eta_{itj}) / \partial \eta$ ,  $\Sigma_{it}(\alpha, d_j) = \text{cov}(y_{it}|d_j)$ , and  $\mu_{it}(\alpha, d_j) = h(\eta_{itj})$ .

If the number  $m$  of quadrature points is large enough, approximation (7.4.7) becomes sufficiently accurate. Thus, as  $n$  and  $m$  tend to infinity the MLEs for  $\alpha$  will be consistent and asymptotically normal under the usual regularity conditions. On the other hand, the number of quadrature points should be as small as possible to keep the numerical effort low. An alternative procedure that may reduce the number of quadrature points is adaptive Gauss-Hermite quadrature (Liu & Pierce, 1994; Pinheiro & Bates, 1995; Hartzel, Liu & Agresti, 2000). Adaptive quadrature is based on the log-likelihood (7.4.2); it first centers the modes with respect to the mode of the function being integrated and in addition scales them according to the curvature.

For high-dimensional integrals Gaussian quadrature techniques are less appropriate since the numerical effort increases exponentially with the dimension of the integral. *Monte Carlo techniques* are then more appropriate since these depend only linearly on the dimension. The simplest Monte Carlo approximation of  $L_i(\alpha)$  is given by

$$L_i(\alpha) \approx L_i^{MC}(\alpha) = \frac{1}{m} \sum_{j=1}^m f(y_i|d_{ij}; \alpha),$$

where the  $j = 1, \dots, m$  random values  $d_{ij}$  are i.i.d. drawings from the mixing density  $g$ . Since  $g$  is assumed to be completely known, these simulations are straightforward and the integrals are just replaced by empirical means evaluated from simulated values of the reparameterized cluster-specific effects  $a_i$ . Replacing  $L_i(\alpha)$  by  $L_i^{MC}(\alpha)$  in (7.4.5) yields

$$s(\alpha) \approx s^{MC}(\alpha) = \sum_{i=1}^n \sum_{j=1}^m c_{ij}^{MC}(\alpha) \frac{\partial \log f(y_i|d_{ij}; \alpha)}{\partial \alpha}, \quad (7.4.11)$$

where the weights  $c_{ij}^{MC}(\alpha)$  are given by

$$c_{ij}^{MC}(\alpha) = \frac{f(y_i|d_{ij}; \alpha)}{\sum_{k=1}^m f(y_i|d_{ik}; \alpha)}, \quad \text{with} \quad \sum_{j=1}^m c_{ij}^{MC}(\alpha) = 1.$$

The components of  $\partial \log f / \partial \alpha$  are defined by (7.4.9) and (7.4.10). Only quadrature points  $d_j$  are replaced by simulated values  $d_{ij}$ . In the general case where  $a_i$  is a vector, the quadrature points  $d_j = (d_{j1}, \dots, d_{js})$  are also vectors with quadrature points as components. Here  $j = (j_1, \dots, j_s)$  is a multiple index with  $j_i \in \{1, \dots, m\}$  for the case of  $m$  quadrature points in each component. Accordingly for Monte Carlo techniques  $d_{ij}$  are vector drawings from the mixing density  $g$ . Then in the linear predictor (7.4.8)  $d_j W_{it} \theta$  is replaced by  $d'_j \otimes W_{it} \theta$  or  $d'_{ij} \otimes W_{it} \theta$ , respectively, and in (7.4.10)  $d_j W'_{it}$  is replaced by  $(d'_j \otimes W_{it})'$  or  $(d'_{ij} \otimes W_{it})'$ , respectively.

The MLEs for  $\alpha$  have to be computed by an iterative procedure such as Newton-Raphson or Fisher scoring. Both algorithms imply calculation of the observed or expected information matrix. Due to the dependence of the weights  $c_{ij}^{GH}$ , resp.,  $c_{ij}^{MC}$ , on the parameters to be estimated, the analytical derivation of information matrices is very cumbersome. As an alternative one might calculate the observed information matrix by numerical differentiation of  $s^{GH}$ , resp.,  $s^{MC}$ . However, numerical inaccuracy often yields MLEs that depend heavily on starting values or do not converge. A direct maximization procedure for cumulative logit and probit models has been considered by Hedeker & Gibbons (1994).

In the examples considered later standard errors are based on the estimated Fisher matrix  $\sum_i s_i(\hat{\alpha}) s_i(\hat{\alpha})'$ , where  $s_i(\hat{\alpha})$  is the contribution of the  $i$ th observation to the score function (compare to Gourieroux & Monfort, 1989). Aitkin (1999) proposes alternatively to compute the standard error as the absolute value of the parameter estimate divided by the square root of the deviance change when the corresponding variable is omitted. Of course, this implies to fit a set of reduced models. The rationale behind this procedure is that the deviance change is approximately equal to the square of the parameter estimated divided by the standard error if the likelihood ratio test and Wald test for the significance of a parameter are equivalent (see also Dietz & Böning, 1995).

### Nonparametric Maximum Likelihood for Finite Mixtures

While Gauss-Hermite quadrature is designed for normally distributed random effects, Monte Carlo techniques may be used for any mixing density  $g$  that is specified in advance. A problem with this specification is that the choice of the mixing distribution may influence the parameter estimates. (e.g., Heckman & Singer, 1984). In order to avoid the specification of a parametric form of the mixing distribution, Aitkin & Francis (1998) and Aitkin (1999) suggested a nonparametric approach based on finite mixtures.

Consider again the Gauss-Hermite approximation of the likelihood that yields the weighted score function (7.4.7). The underlying log-likelihood has the form

$$l = \sum_{i=1}^n \log \left( \sum_j v_j f(y_i | d_j, \alpha) \right), \quad (7.4.12)$$

which may be seen as the likelihood of a mixture distribution with mixture proportions  $v_j$  at mass points  $d_j$ . In Gauss-Hermite quadrature mixture proportions and mass points are known. A different point of view is to regard (7.4.12) as the exact likelihood for the discrete mixing distribution and consider  $v_j$  and  $d_j$  as unknown. For the simple case of a univariate GLM with unidimensional random intercept, the fitted GLM in Gauss-Hermite quadrature approaches is given by

$$\mu_{itj} = h(\eta_{itj}), \quad \eta_{itj} = Z_{it}\beta + \sigma d_j,$$

where  $Z_{it}$  is a row vector. If the  $d_j$  are considered as mass points of a discrete mixing distribution that does not necessarily have unit variance (as is assumed in Gauss-Hermite quadrature), the standard deviation  $\sigma$  has to be omitted. One fits

$$\mu_{itj} = h(\eta_{itj}), \quad \eta_{itj} = Z_{it}\beta + d_j, \quad (7.4.13)$$

where the mass points  $d_j$  themselves determine the variance of the random intercept. For identifiability of the parameters  $d_j$  one has to be aliased with the intercept  $\beta_0$  or the intercept has to be removed from the model. The parameter  $\alpha' = (\beta', \theta')$  is thereby reduced to  $\beta$ .

The parameters to be estimated are  $\beta, v_j$ , and  $d_j$ ,  $j = 1, \dots, m$ . ML estimation is based on

$$\frac{\partial l}{\partial \beta}, \quad \frac{\partial l}{\partial d_j}, \quad \frac{\partial l}{\partial v_j}, \quad j = 1, \dots, m. \quad (7.4.14)$$

From (7.4.7) one obtains the derivatives  $\partial l / \partial \beta$  and  $\partial l / \partial d$ ,  $d' = (d_1, \dots, d_m)$ , in the form

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= \sum_{i=1}^n \sum_{j=1}^m c_{ij}(\beta, d_j) \sum_{t=1}^{T_i} Z'_{it} D_{it}(\beta, d_j) \Sigma_{it}^{-1}(\beta, d_j) (y_{it} - \mu_{it}(\beta, d_j)), \\ \frac{\partial l}{\partial d} &= \sum_{i=1}^n \sum_{j=1}^m c_{ij}(\beta, d_j) \sum_{t=1}^{T_i} \mathbf{1}_j D_{it} \Sigma_{it}^{-1}(\beta, d_j) (y_{it} - \mu_{it}(\beta, d_j)), \end{aligned}$$

where the weight is given by

$$c_{ij}(\beta, d_j) = \frac{v_j f(y_i | d_j, \beta)}{\sum_{k=1}^m v_k f(y_i | d_k, \beta)}$$

and  $\mathbf{1}_j$  is the  $j$ th unit vector. This is the (weighted) likelihood of model (7.4.13) for observations  $\tilde{y}_{itj} = y_{it}$ ,  $t = 1, \dots, T_i$ ,  $j = 1, \dots, m$ .

When considering  $\partial l / \partial v_j$  one has to take into account that for the  $m$  mass points one has  $\sum v_j = 1$  and therefore  $v_m = 1 - \sum_{j \neq m} v_j$ . Thus, the derivative of the log-likelihood (7.4.12) is given by

$$\frac{\partial l}{\partial v_j} = \sum_{i=1}^n \frac{f(y_i|d_j, \beta) - f(y_i|d_m, \beta)}{\sum_s v_s f(y_i|d_s, \beta)} = \sum_{i=1}^n \left( \frac{c_{ij}}{v_j} - \frac{c_{im}}{v_m} \right).$$

The equation  $\partial l / \partial v_j = 0$ ,  $j = 1, \dots, m-1$ , yields

$$\hat{v}_j = \frac{1}{T} \sum_{i=1}^n c_{ij}, \quad T = T_1 + \dots + T_n. \quad (7.4.15)$$

When solving  $\partial l / \partial \beta$  by Fisher scoring, one has to update the mass points  $v_j$  by equation (7.4.15). Instead of direct maximization, the indirect method based on the EM algorithm is used (see Section 7.4.3). The latter has the advantage that the weights  $c_{ij}$  no longer depend on the parameter  $\beta$ . In the nonparametric approach the estimation of the mass points is not to be understood as a valid estimate of the mixing distribution. The finite mixture distribution is only considered as nuisance parameters which might be estimated poorly (see Aitkin, 1999).

## 7.4.2 Posterior Mean Estimation of Random Effects

Predicted values of the random effects  $b_i$  are required to calculate fitted values of the conditional means  $\mu_{it}$ . Since the cluster-specific effects  $b_i$  represent random variables, estimation is based on the posterior density of  $b_i$  given the observations  $Y = (y_1, \dots, y_n)$ . Due to the independence assumptions the posterior of  $b_i$  depends only on  $y_i$ . Using Bayes' theorem the posterior

$$f(b_i|y_i; \beta, Q) = \frac{f(y_i|b_i; \beta)p(b_i; Q)}{\int f(y_i|b_i; \beta)p(b_i; Q)db_i} \quad (7.4.16)$$

is obtained, where  $f(y_i|b_i, \beta) = \prod_{t=1}^{T_i} f(y_{it}|b_i, \beta)$ .

Since the parameters  $\beta$  and  $Q$  are not known, they are replaced by some consistent estimators  $\hat{\beta}$  and  $\hat{Q}$  such that statistical inference on  $b_i$  is based on the empirical Bayes principle. The empirical Bayesian point estimator that is "best" in the mean square error is the posterior mean

$$b_i^m = E(b_i|y_i) = \int b_i f(b_i|y_i; \hat{\beta}, \hat{Q}) db_i.$$

For confidence intervals the posterior covariance matrix

$$V_i^m = \text{cov}(b_i|y_i) = E(b_i b_i' | y_i) - b_i (b_i)'$$

is also required. To calculate the posterior mean  $b_i^m$  or the posterior covariance  $V_i^m$  one has to carry out integrations that have the common structure

$$S(q(b_i)) = \int q(b_i) f(y_i|b_i; \hat{\beta}, \hat{Q}) p(b_i; \hat{Q}) db_i, \quad (7.4.17)$$

where  $q(b_i)$  stands for 1,  $b_i$ , or  $b_i b_i'$ . Unfortunately, the integrals cannot be solved analytically for most situations. Therefore, some approximation technique is required, e.g., numerical or Monte Carlo integration. Details are given in Section 7.4.3.

### 7.4.3 Indirect Maximization Based on the EM Algorithm\*

For indirect maximization of the log-likelihood (7.4.4), we consider an EM algorithm, which is described in its general form in Appendix A.3. If  $Y = (y_1, \dots, y_n)$  denotes the incomplete data, which are observable, and  $A = (a_1, \dots, a_n)$  stands for the unobservable data, which represent the re-parameterized random effects, the EM algorithm is based on the complete data log-density

$$\log f(Y, A; \alpha) = \sum_{i=1}^n \log f(y_i|a_i; \alpha) + \sum_{i=1}^n \log g(a_i). \quad (7.4.18)$$

In the E-step of the  $(p+1)$ th EM-cycle, one has to determine

$$M(\alpha|\alpha^{(p)}) = E\{\log f(Y, A; \alpha)|Y; \alpha^{(p)}\} = \int \log(f(Y, A; \alpha)) f(A|Y; \alpha^{(p)}) dA,$$

which represents the conditional expectation of (7.4.18), given the incomplete data  $Y$  and an estimate  $\alpha^{(p)}$  from the previous EM-cycle. The density  $f(A|Y; \alpha^{(p)})$  denotes the posterior

$$f(A|Y; \alpha^{(p)}) = \frac{\prod_{i=1}^n f(y_i|a_i; \alpha^{(p)}) \prod_{i=1}^n g(a_i)}{\prod_{i=1}^n \int f(y_i|a_i; \alpha^{(p)}) g(a_i) da_i}, \quad (7.4.19)$$

which is obtained after applying Bayes' theorem in connection with the independence assumption (7.2.4). Due to (7.4.18) and (7.4.19) the function  $M(\alpha|\alpha^{(p)})$  simplifies to

$$M(\alpha|\alpha^{(p)}) = \sum_{i=1}^n k_i^{-1} \int [\log f(y_i|a_i; \alpha) + \log g(a_i)] f(y_i|a_i; \alpha^{(p)}) g(a_i) da_i,$$

where

$$k_i = \int f(y_i|a_i; \alpha^{(p)}) g(a_i) da_i$$

is independent from the parameters  $\alpha$  and the reparameterized random effects  $a_i$ . In general, carrying out the E-step is problematic because the integrals cannot be solved analytically. Therefore, numerical or Monte Carlo integration is required. Using a simple Monte Carlo integration, as in (7.4.11), yields the approximation

$$M(\alpha|\alpha^{(p)}) \approx M^{MC}(\alpha|\alpha^{(p)}) = \sum_{i=1}^n \sum_{j=1}^m c_{ij}^{MC} [\log f(y_i|d_{ij}; \alpha) + \log g(d_{ij})], \quad (7.4.20)$$

where the  $j = 1, \dots, m$  weight factors

$$c_{ij}^{MC} = \frac{f(y_i|d_{ij}; \alpha^{(p)})}{\sum_{k=1}^m f(y_i|d_{ik}; \alpha^{(p)})}, \quad \text{with} \quad \sum_{j=1}^m c_{ij}^{MC} = 1, \quad (7.4.21)$$

are completely known and do not depend on the parameters  $\alpha$ . As in (7.4.11) the vectors  $d_{ij}$  are i.i.d. drawings from the mixing density  $g$ .

For normal densities  $g$  the integrals can also be approximated by *Gauss-Hermite quadrature*; see Hinde (1982), Brillinger & Preisler (1983), Anderson & Hinde (1988), and Jansen (1990) for scalar random effects, Anderson & Aitkin (1985) and Im & Gianola (1988) for bivariate random effects, Tutz & Hennevogl (1996) for ordinal models. Using *Gauss-Hermite quadrature* yields the approximation

$$M(\alpha|\alpha^{(p)}) \approx M^{GH}(\alpha|\alpha^{(p)}) = \sum_{i=1}^n \sum_j c_{ij}^{GH} [\log f(y_i|d_j; \alpha) + \log g(d_j)] \quad (7.4.22)$$

with the weight factors

$$c_{ij}^{GH} = \frac{v_j f(y_i|d_j; \alpha^{(p)})}{\sum_k v_k f(y_i|d_k; \alpha^{(p)})}, \quad \sum_j c_{ij}^{GH} = 1,$$

where the sum that is substituted for the  $s$ -dimensional integral is over the multiple index  $j = (j_1, \dots, j_s)$ ,  $j_i \in \{1, \dots, m\}$ ,  $i = 1, \dots, s$ , or the multiple index  $k = (k_1, \dots, k_s)$ , respectively. Instead of drawings  $d_{ij}$  one has the transformed quadrature points  $d_j = (d_{j_1}, \dots, d_{j_s})$  and the weights  $v_j = v_{j_1} \cdots v_{j_s}$ . Again the weight factors  $c_{ij}^{GH}$  do not depend on the parameter  $\alpha$ .

For  $m$  quadrature points in each dimension the sum is over  $m^s$  terms. Therefore, the numerical effort of Gaussian quadrature techniques increases exponentially with the integral dimension. For high-dimensional integrals

Monte-Carlo techniques should be preferred since the sum in (7.4.20) is only over  $m$  terms.

The M-step consists of maximizing  $M(\alpha|\alpha^{(p)})$  with respect to  $\alpha$ . Considering a Monte Carlo approximation of the form (7.4.20), this is equivalent to solving the equation

$$u(\alpha|\alpha^{(p)}) = \frac{\partial M^{MC}(\alpha|\alpha^{(p)})}{\partial \alpha} = \sum_{i=1}^n \sum_{j=1}^m c_{ij}^{MC} \frac{\partial \log f(y_i|d_{ij}; \alpha)}{\partial \alpha} = 0, \quad (7.4.23)$$

where the components of  $\partial \log f / \partial \alpha = (\partial \log f / \partial \beta, \partial \log f / \partial \theta)$  correspond to (7.4.9) and (7.4.10) with the exception that the quadrature points  $d_j$  are replaced by the simulated values  $d_{ij}$ . Note the similarity between (7.4.23) and the direct maximization equation (7.4.11). Both equations differ only in the weight factors, which are now independent from the parameters to be estimated and, therefore, completely known. Additionally the derivative  $\partial \log f / \partial \alpha$  is equivalent to the score function of the GLM

$$E(\tilde{y}_{itj}) = h(\eta_{itj}), \quad \eta_{itj} = Z_{it}\beta + d'_{ij} \otimes W_{it}\theta,$$

with observations  $\tilde{y}_{it1} = \tilde{y}_{it2} = \dots = \tilde{y}_{itm} = y_{it}$ . Therefore,  $u(\alpha|\alpha^{(p)})$  can be interpreted as a weighted score function of such a GLM. Then, given the weights  $c_{ij}^{MC}$ , the solution of (7.4.23) corresponds to a weighted MLE, which can be computed by an iteratively weighted least-squares or Fisher-scoring algorithm. For given weights  $c_{ij}^{MC}$  the expected conditional information matrix has the form

$$U(\alpha|\alpha^{(p)}) = -E \left( \frac{\partial^2 E^{MC}(\alpha|\alpha^{(p)})}{\partial \alpha \partial \alpha'} \right) = \sum_{i=1}^n \sum_{j=1}^m c_{ij}^{MC} F_{ij}(\alpha),$$

where  $F_{ij}(\alpha)$  is partitioned into

$$F_{ij}(\alpha) = \begin{bmatrix} F_{ij}^{\beta\beta} & F_{ij}^{\beta\theta} \\ F_{ij}^{\theta\beta} & F_{ij}^{\theta\theta} \end{bmatrix},$$

with

$$\begin{aligned} F_{ij}^{\beta\beta} &= \sum_{t=1}^{T_i} Z'_{it} D_{it} \Sigma_{it}^{-1} D'_{it} Z_{it}, \\ F_{ij}^{\beta\theta} &= (F_{ij}^{\theta\beta})' = \sum_{t=1}^{T_i} Z'_{it} D_{it} \Sigma_{it}^{-1} D'_{it} (d'_{ij} \otimes W_{it}), \\ F_{ij}^{\theta\theta} &= \sum_{t=1}^{T_i} (d'_{ij} \otimes W_{it})' D_{it} \Sigma_{it}^{-1} D'_{it} (d'_{ij} \otimes W_{it}), \end{aligned}$$

and  $D_{it} = D_{it}(\alpha, d_{ij}) = \partial h(\eta_{itj})/\partial\eta$ ,  $\Sigma_{it} = \Sigma_{it}(\alpha, d_{ij}) = \text{cov}(y_{it}|d_{ij})$ ,  $\mu_{it}(\alpha, d_{ij}) = h(\eta_{itj})$ . Then the M-step consists of the iteration scheme

$$\alpha_{k+1} = \alpha_k + U^{-1}(\alpha_k|\alpha^{(p)})u(\alpha_k|\alpha^{(p)}), \quad (7.4.24)$$

with starting value  $\alpha_0 = \alpha^{(p)}$  and  $k$  as an iteration index.

The following is a sketch of the Monte Carlo version of the EM algorithm (MCEM) with a Fisher-scoring algorithm embedded in each M-step for estimating the fixed parameters of a generalized linear random effects model:

1. Calculate initial values  $\beta^{(0)}$  using an original GLM without random effects, initialize  $\theta^{(0)}$  with an arbitrary starting value, e.g.,  $\theta^{(0)} = 0$ , and set  $\alpha^{(0)} = (\beta^{(0)}, \theta^{(0)})$ .

For  $p = 0, 1, 2, \dots$

2. Approximate  $M(\alpha|\alpha^{(p)})$  by  $M^{MC}(\alpha|\alpha^{(p)})$ ,  
i.e., compute the weights (7.4.21) using the values  $\alpha^{(p)}$ .
3. Carry out the algorithm (7.4.24) to obtain updates  $\alpha^{(p+1)}$ .

In complex designs the difficult step in the algorithm is the  $M$ -step, where the approximation  $M^{MC}(\alpha|\alpha^{(p)})$  is computed. In McCulloch's (1994, 1997) version of the MCEM algorithm the drawings are not independent but from a Markov chain constructed using a Metropolis algorithm (see Appendix A.5). Instead of simple drawings from the mixing density, importance sampling (see Appendix A.5) may be used to improve the approximation (e.g., Booth & Hobert, 1999).

As for every EM algorithm, convergence of the MCEM algorithm is slow. Moreover, computing time depends on  $m$ , the number of simulations. Therefore,  $m$  should be chosen as small as possible. However, to make sure that the MLEs do not depend on  $m$ , an adaptive procedure should be used. That means the number of simulations should be increased successively with increasing EM-cycles. The same has been pointed out by several authors for a Gauss-Hermite version of the EM algorithm; see, e.g., Hinde (1982), Brillinger & Preisler (1983), and Jansen (1990). However, ad-hoc methods for increasing  $m$  may perform rather unsatisfactorily. Booth & Hobert (1999) investigated the problem by explicitly considering the Monte Carlo error, which is connected to the Monte Carlo approximation of the integral. Their method links the choice of the Monte Carlo sample size  $m$  to the Monte Carlo error, thus avoiding a waste of iterations, which occurs for large values of  $m$  at the start and for large Monte Carlo errors in steps close to the optimal value.

An estimator for the asymptotic covariance of the MLEs  $\hat{\alpha}$  is not provided by EM. However, it can be obtained in the same way as in the direct maximization routine, e.g., by numerical differentiation of (7.4.11) with respect to  $\alpha$  and evaluation at  $\hat{\alpha}$ .

Nonparametric maximum likelihood estimation based on the EM algorithm may be performed in the same way with some slight modifications.

The unidimensional model to be fitted is given in (7.4.13) and has the form

$$\mu_{itj} = h(\eta_{itj}), \quad \eta_{itj} = Z_{it}\beta + d_j = \tilde{Z}_{it}\delta,$$

where  $\tilde{Z}_{it} = (Z_{it}, 1'_j)$  with  $1_j$  denoting the  $j$ th unit vector and  $\delta' = (\beta', d_1, \dots, d_m)$ . The conditional information matrix is given by

$$U(\delta|\delta^{(p)}) = \sum_{i=1}^n \sum_{j=1}^m c_{ij} \sum_{t=1}^T \tilde{Z}_{it} D_{it} \Sigma_{it}^{-1} D_{it}' Z_{it}.$$

Moreover,

$$c_{ij} = \frac{v_j f(y_i|d_j, \delta^{(p)})}{\sum_k v_k f(y_i|d_k, \delta^{(p)})}$$

has to be supplemented by  $v_j = \sum_{i=1}^n c_{ij}/T$  from (7.4.15) (see Aitkin, 1999).

#### 7.4.4 Algorithmic Details for Posterior Mean Estimation\*

Next we discuss methods for carrying out the integrations necessary for posterior mean estimation of random effects in Section 7.4.2.

If the random effects are normally distributed, Gauss-Hermite quadrature may be applied. To make sure that the integration variables are orthogonal in the random effects density, the random effects  $b_i$  are replaced by the Cholesky parameterization

$$b_i = \hat{Q}^{1/2} a_i,$$

which was proposed in the previous section. Applying the Gauss-Hermite quadrature as reported in Appendix A.4 yields the approximation

$$S(q(b_i)) \approx \sum_j v_j q(\bar{b}_j) f(y_i|\bar{b}_j; \hat{\beta}, \hat{Q}),$$

where the sum is over the multiple index  $j = (j_1, \dots, j_s)$ ,  $j_i \in \{1, \dots, m\}$ ,  $i = 1, \dots, s$ , and

$$\hat{b}_j = (\hat{b}_{j_1}, \dots, \hat{b}_{j_s}), \quad \text{with} \quad \hat{b}_{j_r} = \hat{Q}^{1/2} d_{j_r},$$

is a vector of the transformed quadrature points  $d_{j_r}$ ,  $r = 1, \dots, s$ , and

$$v_j = \prod_{r=1}^s v_{j_r}, \quad \text{with} \quad \sum_j v_j = 1,$$

is the product of the  $r = 1, \dots, s$  weights  $v_{j_r}$  each associated with the quadrature point  $d_{j_r}$ . To make sure that the resulting approximations of  $b_i^m$ , resp.,  $V_i^m$ , do not depend on the number of quadrature points, the calculations should be repeated with an increasing number of quadrature points until approximations based on different  $m$ 's are nearly equal.

If the random effects density is not normal, Gauss-Hermite quadrature may also be used. Then, as stated in Appendix A.4, we have to match a normal density  $f_N$  with mean  $b_i^m$  and covariance matrix  $V_i^m$  to the posterior (7.4.16), so that

$$S(q(b_i)) = \int \frac{q(b_i) f(y_i|b_i; \hat{\beta}) p(b_i; \hat{Q})}{f_N(b_i; b_i^m, V_i^m)} f_N(b_i; b_i^m, V_i^m) db_i.$$

To determine the unknown parameters  $b_i^m$  and  $V_i^m$ , Gauss-Hermite integration has to be applied iteratively with some starting values for  $b_i^m$  and  $V_i^m$ . A natural choice is the posterior mode  $\hat{b}_i$  and the curvature  $\hat{V}_i$ , which can be computed along the lines of Section 7.3.

An attractive alternative to Gauss-Hermite quadrature is Monte Carlo integration, especially when the dimension of the integral to be approximated is high. The integrals (7.4.17) are then estimated by empirical means, which are obtained from replacing the unobservable random effects by simulated values that are i.i.d. drawings from the random effects density. However, if the cluster size  $T_i$  is small, a large number of drawings is required to obtain consistent estimates of  $b_i^m$  and  $V_i^m$ . More efficient integral estimates, which require fewer drawings, are obtained applying *importance* or *rejection sampling*. Both sampling techniques, which are described in their general form in Appendix A.5, match a density  $g$  to the posterior (7.4.16), where  $g$  should be proportional to the posterior, at least approximately, and it should be easy to sample from  $g$ .

Zellner & Rossi (1984), for example, suggest *importance sampling* (compare to Appendix A.5), which is based on a normal “importance function”  $g$  with the posterior mode  $\hat{b}_i$  as mean and the posterior curvature  $\hat{V}_i$  as covariance matrix. They justify such a choice asymptotically as the cluster size  $T_i$  tends to infinity. Then the posterior becomes normal, so that posterior mode and mean and posterior curvature and covariance coincide asymptotically. The corresponding estimator for integral (7.4.17) is

$$\hat{S}(q(b_i)) = \frac{1}{m} \sum_{j=1}^m q(d_{ij}) w_{ij}, \quad \text{with} \quad w_{ij} = \frac{f(y_i|d_{ij}, \hat{\beta}) p(d_{ij}; \hat{Q})}{g(d_{ij}; \hat{b}_i, \hat{V}_i)},$$

where the simulated values  $d_j = (d_{j1}, \dots, d_{js})$  are i.i.d. drawings from the normal density  $g$  and the resulting estimator for  $b_i^m$  has the form

$$b_i^m = \hat{S}(q(b_i))/\hat{S}(1) = \sum_{j=1}^m d_j c_{ij},$$

$$\text{with } c_{ij} = w_{ij} / \sum_{k=1}^m w_{ik}, \quad \text{and} \quad \sum_{j=1}^m c_{ij} = 1.$$

The quality of this estimator heavily depends on the weights  $c_{ij}$ . If one or a few of them are extreme relative to the others, the estimator  $\hat{b}_i$  is dominated by a single or a few values  $d_j$ . Such a phenomenon especially occurs when the cluster sizes  $T_i$  are small. Then the chosen importance function  $g$  is apparently less appropriate, and importance functions being multimodal are recommended (see, e.g., Zellner & Rossi, 1984).

*Rejection sampling* (see Appendix A.5) has been used by Zeger & Karim (1991) in a Gibbs sampling approach to random effects models. They match a normal density  $g$  with mean  $\hat{b}_i$  and covariance matrix  $c_2 \hat{V}_i$  to the posterior (7.4.16), where the parameter  $c_2 \geq 1$  blows up  $\hat{V}_i$ . The parameters  $\hat{b}_i$  and  $\hat{V}_i$  again denote the posterior mode and curvature, which can be estimated along the lines of Section 7.3.

The parameter  $c_2$  and a further parameter  $c_1$  have to be chosen so that the inequality

$$c_1 g(b_i; \hat{b}_i, c_2 \hat{V}_i) \geq f(b_i | y_i; \hat{\beta}, \hat{Q}) \quad (7.4.25)$$

holds for all values of  $b_i$ . Zeger & Karim (1991) recommend setting  $c_2 = 2$  and choosing  $c_1$  so that the ordinates at the common mode  $\hat{b}_i$  of the density  $g$  and the posterior kernel are equal, e.g.,

$$c_1 = \frac{f(y_i | \hat{b}_i; \hat{\beta}) p(\hat{b}_i; \hat{Q})}{g(b_i; \hat{b}_i, c_2 \hat{V}_i)}.$$

If (7.4.25) is fulfilled, a simulated value  $d_k$  drawn from  $g$  behaves as is distributed according to the posterior (7.4.16) as long as  $d_k$  is not rejected. Let  $u_k$  denote a random number being  $[0,1]$ -distributed; then  $d_k$  is rejected if

$$\frac{f(y_i | d_k; \hat{\beta}) p(d_k; \hat{Q})}{c_1 g(d_k; \hat{b}_i, c_2 \hat{V}_i)} > u_k.$$

Having generated  $j = 1, \dots, m$  accepted drawings  $d_j$ , the posterior mean and the posterior covariance are estimated by

$$b_i^m = \frac{1}{m} \sum_{j=1}^m d_j \quad \text{and} \quad V_i^m = \frac{1}{m} \sum_{j=1}^m (d_j - \hat{b}_i)(d_j - \hat{b}_i)'.$$

Moreover, the  $j = 1, \dots, m$  simulated values  $d_j$  can be used to construct the empirical posterior distribution. Zeger & Karim (1991) propose a fully Bayesian approach, by treating  $\beta, Q$  as random with appropriate priors and estimating them together with random effects by Gibbs sampling.

## 7.5 Examples

**Example 7.3: Ohio children data** (Example 7.1, continued)

Analysis is based on the random effects logit model

$$\log \frac{P(\text{infection})}{P(\text{no infection})} = b_i + \beta_0 + \beta_S x_S + \beta_{A1} x_{A1} + \beta_{A2} x_{A2} + \beta_{A3} x_{A3} + \beta_{SA1} x_S x_{A1} + \beta_{SA2} x_S x_{A2} + \beta_{SA3} x_S x_{A3},$$

where  $x_S$  stands for smoking status in effect coding ( $x_S = 1$ : smoking,  $x_S = -1$ : nonsmoking) and  $x_{A1}$ ,  $x_{A2}$ ,  $x_{A3}$  represent age in effect coding. The parameters  $\beta_0, \beta_S, \beta_{A1}, \beta_{A2}, \beta_{A3}, \dots$  are fixed effects and  $b_i$  stands for a random intercept, where  $b_i \sim N(0, \sigma^2)$ . Apart from the random intercept, the linear predictor is the same as considered in Chapter 6 (Example 6.4).

**Table 7.2.** Random intercept logit model for Ohio children data (effect coding of smoking and age, standard deviations in parentheses)

	Fixed effects	Gauss-Hermite $m = 10$	Monte Carlo $m = 10$	EM-type
$\hat{\beta}_0$	-1.696 (0.062)	-2.797 (0.205)	-2.292 (0.276)	-1.952
$\hat{\beta}_S$	0.136 (0.062)	0.189 (0.110)	0.219 (0.152)	0.140
$\hat{\beta}_{A1}$	0.059 (0.106)	0.088 (0.186)	0.077 (0.260)	0.068
$\hat{\beta}_{A2}$	0.156 (0.103)	0.245 (0.189)	0.213 (0.268)	0.186
$\hat{\beta}_{A3}$	0.066 (0.105)	0.101 (0.191)	0.088 (0.263)	0.078
$\hat{\beta}_{SA1}$	-0.115 (0.106)	-0.178 (0.186)	-0.155 (0.260)	-0.135
$\hat{\beta}_{SA2}$	0.069 (0.103)	0.114 (0.189)	0.099 (0.268)	0.085
$\hat{\beta}_{SA3}$	0.025 (0.105)	0.041 (0.191)	0.035 (0.263)	0.030
$\hat{\sigma}$	— —	2.136 (0.203)	1.817 (0.238)	1.830

Table 7.2 shows the estimation results for the fixed effects model and several estimation procedures for the random effects model. Standard deviations are given in brackets. Following a proposal by Gourieroux & Monfort (1989) standard errors are based on the estimated Fisher matrix  $\sum_i s_i(\hat{\alpha}) s_i(\hat{\alpha})'$ , where  $s_i(\hat{\alpha})$  is the contribution of the  $i$ th observation to the score function (approximated by Gauss-Hermite or Monte Carlo methods).

**Table 7.3.** ML estimates for Ohio children data based on finite mixtures

	$m = 10$		$m = 4$	
$\hat{\beta}_S$	0.167	(0.119)	0.149	(0.109)
$\hat{\beta}_{A1}$	0.091	(0.134)	0.087	(0.133)
$\hat{\beta}_{A2}$	0.245	(0.131)	0.244	(0.131)
$\hat{\beta}_{A3}$	0.103	(0.133)	0.102	(0.133)
$\hat{\beta}_{SA1}$	-0.178	(0.132)	-0.176	(0.132)
$\hat{\beta}_{SA2}$	0.111	(0.130)	0.114	(0.130)
$\hat{\beta}_{SA3}$	0.040	(0.133)	0.042	(0.132)

The estimated effects for the random effects models are larger than for the fixed effect model. This is to be expected because there is a bias of estimates toward zero if the random component is falsely omitted. Comparison of the estimates for the random effects models shows that the Gauss-Hermite procedure yields slightly stronger effects (except for smoking status) and a higher value for the heterogeneity parameter. The estimated standard deviation of the random intercept is quite large; thus heterogeneity should not be ignored.

As far as the inference of smoking is concerned, the same conclusions must be drawn as in Example 6.4. For the fixed effects model the smoking effect seems to be significant. However, the standard deviation is not trustworthy since independent observations are falsely assumed. In the random effects model the smoking effect is not significant, but the estimate is positive signaling a tendency toward increased infection rates. The interpretation of the other effects is similar to Example 6.4.

However, it should be noted that the estimates for random effects models are different from the estimates based on the naive independence assumption. For marginal models the point estimates (in this example) are stable for differing correlation assumptions (see Table 6.8).

For comparison, the nonparametric ML estimates for finite mixtures are shown in Table 7.3 for a mixture based on  $m = 4$  and  $m = 10$  mass points. The estimates are quite similar to the estimates based on the Gauss-Hermite approach.  $\square$

#### **Example 7.4: Bitterness of white wines** (Example 7.2, continued)

Each of the judges tasted eight wines varying with respect to “temperature” ( $x_T = 1$ : low,  $x_T = 0$ : high), “contact” ( $x_C = 1$ : yes,  $x_C = 0$ : no), and “bottle” (first/second). Since “bottle” is not influential, it is omitted in the analysis.

In the first step a fixed effects cumulative logistic model with thresholds  $\alpha_1, \dots, \alpha_4$  may be used:

$$\begin{aligned}
P(y_{it} = 1) &= F(\alpha_1 + \beta_T x_T + \beta_C x_C), \\
P(y_{it} \leq r) &= F(\alpha_1 + \sum_{s=2}^r \exp(\alpha_s) + \beta_T x_T + \beta_C x_C), \quad r = 2, \dots, 4.
\end{aligned} \tag{7.5.1}$$

The covariate effect  $\beta_T$  represents the influence of “low temperature” and  $\beta_C$  the influence of “contact with skin.”

**Table 7.4.** Estimation results of bitterness-of-wine data

	Fixed effects model	Random effects		
		MCEM 10	GHEM 10	EM-type
$\alpha_1$	-5.289 (0.0)	-6.479	-6.496	-6.315
$\alpha_2$	0.974 (0.0)	1.195	1.177	1.150
$\alpha_3$	0.739 (0.0)	0.977	0.956	0.931
$\alpha_4$	0.426 (0.105)	0.631	0.616	0.596
$\beta_T$	2.373 (0.0)	2.994	2.947	2.867
$\beta_c$	1.571 (0.0)	1.925	1.901	1.844
$\sigma^2$	—	1.157	1.496	1.512
log-likelihood	-87.31	-79.86	-82.064	—

Table 7.4 gives the MLEs with corresponding  $p$ -values. The deviance has value 15.87 at 26 df. So, at first glance the model fit looks quite good. However, the goodness-of-fit test is based on the assumption of independent responses  $y_{it}$ . For the responses  $y_{it}$ ,  $t = 1, \dots, 8$ , of the  $i$ th judge such an assumption does not hold. A model that allows each judge to have a specific level for the bitterness of wine includes a random effect varying across judges. In parameterization (7.5.1)  $\alpha_1$  sets the level, whereas  $\alpha_2, \dots, \alpha_r$  are added for higher categories. Thus, the level for judge  $i$  is specified by the addition of a random effect to  $\alpha_1$  in the form

$$P(y_{it} = 1 | \alpha_{1i}) = F(a_i + \alpha_1 + \beta_T x_T + \beta_C x_C), \tag{7.5.2}$$

$$P(y_{it} \leq r | \alpha_{1i}) = F(a_i + \alpha_1 + \sum_{s=2}^r \exp(\alpha_s) + \beta_T x_T + \beta_C x_C), \quad r = 2, \dots, 4,$$

with

$$a_i \stackrel{iid}{\sim} N(0, \sigma^2).$$

We use a maximum likelihood procedure based on the EM algorithm as proposed in Section 7.3 and a 10-point Gauss-Hermite variant of the EM

algorithm and a Monte Carlo version of the EM algorithm with 10 simulations per response  $y_{it}$ . The results are given in Table 7.4. The MLEs of the cumulative logistic model (7.5.1) given in column 1 of Table 7.4 have been used as starting values for the thresholds  $\alpha_1, \dots, \alpha_4$  and for  $\beta_T$  and  $\beta_C$ . In comparison with the cumulative logistic model (7.5.1) the threshold  $\alpha_1$  and the threshold differences  $\alpha_2, \dots, \alpha_4$  and the covariate effects  $\beta_T$  and  $\beta_C$  have changed in the random effects model (7.5.2). As was to be expected  $\beta_T$  and  $\beta_C$  are larger than for the fixed effects model. Moreover, the relatively high estimates of the variance component  $\sigma^2$  indicate heterogeneity across judges. The results are quite similar for Monte Carlo procedure, Gauss Hermite integration, and the empirical Bayes estimate. The data set has been further investigated by Tutz & Hennevogl (1996).

## 7.6 Bayesian Mixed Models

Sections 7.3 and 7.4 described empirical Bayes approaches for mixed models, with unknown fixed effect parameters and variance parameters, but a mixing prior distribution for random effects. In a fully Bayesian setting all parameters are regarded as random.

### Bayesian Generalized Mixed Models

The structural assumption (7.2.2) is now conditional on  $\beta$  and  $b_i$ ,

$$E(y_{it}|\beta, b_i) = h(Z_{it}\beta + W_{it}b_i). \quad (7.6.1)$$

We assume here that random effects  $b_i$  are i.i.d. Gaussian  $r$ -dimensional random variables

$$p(b_i|Q) \sim N(0, Q),$$

but other priors, like mixtures of normals, can also be handled (possibly supplemented by additional mixing parameters). As in Section 2.3, common choices for the prior  $p(\beta)$  are normal distributions or flat, noninformative priors. A standard noninformative prior for the hyperparameter  $Q$  is Jeffreys prior; see Zeger & Karim (1991). However, this choice can lead to improper posteriors; see Hobert & Casella (1996). Therefore, Besag, Green, Higdon & Mengersen (1995) recommend proper but highly dispersed inverted Wishart priors  $Q \sim IW_r(\xi, \Psi)$ , i.e.,

$$p(Q) \propto |Q|^{-\xi-(r+1)/2} \exp(-tr(Q\Psi^{-1})),$$

with carefully selected hyperparameters  $\xi$  and  $\Psi$ . A simpler choice is to impose independent inverse gamma priors to the components of  $b_i$ . Assuming

conditional independence among response variables  $y_{it}|b_i, \beta$ , random effects  $b_i|Q$ , regression parameters  $\beta$ , and the hyperparameter  $Q$ , the posterior distribution can be expressed as

$$p(\beta, b_1, \dots, b_n, Q|Y) \propto \prod_{i=1}^n \prod_{t=1}^{T_i} f(y_{it}|\beta, b_i) p(\beta) \prod_{i=1}^n p(b_i|Q) p(Q).$$

Full conditionals  $p(\beta|\cdot)$ ,  $p(b_i|\cdot)$ ,  $p(D|\cdot)$ , given the data and the rest of parameters “.”, simplify to

$$\begin{aligned} p(\beta|\cdot) &\propto \prod_{i=1}^n \prod_{t=1}^{T_i} f(y_{it}|\beta, b_i) p(\beta), \\ p(b_i|\cdot) &\propto \prod_{t=1}^{T_i} f(y_{it}|\beta, b_i) p(b_i|Q), \\ p(Q|\cdot) &\propto \prod_{i=1}^n f(b_i|Q) p(Q). \end{aligned}$$

The full conditional  $p(Q|\cdot)$  is again inverted Wishart with updated parameters  $\xi + n/2$  and  $\Psi + \frac{1}{2} \sum_{i=1}^n b_i b_i'$ . Standard algorithms for drawing from inverted Wishart distributions (e.g., Ripley, 1987) allow direct implementation of Gibbs updating steps. Zeger & Karim (1991) used Gibbs steps with rejection sampling for updating  $\beta$  and the  $b_i$ 's. Their algorithm involves Fisher scoring and rejection sampling for each updating step. A computationally more efficient noniterative procedure is the weighted least-squares proposal of Gamerman (1997a), which makes only one Fisher scoring step to construct a specific MH proposal for updating  $\beta$  and  $b_i$ ; compare to Section 2.3.2. Alternatively, updating by Metropolis random walk proposals is a simple choice. However, careful tuning of the spread of proposal distributions is necessary.

As an alternative to Gaussian random effects priors, scale mixtures of normals are proposed; see, e.g., Besag, Green, Higdon & Mengersen (1995) and Knorr-Held (1997). This class includes Student-, Laplace-, and other non-Gaussian, symmetric distributions. They are easily incorporated into mixed models, since they are defined hierarchically as well.

## Generalized Additive Mixed Models

We have restricted discussion to uncorrelated random effects. Clayton (1996) describes more general models, including autocorrelated errors for temporal and spatial random effects as well as interactions between various effects. These models are closely related to Bayesian generalized additive mixed models. For univariate responses, these models extend the predictor to the additive, semiparametric form

$$\eta_{it} = f_{(1)}(x_{it1}) + \dots + f_{(p)}(x_{itp}) + z'_{it}\beta + w'_{it}b_i.$$

Generalizations to varying-coefficient mixed models and multivariate responses are obvious. Lin & Zhang (1999) propose approximate inference, using smoothing splines to fit unknown functions and double penalized quasi-likelihood as an extension of the work of Breslow & Clayton (1993), Breslow & Lin (1995) for generalized mixed models. As they point out in the discussion, similarly to approximate inference in GLMMs, there are bias problems, especially with binary data or correlated random effects.

For fully Bayesian inference unknown functions are modelled and estimated by a smoothness prior or basis function approach as described in Section 5.4. Smoothness priors  $p(f_j|\tau_j^2)$  with variance or smoothing parameters  $\tau_j^2$ ,  $j = 1, \dots, p$ , can be defined locally by random walk models (Fahrmeir & Lang, 1999) or globally by smoother matrices (Hastie & Tibshirani, 2000); compare to Section 5.4. Hyperpriors for  $\tau_j^2$  are specified by highly dispersed inverse gamma priors. Assuming conditional independence between observations and all parameters, the posterior now becomes

$$p(f_1, \dots, f_p, \beta, b_1, \dots, b_n, Q, \tau_1^2, \dots, \tau_p^2) \propto \\ \prod_{i=1}^n \prod_{t=1}^{T_i} p(y_{it}|\eta_{it}) \prod_{j=1}^p p(f_j|\tau_j^2) p(\tau_j^2) p(\beta) \prod_{i=1}^n p(b_i|Q) p(Q),$$

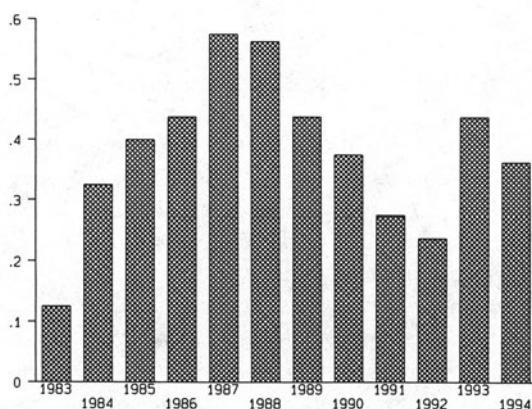
Drawings from the full conditionals  $p(b_i|\cdot)$ ,  $p(\beta|\cdot)$ , and  $p(Q|\cdot)$  can be obtained as for generalized linear mixed models. Posterior samples for variance parameters  $\tau_j^2$  and unknown functions  $f_j$  can be obtained as described in Section 5.4, i.e., by direct Gibbs sampling steps from updated inverse gamma distributions for  $\tau_j^2$ , and by MH-steps with conditional prior proposals (Fahrmeir & Lang, 1999) or the MH algorithm suggested by Hastie & Tibshirani (2000). We prefer the first option for reasons of computational efficiency in the following example.

### **Example 7.5: Longitudinal study on forest damage** (Example 6.5, continued)

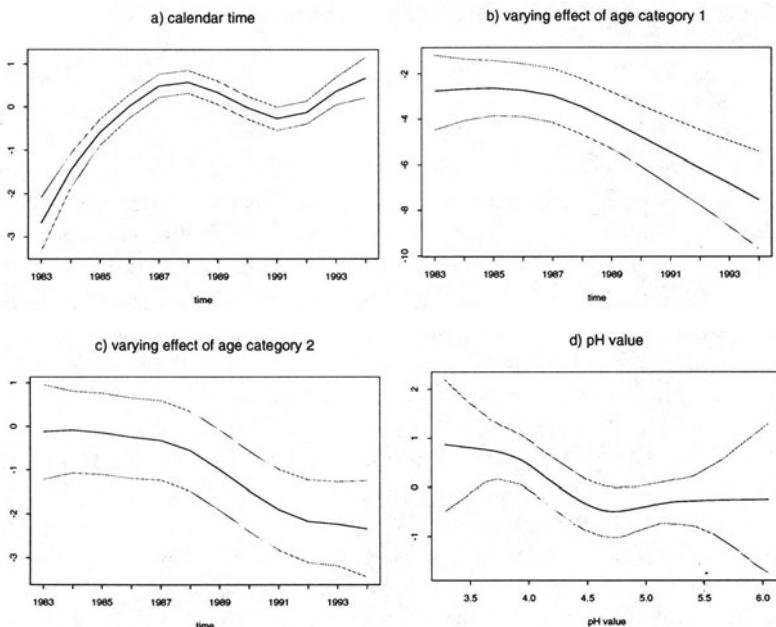
We reanalyze the forest damage data described in Example 6.5, applying now a mixed model instead of a marginal model. In contrast to Example 6.5, we use here a binary indicator for damage state  $y_{it}$  of trees, with  $y_{it} = 1$  for “light or distinct damage” of tree  $i$  in year  $t$ ,  $y_{it} = 0$  for “no damage.” Figure 7.1 shows relative frequencies of the response “damage state” over the years. Again, we can see a clear pattern with a maximum of damage around 1986, while trees seem to recover in the following five years.

Covariates are the same as in Example 6.5:

- A age of tree at the beginning of the study in 1983, measured in three categories “below 50 years” (= 1), between 50 and 120 years (= 2), and above 120 years (reference category);
- CD canopy density at the stand with categories “low” (= 1), “medium” (= 2), and “high” (reference category);



**Figure 7.1.** Relative frequency of the response “damage state” over the years.



**Figure 7.2.** Estimated nonparametric functions for the forest damage data. Shown are the posterior means with 80% credible regions.

pH pH value of the soil near the surface, measured as a metrical covariate with values ranging from a minimum from 3.3 to a maximum of 6.1

In agreement with the marginal model in Example 6.5, we use a logistic varying-coefficient mixed model with predictor

**Table 7.5.** Estimates of constant parameters for the forest damage data.

Covariate	Mean	10% quantile	90% quantile
$CD^1$	3.11	2.15	4.09
$CD^2$	1.35	0.81	1.87

$$\eta_{it} = f_1(t) + f_2(t)A_i^{(1)} + f_3(t)A_i^{(2)} + f_4(pH_{it}) + \beta_1 CD_{it}^{(1)} + \beta_2 CD^{(2)} + b_i,$$

where  $t$  is calendar time in years, and  $A^{(1)}, A^{(2)}, CD^{(1)}, CD^{(2)}$  are dummy variables (in effect coding) for  $A$  and  $CD$ . The impact of calendar time is modelled by a baseline effect  $f_1(t)$  and time-varying effects  $f_2(t), f_3(t)$  of age categories, and the possibly nonlinear effect of pH is also modelled nonparametrically, using second order random walk models as smoothness priors. The tree-specific random effect  $b$  accounts for correlation and unobserved heterogeneity. Figure 7.2 shows posterior mean estimates for the baseline effect  $f_1(t)$ , the time-varying effects  $f_2(t)$  and  $f_3(t)$  of age, and the effect of the pH value  $f_4(pH)$ . Despite different coding and modelling, interpretations are qualitatively the same as with a marginal model analysis: The baseline effect  $f_1(t)$  corresponds to the time trend of old trees over 120 years old. Trees in this age group recovered somewhat after 1986, but then the probability for damage increased again. This is in contrast to the effect of the age groups  $A^{(1)}$  (young trees) and  $A^{(2)}$  (medium age). The estimated curve  $f_2(t)$  is significantly negative and decreases, that is, in comparison to old trees, younger trees have a lower probability of being damaged and they seem to recover better over the years. The effect for trees of medium age in part c) of Figure 7.2 is similar but less pronounced. As we might have expected, low pH values, i.e., acidic soil, have a positive effect on damage probability; it decreases, however, with soil becoming less acidic, until the effect vanishes.

Note that the impact of this effect seems to be very small, because credible intervals are very large, indicating strong uncertainty with the estimated effect.

Estimated effects for canopy density are given in Table 7.5. Again, stands with low ( $CD^{(1)}$ ) or medium ( $CD^{(2)}$ ) density have an increased probability for damage compared to stands with high canopy density.  $\square$

## 7.7 Marginal Estimation Approach to Random Effects Models

Collecting the clustered responses to a multivariate response vector  $y'_i = (y_{i1}, \dots, y_{iT_i})$ , the linear random effects model (7.2.1) can be shown to be equivalent to the linear marginal model

$$y_i \sim N(\nu_i, V_i),$$

where the marginal mean  $\nu_i = E(y_i)$  is easily obtained from (7.1.4) and the marginal covariance  $V_i = \text{cov}(y_i)$  is given in (7.1.5). In contrast to the linear random effects approach, where the cluster-specific mean of  $y_{it}$  is modelled as a function of population-averaged and cluster-specific effects, the marginal approach models the marginal or population-averaged mean of  $y_{it}$  just as a function of population-averaged effects. The essential point is that the population-averaged effect of the covariates measured by the parameters  $\beta$  is the same in both linear random effects models and linear marginal models, so that the distinction between these two approaches is irrelevant. For nonlinear models, however, the distinction between random effects and marginal approaches is important. Marginal approaches to population-averaged models are considered in Sections 3.5.2 and 6.2.2.

Zeger, Liang & Albert (1988) suggest analyzing generalized linear random effects models within that marginal framework based on the marginal mean and marginal covariance structure of the responses  $y_i$ . Let us assume that a generalized linear random effects model, given by (7.2.2), (7.2.3), and (7.2.4), holds. Then

$$\mu'_i = E(y_i|b_i) = (\mu'_{i1}, \dots, \mu'_{iT_i}) \quad (7.7.1)$$

denotes the conditional mean of  $y_i$ , given  $b_i$ , where  $\mu_{it}$  is given by (7.2.2), and

$$\Sigma_i = \text{cov}(y_i|b_i) = \text{diag}(\Sigma_{i1}, \dots, \Sigma_{iT_i}), \quad \text{with} \quad \Sigma_{it} = \phi V(\mu_{it}),$$

denotes the conditional covariance of  $y_i$ , where  $\Sigma_{it}$  corresponds to the covariance function of a simple exponential family. Then the marginal mean of  $y_i$  is obtained by integrating out the random effects  $b_i$  from the conditional mean (7.7.1),

$$\nu_i = (\nu_{i1}, \dots, \nu_{iT_i})' = E(y_i) = \int \mu_i p(b_i; Q) db_i, \quad (7.7.2)$$

where  $p(b_i; Q)$  denotes the mixing density of the random effects  $b_i$ . Only for linear models one gets the simple form  $\nu = h(Z_{it}\beta) = Z_{it}\beta$ . The marginal covariance matrix is given by

$$\begin{aligned} V_i &= \text{cov}(y_i) = \int \Sigma_i p(b_i; Q) db_i + \int (\mu_i - \nu_i)(\mu_i - \nu_i)' p(b_i; Q) db_i \\ &= E \text{ cov}(y_i|b_i) + \text{cov}(\mu_i), \end{aligned} \quad (7.7.3)$$

where  $\Sigma_i = \text{diag}(\Sigma_{i1}, \dots, \Sigma_{iT_i})$ . However, analytical solutions of the integrals are available only for special cases, e.g., linear random effects models. Therefore, Zeger, Liang & Albert (1988) suggest linearizing the response

function  $h$ . Assuming that the random effects  $b_i$  are small, a first-order Taylor series expansion of  $\mu_{it} = h(Z_{it}\beta + W_{it}b_i)$  around  $b_i = 0$  yields

$$\mu_{it} = h(Z_{it}\beta + W_{it}b_i) \approx h(Z_{it}\beta) + \frac{\partial h(Z_{it}\beta)}{\partial \eta'} W_{it}b_i. \quad (7.7.4)$$

Taking expectations with respect to the mixing density  $p$ , one immediately gets the marginal mean approximation

$$\nu_{it} \approx \tilde{\nu}_{it} = h(Z_{it}\beta). \quad (7.7.5)$$

The true marginal mean  $\nu_{it}$  is equivalent to the approximation  $\tilde{\nu}_{it}$  if the response function  $h$  is the identity so that the relationship between the mean  $\mu_{it}$  and  $\beta$ , resp.,  $b_i$  is linear. For non-identical response functions the quality of the marginal mean approximation depends on the magnitude of the variance-covariance components: The larger the variance-covariance components  $Q$ , the larger is the discrepancy between the parameters  $\beta$  of the random effects approach (7.2.2) and those of the marginal approach (7.7.5).

For illustration consider a log-linear Poisson random effects model of the form

$$\mu_{it} = \exp(z'_{it}\beta + w'_{it}b_i), \quad b_i \sim N(0, Q).$$

From (7.7.2) the marginal mean

$$\nu_{it} = \exp(z'_{it}\beta)E\{\exp(w'_{it}b_i)\}$$

is obtained, where the expectation has to be taken with respect to the normal random effects density having mean  $E(b_i) = 0$  and covariance matrix  $\text{cov}(b_i) = Q$ . Since the random variables  $\exp(w'_{it}b_i)$  are log-normal with mean  $\exp(\frac{1}{2}w'_{it}Qw_{it})$ , the marginal mean

$$\nu_{it} = \exp\left(z'_{it}\beta + \frac{1}{2}w'_{it}Qw_{it}\right)$$

is obtained. Apparently the marginal mean approximation  $\tilde{\nu}_{it}$  neglects the effect of the variance-covariance components  $Q$ . Zeger, Liang & Albert (1988) also give an expression for the true marginal mean of a binary logistic random effects model, which also depends on the variance-covariance components. Moreover, Neuhaus, Hauck & Kalbfleisch (1991) showed in a general context that random effects approaches with nonlinear response functions are incompatible with marginal approaches based on  $\tilde{\nu}_{it}$ , and, in contrast to linear models, parameters  $\beta$  have different interpretations.

However, for small or moderate variance-covariance components  $Q$  the marginal mean structure of a generalized linear random effects model is adequately approximated by (7.7.5). Moreover, a first-order approximation of the marginal covariance (7.7.3) is obtained if the components  $\mu_{it}$  of the conditional mean  $\mu_i$  are replaced by approximation (7.7.5) so that

$$V_i \approx \tilde{V}_i = D_i Z_i Q Z_i' D_i + A_i, \quad (7.7.6)$$

where  $A_i = \text{diag}(A_{i1}, \dots, A_{iT_i})$  denotes a block-diagonal matrix, with  $A_{it} = \phi V(\tilde{\nu}_{it})$  evaluated at the marginal mean approximation  $\tilde{\nu}_{it}$ ,  $D_i = \text{diag}(D_{i1}, \dots, D_{iT_i})$  is also block-diagonal, with  $D_{it} = \partial h / \partial \eta_{it}^m$ ,  $\eta_{it}^m = Z_{it}\beta$ , and  $Z_i = (Z_{i1}, \dots, Z_{iT})$  is a “grand” design matrix. The quality of approximation (7.7.5) and estimation of  $\beta$  and  $Q$  by a GEE approach has been studied by Zeger, Liang & Albert (1988).

## 7.8 Notes and Further Reading

When cluster sizes do not depend on clusters, an alternative approach to reduce the number of parameters is to use conditional likelihood methods based on sufficient statistics. This approach is shortly sketched in Chapter 6 (p. 241) and outlined in Conaway (1989, 1990) and Hamerle & Ronning (1992). Similar approaches that are also suited for ordered data and are not considered here have been proposed by Agresti (1993a, 1993b), Agresti (1997), and Agresti & Lang (1993).

An alternative approach to account for heterogeneity that has led to a considerable number of papers is based on conjugate prior distributions. Instead of considering random coefficients, we determine the parameters of the exponential family distribution by conjugate prior distributions yielding beta-binomial, Poisson-gamma, or Dirichlet-multinomial models according to the type of data considered; see, e.g., Williams (1982), Moore (1987), Wilson & Koehler (1991), Brown & Payne (1986), Hausman, Hall & Griliches (1984), and Tsutakawa (1988). Lee & Nelder (1996) consider conjugate models and extend them to allow for more general distributions within a framework of hierarchical generalized linear models.

Waclawiw & Liang (1993) introduced an empirical Bayes technique using estimating functions in the estimation of both the random effects and their variance. Waclawiw & Liang (1994) propose a fully parametric bootstrap method for deriving empirical Bayes confidence intervals.

If random effects are not nested, are not correlated, or do not come in clusters like intercepts and slopes of longitudinal data, the marginal likelihood is given as a high dimensioned integral. Empirical and full Bayes methods for serially and spatially correlated random effects are described in more detail in Chapter 8. Generalized linear mixed model penalized quasi-likelihood approaches have been proposed by Breslow & Clayton (1993), Breslow & Lin (1995), and Lin & Breslow (1996). For alternative derivations see Schall (1991), Wolfinger (1994). Approximations of the likelihood or quantities needed in the EM algorithm have been considered by McCulloch (1997), Booth & Hobert (1999), Quintana, Lia & del Pino (1999), and Clayton & Rasbash (1999).

Bayesian random effects models in this chapter are based on the classical assumption of a normal distribution of random effects. Extensions to non-Gaussian distributions such as discrete mixtures of normals or Student priors can be modelled with an additional stage in the hierarchical model. However, it may be desirable to work with a nonparametric prior distribution. Ibrahim & Kleinman (1998) present a semiparametric Bayesian GLMM, replacing normal priors by a Dirichlet process prior, retaining, however, the linearity of the predictor.

# State Space and Hidden Markov Models

This chapter surveys state space and hidden Markov modelling approaches for analyzing time series or longitudinal data, spatial data, and spatio-temporal data. Responses are generally non-Gaussian, in particular, categorical, counted or nonnegative. State space and hidden Markov models have the common feature that they relate responses to unobserved “states” or “parameters” by an observation model. The states, which may represent, e.g., an unobserved temporal or spatial trend or time-varying covariate effects, are assumed to follow a latent or “hidden” Markov model.

Traditionally, the terms state space and hidden Markov models are mostly used in the context of time series or longitudinal data  $\{y_t\}$ . Then the model consists of an observation model for  $y_t$  given the state  $\alpha_t$  and a Markov chain model for the sequence  $\{\alpha_t\}$  of states. The term state space model is then mainly used for continuous states, and the term hidden Markov model for a finite state space. Given the observations  $y_1, \dots, y_t$  up to  $t$ , estimation of current, future, and past states (“filtering,” “prediction,” and “smoothing”) is a primary goal of inference. Sections 8.1 to 8.4 describe models and inference for time series and longitudinal data, with a focus on exponential family observation models, Gaussian state processes, and smoothing. The survey of Künsch (2000) complements our presentation with respect to discrete states and recent advances in recursive Monte Carlo filtering. The book by MacDonald & Zucchini (1997) on hidden Markov models for discrete-valued time series is another very readable reference.

For approximately normal data, linear state space models and the famous linear Kalman filter have found numerous applications in the analysis of time series (see, e.g., Harvey, 1989; West & Harrison, 1997, for recent treatments) and longitudinal data (e.g., Jones, 1993). Extensions to non-Gaussian time series started with robustifying linear dynamic models and filters (compare, e.g., Martin & Raftery, 1987, and the references therein). Work on exponential family state space models or dynamic generalized mod-

els began only more recently (West, Harrison & Migon, 1985). While the formulation of non-normal state space models in Section 8.2 is straightforward and in analogy to random effects models in the previous chapter, the filtering and smoothing problem, which corresponds to estimation of random effects, becomes harder. Bayesian analysis based on numerical integration will generally require repeated multidimensional integrations and can quickly become computationally intractable. Therefore, recently developed recursive Monte Carlo sampling schemes are often preferable for filtering. Posterior sampling schemes via Markov chain Monte Carlo are particularly attractive for smoothing in complex models for longitudinal and spatial data. We report on work in this area in Section 8.3.2. As an alternative for approximate inference, we consider posterior mode filtering and smoothing (Section 8.3.1). Extensions for longitudinal data are described in Section 8.4.

Section 8.5 considers some extensions to spatial and spatio-temporal data. For spatial data  $\{y_s\}$ , where  $s$  denotes the site or location in space of an observation, the latent or hidden model for the states  $\{\alpha_s\}$  now follows a Markov random field with continuous or discrete state space. For spatio-temporal data, where observations are available across time and space, hidden temporal and spatial Markov models are combined. We use the terms state space and hidden Markov models for all these cases.

## 8.1 Linear State Space Models and the Kalman Filter

As a basis, this section gives a short review of linear state space or dynamic models. Comprehensive treatments can be found, e.g., in Anderson & Moore (1979), Sage & Melsa (1971), Schneider (1986), Harvey (1989), and West & Harrison (1997).

### 8.1.1 Linear State Space Models

In the standard state space form, uni- or multivariate observations  $y_t$  are related to unobserved state vectors  $\alpha_t$  by a *linear observation equation*

$$y_t = Z_t \alpha_t + \varepsilon_t, \quad t = 1, 2, \dots, \quad (8.1.1)$$

where  $Z_t$  is an *observation* or *design matrix* of appropriate dimension, and  $\{\varepsilon_t\}$  is a white noise process, i.e., a sequence of mutually uncorrelated error variables with  $E(\varepsilon_t) = 0$  and  $\text{cov}(\varepsilon_t) = \Sigma_t$ . For *univariate* observations the design matrix reduces to a *design vector*  $z'_t$  and the covariance matrix to the variance  $\sigma_t^2$ . The observation equation (8.1.1) is then in the form of

a dynamic linear regression model with time-varying parameters  $\alpha_t$  or, in the terminology of nonparametric additive models of Chapter 5, a varying-coefficient model with time  $t$  as effect-modifier. The sequence of states is defined by a linear *transition equation*

$$\alpha_t = F_t \alpha_{t-1} + \xi_t, \quad t = 1, 2, \dots, \quad (8.1.2)$$

where  $F_t$  is a *transition matrix*,  $\{\xi_t\}$  is a white noise sequence with  $E(\xi_t) = 0$ ,  $\text{cov}(\xi_t) = Q_t$ , and the initial state  $\alpha_0$  has  $E(\alpha_0) = a_0$  and  $\text{cov}(\alpha_0) = Q_0$ . The *mean* and *covariance structure* of the model is fully specified by assuming that  $\{\varepsilon_t\}$  and  $\{\xi_t\}$  are mutually uncorrelated and uncorrelated with the initial state  $\alpha_0$ .

The *joint* and *marginal distributions* of  $\{y_t, \alpha_t\}$  are completely specified by distributional assumptions on the errors and the initial state. Since linear state space models in combination with the linear Kalman filter and smoother are most useful for analyzing approximately Gaussian data, we assume joint normality throughout this section so that

$$\varepsilon_t \sim N(0, \Sigma_t), \quad \xi_t \sim N(0, Q_t), \quad \alpha_0 \sim N(a_0, Q_0), \quad (8.1.3)$$

and  $\{\varepsilon_t\}, \{\xi_t\}, \alpha_0$  are mutually independent.

It should be remarked that the covariance matrices are allowed to be singular. Therefore, partially exact observations and time-constant states are not excluded by the model. Together with these distributional assumptions, (8.1.1) and (8.1.2) correspond to two-stage linear random effects models in Section 7.1.

In the simplest and basic state space form, the system matrices  $Z_t$ ,  $F_t$ ,  $\Sigma_t$ ,  $Q_t$  and  $a_0$ ,  $Q_0$  are assumed to be deterministic and known. In many applications, however, the covariance matrices  $\Sigma_t$ ,  $Q_t$ , the initial values  $a_0$ ,  $Q_0$ , and in some cases the transition matrices  $F_t$  are unknown wholly or contain unknown hyperparameters, say  $\theta$ , so that

$$\Sigma_t = \Sigma_t(\theta), \quad Q_t = Q_t(\theta), \quad a_0 = a_0(\theta), \quad Q_0 = Q_0(\theta).$$

Moreover, the design matrix may depend on covariates or past observations so that

$$Z_t = Z_t(x_t, y_{t-1}^*), \quad \text{with } y_{t-1}^* = (y_{t-1}, \dots, y_1).$$

The design matrix may be called predetermined since it is known when  $y_t$  is observed. Models where  $Z_t$ , and possibly other system matrices, depend on past observations are termed conditionally Gaussian models. Whereas unknown hyperparameters complicate filtering and smoothing considerably, conditional models pose no further problems if all results are interpreted conditionally.

Comparing with the random effects models of the previous chapter, we see that the main difference is that the sequence  $\{\alpha_t\}$  is no longer i.i.d., but a Markovian process that need not even be stationary.

State space models have their origin in systems theory and engineering, with famous applications in astronautics in the 1960s (see, e.g., Hutchinson, 1984). In this context, the observation equation (8.1.1) describes radar observations  $y_t$ , disturbed by noise, on the state (position, velocity, ...) of a spacecraft, and the transition equation is a linearized and discretized approximation to physical laws of motion in space. Given the measurements  $y_1, \dots, y_t$ , on-line estimation of  $\alpha_t$  ("filtering") and prediction are of primary interest. A main reason for propagating and further developing the state space approach in statistics, and in particular in time series analysis, was that a number of prominent models, e.g., autoregressive-moving-average models, structural time series, and dynamic regression models, can be described and dealt with in a flexible and unifying way. A prominent application in biostatistics is the monitoring of patients and other biometric or ecological processes; see, e.g., Smith & West (1983), Gordon (1986), van Deusen (1989), and the survey in Frühwirth-Schnatter (1991).

In the following we present some simple univariate structural time series models and show how they can be put in state space form. For comprehensive presentations, we refer the reader to Gersch & Kitagawa (1988), Harvey (1989), Kitagawa & Gersch (1996), and West & Harrison (1997). The basic idea is to interpret the decomposition

$$y_t = \tau_t + \gamma_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2), \quad (8.1.4)$$

of a time series into a trend component  $\tau_t$ , a seasonal component  $\gamma_t$ , and an irregular component  $\varepsilon_t$  as the observation equation of a state space model, and to define stochastic trend and seasonal components by recursive transition equations. Simple nonstationary trend models are first- or second-order *random walks* (sometimes shortened as *RW(1)* or *RW(2)*)

$$\tau_t = \tau_{t-1} + u_t, \quad \text{resp.,} \quad \tau_t = 2\tau_{t-1} - \tau_{t-2} + u_t, \quad u_t \sim N(0, \sigma_u^2), \quad (8.1.5)$$

and the *local linear trend model*

$$\begin{aligned} \tau_t &= \tau_{t-1} + \lambda_{t-1} + u_t, \\ \lambda_t &= \lambda_{t-1} + v_t, \quad v_t \sim N(0, \sigma_v^2), \end{aligned} \quad (8.1.6)$$

with mutually independent white noise processes  $\{u_t\}, \{v_t\}$ . If no seasonal component  $\gamma_t$  is present in the model, then (8.1.4), (8.1.5), or (8.1.6) can be put in state space form by defining

$$\alpha_t = \tau_t = 1 \cdot \alpha_{t-1} + u_t, \quad y_t = 1 \cdot \alpha_t + \varepsilon_t, \quad (8.1.7)$$

for the *RW(1)* model and

$$\alpha_t = \begin{bmatrix} \tau_t \\ \tau_{t-1} \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \tau_{t-1} \\ \tau_{t-2} \end{bmatrix} + \begin{bmatrix} u_t \\ 0 \end{bmatrix}, \quad y_t = (1, 0)\alpha_t + \varepsilon_t,$$

for the  $RW(2)$  model. For the local linear trend model, one has

$$\alpha_t = \begin{bmatrix} \tau_t \\ \lambda_t \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \tau_{t-1} \\ \lambda_{t-1} \end{bmatrix} + \begin{bmatrix} u_t \\ v_t \end{bmatrix}, \quad y_t = (1, 0)\alpha_t + \varepsilon_t. \quad (8.1.8)$$

A stochastic seasonal component for quarterly data can be specified in dummy variable form by (e.g., Harvey, 1989, p. 40)

$$\gamma_t = -\gamma_{t-1} - \gamma_{t-2} - \gamma_{t-3} + w_t, \quad w_t \sim N(0, \sigma_w^2). \quad (8.1.9)$$

If the disturbance term  $w_t$  were zero, (8.1.9) reduces to the requirement that seasonal effects sum to zero. By introducing  $w_t$ , the seasonal effects can be allowed to change over time.

Together with one of the trend models, e.g., an  $RW(2)$  model, the transition and observation equation in state space form are

$$\alpha_t = \begin{bmatrix} \tau_t \\ \tau_{t-1} \\ \gamma_t \\ \gamma_{t-1} \\ \gamma_{t-2} \end{bmatrix} = \begin{bmatrix} 2 & -1 & | & 0 & 0 & 0 \\ 1 & 0 & | & 0 & 0 & 0 \\ 0 & 0 & | & -1 & -1 & -1 \\ 0 & 0 & | & 1 & 0 & 0 \\ 0 & 0 & | & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \tau_{t-1} \\ \tau_{t-2} \\ \gamma_{t-1} \\ \gamma_{t-2} \\ \gamma_{t-3} \end{bmatrix} + \begin{bmatrix} u_t \\ 0 \\ w_t \\ 0 \\ 0 \end{bmatrix}$$

and

$$y_t = (1, 0, 1, 0, 0)\alpha_t + \varepsilon_t. \quad (8.1.10)$$

As an alternative to (8.1.9), one may suppose that the seasonal dummy variables follow a random walk (Harrison & Stevens, 1976). A different approach is to model stochastic seasonality in trigonometric form. If there are  $s$  seasons (in the year or another period), then a stochastic seasonal component is the sum

$$\gamma_t = \sum_{j=1}^{[s/2]} \gamma_{jt}$$

of  $[s/2]$  cyclical components defined by

$$\begin{aligned} \gamma_{jt} &= \gamma_{j,t-1} \cos \lambda_j + \tilde{\gamma}_{j,t-1} \sin \lambda_j + w_{jt}, \\ \tilde{\gamma}_{jt} &= -\gamma_{j,t-1} \sin \lambda_j + \tilde{\gamma}_{j,t-1} \cos \lambda_j + \tilde{w}_{jt}, \quad j = 1, \dots, [s/2], \end{aligned}$$

with seasonal frequencies  $\lambda_j = 2\pi j/s$  and mutually independent white noise processes  $\{w_{jt}\}, \{\tilde{w}_{jt}\}$  with a common variance  $\sigma_w^2$  (Harvey, 1989, pp. 40–43). As with the dummy variable form (8.1.9), one obtains a deterministic seasonal component if the disturbances  $w_{jt}, \tilde{w}_{jt}$  are set to zero. With the

use of standard trigonometric identities,  $\gamma_t$  can then be written as a sum of trigonometric terms instead of the recursive form. The component  $\tilde{\gamma}_{jt}$  is only needed for recursive definition of  $\gamma_{jt}$  and is not important for interpretation. Note that for even  $s$  the component  $\gamma_{[s/2],t}$  boils down to

$$\gamma_{[s/2],t} = \gamma_{[s/2],t-1} \cos \lambda_{[s/2]} + w_{[s/2],t}.$$

For quarterly data ( $s = 4$ ,  $\lambda_1 = \pi/2$ ,  $\sin \lambda_1 = 1$ ,  $\cos \lambda_1 = 0$ ) and with a local linear trend model, the state space form becomes

$$\alpha_t = \begin{bmatrix} \tau_t \\ \lambda_t \\ \gamma_{1t} \\ \tilde{\gamma}_{1t} \\ \gamma_{2t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & | & 0 & 0 & | & 0 \\ 0 & 1 & | & 0 & 0 & | & 0 \\ 0 & 0 & | & 0 & 1 & | & 0 \\ 0 & 0 & | & -1 & 0 & | & 0 \\ 0 & 0 & | & 0 & 0 & | & -1 \end{bmatrix} \begin{bmatrix} \tau_{t-1} \\ \lambda_{t-1} \\ \gamma_{1,t-1} \\ \tilde{\gamma}_{1,t-1} \\ \gamma_{2,t-1} \end{bmatrix} + \begin{bmatrix} u_t \\ v_t \\ w_{1t} \\ \tilde{w}_{1t} \\ w_{2t} \end{bmatrix},$$

$$y_t = (1, 0, 1, 0, 1)\alpha_t + \varepsilon_t. \quad (8.1.11)$$

All these state space models are time-invariant, i.e., the system matrices  $F_t$  and  $Z_t$  do not depend on the time index  $t$ . Assuming mutual independence among the error processes, all covariance matrices  $Q$  are diagonal but mostly singular. For example, in (8.1.10)  $Q = \text{diag}(\sigma_u^2, 0, \sigma_w^2, 0, 0)$ , and in (8.1.11)  $Q = \text{diag}(\sigma_u^2, 0, \sigma_w^2, \sigma_w^2, \sigma_w^2)$ . The variances  $\sigma_u^2, \sigma_w^2$  are, generally unknown, hyperparameters of the model. The transition equation is of block-diagonal structure, with blocks corresponding to trend and seasonal components. This block structure is typical for structural time series models, and it reflects the flexibility of the approach: adding or deleting components and blocks corresponding to each other. So other models for trend and seasonality and additional ones for daily effects, calendar effects, global stationary components, etc., can be included; see, e.g., Kitagawa & Gersch (1984), Gersch & Kitagawa (1988), and Harvey (1989).

In addition, covariates and past responses may be incorporated. If their effects are time-invariant and if we delete the seasonal component for simplicity, it leads to

$$y_t = \tau_t + (x'_t, y_{t-1}, \dots, )\beta + \varepsilon_t. \quad (8.1.12)$$

The time-invariance of  $\beta$  can be described by the artificial dynamic relation  $\beta_t = \beta_{t-1} (= \beta)$ . Together with, e.g., an  $RW(2)$  model for  $\tau_t$ , one obtains

$$\alpha_t = \begin{bmatrix} \tau_t \\ \tau_{t-1} \\ \beta_t \end{bmatrix} = \begin{bmatrix} 2 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \tau_{t-1} \\ \tau_{t-2} \\ \beta_{t-1} \end{bmatrix} + \begin{bmatrix} u_t \\ 0 \\ 0 \end{bmatrix},$$

$$y_t = (1, 0, x'_t, y_{t-1}, \dots) \alpha_t + \varepsilon_t. \quad (8.1.13)$$

The model (8.1.12), (8.1.13) can be interpreted as a semiparametric regression model: While the influence of covariates is parameterized in the usual way, the intercept is a trend component specified and estimated *non-parametrically* by an *RW*(2) model, which is the discrete-time analog of a continuous-time cubic spline for  $\tau_t$  (compare Section 5.1). We come back to this analogy in the sequel. Furthermore, continuous-time splines can be treated within the state space framework: Wahba (1978) shows that the assumption of an integrated Wiener process for  $\tau_t$  leads to the common spline methodology, and Wecker & Ansley (1983) and Kohn & Ansley (1987) use this link for spline smoothing by Kalman filter and smoother methods.

Going a step further, one may also admit time-varying covariate effects and seasonality:

$$y_t = \tau_t + \gamma_t + (x'_t, y_{t-1}, \dots) \beta_t + \varepsilon_t. \quad (8.1.14)$$

The simplest choice for  $\{\beta_t\}$  is an *RW*(1) model with mutually independent error components. Then the corresponding block in the transition equation is specified by

$$\beta_t = \beta_{t-1} + w_t, \quad w_t \sim N(0, Q), \quad Q \text{ diagonal.} \quad (8.1.15)$$

A dynamic regression model like (8.1.14) where all coefficients are time-varying may cause some uneasiness. It should be noted, however, that (8.1.15) allows but does not postulate varying parameters  $\beta_t$  if  $Q$  is admitted to be singular and is estimated as a hyperparameter. Variance components in  $Q$  with (estimated) value 0 will then correspond to effects constant in time. Yet one may question the sense of such models and possible causes for time-varying effects. In economic contexts, for example, changing structural and technological conditions, individual behavior, and attitude may be relevant causes. Dynamic modelling allows investigating the (in-) stability of regressions and can lead to better predictions. A second main cause can be misspecification of models, particularly in connection with surrogates, errors in variables, and omitted variables. Thinking positively, time-varying parameters can “swallow” or indicate such effects. In a refined analysis the model can be revised, and dynamic models can be used in this way as a tool for exploratory data analyses.

## 8.1.2 Statistical Inference

As soon as a model can be written in state space form, it provides the key for employing unified methods of statistical inference. Given the observations  $y_1, \dots, y_T$ , estimation of  $\alpha_t$  is the primary goal. This is termed

- filtering for  $t = T$ ,

- smoothing for  $t < T$ ,
- prediction for  $t > T$ .

We first consider the case of known hyperparameters, i.e., system and covariance matrices, initial values, etc., are known or given. For the sequel it is notationally convenient to denote histories of responses, covariates, and states up to  $t$  by

$$y_t^* = (y_1', \dots, y_t')', \quad x_t^* = (x_1', \dots, x_t')', \quad \alpha_t^* = (\alpha_0', \dots, \alpha_t')',$$

where  $y_0^*$ ,  $x_0^*$  are empty. Initial values  $y_0, y_{-1}, \dots$ , needed in autoregressive models, are assumed to be part of the covariates.

### Linear Kalman Filtering and Smoothing

Under the normality assumption, the optimal solution to the *filtering* problem is given by the *conditional* or *posterior mean*

$$a_{t|t} := E(\alpha_t | y_t^*, x_t^*)$$

of  $\alpha_t$  given  $y_t^*$  and  $x_t^*$ . Since the model is linear and Gaussian, the *posterior distribution* of  $\alpha_t$  is also Gaussian,

$$\alpha_t | y_t^*, x_t^* \sim N(a_{t|t}, V_{t|t}),$$

with *posterior covariance matrix*

$$V_{t|t} := E[(\alpha_t - a_{t|t})(\alpha_t - a_{t|t})'].$$

Similarly, the best one-step predictor for  $\alpha_t$ , given observations  $y_{t-1}^*$  up to  $t-1$  only, is

$$a_{t|t-1} := E(\alpha_t | y_{t-1}^*, x_{t-1}^*),$$

and the one-step prediction density is Gaussian,

$$\alpha_t | y_{t-1}^*, x_{t-1}^* \sim N(a_{t|t-1}, V_{t|t-1}),$$

with posterior covariance matrix

$$V_{t|t-1} := E[(\alpha_t - a_{t|t-1})(\alpha_t - a_{t|t-1})'].$$

The famous linear Kalman filter and smoother computes the posterior means and covariance matrices in an efficient recursive way. The usual derivations of the Kalman filter and smoother take advantage of the fact that the posterior distributions are normal. Proofs as in Anderson & Moore (1979) and Harvey (1989) repeatedly apply formulas for expectations and covariance matrices of linear transformations in combination with Bayes' theorem.

*Linear Kalman filter:*

$$\text{Initialization: } a_{0|0} = a_0, V_{0|0} = Q_0$$

For  $t = 1, \dots, T$ :

$$\text{Prediction step: } a_{t|t-1} = F_t a_{t-1|t-1},$$

$$V_{t|t-1} = F_t V_{t-1|t-1} F_t' + Q_t,$$

$$\text{Correction step: } a_{t|t} = a_{t|t-1} + K_t (y_t - Z_t a_{t|t-1}),$$

$$V_{t|t} = V_{t|t-1} - K_t Z_t V_{t|t-1},$$

$$\text{Kalman gain: } K_t = V_{t|t-1} Z_t' [Z_t V_{t|t-1} Z_t' + \Sigma_t]^{-1}.$$

Given the observations  $y_1, \dots, y_{t-1}$ , the prediction step updates the last filter value  $a_{t-1|t-1}$  to the one-step-prediction  $a_{t|t-1}$  according to the linear transition equation (8.1.2). As soon as  $y_t$  is observed,  $a_{t|t-1}$  is corrected additively by the one-step prediction error, optimally weighted by the Kalman gain  $K_t$ , to obtain the new filtering estimate  $a_{t|t}$ .

Mathematically equivalent variants, which are in some situations of computational advantage, are information filters, square root filters, etc. (see, e.g., Schneider, 1986). Let us only take a look at the correction step. Applying the matrix inversion lemma (e.g., Anderson & Moore, 1979), one can rewrite it as

$$\begin{aligned} V_{t|t} &= \left( V_{t|t-1}^{-1} + Z_t' \Sigma_t^{-1} Z_t \right)^{-1}, \\ a_{t|t} &= a_{t|t-1} + V_{t|t} Z_t' \Sigma_t^{-1} (y_t - Z_t a_{t|t-1}). \end{aligned} \quad (8.1.16)$$

This exhibits another interpretation: Information increments  $Z_t' \Sigma_t^{-1} Z_t$  are added to the currently available information  $V_{t|t-1}^{-1}$  to obtain the updated information  $V_{t|t}^{-1}$ , and the weighting Kalman gain is closely related to it.

The *smoother* for  $\alpha_t$  given all observations  $y_T^* = (y_1, \dots, y_T)$  and  $x_T^*$

$$a_{t|T} := E(\alpha_t | y_T^*, x_T^*),$$

and again the posterior is normal,

$$\alpha_t | y_T^*, x_T^* \sim N(a_{t|T}, V_{t|T}),$$

with

$$V_{t|T} := E[(\alpha_t - a_{t|T})(\alpha_t - a_{t|T})'].$$

Smoothers are usually obtained in subsequent backward steps, proceeding from  $T$  to 1. The fixed interval smoother given later is the traditional form of smoothing; see, e.g., Anderson & Moore (1979). Recently, faster variants for smoothing have been developed by De Jong (1989), Kohn & Ansley (1989), and Koopman (1993).

The (“fixed-interval”) *smoother* consists of backward recursions for  $t = T, \dots, 1$ :

$$\begin{aligned} a_{t-1|T} &= a_{t-1|t-1} + B_t(a_{t|T} - a_{t|t-1}), \\ V_{t-1|T} &= V_{t-1|t-1} + B_t(V_{t|T} - V_{t|t-1})B_t', \end{aligned}$$

with

$$B_t = V_{t-1|t-1}F_t'V_{t|t-1}^{-1}.$$

In each step, the smoothing estimate  $a_{t-1|T}$  is obtained from the filtering estimate  $a_{t-1|t-1}$  by adding the appropriately weighted difference between the smoothing estimate  $a_{t|T}$  of the previous step and the prediction estimate  $a_{t|t-1}$ .

In the following, we will sketch the lines of argument for a derivation of Kalman filtering and smoothing, which corresponds to the historically first derivation (Thiele, 1980; Lauritzen, 1981), makes the relationship to spline smoothing methods and penalized least squares methods (Section 5.1) clearer, and is of importance for posterior mode estimation in non-Gaussian state space models. A recent review of the connections between dynamic or state space models and semiparametric models for Gaussian and non-Gaussian response is given by Fahrmeir & Knorr-Held (2000).

### Kalman Filtering and Smoothing as Posterior Mode Estimation\*

The starting point is the joint conditional density  $p(\alpha_0, \alpha_1, \dots, \alpha_T | y_T^*, x_T^*)$ , i.e., the filtering and smoothing problems are treated simultaneously. Since the posterior is normal, posterior means and posterior modes are equal and can therefore be obtained by maximizing the posterior density. Repeated application of Bayes’ theorem, thereby making use of the model assumptions in (8.1.1), (8.1.2), and taking logarithms shows that this maximization is equivalent to minimization of the penalized least-squares criterion

$$\begin{aligned} PLS(\alpha) &= \sum_{t=1}^T (y_t - Z_t \alpha_t)' \Sigma_t^{-1} (y_t - Z_t \alpha_t) + (\alpha_0 - a_0)' Q_0^{-1} (\alpha_0 - a_0) \\ &\quad + \sum_{t=1}^T (\alpha_t - F_t \alpha_{t-1})' Q_t^{-1} (\alpha_t - F_t \alpha_{t-1}) \end{aligned} \tag{8.1.17}$$

with respect to  $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_T)$ . For simplicity, we have assumed that the covariance matrices  $\Sigma_t, Q_t$  are nonsingular. One may, however, drop this restriction.

As an example, consider the trend model  $y_t = \tau_t + \varepsilon_t$  with  $\tau_t$  as an  $RW(2)$  model. Setting  $\lambda := \sigma_\varepsilon^2 / \sigma_u^2$  and choosing a diffuse prior for  $\alpha_0$ , (8.1.17) reduces to the minimization of

$$\sum_{t=1}^T (y_t - \tau_t)^2 + \lambda \sum_{t=1}^T (\tau_t - 2\tau_{t-1} + \tau_{t-2})^2. \quad (8.1.18)$$

This criterion is exactly the *penalized least-squares criterion* of Whittaker (1923) for “optimal smoothing” of a trend component. Minimizing (8.1.18) tries to hold the balance between fit of the data and smoothness of the trend, expressed by the quadratic penalty term on the right side of (8.1.18). The *smoothness parameter*  $\lambda$  weights the two competing goals *data fit* and *smoothness*. Incorporation of further components results in more general discrete time-smoothing methods (e.g., Schlicht, 1981; Hebbel & Heiler, 1987; Pauly, 1989). Passing over to continuous-time smoothing leads to common continuous time splines (Reinsch, 1967; Wahba, 1978; Kohn & Ansley, 1987). In this context the following Bayesian interpretation of state space models seems quite natural: The transition model (8.1.2) defines a prior distribution for the sequence  $\{\alpha_t\}$  of states and therefore is sometimes termed smoothness-prior. One may, however, also forget this smoothness-prior and the transition model and start in a model-free manner right from criterion (8.1.17). To rewrite (8.1.17) in matrix notation, we define  $y_0 := a_0$ ,  $Z_0 := I$  and introduce the “stacked” observation vector

$$y' = (y'_0, y'_1, \dots, y'_T),$$

the block-diagonal “grand” design matrix

$$Z = \text{diag}(Z_0, Z_1, \dots, Z_T),$$

and the block-diagonal weight matrix

$$W = \text{diag}(Q_0^{-1}, \Sigma_1^{-1}, \dots, \Sigma_T^{-1}).$$

Then the penalized least-squares criterion can be written as

$$PLS(\alpha) = (y - Z\alpha)'W(y - Z\alpha) + \alpha'K\alpha.$$

The “penalty matrix”  $K$  is symmetric and block-tridiagonal, with blocks easily obtained from (8.1.17):

$$K = \begin{bmatrix} K_{00} & K_{01} & & & & 0 \\ K_{10} & K_{11} & K_{12} & & & \\ & K_{21} & \ddots & \ddots & & \\ & & \ddots & \ddots & K_{T-1,T} & \\ 0 & & & K_{T,T-1} & K_{TT} & \end{bmatrix},$$

with

$$\begin{aligned}
K_{t-1,t} &= K'_{t,t-1}, \quad 1 \leq t \leq T, \\
K_{00} &= F'_1 Q_1^{-1} F_1, \\
K_{tt} &= Q_t^{-1} + F'_{t+1} Q_{t+1}^{-1} F_{t+1}, \quad 1 \leq t \leq T, \\
F_{T+1} &= 0, \\
K_{t-1,t} &= -F'_t Q_t^{-1}, \quad 1 \leq t \leq T.
\end{aligned}$$

Setting the first derivatives  $2(-Z'W(y - Z\alpha) + K\alpha)$  of  $PLS(\alpha)$  to zero and solving for the maximizer  $\hat{\alpha} = (a_{0|T}, \dots, a_{T|T})$ , we obtain

$$\hat{\alpha} = (Z'WZ + K)^{-1} Z'Wy. \quad (8.1.19)$$

Since the posterior distribution is normal, the posterior mode  $\hat{\alpha}$  coincides with the posterior mean  $(a_{0|T}, \dots, a_{T|T})$ , which is obtained from the Kalman filter and smoother. It computes  $\hat{\alpha}$  without explicitly inverting  $Z'WZ + K$  but making efficient use of its block-banded structure. The penalized least squares approach sketched earlier indicates the close relationship to non- and semiparametric smoothing by splines; see Chapter 5 and Hastie & Tibshirani (1990): Instead of nonparametrically smoothing unknown *covariate functions*, we now consider smoothing unknown states as a *function of (discrete) time*.

### Unknown Hyperparameters

In an empirical Bayesian framework hyperparameters  $\theta$ , such as initial values or covariance matrices, are considered as unknown constants. Under the normality assumption the method of maximum likelihood is then a natural choice for estimation. Two variants are commonly in use. The *direct* method is based on the factorization

$$L(y_1, \dots, y_T; \theta) = \prod_{t=1}^T p(y_t | y_{t-1}^*; \theta)$$

of the likelihood, i.e., the joint density of  $y_1, \dots, y_t$  as a function of  $\theta$ , into the product of conditional densities  $p(y_t | y_{t-1}^*; \theta)$  of  $y_t$  given the data  $y_{t-1}^* = (y_1, \dots, y_{t-1})$  up to time  $t-1$ . Under the normality assumption, the joint and the conditional densities are normal, and corresponding conditional first and second moments can be computed for given  $\theta$  by the Kalman filter and one-step predictions

$$y_{t|t-1} = E(y_t | y_{t-1}^*) = Z_t a_{t|t-1}$$

and corresponding one-step prediction error covariance matrices

$$Z_t V_{t|t-1} Z_t' + \Sigma_t.$$

The likelihood can therefore be computed for any fixed  $\theta$  and can be maximized by numerical algorithms. Since this direct method becomes less favorable in the non-Gaussian situation, we do not present details but refer the reader to Harvey (1989, p. 127).

*Indirect* maximization starts from the joint likelihood

$$L(y_1, \dots, y_T, \alpha_0, \dots, \alpha_T; \theta)$$

of time series observations and unobserved states and uses the EM principle for ML estimation of  $\theta$ . The resulting algorithm, which will be used in modified form in Section 8.3.1, is described ahead.

Other estimation procedures are generalized least squares (Pauly, 1989) and special methods for time-invariant or stationary models (Aoki, 1987; Harvey, 1989, p. 191) and Bayes methods, where  $\theta$  is treated as a stochastic parameter with some prior distribution. Full Bayes methods consider  $\theta$  as a random variable with some prior distribution. Traditional estimation procedures involve the multiprocess Kalman filter or numerical integration; see, e.g., the survey in Frühwirth-Schnatter (1991). A Gibbs sampling approach was first suggested by Carlin, Polson & Stoffer (1992). Their “single move” sampler updates each  $\alpha_t$  separately by drawing from the posterior  $p(\alpha_t | \cdot)$  of  $\alpha_t$  given the data and all other parameters. Convergence and mixing behavior can be considerably improved by the “block-move” samplers developed by Carter & Kohn (1994a), Frühwirth-Schnatter (1994), Shephard (1994), and De Jong & Sheppard (1995). They sample from the posterior  $p(\alpha | \cdot)$  of the entire vector  $\alpha = (\dots, \alpha_t, \dots)$  of states, using Kalman filtering and backward sampling. Assigning inverse gamma or Wishart priors to unknown variances of errors in the observation or transition model, posterior samples one can obtain by drawings from inverse gamma or Wishart distributions with updated parameters. Their procedure can be extended to so-called conditional Gaussian state-space models, where error distributions are discrete or continuous mixtures of normals; see, e.g., Carter & Kohn (1994b). However, these multi-move samplers are less useful for “fundamentally” non-Gaussian models, when responses are discrete or nonnegative as in dynamic generalized linear models (Section 8.2.1).

### EM Algorithm for Estimating Hyperparameters\*

Let us consider the case of a univariate time-invariant state space model with unknown vector of hyperparameters  $\theta = (\sigma^2, Q, a_0, Q_0)$ . The joint log-likelihood of the complete data, in the terminology of the EM principle, is, apart from additive constants not containing elements  $\theta$ , given by

$$\begin{aligned}
l(y_1, \dots, y_T, \alpha_0, \dots, \alpha_T; \theta) &= -\frac{T}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^T (y_t - z'_t \alpha_t)^2 \\
&\quad - \frac{T}{2} \log(\det Q) - \frac{1}{2} \sum_{t=1}^T (\alpha_t - F_t \alpha_{t-1})' Q^{-1} (\alpha_t - F_t \alpha_{t-1}) \\
&\quad - \frac{1}{2} \log(\det Q_0) - \frac{1}{2} (\alpha_0 - a_0)' Q_0^{-1} (\alpha_0 - a_0).
\end{aligned} \tag{8.1.20}$$

In the  $p$ th cycle of the algorithm, the E-step consists of computing

$$M(\theta|\theta^{(p)}) := E \left[ l(y_1, \dots, y_T, \alpha_0, \dots, \alpha_T; \theta) | y_1, \dots, y_T, \theta^{(p)} \right], \tag{8.1.21}$$

the conditional expectation of the log-likelihood given the observations and the current iterate  $\theta^{(p)}$ . The next iterate  $\theta^{(p+1)}$  is obtained as the maximizer of  $M(\theta|\theta^{(p)})$  with respect to  $\theta$ . In the present situation, conditional expectations of quadratic forms, such as  $E(\alpha_t \alpha'_t | y_1, \dots, y_T; \theta^{(p)})$ , appearing in (8.1.21) can be computed by running the Kalman filter and smoother fixed at  $\theta^{(p)}$ , and the maximization problem can be solved analytically. Detailed derivations proceed along similar lines as in Section 7.1.2. The resulting iterative algorithm (compare to Los, 1984; Schneider, 1986; Kirchen, 1988; Harvey, 1989) jointly estimates  $a_{t|T}, V_{t|T}$ , and  $\theta$  as follows:

1. Choose starting values  $\sigma_{(0)}^2, Q^{(0)}, Q_0^{(0)}, a_0^{(0)}$ . For  $p = 0, 1, 2, \dots$
2. Smoothing: Compute  $a_{t|T}^{(p)}, V_{t|T}^{(p)}, t = 1, \dots, T$ , by linear Kalman filtering and smoothing, with unknown parameters replaced by their current estimates  $\sigma_{(p)}^2, Q^{(p)}, Q_0^{(p)}, a_0^{(p)}$ .
3. EM-step: Compute  $\sigma_{(p+1)}^2, Q^{(p+1)}, Q_0^{(p+1)}, a_0^{(p+1)}$  by

$$\begin{aligned}
a_0^{(p+1)} &= a_{0|T}^{(p)}, \quad Q_0^{(p+1)} = V_{0|T}^{(p)}, \\
\sigma_{(p+1)}^2 &= \frac{1}{T} \sum_{t=1}^T \left[ \left( y_t - z'_t \alpha_{t|T}^{(p)} \right)^2 + z'_t V_{t|T}^{(p)} z_t \right], \\
Q^{(p+1)} &= \frac{1}{T} \sum_{t=1}^T \left[ \left( a_{t|T}^{(p)} - F_t a_{t-1|T}^{(p)} \right) \left( a_{t|T}^{(p)} - F_t a_{t-1|T}^{(p)} \right)' + V_{t|T}^{(p)} \right. \\
&\quad \left. - F_t B_t^{(p)} V_{t|T}^{(p)} - V_{t|T}^{(p)'} B_t^{(p)'} F_t' + F_t V_{t-1|T}^{(p)} F_t' \right],
\end{aligned}$$

with  $B_t^{(p)}$  defined as in the smoothing steps.

## 8.2 Non-Normal and Nonlinear State Space Models

Models with nonlinear observation and transition equation and additive Gaussian errors have been used early in engineering applications (Jazwinski, 1970; Sage & Melsa, 1971; Anderson & Moore, 1979). Since the linear Kalman filter is no longer applicable, various approximative filters have been proposed (extended Kalman filter, second-order filter, Edgeworth-expansions, Gaussian sum filter, etc.), which work satisfactorily in many situations. Non-Gaussian errors explicitly appear for the first time in robustified linear state space models, assuming distributions with heavier tails and suggesting approximate conditional mean (ACM) filters; see, e.g., Martin (1979), West (1981), and Martin & Raftery (1987). Non-normality becomes even more obvious for fundamentally non-Gaussian time series of counts, qualitative data, or nonnegative data. Figure 8.1 (Section 8.3.1) shows a time series of categorized daily rainfall data from the Tokyo area for the years 1983 and 1984 (Kitagawa, 1987). For each day it is recorded whether or not rainfall over 1 mm occurred. For each calendar day  $t, t = 1, \dots, 366$ ,  $y_t = 0, 1$ , or 2 means that there was no rain this day in both years, rain in one of the years, or rain in both years, respectively. Compared to metrical time series it becomes apparently more difficult to discover trends or seasonal effects for the probability of rainfall, and some kind of smoothing would surely be helpful. The next section introduces models for such kinds of non-normal time series.

### 8.2.1 Dynamic Generalized Linear Models

Let us rewrite the Gaussian linear observation equation (8.1.1) as

$$y_t | \alpha_t, y_{t-1}^*, x_t^* \sim N(\eta_t = Z_t \alpha_t, \Sigma_t),$$

where the design matrix  $Z_t$  is a function of past responses and covariates. An obvious generalization for the exponential family framework is the following *observation model*.

The conditional density  $p(y_t | \alpha_t, y_{t-1}^*, x_t^*)$  is of the simple (uni- or multivariate) exponential family type with (conditional) mean

$$E(y_t | \alpha_t, y_{t-1}^*, x_t^*) = \mu_t = h(\eta_t), \quad \eta_t = Z_t \alpha_t, \quad t = 1, 2, \dots, \quad (8.2.1)$$

where  $h$  is one of the common response functions,  $\eta_t$  is the linear predictor, and  $Z_t$  is a function of covariates and, possibly, of past responses. Modelling of  $Z_t$  is performed along the lines in previous chapters.

Equation (8.2.1) together with the exponential family assumption replaces the observation equation in linear normal models. For parameter

transitions, we retain a *linear Gaussian transition model*, as long as this is compatible with the observation model:

$$\alpha_t = F_t \alpha_{t-1} + \xi_t, \quad t = 1, 2, \dots \quad (8.2.2)$$

The error process  $\{\xi_t\}$  is Gaussian white noise,  $\xi_t \sim N(0, Q_t)$ , with  $\xi_t$  independent of  $y_{t-1}^*, x_t^*$  and of  $\alpha_0 \sim N(a_0, Q_0)$ .

For univariate time series the linear predictor reduces to  $\eta_t = z_t' \alpha_t$ . We can define structural models with trend  $\tau_t$ , seasonal component  $\gamma_t$ , and covariates in complete analogy as in Section 8.1.1 by decomposing the linear predictor into

$$\eta_t = \tau_t + \gamma_t + x_t' \beta_t, \quad (8.2.3)$$

with transition models for  $\tau_t$  and  $\gamma_t$  in state space forms like (8.1.6) to (8.1.11). For example, a *binary dynamic logit model* with a trend component, a covariate, and past response is defined by

$$P(y_t = 1 | a_t) = \exp(\tau_t + \beta_{1t} x_t + \beta_{2t} y_{t-1}) / (1 + \exp(\tau_t + \beta_{1t} x_t + \beta_{2t} y_{t-1})) \quad (8.2.4)$$

together with a transition model such as (8.1.13) or (8.1.15), for  $\alpha_t = (\tau_t, \beta_t)$ , and a *dynamic log-linear Poisson* model similarly by

$$\lambda_t = E(y_t | \alpha_t) = \exp(z_t' \alpha_t). \quad (8.2.5)$$

For univariate dynamic models, a somewhat different kind of modelling is proposed by West, Harrison & Migon (1985). Instead of an explicit transition equation, they assume conjugate prior-posterior distributions for the natural parameter and impose a linear prediction equation for  $\alpha_t$ , involving the discount concept to circumvent estimation of unknown error covariance matrices. The observation equation is regarded only as a “guide relationship” between the natural parameter and  $z_t' \alpha_t$  to determine prior-posterior parameters by a “method of moments.” There is no general way to extend their method of moments (or some approximation by mode and curvature) to multivariate responses. The problems connected with the conjugate prior-posterior approach can be avoided, at least for the filtering problem, for a far simpler class of models, where only the “grand mean” or “level” varies over time while covariate effects are kept constant (Harvey & Fernandes, 1988; Harvey, 1989, Ch. 6.6; Smith & Miller, 1986). Dynamic generalized linear models are also described in Lindsey (1993).

The modelling approach in this section is analogous to corresponding two-stage models for random effects in Chapter 7 and has been proposed, e.g., in Fahrmeir (1992a, 1992b) and Fahrmeir & Kaufmann (1991). A recent survey is given by Ferreira & Gamerman (1999).

## Categorical Time Series

An interesting class of multivariate dynamic models is those for *time series of multicategorical or multinomial responses*. If  $k$  is the number of categories, responses  $y_t$  can be described by a vector  $y_t = (y_{t1}, \dots, y_{tq})$ , with  $q = k - 1$  components. If only one multicategorical observation is made for each  $t$ , then  $y_{tj} = 1$  if category  $j$  has been observed, and  $y_{tj} = 0$  otherwise,  $j = 1, \dots, q$ . If there are  $n_t$  independent repeated responses at  $t$ , then  $y_t$  is multinomial with repetition number  $n_t$ , and  $y_{tj}$  is the absolute (or relative) frequency for category  $j$ . For known  $n_t$ , multinomial models are completely determined by corresponding (conditional) response probabilities  $\pi_t = (\pi_{t1}, \dots, \pi_{tq})$ , specified by  $\pi_t = h(Z_t \alpha_t)$ . In this way one obtains dynamic versions of the multinomial models for unordered or ordered categories of Chapter 3. The basic idea is to model the components  $\eta_{tj}, j = 1, \dots, q$ , of the multivariate predictor  $\eta_t = Z_t \alpha_t$  in (8.2.1) in analogy to (8.2.3).

A *dynamic multivariate logistic model* with trend and covariate effects can be specified by

$$\pi_{tj} = \frac{\exp(\eta_{tj})}{1 + \sum_{r=1}^q \exp(\eta_{tr})}, \quad (8.2.6)$$

with

$$\eta_{tj} = \tau_{tj} + x'_t \beta_{tj}, \quad j = 1, \dots, q. \quad (8.2.7)$$

Model (8.2.6) also arises by applying the principle of random utility to underlying linear dynamic models with stochastic trends as in Section 3.3. It can be put in state space form (8.2.1), (8.2.2) along similar lines as in Section 8.1.1. As an example consider the case  $q = 2$  and let  $\tau_{tj}, j = 1, 2$ , be defined by second-order random walks and the time-varying covariate effects by first-order random walks

$$\beta_{tj} = \beta_{t-1,j} + \xi_{tj}, \quad \xi_{tj} \sim N(0, Q_j).$$

The degenerate case  $Q_j = 0$  corresponds to covariate effects constant in time. Setting

$$\alpha'_t = (\tau_{t1}, \beta'_{t1}, \tau_{t2}, \beta'_{t2})$$

and assuming mutual independence, the transition equation is

$$\begin{bmatrix} \tau_{t1} \\ \tau_{t-1,1} \\ \beta_{t1} \\ \tau_{t2} \\ \tau_{t-1,2} \\ \beta_{t2} \end{bmatrix} = \begin{bmatrix} 2 & -1 & | & 0 & 0 & 0 & | & 0 \\ 1 & 0 & | & 0 & 0 & 0 & | & 0 \\ 0 & 0 & | & I & 0 & 0 & | & 0 \\ 0 & 0 & | & 0 & 2 & -1 & | & 0 \\ 0 & 0 & | & 0 & 1 & 0 & | & 0 \\ 0 & 0 & | & 0 & 0 & 0 & | & I \end{bmatrix} \begin{bmatrix} \tau_{t-1,1} \\ \tau_{t-2,2} \\ \beta_{t-1,1} \\ \tau_{t-1,2} \\ \tau_{t-2,2} \\ \beta_{t-1,2} \end{bmatrix} + \begin{bmatrix} u_{t1} \\ 0 \\ \xi_{t1} \\ u_{t2} \\ 0 \\ \xi_{t2} \end{bmatrix}. \quad (8.2.8)$$

The observation model (8.1.1) is then obtained with

$$Z_t = \begin{bmatrix} 1 & 0 & x'_t & | & 0 & 0 & 0 \\ 0 & 0 & 0 & | & 1 & 0 & x'_t \end{bmatrix} \quad (8.2.9)$$

and the response function of the multivariate logistic model according to (8.2.6). The transition matrix is block-diagonal with blocks corresponding to alternative-specific trends and covariate effects. Apparently, other trend models or additional seasonal components that have been discussed in Section 8.1.1 can be included by changing or adding corresponding elements in the linear predictor and blocks in (8.2.8) in the same way as for univariate normal linear state space models.

The simplest models for ordered categories are *dynamic cumulative models*. They can be derived from a threshold mechanism for an underlying linear dynamic model as in Section 3.4. The resulting (conditional) response probabilities are

$$\pi_{tj} = F(\tau_{tj} + x'_t \beta_t) - F(\tau_{t,j-1} + x'_t \beta_t), \quad j = 1, \dots, q, \quad (8.2.10)$$

with ordered threshold parameters

$$-\infty = \tau_{t0} < \tau_{t1} < \dots < \tau_{tq} < \infty,$$

a global covariate effect  $\beta_t$ , and  $F$  a known distribution function, e.g., the logistic one. The order restriction guarantees that the probabilities in (8.2.10) are nonnegative. If thresholds vary according to one of the stochastic trend and seasonal models, this ordering can be destroyed with positive probability. This is not a problem in practice as long as thresholds are clearly separated and variances of the errors are small. However, the problem can be overcome by the same reparameterization as for random effects models: Introducing the parameter vector  $\tilde{\tau}_t = (\tilde{\tau}_{t1}, \dots, \tilde{\tau}_{tq})'$ , thresholds may be reparameterized by

$$\tau_{t1} = \tilde{\tau}_{t1}, \quad \tau_{tr} = \tilde{\tau}_{t1} + \sum_{s=2}^r \exp(\tilde{\tau}_{ts}), \quad r = 2, \dots, q,$$

or equivalently by

$$\tilde{\tau}_{t1} = \tau_{t1}, \quad \tilde{\tau}_{tr} = \log(\tau_{tr} - \tau_{t,r-1}). \quad (8.2.11)$$

Then  $\tilde{\tau}_t$  may vary without restriction.

Dynamic cumulative models can be written in state space form along the previous lines. In the simplest case thresholds and covariate effects obey a first-order random walk or are partly constant in time. Then, in original parameterization,

$$\alpha'_t = (\tau_{t1}, \dots, \tau_{tq}, \beta'_t) = \alpha'_{t-1} + \xi'_t,$$

$$Z_t = \begin{bmatrix} 1 & & x'_t \\ & \ddots & \vdots \\ & & 1 & x'_t \end{bmatrix},$$

and the response function is the common one. Thresholds with other trend and seasonal components can again be modelled as in Section 8.1.1 and are incorporated by appropriate modifications of  $\alpha_t$  and  $Z_t$ .

Dynamic versions of other models for ordered categories discussed in Section 3.3, such as sequential models, can be designed with analogous reasoning.

To specify the models completely in terms of densities, additional basic assumptions on the conditional densities of responses and covariates are required.

- (A1) Conditional on  $\alpha_t$  and  $(y_{t-1}^*, x_t^*)$ , current observations  $y_t$  are independent of  $\alpha_{t-1}^*$ , i.e.,

$$p(y_t | \alpha_t^*, y_{t-1}^*, x_t^*) = p(y_t | \alpha_t, y_{t-1}^*, x_t^*), \quad t = 1, 2, \dots$$

This conditional independence assumption is implied in linear models by the assumptions on the error structure.

- (A2) Conditional on  $y_{t-1}^*, x_{t-1}^*$ , covariates  $x_t$  are independent of  $\alpha_{t-1}^*$ , i.e.,

$$p(x_t | \alpha_{t-1}^*, y_{t-1}^*, x_{t-1}^*) = p(x_t | y_{t-1}^*, x_{t-1}^*), \quad t = 1, 2, \dots$$

Loosely speaking, (A2) means that the covariate process contains no information on the parameter process. It can be omitted for deterministic covariates.

The next assumption is implied by the transition model (8.2.2) and the assumption on the error sequence, but is restated for completeness.

- (A3) The parameter process is Markovian, i.e.,

$$p(\alpha_t | \alpha_{t-1}^*, y_{t-1}^*, x_t^*) = p(\alpha_t | \alpha_{t-1}), \quad t = 1, 2, \dots$$

## 8.2.2 Nonlinear and Nonexponential Family Models\*

As for static GLMs, one may drop the assumption that the conditional density  $p(y_t | \alpha_t, y_{t-1}^*, x_t^*)$  depends on the conditioning variables in the form of a linear predictor  $\eta_t = Z_t \alpha_t$ . If we retain the assumption that the conditional density is of the exponential family type, the observation model is given by the general specification

$$\mu_t = h_t(\alpha_t, y_{t-1}^*, x_t^*) \tag{8.2.12}$$

for the conditional mean. As long as this is sensible and compatible with the model, we may retain a linear Gaussian transition model, but we may also generalize it, for example, to a nonlinear Gaussian transition equation

$$\alpha_t = f_t(\alpha_{t-1}) + \xi_t. \quad (8.2.13)$$

This family of state space models includes the standard form of nonlinear Gaussian state space models (e.g., Anderson & Moore, 1979; Sage & Melsa, 1971) with observation equation

$$y_t = h_t(\alpha_t) + \varepsilon_t. \quad (8.2.14)$$

One may go a step further and drop the exponential family assumption for the observation model. All such models can be written in the form of the following *general state space model*.

The observation model is specified by (conditional) *observation densities*

$$p(y_t | \alpha_t, y_{t-1}^*, x_t^*), \quad t = 1, 2, \dots, \quad (8.2.15)$$

and the transition model by *transition densities*

$$p(\alpha_t | \alpha_{t-1}), \quad t = 1, 2, \dots. \quad (8.2.16)$$

Both densities will generally depend on a vector  $\theta$  of hyperparameters. To specify the model completely in terms of joint densities, we suppose that the basic assumptions (A1), (A2), and (A3) of Section 8.2.1 hold.

For example, in the case of nonlinear Gaussian models defined by observation and transition equations as in (8.2.14) and (8.2.13),

$$\begin{aligned} p(y_t | \alpha_t, y_{t-1}^*, x_t^*) &\sim N(h_t(\alpha_t), \sigma_t^2), \\ p(\alpha_t | \alpha_{t-1}) &\sim N(f_t(\alpha_{t-1}), Q_t). \end{aligned}$$

More generally, errors  $\{\varepsilon_t\}$  and  $\{\xi_t\}$  may be non-Gaussian white noise processes with densities  $d(\varepsilon)$  and  $e(\xi)$ . Then observation and transition densities are given by

$$p(y_t | \alpha_t, y_{t-1}^*, x_t^*) \sim d(y_t - h_t(\alpha_t)), \quad (8.2.17)$$

$$p(\alpha_t | \alpha_{t-1}) \sim e(\alpha_t - f_t(\alpha_{t-1})). \quad (8.2.18)$$

Choosing densities  $d$  and  $e$  with heavy tails, e.g., Cauchy, Student, or mixtures of normals, one obtains robustified linear or nonlinear models.

## 8.3 Non-Normal Filtering and Smoothing

Estimation is based on posterior densities like  $p(\alpha_0, \alpha_1, \dots, \alpha_T | y_T^*)$  or  $p(\alpha_t | y_t^*)$  for smoothing, or  $p(\alpha_t | y_{t-1}^*)$ ,  $p(\alpha_t | y_t^*)$  for prediction and filtering. One may distinguish three approaches: (i) conjugate prior-posterior

analyses, trying to solve necessary integrations in Bayes' theorem analytically, perhaps making additional approximations, (ii) full Bayes or at least posterior mean analyses based on numerical integration or (Markov Chain) Monte Carlo methods, (iii) posterior mode estimation, avoiding integration or sampling procedures. Type (i) has already been briefly discussed in Section 8.2. More detailed expositions can be found in the literature cited there. It should be noted that in smoothing there is no merit in assuming a conjugate prior. For example, with a log-linear Poisson observation model, calculation of the smoothing density cannot be simplified, even if a gamma distribution is assumed for  $\lambda_t$ . Ferreira & Gamerman (1999) present an overview of different statistical methodologies proposed to deal with dynamic generalized linear models from a Bayesian viewpoint.

We will consider the other two estimation methods for dynamic generalized linear models with exponential family observation densities  $p(y_t|\eta_t)$ ,  $\eta_t = Z_t\alpha_t$  of the form (8.2.1) and Gaussian linear transition models of the form (8.2.2). The focus will be on dynamic Poisson, binomial, and multinomial models, which will be used in the applications. It should be remarked, however, that most of the material in this section can be extended to nonlinear and nonexponential family state space models.

### 8.3.1 Posterior Mode Estimation

We first consider estimation of  $\alpha_T^* = (\alpha_0, \dots, \alpha_T)$  for known or given hyperparameters such as initial values and covariance matrices  $Q_t$  in the parameter model (8.2.2).

Repeated application of Bayes' theorem yields

$$\begin{aligned} p(\alpha_T^*|y_T^*, x_T^*) &= \prod_{t=1}^T p(y_t|\alpha_t^*, y_{t-1}^*, x_t^*) \cdot \prod_{t=1}^T p(\alpha_t|\alpha_{t-1}^*, y_{t-1}^*, x_t^*) \\ &\quad \cdot \prod_{t=1}^T p(x_t|\alpha_{t-1}^*, y_{t-1}^*, x_{t-1}^*)/p(y_T^*, x_T^*) \cdot p(\alpha_0). \end{aligned}$$

Using assumptions (A1), (A2), and (A3), we obtain

$$p(\alpha_T^*|y_T^*, x_T^*) \propto \prod_{t=1}^T p(y_t|\alpha_t, y_{t-1}^*, x_t^*) \prod_{t=1}^T p(\alpha_t|\alpha_{t-1}) \cdot p(\alpha_0).$$

Maximization of the conditional density is thus equivalent to maximizing the log-posterior

$$\begin{aligned} PL(\alpha_T^*) &= \sum_{t=1}^T l_t(\alpha_t) - \frac{1}{2}(\alpha_0 - a_0)' Q_0^{-1}(\alpha_0 - a_0) \\ &\quad - \frac{1}{2} \sum_{t=1}^T (\alpha_t - F_t \alpha_{t-1})' Q_t^{-1}(\alpha_t - F_t \alpha_{t-1}), \end{aligned} \tag{8.3.1}$$

where  $l_t(\alpha_t) = \log p(y_t | \eta_t = Z_t \alpha_t)$  are the log-densities of the observation model (8.2.1). The log-posterior (8.3.1) is a *penalized (log-) likelihood criterion*. Compared to the penalized-least-squares criterion (8.1.17), least-squares distances implied by the Gaussian observation model are replaced by Kullback-Leibler distances. Posterior mode smoothers

$$\hat{\alpha} = (a_{0|T}, \dots, a_{t|T}, \dots, a_{T|T})$$

are maximizers of (8.3.1).

As an example, consider a binary logit model (8.2.4), excluding  $x_t$  and  $y_{t-1}$ . Choosing an  $RW(2)$ -model with a diffuse prior for the initial value  $\tau_0$ , (8.3.1) becomes

$$\begin{aligned} PL(\tau_T^*) &= \sum_{t=1}^T [y_t \log \pi_t(\tau_t) + (1 - y_t) \log(1 - \pi_t(\tau_t))] \\ &\quad - \frac{1}{2\sigma_\tau^2} \sum_{t=1}^T (\tau_t - 2\tau_{t-1} + \tau_{t-2})^2, \end{aligned} \tag{8.3.2}$$

with  $\pi_t(\tau_t) = \exp(\tau_t) / [1 + \exp(\tau_t)]$ . Comparison with (8.1.18) shows that essentially the sum of squares  $\sum (y_t - \tau_t)^2$  is replaced by the sum of binomial log-likelihood contributions.

The penalized log-likelihood criterion (8.3.1) is a discrete-time analog to the criteria of Green & Yandell (1985), O'Sullivan, Yandell & Raynor (1986), and Green (1987) for spline smoothing in non- or semiparametric generalized linear models (compare with Section 5.3). This corresponds to the relationship between Kalman smoothing and spline smoothing in linear models, mentioned in Section 8.1.

Numerical maximization of the penalized log-likelihood (8.3.1) can be achieved by various algorithms. Fahrmeir (1992a) suggests the generalized extended Kalman filter and smoother as an approximative posterior mode estimator in dynamic generalized linear models. Fahrmeir & Kaufmann (1991) develop iterative forward-backward Gauss-Newton (Fisher-scoring) algorithms. Gauss-Newton smoothers can also be obtained by iterative application of linear Kalman filtering and smoothing to a “working” model, similarly as Fisher scoring in static generalized linear models can be performed by iteratively weighted least squares applied to “working” observations, see Fahrmeir & Wagenpfeil (1997) and closely related suggestions by Singh & Roberts (1992) and Durbin & Koopman (1993). This is also described in more detail below.

### Generalized Extended Kalman Filter and Smoother\*

This algorithm can be derived in a straightforward but lengthy way as an approximate posterior mode estimator by extending Sage's and Melsa's

(1971, p. 447) arguments for maximum posterior estimation in nonlinear systems from conditionally Gaussian to exponential family observations. To avoid unnecessary repetition, we omit any details. Basically, the filter is derived as a gradient algorithm via the discrete maximum principle, replacing Gaussian log-likelihoods and derivatives by corresponding terms for non-Gaussian observations. In Taylor expansions, unknown parameters have to be replaced by currently available estimates. This also concerns observation covariance matrices  $\Sigma_t(\alpha_t)$ , in contrast to the Gaussian case, where  $\Sigma_t$  is assumed to be known. The same final result can also be obtained by using linear Bayes arguments or by linearizing the observation equations around the current estimates. For simplicity we assume a linear transition equation, but more general models could also be treated. Note that in the following, filter and smoother steps  $a_{t|t}, V_{t|t}, a_{t|t-1}, V_{t|t-1}, a_{t|T}, V_{t|T}$  are numerical approximations to posterior modes and curvatures (inverses of negative second derivatives of corresponding log-posteriors). For linear Gaussian models they coincide with posterior means and covariance matrices.

*Filter steps:*

For  $t = 1, 2, \dots$

(1) Prediction

$$\begin{aligned} a_{t|t-1} &= F_t a_{t-1|t-1}, \quad a_{0|0} = a_0, \\ V_{t|t-1} &= F_t V_{t-1|t-1} F_t' + Q_t, \quad V_{0|0} = Q_0. \end{aligned}$$

(2) Correction (scoring form)

$$\begin{aligned} V_{t|t} &= \left( V_{t|t-1}^{-1} + R_t \right)^{-1}, \\ a_{t|t} &= a_{t|t-1} + V_{t|t} r_t, \end{aligned}$$

where  $r_t = \partial l_t / \partial \alpha_t$  and  $R_t = -E(\partial^2 l_t / \partial \alpha_t \partial \alpha_t')$  are the score function and (expected) information matrix contribution of observation  $y_t$ , however, evaluated at the prediction estimate  $a_{t|t-1}$ , i.e.,

$$r_t = Z_t' D_t \Sigma_t^{-1} (y_t - h(Z_t a_{t|t-1})), \quad (8.3.3)$$

$$R_t = Z_t' D_t \Sigma_t^{-1} D_t' Z_t, \quad (8.3.4)$$

with the first derivative  $D_t = \partial h / \partial \eta_t$  and the covariance matrix  $\Sigma_t$  of  $y_t$  evaluated at  $a_{t|t-1}$ .

An alternative form of the correction step can be obtained in this case by an application of the matrix inversion lemma:

(2)\* Correction (Kalman gain form)

$$\begin{aligned} a_{t|t} &= a_{t|t-1} + K_t (y_t - h_t(Z a_{t|t-1})), \\ V_{t|t} &= (I - K_t D_t' Z_t) V_{t|t-1}, \end{aligned}$$

with the Kalman gain

$$K_t = V_{t|t-1} Z_t' D_t \left( D_t' Z_t V_{t|t-1} Z_t' D_t + \Sigma_t \right)^{-1},$$

and  $D_t, \Sigma_t$  evaluated at  $a_{t|t-1}$ .

Both forms (2) and (2)\* of the correction steps are mathematically equivalent, as can be shown by an application of the “matrix inversion lemma.” The Kalman gain form (2)\* is the more familiar form of extended Kalman filtering for nonlinear Gaussian state space models.

The correction steps in scoring form are more general since they also apply to nonexponential observation models. They can be interpreted as follows: The inverse  $V_{t|t-1}^{-1}$  is the (estimated) information on  $\alpha_t$  given  $y_{t-1}^*$ . The matrix  $R_t$  is the information on  $\alpha_t$  contributed by the new observation  $y_t$ , and the sum  $V_{t|t-1}^{-1} + R_t$  is the information on  $\alpha_t$  given  $y_t^*$ . Inversion gives the (estimated) covariance matrix  $V_{t|t}$ . Thus, the correction step (2) has just the form of a single Fisher-scoring step.

This observation suggests introducing an additional iteration loop in the correction steps, with  $a_{t|t-1}$  as a starting value. Such additional local iterations may be useful if  $a_{t|t}$  is comparably far from  $a_{t|t-1}$ . In the applications of this section, additional iterations do not lead to any relevant differences in the estimates for binomial models. They are useful, however, for time series of counts where a number of observations equal to or near zero are followed by large values, as in the telephone data example. Alternatively the estimates can be improved by additional Fisher-scoring iterations; see below.

#### *Smoothening steps:*

For  $t = T, \dots, 1$

$$\begin{aligned} a_{t-1|T} &= a_{t-1|t-1} + B_t(a_{t|T} - a_{t|t-1}), \\ V_{t-1|T} &= V_{t-1|t-1} + B_t(V_{t|T} - V_{t|t-1})B_t', \end{aligned}$$

where

$$B_t = V_{t-1|t-1} F_t' V_{t|t-1}^{-1}. \quad (8.3.5)$$

These smoother steps run through as in the linear case.

#### **Gauss-Newton and Fisher-Scoring Filtering and Smoothing\***

A maximizer of the penalized log-likelihood  $PL(\alpha_T^*)$  with generally better approximation quality can be found by Gauss-Newton or Fisher-scoring iterations. We will show that this can be achieved by applying linear Kalman filtering and smoothing to a “working model” in each Fisher-scoring iteration. A different, though mathematically equivalent, form of the algorithm

is derived in Fahrmeir & Kaufmann (1991). To simplify notation, we will write  $\alpha$  for  $\alpha_T^* = (\alpha_0, \dots, \alpha_T)$ . Then the penalized log-likelihood criterion (8.3.1) can be written as

$$PL(\alpha) = l(\alpha) - \frac{1}{2}\alpha'K\alpha,$$

where

$$l(\alpha) = \sum_{t=0}^T l_t(\alpha_t),$$

$l_0(\alpha_0) := -(\alpha_0 - a_0)'Q_0^{-1}(\alpha_0 - a_0)/2$ , and the penalty matrix  $K$  is the same as in Section 8.1.2. Similarly as in that section, we define the “stacked” observation vector

$$y' = (a'_0, y'_1, \dots, y'_T),$$

the vector of expectations

$$\mu(\alpha)' = (\alpha'_0, \mu'_1(\alpha_1), \dots, \mu'_T(\alpha_T)),$$

with  $\mu_t(\alpha_t) = h(Z_t\alpha_t)$ , the block-diagonal covariance matrix

$$\Sigma(\alpha) = \text{diag}(Q_0, \Sigma_1(\alpha_1), \dots, \Sigma_T(\alpha_T)),$$

the block-diagonal design matrix

$$Z = \text{diag}(I, Z_1, \dots, Z_T),$$

and the block-diagonal matrix

$$D(\alpha) = \text{diag}(I, D_1(\alpha_1), \dots, D_T(\alpha_T)),$$

where  $D_t(\alpha_t) = \partial h(\eta_t)/\partial\eta$  is the first derivative of the response function  $h(\eta)$  evaluated at  $\eta_t = Z_t\alpha_t$ . Then the first derivative of  $PL(\alpha)$  is given by

$$u(\alpha) = \partial PL(\alpha)/\partial\alpha = Z'D(\alpha)\Sigma^{-1}(\alpha)(y - \mu(\alpha)) - K\alpha.$$

The expected information matrix is

$$U(\alpha) = -E(\partial^2 PL(\alpha)/\partial\alpha\partial\alpha') = Z'W(\alpha)Z + K,$$

with the weight matrix  $W(\alpha) = D(\alpha)\Sigma^{-1}(\alpha)D'(\alpha)$ . The expressions for first and expected second derivatives of  $l(\alpha)$  are obtained as in Sections 2.2.1 and 3.4.1. A Fisher-scoring step from the current iterate  $\alpha^0$ , say, to the next iterate  $\alpha^1$  is then

$$(Z'W(\alpha^0)Z + K)(\alpha^1 - \alpha^0) = Z'D(\alpha^0)\Sigma^{-1}(\alpha^0)(y - \mu(\alpha^0)) - K\alpha^0.$$

This can be rewritten as

$$\alpha^1 = (Z'W(\alpha^0)Z + K)^{-1}Z'W(\alpha^0)\tilde{y}^0,$$

with “working” observation

$$\tilde{y}^0 = D^{-1}(\alpha^0)(y - \mu(\alpha^0)) + Z\alpha^0.$$

Compared with (8.1.19) in Section 8.1.2, it can be seen that one iteration step can be performed by applying the linear Kalman smoother with “working” weight  $W = W(\alpha^0)$  to the “working” observation  $\tilde{y}^0$ . It is recommended that the iterations be initialized with the generalized extended Kalman smoother of Fahrmeir (1992a). Iterations will often stop after very few steps.

Comparison with spline smoothing in Section 5.5 sheds further light on the close relationship between nonparametric approaches and state space modelling: If we use posterior mode smoothing or start directly from the penalized log-likelihood criterion, then this is a kind of discrete-time spline smoothing for trend, seasonality, and covariate effects in dynamic generalized linear models. In contrast to generalized additive models (Section 5.3.2), no inner backfitting loop is required in the Fisher-scoring iterations.

### Estimation of Hyperparameters\*

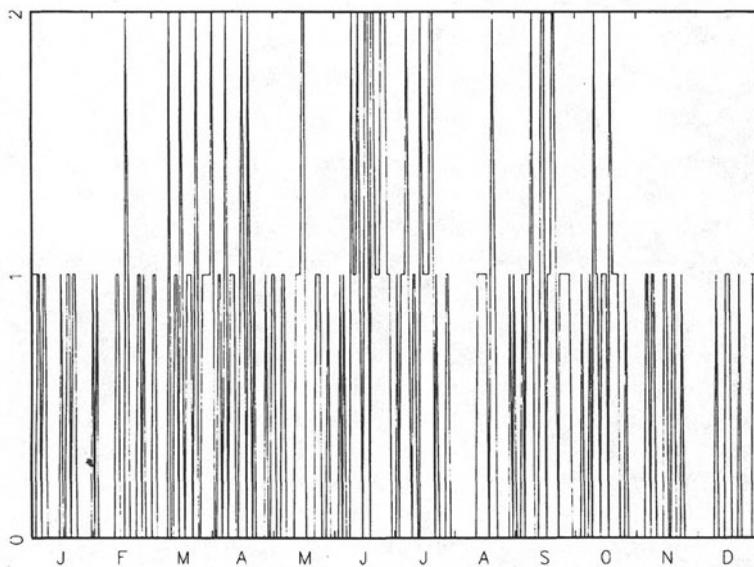
We consider only the situation of unknown initial values  $\alpha_0, Q_0$  and unknown error covariance matrix  $Q_t = Q$  (independent of  $t$ ). In analogy to the related situation in random effects models (Chapter 7), we suggest using an EM-type algorithm, which replaces posterior means and covariance matrices by posterior modes and curvatures obtained from one of the filtering and smoothing algorithms. The resulting EM-type algorithm is then formally identical with the EM algorithm in Section 8.1.2 (omitting the estimation of  $\sigma^2$ ). It has been studied in some detail by Goss (1990), Fahrmeir & Wagenpfeil (1997), and Wagenpfeil (1996), including comparisons with cross-validation. It is used in the following examples.

### Some Applications

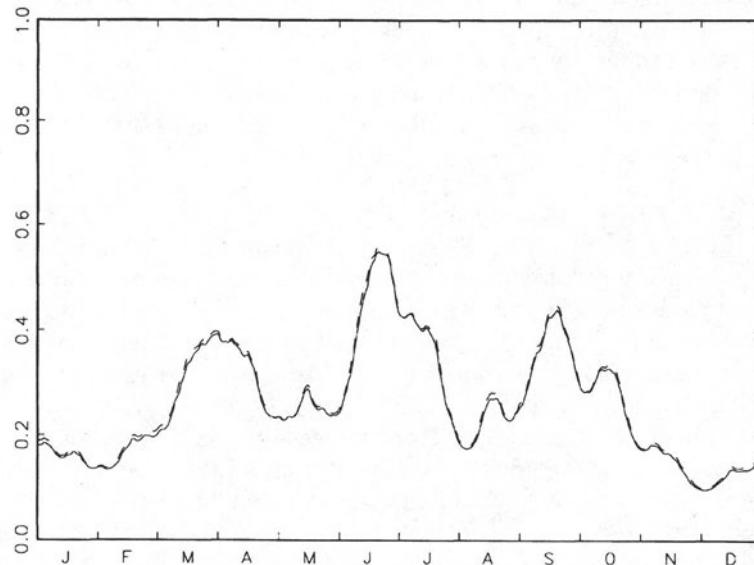
For illustration and comparison, we apply simple dynamic models and posterior mode filtering and smoothing to discrete-valued time series already analyzed in the literature.

#### **Example 8.1: Rainfall data** (Example 5.2, continued)

Figure 8.1 displays the number of occurrences of rainfall in the Tokyo area for each calendar day during the years 1983–1984. To obtain a smooth estimate of the probability  $\pi_t$  of occurrence of rainfall on calendar day  $t, t = 1, \dots, 366$ , Kitagawa (1987) chose the following simple dynamic binomial logit model:



**Figure 8.1.** Number of occurrences of rainfall in the Tokyo area for each calendar day during 1983–1984.



**Figure 8.2.** Smoothed probabilities  $\hat{\pi}_t$  of daily rainfall, obtained by generalized Kalman (---) and Gauss-Newton smoothing (—).

$$\begin{aligned} y_t &\sim \begin{cases} B(1, \pi_t), & t = 60 \text{ (February 29)} \\ B(2, \pi_t), & t \neq 60, \end{cases} \\ \pi_t &= h(\alpha_t) = \exp(\alpha_t)/(1 + \exp(\alpha_t)), \\ \alpha_{t+1} &= \alpha_t + \xi_t, \quad \xi_t \sim N(0, \sigma^2), \end{aligned}$$

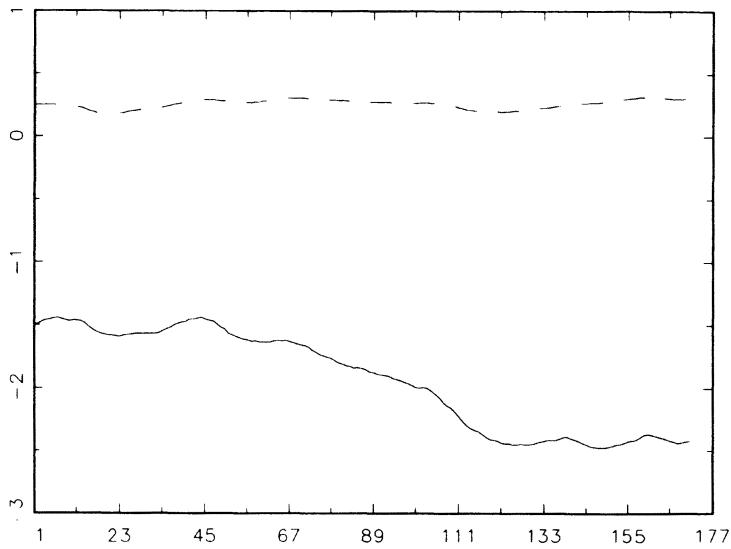
so that  $\pi_t = P$  (rain on day  $t$ ) is reparameterized by  $\alpha_t$ . Figure 8.2 shows corresponding smoothed estimates  $\hat{\pi}_t = h(\hat{\alpha}_{t|366})$  together with pointwise confidence bands ( $\hat{\pi}_t \pm \hat{\sigma}_t$ ) based on generalized Kalman smoothing and the Gauss-Newton smoother, which was initialized by the Kalman smoother and stopped after two iterations. Starting values and the random walk variance were estimated by the EM-type algorithm as  $\hat{\alpha}_0 = -1.51$ ,  $\hat{q}_0 = 0.0019$ ,  $\hat{\sigma}^2 = 0.032$ . In this example, generalized Kalman smoothing and Gauss-Newton smoothing lead to more or less the same pattern for the estimated probability of rainfall for calendar days. Whereas it is difficult to detect such a smooth pattern in Figure 8.1 by visual inspection, Figure 8.2 gives the impression of a distinct seasonal pattern: There are wet seasons in spring (March, April) and fall (September, October), June is the month with the highest probability for rainfall, and winters are dry. Such a particular pattern may be compared to similar curves for former years to see if there is any significant climatical change, or with curves for other geographical regions. Compared to Kitagawa's (1987) posterior mean smoother based on approximate numerical integration, there are only minor departures for the initial time period. Compared to cubic spline smoothing (Example 5.5), the curve has a similar pattern as the curve of Figure 5.4. Use of an  $RW(2)$  model produces a posterior mode smoother that is almost indistinguishable from the cubic spline smoother of Figure 5.3. This is not astonishing due to the close relationship between the two approaches.  $\square$

### Example 8.2: Advertising data

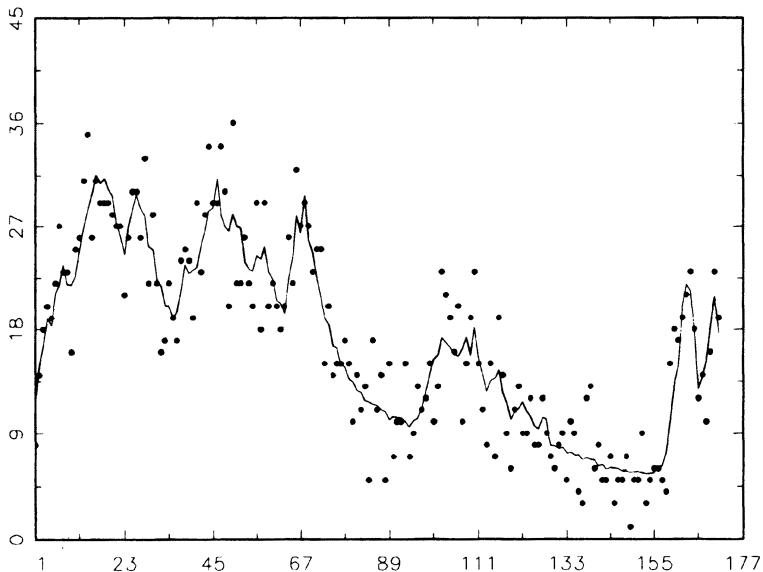
West, Harrison & Migon (1985) analyzed binomial advertising data by a logit “guide relationship” together with a random walk model and subjectively chosen discount factors. The data are weekly counts  $y_t$  of the number of people, out of a sample of  $n_t = 66$  for all  $t$ , who give a positive response to the advertisement of a chocolate bar. As a measure of advertisement influence, an “adstock coefficient” serves as a covariate  $x_t$ . The following questions might be of interest: Is there any general trend over time concerning positive response to advertising? How large is the effect of advertising, measured by the covariate? Is this effect stable over time or does it change? We reanalyze the data by the following closely related dynamic logit model, with  $n_t = 66$ :

$$\pi_t = h(\tau_t + x_t \beta_t), \quad \alpha_{t+1} = \alpha_t + \xi_t,$$

with  $\alpha_t = (\tau_t, \beta_t)'$  and  $\text{cov } \xi_t = \text{diag}(\sigma_0^2, \sigma_1^2)$ . The posterior mode smoothers in Figure 8.3 show a slight decrease of the grand mean parameter, whereas the positive advertising effect is more or less constant in time. This means



**Figure 8.3.** Smoothed trend (lower line) and advertising effect.



**Figure 8.4.** Advertising data and fitted values (—).

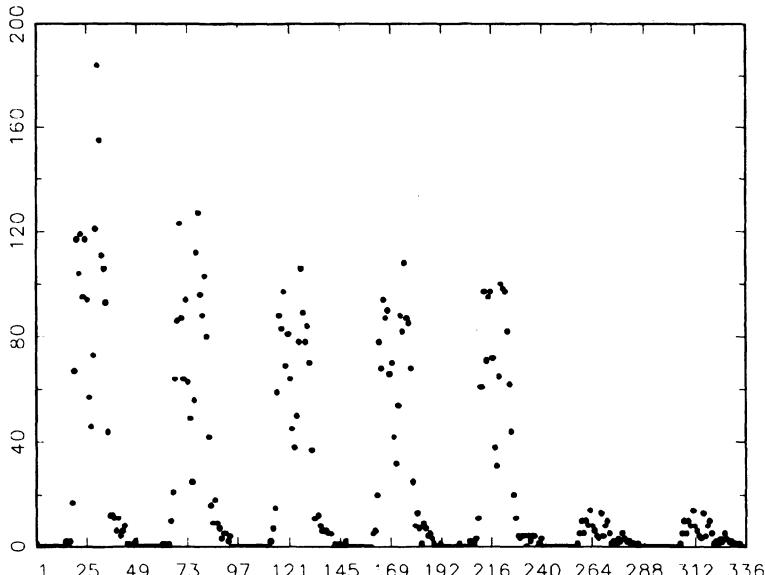
that a certain amount of advertising, measured by  $x_t$ , has same positive, though not very large, effect regardless of whether we are at the beginning or end of the advertising campaign. However, there is a general negative and decreasing trend over time, so that additional advertisement efforts are necessary to keep the probability of positive response at least at a constant level. The variance components were estimated as  $\hat{\sigma}_0^2 = 0.0025$ ,  $\hat{\sigma}_1^2 = 0.0002$ .

The data  $y_t$  together with fitted values  $\hat{y}_t$  are again in good agreement with the results of West, Harrison & Migon (1985) (Figure 8.4). Our fitted values are somewhat nearer to the data at the beginning of the observation period.  $\square$

### Example 8.3: Phone calls

The data, analyzed again in West, Harrison & Migon (1985), consist of counts of phone calls, registered within successive periods of 30 minutes, at the University of Warwick, from Monday, Sept.6, 1982, 0.00 to Sunday, Sept.12, 1982, 24.00. The data in Figure 8.5 show great variability, since counts are also registered during the nights and weekends. In particular, night hours often lead to longer periods of zero counts. For time series data of this kind, application of common linear state space methods will not be appropriate. Therefore, according to the suggestion of West, Harrison & Migon (1985), we analyze the data with a dynamic log-linear Poisson model (8.2.5), including a trend and seasonal component of the type in (8.1.11):

$$y_t \sim Po(\exp(z_t' \alpha_t)),$$



**Figure 8.5.** Number of phone calls for half-hour intervals.

$$z'_t = (1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0), \quad \alpha'_t = (\tau_t, \gamma_{1t}, \dots, \gamma_{10,t}), \\ \alpha_t = F\alpha_{t-1} + \xi_t, \quad \xi_t \sim N(0, Q),$$

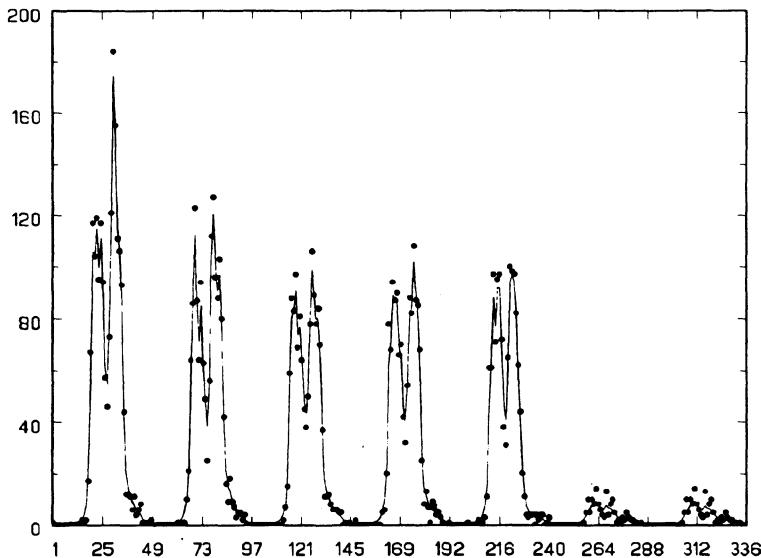
with

$$F = \begin{bmatrix} 1 & & & & 0 \\ & F_1 & & & \\ & & \ddots & & \\ & & & 0 & F_5 \end{bmatrix}, \\ F_i = \begin{bmatrix} \cos ip & \sin ip \\ -\sin ip & \cos ip \end{bmatrix}, \quad i = 1, \dots, 5, \quad p = \pi/24,$$

and  $Q$  diagonal. The following variance component estimates corresponding to  $\tau, \gamma_1, \dots, \gamma_{10}$  were obtained after 64 iterations of the EM-type algorithm with a relative stop criterion of  $\varepsilon \leq 0.01$ :  $\hat{q}_0 = 0.05, \hat{q}_1 = \hat{q}_2 = 0.003, \hat{q}_3 = \dots = \hat{q}_8 = 0.0001, \hat{q}_9 = \hat{q}_{10} = 0.00009$ . Fitted values obtained by Gauss-Newton smoothing are shown in Figure 8.6. Peaks in the morning and afternoon hours and lows during lunch and night are clearly reproduced. As one would perhaps expect, Monday—the day after the weekend—has the highest rate of telephone calls. The daily pattern remains very similar for the whole week, but the level differs: Tuesday has a higher level than Wednesday, Thursday, and Friday, while there are many fewer calls on Saturday and Sunday. (The last part of the graph is predicted counts for next Friday.)  $\square$

### 8.3.2 Markov Chain Monte Carlo and Integration-based Approaches

This section describes procedures for posterior mean and full Bayesian inference in dynamic generalized linear models. Recently developed MCMC techniques are particularly useful for posterior smoothing, i.e., generating samples from  $p(\alpha_t|y_T^*)$ ,  $t = 1, \dots, T$ , and from posteriors for hyperparameters given all observations. They can deal with high-dimensional state vectors in complex models and can be adopted as building blocks for MCMC inference in other situations like estimation of unknown functions in generalized additive (mixed) models (Section 5.4 and 7.5), analysis of longitudinal and time-space data (Sections 8.4 and 8.5), discrete-time survival and event history data (Section 9.4), and time-space data. However, they are generally less suitable for real-time filtering or prediction problems, as, for example, in on-line monitoring of patients during operation. For lower-dimensional state vectors, integration-based approaches proceeding recursively in time are then more appropriate. Direct numerical integration techniques are only



**Figure 8.6.** Observed and fitted (—) values.

feasible for dimensions up to 3 or 4. Otherwise, more recently developed recursive Monte Carlo filters have to be used.

In our applications, off-line smoothing is of prime interest, so that we put more emphasis on MCMC inference.

### MCMC Inference

In contrast to Gaussian linear state space models, full conditionals  $p(\alpha|\cdot)$  for state parameters are no longer Gaussian for dynamic generalized linear models with fundamentally non-Gaussian responses. Therefore, direct single or multi-move Gibbs sampling is not possible. Fahrmeir, Hennevogl & Klemme (1992) extended the single-move Gibbs sampler of Carlin, Polson & Stoffer (1992) to dynamic generalized linear models by combining it with rejection sampling; compare with Section 8.3.2 of the first edition of this book. For binary and categorical responses, a single-move Gibbs sampler was suggested by Carlin & Polson (1992), based on thresholding mechanisms for latent linear models with continuous response. As for Gaussian observations, these single-move methods may have poor performance when parameters  $\alpha_t$  are highly correlated in the posterior. Multi-move samplers for categorical response based on latent Gaussian state space are developed by Fahrmeir & Lang (2000) in the more general context of space-time models; see also Section 5.4. However, this approach is not possible for other types of response.

To overcome this difficulty, Gamerman (1998) reparameterizes the model to a priori independent errors of the transition model and proposes sampling

with Gaussian Hastings proposals based on ideas from posterior mode estimation. Shephard & Pitt (1997) divide  $\alpha$  into several subvectors or “blocks,” to construct “block moves” as an intermediate strategy between single moves and multi-moves, because the latter can result in very low acceptance rates of MH steps. They use several Fisher scoring type iterations for every updating step to compute first and second moments of a Gaussian Hastings proposal that is matched to the block through an analytic Taylor expansion.

Knorr-Held (1997, 1999) introduces a specific hybrid MH block move algorithm with conditional prior proposals. Basically, he uses the full conditional prior distribution  $p(\alpha_{r,s} | \alpha_{\neq r,s}, Q)$  to generate a proposal value  $\alpha_{r,s}^*$  for blocks  $\alpha_{r,s} = (\alpha_r, \dots, \alpha_s)$ ,  $1 \leq r < s \leq T$ . This algorithm is computationally very efficient, since it avoids any Fisher scoring steps and all proposals are drawn from Gaussian distributions with known moments. The methodology is particularly attractive in situations where the prior is relatively strong compared to the likelihood. Typical examples include smoothness priors for categorical data. Tuning block sizes ensures good mixing and convergence behavior. We will use this block-move algorithm in all examples.

The Gaussian conditional priors  $p(\alpha_{rs} | \alpha_{\neq rs}, Q)$  are derived from the multivariate Gaussian prior

$$p(\alpha | Q) \propto \exp\left(-\frac{1}{2}\alpha' K \alpha\right)$$

obtained from the transition model.

For non-singular  $Q$ , the general form of the penalty matrix  $K$  is given as in Section (8.1.2). For example, the penalty matrix for an  $RW(1)$  model with a diffuse initial prior and scalar  $Q$  is

$$K = \frac{1}{Q} \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{pmatrix}. \quad (8.3.6)$$

If  $Q$  is singular, as for higher-order autoregressive transition models,  $K$  can be derived directly. The penalty matrix for an  $RW(2)$  model with diffuse initial priors is derived as

$$K = \frac{1}{Q} \begin{pmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & 1 & -4 & 6 & -4 & 1 & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & 1 & -4 & 6 & -4 & 1 \\ & & & & 1 & -4 & 6 & -4 & 1 \\ & & & & & 1 & -4 & 5 & -2 \\ & & & & & & 1 & -2 & 1 \end{pmatrix}. \quad (8.3.7)$$

Note that  $K$  is singular with diffuse initial priors.

A partition of the vector  $\alpha$  as

$$\begin{aligned} \alpha &= (\alpha'_1, \dots, \alpha'_{r-1}, \alpha'_r, \dots, \alpha'_s, \alpha'_{s+1}, \dots, \alpha'_T)' \\ &= (\alpha'_{1,r-1}, \alpha'_{rs}, \alpha'_{s+1,T})' \end{aligned}$$

implies a corresponding partition

$$K = \begin{bmatrix} & K'_{1,r-1} \\ K_{1,r-1} & K_{rs} & K_{s+1,T} \\ & K'_{s+1,T} \end{bmatrix}$$

of the penalty matrix. Then the following result can be shown by simple matrix algebra: The conditional prior of the block  $\alpha_{rs}$  is Gaussian,

$$\alpha_{rs} | \alpha_{1,r-1}, \alpha_{s+1,T}, Q \sim N(\mu_{rs}, \Sigma_{rs}),$$

with moments

$$\mu_{rs} = \begin{cases} -K_{rs}^{-1} K_{s+1,T} \alpha_{s+1,T} & \text{if } r = 1, \\ -K_{rs}^{-1} K_{1,r-1} \alpha_{1,r-1} & \text{if } s = T, \\ -K_{rs}^{-1} (K_{1,r-1} \alpha_{1,r-1} + K_{s+1,T} \alpha_{s+1,T}) & \text{otherwise,} \end{cases}$$

$$\text{and } \Sigma_{rs} = K_{rs}^{-1}.$$

Using the band structure of  $K$ , inverses  $K_{rs}^{-1}$  can be computed efficiently and in advance of sampling.

A block-move MH step with conditional prior proposal to update the full conditional

$$p(\alpha_{rs} | \cdot) \propto \prod_{t=r}^s p(y_t | \alpha_t) p(\alpha_{rs} | \alpha_{1,r-1}, \alpha_{s+1,T}, Q)$$

is defined as follows:

Draw a proposal  $\alpha_{rs}$  from  $N(\mu_{rs}), \Sigma_{rs}$ , with  $\mu_{rs}, \Sigma_{rs}$  evaluated at current values  $\alpha$  and  $Q$  of the chain, and accept it with probability

$$\min \left\{ 1, \frac{\prod_{t=s}^r p(y_t | \alpha_t^*)}{\prod_{t=r}^s p(y_t | \alpha_t)} \right\}.$$

This simple form of acceptance probability results since proposal densities cancel out. Thus, only computation of likelihoods is needed and any computer-intensive Fisher scoring steps can be avoided.

Typically, bigger block sizes cause smaller acceptance rates. Best mixing and convergence behavior are obtained with acceptance rates around 0.5. Knorr-Held (1997, 1999) discusses several blocking strategies.

For a fully Bayesian analysis, highly dispersed Wishart or gamma priors are assigned to  $Q$  or variances. As in Section 7.5, posteriors are Wishart or inverse gamma distributions with updated parameters, so that direct Gibbs sampling steps are possible. Together, the following hybrid MCMC technique is obtained:

- (i) Partition  $\alpha$  into blocks  $(\alpha'_{1u}, \dots, \alpha'_{rs}, \dots, \alpha'_{vT})'$ .
- (ii) For each block, draw samples from full conditionals

$$p(\alpha_{1u} | \cdot), \dots, p(\alpha_{rs} | \cdot), \dots, p(\alpha_{vT} | \cdot)$$

by MH steps with conditional prior proposals.

- (iii) Draw samples for  $Q$  or variances from posterior inverse Wishart or gamma priors.
- (iv) Repeat steps (i) to (iii) until convergence is reached.

This hybrid MCMC scheme can easily be extended to dynamic models

$$\eta_t = Z_t \alpha_t + W_t \beta$$

with time-constant effects, imposing Gaussian or flat priors for  $\beta$  and adding a further MH update step as outlined in Section 2.3.2.

### Integration-based Approaches

Suppressing hyperparameters, let  $p(\alpha_t | y_{t-1}^*)$  and  $p(\alpha_t | y_t^*)$  denote prediction and filtering densities of  $\alpha_t$ ,  $t = 1, 2, \dots$ , given observations up to time  $t-1$  or  $t$ , respectively. Assume that the filtering density  $p(\alpha_{t-1} | y_{t-1}^*)$  is known at time  $t-1$ . Using the model assumptions, the following integral recursions for prediction and filtering are easily derived:

$$\begin{aligned}
p(\alpha_t | y_{t-1}^*) &= \int p(\alpha_t | \alpha_{t-1}, y_{t-1}^*) p(\alpha_{t-1} | y_{t-1}^*) d\alpha_{t-1} \\
&= \int p(\alpha_t | \alpha_{t-1}) p(\alpha_{t-1} | y_{t-1}^*) d\alpha_{t-1}, \\
p(\alpha_t | y_t^*) &= \frac{p(y_t | \alpha_t, y_{t-1}^*) p(\alpha_t | y_{t-1}^*)}{p(y_t | y_{t-1}^*)},
\end{aligned}$$

with

$$p(y_t | y_{t-1}^*) = \int p(y_t | \alpha_t) p(\alpha_t | y_{t-1}^*) d\alpha_t.$$

Similarly, backward integrations for smoothing densities can be obtained; see Kitagawa (1987) or Hürzeler (1998) and Künsch (2000). The prediction densities are also the key for likelihood estimation of hyperparameters because the joint density or likelihood is

$$p(y_1, \dots, y_T) = \prod_{t=1}^T p(y_t | y_{t-1}^*) = \prod_{t=1}^T p(y_t | \alpha_t) p(\alpha_t | y_{t-1}^*) d\alpha_t.$$

Numerical approximation of the integral recursions started with Kitagawa (1987). However, due to its complexity and numerical effort it is generally only applicable to dynamic generalized linear models with univariate response  $y_t$  and scalar or very low-dimensional  $\alpha_t$ . Hodges & Hale (1993) and Tanizaki (1993) modified and improved Kitagawa's original method, but all these methods are only feasible in low-dimensional state space models. For hidden Markov models with a finite state space, all the integrals in filtering and smoothing recursions are simple sums. Künsch (2000) gives recursions in a numerically stable form. If  $M$  is the number of states, these recursions require  $O(TM^2)$  operations in the case of known hyperparameters.

An approximate solution to the prediction and filter problem in dynamic generalized linear models with linear Gaussian transition models has been given by Schnatter (1992) and Frühwirth-Schnatter (1991). In contrast to Kitagawa (1987), the prediction step is carried out only approximately so that the conditional densities  $p(\alpha_t | y_{t-1}^*)$  are available in closed form. More precisely, the density  $p(\alpha_t | y_{t-1}^*)$  is approximated by an  $N(a_{t|t-1}, V_{t|t-1})$ -density with mean  $a_{t|t-1} = F_t a_{t-1|t-1}^m$  and covariance matrix  $V_{t|t-1} = F_t V_{t-1|t-1}^m F_t' + Q_t$ , where  $a_{t-1|t-1}^m$  and  $V_{t-1|t-1}^m$  are posterior means and variances obtained from the approximate filtering density at time  $t - 1$ . Due to the approximation, numerical effort no longer increases exponentially with time. To solve the integrations, Gauss-Hermite quadrature may be applied as in Schnatter (1992) or Monte Carlo integration with importance sampling as in Hennevogl (1991). Both methods require filter posterior modes  $a_{t|t}$  and curvatures  $V_{t|t}$  that can be obtained easily and quickly by the generalized extended Kalman filter described in the previous section.

As an alternative to numerical integration, various Monte Carlo approximations have been suggested; see, e.g., Müller (1991), Gordon, Salmond & Smith (1993), and Kitagawa (1996). Hürzeler (1998), and Hürzeler & Künsch (1998) recursively approximate the integral recursions by means of conditional densities  $p(\alpha_t | \alpha_{t-1|t-1}^{(j)})$  of the filter sample  $\alpha_{t-1|t-1}^{(j)}$ ,  $j = 1, \dots, N$ , at time  $t - 1$ . Filter updates  $\alpha_{t|t}^{(j)}$  are obtained via rejection sampling. Practical feasibility is demonstrated through several examples with real and artificial data. Other recent proposals are made in Pitt & Shephard (1999), in Durbin & Koopman (1993), and in the book edited by Doucet, Freitas & Gordon (2000); see Künsch (2000) for a recent review. The methods are often called *particle filters*.

The following examples are taken from Knorr-Held (1997) and Fahrmeir & Knorr-Held (2000).

**Example 8.4: Rainfall data** (Example 8.1, continued)

Recall the rainfall data  $y_t$ ,  $t = 1, \dots, 366$ , which were analyzed in Example 8.1 by a dynamic binomial logit model of the form

$$y_t \sim \begin{cases} B(1, \pi_t), & t = 60 \text{ (February 29)} \\ B(2, \pi_t), & t \neq 60, \end{cases}$$

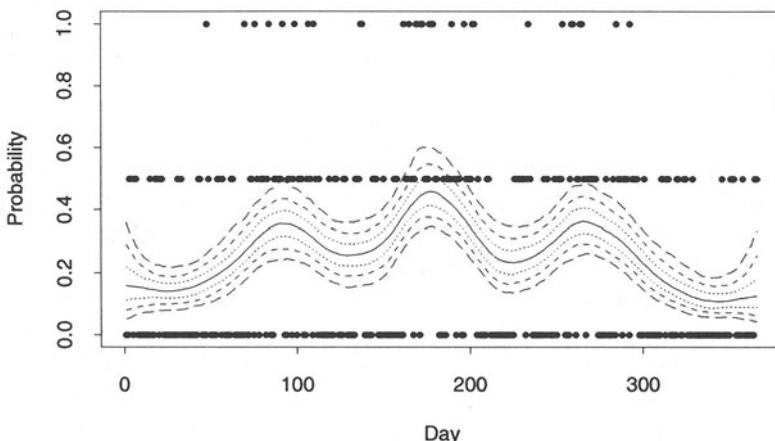
$$\pi_t = \exp(\alpha_t) / (1 + \exp(\alpha_t)),$$

together with a scalar random walk model  $\alpha_{t+1} = \alpha_t + \xi_t$ . Smoothed posterior mode estimates of  $\pi_t$  were obtained by generalized extended Kalman filtering and smoothing, combined with an EM-type algorithm for estimating the unknown variance  $\sigma^2 = \text{var}(\xi_t)$  of the *RW*(1) model. Here we reanalyze the data with a fully Bayesian MCMC approach using conditional prior proposals.

To obtain smoother estimates for  $\pi_t$  and for comparison with cubic spline smoothing (Example 5.4), we assume an *RW*(2) or second-order difference prior  $\alpha_t = 2\alpha_{t-1} - \alpha_{t-2} + \xi_t$ ,  $\text{var}(\xi_t) = \sigma^2$ . One of the advantages of MCMC is the possibility to calculate exact posterior distributions of functionals of parameters. For the Tokyo rainfall data, the posterior estimates of the probabilities  $\pi_t$  are of main interest. Instead of plugging in an estimate for  $\{\alpha_t\}$ , we calculate posterior samples from  $\pi_t = \exp(\alpha_t) / 1 + \exp(\alpha_t)$ , using the original samples from  $p(\alpha|y)$ . The posterior distributions  $p(\pi|y)$  can now be explored in detail without any approximation. In contrast, posterior mode or spline estimates do not have this feature. Here plug-in estimates, especially confidence bands, are typically biased due to nonlinearity.

Figure 8.7 shows the posterior estimates of the probabilities  $\{\pi_t\}$  for the Tokyo rainfall data, calculated by a conditional prior block MCMC algorithm. A highly dispersed but proper inverse gamma hyperprior with  $a = 1$ ,  $b = 0.00005$  was assigned to  $\sigma^2$ . This prior has a mode at 0.000025.

The estimated posterior median was 0.00001. The pattern in Figure 8.7 with peaks for wet seasons reflects the climate in Tokyo. In Example 5.2 the probabilities  $\{\pi\}$  were fitted by a cubic smoothing spline, with the smoothing parameter estimated by a generalized cross-validation criterion. This criterion had two local minima, at  $\lambda = 32$  and  $\lambda = 4064$ . The smoothing spline for  $\lambda = 4064$  is quite close to the posterior median fit, whereas the smoothing spline for  $\lambda = 32$  is much rougher. Such rougher posterior median estimates are also obtained if the parameter  $b$  for the inverse gamma prior is set to higher values. For example, with  $a = 1, b = 0.005$ , the prior mode equals 0.0025. This prior is in favor of larger values for  $\sigma^2$ , so that posterior median estimates for  $\{\pi_t\}$  become rougher. These results correspond to empirical evidence experienced in other applications: If smoothing and variance parameters are properly adjusted, posterior mean and medians are often rather close to posterior modes or penalized likelihood estimates. Also, estimation of hyperparameters by cross-validation or an EM-type algorithm can be helpful for the choice of parameters of the hyperprior in a fully Bayesian model. Similar evidence is provided by the next example.  $\square$



**Figure 8.7.** Tokyo rainfall data. Data and fitted probabilities (posterior median within 50%, 80%, and 95% credible regions). The data are reproduced as relative frequencies with values 0, 0.5, and 1.

### Example 8.5: Weekly incidence of AHC

As a second example, we consider a time series of counts  $y_t$  of the weekly incidence of acute hemorrhagic conjunctivitis (AHC) in the Chiba prefecture in Japan during 1987. Kashiwagi & Yanagimoto (1992) analyze these data, assuming a log-linear Poisson model

$$y_t | \lambda_t \sim P_0(\lambda_t), \quad \lambda_t = \exp(\alpha_t),$$

and a first-order random walk prior for  $\alpha$ . They obtain a posterior mean estimate based on numerical integrations similar to that in Kitagawa (1987). As in Example 8.5, samples for  $\alpha_t$  obtained from MCMC simulation are used to calculate posterior samples for  $\lambda_t$ .

Estimates for the AHC data are shown in Figure 8.8(a) and (b) for both first- and second-order random walk priors. The posterior distribution of the intensities  $\{\lambda_t\}$  shows a peak around week 33 similar to the results of Kashiwagi & Yanagimoto (1992). Compared to the model with second-order random walk priors, estimates in Figure 8.8(a) are somewhat rougher and the peak around week 33 is lower and flatter. This reflects the fact that first-order random walk priors favor horizontal, locally straight lines. Figure 8.8(c) shows a Bayesian cubic spline-type estimate with a continuous-time prior from Biller & Fahrmeir (1997). As was to be expected with equally spaced observations, these estimates are in close agreement with those in Figure 8.8(b). Figure 8.8(d) displays the cubic smoothing spline, which is the posterior mode estimator from the Bayesian point of view. As with the rainfall data example, it is again quite close to the posterior median in Figure 8.8(c).  $\square$

## 8.4 Longitudinal Data

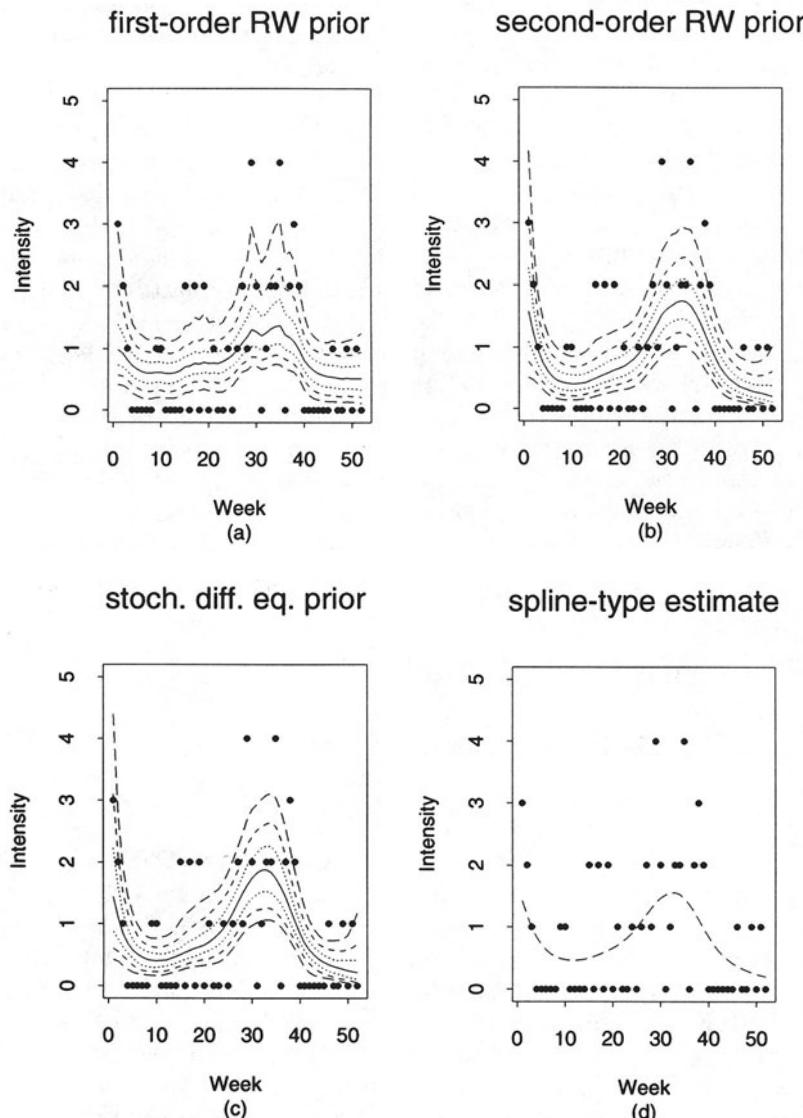
Let the longitudinal or panel data consist of observations

$$(y_{it}, x_{it}), \quad i = 1, \dots, n, \quad t = 1, \dots, T,$$

for a population of  $n$  units observed across time. Fixed effects models for such data are the subject of Section 6.2. Models with random effects across units  $i$  (Chapter 7) are sensible alternatives, particularly if  $T$  is small compared to  $n$ . The state space modelling approach to longitudinal data allows, in principle, introducing random effects across units and time by including them both in a “large” state vector. For normal data, models of this kind have been proposed and studied previously; see, e.g., Rosenberg (1973), Hsiao (1986), and in particular Jones (1993). It should be noted that the material in Section 8.4 is written for the case of equally spaced time points. For linear state space models, nonequally spaced time points are treated, e.g., in Jones (1993). State space modelling of non-Gaussian longitudinal data has been developed more recently. Due to their flexibility, MCMC approaches are particularly appealing.

### 8.4.1 State Space Modelling of Longitudinal Data

In the sequel it will be convenient to collect individual observations in “panel waves”



**Figure 8.8.** AHC data. Data and fitted probabilities (posterior median within 50%, 80%, and 95% credible regions).

$$y'_t = (y'_{1t}, \dots, y'_{nt}), \quad x'_t = (x'_{1t}, \dots, x'_{nt}), \quad t = 1, \dots, T,$$

and to denote “histories” up to  $t$  by  $y_t^* = (y_1, \dots, y_t)$ ,  $x_t^* = (x_1, \dots, x_t)$  as before. In view of the longitudinal data situation, it is natural to consider models for individual responses  $y_{it}$  conditional on the predetermined observations  $y_{t-1}^*$ ,  $x_t^*$  of past responses and of past and current covariates, and on

individual parameter vectors  $\alpha_{it}$ . Generally, these parameter vectors may contain parameters that are constant over units (“cross-fixed”) or time, and parameters that vary over units (“cross-varying”) or time. Within the linear exponential family framework, a corresponding *observation model* is given by the following:

The conditional densities  $p(y_{it}|\alpha_{it}, y_{t-1}^*, x_t^*)$  are of the simple exponential family type with mean

$$E(y_{it}|\alpha_{it}, y_{t-1}^*, x_t^*) = \mu_{it} = h(\eta_{it}) \quad (8.4.1)$$

and linear predictor

$$\eta_{it} = Z_{it}\alpha_{it}, \quad (8.4.2)$$

where the design matrix  $Z_{it}$  is a function of covariates and, possibly, past responses.

However, without adding further structural assumptions on the variation of effects over time and individuals, the estimation of  $\alpha_{it}$  will generally not be possible.

A straightforward extension of dynamic models for time series can be obtained by the assumption that parameters do not vary over units. Then the predictor in (8.4.2) reduces to

$$\eta_{it} = Z_{it}\alpha_t. \quad (8.4.3)$$

If we assume that individual responses are conditionally independent, a dynamic model for longitudinal data is obtained by supplementing the observation model (8.4.1), (8.4.3) with Gaussian transition models as smoothness priors for  $\alpha$  as in Section 8.3. Just as for time series, some subvector of  $\alpha_t$  may indeed be constant. This can be made explicit by rewriting the predictor (8.4.3) in additive form  $\eta_{it} = Z_{it}\alpha_t + V_{it}\beta$ .

Observation models of the form (8.4.1), (8.4.3) may be appropriate if heterogeneity among units is sufficiently described by observed covariates. This will not always be the case, in particular for larger cross-sections. A natural way to deal with this problem is an additive extension of the linear predictor to

$$\eta_{it} = Z_{it}\alpha_t + V_{it}\beta + W_{it}b_i, \quad (8.4.4)$$

where  $b_i$  are unit-specific parameters and  $W_{it}$  is an appropriate design matrix. A *dynamic generalized linear mixed model* (DGLMM) is obtained with usual transition models for  $\alpha$  and a random effects model for the unit-specific parameters. As in Chapter 7, a common assumption is that the  $b_i$ 's are i.i.d. Gaussian,

$$b_i \sim N(0, D), \quad (8.4.5)$$

with covariance matrix  $D$ .

In principle, all models above can be put in state space form by appropriate specifications of “panel wave” parameter vectors  $\alpha_t = (\alpha'_{t1}, \dots, \alpha'_{tn})'$  and associated transition models. Therefore, it seems that the filtering and smoothing algorithms of the previous section can be applied to the sequence  $\{y_t, x_t\}$  of panel waves to estimate the sequence  $\{\alpha_t = (\alpha_{1t}, \dots, \alpha_{nt})\}$ . However, the dimension of  $\alpha_t$  and of (8.4.2) is now  $\dim(\alpha_{it})$  times  $n$ , the size of the cross section. Without further restrictions or simplification, the computational amount becomes infeasible even for moderate  $n$ . A first attempt to decompose the filtering problem into  $n$  parallel approximate filtering algorithms  $\alpha_{it}$ ,  $i = 1, \dots, n$ , has been made in Fahrmeir, Kaufmann & Morawitz (1989), but the quality of approximations involved is difficult to assess.

In the next section we confine attention to dynamic mixed models of the form (8.4.1), (8.4.4), including the model (8.4.3) as a special case.

## 8.4.2 Inference For Dynamic Generalized Linear Mixed Models

As for time series, we may consider approximate inference by penalized likelihood or posterior mode estimation and fully Bayesian inference via MCMC.

For model (8.4.3) without random effects, the *posterior mode approach* maximizes

$$\begin{aligned} PL(\alpha) := & \sum_{t=1}^T \sum_{i=1}^n l_{it}(\alpha_t) - \frac{1}{2}(\alpha_0 - a_0)' Q_0^{-1}(\alpha_0 - a_0) \\ & - \frac{1}{2} \sum_{t=1}^T (\alpha_t - F_t \alpha_{t-1})' Q_t^{-1}(\alpha_t - F_t \alpha_{t-1}), \end{aligned} \tag{8.4.6}$$

where  $l_{it}(\alpha_t)$  is (conditional) log-likelihood contribution of observation  $y_{it}$ . Computationally efficient solutions can be obtained by adopting extended or iterative Kalman-type smoothers for time series; see the first edition of this book or Wagenpfeil (1996) for details. If the size  $n$  of the cross section is large enough, asymptotic theory for ML estimation in cross sections also applies to posterior mode estimation of  $\alpha_t$ , which is constant across units  $i$  at time  $t$ . Then the posterior distribution of  $\alpha_t$  is approximately normal under rather mild conditions. As a consequence, posterior modes  $a_t$  provide reasonable approximations to posterior expectations of  $\alpha_t$ .

For dynamic mixed models (8.4.4), a further penalty term

$$\sum_{i=1}^n b_i' D b_i,$$

corresponding to the Gaussian prior (8.4.5), has to be added to (8.4.6). An algorithmic solution for the resulting joint posterior mode or penalized likelihood estimates  $\hat{\alpha}$ ,  $\hat{\beta}$ , and  $\hat{b}$  is worked out in Biller (2000c). He uses backfitting by cycling between posterior mode estimation for random effects (Section 7.3) and an iterative Kalman smoother combined with an EM-type algorithm for estimating hyperparameters. However, computation times can become large; also, significant bias can occur with sparse data; see Breslow & Clayton (1993), and Breslow & Lin (1995).

Fully Bayesian MCMC techniques are an attractive alternative for dynamic mixed models through their model flexibility. The additional parameters  $b_1, \dots, b_n$  are added to the set of unknown parameters and are updated as in Section 7.6 with some well-designed proposals, for example, with Metropolis random walk proposals, in every MCMC cycle. In addition, a hyperprior for  $D$  has to be introduced. Again the usual choice is an inverted Wishart distribution

$$p(D) \propto |D|^{-\zeta - (m+1)/2} \exp(-\text{tr}(\Psi D^{-1}))$$

with parameters  $\zeta > (m-1)/2$  and  $|\Psi| > 0$ ; here  $m$  is the dimension of  $b_i$ . A Gibbs step can then be used to update  $D$ . Time-varying and fixed effects are modelled and estimated by sampling from their full conditionals as described in (8.3.2), modifying likelihood contributions  $p(y_{it}|\alpha_t^*)$  in the acceptance probability (8.3.2) to  $\prod_{i=1}^n p(y_{it}|\alpha_t^*)$ .

Knorr-Held (1997) and Fahrmeir & Knorr-Held (2000) provide details and present the following example.

### **Example 8.6: Business test** (Example 6.3, continued)

We apply the method to monthly IFO business test data collected in the industrial sector “Steine und Erden”, for the period from January 1980 to December 1990. Firms in this sector manufacture initial products for the building trade industry.

The response variable is formed by the production plans  $P_t$ . Its conditional distribution is assumed to depend on the covariates “orders in hand”  $O_t$  and “expected business condition”  $G_t$ , and on the production plans  $P_{t-1}$  of the previous month. No interaction effects are included. This choice is motivated by previous results of König, Nerlove & Oudiz (1981) and Nerlove (1983, p. 1273), applying log-linear probability models to a large panel of branches. Attempts to carry out analyses for subgroups or smaller industry branches separately, which should also be of scientific interest, often run into problems due to the nonexistence of estimates. The reason is a particular property of this longitudinal data set: If in doubt, firms seem to have a conservative tendency and prefer the “no change” (“=” category. For multivariate categorical analyses, this results in a data pattern with a majority of entries in certain combinations whereas data are rather sparse in others. As a consequence of such data sparseness, methods for analyzing

time-varying effects by sequentially fitting models for cross sections (e.g., Stram, Wei & Ware, 1988) will often break down if applied to branches.

In the following each trichotomous ( $k = 3$ ) variable is described by two ( $q = 2$ ) dummy variables, with “–” as the reference category. Thus (1,0), (0,1), and (0,0) stand for the responses +, =, and –, respectively. The relevant dummies for “+” and “=” are shortened by  $P_t^+$ ,  $P_t^=$ , etc. A cumulative logistic model is used due to the ordinal nature of the response variable: Let  $P_{it} = 1$  and  $P_{it} = 2$  denote the response categories “increase” and “no change,” respectively. Then

$$P(P_{it} \leq j) = F(\theta_{itj} + x'_{it}\beta_t), \quad j = 1, 2,$$

is assumed with  $x_{it} = (G_{it}^+, G_{it}^=, P_{it}^+, P_{it}^=, A_{it}^+, A_{it}^=)^t$  and  $F(x) = 1/(1 + \exp(-x))$ .

We decompose both threshold parameters  $\theta_{it1}$  and  $\theta_{it2}$  into trend parameters  $\tau_t$ , seasonal parameters  $\gamma_t$ , and unit-specific parameters  $b_i$ , one for each threshold:

$$\theta_{itj} = \tau_{tj} + \gamma_{tj} + b_{ij}, \quad j = 1, 2.$$

Note that the threshold parameters have to follow the restriction  $\theta_{it1} < \theta_{it2}$  for all combinations of  $t$  and  $i$ . A seasonal model (8.1.9) with period 12 was chosen for the seasonal parameters of both thresholds. First-order random walk priors are assigned to all covariate effect parameters  $\beta_t$  and to both trend parameters  $\tau_{t1}, \tau_{t2}$ . All time-changing parameters are assumed to be mutually independent with proper but highly dispersed inverse gamma hyperpriors ( $a = 1, b = 0.005$ ). The firm-specific parameters  $b_i = (b_{i1}, b_{i2})'$  are assumed to follow a Gaussian distribution with mean zero and dispersion  $D$ . We used the parameter values  $\zeta = 1$  and  $\Psi = \text{diag}(0.005, 0.005)$  for the inverted Wishart hyperprior specification for  $D$ .

This model can be written as a dynamic mixed model with

$$\pi_{it} = h(\eta_{it}) = h(Z_{it}\alpha_t + W_{it}b_i),$$

where  $\alpha'_t = (\tau_{t1}, \gamma_{t1}, \tau_{t2}, \gamma_{t2}, \beta'_t)$ ,

$$W_{it} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and

$$Z_{it} = \begin{pmatrix} 1 & 1 & 0 & 0 & x'_{it} \\ 0 & 0 & 1 & 1 & x'_{it} \end{pmatrix}.$$

The response variable  $y_{it}$  is multinomially distributed

$$y_{it} \sim M_2(1, \pi_{it}),$$

where  $y_{it} = (1, 0)', (0, 1)',$  or  $(0, 0)'$ , if the first (+), second (=), or third (-) category is observed. The link function  $h$  is given by

$$h(\eta_{it}) = \begin{pmatrix} F(\eta_{it1}) \\ F(\eta_{it2}) - F(\eta_{it1}) \end{pmatrix}.$$

Figure 8.9 displays the temporal pattern of the trend parameters  $\tau_{tj}$ ,  $j = 1, 2$ , and of both threshold parameters  $\theta_{tj} = \tau_{tj} + \gamma_{tj}$ ,  $j = 1, 2$ . The first trend parameter is slightly decreasing while the second remains constant over the whole period. A distinct seasonal pattern can be seen with higher probabilities of positive response in spring and negative response in fall. However, firm-specific deviations from this pattern are substantial as Figure 8.10 shows. Here, posterior median estimates of the first and second firm-specific parameter  $b_{i1}$  and  $b_{i2}$  are plotted against each other for all 51 firms. Interestingly, these two parameters are often highly negatively correlated. The estimated dispersion matrix of the random effect distribution is

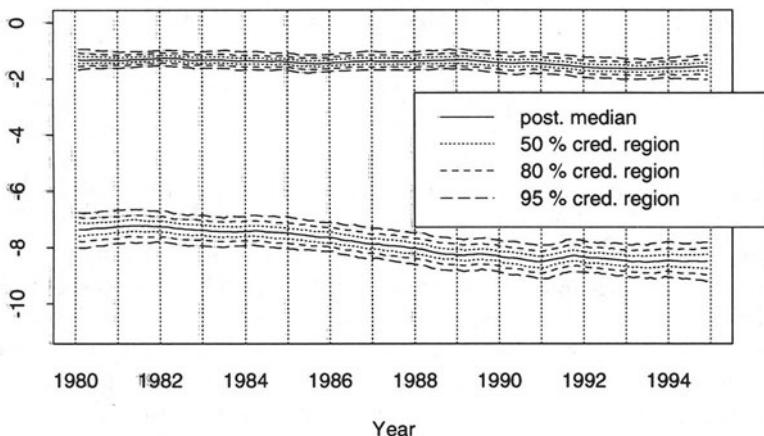
$$\hat{D} = \begin{pmatrix} 0.78 & -0.28 \\ -0.28 & 0.23 \end{pmatrix},$$

and the estimated correlation, based on posterior samples of the corresponding functional of  $D$ , is  $-0.67$ . Both estimates are posterior median estimates. It seems that some firms are more conservative in their answers and often choose “no change” for the response variable, relative to the overall frequencies. Such firms have negative values for  $b_{i1}$  and positive values for  $b_{i2}$ . Other firms avoid the category “no change” and answer often more extremely with “decrease” or “increase.” For these firms,  $b_{i1}$  is positive and  $b_{i2}$  negative.

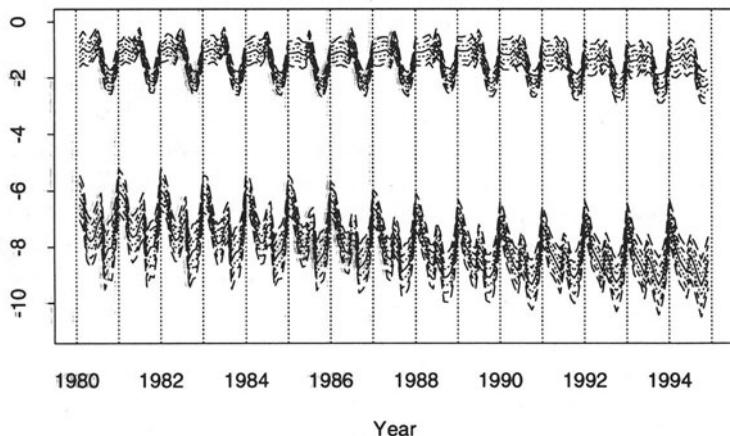
The estimated patterns of time-dependent covariate effects (Figure 8.11) show an interesting temporal pattern, in particular the effect of the dummy  $G+$  (Figure 8.12), which stands for expected improved business conditions, relative to  $G-$ : A distinct low can be seen at the end of 1981, when the German economy was shaken by a recession. In 1982 a new government under the leadership of Chancellor Helmut Kohl was established. From that time onward the effect increased until 1989/1990, with some additional variation, and can be interpreted as a growing trust in the government.

The peak in 1989/1990 coincides with the German reunification, which was expected to have a catalytic effect on the economy due to the sudden opening of the market in the former East Germany. In the years 1986, 1990 and 1994, parliament elections were held in fall. In these years the effect is always decreasing toward the end of the year, which may be due to the uncertainty regarding the election results.  $\square$

### Estimates of trend components



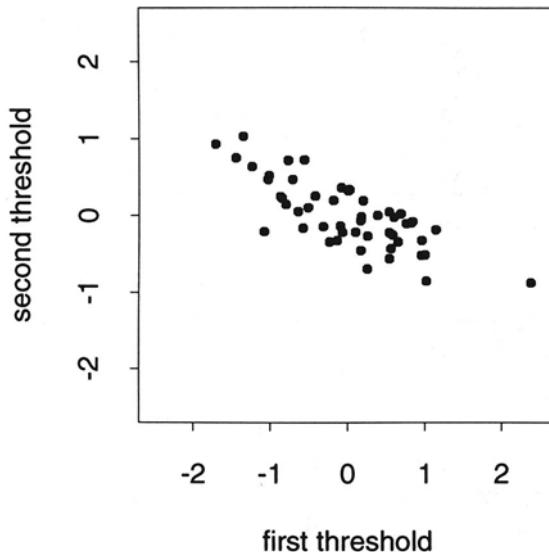
### Estimates of threshold parameters



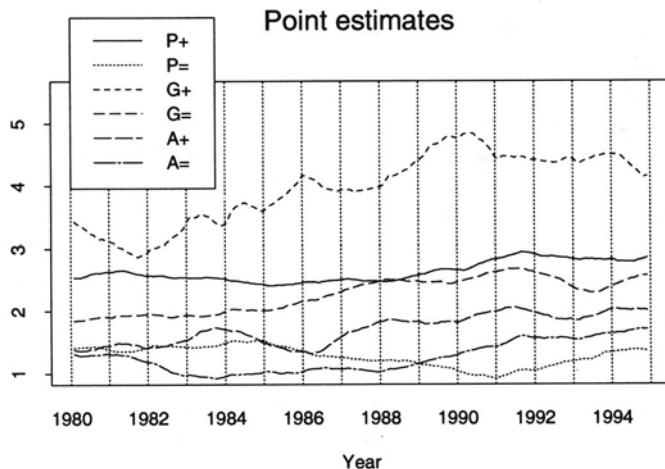
**Figure 8.9.** Estimates of trends and thresholds. Dashed vertical lines represent the month January of each year.

## 8.5 Spatial and Spatio-temporal Data

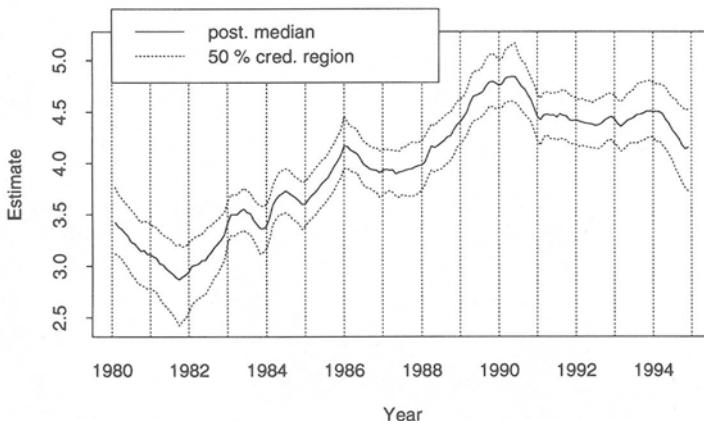
In various applications, individuals or units belong to a specific spatial site or location  $s \in \{1, \dots, S\}$ . In image analysis, units are pixels,  $s$  denotes the site of a pixel on a regular grid  $\{1, \dots, S\}$ , and the observed responses are signals measured at each pixel. In disease mapping, units are districts or



**Figure 8.10.** Plot of the estimates of  $b_{i1}$  against  $b_{i2}$  for each unit.



**Figure 8.11.** Estimated time-changing covariate effects. Dashed vertical lines represent the month January of each year.



**Figure 8.12.** Estimated time-changing covariate effect of  $G^+$ . Dashed vertical lines represent the month January of each year.

regions  $s \in \{1, \dots, S\}$  of a country, and the observed response is the number of deaths from a disease in district  $s$  within a year. In geographical statistics, for example, in an unemployment study, the district  $s \in \{1, \dots, S\}$ , where individual  $i$  lives is known, together with covariates  $x_i$  characterizing  $i$ . In analogy to time series or longitudinal data, it may be necessary to include parameters that model spatial correlation. The general idea for formulation of appropriate spatial priors is that units close to each other are more alike than two arbitrary units. This calls for a generalization of random walk or autoregressive priors to situations with two or more dimensions. Since space is unordered, whereas time is ordered, Markov processes are replaced by Markov random fields. In this section we first consider simple spatial priors as a generalization of random walk models. More details on the theory of Markov random fields and their application in spatial statistics can be found in Besag (1974), Cressie (1993), and Besag & Kooperberg (1995).

Spatio-temporal data are obtained if a sequence of images is observed as in functional magnetic resonance imaging, if the number of deaths from a disease in districts  $s \in \{1, \dots, S\}$  is observed in successive years  $t = 1, \dots, T$ , or if the employment status of individuals living in district  $s_i$  is observed for each month  $t = 1, \dots, T$ . Hidden Markov models for such space-time data can be constructed by appropriate combination of state space models for time series or longitudinal data and spatial models.

In the following, the basic concepts are explained with simple models. Let  $y_s$ ,  $s = 1, \dots, S$ , denote observations on the spatial array  $\{1, \dots, S\}$ . Within the generalized linear modelling framework we assume that the (conditional) mean of  $y_s$  is linked to a predictor  $\eta_s$  by  $E(y_s | \eta_s) = h(\eta_s)$ . The simplest extension to the common linear predictor is then

$$\eta_s = w'_s \beta + \theta_s, \quad (8.5.1)$$

where  $\theta_s$  is a spatial random effect for pixel or region  $s$  and  $w'_s \beta$  is the usual linear part of the predictor, with covariate vectors  $w_s$  characterizing site  $s$ . For  $\beta$  usually a diffuse prior is chosen. For  $\theta_s$ ,  $s \in \{1, \dots, S\}$ , an intuitive generalization of simple first-order random walk priors

$$p(\alpha | \tau_\alpha^2) \propto \exp \left( -\frac{1}{2} \sum_{t=2}^T \frac{(\alpha_t - \alpha_{t-1})^2}{\tau_\alpha^2} \right) \quad (8.5.2)$$

for a time trend  $\alpha = (\alpha_1, \dots, \alpha_T)$  is as follows: Replace neighbors  $t$  and  $t-1$  in one dimension by spatial neighbors or adjacent units. Denoting  $s \sim j$ , if unit  $s$  is adjacent to unit  $j$ , a spatial generalization of (8.5.2) for the block  $\theta = (\theta_1, \dots, \theta_s)$  of spatial effects is the Markov random field prior

$$p(\theta | \tau_\theta^2) \propto \exp \left( -\frac{1}{2} \sum_{s \sim j} \frac{(\theta_s - \theta_j)^2}{\tau_\theta^2} \right) \quad (8.5.3)$$

$$= \exp \left( -\frac{1}{2\tau_\theta^2} \theta' K_\theta \theta \right). \quad (8.5.4)$$

The elements of the penalty matrix  $K_\theta = (k_{sj})$  are given by

$$k_{sj} = \begin{cases} m_s, & \text{if } j = s, \\ -1, & \text{if } s \sim j, \\ 0, & \text{otherwise,} \end{cases}$$

where  $m_s$  is the number of units adjacent to  $s$ .

Following the terminology of Künsch (1987) and Besag, York & Mollié (1991), such a prior is called a Gaussian intrinsic autoregression. From (8.5.3) we can derive the conditional distribution of  $\theta_s$  given the rest  $\theta_{j \neq s}$  of parameters as

$$\theta_s | \theta_{j \neq s} \sim N \left( \sum_{j \in \delta_s} \theta_j / m_s, \tau_\theta^2 / m_s \right), \quad (8.5.5)$$

where  $j \in \delta_s$  denotes that site  $j$  is a neighbor of site  $s$ . The conditional representation (8.5.5) shows the Markov property of the random field  $\theta$ , because the conditional distribution of  $\theta_s | \theta_{j \neq s}$  depends only on the neighbors. Just as  $\tau_\alpha^2$  for the random walk model (8.5.2), the variance  $\tau_\theta^2$  controls spatial smoothness of  $\theta$ . For a fully Bayesian analysis, we assign an inverse Gamma prior  $\tau_\theta^2 \sim \text{IG}(a_\theta, b_\theta)$ .

A number of variations and modifications are conceivable for the priors (8.5.3, 8.5.5). First, there are different possibilities for defining the neighborhood relationship  $s \sim j$ . For example, on a regular grid as in image

analysis, we may take only the four nearest pixels in the  $x$ - and  $y$ -directions as neighbors, or we may additionally consider the next four neighbors in diagonal directions. Second, other priors are possible, e.g., priors based on weighted distances  $w_{sj}(\theta_s - \theta_j)^2$ , on other distances such as  $|\theta_s - \theta_j|$ , or on priors for 0–1 parameters  $\theta_s$ ; see Besag, York & Mollie (1991), and Besag, Green, Higdon & Mengersen (1995).

Furthermore, various extensions of the simple predictor (8.5.1) are conceivable. For example, we may add spatially uncorrelated, exchangeable random effects  $\phi = (\phi_1, \dots, \phi_s)'$  with prior

$$\phi | \tau_\phi \sim N(0, \tau_\phi^2 I)$$

to (8.5.1). In the predictor

$$\eta_s = w'_s \beta + \theta_s + \phi_s, \quad (8.5.6)$$

$\theta_s$  and  $\phi_s$  represent spatially structured and unstructured effects of site  $s$ , respectively.

Assume now that individual data are given by

$$(y_i, x_i, w_i, s_i), \quad i = 1, \dots, n,$$

where  $x_i$  is a vector of continuous covariates whose effects are to be modelled nonparametrically,  $w_i$  is a further vector of covariates with linear effects, and  $s_i \in \{1, \dots, S\}$  denotes the site or location of unit  $i$ . For example, in the following application  $y_i$  is the rent for flat  $i$  in Munich,  $x_{i1}$  and  $x_{i2}$  are floor space and year of construction,  $w_i$  is a vector of binary indicators, and  $s_i$  is the subquarter in Munich where the flat is located. An appropriate predictor might be in the form

$$\eta_i = \sum_{j=1}^P f_{(j)}(x_{ij}) + w'_i \beta + \theta_{s_i} + b_{s_i}, \quad (8.5.7)$$

of generalized additive mixed models with an exchangeable random effect  $b_{s_i}$ , as in Section 7.6 and an additional, spatially structured effect  $\theta_{s_i}$ . As in Section 5.4, equation (5.4.12), we assume smoothness priors

$$p(f_j | \tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2} f'_j K_j f_j\right), \quad j = 1, \dots, p,$$

for the vectors  $f_j$  of function evaluations.

Inference can be based, at least in principle, on approximate penalized likelihood approaches, as in Breslow & Clayton (1993), and Lin & Zhang (1999), or on fully Bayesian approaches via MCMC. While penalized likelihood approaches are still feasible in simpler models, fully Bayesian approaches are better suited for complex models like (8.5.7). Since the priors

for functions  $f_j$  and the prior (8.5.3) for the spatial effect  $\theta$  have the same structure with sparse penalty matrices  $K_j$  and  $K_\theta$ , respectively, posterior sampling via MCMC as described in Sections 5.4 and 7.6 can be extended to semiparametric models (8.5.7) with spatial effects by adding further full conditionals to the sampling algorithms in those sections.

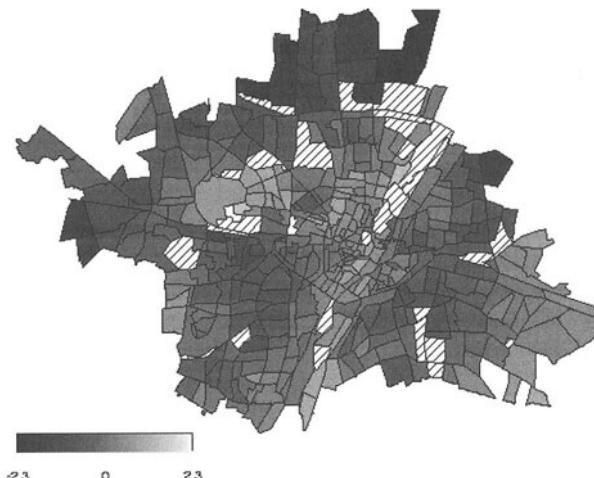
**Example 8.7: Rents for flats** (Example 5.9, continued)

We reanalyze the data on rents for flats, but without using the evaluation of experts for location of a flat into one of the three categories average, good, and top. Instead we let the data speak and consider location as a spatial covariate  $s$ , where  $s \in \{1, \dots, S\}$  denotes the subquarter in Munich where the flat is located. Compared to Example 5.9, we now choose a Gaussian model with predictor

$$\eta = \alpha + f_{(1)}(F) + f_{(2)}(A) + w'\beta + f_{(3)}(\theta_s),$$

where the parametric effects  $\beta_1 L_1 + \beta_2 L_2$  of location are replaced by the spatially structured effect  $f_{(3)}(\theta_s)$ , modelled by the Markov random field (8.5.3).

Since the estimates for  $f_1(F)$ ,  $f_2(A)$ , and  $w'\beta$  are quite close to those already given in Example 5.9, we show here only a map of Munich (see Figure 8.13), displaying subquarters and the posterior means of corresponding spatial effects.  $\square$



**Figure 8.13.** Posterior means of spatial effects.

Extensions of state space modelling for spatio-temporal data  $y_{st}$ ,  $s = 1, \dots, S, t = 1, \dots, T$ , have gained much interest recently. For example, in

disease mapping,  $y_{st}$  is the number of deaths for district  $s$  during year  $t$ . If  $n_{st}$  is the number of persons at risk, then a binomial model  $y_{st} \sim B(\pi_{st}, n_{st})$ , conditional on a predictor  $\eta_{st}$ , is a reasonable assumption. Knorr-Held & Besag (1998) assume a logit model

$$\pi_{st} = \exp(\eta_{st}/(1 + \exp(\eta_{st})))$$

and extend the predictor (8.5.7) additively to

$$\eta_{st} = \mu + \alpha_t + \gamma_t + \theta_s + \phi_s. \quad (8.5.8)$$

In (8.5.8)  $\mu$  is an overall risk lever (or more generally a linear term  $w'_s \beta$ ),  $\alpha_t$ ,  $t = 1, \dots, T$ , is a time trend following a first- or second-order random walk,  $\gamma_t$  are additional temporal but unstructured i.i.d  $N(0, \tau_\gamma^2)$  effects, and  $\theta_s$  and  $\phi_s$  are spatially structured and unstructured effects, respectively. Therefore, each of the four blocks  $\alpha = (\alpha_1, \dots, \alpha_T)', \gamma = (\gamma_1, \dots, \gamma_T)', \theta = (\theta_1, \dots, \theta_S)',$  and  $\phi = (\phi_1, \dots, \phi_S)'$  has a multivariate Gaussian prior with mean zero and appropriate structure matrix  $K/\tau^2$ . With  $RW(1)$  and  $RW(2)$  models for  $\alpha$ , the structure or penalty matrix  $K_\alpha$  has the banded form in (8.3.6) or (8.3.7),  $K_\theta$  is as in (8.5.3), and  $K_\gamma = K_\phi = I$ . With inverse gamma priors for variances, posterior sampling by MCMC schemes can be extended to spatio-temporal models.

The above formulation is separable in space and time and requires appropriate expansion to allow for space-time interactions. The most natural way is to add an interaction effect  $\delta_{st}$  to (8.5.8) so that

$$\eta_{st} = \mu + \alpha_s + \gamma_t + \theta_s + \phi_s + \delta_{st}, \quad (8.5.9)$$

and assume a Gaussian prior with structure matrix  $K_\delta$  for the vector  $\delta = (\delta_{11}, \dots, \delta_{ST})'$ . Knorr-Held (2000) follows a suggestion of Clayton (1996) and specifies  $K_\delta$  as the Kronecker product of the structure matrices of those main effects that are assumed to interact. For example, if the temporal main effect  $\alpha$  and the unstructured spatial effect  $\phi$  interact, this leads to

$$\begin{aligned} p(\delta | \tau_\delta^2) &\propto \exp\left(-\frac{1}{2\tau_\delta^2} \delta' K_\alpha \otimes K_\phi \delta\right) \\ &= \exp\left(-\frac{1}{2\tau_\delta^2} \sum_{s=1}^S \sum_{t=2}^T (\delta_{st} - \delta_{s,t-1})^2\right). \end{aligned}$$

This model will be suitable if temporal trends are different for all sites but do not have any spatial structure.

The most complex form of interaction arises between the main effects  $\alpha$  and  $\theta$ . Then  $\delta$  can no longer be factored into independent blocks, and the prior can be written as

$$p(\delta | \tau_\delta^2) \propto \exp\left(-\frac{1}{2\tau_\delta^2} \sum_{t=2}^T \sum_{s \sim j} (\delta_{st} - \delta_{jt} - \delta_{s,t-1} - \delta_{j,t-1})^2\right).$$

Details with computational issues and an application to the Ohio lung cancer data can be found in Knorr-Held (2000).

Similar extensions are possible for individual space-time data

$$(y_{it}, x_{it}, w_{it}, s_i) \quad i = 1, \dots, n; \quad t = 1, \dots, T,$$

where  $y_{it}$  is the response of individual  $i$  at time  $t$ , living in district  $s_i$ , and having covariate values  $x_{it}, w_{it}$  that may be time-dependent or time-constant. In this case model (8.5.7) is expanded again by adding temporal effects and, possibly, interaction effects. We will consider an application to unemployment data in Section 9.4.

## 8.6 Notes and Further Reading

Classical state space models and hidden Markov models have a long tradition in engineering, time series analysis, and speech recognition. Extensions to more complex situations have become a focus of interest more recently. A main reason is their flexibility to deal with many nonlinear and non-Gaussian situations in a number of challenging applications, for example, molecular biology and genetics, image analysis, and spatial econometrics. We could touch on only some of the recent developments, in particular for spatial and spatio-temporal data. The comprehensive book by Cressie (1993) covers many topics in spatial statistics beyond the Markov random field approach. Winkler (1995) is a valuable source for image analysis.

# 9

## Survival Models

In recent years the analysis of survival time, lifetime, or failure time data has received considerable attention. The methodology applies in medical trials, where survival is of primary interest, and in reliability experiments, where failure time is the duration of interest. We will mostly refer to survival time although, in principle, situations where the time until the occurrence of some event is of interest are considered.

There is a considerable number of excellent books on survival analysis. In particular, the case of continuous time is treated extensively in Lawless (1982), Kalbfleisch & Prentice (1980), Blossfeld, Hamerle & Mayer (1989), and Lancaster (1990). A standard reference for the counting process approach is Andersen, Borgan, Gill & Keiding (1993). In the following, models for continuous survival time, which are covered in these just-mentioned books, are only sketched.

In applications time is often measured as a discrete variable. For example, in studies on the duration of unemployment (see Example 9.1) the time of unemployment is most often given in months. Here the focus is on models that may be estimated similarly to generalized linear models. Extensive treatment is given for the case of grouped or discrete survival time.

### 9.1 Models for Continuous Time

#### 9.1.1 Basic Models

Survival time is considered a nonnegative random variable  $T$ . For  $T$  continuous, let  $f(t)$  denote the probability density function and  $F(t)$  denote the corresponding distribution function given by

$$F(t) = P(T \leq t) = \int_0^t f(u)du.$$

The probability of an individual surviving until time  $t$  is given by the so-called *survival function*

$$S(t) = P(T > t) = 1 - F(t),$$

which is sometimes also called the *reliability function*. A basic concept in the analysis of survival time is the *hazard function*  $\lambda(t)$ , which is defined as the limit

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \quad (9.1.1)$$

The hazard function measures the instantaneous rate of death or failure at time  $t$  given that the individual survives until  $t$ . It is sometimes also useful to consider the cumulative hazard function

$$\Lambda(t) = \int_0^t \lambda(u) du.$$

The connection between distribution function, density, hazard function, and cumulative hazard function is given by the equations

$$\begin{aligned} \lambda(t) &= \frac{f(t)}{S(t)}, \\ S(t) &= \exp \left( - \int_0^t \lambda(u) du \right) = \exp(-\Lambda(t)), \\ f(t) &= \lambda(t) \exp \left( - \int_0^t \lambda(u) du \right) = \lambda(t) \exp(-\Lambda(t)) \end{aligned}$$

(see, e.g., Lawless, 1982). The distribution of  $T$  is determined completely by one of these quantities. Basic models illustrate the relationships among these basic concepts.

## Exponential Distribution

Assume a constant hazard function

$$\lambda(t) = \lambda, \quad t \geq 0,$$

where  $\lambda > 0$ . Then the probability density function is given by

$$f(t) = \lambda e^{-\lambda t},$$

which is the density of an exponential distribution with parameter  $\lambda$ . The expectation is given by  $1/\lambda$  and the variance by  $1/\lambda^2$ .

## Weibull Distribution

Let the hazard function be determined by

$$\lambda(t) = \lambda\alpha(\lambda t)^{\alpha-1},$$

where  $\lambda > 0, \alpha > 0$  are parameters sometimes referred to as shape and scale parameters. Equivalently, one may consider the Weibull density function

$$f(t) = \lambda\alpha(\lambda t)^{\alpha-1} \exp(-(\lambda t)^\alpha) \quad (9.1.2)$$

or the survival function

$$S(t) = \exp(-(\lambda t)^\alpha).$$

Characterized by two parameters, the distribution is more flexible than the exponential model, which is included as the special case where  $\alpha = 1$ . The hazard function is increasing for  $\alpha > 1$ , decreasing for  $\alpha < 1$ , and constant for  $\alpha = 1$ . Expectation and variance are given by

$$\begin{aligned} E(T) &= \Gamma\left(\frac{1+\alpha}{\alpha}\right) / \lambda, \\ \text{var}(T) &= \left( \Gamma\left(\frac{\alpha+2}{\alpha}\right) - \Gamma\left(\frac{\alpha+1}{\alpha}\right)^2 \right) / \lambda^2. \end{aligned}$$

If  $T$  is Weibull distributed with parameters  $\lambda$  and  $\alpha$ , the transformation  $Y = \log T$  has the extreme-value distribution with density function

$$f(y) = \frac{1}{\sigma} \exp[(y-u)/\sigma - \exp((y-u)/\sigma)] \quad (9.1.3)$$

and distribution function

$$F(y) = 1 - \exp[-\exp((y-u)/\sigma)],$$

where  $u = -\log \lambda$  and  $\sigma = 1/\alpha$  are location and scale parameters. For  $u = 0, \sigma = 1$ , (9.1.3) is the standard (minimum) extreme-value function.

### Piecewise Exponential Model

Let the time axis  $[a_0, \infty)$  be divided into  $k$  intervals  $[a_0, a_1], (a_1, a_2], \dots, (a_{s-1}, a_s], \dots, (a_q, \infty)$ ,  $q = k - 1$ . A nonnegative random variable  $T \geq 0$  is piecewise exponential on the grid  $a' = (a_0, a_1, \dots, a_q)$ , for short  $T \sim \text{PE}(\lambda, a)$  if the hazard rate is piecewise constant,

$$\lambda(t) = \begin{cases} \lambda_s, & t \in (a_{s-1}, a_s], \quad s = 1, \dots, q, \\ \lambda_k, & t > a_q. \end{cases}$$

A piecewise exponential model for survival data assumes that survival times  $T_1, \dots, T_n$  are i.i.d.  $\text{PE}(\lambda, a)$  distributed. For a given grid, the parameters  $\lambda' = (\lambda_1, \dots, \lambda_s, \dots, \lambda_q)$  are unknown and have to be estimated from observed survival times.

The survival and density function are characterized by

$$\begin{aligned} S(t|T \geq a_{s-1}) &= \exp(-\lambda_s(t - a_{s-1})), \quad t \in (a_{s-1}, a_s], \\ f(t|T \geq a_{s-1}) &= \lambda_s \exp(-\lambda_s(t - a_{s-1})), \quad t \in (a_{s-1}, a_s]. \end{aligned}$$

#### 9.1.2 Parametric Regression Models

Now let  $x' = (x_1, \dots, x_p)$  denote a set of (time-independent) covariates that influence the lifetime  $T$ . Then one has to consider the distribution of  $T$ , given  $x$ , and thus the population density  $f(t)$ , hazard function  $\lambda(t)$ , and survival function  $S(t)$  become density  $f(t|x)$ , hazard function  $\lambda(t|x)$ , and survival function  $S(t|x)$ , respectively.

#### Location-Scale Models for $\log T$

There are several approaches to the construction of regression models for lifetime data. *Location-scale models* for the log-lifetime  $Y = \log T$  have the form

$$\log T = \mu(x) + \sigma \varepsilon, \tag{9.1.4}$$

where  $\sigma$  is a constant scale parameter and  $\varepsilon$  is a noise variable independent of  $x$ . If  $\varepsilon$  follows the standard extreme-value distribution (9.1.3) with  $u = 0, a = 1$ , then  $Y = \log T$  has density function

$$f(y|x) = \frac{1}{\sigma} \exp \left[ \frac{y - \mu(x)}{\sigma} - \exp \left( \frac{y - \mu(x)}{\sigma} \right) \right],$$

and the lifetime  $T$  is Weibull distributed with density function

$$f(t|x) = \frac{1}{\sigma} \exp[-\mu(x)] [t \exp(-\mu(x))]^{1/\sigma-1} \exp\left[-(t \exp(-\mu(x)))^{1/\sigma}\right]. \quad (9.1.5)$$

That means the shape parameter  $\lambda$  in (9.1.2) is specified by  $\lambda = \exp(-\mu(x))$  and the scale parameter  $\alpha$  in (9.1.2) equals a constant  $\alpha = 1/\sigma$ . Only the shape parameter is influenced by the covariates, whereas the scale parameter is independent of  $x$ . Models where  $\alpha$  is normally distributed or follows a log-gamma distribution are considered extensively in Lawless (1982).

### Proportional Hazards Models

The Weibull distribution model with covariates given by (9.1.5) has the hazard function

$$\lambda(t|x) = \frac{\exp(-\mu(x))}{\sigma} [t \exp(-\mu(x))]^{1/\sigma-1}$$

or, equivalently,

$$\lambda(t|x) = \frac{t^{1/\sigma-1}}{\sigma} \exp\left[\frac{-\mu(x)}{\sigma}\right]. \quad (9.1.6)$$

For two subpopulations or individuals characterized by  $x_1$  and  $x_2$ , it is immediately seen that

$$\frac{\lambda(t|x_2)}{\lambda(t|x_1)} = \exp\left[\frac{\mu(x_2) - \mu(x_1)}{\sigma}\right].$$

That means that the ratio  $\lambda(t|x_1)/\lambda(t|x_2)$  does not depend on time  $t$ . If the hazard for the first individual is twice the hazard for the second individual after 1 year, the ratio is the same after 2 years, 2.5 years, etc; the ratio of risks is the same at any time. Models that have this property are called *proportional hazards models*. The Weibull distribution model is a location-scale model of type (9.1.4) and a proportional hazards model. However, in general, location-scale models for  $\log T$  do not show proportional hazards. Assuming, for example, a normal distribution for  $\varepsilon$  in (9.1.4) yields hazards that are not proportional over time. A very general proportional hazards model due to Cox (1972) is given by

$$\lambda(t|x) = \lambda_0(t) \exp(x'\gamma), \quad (9.1.7)$$

where the baseline hazard function  $\lambda_0(t)$  is assumed to be the same for all observations but is not assumed to be known. In contrast to the Weibull model (9.1.6), no specific structure is assumed for the baseline hazard. Estimation of  $\gamma$  may be based on marginal likelihood (Kalbfleisch & Prentice, 1973) or on the concept of partial likelihood (Cox, 1972, 1975; Tsiatis, 1981; Prentice & Self, 1983).

## Linear Transformation Models and Binary Regression Models

Linear transformation models have the general form

$$h(T) = x'\gamma + \varepsilon, \quad (9.1.8)$$

where  $h$  is an increasing continuous function and  $\varepsilon$  is a random error variable with distribution function  $F_\varepsilon$ . Obviously, location-scale models for the log-lifetime are special cases where  $h = \log$ . It is immediately seen that (9.1.8) is equivalent to

$$P(T \leq t|x) = F_\varepsilon(h(t) - x'\gamma)$$

for all  $t$ . For the logistic distribution function  $F_\varepsilon(z) = 1/(1 + \exp(-z))$ , one gets the *proportional odds model*

$$\log \frac{P(T \leq t|x)}{P(T > t|x)} = h(t) - x'\gamma,$$

which for fixed  $t$  may be considered a binary response model with response

$$Y_t = \begin{cases} 1 & \text{if } T \leq t, \\ 0 & \text{if } T > t. \end{cases}$$

The connection between Cox's proportional hazards model and linear transformation models is seen if for  $F_\varepsilon$  the extreme-value distribution  $F_\varepsilon(z) = 1 - \exp(-\exp(z))$  is assumed. Then one has

$$\log(-\log P(T > t|x)) = h(t) - x'\gamma. \quad (9.1.9)$$

Let  $h(t)$  be defined by

$$h(t) = \log \int_0^t \lambda_0(s)ds,$$

where  $\lambda_0(t)$  is a function fulfilling  $\lambda_0 \geq 0$ . Then one gets for the hazard function

$$\lambda(t|x) = \lambda_0(t) \exp(-x'\gamma), \quad (9.1.10)$$

where  $\lambda_0$  is the baseline hazard function. Equivalently, one has

$$S(t|x) = S_0(t)^{\exp(-x'\gamma)},$$

where  $S_0(t) = \exp\left(-\int_0^t \lambda_0(s)ds\right)$  is the baseline survival function. Obviously the model is equivalent to the Cox model (9.1.7); only the sign of the parameter vector  $\gamma$  has changed. Therefore, the Cox model is a linear transformation model with unknown transformation  $h$  given the baseline hazard function  $\lambda_0(s)$  is unspecified (see also Doksum & Gasko, 1990). In fact, the class of models defined by (9.1.10) is invariant under the group of differentiable strictly monotone increasing transformations on  $t$ . If  $g$  is a transformation, the hazard of  $t' = g(t)$  given by  $\lambda_0(g^{-1}(t'))\partial g^{-1}(t')/\partial t' \exp(-x'\gamma)$  is again of type (9.1.10) (see Kalbfleisch & Prentice, 1973).

### 9.1.3 Censoring

In survival analysis, most often only a portion of the observed times can be considered exact lifetimes. For the rest of the observations one knows only that the lifetime exceeds a certain value. This feature is referred to as censoring. More specifically a lifetime is called right-censored at  $t$  if it is known that the life time is greater than or equal to  $t$  but the exact value is not known. There are several types of censoring due to the sampling situation.

#### Random Censoring

The concept of random censoring is often assumed to hold in observation studies over time. It is assumed that each individual (unit)  $i$  in the study has a lifetime  $T_i$  and a censoring time  $C_i$  that are independent random variables. The observed time is given by  $t_i = \min(T_i, C_i)$ . It is often useful to introduce an indicator variable for censoring by

$$\delta_i = \begin{cases} 1 & \text{if } T_i < C_i, \\ 0 & \text{if } T_i \geq C_i. \end{cases}$$

The data may now be represented by  $(t_i, \delta_i)$ . Let  $f_c$ ,  $S_c$  denote the density function and survival function for the censoring variable  $C$ . Then the likelihood for an uncensored observation  $(t_i, \delta_i = 1)$  is given by

$$f_i(t_i)S_c(t_i)$$

as the product of the lifetime density at  $t_i$  and the probability for censoring time greater than  $t_i$  is given by  $P(C_i > t_i) = S_c(t_i)$ . For a censored observation  $(t_i, \delta_i = 0)$ , the likelihood is given by

$$f_c(t_i)S_i(t_i)$$

as the product of the censoring density at  $t_i$  and the probability of lifetimes greater than  $t_i$  is given by  $P(T_i > t_i) = S_i(t_i)$ . Combined into a single expression, the likelihood for observation  $(t_i, \delta_i)$  is given by

$$[f_i(t_i)S_c(t_i)]^{\delta_i} [f_c(t_i)S_i(t_i)]^{1-\delta_i} = [f_i(t_i)^{\delta_i} S_i(t_i)^{1-\delta_i}] [f_c(t_i)^{1-\delta_i} S_c(t_i)^{\delta_i}].$$

The likelihood for the sample  $(t_i, \delta_i), i = 1, \dots, n$ , is given by

$$L = \prod_{i=1}^n [f_i(t_i)^{\delta_i} S_i(t_i)^{1-\delta_i}] [f_c(t_i)^{1-\delta_i} S_c(t_i)^{\delta_i}].$$

If the censoring time is not determined by parameters of interest, i.e., censoring is noninformative, the likelihood may be reduced to

$$L = \prod_{i=1}^n f_i(t_i)^{\delta_i} S(t_i)^{1-\delta_i}.$$

### Type I Censoring

Sometimes life test experiments have a fixed observation time. Exact lifetimes are known only for items that fail by this fixed time. All other observations are right censored. More generally, each item may have a specific censoring time  $C_i$  that is considered fixed in advance. The likelihood for observation  $(t_i, \delta_i)$  is given by

$$L_i = f_i(t_i)^{\delta_i} S_i(C_i)^{1-\delta_i}.$$

Since if  $\delta_i = 0$  the observation  $t_i = \min\{T_i, C_i\}$  has value  $t_i = C_i$ , the likelihood is equivalent to the reduced likelihood for random censoring. In fact, Type I censoring may be considered a special case of random censoring when degenerate censoring times are allowed.

Alternative censoring schemes like Type II censoring, where only a fixed proportion of observations is uncensored, and more general censoring schemes are considered in detail in Lawless (1982).

### 9.1.4 Estimation

If covariates are present the data are given by triples  $(t_i, \delta_i, x_i)$ ,  $i \geq 1$ , where  $t_i$  is the observed time,  $\delta_i$  is the indicator variable, and  $x_i$  denotes the vector of covariates of the  $i$ th observation. Aitkin & Clayton (1980) show how parametric survival models such as the exponential model, the Weibull model, and the extreme-value model are easily estimated within the framework of generalized linear models. Consider the general proportional hazards model

$$\lambda(t|x) = \lambda_0(t) \exp(x'\gamma)$$

with survival function

$$S(t|x) = \exp(-\Lambda_0(t) \exp(x'\gamma))$$

and density

$$f(t|x) = \lambda_0(t) \exp(x'\gamma - \Lambda_0(t) \exp(x'\gamma)),$$

where  $\Lambda_0(t) = \int_0^t \lambda_0(u) du$ . Assuming random censoring, which is noninformative, one gets the likelihood

$$\begin{aligned} L &= \prod_{i=1}^n f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \\ &= \prod_{i=1}^n [\lambda_0(t_i) \exp(x_i'\gamma - \Lambda_0(t_i) \exp(x_i'\gamma))]^{\delta_i} \exp(-\Lambda_0(t_i) \exp(x_i'\gamma))^{1-\delta_i} \\ &= \prod_{i=1}^n \mu_i^{\delta_i} e^{-\mu_i} \left( \frac{\lambda_0(t_i)}{\Lambda_0(t_i)} \right)^{\delta_i}, \end{aligned}$$

where  $\mu_i = \Lambda_0(t_i) \exp(x_i'\gamma)$ . The second term depends only on the baseline hazard function and does not involve the parameter  $\gamma$ . The first term is equivalent to the kernel of the likelihood function of Poisson variates  $\delta_i \sim P(\mu_i)$ . The corresponding log-linear Poisson model, which has the same likelihood function, is given by

$$\log(\mu_i) = \log \Lambda_0(t_i) + x_i'\gamma,$$

which is a linear model with a constant  $\log \Lambda_0(t_i)$ . For specific models considered in the following, the log-likelihood may be maximized in the same way as for GLMs.

### Exponential Model

For the exponential model we have  $\lambda(t) = \lambda_0 \exp(x'\gamma)$  with baseline hazard function  $\lambda_0(t) = \lambda_0$  and  $\Lambda_0(t) = \lambda_0 t$ . The second term in the likelihood  $\lambda_0(t)/\Lambda_0(t) = 1/t$  does not depend on any further parameters. Thus, maximization of the first term will do. This is equivalent to estimate  $\gamma$  for the Poisson model

$$\log(\mu_i) = \log(\lambda_0 t_i) + x'_i \gamma.$$

By taking  $\log \lambda_0$  into the linear term, one considers the model

$$\log(\mu_i) = \log t_i + x'_i \gamma. \quad (9.1.11)$$

Consequently, the log-linear model may be fitted where  $\log t_i$  is included in the regression model with a known coefficient of 1. In GLIM terminology such a variable is called offset (Aitkin, Anderson, Francis & Hinde, 1989).

### Weibull Model

For the Weibull model one has  $\lambda_0(t) = t^{1/\sigma-1}/\sigma$  and  $\Lambda_0(t) = t^{1/\sigma}$ . Now, the second term in the likelihood  $\lambda_0(t)/\Lambda_0(t) = 1/(t\sigma)$  depends on the unknown parameter  $\sigma$ . Instead of  $\sigma$  one may use  $\alpha = 1/\sigma$ . The log-likelihood now is given by

$$l = \sum_{i=1}^n (\delta_i \log(\mu_i) - \mu_i) - \sum_{i=1}^n \delta_i \log t_i + \delta \log \alpha,$$

where  $\delta = \sum_{i=1}^n \delta_i$  is the number of uncensored observations. The Poisson model corresponding to the first term is given by

$$\log(\mu_i) = \alpha \log t_i + x'_i \gamma.$$

From the likelihood equations

$$\begin{aligned} \frac{\partial l}{\partial \gamma} &= \sum_i x_i (\delta_i - \mu_i) = 0, \\ \frac{\partial l}{\partial \alpha} &= \sum_i \log(t_i) (\delta_i - \mu_i) + \frac{\delta}{\alpha} = 0, \end{aligned}$$

one gets for the maximum likelihood estimate

$$\hat{\alpha} = \left( \frac{1}{\delta} \sum_i \log(t_i)(\mu_i - \delta_i) \right)^{-1}. \quad (9.1.12)$$

Aitkin & Clayton (1980) propose estimating  $\gamma$  and  $\alpha$  iteratively by starting with  $\hat{\alpha}^{(0)} = 1$ , i.e., the exponential model. The fitting of model (9.1.11) yields estimates  $\hat{\mu}_i^{(0)}$ . Inserting  $\hat{\mu}_i^{(0)}$  in (9.1.12) yields an estimate  $\tilde{\alpha}^{(0)}$ . Now the Poisson model

$$\log(\mu_i) = \hat{\alpha}^{(1)} \log t_i + x\gamma$$

with offset  $\hat{\alpha}^{(1)} \log t_i$  is fitted where  $\hat{\alpha}^{(1)} = (\hat{\alpha}^{(0)} + \tilde{\alpha}^{(0)})/2$ . This process is continued until convergence. According to Aitkin & Clayton (1980), the damping of the successive estimates of  $\lambda$  improves the convergence.

For the fitting of models based on the extreme-value distributions, see Aitkin & Clayton (1980).

### Piecewise Exponential Model

Defining the sequence  $\gamma_{0s} = \log \lambda_s$  of baseline parameters,  $s = 1, \dots, q$ , the piecewise exponential regression model for survival data with covariates is given by hazard rates

$$\lambda_i(t) = \exp(\gamma_{0s} + x'_i \gamma), \quad t \in (a_{s-1}, a_s], \quad s = 1, \dots, q.$$

With event indicators  $y_{is}$ ,  $s = 1, \dots, q$ ,

$$y_{is} = \begin{cases} 1, & \text{individual } i \text{ fails in } (a_{s-1}, a_s], \\ 0, & \text{individual } i \text{ survives or censored in } (a_{s-1}, a_s], \end{cases}$$

the log-likelihood for  $\theta' = (\gamma_{01}, \dots, \gamma_{0q}, \gamma)$  is

$$l(\theta) = \sum_{s=1}^q \sum_{i \in R_s} (y_{is} \eta_{is} - \Delta_{is} \exp(\eta_{is})). \quad (9.1.13)$$

Here  $R_s$  is the risk set in  $(a_{s-1}, a_s]$ ,  $\eta_{is} = \gamma_{0s} + x'_i \gamma$ , and

$$\Delta_{is} = \max\{0, \min\{a_s - a_{s-1}, t_i - a_{s-1}\}\}$$

is the observed survival time of individual  $i$  in  $(a_{s-1}, a_s]$ .

Formula (9.1.13) can be derived from the general likelihood for survival data, using temporal factorization (Arjas, 1989, Section 4.2), and from the expression for the density and survival function of the exponential model, or from the general form of counting process likelihoods (Andersen & Borgan, 1985; Andersen, Borgan, Gill & Keiding, 1993).

The log-likelihood (9.1.13) is in the form of the log-likelihood of a log-linear Poisson model with offset  $\Delta_{is}$  and can therefore be fitted with software for GLMs.

## 9.2 Models for Discrete Time

Often time cannot be observed continuously; it is only known to lie between a pair of consecutive follow-ups. Data of this kind are known as interval censored. Since many ties occur, these data cause problems when partial likelihood methods for continuous-time models are used.

In the following let time be divided into  $k$  intervals  $[a_0, a_1), [a_1, a_2), \dots, [a_{q-1}, a_q), [a_q, \infty)$  where  $q = k - 1$ . Often for the first interval  $a_0 = 0$  may be assumed and  $a_q$  denotes the final follow-up. Instead of observing continuous time, one observes the discrete time  $T \in \{1, \dots, k\}$  where  $T = t$  denotes failure within the interval  $[a_{t-1}, a_t)$ . The *discrete hazard function* is given by

$$\lambda(t|x) = P(T = t|T \geq t, x), \quad t = 1, \dots, q, \quad (9.2.1)$$

which is a conditional probability for the risk of failure in interval  $[a_{t-1}, a_t)$  given the interval is reached.

The *discrete survival function* for surviving interval  $[a_{t-1}, a_t)$  is given by

$$S(t|x) = P(T > t|x) = \prod_{i=1}^t (1 - \lambda(i|x)). \quad (9.2.2)$$

Alternatively one may consider the probability for reaching interval  $[a_{t-1}, a_t)$  as a survival function. With

$$\tilde{S}(t|x) = P(T \geq t|x) = \prod_{i=1}^{t-1} (1 - \lambda(i|x)), \quad (9.2.3)$$

one gets  $\tilde{S}(t|x) = S(t-1|x)$ . The unconditional probability for failure in interval  $[a_{t-1}, a_t)$  is given by

$$P(T = t|x) = \lambda(t|x) \prod_{i=1}^{t-1} (1 - \lambda(i|x)) = \lambda(t|x) \tilde{S}(t|x). \quad (9.2.4)$$

Assuming covariates that do not depend on time, the data are given in the form  $(t_i, x_i, \delta_i)$ , where  $\delta_i$  is the indicator variable for censoring given by

$$\delta_i = \begin{cases} 1, & \text{failure in interval } [a_{t_i-1}, a_{t_i}), \\ 0, & \text{censoring in interval } [a_{t_i-1}, a_{t_i}). \end{cases}$$

In cases where the intervals depend on the individuals, e.g., if individuals miss visits in a periodic follow-up, the concept of discrete survival time must be somewhat modified (Finkelstein, 1986).

### 9.2.1 Life Table Estimates

A simple way to describe survival data for the total sample or for subpopulations (without reference to covariates) is by the life table. The method is described for discrete time, but it is also a useful nonparametric estimate for continuous-time observations after defining intervals. Let

- $d_r$  denote the number of observed lifetimes in the interval  $[a_{r-1}, a_r)$  (deaths), and
- $w_r$  denote the numbers of censored observations in the interval  $[a_{r-1}, a_r)$  (withdrawals).

The number of observations at risk in the  $r$ th interval is given by

$$n_r = n_{r-1} - d_{r-1} - w_{r-1}, \quad (9.2.5)$$

where with  $d_0 = w_0 = 0$  we have  $n_1 = n$ . Without censoring, the natural estimate for  $\lambda(t|x)$  is given by

$$\hat{\lambda}_t = \frac{d_t}{n_t}. \quad (9.2.6)$$

For  $w_t > 0$  the so-called standard life table estimate takes the withdrawals into account by

$$\hat{\lambda}_t = \frac{d_t}{n_t - w_t/2}. \quad (9.2.7)$$

The latter estimate considers withdrawals being under risk for half the interval. Under censoring, the first estimate (9.2.6) is appropriate if all withdrawals are assumed to occur at the end of the interval  $[a_{t-1}, a_t)$ . If withdrawals are assumed to occur right at the beginning of the interval  $[a_{t-1}, a_t)$ , the appropriate choice is

$$\hat{\lambda}_t = \frac{d_t}{n_t - w_t}. \quad (9.2.8)$$

The standard life table estimate is a compromise between (9.2.6) and (9.2.8). Based on (9.2.2) the probability for surviving beyond  $a_t$  may be estimated by

$$\hat{S}(t) = \prod_{i=1}^t (1 - \hat{\lambda}_i). \quad (9.2.9)$$

Consequently, the estimated probability for failure in interval  $[a_{t-1}, a_t]$  is given by

$$\hat{P}(T = t) = \hat{\lambda}_t \prod_{i=1}^{t-1} (1 - \hat{\lambda}_i).$$

Without censoring ( $w_t = 0$ ) the number of deaths is multinomially distributed with  $(d_1, \dots, d_q) \sim M(n, (\pi_1, \dots, \pi_q))$ , where  $\pi_t = P(T = t)$ . Using (9.2.5) and (9.2.9) yields the simple estimate

$$\hat{S}(t) = \frac{n - d_1 - \dots - d_t}{n},$$

which is the number of individuals surviving beyond  $a_t$  divided by the sample size. Since  $n - d_1 - \dots - d_t \sim B(n, S(t))$ , expectation and variance are given by

$$\begin{aligned} E(\hat{S}(t)) &= S(t), \\ \text{var}(\hat{S}(t)) &= S(t)(1 - S(t))/n, \end{aligned}$$

and the covariance for  $t_1 < t_2$  is given by

$$\text{cov}(\hat{S}(t_1), \hat{S}(t_2)) = \frac{(1 - S(t_1))S(t_2)}{n}.$$

For  $\hat{\lambda}_t$  one gets

$$\begin{aligned} E(\hat{\lambda}_t) &= \lambda(t), \\ \text{var}(\hat{\lambda}_t) &= \lambda(t)(1 - \lambda(t))E(1/n_t), \end{aligned}$$

where it is assumed that  $n_t > 0$ . In the censoring case ( $w_t > 0$ ) we have a multinomial distribution  $(d_1, w_1, \dots, d_q, w_q) \sim M(n, (\pi_1^d, \pi_1^w, \dots, \pi_q^d, \pi_q^w))$ . Considering continuous lifetime  $T$  and censoring time  $C$ , the probabilities are given by

$$\pi_r^d = P(T \in [a_{r-1}, a_r], T \leq C), \quad \pi_r^w = P(C \in [a_{r-1}, a_r], C < T).$$

Since the frequencies  $(d_1/n, w_1/n, \dots, w_q/n)$  are asymptotically normally distributed, the standard life table estimate  $\hat{\lambda}_t = d_t/(n_t - w_t/2)$  is also asymptotically normal with expectation

$$\lambda_t^* = \frac{\pi_t^d}{\pi_t^0 + \pi_t^w/2},$$

where  $\pi_t^0 = E(n_t/n)$ . However, only in the case without withdrawals  $\lambda_t^* = \lambda(t)$ . Thus, the standard life table estimate is not a consistent estimate. Lawless (1982) derives for the asymptotic variance for the random censorship model

$$\widehat{\text{var}}(\hat{\lambda}_t) = \frac{1}{n} (\lambda_t^* - \lambda_t^{*2}) \frac{\pi_t^0 - \pi_t^w/4}{(\pi_t^0 - \pi_t^w/2)(\pi_t^0 - \pi_t^w/2)}.$$

For the covariance  $\text{cov}(\hat{\lambda}_{t_1}, \hat{\lambda}_{t_2}) = 0$  holds asymptotically. Since  $(n_t - w_t/2)/n$  converges to  $\pi_t^0 - \pi_t^w/2$ , the usual estimate

$$\widehat{\text{var}}(\hat{\lambda}_t) = \frac{\hat{\lambda}_t - \hat{\lambda}_t^2}{n_t - w_t/2}$$

will overestimate  $\text{var}(\hat{\lambda}_t)$  if  $\lambda_t$  and  $\lambda_t^*$  are not too different. For  $\hat{S}(t)$  Lawless (1982) derives for large sample sizes

$$\text{var}(\hat{S}(t)) = S^*(t)^2 \sum_{i=1}^t \frac{\text{var}(1 - \hat{\lambda}_t)}{(1 - \lambda_t^*)^2},$$

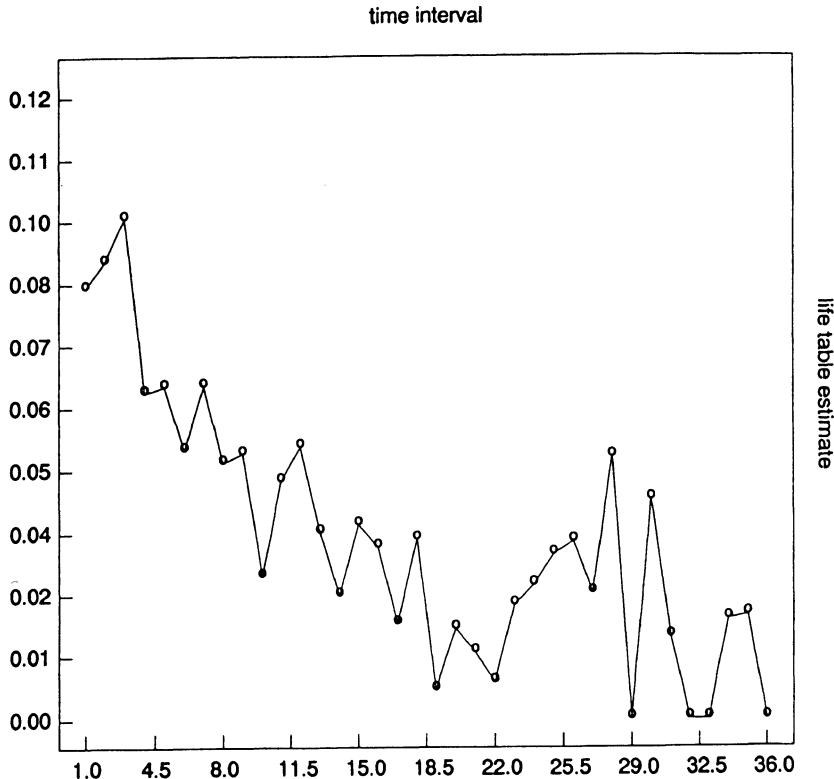
where  $S^*(t) = \prod_{i=1}^t (1 - \lambda_i^*)$ . Approximating  $\text{var}(1 - \hat{\lambda}_t)$  by  $\hat{\lambda}_t(1 - \hat{\lambda}_t)/(n_t - w_t/2)$  and  $S^*(t), \lambda_t^*$  by  $\hat{S}(t), \hat{\lambda}_t$  yields Greenwood's (1926) often-used formula

$$\text{var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{i=1}^t \frac{\hat{\lambda}_t}{(1 - \hat{\lambda}_t)(n_t - w_t/2)}$$

as an approximation.

### Example 9.1: Duration of unemployment

The data set comprises 1669 unemployed persons who are observed from January 1983 until December 1988 in the socioeconomic panel in Germany (Hanefeld, 1987). Time is measured in months. As absorbing state only employment in a full-time job is considered. All other causes for ending unemployment (employment in a part-time job or going back to school) are considered censoring. Figure 9.1 shows the estimated hazard rates based on (9.2.8). Although showing the typical picture of unemployment data (short increase, slow decrease), the estimate is quite irregular. In particular, when the local sample size is small (for long-time unemployment), the curve is quite jagged. Figure 9.2 shows the corresponding survival function based on the life table estimate. The rather smooth survival function gives an estimate for the probability of still being unemployed after  $t$  months.  $\square$



**Figure 9.1.** Life table estimate for unemployment data.

### 9.2.2 Parametric Regression Models

In this section we consider the case where discrete lifetime  $T_i$  depends on a vector of covariates  $x_i$ .

#### The Grouped Proportional Hazards Model

The proportional hazards or Cox model (9.1.7) for continuous time is given by

$$\lambda_c(t|x) = \lambda_0(t) \exp(x'\gamma), \quad (9.2.10)$$

where  $\lambda_c(t|x)$  stands for the continuous hazard function. If time  $T$  is considered a discrete random variable where  $T = t$  denotes failure within the interval  $[a_{t-1}, a_t]$ , the assumption of (9.2.10) yields the grouped proportional hazards model

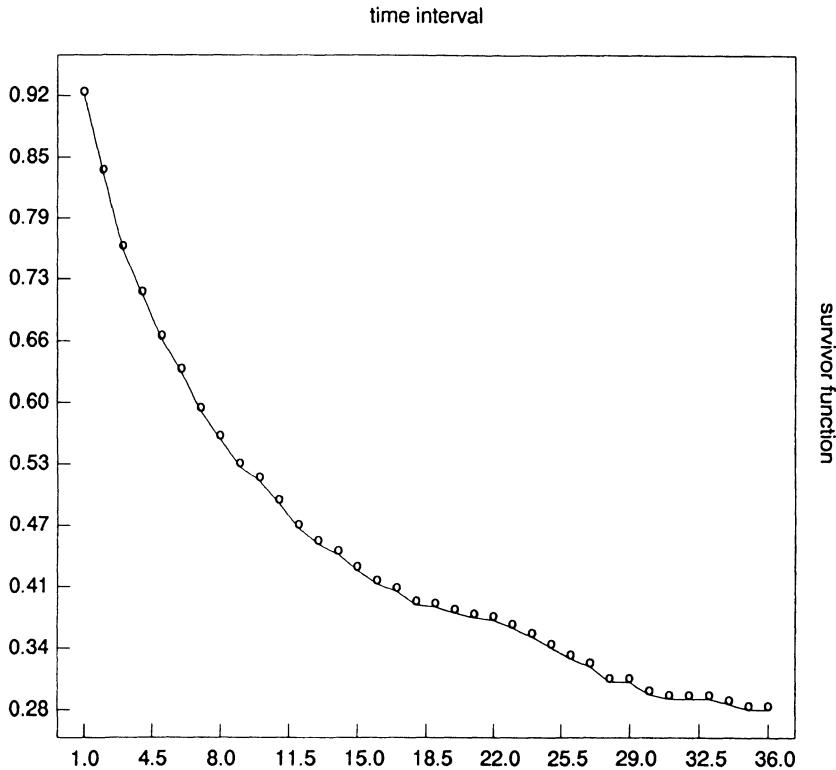


Figure 9.2. Estimated survival function for unemployment data.

$$\lambda(t|x) = 1 - \exp(-\exp(\gamma_t + x'\gamma)), \quad (9.2.11)$$

where the parameters

$$\gamma_t = \log(\exp(\theta_t) - \exp(\theta_{t-1})) \quad \text{with} \quad \theta_t = \log \int_0^{a_t} \lambda_0(u) du$$

are derived from the baseline hazard function  $\lambda_0(u)$  (see Kalbfleisch & Prentice, 1973, 1980). It should be noted that the parameter vector  $\gamma$  is unchanged by the transition to the discrete version. This means, as far as the influence of covariates  $x$  is concerned, the discrete model allows the same analysis as the proportional hazards model. Alternative formulations of (9.2.11) are given by

$$\log(-\log(1 - \lambda(t|x))) = \gamma_t + x'\gamma \quad (9.2.12)$$

and

$$\log(-\log(P(T > t|x))) = \theta_t + x'\gamma. \quad (9.2.13)$$

Since the hazard rate is given by  $\lambda(t|x) = P(T = t|T \geq t, x)$ , it is seen that the grouped proportional hazards model (9.2.11) is a sequential model with distribution function  $F(u) = 1 - \exp(-\exp(u))$  (see (3.3.18) in Chapter 3).

In general, grouping implies a loss of information. The extent of the loss of information is considered, e.g., by Gould & Lawless (1988) for a special model.

### A Generalized Version: The Model of Aranda-Ordaz

Instead of considering the multiplicative model (9.2.10), one may start from an additive form for continuous time by assuming

$$\lambda_c(t|x) = \lambda_0(t) + x'\gamma.$$

Then for the discrete time  $T$  one can derive

$$-\log(1 - \lambda(t|x)) = \rho_t - \rho_{t-1} + (a_t - a_{t-1})x'\gamma,$$

where  $\rho_t = \int_0^{a_t} \lambda_0(u)du$ . If intervals are equidistant, i.e.,  $\Delta = a_t - a_{t-1}$ , the discrete model has the form

$$-\log(1 - \lambda(t|x)) = \delta_t + x'\gamma, \quad (9.2.14)$$

where  $\Delta$  is absorbed into the parameter vector  $\gamma$  and  $\delta_t = \rho_t - \rho_{t-1}$ . Aranda-Ordaz (1983) proposed a general model family that includes the grouped Cox model (9.2.11) as well as the grouped version of the additive model (9.2.14). The model family is given by

$$\begin{aligned} \log(-\log(1 - \lambda(t|x))) &= \gamma_t + x'\gamma \quad \text{for } \alpha = 0, \\ [\{-\log(1 - \lambda(t|x))\}^\alpha - 1]/\alpha &= \gamma_t + x'\gamma \quad \text{for } \alpha \neq 0. \end{aligned} \quad (9.2.15)$$

For  $\alpha = 0$  one gets the grouped Cox model; for  $\alpha = 1$  one gets model (9.2.14) with  $\delta_t = 1 + \gamma_t$ . Thus, (9.2.15) includes both cases. The model may also be written in the form

$$\lambda(t|x) = F_\alpha(\gamma_t + x'\gamma),$$

where  $F_\alpha$  is the distribution function

$$F_\alpha(u) = \begin{cases} 1 - \exp(-(1 + \alpha u)^{1/\alpha}) & \text{for } u \in [-1/\alpha, \infty), \\ 0 & \text{otherwise,} \end{cases}$$

that depends on the additional parameter  $\alpha$ . In this form it becomes clear that the grouped proportional hazards model is the limiting case  $\alpha \rightarrow 0$ . However, it also becomes clear that the range of parameters  $\gamma$  is restricted here. In order to avoid problems, the linear predictor should fulfill  $\gamma_t + x'\gamma > -1/\alpha$ .

### The Logistic Model

An alternative model, which Thompson (1977) has considered, is the logistic model for the discrete hazard

$$\lambda(t|x) = \frac{\exp(\gamma_t + x'\gamma)}{1 + \exp(\gamma_t + x'\gamma)}.$$

The model differs only slightly from the discrete logistic model given by Cox (1972). Thompson (1977) also shows that the model is very similar to the proportional hazards model if the grouping intervals become short.

### Sequential Model and Parameterization of the Baseline Hazard

The common structure of the discrete survival models of the previous sections is that of the sequential model in Section 3.3.4. The discrete hazard has the form

$$\lambda(t|x) = F(\gamma_{0t} + x'\gamma),$$

where  $F$  is a distribution function that for the model of Aranda-Ordaz depends on an additional parameter. In ordinal regression models the number of response categories is usually very small. However, for survival models the number of time intervals may be very large, e.g., if the data are given in months. Thus, the number of parameters  $\gamma_{01}, \dots, \gamma_{0q}$  that represent the baseline hazard may be dangerously high. An approximation of the baseline hazard function used, e.g., by Mantel & Hankey (1978), is given by a polynomial of degree  $s$ :

$$\gamma_{0t} = \sum_{i=0}^s \alpha_i t^i. \quad (9.2.16)$$

An alternative approach used by Efron (1988) for the case without covariates is based on simple regression splines. He considered a cubic-linear spline of the form

$$\gamma_{0t} = \alpha_0 + \alpha_1 t + \alpha_2(t - t_c)_-^2 + \alpha_3(t - t_c)_-^3,$$

where  $(t - t_c)_- = \min\{(t - t_c), 0\}$ . Here  $t_c$  is a cut-off point chosen from the data; the baseline hazard function is cubic before  $t_c$  and linear after  $t_c$ . The reason for the approach is simple: For most survival data there are many data available at the beginning and thus a more complicated structure may be fitted; for higher values of  $t$  the data become sparse and a simple (linear) structure is fitted. A more general model allows the parameter vector  $\gamma$  to depend on time. In

$$\lambda(t|x) = F(\gamma_{0t} + x'\gamma_t) \quad (9.2.17)$$

the parameter  $\gamma_t$  varies over time. Of course, by considering  $x' = (x'_1, x'_2)$  the second term  $x'\gamma_t$  may be split up in  $x'_1\tilde{\gamma} + x'_2\tilde{\gamma}_t$ , where the weight on the first subvector  $x'_1$  is time-independent (global) and the weight on subvector  $x'_2$  is time-dependent. In the context of sequential models (see Section 3.3.4) time-dependent weighting is called category-specific, where categories now refer to discrete time. For ML estimation the number of time intervals and the number of covariates with time-varying weights must be small compared to the number of observations. Alternative estimation procedures that handle the case of time-varying parameters more parsimoniously are considered in Section 9.4.

### 9.2.3 Maximum Likelihood Estimation

In the following the model is considered in the simple form

$$\lambda(t|x_i) = F(z'_{it}\beta), \quad (9.2.18)$$

where, e.g., model (9.2.17) is given by  $z_{it} = (0, \dots, 1, x_i, \dots, 0)$  and  $\beta = (\gamma_{01}, \gamma_1, \dots, \gamma_{0q}, \gamma_q)$ . The data are given by  $(t_i, \delta_i, x_i)$ ,  $i = 1, \dots, n$ , where discrete time  $t_i = \min\{T_i, C_i\}$  is the minimum of survival time  $T_i$  and censoring time  $C_i$ . The indicator variable  $\delta_i$  is determined by

$$\delta_i = \begin{cases} 1 & \text{if } T_i < C_i, \\ 0 & \text{if } T_i \geq C_i. \end{cases}$$

Assuming independence of  $C_i$  and  $T_i$  (random censoring) the probability of observing  $(t_i, \delta_i = 1)$  is given by

$$P(T_i = t_i, \delta_i = 1) = P(T_i = t_i)P(C_i > t_i). \quad (9.2.19)$$

The probability of censoring at time  $t_i$  is given by

$$P(C_i = t_i, \delta_i = 0) = P(T_i \geq t_i)P(C_i = t_i). \quad (9.2.20)$$

In (9.2.19) and (9.2.20) it is assumed that a failure in interval  $[a_{t_{i-1}}, a_{t_i})$  implies a censoring time beyond  $a_{t_{i-1}}$  and censoring in interval  $[a_{t_{i-1}}, a_{t_i})$  implies survival beyond  $a_{t_{i-1}}$ . Thus, implicit censoring is assumed to occur at the beginning of the interval. If censoring is assumed to occur at the end of the interval, (9.2.19) and (9.2.20) have to be substituted by

$$\begin{aligned} P(T_i = t_i, \delta_i = 1) &= P(T_i = t_i)P(C_i \geq t_i), \\ P(C_i = t_i, \delta_i = 0) &= P(C_i = t_i)P(T_i \geq t_i). \end{aligned} \quad (9.2.21)$$

Combining (9.2.19) and (9.2.20) yields the likelihood contribution of observation  $(t_i, \delta_i)$ :

$$L_i = P(T_i = t_i)^{\delta_i} P(T_i \geq t_i)^{1-\delta_i} P(C_i > t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}.$$

If the factor  $c_i = P(C_i > t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}$  that represents the censoring contributions does not depend on the parameters determining the survival time (noninformative in the sense of Kalbfleisch & Prentice, 1980), the likelihood contribution reduces to

$$L_i = c_i P(T_i = t_i)^{\delta_i} P(T_i \geq t_i)^{1-\delta_i}.$$

Including covariates and using the definition of the discrete hazard function, one gets

$$L_i = c_i \lambda(t_i|x_i)^{\delta_i} \prod_{j=1}^{t_i-1} (1 - \lambda(j|x_i)). \quad (9.2.22)$$

It is useful to consider (9.2.22) in a different form. For  $\delta_i = 0$  (9.2.22) may be written by

$$L_i \propto \prod_{j=1}^{t_i-1} \lambda(j|x_i)^{y_{ij}} (1 - \lambda(j|x_i))^{1-y_{ij}},$$

where  $y_i = (y_{i1}, \dots, y_{i,t_i-1}) = (0, \dots, 0)$ . For  $\delta_i = 1$  (9.2.22) may be written by

$$L_i \propto \prod_{j=1}^{t_i} \lambda(j|x_i)^{y_{ij}} (1 - \lambda(j|x_i))^{1-y_{ij}},$$

where  $y_i = (y_{i1}, \dots, y_{it_i}) = (0, \dots, 0, 1)$ . Here  $y_{ij}$  stands for the transition from interval  $[a_{j-1}, a_j]$  to  $[a_j, a_{j+1}]$  given by

$$y_{ij} = \begin{cases} 1, & \text{individual fails in } [a_{j-1}, a_j), \\ 0, & \text{individual survives in } [a_{j-1}, a_j), \end{cases}$$

$j = 1, \dots, t_i$ . That means the total log-likelihood for model  $\lambda(t|x_i) = F(z'_{it}\beta)$  given by

$$l \propto \sum_{i=1}^n \sum_{j=1}^{t_i-(1-\delta_i)} (y_{ij} \log \lambda(j|x_i) + (1 - y_{ij}) \log(1 - \lambda(j|x_i)))$$

is identical to the log-likelihood of the  $\sum_i (t_i - 1 + \delta_i)$  observations  $y_{11}, \dots, y_{1,t_1-(1-\delta_1)}, y_{21}, \dots, y_{n,t_n-(1-\delta_n)}$  from the binary response model  $P(y_{ij} = 1|x_i) = F(z'_{ij}\beta)$ . Thus, ML estimates may be calculated in the same way as for generalized linear models. The vector of binary responses and the design matrix are given by

$$y = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1,t_1-(1-\delta_1)} \\ y_{21} \\ \vdots \\ y_{n,t_n-(1-\delta_n)} \end{bmatrix}, \quad Z = \begin{bmatrix} z'_{11} \\ \vdots \\ z'_{1,t_1-(1-\delta_1)} \\ z'_{21} \\ \vdots \\ z'_{n,t_n-(1-\delta_n)} \end{bmatrix}.$$

It should be noted that  $z_{it}$  in matrix  $Z$  for the simplest case does not depend on  $t$ . Alternatively, the observations may be reordered, yielding the likelihood function

$$l = \sum_{t=1}^q \sum_{i \in R_t} (y_{it} \log \lambda(t|x_i) + (1 - y_{it}) \log(1 - \lambda(t|x_i))), \quad (9.2.23)$$

where  $R_t = \{i : t \leq t_i - (1 - \delta_i)\}$  is the risk set, i.e., the set of individuals who are at risk in interval  $[a_{t-1}, a_t]$ .

If censoring is assumed to occur at the end of the interval, (9.2.20) has to be substituted by (9.2.21). Then the likelihood has the form

$$\begin{aligned}
L_i &= c_i P(T_i = t_i)^{\delta_i} P(T_i > t_i)^{1-\delta_i} \\
&= c_i \lambda(t_i|x_i)^{\delta_i} \prod_{j=1}^{t_i-1} (1 - \lambda(j|x_i))^{\delta_i} \prod_{j=1}^{t_i} (1 - \lambda(j|x_i))^{1-\delta_i} \\
&= c_i \lambda(t_i|x_i)^{\delta_i} (1 - \lambda(t_i|x_i))^{1-\delta_i} \prod_{j=1}^{t_i-1} (1 - \lambda(j|x_i)) \\
&= c_i \prod_{j=1}^{t_i} \lambda(j|x_i)^{y_{ij}} (1 - \lambda(j|x_i))^{1-y_{ij}},
\end{aligned}$$

where

$$y'_i = (y_{i1}, \dots, y_{it_i}) = \begin{cases} (0, \dots, 0) & \text{if } \delta_i = 0, \\ (0, \dots, 1) & \text{if } \delta_i = 1. \end{cases}$$

The log-likelihood is now given by

$$\begin{aligned}
l &= \sum_{i=1}^n \sum_{j=1}^{t_i} (y_{ij} \log(\lambda(j|x_i)) + (1 - y_{ij}) \log(1 - \lambda(j|x_i))) \\
&= \sum_{t=1}^q \sum_{i \in R'_t} (y_{it} \log(\lambda(t|x_i)) + (1 - y_{it}) \log(1 - \lambda(t|x_i))),
\end{aligned}$$

where the risk set  $R'_t = \{i | t \leq t_i\}$  in interval  $[a_{t-1}, a_t]$  now includes the individuals who are censored in this interval. In fact, the individuals are under risk in  $[a_{t-1}, a_t]$  since censoring is considered to happen at the end of the interval. Under this assumption the response vector and design matrix have to be modified accordingly, now yielding  $\sum_i t_i$  observations for the pseudo-model  $P(y_{ij} = 1|x_i) = F(z'_{ij}\beta)$ .

**Example 9.2: Duration of unemployment** (Example 9.1, continued)  
For the sample of unemployed persons considered in Example 9.1 several covariates are observed. Interesting covariates considered in the following are

Gender (dichotomous, 1: male, 2: female)  
 Education (four categories, 1: low, 2: medium, 3: high, 4: university)  
 Nationality (dichotomous, 1: German, 2: other)  
 Illness (dichotomous, 1: yes, 2: no)  
 Previous employment level (five categories, 1: full time, 2: part time, 3: job training, 4: school/university, 5: other)  
 Country (ten categories, 1: Schleswig-Holstein, 2: Hamburg, 3: Niedersachsen, 4: Bremen, 5: Nordrhein-Westfalen, 6: Hessen, 7: Rheinland-Pfalz/Saarland, 8: Baden-Württemberg, 9: Bayern, 10: Berlin)  
 Age (four categories, 1: below 30, 2: 31 to 40, 3: 41 to 50, 4: above 50)  
 Beginning of unemployment (four categories, 1: January to March, 2: April to June, 3: July to September, 4: October to December)

Table 9.1 shows the estimated values for the logistic model and the grouped proportional hazards model. The covariates are given in effect coding. All of the variables except educational level have very small  $p$ -values in at least some categories. For example, within the country effect most of the countries show no deviation from the baseline level of zero effect; only the countries 1, 8, 9 show a strong deviation. In particular, unemployed persons from country 9 have improved chances of ending their unemployment early. Both models, the logistic model and the grouped proportional hazards model, yield almost the same parameter estimates. Also, the fit of the model is almost identical.  $\square$

### 9.2.4 Time-varying Covariates

Maximum likelihood estimation has been considered for the model  $\lambda(t|x_i) = F(z'_{it}\beta)$ , where  $z_{it}$  is built from the covariates  $x_i$ . For the models considered in previous sections it is implicitly assumed that  $z_{it}$  depends on time  $t$  merely by incorporating the baseline hazard. However, more generally  $z_{it}$  may be a vector that incorporates covariates varying over time.

Let  $x_{i1}, \dots, x_{it}$  denote the sequence of observations of covariates for the  $i$ th unit until time  $t$ , where  $x_{it}$  is a vector observed at the beginning of interval  $[a_{t-1}, a_t)$  or is fixed at discrete time  $t$ . In interval  $[a_{t-1}, a_t)$  the “history” of covariates

$$x'_i(t) = (x'_{i1}, \dots, x'_{it})$$

may influence the hazard rate in the model

$$\lambda(t|x_i(t)) = P(T = t|T \geq t, x_i(t)) = F(z'_{it}\beta), \quad (9.2.24)$$

**Table 9.1.** Duration of unemployment

	Grouped proportional hazards model	<i>p</i> -values	Logistic model	<i>p</i> -values
1	−3.391	0.000	−3.358	0.000
POLY(1)	0.008	0.821	0.002	0.952
POLY(2)	−0.005	0.123	−0.004	0.164
POLY(3)	0.000	0.135	0.000	0.169
GENDER	0.382	0.000	0.403	0.000
EDUCATION(1)	−0.139	0.364	−0.149	0.357
EDUCATION(2)	−0.072	0.317	−0.075	0.324
EDUCATION(3)	0.178	0.025	0.190	0.024
ILLNESS	−0.105	0.030	−0.111	0.029
NATIONALITY	0.233	0.000	0.244	0.000
LEVEL(1)	0.397	0.000	0.408	0.000
LEVEL(2)	−0.373	0.026	−0.380	0.027
LEVEL(3)	0.461	0.000	0.484	0.000
LEVEL(4)	0.032	0.801	0.027	0.841
COUNTRY(1)	−0.238	0.235	−0.257	0.220
COUNTRY(2)	−0.556	0.055	−0.580	0.052
COUNTRY(3)	−0.039	0.717	−0.043	0.702
COUNTRY(4)	−0.320	0.211	−0.333	0.211
COUNTRY(5)	−0.018	0.829	−0.024	0.791
COUNTRY(6)	0.059	0.629	0.063	0.622
COUNTRY(7)	0.135	0.304	0.143	0.308
COUNTRY(8)	0.309	0.001	0.327	0.001
COUNTRY(9)	0.406	0.000	0.425	0.000
AGE(1)	0.626	0.000	0.652	0.000
AGE(2)	0.352	0.000	0.361	0.000
AGE(3)	0.227	0.011	0.226	0.014
MONTH(1)	0.004	0.936	0.002	0.961
MONTH(2)	−0.144	0.045	−0.155	0.041
MONTH(3)	−0.060	0.352	−0.061	0.372

where  $z_{it}$  is composed from  $x_i(t)$ . There are many possibilities of specifying the vector  $z_{it}$ . A simple way where only the vector observed at time  $t$  is of influence is given by

$$z'_{it}\beta = \gamma_{0t} + x'_{it}\gamma,$$

where  $z'_{it} = (0, \dots, 1, \dots, 0, x'_{it}), \beta' = (\gamma_{01}, \dots, \gamma_{0q}, \gamma')$ . If the parameter varies over time, one may specify

$$z'_{it}\beta = \gamma_{0t} + x'_{it}\gamma_t,$$

where  $z'_{it} = (0, \dots, 1, \dots, 0, 0, \dots, x'_{it}, \dots, 0), \beta' = (\gamma_{01}, \dots, \gamma_{0q}, \gamma'_1, \dots, \gamma'_q)$ . Of course, time lags may be included by

$$z'_{it}\beta = \gamma_{0t} + x'_{it}\gamma_0 + \dots + x'_{i,t-r}\gamma_{-r},$$

where

$$\begin{aligned} z'_{it} &= (0, \dots, 1, \dots, 0, x'_{it}, x'_{i,t-1}, \dots, x'_{i,t-r}), \\ \beta' &= (\gamma_{01}, \dots, \gamma_{0q}, \gamma'_0, \gamma'_{-1}, \dots, \gamma'_{-r}). \end{aligned}$$

In  $z_{it}$ , characteristics of the interval (e.g., the length of the interval) may be included by

$$z'_{it}\beta = \gamma_{0t} + (a_t - a_{t-1})\gamma, \quad (9.2.25)$$

where  $z'_{it} = (0, \dots, 1, \dots, 0, a_t - a_{t-1}), \beta' = (\gamma_{01}, \dots, \gamma_{0q}, \gamma)$ .

For time-dependent covariates one may distinguish between two types: *external* and *internal* covariates (Kalbfleisch & Prentice, 1980). External covariates are not directly involved with failure. The components of  $z_{it}$  that refer to the coding of the interval (i.e., the baseline hazard) as well as the difference  $a_t - a_{t-1}$  in (9.2.25) are fixed in advance; they are not determined by the failure mechanism. If  $x_{i1}, \dots, x_{it}$  is the output of a stochastic process, it may be considered external if the condition

$$\begin{aligned} P(x_{i,t+1}, \dots, x_{iq} | x_i(t), y_i(t)) &= P(x_{i,t+1}, \dots, x_{iq} | x_i(t)), \\ t &= 1, \dots, q, \end{aligned} \quad (9.2.26)$$

holds where  $y_{ij}$  denotes failure in interval  $[a_{j-1}, a_j]$  and  $y_i(t) = (y_{i1}, \dots, y_{it})$ . Equation (9.2.26) means that the path of the covariate process is not influenced by failure. A consequence of (9.2.26) is that

$$P(y_{it}|x_i(t), y_i(t-1)) = P(y_{it}|x_i(q), y_i(t-1)),$$

and therefore conditioning may be done on the whole path  $x_i(q)$  by

$$\begin{aligned}\lambda(t|x_i(t)) &= P(y_{it} = 1|x_i(t), y_{i1} = 0, \dots, y_{i,t-1} = 0) \\ &= P(y_{it} = 1|x_i(q), y_{i1} = 0, \dots, y_{i,t-1} = 0) \\ &= \lambda(t|x_i(q)).\end{aligned}$$

For external covariates under mild restrictions, the likelihood (9.2.23) may still be used by substituting time-dependent covariates  $x_i(t)$  for the time-independent covariate  $x_i$ . For a derivation, see the upcoming subsection “Maximum Likelihood Estimation\*.”

### Internal Covariates\*

Internal covariates carry with their observed path information on failure such as characteristics of an individual that can be observed only as long as the individual is in the study and alive. Now the hazard function only incorporates the path until time  $t$ .

$$\lambda(t|x(t)) = P(y_{it} = 1|x(t), y_{i1} = 0, \dots, y_{i,t-1} = 0).$$

Condition (9.2.26) cannot be assumed to hold since covariates  $x_{i,t+1}, \dots, x_{iq}$  may no longer have any meaning if a patient dies in interval  $[a_{t-1}, a_t]$ . In particular, for the latter type of covariates the simple connection between survival function and hazard function given at the beginning of Section 9.2 no longer holds. One may consider a type of “survival function” by defining

$$S(t|x(t)) = P(T > t|x(t)),$$

where  $T > t$  denotes the sequence  $y_{i1} = 0, \dots, y_{it} = 0$ . Alternatively, one may write

$$S(t|x(t)) = \prod_{s=1}^t P(T > s|T \geq s, x(t)).$$

However, since the factor  $P(T > s|T \geq s, x(t))$  includes the history  $x(t)$ , it is not equal to  $1 - \lambda(s|x(s))$ , which includes only the history until time  $s$ .

Consider the general formula

$$P(A_s|B_s \cap C_s) = P(C_s|A_s \cap B_s)P(A_s|B_s)/P(C_s|B_s)$$

with  $A_s = \{T > s\}, B_s = \{T \geq s, x(s)\}, C_s = \{x_{s+1}, \dots, x_t\}$ . Then one gets

$$P(T > s | T \geq s, x(t)) = P(T > s | T \geq s, x(s))q_s = (1 - \lambda(s|x(s)))q_s,$$

where  $q_s = P(x_{s+1}, \dots, x_t | T > s, x(s)) / P(x_{s+1}, \dots, x_t | T \geq s, x(s))$ . For external covariates  $q_s = 1$  holds and the survival function is again given by

$$S(t|x(t)) = S(t|x(q)) = \prod_{s=1}^t (1 - \lambda(s|x(s))).$$

### Maximum Likelihood Estimation\*

The data are given by  $(t_i, \delta_i, x_i(t_i)), i = 1, \dots, n$ , where  $t_i$  is again the minimum  $t_i = \min\{T_i, C_i\}$ . The probability for the event  $\{t_i, \delta_i = 1, x_i(t_i)\}$  is given by

$$\begin{aligned} P(t_i, \delta_i = 1, x_i(t_i)) &= P(T_i = t_i, C_i > t_i, x_i(t_i)) \\ &= P(T_i = t_i, C_i > t_i, x_i(t_i) | H_{i,t_i-1}) \prod_{s=1}^{t_i-1} P(T_i > s, C_i > s, x_i(s) | H_{i,s-1}), \end{aligned}$$

where  $H_{i,s} = \{T_i > s, C_i > s, x_i(s)\}, s = 1, \dots, t_i - 1, H_{i0} = \{T_i > 0, C_i > 0\}$  is a sequence of “histories” fulfilling  $H_{i,s+1} \subset H_{i,s}$ .

By using the simple formula  $P(A \cap B | C) = P(A | B \cap C)P(B | C)$ , the above may be rewritten as

$$\begin{aligned} P(t_i, \delta_i = 1, x_i(t_i)) &= P(T_i = t_i | C_i > t_i, x_i(t_i), H_{i,t_i-1})P(x_i(t_i), C_i > t_i | H_{i,t_i-1}) \\ &\quad \cdot \prod_{s=1}^{t_i-1} P(T_i > s | C_i > s, x_i(s), H_{i,s-1})P(x_i(s), C_i > s | H_{i,s-1}). \end{aligned}$$

Assuming

$$P(T_i = s | T_i > s - 1, C_i > s, x_i(s)) = P(T_i = s | T_i > s - 1, x_i(s)), \quad (9.2.27)$$

which holds for  $T_i, C_i$  independent, one gets

$$\begin{aligned} P(T_i = s | C_i > s, x_i(s), H_{i,s-1}) &= P(T_i = s | T_i > s - 1, C_i > s, x_i(s)) \\ &= P(T_i = s | T_i > s - 1, x_i(s)) \\ &= \lambda(s | x_i(s)), \end{aligned}$$

and therefore

$$\begin{aligned} P(t_i, \delta_i = 1, x_i(t_i)) &= \lambda(t_i | x_i(t_i)) \prod_{s=1}^{t_i-1} [1 - \lambda(s | x_i(s))] \\ &\quad \cdot \prod_{s=1}^{t_i} P(x_i(s), C_i > s | H_{i,s-1}). \end{aligned} \quad (9.2.28)$$

If censoring takes place at the beginning of the interval, the interesting probability is that for the event  $\{t_i, \delta_i = 0, x_i(t_i-1)\}$ . The same probability has to be computed if data are discrete time points, because then  $t_i, \delta_i = 0$  implies that the past covariates can be observed only until time  $t_i - 1$ . Assuming (9.2.26) holds, it is given by

$$\begin{aligned} P(t_i, \delta_i = 0, x_i(t_i-1)) &= P(T_i \geq t_i, C_i = t_i, x_i(t_i-1)) \\ &= P(C_i = t_i | H_{i,t_i-1}, x_i(t_i-1)) P(x_i(t_i-1) | H_{i,t_i-1}) \\ &\quad \cdot \prod_{s=1}^{t_i-1} P(T_i > s | C_i > s, x_i(s), H_{i,s-1}) P(C_i > s, x_i(s) | H_{i,s-1}) \\ &= P(C_i = t_i | H_{i,t_i-1}) \\ &\quad \cdot \prod_{s=1}^{t_i-1} (1 - \lambda(s | x_i(s))) P(C_i > s, x_i(s) | H_{i,s-1}). \end{aligned} \quad (9.2.29)$$

Equations (9.2.28) and (9.2.29) yield

$$P(t_i, \delta_i, x_i(t_i)^{\delta_i}, x_i(t_i-1)^{1-\delta_i}) = \lambda(t_i | x_i(t_i))^{\delta_i} \prod_{s=1}^{t_i-1} (1 - \lambda(s | x_i(s))) Q_i,$$

where

$$\begin{aligned} Q_i &= P(C_i = t_i | H_{i,t_i-1})^{1-\delta_i} P(C_i > t_i, x_i(t_i) | H_{i,t_i-1})^{\delta_i} \\ &\quad \cdot \prod_{s=1}^{t_i-1} P(C_i > s, x_i(s) | H_{i,s-1}). \end{aligned} \quad (9.2.30)$$

If the factor  $Q_i$  is noninformative, i.e., it does not depend on the parameters determining survival time, then the likelihood is the same as (9.2.23), which was derived for time-independent covariates.

For  $n$  observations  $(t_i, x_i(t_i), \delta_i)$  one gets for the log-likelihood

$$l \propto \sum_{i=1}^n \sum_{r=1}^{t_i-(1-\delta_i)} \left[ y_{ir} \log \lambda(r | x_i(r)) + (1 - y_{ir}) \log(1 - \lambda(r | x_i(r))) \right]$$

or, equivalently,

$$l \propto \sum_{t=1}^q \sum_{i \in R_t} \left[ y_{it} \log \lambda(t | x_i(t)) + (1 - y_{it}) \log(1 - \lambda(t | x_i(t))) \right], \quad (9.2.31)$$

where  $R_t = \{i : t < t_i - (1 - \delta_i)\}$  is the risk set.

ML estimates may be estimated within the framework of generalized linear models with the response and design matrix as given in Section 9.2.3. The only difference is that  $z_{it}$  now includes time-varying components. For the consideration of time-dependent covariates, see also Hamerle & Tutz (1989).

## 9.3 Discrete Models for Multiple Modes of Failure

In the preceding sections methods have been considered for observations of failure or censoring where it is assumed that there is only one type of failure event. Often one may distinguish between several distinct types of terminating events. For example, in a medical treatment study the events may stand for several causes of death. In studies on unemployment duration one may distinguish between full-time and part-time jobs that end the unemployment duration. In survival analysis, models for this type of data are often referred to as competing risks models. We will use this name although in the case of unemployment data competing chances would be more appropriate. Most of the literature for competing risks considers the case of continuous time (e.g., Kalbfleisch & Prentice, 1980).

### 9.3.1 Basic Models

Let  $R \in \{1, \dots, m\}$  denote the distinct events of failure or causes. Considering discrete time  $T \in \{1, \dots, q + 1\}$ , the *cause-specific hazard function* resulting from cause or risk  $r$  is given by

$$\lambda_r(t|x) = P(T = t, R = r | T \geq t, x),$$

where  $x$  is a vector of time-independent covariates. The *overall hazard function* for failure regardless of cause is given by

$$\lambda(t|x) = \sum_{r=1}^m \lambda_r(t|x) = P(T = t | T \geq t, x).$$

The survival function and unconditional probability of a terminating event are given as in the simple case of one terminating event by

$$S(t|x) = P(T > t|x) = \prod_{i=1}^t (1 - \lambda(i|x))$$

and

$$P(T = t|x) = \lambda(t|x)S(t-1|x).$$

For an individual reaching interval  $[a_{t-1}, a_t]$ , the conditional response probabilities are given by

$$\lambda_1(t|x), \dots, \lambda_m(t|x), 1 - \lambda(t|x),$$

where  $1 - \lambda(t|x)$  is the probability for survival. Modelling these events may be based on the approach for mult categorial responses outlined in Chapter 3. A candidate for unordered events is the multinomial logit model

$$\lambda_r(t|x) = \frac{\exp(\gamma_{0tr} + x'\gamma_r)}{1 + \sum_{i=1}^m \exp(\gamma_{0ti} + x'\gamma_i)}, \quad (9.3.1)$$

for  $r = 1, \dots, m, t = 1, \dots, q$ . In model (9.3.1) the parameters  $\gamma_{01j}, \dots, \gamma_{0qj}$  represent the cause-specific baseline hazard function and  $\gamma_r$  is the cause-specific weight. Like the single-event case considered in Section 9.2, the baseline hazard function may be simplified by using a polynomial approximation and the weight  $\gamma_r$  may depend on time. Then the influence term  $\gamma_{0tr} + x'\gamma_r$  is substituted by the more general term  $\eta_r = z_{tr}'\gamma$ . The general form of (9.3.1) is given by

$$\lambda_r(t|x) = h_r(Z_t\beta), \quad (9.3.2)$$

where  $h_r$  is the local link function for responses in interval  $[a_{t-1}, a_t]$  and  $Z_t$  is a design matrix composed of  $x$  and depending on time  $t$ . The simple model (9.3.1) has the logit response function

$$h_r(\eta_1, \dots, \eta_m) = \frac{\exp(\eta_r)}{1 + \sum_{i=1}^m \exp(\eta_i)}$$

and design matrix

$$Z_t = \begin{bmatrix} 0 & 1 & & & 0 & x' \\ & & 1 & & & x' \\ & & & \ddots & & \ddots \\ 0 & & & & 1 & 0 & & x' \end{bmatrix}, \quad (9.3.3)$$

where the parameter vector is given by

$$\beta' = (\gamma_{011}, \dots, \gamma_{01m}, \gamma_{021}, \dots, \gamma_{0qm}, \gamma'_1, \dots, \gamma'_m).$$

For modelling the conditional response in the interval  $[a_{t-1}, a_t)$ , alternative models from Chapter 3 may be more appropriate. If the events are ordered, ordinal models like sequential or cumulative models may yield a simpler structure with fewer parameters; see Tutz (1995a).

A quite different approach to parametric models is based on models derived for the continuous case. A proportional hazards model in which the cause-specific hazard function at continuous time  $t$  depends on  $x$  is given by

$$\lambda_r(t) = \lambda_{0r}(t) \exp(x' \gamma_r), \quad (9.3.4)$$

for  $r = 1, \dots, m$  (e.g., Kalbfleisch & Prentice, 1980). Derivation of the discrete model for observations in intervals yields parametric models at least in special cases. If the baseline hazard function does not depend on cause  $r$ , the model has the simpler form

$$\lambda_r(t) = \lambda_0(t) \exp(x' \gamma_r).$$

For this model the discrete hazard function where  $T = t$  denotes failure in the interval  $[a_{t-1}, a_t)$  may be derived by

$$\lambda_r(t|x) = \frac{\exp(\gamma_{0t} + x' \gamma_r)}{\sum_{j=1}^m \exp(\gamma_{0t} + x' \gamma_j)} \left\{ 1 - \exp \left( - \sum_{j=1}^m \exp(\gamma_{0t} + x' \gamma_j) \right) \right\}, \quad (9.3.5)$$

where the baseline hazard function is absorbed into the parameters  $\gamma_{0t} = \log(\int_{a_{t-1}}^{a_t} \lambda_0(t) dt)$ . For the derivation, see Hamerle & Tutz (1989). Model (9.3.5) has the general form (9.3.2), where the response function is given by

$$h_r(\eta_1, \dots, \eta_m) = \frac{\exp(\eta_r)}{\sum_{j=1}^m \exp(\eta_j)} \left\{ 1 - \exp \left( - \sum_{j=1}^m \exp(\eta_j) \right) \right\}$$

and the design matrix has the form (9.3.3). If the covariates are stochastic processes  $x_{i1}, \dots, x_{it}$ , cause-specific and global hazard functions have the forms

$$\begin{aligned} \lambda_r(t|x_i(t)) &= P(T_i = t, R_i = r | T_i \geq t, x_i(t)), \\ \lambda(t|x_i(t)) &= \sum_{r=1}^m \lambda_r(t|x_i(t)), \end{aligned}$$

where  $x_i(t) = (x_{i1}, \dots, x_{it})$  is the sequence of observations until time  $t$ . The model for the hazard function has the form (9.3.2), where the design matrix  $Z_t$  is a function of  $t$  and  $x(t)$ .

### 9.3.2 Maximum Likelihood Estimation

The data are given by  $(t_i, r_i, \delta_i, x_i)$ , where  $r_i \in \{1, \dots, m\}$  indicates the terminating event. We consider the case of random censoring with  $t_i = \min\{T_i, C_i\}$ , and censoring at the beginning of the interval. First the simpler case of time-independent covariates  $x$  is treated. The likelihood contribution of the  $i$ th observation for model (9.3.2) is given by

$$L_i = P(T_i = t_i, R_i = r_i)^{\delta_i} P(T_i \geq t_i)^{1-\delta_i} P(C_i > t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}.$$

For noninformative censoring it may be reduced to

$$\begin{aligned} L_i &= P(T_i = t_i, R_i = r_i|x_i)^{\delta_i} P(T_i \geq t_i|x_i)^{1-\delta_i} \\ &= \lambda_{r_i}(t_i|x_i)^{\delta_i} P(T_i \geq t_i|x_i) \\ &= \lambda_{r_i}(t_i|x_i)^{\delta_i} \prod_{t=1}^{t_i-1} \left(1 - \sum_{r=1}^m \lambda_r(t|x_i)\right). \end{aligned} \quad (9.3.6)$$

An alternative form of the likelihood is based on dummy variables given by

$$y_{itr} = \begin{cases} 1, & \text{failure of type } r \text{ in interval } [a_{t-1}, a_t), \\ 0, & \text{no failure in interval } [a_{t-1}, a_t), \end{cases}$$

for  $r = 1, \dots, m$ . Given that an individual reaches interval  $[a_{t-1}, a_t)$ , the response is multinomial with  $y_{it} = (y_{it1}, \dots, y_{itm}) \sim M(1, \lambda_1(t|x_i), \dots, \lambda_m(t|x_i))$ . Therefore, the dummy variable  $y_{it,m+1} = 1 - y_{it1} - \dots - y_{itm}$  has value 1 if individual  $i$  does not fail in interval  $[a_{t-1}, a_t)$  and  $y_{it,m+1} = 0$  if individual  $i$  fails in  $[a_{t-1}, a_t)$ . The likelihood (9.3.6) for the  $i$ th observation has the form

$$\begin{aligned} L_i &= \prod_{r=1}^m \lambda_r(t_i|x_i)^{\delta_i y_{itr}} \left(1 - \sum_{r=1}^m \lambda_r(t_i|x_i)\right)^{\delta_i y_{it,m+1}} \\ &\quad \cdot \prod_{t=1}^{t_i-1} \left\{ \prod_{r=1}^m \lambda_r(t|x_i)^{y_{itr}} \right\} \left\{1 - \sum_{r=1}^m \lambda_r(t|x_i)\right\}^{y_{it,m+1}} \\ &= \prod_{t=1}^{t_i-1+\delta_i} \left\{ \prod_{r=1}^m \lambda_r(t|x_i)^{y_{itr}} \right\} \left\{1 - \sum_{r=1}^m \lambda_r(t|x_i)\right\}^{y_{it,m+1}}. \end{aligned}$$

This means the likelihood for the  $i$ th observation is the same as that for the  $t_i - 1 + \delta_i$  observations  $y_{i1}, \dots, y_{i,t_i-1+\delta_i}$  of the multicategorical model  $P(Y_{it} = r) = h_r(Z_t \beta)$ , where  $Y_{it} = r$  if  $y_{itr} = 1$ . Thus, as in the single-cause

model, ML estimates may be calculated within the framework of generalized linear models. If  $\delta_i = 0$ , we have the  $t_i - 1$  observation vectors  $y_{i1}, \dots, y_{i,t_i-1}$ , and if  $\delta_i = 1$  we have the  $t_i$  observation vectors  $y_{i1}, \dots, y_{i,t_i}$ , yielding a blown-up design. For the  $i$ th observation, the response and design matrices are given by

$$\begin{bmatrix} y_{i1} \\ \vdots \\ y_{i,t_i-1+\delta_i} \end{bmatrix}, \quad \begin{bmatrix} Z_1 \\ \vdots \\ Z_{t_i-1+\delta_i} \end{bmatrix}.$$

The total log-likelihood is given by

$$\begin{aligned} l &= \sum_{i=1}^n \sum_{t=1}^{t_i-1+\delta_i} \left( \sum_{r=1}^m y_{itr} \log \lambda_r(t|x_i) + y_{it,m+1} \log \left( 1 - \sum_{r=1}^m \lambda_r(t|x_i) \right) \right) \\ &= \sum_{t=1}^q \sum_{i \in R_t} \left( \sum_{r=1}^m y_{itr} \log \lambda_r(t|x_i) + y_{it,m+1} \log \left( 1 - \sum_{r=1}^m \lambda_r(t|x_i) \right) \right), \end{aligned} \quad (9.3.7)$$

where in the latter form  $R_t = \{i | t_i - 1 + \delta_i \geq t\}$  is the number of individuals under risk in the interval  $[a_{t-1}, a_t]$ .

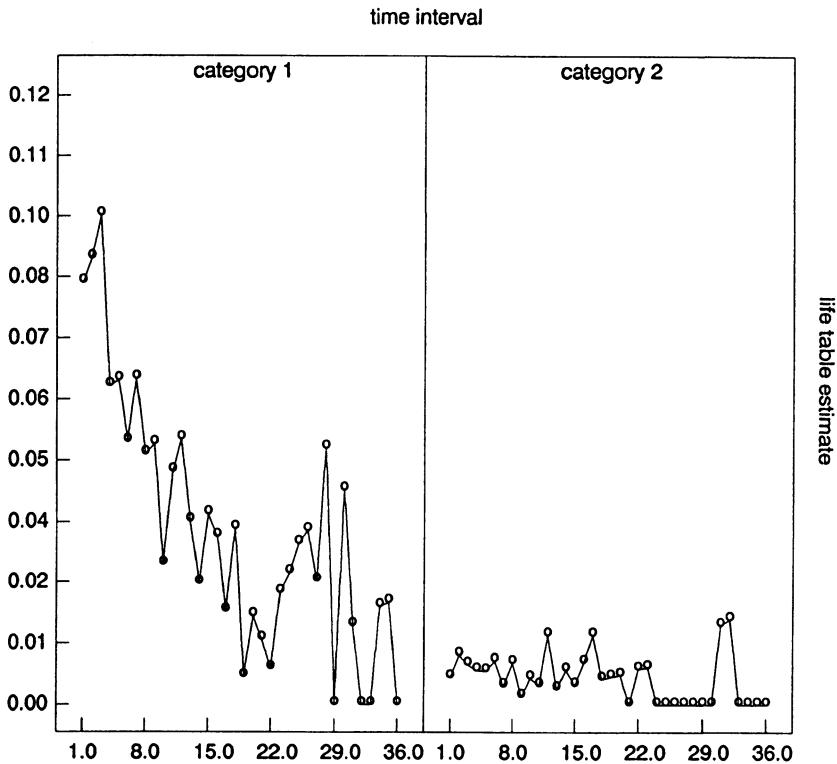
**Example 9.3: Duration of unemployment** (Example 9.2, continued)  
The duration data considered in Example 9.2 are now considered for the case of multiple modes of failure. Instead of the single absorbing event full-time job the causes full-time job or part-time job are considered. Figure 9.3 shows the simple life-time table estimate for these causes. The figure shows that the transition from unemployment to full-time job is dominant; transition to part-time jobs has a rather low value at all times.

Table 9.2 shows the fitted values for the parameters of the logistic model where the polynomial term for the baseline hazard and the explanatory variables have cause-specific weights. Comparison with Table 9.1 shows that the full-time parameters are almost the same as for the duration model considering only the absorbing event full-time job. This is easily explained by looking at the log-likelihood (9.3.7), which for the logistic model may be separated into the sum of log-likelihoods for models specifying each cause separately. However, this holds only for the natural link function underlying the logistic model.

Table 9.2 shows that effects for full-time jobs are often quite different from effects for part-time jobs; sometimes even the sign is different. For example, the effects of gender are 0.403 and  $-0.780$ , showing that men have a higher hazard rate than women with respect to full-time jobs but a lower hazard rate with respect to part-time jobs. That might be due to the strong preference of men for full-time jobs. Similar effects are found for the variable “previous level of employment.”  $\square$

**Table 9.2.** Estimates of cause-specific logistic model for duration of unemployment data.

	Full-time	<i>p</i> -value	Part-time	<i>p</i> -value
1	−3.361	0.000	−5.355	0.000
POLY(1)	0.002	0.944	−0.195	0.168
POLY(2)	−0.004	0.161	0.023	0.112
POLY(3)	0.000	0.167	−0.000	0.101
GENDER	0.403	0.000	−0.780	0.000
EDUCATION(1)	−0.150	0.356	−0.472	0.312
EDUCATION(2)	−0.073	0.338	−0.290	0.200
EDUCATION(3)	0.191	0.023	0.276	0.252
ILLNESS	−0.111	0.029	0.059	0.673
NATIONALITY	0.243	0.000	0.886	0.000
LEVEL(1)	0.408	0.000	−0.690	0.001
LEVEL(2)	−0.381	0.027	0.564	0.015
LEVEL(3)	0.484	0.000	0.203	0.503
LEVEL(4)	0.028	0.834	0.334	0.252
COUNTRY(1)	−0.262	0.211	0.210	0.594
COUNTRY(2)	−0.577	0.054	0.686	0.220
COUNTRY(3)	−0.044	0.693	0.249	0.383
COUNTRY(4)	−0.332	0.213	0.321	0.636
COUNTRY(5)	−0.024	0.790	−0.609	0.030
COUNTRY(6)	0.063	0.626	−0.259	0.523
COUNTRY(7)	0.142	0.309	−1.104	0.095
COUNTRY(8)	0.328	0.001	−0.021	0.945
COUNTRY(9)	0.426	0.000	−0.090	0.770
AGE(1)	0.653	0.000	0.456	0.040
AGE(2)	0.361	0.000	0.149	0.566
AGE(3)	0.226	0.015	0.058	0.828
MONTH(1)	0.003	0.959	−0.068	0.732
MONTH(2)	−0.155	0.040	0.223	0.292
MONTH(3)	−0.061	0.369	0.149	0.452



**Figure 9.3.** Life table estimate for duration of unemployment with causes full-time job (category 1) or part-time job (category 2).

## 9.4 Smoothing in Discrete Survival Analysis

Smoothing of hazard rates may be based on various approaches. In this subsection, we consider simple smoothed life table estimates as well as smoothing based on state space or dynamic modelling techniques, which is closely related to discrete spline smoothing and smoothing by local likelihood methods.

### 9.4.1 Smoothing Life Table Estimates

In the simple case of no covariates, the life table estimate (9.2.8) is of the simple form

$$\hat{\lambda}(t) = \frac{\text{number of failures in } [a_{t-1}, a_t]}{\text{number of individuals under risk in } [a_{t-1}, a_t]}. \quad (9.4.1)$$

As seen in Figure 9.1, this estimate may be quite unsteady; in particular for large  $t$  where the number of individuals under risk is low, there is much noise in the data. The simple relative frequency (9.4.1) is not suited for discovering the underlying form of the hazard function.

A smoothed version of the life table estimate may be based on a polynomial or spline function fit or on smoothing procedures for discrete responses as considered in Section 5.2. One approach outlined in Section 5.2 is localizing, which is considered in the following. Data are given as discrete survival times  $t_i = \min(T_i, C_i)$  and censoring indicators  $\delta_i$ . As is Section 9.2.3, one uses the binary sequence

$$y_{it} = \begin{cases} 1 & \text{if } t = t_i \text{ and } \delta_i = 1, \\ 0 & \text{otherwise,} \end{cases}$$

for  $i = 1, \dots, n$ ,  $t = 1, \dots, t_i - (1 - \delta_i)$ . That means for fixed  $t$  one has observations only from the risk set  $R_t = \{i : t < t_i - (1 - \delta_i)\}$ . The corresponding likelihood as given in (9.2.23) has the form of a binary response model

$$l = \sum_{t=1}^q \sum_{i \in R_t} y_{it} \log \lambda(t) + (1 - y_{it}) \log(1 - \lambda(t)).$$

The local polynomial estimate at time  $t$  results from maximizing

$$l_t = \sum_{s=1}^q \sum_{i \in R_t} \{y_{is} \log \lambda_{ts} + (1 - y_{is}) \log(1 - \lambda_{ts})\} w_\gamma(t, s),$$

where  $\lambda_{ts} = F(\beta_0 + (s - t)\beta_1 + (s - t)^2\beta_2 + \dots)$  is a polynomial model centered at  $t$  and  $w_\gamma(t, s)$  is a weight function depending on the smoothing parameter  $\gamma$ . The weight function  $w_\gamma(t, s)$ , which is usually a function decreasing with  $(t - s)$ , makes  $l_t$  a local likelihood. Maximizing  $l_t$  within the usual framework of GLMs yields the estimate  $\hat{\lambda}_t = F(\hat{\beta}_0)$ . For the simple case of local constant fitting and kernel weights  $w_\gamma(t, s) \propto K((t - s)/\gamma)$ , one obtains the smoothed estimate

$$\hat{\lambda}(t) = \sum_{s=1}^q \sum_{i \in R_s} y_{is} w_\lambda(t, s) = \sum_{s=1}^q \hat{\lambda}_s (|R_s| w_\lambda(t, s)), \quad (9.4.2)$$

where  $\hat{\lambda}_s = \sum_{i \in R_s} y_{is} / |R_s|$  is the simple life table estimate from Section 9.2.1 where the number of failures is divided by the number of individuals at risk. Thus,  $\hat{\lambda}(t)$  is a weighted sum of the simple life table estimate. For  $\gamma \rightarrow 0$  one obtains the life table estimate; for  $\gamma \rightarrow \infty$  one obtains the ultrasmooth estimate  $\hat{\lambda}(1) = \dots = \hat{\lambda}(q)$ . Alternative smoothed life table estimates that incorporate categorical variables have been considered by Copas & Haberman (1983), Tutz & Pritscher (1996), and Congdon (1993).

### 9.4.2 Smoothing with Covariates

Data are given as before by observed discrete survival times  $t_i = \min(T_i, C_i)$ , censoring indicators  $\delta_i$ , and possibly time-varying covariates  $x_{i1}, \dots, x_{it}, \dots, i = 1, \dots, n$ . With the binary indicator sequences  $y_{it}$ ,  $i \in R_t$ ,  $t = 1, \dots, t_i$ , the data are given in the form of longitudinal binary observations

$$(y_{it}, x_{it}), \quad t = 1, \dots, t_i, \quad i = 1, \dots, n,$$

and the hazard function is the probability of failure,

$$\lambda(t|x_i(t)) = P(y_{it} = 1|x_i(t)) = h(\eta_{it}) \quad (9.4.3)$$

with appropriate non- or semiparametric predictor  $\eta_{it}$  and  $h$  a response function for a binary regression model, e.g., a logit model or a grouped Cox model. The basic form of the predictor is

$$\eta_{it} = \gamma_{0t} + x'_{it}\gamma,$$

where the sequence  $\{\gamma_{0t}, t = 1, 2, \dots\}$  represents the baseline effect. Various extensions of this basic semiparametric predictor are conceivable. For example, with

$$\eta_{it} = \gamma_{0t} + z'_{it}\gamma_t + w'_{it}\beta, \quad (9.4.4)$$

we are modelling time-varying effects of covariate  $z_{it}$ , such as the effect of a therapy changing with duration or survival time, whereas  $w_{it}$  are covariates with an effect that remains constant over time.

As mentioned at the end of Section 9.2.2, “static” models for discrete time which treat baseline hazard coefficients and covariate parameters as “fixed effects” are appropriate if the number of intervals is comparably small. In situations with many intervals, but not enough to apply models for continuous time, such unrestricted modelling and fitting of hazard functions often lead to the nonexistence and divergence of ML estimates due to the large number of parameters. This difficulty in real data problems becomes even more apparent if covariate effects are also assumed to be time-varying, as, for example, the effect of a certain therapy in a medical context or the effect of financial support while unemployed on the duration of unemployment.

To avoid such problems, one may try a more parsimonious parameterization by specifying certain functional forms, e.g., (piecewise) polynomials, for time-varying coefficients. However, simply imposing such parametric functions can conceal spikes in the hazard function or other unexpected patterns. Purely nonparametric methods to overcome such problems have been discussed by Huffer & McKeague (1991), based on Aalen’s (1980, 1989) additive risk model for continuous time and Fan & Gijbels (1994), Li & Doss (1995), and Abrahamowicz, MacKenzie & Esaile (1996).

If further time scales, such as calendar time  $d$ , or metrical covariates, such as age, are measured,

$$\eta_{it} = \gamma_0 t + f_1(d_{it}) + f_2(a_i) + z'_{it} \gamma_t + w'_{it} \beta \quad (9.4.5)$$

may be a potential additive predictor. Here, the function  $f_1(d)$  is the calendar time trend,  $d_{it}$  denotes calendar time of individual  $i$  at duration time  $t$ , and  $f_2(a)$  is the effect of age, with  $a_i$  being the age of individual  $i$  at the beginning of duration time.

In unified notation, the general form of the predictor is

$$\eta_{it} = f_0(t) + \sum_{j=1}^p z_{itj} f_j(x_{it}) + w'_{it} \beta, \quad (9.4.6)$$

leading to varying-coefficient models, where the effect modifier is often a time scale and  $z'_{itj} = (z_{it1}, \dots, z_{itp})$  is a vector of covariates with varying coefficients.

In principle, any of the basic concepts for non- and semiparametric function estimation, that is, basis function expansion, localization, penalization, and Bayesian approaches described in Chapter 5 and later, can be applied. Extensions to semiparametric discrete-time models for multiple models of failure are conceptually straightforward: Binary event indicators are replaced by categorical indicators

$$y_{itr} = \begin{cases} 1 & \text{if } t = t_i, \delta_i = 1 \text{ and failure is of type } r, \\ 0 & \text{otherwise,} \end{cases}$$

and mult categorial response models are used instead of binary ones.

### 9.4.3 Dynamic Discrete-Time Survival Models

In their basic form, dynamic discrete-time models are semiparametric Bayesian approaches based on state space models to estimate time-varying baseline and covariate effects together with time-constant effects in a semiparametric predictor of the form (9.4.4). As with longitudinal data in Section 8.4, approximate methods such as posterior mode estimation or fully Bayesian approach are possible.

#### Posterior Mode Smoothing

All parameters are gathered in a state vector  $\alpha_t$ , and we consider the hazard rate model (9.4.3), (9.4.4) together with a linear and Gaussian transition equation

$$\alpha_t = F_t \alpha_{t-1} + \xi_t, \quad t = 1, 2, \dots, \quad (9.4.7)$$

with the usual assumptions, including  $\xi_t \sim N(0, Q)$ ,  $\alpha_0 \sim N(a_0, Q_0)$ .

All assumptions made for ML estimation in Sections 9.2.2 and 9.2.3 are supposed to hold conditionally, given the sequence  $\{\alpha_t\}$  of states. For time-varying covariates we assume that the factors  $Q_i$  in (9.2.30) are noninformative. This condition replaces assumption (A2) of Section 8.2.1. Relying on the state space approach and proceeding as in Section 8.4, one arrives at the penalized log-likelihood criterion

$$\begin{aligned} PL(\alpha) = & \sum_{t=1}^q \sum_{i \in R_t} l_{it}(\alpha_t) - \frac{1}{2} (\alpha_0 - a_0)' Q_0^{-1} (\alpha_0 - a_0) \\ & - \frac{1}{2} \sum_{t=1}^q (\alpha_t - F_t \alpha_{t-1})' Q_t^{-1} (\alpha_t - F_t \alpha_{t-1}), \end{aligned}$$

where

$$l_{it}(\alpha_t) = y_{it} \log \lambda(t|x_i(t)) + (1 - y_{it}) \log(1 - \lambda(t|x_i(t))),$$

where  $\lambda(t|x_i(t)) = F(z'_{it}\alpha_t)$  is the individual log-likelihood contribution as in Sections 9.2.2 and 9.2.3. The risk set is  $R_t = \{i : t \leq t_i\}$  if censoring is considered to happen at the end of the interval  $[a_{t-1}, a_t]$ . As a very simple example, consider a logit model

$$\lambda(t) = \exp(\alpha_t) / (1 + \exp(\alpha_t))$$

for the hazard function without covariates. Together with a first-order random walk model for  $\alpha_t$ , the penalized log-likelihood criterion becomes

$$L = \sum_{t=1}^q \sum_{i \in R_t} l_{it}(\alpha_t) - \frac{1}{2\sigma_0^2} (\alpha_0 - a_0)^2 - \frac{1}{2\sigma^2} \sum_{t=1}^q (\alpha_t - \alpha_{t-1})^2.$$

The last term penalizes large deviations between successive baseline parameters and leads to smoothed estimates. Other stochastic models for the states  $\alpha_t$ , such as second-order random walk models, lead to other forms of the penalty term, but with a similar effect.

For competing risk models, the binomial likelihood contributions have to be replaced by multinomial ones, with a mult categorial model for hazard rates. Posterior mode smoothers maximize the penalized log-likelihood criterion (9.4.5). They can be obtained by modifications of extended or iterative Kalman filtering and smoothing of Chapter 8; see Fahrmeir (1994) and Fahrmeir & Wagenpfeil (1996) for details. A data-driven choice of hyperparameters is possible in combination with an EM-type algorithm. Biller (2000b) extends this approach to models with frailty effects.

If one is not willing to rely on the Bayesian smoothness priors implied by the transition model (9.4.7), one may also start directly from the penalized likelihood criterion. Then the technique may be viewed as discrete-time spline smoothing for survival models.

## Fully Bayesian Inference via MCMC

Alternatively, *fully Bayesian inference* can be based on MCMC techniques for longitudinal data (Section 8.4). For models with time-varying effects as in (9.4.4), this is described in Fahrmeir & Knorr-Held (1997), with an application to unemployment data from the German socio-economic panel GSOEP.

Predictors of the form (9.4.4) are appropriate if only effects of *one time scale*, namely duration time, are modelled and estimated nonparametrically. In the presence of additional time scales, such as calendar time and age, or further nonlinear covariate effects as in (9.4.5) and (9.4.6), fully Bayesian semiparametric inference via MCMC as described in Section 5.4 is more flexible. In addition, individual or group-specific random effects (“frailty”) can be added to the predictor (9.4.6) as in Chapter 8. Posterior samples can be drawn as described there and in more detail in Fahrmeir & Lang (1999). Of course, other semiparametric Bayesian approaches mentioned in Section 5.4 can also be adapted to discrete-time survival and competing risks data.

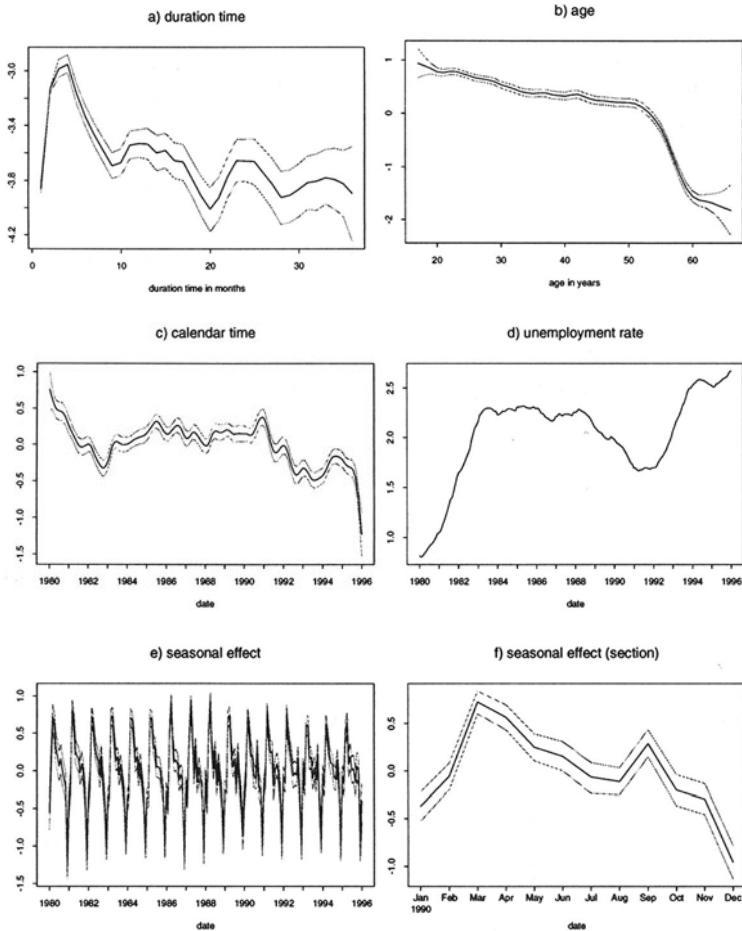
### Example 9.4: Duration of unemployment: A spatio-temporal analysis

Dynamic discrete-time duration models were applied to monthly unemployment data from the socioeconomic panel in Fahrmeir & Wagenpfeil (1996) for simultaneous posterior mode estimation of time-varying baseline and covariate effects via iterative Kalman filtering and smoothing. A full Bayesian analysis via MCMC is given in Fahrmeir & Knorr-Held (1997). Here we analyze monthly unemployment data from the official register of the German Federal Employment. The sample consists of about 6300 males having full-time jobs in West Germany, with unemployment periods in the period from 1980 to 1995. In addition to the usual covariates and the calendar time of the beginning and end of unemployment spells, the district in which the unemployed live is available. In contrast to data from the socioeconomic panel, this allows spatio-temporal modelling on a small regional scale.

Our analysis is based on the following covariates:

- $D$  calendar time measured in months,
- $A$  age (in years) at the beginning of unemployment,
- $N$  nationality, dichotomous with categories “German” and “foreigner” (= reference category),
- $U$  unemployment compensation, trichotomous with categories “unemployment benefit” (= reference category), “unemployment assistance” ( $U_1$ ), and “subsistence allowance” ( $U_2$ ),
- $C$  district in which the unemployed live.

Note that calendar time  $D$  and unemployment compensation  $U$  are both duration-time-dependent covariates. With the event indicators



**Figure 9.4.** Estimated nonparametric functions and seasonal effect. Shown are the posterior means within 80% credible regions.

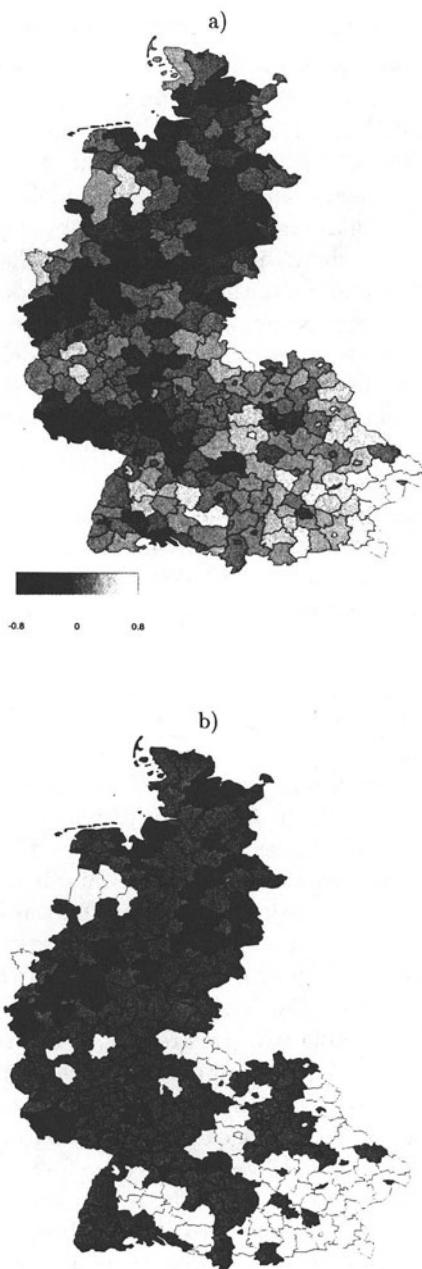
$$y_{it} = \begin{cases} 1, & \text{individual } i \text{ gets a full-time job after month } t, \\ 0, & \text{otherwise,} \end{cases}$$

the hazard function is modelled by a logit model

$$\lambda(t|x_i(t)) = P(y_{it} = 1|\eta_{it}) = \exp(\eta_{it})/(1 + \exp(\eta_{it})),$$

with semiparametric predictor

$$\begin{aligned} \eta_{it} = & f_{(1)}(t) + f_{(2)}^{Tr}(D_{it}) + f_{(3)}^S(D_{it}) + f_{(4)}(A_i) \\ & + f_{(5)}(C_i) + \beta_1 N_i + \beta_2 U_{it}^1 + \beta_3 U_{it}^2. \end{aligned}$$



**Figure 9.5.** Posterior mean and posterior probabilities of the district-specific effect.

Bayesian spatio-temporal inference via MCMC is carried out as described in Section 8.5, (8.1.9). The baseline effect  $f_{(1)}(t)$ , the calendar time trend  $f_{(2)}^{Tr}(D_{it})$ , and the effect of age  $f_{(4)}(A_i)$  are estimated nonparametrically using second-order random walks. For the seasonal component  $f_{(3)}^S(D_{it})$  we choose the smoothness prior (8.1.9) to allow for a flexible time-varying seasonal effect. For the spatial covariate “district” we choose the (Gaussian) Markov random field prior (8.5.3). The influence of the categorical covariates nationality, and unemployment compensation are modelled as fixed effects with diffuse priors. The estimation results of the nonparametric terms and the seasonal component are shown in Figure 9.4a)–f). The baseline effect (Panel a)) increases for the first 4 months and then slopes downward. Therefore, the probability of finding a new job is best in the first few months of unemployment. The effect of age in panel b) is slowly declining until age 52, dramatically declining for people older than 53. Panel c) displays the calendar time trend. For comparison with the estimated trend, the absolute number of unemployed people in Germany from 1980 to 1995 is shown in panel d). Not surprisingly, a declining calendar time trend corresponds to an increase in the unemployment rate, and vice versa. So the estimated calendar time trend accurately reflects the economic trend of the labor market in Germany. The extreme decline toward 1996 is a boundary effect, which should not be interpreted. The estimated seasonal pattern (panel e)) is relatively stable over the observation period. For better insight, a section of the seasonal pattern for 1988 is displayed in panel f). It shows typical peaks in spring and autumn, a global minimum in winter, and a local minimum in July and August. Low hiring rates in summer can be explained by the distribution of holidays and vacations. In Figure 9.5a) the estimated posterior mean of the district-specific effect  $f_{(5)}(C_i)$  is displayed, showing a strong spatial pattern, with better chances of getting a new job in the southern part of West Germany, and lower chances in the middle and the north.

The dark spots in the map mark areas that are known for their economical problems during the 1980s and 1990s. This becomes even more obvious with Figure 9.5b), showing areas with strictly positive (white) and strictly negative (black) credible intervals. Areas where the credible interval contains zero appear in gray. Table 9.3 gives results of the remaining effects. Germans have improved job chances compared to foreigners, but the effects are not overwhelmingly large. The estimate of  $-0.72$  for the subsistence effect is more or less zero. Due to effect coding, the effect of insurance based unemployment benefits is  $0.72 = 0.72 + 0$  and is therefore clearly positive. At first sight, this result seems to contradict the widely held conjecture about the negative side effects of unemployment benefits. However, it may be that the variable “unemployment benefit” also acts as a surrogate variable for those who have worked and therefore contributed regularly to the insurance system in the past.  $\square$

covariate	mean	10% quantile	90% quantile
$N$	0.13	0.09	0.16
$U^1$	0	-0.04	0.04
$U^2$	-0.72	-0.79	-0.66

**Table 9.3.** Estimates of constant parameters in the unemployment data

### Example 9.5: Head and neck cancer

Efron (1988) considers a head and neck cancer study where time is discretized by one-month intervals. Table 9.4 shows the data (from Efron, 1988, Table 1). Figure 9.6 shows a cubic spline fit following (9.2.16), where the cut-off point is chosen by  $t_c = 11$  as suggested by Efron (1988).

Moreover, posterior mode smoothing based on the generalized Kalman filter and smoother is given in Figure 9.6. The underlying model is the simple logistic model  $\lambda(t) = \exp(\alpha_t)/(1 + \exp(\alpha_t))$  and first-order random walk  $\alpha_t = \alpha_{t-1} + \xi_t$ . Alternatively, the data are smoothed by the categorical kernel estimate (9.4.2) based on Nadaraya-Watson weights with normal kernels and the Aitchison and Aitken kernel as discrete kernel. The smoothing parameters chosen by cross-validation are  $\nu_0^* = 0.5$  for time-dependent smoothing and  $\lambda = 0.97$ . In Figure 9.7 the life table estimate is also given by full cubicles. The life table estimate is quite jagged, whereas the nonparametric estimate based on a cross-validated smoothing parameter is a rather smooth function. It is seen that for this simple data set the estimates yield rather similar results: After a short increase, the hazard function decreases. Some difference is seen in the decreasing branch. The spline fit is cubic for  $t \leq 11$  and linear for  $t \geq 11$ . Consequently, there is a steady decrease for  $t \geq 11$ . Posterior mode smoothing, in particular the smoother kernel estimate, shows that the decline is very slow for  $11 \leq t \leq 18$  and steeper for  $t \geq 18$ . These estimates are more locally determined and are closer to the relative frequencies, which in unsmoothed form give no feeling for the underlying hazard function.  $\square$

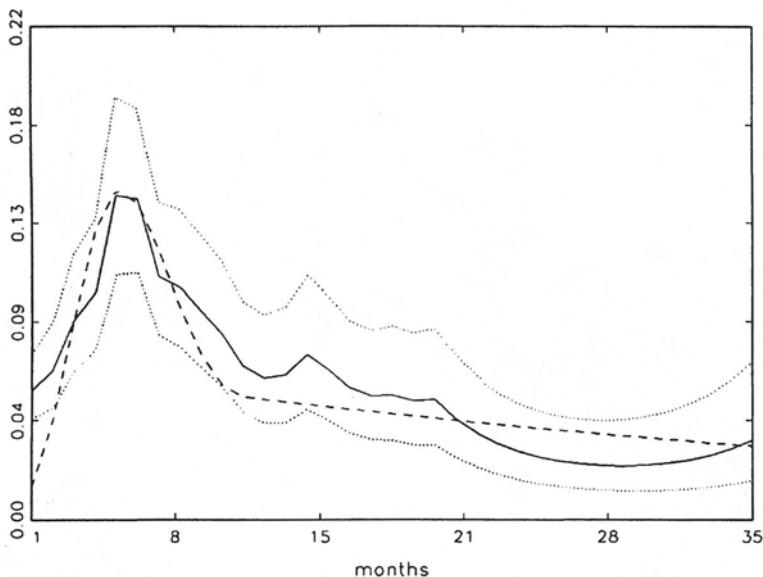
## 9.5 Remarks and Further Reading

In this chapter we focused on parametric discrete-time survival models and on some semiparametric extensions. Other semi- or nonparametric approaches like spline smoothing or CART can also be adapted to the situation of duration and, more generally, event history data; see, for example, Klinger, Dannegger & Ulm (2000) for recent work.

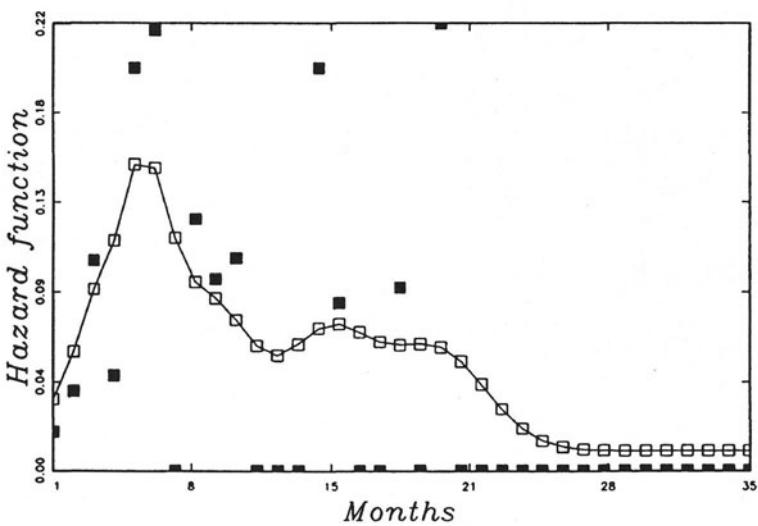
For continuous time, counting process approaches (Andersen, Borgan, Gill & Keiding, 1993) offer attractive nonparametric alternatives to the smoothing methods considered here.

**Table 9.4.** Head and neck cancer (Efron, 1988)

Month	Patients at risk	Deaths	Withdrawals	Month	Patients at risk	Deaths	Withdrawals
1	51	1	0	25	7	0	0
2	50	2	0	26	7	0	0
3	48	5	1	27	7	0	0
4	42	2	0	28	7	0	0
5	40	8	0	29	7	0	0
6	32	7	0	30	7	0	0
7	25	0	1	31	7	0	0
8	24	3	0	32	7	0	0
9	21	2	0	33	7	0	0
10	19	2	1	34	7	0	0
11	16	0	1	35	7	0	0
12	15	0	0	36	7	0	0
13	15	0	0	37	7	1	1
14	15	3	0	38	5	1	0
15	12	1	0	39	4	0	0
16	11	0	0	40	4	0	0
17	11	0	0	41	4	0	1
18	11	1	1	42	3	0	0
19	9	0	0	43	3	0	0
20	9	2	0	44	3	0	0
21	7	0	0	45	3	0	1
22	7	0	0	46	2	0	0
23	7	0	0	47	2	1	1
24	7	0	0				



**Figure 9.6.** Cubic-linear spline fit for head and neck cancer data (---) and posterior mode smoother (—) with  $\pm$  standard deviation confidence bands.



**Figure 9.7.** Smoothed kernel estimate for head and neck cancer data.

# Appendix A

## A.1 Exponential Families and Generalized Linear Models

### Simple Exponential Families

We say that a  $q$ -dimensional random variable  $y$ , or more exactly its distribution resp. density, belongs to a simple exponential family if its discrete or continuous density with respect to a  $\sigma$ -finite measure has the form

$$f(y|\theta, \lambda) = \exp\{[y'\theta - b(\theta)]/\lambda + c(y, \lambda)\}, \quad (\text{A.1.1})$$

with  $c(y, \lambda) \geq 0$  and measurable. Jorgensen (1992) calls (A.1.1) an exponential dispersion model. The  $q$ -dimensional parameter  $\theta \in \Theta \subset \mathbb{R}^q$  is the natural parameter of the family and  $\lambda > 0$  is a *nuisance* or *dispersion* parameter. For given  $\lambda$ , we will generally assume that  $\Theta$  is the natural parameter space, i.e., the set of all  $\theta$  satisfying  $0 < \int \exp\{y'\theta/\lambda + c(y, \lambda)\} dy < \infty$ . Then  $\Theta$  is convex, and in the interior  $\Theta^0$ , assumed to be nonvoid, all derivatives of  $b(\theta)$  and all moments of  $y$  exist. In particular,

$$E_\theta(y) = \mu(\theta) = \frac{\partial b(\theta)}{\partial \theta}, \quad (\text{A.1.2})$$

$$\text{cov}_\theta(y) = \Sigma(\theta) = \lambda \frac{\partial^2 b(\theta)}{\partial \theta \partial \theta'}. \quad (\text{A.1.3})$$

The covariance matrix  $\Sigma(\theta)$  is supposed to be positive definite in  $\Theta^0$  so that  $\mu : \Theta^0 \rightarrow M = \mu(\Theta^0)$  is injective. Inserting the inverse function  $\theta(\mu)$  into  $\partial^2 b(\theta) / \partial \theta \partial \theta'$  gives the variance function

$$v(\mu) = \frac{\partial^2 b(\theta(\mu))}{\partial \theta \partial \theta'}$$

and

$$\text{cov}(y) = \lambda v(\mu)$$

as a function of  $\mu = E y$  and the dispersion parameter  $\lambda$ . As a generalization of (A.1.1), Zhao, Prentice & Self (1992) introduced “partly” exponential models, a very broad class of distributions.

## Generalized Linear Models

Generalized linear models (GLMs) for independent  $q$ -dimensional observations  $y_1, \dots, y_n$  and covariates  $x_1, \dots, x_n$  are characterized by the following structure:

- (i) The  $\{y_i\}$  are independent with densities from simple exponential families

$$f(y_i|\theta_i, \lambda_i), \quad i = 1, \dots, n,$$

with  $\lambda_i = \phi/\omega_i$ , where  $\omega_i$  are *known* weights and  $\phi$  a, possibly unknown, nuisance parameter that is *constant* across observations.

- (ii) The covariate vector  $x_i$  influences  $y_i$  in the form of a  $q$ -dimensional *linear predictor*

$$\eta_i = Z_i\beta, \quad (\text{A.1.4})$$

where  $\beta$  is a  $p$ -dimensional parameter out of an admissible set  $B \subset \mathbf{R}^p$ , and  $Z_i = Z(x_i)$ , the  $(q \times p)$ -*design matrix*, is a function of the covariates.

- (iii) The linear predictor  $\eta_i$  is related to the mean  $\mu_i = \mu(\theta_i)$  by the *response function*  $h : \mathbf{R}^q \rightarrow M$ ,

$$\mu_i = h(\eta_i) = h(Z_i\beta). \quad (\text{A.1.5})$$

If the inverse  $g = h^{-1} : M \rightarrow \mathbf{R}^q$  of  $h$  exists, then

$$g(\mu_i) = \eta_i = Z_i\beta, \quad (\text{A.1.6})$$

and  $g$  is called the *link function*.

For some theoretical purposes it is more convenient to relate the linear predictor  $\eta_i$  to the natural parameter  $\theta_i$  (instead of  $\mu_i = \mu(\theta_i)$ ) by  $u = (g \circ \mu)^{-1} = \mu^{-1} \circ h$ , i.e.,

$$\theta_i = u(Z_i\beta) = \mu^{-1}(h(Z_i\beta)), \quad (\text{A.1.7})$$

as in the original definition of Nelder & Wedderburn (1972).

Of special importance are *natural link functions*  $g = \mu^{-1}$  and  $u$  the identity. Then we obtain a linear model  $\theta_i = Z_i\beta$  for the natural parameter. Natural link functions are, e.g., the logit function  $\log(\mu/(1 - \mu))$  for the binomial distribution and the  $\log(\mu)$  function for the Poisson distribution.

In the above definitions, covariates  $x_i$  and, as a consequence, design matrices  $Z_i$  are tacitly assumed to be known constants. For *stochastic* regressors  $x_i$ , all definitions have to be understood conditionally: The  $y_1, \dots, y_n$  are conditionally independent, and  $f(y_i|\theta_i, \lambda_i), \mu(\theta_i)$ , etc., are conditional densities, means, and so forth.

Inserting (A.1.7) in  $\mu(\theta_i)$  gives back (A.1.5), and insertion in  $\Sigma(\theta_i)$  yields  $\text{cov}(y)$  as a function of  $\beta$ . To stress dependence on  $\beta$ , we will often write

$$E_\beta y_i = \mu_i(\beta), \quad \text{cov}_\beta y_i = \Sigma_i(\beta). \quad (\text{A.1.8})$$

## Log-likelihood, Score Function, Expected and Observed Information

For the following we assume that

- (i) the admissible parameter space  $B$  is open,
- (ii)  $h(Z_i\beta) \in M = \mu(\Theta^0)$ ,  $i = 1, 2, \dots$ , for all  $\beta \in B$ ,
- (iii)  $h$ ,  $g$ , and  $u$  are twice continuously differentiable,  $\det(\partial g/\partial\eta) \neq 0$ ,
- (iv)  $\sum_{i=1}^n Z_i Z'_i$  has full rank for  $n \geq n_0$ , say.

Condition (ii) is necessary to have a well-defined GLM for all  $\beta$ . Conditions (i) and (iii) guarantee that the second derivatives of the log-likelihood are continuous. The rank condition and  $\det(\partial g/\partial\eta) \neq 0$  will ensure that the expected information matrix is positive definite for all  $\beta, n \geq n_0$ .

The log-likelihood contribution of observation  $y_i$  for  $\beta$  is, up to a constant not dependent on  $\beta$ ,

$$l_i(\beta) = [y'_i \theta_i - b(\theta_i)]/\lambda_i.$$

The individual score function  $s_i(\beta) = \partial l_i(\beta)/\partial\beta$  is obtained by differentiation, using (A.1.2) and (A.1.3):

$$\begin{aligned} s_i(\beta) &= \frac{\partial h(Z_i\beta)}{\partial\beta} \Sigma_i^{-1}(\beta)(y_i - \mu_i(\beta)) \\ &= Z'_i D_i(\beta) \Sigma_i^{-1}(\beta)(y_i - \mu_i(\beta)), \end{aligned} \tag{A.1.9}$$

with  $D_i(\beta) = \partial h(\eta)/\partial\eta$  the Jacobian of  $h(\eta)$ , evaluated at  $\eta_i = Z_i\beta$ .

An equivalent form is

$$s_i(\beta) = Z'_i W_i(\beta) \frac{\partial g(\mu_i)}{\partial\mu'}(y_i - \mu_i(\beta)), \tag{A.1.10}$$

with the “weight matrix”

$$W_i(\beta) = \left[ \frac{\partial g(\mu_i)}{\partial\mu'} \Sigma_i(\beta) \frac{\partial g(\mu_i)}{\partial\mu} \right]^{-1} = D_i(\beta) \Sigma_i^{-1}(\beta) D_i(\beta)'.$$

From (A.1.9) and (A.1.10) it is easily seen that  $E_\beta s_i = 0$ , as is common in ML estimation under regularity conditions.

The contribution of  $y_i$  to the Fisher information or expected information is

$$\begin{aligned} F_i(\beta) &= \text{cov}_\beta(s_i(\beta)) = E_\beta(s_i(\beta)s'_i(\beta)) \\ &= Z'_i D_i(\beta) \Sigma_i^{-1}(\beta) D_i(\beta)' Z_i = Z'_i W_i(\beta) Z_i. \end{aligned} \tag{A.1.11}$$

Further differentiation shows that the observed information of  $y_i$  is

$$F_{i,obs}(\beta) = \frac{-\partial^2 l_i(\beta)}{\partial \beta \partial \beta'} = F_i(\beta) - R_i(\beta),$$

with

$$R_i(\beta) = \sum_{r=1}^q Z_i' U_{ir}(\beta) Z_i (y_{ir} - \mu_{ir}(\beta)), \quad (\text{A.1.12})$$

where  $U_{ir}(\beta) = \partial^2 u_r(Z_i \beta) / \partial \eta \partial \eta'$ , and  $u_r(\eta)$ ,  $y_{ir}$  and  $\mu_{ir}(\beta)$  are the components of  $u(\eta)$ ,  $y_i$ ,  $\mu_i(\beta)$ . It is easily seen that

$$E_\beta(F_{i,obs}(\beta)) = F_i(\beta), \quad E_\beta(R_i(\beta)) = 0.$$

For natural link functions, where  $\theta_i = \eta_i = Z_i \beta$ , the expressions for  $s_i(\beta)$  and  $F_i(\beta)$  simplify, since  $D_i(\beta) = \partial^2 b(\theta) / \partial \theta \partial \theta'$  evaluated at  $\theta_i = Z_i \beta = \sum_i (\beta) \omega_i / \phi$ . Also, expected and observed information matrices coincide.

The formula for the score function on the right side of (A.1.9) was obtained by applying the chain rule for differentiation to  $\mu(\beta) = h(Z\beta)$ , so that  $\partial \mu / \partial \beta = Z'D(\beta)$  with  $D(\beta) = \partial h / \partial \eta$ , evaluated at  $\eta = Z\beta$ . Defining directly the first derivative of  $\mu(\beta)$  by

$$M(\beta) := \partial \mu / \partial \beta,$$

(A.1.9) becomes

$$s_i(\beta) = M_i(\beta) \Sigma_i^{-1}(\beta) (y_i - \mu_i(\beta)), \quad (\text{A.1.13})$$

and the Fisher information is

$$F_i(\beta) = M_i(\beta) \Sigma_i^{-1}(\beta) M_i'(\beta). \quad (\text{A.1.14})$$

For generalized linear models these formulas are, of course, equivalent to (A.1.9) and (A.1.11), because  $M(\beta) = Z'D(\beta)$ . However, (A.1.13) and (A.1.14) remain valid in the more general case of nonlinear exponential family models, where the assumption of a linear predictor is dropped and the mean is assumed to be a general nonlinear function

$$\mu(\beta) = \mu(x; \beta)$$

of covariates and parameters.

## A.2 Basic Ideas for Asymptotics

In the main text we made repeated use of statements like “under appropriate regularity conditions (quasi-)maximum likelihood estimators are consistent and asymptotically normal,” “test statistics are asymptotically  $\chi^2$ -distributed,” etc. This appendix briefly describes the line of arguments that lead to such asymptotic results. It may be of interest for more mathematically oriented readers but is not a necessary requirement for reading the text.

For a compact notation the stochastic versions  $o_p$  and  $O_p$  of the “Landau” symbols  $o$  and  $O$  are convenient. Let  $x_n$  be a sequence of deterministic vectors and  $r_n$  a sequence of positive real numbers. Then

$$x_n = o(r_n) \Leftrightarrow \|x_n\|/r_n < c, n \geq N$$

for all  $c > 0$  and sufficiently large  $N$ , and

$$x_n = O(r_n) \Leftrightarrow \|x_n\|/r_n < C \quad \text{for all } n \geq N$$

for some constant  $C$  and sufficiently large  $N$ . Obviously,  $\|x_n\| = o(r_n)$  is equivalent to  $\|x_n\|/r_n \rightarrow 0$  for  $n \rightarrow \infty$ . The “p-versions” are

$$x_n = o_p(r_n) \Leftrightarrow P(\|x_n\|/r_n < c) \geq 1 - \varepsilon, \quad n \geq N,$$

for every  $c > 0, \varepsilon > 0$ , and sufficiently large  $N$ , and

$$x_n = O_p(r_n) \Leftrightarrow \text{for every } \varepsilon > 0 \text{ there is a } C > 0$$

with

$$P(\|x_n\|/r_n < C) \geq 1 - \varepsilon, \quad n \geq N,$$

for sufficiently large  $N$ .

Obviously,  $x_n = o_p(r_n)$  is equivalent to  $\|x_n\|/r_n \xrightarrow{p} 0$  (in probability), with the special case  $x_n \xrightarrow{p} 0$  for  $r_n \equiv 1$ . If  $x_n = O_p(1)$ , then  $x_n$  is said to be bounded in probability. For equivalent definitions and useful properties of  $o_p, O_p$  see, e.g., Prakasa Rao (1987).

Let us first discuss asymptotics of maximum likelihood inference for independent but, in general, not identically distributed observations  $y_1, \dots, y_n$ . We consider local MLEs  $\hat{\beta}_n$  for  $\beta$  in the interior of the parameter space  $B$ , i.e., MLEs that are local maximizers of the log-likelihood or, equivalently, corresponding roots of the ML equations. In the following we tacitly assume that the dispersion parameter  $\phi$  is known. However, results remain valid if  $\phi$  is replaced by a consistent estimate. Let

$$\begin{aligned}
\ell_n(\beta) &= \sum_{i=1}^n \frac{y_i' \theta(\mu_i) - b(\theta(\mu_i))}{\phi} \omega_i, \quad \mu_i = h(Z_i \beta), \\
s_n(\beta) &= \sum_{i=1}^n Z_i' D_i(\beta) \Sigma_i^{-1}(\beta) (y_i - \mu_i), \\
F_n(\beta) &= \sum_{i=1}^n Z_i' D_i(\beta) \Sigma_i^{-1}(\beta) D_i'(\beta) Z_i, \\
H_n(\beta) &= -\partial^2 \ell_n(\beta) / \partial \beta \partial \beta'
\end{aligned}$$

denote the log-likelihood, score function, expected and observed information matrix of the sample  $y_1, \dots, y_n$ . The index  $n$  is introduced to make the dependence of likelihoods, MLEs, etc., on sample size  $n$  explicit. Note that in this section  $\ell_n(\beta)$ ,  $s_n(\beta)$ , etc., denote log-likelihoods, score functions of the whole sample  $y_1, \dots, y_n$ , and not corresponding individual contributions of observation  $y_n$ .

Given a finite sample of size  $n$ , there are situations where no maximum of  $\ell_n(\beta)$  and no root of  $s_n(\beta)$  exist in the interior of  $B$ . Furthermore, local and global maxima need not coincide in general. However, for many important models local and global maxima are identical and unique if they exist; see Chapter 2 for references. For asymptotic considerations only *asymptotic existence*, i.e.,

$$P(s_n(\hat{\beta}_n) = 0, H_n(\hat{\beta}_n) \text{ p.d.}) \rightarrow 1$$

for  $n \rightarrow \infty$  is required.

We give a short outline of standard  $n^{1/2}$ -asymptotics. For this case typical “regularity assumptions” are weak conditions on third derivatives of  $\ell_n(\beta)$ , existence of third moments of  $y$ , and in particular convergence of  $F_n(\beta)/n = \text{cov } s_n(\beta)/n = E H_n(\beta)/n$  to a p.d. limit, say

$$F_n(\beta)/n \rightarrow F(\beta). \tag{A.2.1}$$

Then the following asymptotic results hold under regularity assumptions:  
Asymptotic normality of the score function,

$$n^{-1/2} s_n(\beta) \xrightarrow{d} N(0, F(\beta)), \tag{A.2.2}$$

asymptotic existence and consistency of  $\hat{\beta}_n$ , and asymptotic normality of the (normed) MLE,

$$n^{1/2} (\hat{\beta} - \beta) \xrightarrow{d} N(0, F(\beta)^{-1}). \tag{A.2.3}$$

We now give a brief outline of how to prove these statements (compare with McCullagh, 1983). The score function  $s_n(\beta)$  is the sum of individual contributions  $s_i(\beta)$  with  $E s_i(\beta) = 0$  and finite variance. The total variance is  $\text{cov } s_n(\beta) = F_n(\beta)$ . Because  $F_n(\beta)/n \rightarrow F(\beta)$  no finite set of individual contributions is dominant, and some central limit theorem implies

$$n^{-1/2} s_n(\beta) \sim N(0, F_n(\beta)/n) + O_p(n^{-1/2})$$

and (A.2.2).

Proofs for consistency can be based on the following: Because  $E y_i = \mu_i(\beta)$  for the “true” parameter  $\beta$ , we have

$$E s_n(\beta) = \sum_{i=1}^n Z'_i D_i(\beta) \Sigma_i^{-1}(\beta) (E y_i - \mu_i(\beta)) = 0$$

where the “true”  $\beta$  is a root of the *expected* score function  $E s_n(\beta)$ . The MLE  $\hat{\beta}_n$  is a root of the “observed” score function

$$s_n(\beta) = \sum_{i=1}^n Z'_i D_i(\beta) \Sigma_i^{-1}(\beta) (y_i - \mu_i(\beta)).$$

By some law of large numbers

$$[s_n(\beta) - E s_n(\beta)]/n \longrightarrow 0$$

in probability. With some additional arguments one obtains  $\hat{\beta}_n \rightarrow \beta$  in probability. Other proofs make more direct use of the asymptotic behavior of  $\ell_n(\beta)$  in small neighborhoods of  $\beta$ .

Asymptotic normality (A.2.3) of  $\hat{\beta}$  is shown as follows: Expansion of the ML equations  $s_n(\hat{\beta}_n) = 0$  about  $\beta$  gives

$$s_n(\beta) = \bar{H}_n(\hat{\beta}_n - \beta),$$

where  $\bar{H}_n = H_n(\bar{\beta})$  for some  $\bar{\beta}$  between  $\hat{\beta}_n$  and  $\beta$ , and

$$\hat{\beta}_n - \beta = \bar{H}_n^{-1} s_n(\beta) = F_n^{-1}(\beta) s_n(\beta) + O_p(n^{-1}). \quad (\text{A.2.4})$$

Multiplying both sides by  $n^{1/2}$  and using (A.2.1) and (A.2.2), one obtains the  $n^{1/2}$ -asymptotic normality result (A.2.3).

Asymptotic  $\chi^2$ -distributions for the (log-)likelihood ratio, the Wald and score statistic are again obtained by expansion of  $\ell_n(\beta)$  in a Taylor series about  $\hat{\beta}_n$ :

$$\begin{aligned}\lambda_n &= -2(\ell_n(\beta) - \ell_n(\hat{\beta}_n)) = (\hat{\beta}_n - \beta)' \bar{H}_n(\hat{\beta}_n - \beta) \\ &= (\hat{\beta}_n - \beta)' \hat{F}_n(\hat{\beta}_n) + O_p(n^{-1/2}) = s_n'(\beta) \hat{F}_n^{-1}(\beta) s_n(\beta) + O_p(n^{-1/2}).\end{aligned}$$

As  $n^{1/2}(\hat{\beta} - \beta)$  and  $n^{-1/2}s_n(\beta)$  are asymptotically normal, it follows that the likelihood ratio statistic  $\lambda_n$  for testing  $H_0 : \beta = \beta_0$  against  $H_1 : \beta \neq \beta_0$  is asymptotically  $\chi^2$  and asymptotically equivalent to the Wald and score statistic. For composite null hypotheses,  $\lambda_n = -2(\ell_n(\hat{\beta}_n) - \ell_n(\beta_0))$  can be expressed as the difference of two quadratic forms, and this difference can be shown to be asymptotically  $\chi^2$  with the correct degree of freedoms; see, e.g., Cox & Hinkley (1974) for details.

The standard  $n^{1/2}$ -approach to first-order asymptotics sketched above mimics corresponding proofs for i.i.d. observations. Apart from certain domination conditions, the convergence assumption  $F_n(\beta)/n \rightarrow F(\beta)$  in (A.2.1) is the really crucial one. This assumption will hold in the situation of “stochastic regressors,” where pairs  $(y_i, x_i)$  are independent realizations of  $(y, x)$ , and  $n^{1/2}$ -asymptotics are appropriate. However, the convergence condition (A.2.1) will typically not hold in the situation of trending regressors, for example, in planned experiments. Based on the same general ideas but applied more carefully, asymptotic results can be obtained that allow for considerably more heterogeneity of the observations. In particular, convergence conditions like (A.2.1) can be completely avoided if matrix normalization by square roots of the (expected or observed) information matrix is used instead of  $n^{1/2}$ -normalization as in (A.2.3). Under rather weak assumptions one obtains asymptotic normality of normed score functions and MLEs in the form

$$\begin{aligned}F_n^{-1/2}(\beta)s_n(\beta) &\xrightarrow{d} N(0, I), \\ F_n^{T/2}(\hat{\beta}_n)(\hat{\beta}_n - \beta) &\xrightarrow{d} N(0, I),\end{aligned}$$

where  $F_n^{-1/2}$  is the inverse of a left square root  $F_n^{1/2}$  of  $F_n$  and  $F_n^{T/2} = (F_n^{1/2})^T$  is the corresponding right square root. For details we refer the reader to Fahrmeir & Kaufmann (1985, 1986) and Fahrmeir (1987b, 1988) for results of this kind.

Next let us briefly discuss asymptotics for quasi-likelihood models in Section 2.3 and later in the text, where it is assumed that the mean is still correctly specified, i.e.,

$$\mu_i = E y_i = h(Z_i \beta),$$

but the variance function  $\Sigma_i = \Sigma(\mu_i)$  may be different from the true variance  $S_i = \text{cov } y_i$ . Going through the outline of asymptotics for genuine likelihood inference above, it can be seen that most arguments go through unchanged up to the following modification: The quasi-score function  $s_n(\beta)$  is asymptotically normal, but with

$$\text{cov } s_n(\beta) = V_n(\beta) = \sum_{i=1}^n Z_i D_i \Sigma_i^{-1} S_i \Sigma_i^{-1} D_i' Z_i'.$$

The expected “quasi-information” is still given by the common form

$$F_n(\beta) = \sum_{i=1}^n Z_i' D_i \Sigma_i^{-1} D_i' Z_i$$

so that

$$V_n(\beta) = F_n(\beta) \quad \text{if } \Sigma_i = S_i.$$

If we assume

$$V_n(\beta)/n \rightarrow V(\beta) \quad \text{p.d.},$$

then

$$n^{-1/2} s_n(\beta) \xrightarrow{d} N(0, V(\beta))$$

instead of (A.2.2). Inserting this result in (A.2.4), one obtains asymptotic normality

$$n^{1/2} (\hat{\beta}_n - \beta) \xrightarrow{d} N(0, F(\beta)^{-1} V(\beta) F(\beta)^{-1})$$

of the QLME with an adjusted asymptotic covariance matrix in the form of a sandwich matrix. Replacing the sandwich matrix by an estimate, one obtains

$$\hat{\beta}_n \xrightarrow{a} N(\beta, \hat{F}_n^{-1} \hat{V}_n \hat{F}_n^{-1}),$$

where  $\hat{F}_n = F_n(\hat{\beta}_n)$  and

$$\hat{V}_n = \sum_{i=1}^n Z_i' \hat{D}_i \hat{\Sigma}_i^{-1} (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' \hat{\Sigma}_i^{-1} \hat{D}_i' Z_i,$$

with  $\hat{D}_i = D_i(\hat{\beta}_n)$ ,  $\hat{\Sigma}_i = \Sigma_i(\hat{\beta}_n)$ ,  $\hat{\mu}_i = \mu_i(\hat{\beta}_n)$ .

Corresponding results with rigorous proofs for the general approach using matrix normalization can be found in Fahrmeir (1990). For dependent observations as in generalized autoregressive linear models (Section 6.1), the main ideas go through again under the following modification: The sequence  $s_i(\beta)$  of individual score function distributions is no longer a sequence of independent random variables, but a martingale difference sequence. Applying laws and limit theorems for martingales, consistency, and asymptotic normality of the MLE can be shown under appropriate assumptions; see Kaufmann (1987) for detailed proofs and, e.g., the survey in Fahrmeir (1987a).

## A.3 EM Algorithm

The EM algorithm is a general iterative method to obtain maximum likelihood estimators in incomplete data situations. It was first proposed by Hartley (1958) and was generalized by Dempster, Laird & Rubin (1977).

Let  $y \in \mathbf{R}^n$  denote a vector of observed data and  $z \in \mathbf{R}^m$  a vector of unobservable data. Then the hypothetical complete data are given by  $(y, z)$  and the incomplete data that were observed are given by  $y$ . Furthermore, let  $f(y, z; \theta)$  denote the joint density of the complete data depending on an unknown parameter vector  $\theta \in \Theta$ , and  $k(z|y; \theta)$  is the conditional density of the unobservable data  $z$ , given the observed data  $y$ , which also depends on  $\theta$ .

To obtain the maximum likelihood estimator (MLE) for  $\theta$ , the marginal log-likelihood

$$l(\theta) = \log \int f(y, z; \theta) dz \quad (\text{A.3.1})$$

is usually maximized directly. Indirect maximization of (A.3.1) by the EM algorithm avoids the numerical evaluation of the integral by considering

$$l(\theta) = \log f(y, z; \theta) - \log k(z|y; \theta) \quad (\text{A.3.2})$$

instead of (A.3.1). The problem in maximizing (A.3.2) is that  $z$  is unobservable. Therefore, expectations are taken on both sides of (A.3.2) with respect to the conditional density  $k(z|y; \theta_0)$ ,  $\theta_0 \in \Theta$ , so that

$$l(\theta) = M(\theta|\theta_0) - H(\theta|\theta_0)$$

is obtained, where

$$\begin{aligned} M(\theta|\theta_0) &= E\{\log f(y, z; \theta)|y; \theta_0\} \\ &= \int \log f(y, z; \theta) k(z|y; \theta_0) dz, \\ H(\theta|\theta_0) &= E\{\log k(z|y; \theta)|y; \theta_0\} = \int \log k(z|y; \theta) k(z|y; \theta_0) dz. \end{aligned} \quad (\text{A.3.3})$$

Then the EM algorithm maximizes  $l(\theta)$  iteratively by maximizing  $M(\theta|\theta_0)$  with respect to  $\theta$ , where  $\theta_0$  is given at each cycle of the iteration. In contrast to the integral in (A.3.1), evaluation of the integral in (A.3.3) is straightforward for many applications (see, e.g., Dempster, Laird & Rubin, 1977). If  $\theta^{(0)}$  denotes a starting value for  $\theta$ , the  $(p+1)$ -th cycle of the EM algorithm consists of the following two steps for  $p = 0, 1, \dots$ :

*E(xpectation)-step:* Compute the expectation  $M(\theta|\theta^{(p)})$  given by (A.3.3)

*M(maximizing)-step:* Determine  $\theta^{(p+1)}$  by  $M(\theta|\theta^{(p)}) \rightarrow \max_{\theta}$ .

The EM algorithm has the desirable property that the log-likelihood  $l$  always increases or stays constant at each cycle: Let  $\hat{\theta}$  maximize  $M(\theta|\theta_0)$ , given  $\theta_0$ . Then we have  $M(\hat{\theta}|\theta_0) \geq M(\theta|\theta_0)$  for all  $\theta$  by definition and  $H(\theta|\theta_0) \leq H(\theta_0|\theta_0)$  for all  $\theta$  by Jensen's inequality, so that

$$l(\hat{\theta}) \geq l(\theta_0).$$

Convergence of the log-likelihood sequence  $l(\theta^{(p)})$ ,  $p = 0, 1, 2, \dots$ , against a global or local maximum or a stationary point  $l_*$  is ensured under weak regularity conditions concerning  $\Theta$  and  $l(\theta)$  (see, e.g., Dempster, Laird & Rubin, 1977). However, if more than one maximum or stationary point exists, convergence against one of these points depends on the starting value. Moreover, convergence of the log-likelihood sequence  $l(\theta^{(p)})$ ,  $p = 0, 1, 2, \dots$ , against  $l_*$  does not imply the convergence of  $(\theta^{(p)})$  against a point  $\theta_*$ , as Wu (1983) and Boyles (1983) pointed out. In general, convergence of  $(\theta^{(p)})$  requires stronger regularity conditions, which are ensured in particular for complete data densities  $f(y, z; \theta)$  of the simple or curved exponential family. For finite mixtures of densities from the exponential family, see Redner & Walker (1984).

The rate of convergence depends on the relative size of the unobservable information on  $\theta$ . If the information loss due to the missing  $z$  is a small fraction of the information in the complete data  $(y, z)$ , the algorithm converges rapidly. On the other hand, the rate of convergence becomes rather slow for parameters  $\theta$  near the boundary of  $\Theta$  and an estimator for the variance-covariance matrix of the MLE for  $\theta$ , e.g., the observed or expected information on  $\theta$  in the observed data  $y$ , is not provided by the EM algorithm. Newton-Raphson or other gradient methods that maximize (A.3.1) directly are generally faster and yield an estimator for the variance-covariance matrix of the MLE. However, the EM algorithm is simpler to implement and numerically more stable. An estimate for the variance-covariance matrix of the MLE is obtained if an additional analysis is applied after the last cycle of the EM algorithm (Louis, 1982). The method can also be used to speed up the EM algorithm (see also Meilijson, 1989). However, for complex complete data structures the procedure is rather cumbersome.

## A.4 Numerical Integration

Numerical integration techniques approximate integrals that cannot be solved analytically. Simpson's rule and quadrature methods are prominent among the techniques for univariate integrals (see Davis & Rabinowitz, 1975). These methods are based on reexpressing a regular function

$f(x) : \mathbf{R} \rightarrow \mathbf{R}$  as the product of a *weight function*  $w(x) : \mathbf{R} \rightarrow \mathbf{R}_+$  and another function  $g(x) : \mathbf{R} \rightarrow \mathbf{R}$ ,

$$f(x) = w(x)g(x).$$

Then, most numerical integration methods approximate an integral by a discrete summation,

$$\int_{\mathbf{R}} f(x)dx \approx \sum_{i=1}^k w_i g(x_i), \quad (\text{A.4.1})$$

where the points  $x_i$  are called nodes, the  $w_i$  are the weights, and the nodes and weights together constitute an integration rule.

Following Stroud (1971) and Davis & Rabinowitz (1984), an integration rule should have at least the following properties to reduce the numerical effort and integration error in (A.4.1) as far as possible:

- Nodes and weights should be easily found and calculated,
- Nodes should lie in the region of integration,
- Weights should all be positive.

Such integration rules are available as long as the weight function  $w(x)$  is known.

### Univariate Gauss-Hermite Integration

For most statistical integration problems the normal density

$$w_N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

with mean  $\mu$  and variance  $\sigma^2$  can be deduced as a weight function.

Such a weight function is the basis of the Gauss-Hermite integration rule. If  $g(x), x \in \mathbf{R}$ , denotes a regular function, the Gauss-Hermite rule approximates integrals of the form

$$\int_{-\infty}^{+\infty} \exp(-x^2)g(x)dx \quad (\text{A.4.2})$$

by the sum

$$\sum_{i=1}^k w_i g(x_i), \quad (\text{A.4.3})$$

where the node  $x_i$  is the  $i$ th zero of the Hermite polynomial having degree  $k$ ,  $H_k(x)$ , and the weight  $w_i$  depends on the number  $k$  of nodes and the Hermite polynomial  $H_{k-1}(x)$  evaluated at  $x_i$ . Tables for the nodes and weights can be found in Stroud & Secrest (1966) and Abramowitz & Stegun (1972). As long as  $g(x)$  is a polynomial of maximal degree  $2k - 1$ , the sum (A.4.3) delivers the exact value of (A.4.2). That means approximation (A.4.3) becomes arbitrarily accurate when the number  $k$  of nodes is increased.

Let us consider now, more generally, the function

$$f(x) = w_N(x; \mu, \sigma^2)g(x). \quad (\text{A.4.4})$$

The simple substitution  $x = \sqrt{2}\sigma z + \mu$  yields the identity

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{+\infty} \pi^{-1/2} \exp(-z^2)g(\sqrt{2}\sigma z + \mu)dz,$$

which can be solved by Gauss-Hermite integration in the following way:

$$\int_{-\infty}^{+\infty} f(x)dx \approx \sum_{i=1}^k v_i g(\sqrt{2}\sigma x_i + \mu), \quad (\text{A.4.5})$$

where  $v_i = \pi^{-1/2}w_i$  is the transformed weight and  $x_i$  is the tabulated  $i$ th zero of the Hermite polynomial. Thus, the integral of a function is well approximated by Gauss-Hermite if  $f(x)$  has the form (A.4.4), where  $g(x)$  stands for a polynomial in  $x$ .

Within this book most integration problems can be traced back to the Bayesian paradigm that can be characterized in the following way (see, e.g., Naylor & Smith, 1982; Smith et al., 1985; Shaw, 1988): Suppose we have data  $y$  together with a probability model  $f(y|\theta)$  indexed by a parameter  $\theta \in \mathbb{R}$  with prior density  $q(\theta)$ . Suppose further that the resulting posterior density

$$p(\theta|y) = \frac{f(y|\theta)q(\theta)}{\int_{\mathbb{R}} f(y|\theta)q(\theta)d\theta} \quad (\text{A.4.6})$$

is analytically intractable, since the integral that also determines the marginal density of  $y$  cannot be solved analytically. As a consequence no closed-form solutions are available for the first two posterior moments

$$E(\theta^r|y) = \int_{\mathbb{R}} \theta^r p(\theta|y)d\theta, \quad r = 1, 2, \quad (\text{A.4.7})$$

which are assumed to exist. For the application of the Gauss-Hermite rule to (A.4.6) and (A.4.7), we consider in the following the integral function

$$S(a(\theta)) = \int_{\mathbb{R}} a(\theta) f(y|\theta) q(\theta) d\theta$$

with  $a(\theta) = 1, \theta$ , or  $\theta^2$ .  $S(a(\theta))$  covers all interesting measures. For example,  $S(1)$  corresponds to the marginal density of  $y$  and  $S(\theta)/S(1)$  denotes the posterior mean.

Application of the Gauss-Hermite integration rule to  $S(a(\theta))$  is straightforward when the prior  $q(\theta)$  corresponds to a normal density with known mean  $\mu$  and known variance  $\sigma^2$ . Since  $a(\theta)$  is a polynomial of degree at most 2, the numerical accuracy of the approximation depends only on the polynomial degree of the likelihood  $f(y|\theta)$  and the number of nodes used.

If the prior  $q(\theta)$  is non-normal, iterative use of the Gauss-Hermite rule is recommended by various authors (see, e.g., Naylor & Smith, 1982; Smith et al., 1985). The procedure is based on the Bayesian analog of the central limit theorem, indicating that the posterior may be well approximated by a normal density multiplied by a polynomial in  $\theta$ . Expanding the posterior kernel by such a normal component with mean  $\mu$  and variance  $\sigma^2$ , say  $w_N(\theta; \mu, \sigma^2)$ , yields

$$S(a(\theta)) = \int_{\mathbb{R}} a(\theta) h(\theta) w_N(\theta; \mu, \sigma^2) d\theta$$

with  $h(\theta) = f(y|\theta)q(\theta)/w_N(\theta; \mu, \sigma^2)$ . Given  $\mu$  and  $\sigma^2$ , the Gauss-Hermite rule (A.4.5) could be applied, provided  $a(\theta)h(\theta)$  is at least approximately a polynomial in  $\theta$ . In most practical situations, however,  $\mu$  and  $\sigma^2$  are unknown. A crude, but sometimes satisfactory, procedure is to replace  $\mu$  by the posterior mode and  $\sigma^2$  by the posterior curvature. Both are easily estimated by maximizing the log-posterior

$$\ell(\theta) = \log f(y|\theta) + \log q(\theta)$$

with respect to  $\theta$ . The procedure can be improved by the following iterative process:

1. Choose initial values  $\mu_0$  and  $\sigma_0^2$  for  $\mu$  and  $\sigma^2$ .
2. For some chosen  $k$ , approximate  $S(\theta)$  by

$$S_p(a(\theta)) = \sum_{i=1}^k a(\theta_{p,i}) h_p(\theta_{p,i}) v_i,$$

with nodes  $\theta_{p,i} = \sqrt{2}\sigma_{p-1}x_i + \mu_{p-1}$ , weights  $v_i = \pi^{-1/2}w_i$ , and the function  $h_p(\theta) = f(y|\theta)q(\theta)/w_N(\theta; \mu_{p-1}, \sigma_{p-1}^2)$ .

3. Calculate updates of  $\mu$  and  $\sigma^2$  according to

$$\mu_p = S_p(\theta)/S_p(1) \quad \text{and} \quad \sigma_p^2 = S_p(\theta^2)/S_p(1) - \mu_p^2.$$

4. Repeat steps 2 and 3 as long as the changes in updated moments are not small enough.

This procedure gives approximations for the posterior mean and the posterior variance at each cycle  $p$ . These approximations are used to construct the nodes  $\theta_{p,i}$  and the function  $h_p$  for the next cycle. With an increasing number of cycles, the number  $k$  of nodes should be successively increased to obtain stable values for  $\mu$  and  $\sigma^2$  at each size of  $k$ .

### Multivariate Gauss-Hermite Integration

Let us now consider the case of  $m$ -dimensional  $x = (x_1, \dots, x_m)$  with a regular function  $f(x) : \mathbb{R}^m \rightarrow \mathbb{R}$ , weight function  $w(x) : \mathbb{R}^m \rightarrow \mathbb{R}_+$ , and another function  $g(x) : \mathbb{R}^m \rightarrow \mathbb{R}$ , so that  $f(x) = w(x)g(x)$ . A Cartesian product rule based on (A.4.3) may be applied to approximate the  $m$ -dimensional integral

$$\int_{\mathbb{R}^m} f(x) dx = \int_{\mathbb{R}} \dots \int_{\mathbb{R}} w(x_1, \dots, x_m) g(x_1, \dots, x_m) dx_1 \dots dx_m,$$

provided that the weight function is given by

$$w(x) = \exp(-x'x) = \exp(-x_1^2) \cdot \dots \cdot \exp(-x_m^2).$$

Then the univariate Gauss-Hermite rule applies to each of the components of  $x$  in turn. Using  $k_r$  nodes in the  $r$ th dimension,  $r = 1, \dots, m$ , yields the approximation

$$\int_{\mathbb{R}^m} f(x) dx \approx \sum_{i_1=1}^{k_1} w_{i_1}^{(1)} \dots \sum_{i_m=1}^{k_m} w_{i_m}^{(m)} g\left(x_{i_1}^{(1)}, \dots, x_{i_m}^{(m)}\right), \quad (\text{A.4.8})$$

where  $x_{i_r}^{(r)}$  is the  $i_r$ th zero of the Hermite polynomial with degree  $k_r$  and  $w_{i_r}^{(r)}$  is the corresponding weight. The Cartesian product rule (A.4.8) has  $k = \prod_{r=1}^m k_r$  nodes  $x_i = (x_{i_1}^{(1)}, \dots, x_{i_m}^{(m)})$ ,  $i = (i_1, \dots, i_m)$ , and is exact as long as  $g$  is a polynomial containing terms up to  $x_r^{2k_r-1}$  for each dimension  $r = 1, \dots, m$ . Unfortunately, the number  $k$  of nodes increases exponentially with the number  $m$  of dimensions. So Cartesian product rules are less appropriate to approximate high-dimensional integrals. In practice Cartesian product rules work efficiently up to 5- or 6-dimensional integrals.

If we consider, more generally, the multivariate extensions of the Bayesian paradigm that is based on the posterior density

$$p(\theta|y) = \frac{f(y|\theta)q(\theta)}{\int f(y|\theta)q(\theta)d\theta}$$

with parameter  $\theta = (\theta_1, \dots, \theta_m) \in \mathbb{R}^m$ , data  $y = (y_1, \dots, y_m)$ , likelihood  $f(y|\theta)$ , and prior density  $q(\theta)$ , all integration problems can be traced back to

$$S(a(\theta)) = \int_{\mathbb{R}^m} a(\theta) f(y|\theta) q(\theta) d\theta. \quad (\text{A.4.9})$$

For example, posterior mean and covariance are given by

$$E(\theta|y) = S(\theta)/S(1) \quad \text{and} \quad \text{cov}(\theta|y) = S(\theta\theta')/S(1) - E(\theta|y)E(\theta|y)'.$$

To apply the Cartesian Gauss-Hermite product rule to  $S(a(\theta))$  we assume, in analogy to the univariate case, that the posterior density can be well approximated by the product of a multivariate normal density and a polynomial in  $\theta$ . Expansion of the posterior kernel by a multivariate normal density  $w_N(\theta; \mu, \Sigma)$  with mean  $\mu = (\mu_1, \dots, \mu_m)$  and variance-covariance matrix  $\Sigma$  yields

$$S(a(\theta)) = \int_{\mathbb{R}^m} a(\theta) h(\theta) w_N(\theta; \mu, \Sigma) d\theta.$$

The substitution

$$\theta = \sqrt{2} \Sigma^{1/2} z + \mu$$

with the left Cholesky square root  $\Sigma^{1/2}$  has the desirable property that the components of  $z$  are nearly orthogonal and that

$$S(a(\theta)) = \int_{\mathbb{R}^m} a(z) h(z) \pi^{-m/2} \exp(-z'z) dz$$

has a weight function that is required to use the Cartesian product rule of the Gauss-Hermite type.

Then (A.4.9) can be approximated by an extension of the iterative integration scheme that was proposed for the univariate case. The steps are as follows:

1. Choose initial values  $\mu_0$  and  $\Sigma_0$  for  $\mu$  and  $\Sigma$ , and set  $p = 1$ .
2. For some chosen  $k_1, \dots, k_m$ , approximate  $S(a(\theta))$  by

$$S_p(a(\theta)) = \sum_{i_1=1}^{k_1} v_{i_1}^{(1)} \cdots \sum_{i_m=1}^{k_m} v_{i_m}^{(m)} a(\theta_{p,i}) h_p(\theta_{p,i})$$

where for the multiple index  $i = (i_1, \dots, i_m)$ , the nodes are given by

$$\theta_{p,i} = \sqrt{2} \Sigma_{p-1}^{1/2} z_i + \mu_{p-1}$$

with  $z_i = (z_{i_1}^{(1)}, \dots, z_{i_m}^{(m)})$ ,  $z_{i_r}^{(r)}$ ,  $r = 1, \dots, m$ , denoting the tabled nodes of univariate Gauss-Hermite integration of order  $k_r$ . The corresponding weights are given by  $v_{i_r}^{(r)} = \pi^{-1/2} w_{i_r}^{(r)}$  and  $h_p(\theta) = f(y|\theta)q(\theta)/w_N(\theta; \mu_{p-1}, \Sigma_{p-1})$ .

3. Calculate updates of  $\mu$  and  $\Sigma$  according to  $\mu_p = S_p(\theta)/S_p(1)$  and  $\Sigma_p = S_p(\theta\theta')/S_p(1) - \mu_p\mu_p'$ .
4. Repeat steps 2 and 3 and set  $p = p + 1$  as long as the changes of the updated posterior moments are not small enough.

Examples for the efficiency of the above scheme can be found in Naylor & Smith (1982) and Smith et al. (1985), among others. For a prior density  $q(\theta)$  with known mean  $E(\theta)$  and known variance-covariance matrix  $\text{cov}(\theta)$ , Hennevogl (1991) recommends using

$$\mu_0 = E(\theta) \quad \text{and} \quad \Sigma_0 = \text{cov}(\theta)$$

as starting values. Alternatively, posterior mode and curvature, which are easy and fast to compute, may be used.

## A.5 Monte Carlo Methods

### Importance Sampling

The simplest Monte Carlo method for computing integrals of the form

$$I = \int_{-\infty}^{+\infty} g(x)f(x)dx,$$

where  $g$  is a continuous function and  $f$  is a density, consists of approximating  $I$  by the arithmetic mean

$$\hat{I} = \frac{1}{m} \sum_{j=1}^m g(x_j),$$

where  $x_j, j = 1, \dots, m$ , are i.i.d. drawings from the density  $f(x)$ . A better and more efficient approximation of  $I$  can be obtained by importance sampling: Rewriting  $I$  as

$$I = \int_{-\infty}^{+\infty} g(x) \frac{f(x)}{\phi(x)} \phi(x) dx,$$

with the density  $\phi(x)$  as the importance function, one may also approximate  $I$  by

$$\bar{I} = \frac{1}{m} \sum_{j=1}^m g(x_j) \frac{f(x_j)}{\phi(x_j)},$$

where the  $x_i$ 's are now drawn from  $\phi(x)$ . It can be shown (Ripley, 1987) that this approximation by importance sampling becomes quite good if  $g(x)f(x)/\phi(x)$  is nearly constant. An optimal choice would be

$$\phi(x) \propto |g(x)|f(x);$$

however,  $\phi(x)$  should also allow fast and simple random number generation. In a Bayesian context, integrals are often of the form

$$I = \int a(x) f(y|x) f(x) dx,$$

where  $f(y|x)f(x)$  is proportional to the posterior density  $f(x|y)$ . Since this posterior becomes approximately normal for larger numbers of observations  $y$ , the importance sampling function is often chosen as a (multivariate) normal  $N(\mu, \Sigma)$ , where  $\mu$  and  $\Sigma$  are matched to the posterior mean and covariance of  $f(x|y)$  or some approximation, such as the mode and curvature of  $f(x|y)$ .

## Rejection Sampling

Rejection sampling is a technique for generating random numbers from a density  $f(x)$  without drawing directly from  $f(x)$  itself. A typical application arises in Bayesian inference where  $f(x)$  itself is not available but only an un-normalized “density”  $g(x)$  (often the numerator in Bayes’ theorem) is available:

$$g(x) \propto f(x).$$

If there are a density  $h(x)$  available for drawing random numbers and a constant  $C$  such that

$$g(x) \leq C \cdot h(x)$$

for all  $x$ , then the following rejection algorithm generates random numbers from  $f(x)$ :

1. Generate a random number  $x^*$  from  $h(x)$ .
2. Accept  $x^*$  as a random number from  $f(x)$  with probability

$$\pi(x^*) = \frac{g(x^*)}{Ch(x^*)}.$$

Step 2 is usually implemented by drawing  $u$  from the uniform distribution on  $[0, 1]$ , and  $x^*$  is accepted if  $u \leq \pi(x^*)$  and rejected otherwise. Repeated application generates a random sample from  $f(x)$ . In order not to waste too many drawings from  $h(x)$ , the “envelope” function  $C \cdot h(x)$  should be as close as possible to  $f(x)$ . More details, such as formal proofs or choice of the envelope function, can be found in Devroye (1986) and Ripley (1987).

### Gibbs Sampling and Markov Chain Monte Carlo (MCMC)

MCMC techniques have revolutionized Bayesian inference. Bayesian inference starts with a prior distribution  $f(\theta)$  for an unknown parameter vector  $\theta$ . Given observed data  $Y$ , the posterior  $p(\theta|Y)$  is determined by Bayes’ theorem,

$$p(\theta|Y) = \frac{f(Y|\theta)f(\theta)}{\int f(Y|\theta)f(\theta)d\theta} \propto f(Y|\theta)f(\theta),$$

where  $f(Y|\theta)$  is the likelihood. For a high dimension of  $\theta$ , numerical integration or classical Monte Carlo methods, such as importance or rejection sampling, are often computationally infeasible. Here MCMC methods are more appropriate.

In the following, let  $p(\theta)$  denote the posterior distribution of  $\theta$ , suppressing the conditioning on the data  $Y$  notationally. Most MCMC methods split  $\theta$  into components  $\theta = (\theta'_1, \dots, \theta'_T, \dots, \theta'_H)'$  of subvectors of possibly differing dimension. The posterior  $p(\theta)$ , typically high-dimensional and rather complicated, is not needed; only so called *full conditionals*

$$p(\theta_T|\theta_{-T}), \quad T = 1, \dots, H,$$

of subvectors given the remaining components

$$\theta_{-T} = (\theta'_1, \dots, \theta'_{T-1}, \theta'_{T+1}, \dots, \theta'_H)'$$

are needed.

The *Gibbs sampler*, introduced by Geman & Geman (1984) and probably the most prominent member of MCMC algorithms, iteratively updates all components  $\theta_T$  by drawing samples from their full conditionals. The Gibbs sampler consists of the following updating scheme.

- (i) Choose values  $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_T^{(0)}, \dots, \theta_H^{(0)})$ .

(ii) For  $k = 1, 2, \dots$ :

- draw  $\theta_1^{(k)}$  from  $p(\theta_1 | \theta_{-1}^{(k)})$ , ...,
- draw  $\theta_T^{(k)}$  from  $p(\theta_T | \theta_{-T}^{(k)})$ , ...,
- draw  $\theta_H^{(k)}$  from  $p(\theta_H | \theta_{-H}^{(k)})$ .

Markov chain theory says that under mild conditions  $\theta^{(k)}$  converges in distribution to  $\theta$  for  $k \rightarrow \infty$  (Geman & Geman, 1984; Tierney, 1994). Repeating the updating steps often enough, joint and marginal densities, their moments, and other characteristics may be estimated from this sample, e.g., by density smoothers or empirical moments.

In practice, of course, one also has to choose a termination criterion for stopping the Gibbs iterations for some finite  $k$ . For this question and other details we refer the reader to Gelfand & Smith (1990), Gilks, Richardson & Spiegelhalter (1996, Ch. 8), Gelman, Carlin, Stern & Rubin (1995, Ch. 11), and Gamerman (1997b, Ch. 5), among others.

Gibbs sampling is appropriate if drawings from the full conditionals are computationally cheap, for example if the full conditionals are standard distributions like (multivariate) normals, (inverse) gamma, etc. Often, however, the full conditionals are not easily available, in particular the normalizing constant is often unknown. For scalar or low-dimensional components of  $\theta$ , Gibbs sampling can be combined with the ideas of rejection sampling to *adaptive rejection sampling* (Gilks & Wild, 1992).

General MCMC techniques with Metropolis-Hastings (MH) updating are often efficient tools in this situation. MH steps are typically easier to implement and often make an MCMC algorithm more efficient in terms of CPU time. An MH step proposes a new value for a given component and accepts it with a certain probability. A Gibbs step turns out to be a special case where the proposal is always accepted.

Let  $p(\theta_T | \theta_{-T})$  be the full conditional of a component  $\theta_T$  of  $\theta$ , given the rest of the components, denoted by  $\theta_{-T}$ . To update  $\theta_T = \theta_T^{(k)}$  in iteration step  $k$ , it is sufficient to generate a proposal  $\theta'_T$  from an arbitrarily chosen transition kernel  $P(\theta_T \rightarrow \theta'_T; \theta_{-T})$  and accept the generated proposal with probability

$$\delta = \min \left\{ 1, \frac{p(\theta'_T | \theta_{-T}) P(\theta'_T \rightarrow \theta_T; \theta_{-T})}{p(\theta_T | \theta_{-T}) P(\theta_T \rightarrow \theta'_T; \theta_{-T})} \right\};$$

otherwise, leave  $\theta_T$  unchanged. This is the Hastings algorithm used for updating full conditionals. Only a ratio of the full conditional of  $\theta_T$  enters in  $\delta$ , so  $p(\theta_T | \theta_{-T})$  needs to be known only up to a multiplicative constant and does not need to be normalized, a very convenient fact for implementation. Note that both the current state  $\theta_T$  and the proposed new state  $\theta'_T$ , as well as the current states of the other components  $\theta_{-T}$ , affect  $\delta$ .

Gibbs sampling corresponds to the specific choice

$$P(\theta_T \rightarrow \theta'_T; \theta_{-T}) = p(\theta'_T | \theta_{-T})$$

so that  $\delta$  becomes 1 and therefore all proposals are accepted. Here the current state of  $\theta_T$  does not affect the new one,  $\theta'_T$ .

There is a great flexibility in the choice of the transition kernel  $P$ . Common choices are random walk Metropolis proposals and (conditional) independence proposals (Tierney, 1994).

Random walk Metropolis proposals are generated from a distribution, that is symmetric about the current value  $\theta_T$ . Gaussian or rectangular distributions are often used. In contrast, conditional independence proposals do not depend on the current state of  $\theta_T$ ; they may depend, however, on the current values of  $\theta_{-T}$ . As we have seen, the Gibbs sampling kernel is a specific conditional independence proposal. However, it is crucial that for a chosen transition kernel  $P$ , the acceptance probability  $\delta$  not be too small (in average) and that both convergence and mixing behavior of the whole simulated Markov chain be satisfactory. A well-mixing Markov chain is moving rapidly throughout the support of the stationary distribution  $p(\theta)$ .

Somewhat surprising is the fact that one is allowed to use hybrid procedures – that is, to use different versions of Hastings proposals for updating different components of  $\theta$ . One strategy is to sample from the full conditionals that is a “Gibbs step,” as long as this is easy and fast. If not, a specific Hastings step with a simple proposal distribution mostly works in CPU time. As long as all components are updated in a deterministic or even random order (which may ensure better mixing of the chain), the chain converges to the posterior. Assessment of convergence using diagnostic measures is described in Cowles & Carlin (1996) and Mengersen, Robert & Guihenneuc-Jouyaux (1998).

For more details, we refer to the literature already cited and to Besag, Green, Higdon & Mengersen (1995).

# Appendix B

## Software for Fitting Generalized Linear Models and Extensions

This section informs the reader about software that can be used for fitting GLMs and extensions. It includes not only software specialized for GLMs, but also well-known, general-purpose statistical packages and programming environments with general statistical features. The following list gives a selection of packages that will be described. This selection is based on a subjective choice and does not give a complete overview of all packages available. Further information can be obtained from the distributors, from manuals that come with package and, last but not least, from the Internet.

### **General-purpose statistical packages**

SAS: statistical package for all fields of applications

SPSS/PC+: statistical package with special features for social sciences

BMDP: statistical package with special features for biometrical applications

GENSTAT: statistical package for all fields of applications

STATA: statistical package with special features for econometric applications

### **Programming environments with statistical features**

GLIM: programming environment for fitting GLMs

S-Plus: programming environment with many built-in functions for statistical analysis (one of which is for fitting GLMs)

GAUSS: programming language with add-on modules for fitting GLMs

XPLORE: programming environment for exploratory regression and data analysis

### **Specialized packages**

EGRET: package designed for epidemiological applications

LIMDEP: package designed for econometrical applications

BUGS: package designed for Bayesian inference

BayesX: package designed for Bayesian inference

## SAS

SAS (Statistical Analysis System) is a statistical package with a particularly wide range of facilities for data management and data analysis. The SAS base system includes tools for data management, report writing, and descriptive statistics and is rather limited in terms of statistical analysis. Other SAS software products have to be integrated to provide one total analysis system. The most important is the SAS/STAT software, which includes a powerful set of statistical analysis procedures. Other software products include SAS/GRAph for generating high-quality graphics; SAS/IML, which provides an interactive matrix programming language for writing procedures in SAS; SAS/ETS for econometrical applications; and SAS/QC for statistical methods in quality control.

*Metrical* responses can be modelled within SAS/STAT using single or multiple linear regression, nonlinear regression, and general linear models that may include random effects. Several methods are available for variable selection. A new procedure MIXED can fit mixed linear models, i.e., general linear models that include both fixed effects and random effects. It is part of the SAS/STAT software in release 6.07 and later versions. Survival models in SAS/STAT include linear location-scale models (see Section 9.1.2) where the distribution of the noise can be taken from a class of distributions that includes the extreme-value, logistic, exponential, Weibull, log-normal, log-logistic, and gamma distributions. The data may be left-, right-, or interval-censored.

*Categorical* responses can be fitted using several procedures of SAS/STAT. These procedures are CATMOD, PROBIT, LOGISTIC, and GENMOD.

CATMOD is a procedure for CAtegorical data MODelling. It includes logit models for multinomial responses and cumulative logit response functions to take into account ordered categories. Additionally, marginal probabilities can be specified as response functions and simple models for repeated measurements can be fitted using a “repeated” statement. It is possible to define response functions by using transformations and matrix operations within the “response” statement or by reading them directly together with their covariance matrix from a data set.

PROBIT can fit models for binary and multinomial responses. Probit, logit, and complementary log-log links are available. There is an option to take overdispersion into account. Deviances are not given along with the standard output, but they can be easily calculated from the quantities given.

LOGISTIC is a new procedure available in SAS/STAT version 6.04 and later. Cumulative logistic models for ordered responses can be fitted. The procedure includes facilities for automatic variable selection and a good range of statistics for model checking.

In Release 6.12 the GENMOD procedure in SAS/STAT introduced the capacity for Generalized Estimation Equations (Sections 3.5.2, 6.2.2). The

GEE analysis is implemented with a REPEATED statement in which the working correlation matrix is specified. Available correlation structures are autoregressive (1), exchangeable, independent and unstructured, among others. Since version 8, GEE is available also for ordinal data based on cumulative-type models.

From version 6.08 a new module "SAS/INSIGHT" based on Windows will include basic generalized models for metrical, binary, and count data. A new procedure GENMOD for fitting generalized linear models is planned and will be included as an experimental version in SAS version 6.08.

SAS Institute Inc.  
Cary, North Carolina 27512-8000, USA

SAS Software Ltd.  
Wittoning House  
Henley Road, Medmenham  
Marlow SL7 2EB, Bucks, U.K.  
Information: <http://www.sas.com>

## SPSS

SPSS (Statistical Package for Social Scientists) is a statistical package with a wide range of statistical directives especially designed to meet the needs of social scientists. The base system can be extended by add-on modules. SPSS can perform many analyses commonly used in social science. However, there is no programming language included to allow for users' own programs.

The "Regression Models" module includes parametric binary and multinomial logistic regression. The procedure offers stepwise methods for variable selection. Binary logit or probit models and logistic models for ordinal outcomes can be analyzed with procedures that are part of the "Advanced Models" module. This module also has capabilities for parametric generalized linear models and for Cox regression.

However, none of the more advanced approaches beyond parametric GLMs is implemented in the current version.

SPSS Inc.  
444 North Michigan Avenue  
Chicago, Illinois 60611, USA

Information: <http://www.spss.com>

## BMDP

BMDP (BioMeDical Package) is a general-purpose statistical package that consists of a collection of independent programs for statistical analysis, some of which was particularly designed for the needs of biological and medical researchers. It is accompanied by a data manager, but does not include a programming language. The graphic facilities are limited.

*Metrical* response modelling includes multivariate regression, Ridge-regression, nonlinear and polynomial regression. Variable selection is available. Time series can be analyzed using ARIMA models and spectral analysis. Survival models include Cox's proportional hazards model and log-linear models with error-term distributions such as Weibull, exponential, log-normal, and log-logistic. Variable selection is available for all models.

The program LR (Logistic Regression) fits logit models for *binary* responses. LR can be used to either fit a specified model or perform variable selection. It provides a good range of diagnostics. Deviances are given and predicted values easily obtained. LR cannot take overdispersion into account.

*Multicategorical* responses can be modelled using the program PR (Poly-chotomous Logistic Regression). The program includes the logistic model for multinomial responses and the cumulative logit model for ordered responses. The features of PR are similar to LR. In particular, variable selection is also available.

LE is a program for estimating the parameters that maximize likelihood functions and may be used to fit models. LE uses an iterative Newton-Raphson algorithm where the (log) density and initial values for the parameters must be provided. The manual illustrates LE with a logit model for binary responses and a cumulative logistic model for ordered responses.

BMDP Statistical Software Inc.  
1440 Sepulveda Blvd.  
Los Angeles, California 90025, USA

BMDP Statistical Software Ltd.  
Cork Technology Park  
Cork, Ireland

## GENSTAT

GENSTAT 5 (GENeral STATistical package) is a general-purpose statistical package that supplies a wide range of standard directives, but also includes a flexible command language enabling the user to write programs. FORTRAN

programs may also be included via the OWN directive. GENSTAT has a wide range of data-handling facilities and provides high-resolution graphics.

*Metrical* responses can be modelled based on simple or multiple linear regression, standard nonlinear curves, or general nonlinear regression. Time series can be analyzed using models such as AR and ARIMA.

The following distributions are included for fitting *generalized linear models*: binomial, Poisson, normal, gamma, inverse normal. Implemented link functions are identity, log, logit, reciprocal, power, square root, and complementary log-log.

GENSTAT provides a good range of facilities for model checking. Deviances are given along with the standard output. Overdispersion can be modelled. A separate command yields predicted values along with their standard errors.

There is a link from GENSTAT to GLIM, since GLIM macros can be translated into GENSTAT procedures (and vice versa).

Numerical Algorithms Group Inc.  
1400 Opus Place, Suite 200  
Downers Grove, Illinois 60515-5702, USA

Numerical Algorithms Group Ltd.  
Wilkinson House  
Jordan Hill Road  
Oxford OX2 8DR, U.K.

Information: <http://www.nag.co.uk>

## STATA

STATA is a statistical package with a very broad range of statistical, graphical, and data-management capabilities and is also fully programmable. It is useful for all disciplines, but, as a particular feature, a lot of inferential procedures are implemented that are very popular among econometricians. It also includes many models within the GLM framework and its extensions that other packages do not provide. Among others, it includes a variety of count data models beyond the usual Poisson family, GEE methods for panel data and various random effects models; see the list below. On the other side, the current version does not allow one to fit non- or semiparametric generalized regression models. The following list provides a selection of inferential tools for GLMs and related models.

- Generalized linear models: The allowable distributional families are Gaussian, inverse Gaussian, gamma, binomial, Poisson, and negative binomial. They can be combined with a variety of link functions, for

example, logit, probit, complementary log-log, and power link functions for binomial responses. For categorical response, the multinomial and the cumulative logit model, as well as the ordered probit models are available.

- Count data models: Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial.
- GEE methods: Estimation is based on the original GEE1 approach of Liang & Zeger (1986). Allowable distribution families are Gaussian, binomial, Poisson, and gamma, combined with appropriate link functions. All working correlations described in Chapters 3 and 6 are implemented.
- Random effects models: random effects probit, logistic, complementary log-log and Poisson regression, gamma random effects Poisson and beta random effects negative binomial regression.

Information: <http://www.stata.com>

## GLIM

GLIM4 (Generalized Linear Interactive Modelling) is an interpretive language especially designed for fitting generalized linear models. It was developed at the Centre for Applied Statistics, University of Lancaster, U.K. GLIM4 includes comprehensive facilities for manipulating and displaying data. The interpretive character of the language is adjusted to the needs of iterative statistical model building. GLIM4 supports developing, fitting, and checking statistical models.

GLIM4 is essentially designed for statistical modelling with univariate responses. The following exponential families are included as standard options in GLIM4: normal, binomial, gamma, inverse normal, and Poisson. Additionally, it is possible to specify distributions. The multinomial distribution is included, because it can easily be modelled using a relationship between the Poisson and the multinomial distribution. This “Poisson trick” is described, e.g., in Aitkin et al. (1989) or in the GLIM4 manual. GLIM4 includes various link functions as standard options, but again it is possible to specify link functions.

GLIM4 is extensible in many respects. The user can write macros using the GLIM4 language or include FORTRAN routines. A macro library comes with the package. This library contains, e.g., macros to fit generalized additive models or others to fit survival models with distributions such as Weibull, extreme-value, or (log) logistic. Further macros published in the GLIM4 newsletters are available on an archive server “statlib” which can be reached via electronic mail (address: [statlib@lib.stat.cmu.edu](mailto:statlib@lib.stat.cmu.edu)) or FTP (ftp to lib.stat.cmu.edu (128.2.241.142) and login with user name statlib).

The user has access to the system settings of GLIM4 and most of the data structures derived in model fitting and calculations. GLIM4 uses a

weighted-least-squares algorithm for parameter estimation. It allows user-defined macros to be called at various stages of the fitting algorithm. Additionally, any structure available to the user can be modified during the fit. This allows the user to modify the iteratively reweighted-least-squares algorithm to perform special actions for specialized and nonstandard statistical models.

NAG Inc.  
1400 Opus Place, Suite 200  
Downers Grove, Illinois 60515-5702, USA

NAG Ltd.  
Wilkinson House  
Jordan Hill Road  
Oxford OX2 8DR, U.K.  
Information: <http://www.nag.co.uk>

## S-Plus

The S-PLUS programming environment includes a wide range of built-in functions for statistical analysis. Most of these functions can be easily extended using the S-PLUS programming language. Additionally, S-PLUS provides an interface to FORTRAN and C, thus increasing the extensibility of S-PLUS. S-PLUS has very flexible graphic facilities.

For *metrical* responses the functions available include multiple regression models, nonlinear models, local regression models, tree-based models, and regression models using kernel smoothers and other smoothing techniques. Survival times can be fitted using the Cox model or a counting process extension to the Cox model. Time series data can be analyzed using AR models, ARIMA models, or spectral analysis.

*Generalized linear models* can be fitted using the function `glm()`. Models can be fitted for the Gaussian, binomial, Poisson, gamma and inverse Gaussian distributions. The family-statement in `glm()` also includes an option “quasi” for the fitting of quasi-likelihood models. Standard links available are identity, log, logit, probit, sqrt, inverse, and log-log. The function `step.glm()` allows stepwise variable selection. The `glm()`-code is available on the archive server statlib, allowing the user to extend the function to nonstandard likelihoods. Simpler ways of extending the capabilities of the `glm()` function and examples for its use are given in Chambers & Hastie (1992). That book also describes the use of the function `gam()`, which fits generalized additive models for the models listed above, using smoothing splines or local regression.

S-PLUS provides a very good range of diagnostics for model checking. Deviances and fitted values are calculated automatically.

There are many S-functions written by researchers that extend the facilities of S-PLUS considerably. Some of these functions are available on the statlib server. Among those are the function *logist*, which fits ordinal logistic regression models, the function *net()*, which fits feedforward neural nets, and the function *fda()*, which fits flexible discriminant functions and provides access to a MARS version.

Statistical Sciences  
52 Sandfield Road  
Headington  
Oxford OX3 7RJ, U.K.

Information <http://www.splus.mathsoft.com>

## GAUSS

GAUSS is a matrix programming language including standard matrix operations, but also a wide range of statistical and data-handling procedures. GAUSS provides good graphic facilities. There are interfaces to FORTRAN, C, and PASCAL.

There is an add-on module “Quantal Response,” which includes routines to fit the following *generalized linear models*: binomial probit, multinomial logit, ordered cumulative logit, and log-Poisson models. Deviances and predicted values are calculated automatically. Other measures for model checking are easy to compute but not included as a standard.

Further facilities for modelling *count* and *duration* data are available in the module “Count.” It includes models for count responses such as Poisson, negative binomial, hurdle Poisson, and seemingly unrelated Poisson regression models. Models for duration data included are exponential, exponential-gamma, and Pareto duration models. All models (count and duration) allow truncation or censoring to be taken into account.

GAUSS may be used in general to compute sophisticated models that are not available in standard statistical packages. The module “Maximum-Likelihood,” which includes seven algorithms to maximize user-defined log-likelihoods and may reduce the effort in programming models.

Aptech Systems Inc.  
23804 S.E. Kent-Kangley Road  
Maple Valley, Washington 98038, USA

## Xplore

XploRe is a statistical computing environment, including a great variety of non- and semiparametric methods, interactive graphics for visual presentation and exploration, and an object-oriented programming language.

Metric responses can be analyzed using a wide range of modern regression techniques such as kernel estimation, spline smoothing, ACE, projection pursuit regression, wavelets, and neural networks.

A wide range of methods for generalized linear models and semiparametric extensions is stored in several libraries. They include most of the common univariate generalized linear models, mult categorial response models, generalized partial linear and additive models. The GAM library offers various methods for fitting and testing, such as Nadaraya-Watson, local linear and quadratic backfitting smoothers, integration-based estimators, possibilities for estimating and testing models with interaction terms, and estimates of derivatives of regression functions.

Information:

XploRe Systems

W. Härdle

Institut für Statistik und Ökonometrie

FB Wirtschaftswissenschaften

Humboldt Universität zu Berlin

Spandauer Str. 1 D-10178 Berlin, Germany

<http://www.xplore-stat.de>

## EGRET

EGRET (Epidemiological, Graphics, Estimation, and Testing program) is a package specializing in the analysis of epidemiological studies and survival models. There are a module for data definition and a module for data analysis. The facilities for data management and data manipulation within the package are limited.

EGRET includes logit models for binary responses, conditional logit models for modelling data from matched case-control studies (see Collett, 1991, Section 7.7.1), logit models with random effects and beta-binomial regression. Survival models include models with exponential and Weibull noise distribution and Cox's proportional hazards model as well as Cox regression with time-dependent covariates. Poisson regression can be used for modelling cohort data. Anticipated upgrades of EGRET, which may by now be available, will include logit models for mult categorial data, logit models with nested random effects, Poisson regression models with random effects,

and random effects regression models with asymmetric mixing distributions.

Statistics & Epidemiology Research Corporation  
909 Northeast 43rd Street, Suite 310  
Seattle, Washington 98105, USA

## LIMDEP

LIMDEP (LIMited DEPendent Variables) is a package originally designed for applications in econometrics. Most of the procedures included are designed for metrical regression and time series analysis. However, there are also many generalized linear models for categorical responses that can be fitted within LIMDEP. The package also includes features for nonlinear optimization, evaluation of integrals, and minimizing nonlinear functions.

*Metrical* responses can be fitted using linear models with fixed and random effects. Further models are Tobit models for censored responses, stochastic frontier models, and seemingly unrelated regression models. Time series can be analyzed using AR, ARIMA and ARMAX models. Survival models include models with Weibull, exponential, normal, logistic and gamma noise distributions, but also Cox's proportional hazards model. The data may be censored in various ways.

Important generalized linear models for *categorical* responses included in LIMDEP are the logit model for binomial and multinomial responses and the cumulative logit model for ordered multicategorical responses. For binary data a probit model can be fitted that can include random effects. Additionally, there is a feature to fit a bivariate probit. Nonparametric models for binary data are also available.

*Count* data can be fitted using Poisson or negative binomial models. Censoring or truncation of the data can be taken into account.

Econometric Software Inc.  
46 Maple Avenue  
Bellport, New York 11713, USA

## BUGS

BUGS (Bayesian Inference Using Gibbs Sampling) is a program for the Bayesian analysis of complex statistical models using Markov Chain Monte Carlo methods. It is developed jointly by the MRC Biostatistics Unit, Cambridge, and the Imperial College School of Medicine, London.

Typical applications include generalized linear mixed models with hierarchical, temporal, and spatial random effects; latent variable models; frailty models; models with measurement error, and missing data problems. A good overview of statistical problems that BUGS can solve is provided by a collection of examples obtainable from the web site.

The “classic” BUGS version generally uses univariate Gibbs sampling and a simple Metropolis-within-Gibbs sampling routine when necessary. WinBUGS has a more sophisticated univariate Metropolis sampler.

The “classic” BUGS program uses a text-based model description and a command-line interface. WinBUGS also offers a graphical user interface. A menu-driven set of S-Plus functions (CODA) is available to calculate convergence diagnostics and graphical statistical summaries of posterior samples.

#### Information:

MRC Biostatistics Unit, Institute of Public Health,  
Robinson Way, Cambridge CB2 2SR, U.K.

email: [bugs@mrc-bsu.cam.ac.uk](mailto:bugs@mrc-bsu.cam.ac.uk)

[ftp.mrc-bsu.ac.uk](ftp://mrc-bsu.ac.uk)

<http://www.mrc-bsu.cam.ac.uk/bugs>

## BayesX

BayesX is an object-oriented package for Bayesian inference based on Markov Chain Monte Carlo (MCMC) inference techniques, developed at the Department of Statistics at the University of Munich. Most Bayesian computations in the examples of this book have been carried out with BayesX.

The main feature of BayesX is very powerful tools for Bayesian semi-parametric regression within the GLM framework as described in Section 5.4 and relevant parts of Chapters 7, 8, and 9.

Beyond Bayesian inference for usual parametric GLMs, the current version includes possibilities to estimate nonlinear effects of metrical covariates, trends and seasonal patterns of time scales, and uncorrelated or spatially structured random effects. Predictors may include varying-coefficient terms and nonparametric interaction terms of two metrical covariates. The response can be either Gaussian, gamma, binomial, Poisson, ordered, or unordered categorical. For categorical responses, cumulative or multinomial logistic models as well as probit models based on latent Gaussian models are implemented.

The following additional features will be incorporated or will already be available when this book is published:

functions for prediction, options for two-dimensional surface smoothing, estimation of unsMOOTH functions and surfaces, adaptive B-spline regression, additional functions for handling data sets, and geographical maps.

BayesX runs under Windows and is available without charge for non-commercial usage.

Information:

Stefan Lang and Andreas Brezger  
Department of Statistics, University of Munich  
Ludwigstr. 33, 80539 Munich  
email: lang@stat.uni-muenchen.de  
andib@stat.uni-muenchen.de

# Bibliography

- AALLEN, O. O. (1980). A Model for Nonparametric Regression Analysis of Life-Times. *Statistics in Medicine* 2, 1–25.
- AALLEN, O. O. (1989). A Linear Regression Model for the Analysis of Life-Times. *Statistics in Medicine* 8, 907–925.
- ABRAHAMOWICZ, M., MACKENZIE, T., AND ESDAILE, J. M. (1996). Time-dependent Hazard Ratio: Modeling and Hypothesis Testing with Application in Lupus Nephritis. *J. Am. Stat. Ass.* 91, 1432–1440.
- ABRAMOWITZ, M. AND STEGUN, I. (1972). *Handbook of Mathematical Functions*. New York: Dover.
- AGRESTI, A. (1984). *Analysis of Ordinal Categorical Data*. New York: Wiley.
- AGRESTI, A. (1990). *Categorical Data Analysis*. New York: Wiley.
- AGRESTI, A. (1992). A Survey of Exact Inference for Contingency Tables. *Statistical Science* 7, 131–177.
- AGRESTI, A. (1993a). Computing Conditional Maximum Likelihood Estimates for Generalized Rasch Models Using Simple Loglinear Models with Diagonal Parameters. *Scandinavian Journal of Statistics* 20, 63–72.
- AGRESTI, A. (1993b). Distribution-free Fitting of Logit Models with Random Effects for Repeated Categorical Responses. *Statistics in Medicine* 12, 1969–1988.
- AGRESTI, A. (1997). A Model for Repeated Measurements of a Multivariate Binary Response. *Journal of the American Statistical Association* 92, 315–321.
- AGRESTI, A. (1999). Modelling Ordered Categorical Data: Recent Advances and Future Challenges. *Statistics in Medicine* 18, 2191–2207.
- AGRESTI, A. AND LANG, J. B. (1993). A Proportional Odds Model with Subject-Specific Effects for Repeated Ordered Categorical Responses. *Biometrika* 80, 527–534.
- AITKEN, C. G. G. (1983). Kernel Methods for the Estimation of Discrete Distributions. *J. Statist. Comput. Simul.* 16, 189–200.
- AITKIN, M. (1999). A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models. *Biometrics* 55, 117–128.
- AITKIN, M., ANDERSON, D., FRANCIS, B., AND HINDE, J. (1989). *Statistical*

- Modelling in GLIM*. Oxford: Clarendon Press.
- AITKIN, M. AND CLAYTON, D. (1980). The Fitting of Exponential, Weibull and Extreme Value Distributions to Complex Censored Survival Data Using GLIM. *Applied Statistics* 29, 156–163.
- AITKIN, M. AND FRANCIS, B. J. (1998). Fitting Generalized Linear Variance Component Models by Nonparametric Maximum Likelihood. *The GLIM Newsletter* 29 (in press).
- AITKIN, M. AND LONGFORD, N. T. (1986). Statistical Modelling Issues in School Effectiveness Studies. *Journal of the Royal Statistical Society A* 149, 1–43.
- ALBERT, J. AND CHIB, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association* 88, 669–679.
- ALBERT, J. AND CHIB, S. (1995). Bayesian Residual Analysis for Binary Response Regression Models. *Biometrika* 82, 747–759.
- ALBERT, J. AND CHIB, S. (1997). Bayesian Tests and Model Diagnostics in Conditionally Independent Hierarchical Models. *Journal of the American Statistical Association* 92, 916–925.
- ALBERT, J. H. (1988). Computational Methods Using a Bayesian Hierarchical Generalized Linear Model. *Journal of the American Statistical Association* 83, 1037–1044.
- ANDERSEN, E. B. (1973). *Conditional Inference and Models for Measuring*. Copenhagen: Metalhygiejnish Forlag.
- ANDERSEN, E. B. (1980). *Discrete Statistical Models with Social Science Applications*. Amsterdam: North-Holland.
- ANDERSEN, P. K. AND BORGAN, Ø. (1985). Counting Process Models for Life History Data: A Review (with Discussion). *Scand. J. Statistics* 12, 97–158.
- ANDERSEN, P. K., BORGAN, Ø., GILL, R., AND KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Berlin; New York: Springer-Verlag.
- ANDERSON, B. AND MOORE, J. (1979). *Optimal Filtering*. Englewood Cliffs, NJ: Prentice-Hall.
- ANDERSON, D. A. AND AITKIN, M. (1985). Variance Component Models with Binary Response: Interviewer Variability. *Journal of the Royal Statistical Society Ser. B* 47, 203–210.
- ANDERSON, D. A. AND HINDE, J. P. (1988). Random Effects in Generalized Linear Models and the EM Algorithm. *Comm. Statist. A – Theory Methods* 17, 3847–3856.
- ANDERSON, J. A. (1972). Separate Sample Logistic Discrimination. *Biometrika* 59, 19–35.
- ANDERSON, J. A. (1984). Regression and Ordered Categorical Variables. *Journal of the Royal Statistical Society B* 46, 1–30.
- ANDREWS, D. W. K. (1987). Asymptotic Results for Generalized Wald Tests. *Econometric Theory* 3, 348–358.
- ANDREWS, D. W. K. (1988). Chi-Square Diagnostic Tests for Econometric Mod-

- els: Theory. *Econometrica* 56, 1419–1453.
- AOKI, M. (1987). *State Space Modelling of Time Series*. Heidelberg: Springer-Verlag.
- ARANDA-ORDAZ, F. J. (1983). An Extension of the Proportional-Hazard-Model for Grouped Data. *Biometrics* 39, 109–118.
- ARJAS, E. (1989). Survival Models and Martingale Dynamics (with Discussion). *Scand. J. Statistics* 16, 177–225.
- ARMINGER, G. AND KÜSTERS, U. (1985). Simultaneous Equation Systems with Categorical Observed Variables. In R. Gilchrist, B. Francis & J. Whit-taker (Eds.), *Generalized Linear Models. Lecture Notes in Statistics*. Berlin: Springer-Verlag.
- ARMINGER, G. AND SCHOENBERG, R. J. (1989). Pseudo Maximum Likelihood Es-timation and a Test for Misspecification in Mean and Covariance Structure Models. *Psychometrika* 54, 409–425.
- ARMINGER, G. AND SOBEL, M. (1990). Pseudo Maximum Likelihood Estimation of Mean- and Covariance Structures with Missing Data. *Journal of the American Statistical Association* 85, 195–203.
- ARMSTRONG, B. AND SLOAN, M. (1989). Ordinal Regression Models for Epidemi-ologic Data. *American Journal of Epidemiology* 129, 191–204.
- ARTES, R. AND JORGENSEN, B. (2000). Longitudinal Data Estimating Equations for Dispersion Models. *Scandinavian Journal of Statistics* 27, 321–334.
- ASHBY, M., NEUHAUS, J., HAUCK, W., BACCHETTI, P., HEILBRON, D., JEWELL, N., SEGAL, M., AND FUSARO, R. (1992). An Annotated Bibliography of Methods for Analysing Correlated Categorical Data. *Statistics in Medi-cine* 11, 67–99.
- ATKINSON, A. AND RIANI, M. (2000). *Robust Diagnostic Regression Analysis*. New York: Springer-Verlag.
- AZZALINI, A. (1983). Maximum Likelihood Estimation of Order m for Stationary Stochastic Processes. *Biometrika* 70, 381–387.
- AZZALINI, A. AND BOWMAN, A. W. (1993). On the Use of Nonparametric Re-gression for Checking Linear Relationships. *Journal of the Royal Statistical Society B* 55, 549–557.
- AZZALINI, A., BOWMAN, A. W., AND HÄRDLE, W. (1989). On the Use of Non-parametric Regression for Linear Models. *Biometrika* 76, 1–11.
- BARNHART, H. AND SAMPSON, A. (1994). Overviews of Multinomial Models for Ordinal Data. *Comm. Stat-Theory & Methods* 23(12), 3395–3416.
- BARTHOLOMEW, D. J. (1980). Factor Analysis for Categorical Data. *Journal of the Royal Statistical Society B* 42, 293–321.
- BELSLEY, D. A., KUH, E., AND WELSCH, R. E. (1980). *Regression Diagnostics*. New York: Wiley.
- BEN-AKIVA, M. E. AND LERMAN, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press.
- BENEDETTI, J. K. (1977). On the Nonparametric Estimation of Regression Func-

- tions. *Journal of the Royal Statistical Society B* 39, 248–253.
- BENJAMIN, M., RIGBY, R., AND STASINOPoulos, M. (2000). Generalized Autoregressive Moving Average Models. *Journal of the American Statistical Association* (under revision).
- BERHANE, K. AND TIBSHIRANI, R. (1998). Generalized Additive Models for Longitudinal Data. *The Canadian Journal of Statistics* 26, 517–535.
- BESAG, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems (with Discussion). *Journal of the Royal Statistical Society B* 36, 192–236.
- BESAG, J., GREEN, P. J., HIGDON, D., AND MENGERSEN, K. (1995). Bayesian Computation and Stochastic Systems. *Statistical Science* 10, 3–66.
- BESAG, J. AND KOOPERBERG, C. (1995). On Conditional and Intrinsic Autoregressions. *Biometrika* 82, 733–746.
- BESAG, J. E., YORK, J. C., AND MOLLIE, A. (1991). Bayesian Image Restoration with Two Applications in Spatial Statistics (with Discussion). *Annals of the Institute of Statistical Mathematics* 43, 1–59.
- BHAPKAR, V. P. (1980). ANOVA and MANOVA: Models for Categorical Data. In P. R. Krishnaiah (Ed.), *Handbk. Statist.*, Volume Vol. 1: Anal. Variance, pp. 343–387. Amsterdam; New York: North-Holland/Elsevier.
- BICKEL, P. (1983). Minimax Estimation of the Mean of a Normal Distribution Subject to Doing Well at a Point. In M. Rizvi, J. Rustagi & D. Siegmund (Eds.), *Recent Advances in Statistics*, pp. 511–528. New York: Academic Press.
- BILLER, C. (2000a). Adaptive Bayesian Regression Splines in Semiparametric Generalized Linear Models. *Journal of Computational and Graphical Statistics* 9, 122–140.
- BILLER, C. (2000b). Discrete Duration Models Combining Dynamic and Random Effects. *Lifetime Data Analysis* (to appear).
- BILLER, C. (2000c). Posterior Mode Estimation in Dynamic Generalized Linear Mixed Models. Revised for *Allgemeines Statistisches Archiv (Journal of the German Statistical Association)*.
- BILLER, C. AND FAHRMEIR, L. (1997). Bayesian Spline-type Smoothing in Generalized Regression Models. *Computational Statistics* 12, 135–151.
- BIRCH, M. (1963). Maximum Likelihood in Three-way Contingency Tables. *Journal of the Royal Statistical Society B* 25, 220–233.
- BISHOP, Y., FIENBERG, S., AND HOLLAND, P. (1975). *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press.
- BLOSSFELD, H. P., HAMERLE, A., AND MAYER, K. U. (1989). *Event History Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- BOCK, R. D. AND AITKIN, M. (1981). Marginal Maximum Likelihood Estimation of Item Parameters: An Application of an EM Algorithm. *Psychometrika* 46, 443–459.
- BONNEY, G. (1987). Logistic Regression for Dependent Binary Observations.

- Biometrics* 43, 951–973.
- BOOTH, J. G. AND HOBERT, J. P. (1999). Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm. *J. R. Statist. Soc. B* 61, 265–285.
- BÖRSCH-SUPAN, A. (1990). On the Compatibility of Nested Logit Models with Utility Maximization. *Journal of Econometrics* 43, 373–388.
- BOYLES, R. A. (1983). On the Covergence of the EM Algorithm. *Journal of the Royal Statistical Society B* 45, 47–50.
- BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., AND STONE, J. C. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth.
- BRESLOW, N. E. (1984). Extra-Poisson Variation in Log-Linear Models. *Applied Statistics* 33, 38–44.
- BRESLOW, N. E. AND CLAYTON, D. G. (1993). Approximate Inference in Generalized Linear Mixed Model. *Journal of the American Statistical Association* 88, 9–25.
- BRESLOW, N. E. AND LIN, X. (1995). Bias Correction in Generalized Linear Mixed Models with a Single Component of Dispersion. *Biometrika* 82, 81–91.
- BRILLINGER, D. R. AND PREISLER, M. K. (1983). Maximum Likelihood Estimation in a Latent Variable Problem. In T. Amemiya, S. Karlin & T. Goodman (Eds.), *Studies in Econometrics, Time Series, and Multivariate Statistics*, pp. 31–65. New York: Academic Press.
- BROWN, P. J. AND PAYNE, C. D. (1986). Aggregate Data. Ecological Regression and Voting Transitions. *Journal of the American Statistical Association* 81, 452–460.
- BRUMBACK, B., RYAN, L., SCHWARTZ, J., NEAS, L., STARK, P., AND BURGE, H. (2000). Transitional Regression Models, with Application to Environmental Time Series. *Journal of the American Statistical Assoc.* 95, 16–27.
- BUJA, A., HASTIE, T., AND TIBSHIRANI, R. (1989). Linear Smoothers and Additive Models. *Annals of Statistics* 17, 453–510.
- BUSE, A. (1982). The Likelihood Ratio, Wald and Lagrange Multiplier Test: An Expository Note. *The American Statistician* 36, 153–157.
- CAMERON, A. AND TRIVEDI, K. (1998). *Regression Analysis of Count Data. Econometric Society Monographs No. 30*. Cambridge: Cambridge University Press.
- CAMERON, A. AND TRIVEDI, P. K. (1986). Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests. *Journal of Applied Econometrics* 1, 29–53.
- CARLIN, B. AND CHIB, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society B* 57, 473–484.
- CARLIN, B. AND POLSON, N. (1992). Monte Carlo Bayesian Methods for Discrete Regression Models and Categorical Time Series. In J. Bernardo, J. Berger, A. Dawid & A. Smith (Eds.), *Bayesian Statistics 4*, pp. 577–586. Oxford:

University Press.

- CARLIN, B. P., POLSON, N. G., AND STOFFER, D. S. (1992). A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modelling. *Journal of the American Statistical Association* 87, 493–500.
- CARROLL, R. J., FAN, J., GIJBELS, I., AND WAND, M. P. (1997). Generalized Partially Linear Single-index Models. *Journal of the American Statistical Association* 92, 477–489.
- CARROLL, R. J., KÜCHENHOFF, H., LOMBARD, F., AND STEFANSKI, L. A. (1996). Asymptotics for the SIMEX Estimator in Nonlinear Measurement Error Models. *Journal of the American Statistical Assoc.* 91, 242–250.
- CARROLL, R. J. AND PEDERSON, S. (1993). On Robustness in the Logistic Regression Model. *Journal of the Royal Statistical Society B* 55, 693–706.
- CARROLL, R. J., RUPPERT, D., AND STEFANSKI, L. A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.
- CARROLL, R. J., RUPPERT, D., AND WELSH, A. H. (1998). Local Estimating Equations. *Journal of the American Statistical Association* 93, 214–227.
- CARTER, C. K. AND KOHN, R. (1994a). On Gibbs Sampling for State Space Models. *Biometrika* 81, 541–553.
- CARTER, C. K. AND KOHN, R. (1994b). Robust Bayesian Nonparametric Regression. In W. Härdle & M. G. Schimek (Eds.), *Statistical Theory and Computational Aspects of Smoothing*, pp. 128–148. Heidelberg: Physica Verlag.
- LE CESSIE, S. AND VAN HOUWELINGEN, J. C. (1991). A Goodness-of-Fit Test for Binary Regression Models, Based on Smoothing Methods. *Biometrics* 47, 1267–1282.
- CHAMBERS, J. M. AND HASTIE, T. J. (1992). *Statistical Models in S*. Pacific Grove, CA: Wadsworth Brooks/Cole.
- CHEN, M.-H. AND DEY, D. K. (1999). Bayesian Analysis for Correlated Ordinal Data Models. In D. Dey, S. Gosh & B. Mallick (Eds.), *Generalized Linear Models: A Bayesian Perspective*, Chapter 8, pp. 135–162. New York: Marcel Dekker.
- CHIB, S. (1995). Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association* 90, 1313–1321.
- CHIB, S. (1999). Bayesian Methods for Correlated Binary Data. In D. Dey, S. Gosh & B. Mallick (Eds.), *Generalized Linear Models: A Bayesian Perspective*. New York: Marcel Dekker.
- CHRISTENSEN, R. (1997). *Log-linear Models and Logistic Regression*. New York: Springer-Verlag.
- CLAYTON, D. AND RASBASH, J. (1999). Estimation in Large Crossed Random-Effect Models by Data Augmentation. *J. R. Statist. Soc. A* 162, 425–436.
- CLAYTON, D. G. (1996). Generalized Linear Mixed Models. In W. R. Gilks, S. Richardson & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

- CLEVELAND, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* 74, 829–836.
- CLEVELAND, W. S. AND LOADER, C. (1996). Smoothing by Local Regression: Principles and Methods. In W. Härdle & M. Schimek (Eds.), *Statistical Theory and Computational Aspects of Smoothing*, pp. 10–49. Heidelberg: Physica-Verlag.
- CLOGG, C. (1982). Some Models for the Analysis of Association in Multiway Crossclassifications Having Ordered Categories. *Journal of the American Statistical Association* 77, 803–815.
- COLLETT, D. (1991). *Modelling Binary Data*. London: Chapman & Hall.
- CONAWAY, M. R. (1989). Analysis of Repeated Categorical Measurements with Conditional Likelihood Methods. *Journal of the American Statistical Association* 84, 53–62.
- CONAWAY, M. R. (1990). A Random Effects Model for Binary Data. *Biometrics* 46, 317–328.
- CONGDON, P. (1993). Statistical Graduation in Local Demographic Analysis and Projection. *J. Roy. Statist. Soc. A* 156(2), 237–270.
- CONOLLY, M. A. AND LIANG, K. (1988). Conditional Logistic Regression Models for Correlated Binary Data. *Biometrika* 75, 501–506.
- COOK, R. D. (1977). Detection of Influential Observations in Linear Regression. *Technometrics* 19, 15–18.
- COOK, R. D. AND WEISBERG, S. (1982). *Residuals and Influence in Regression*. London: Chapman & Hall.
- COPAS, J. B. (1988). Binary Regression Models for Contaminated Data (with Discussion). *Journal of the Royal Statistical Society B* 50, 225–265.
- COPAS, J. B. AND HABERMAN, S. (1983). Non-Parametric Graduation Using Kernel Methods. *J. Inst. Act.* 110, 135–156.
- CORBEIL, R. R. AND SEARLE, S. R. (1976). A Comparison of Variance Component Estimators. *Biometrics* 32, 779–791.
- COWLES, M. K. AND CARLIN, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association* 91, 883–904.
- Cox, C. (1988). Multinomial Regression Models Based on Continuation Ratios. *Statistics in Medicine* 7, 433–441.
- Cox, D. (1970). *The Analysis of Binary Data*. London: Chapman & Hall.
- Cox, D. R. (1961). Tests of Separate Families of Hypotheses. In *Proc. of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1. Berkeley: University of California Press.
- Cox, D. R. (1962). Further Results on Tests of Separate Families of Hypotheses. *Journal of the Royal Statistical Society B* 24, 406–424.
- Cox, D. R. (1972). Regression Models and Life Tables (with Discussion). *Journal of the Royal Statistical Society B* 34, 187–220.

- Cox, D. R. (1975). Partial Likelihood. *Biometrics* 62, 269–275.
- Cox, D. R. AND HINKLEY, D. V. (1974). *Theoretical Statistics*. London: Chapman & Hall.
- Cox, D. R. AND WERMUTH, N. (1996). *Multivariate Dependencies*. London, New York: Chapman & Hall.
- CRAVEN, P. AND WAHBA, G. (1979). Smoothing Noisy Data with Spline Functions. *Numerische Mathematik* 31, 377–403.
- CRESSIE, N. (1993). *Statistics for Spatial Data, Rev.* New York: Wiley.
- CRESSIE, N. AND READ, I. (1984). Multinomial Goodness-of-Fit Tests. *Journal of the Royal Statistical Society B* 46, 440–464.
- CZADO, C. (1992). On Link Selection in Generalized Linear Models. In S. L. N. in Statistics (Ed.), *Advances in GLIM and Statistical Modelling*. Springer-Verlag. 78, 60–65.
- DAGANZO, C. (1980). *Multinomial Probit*. New York: Academic Press.
- DALE, J. (1986). Asymptotic Normality of Goodness-of-Fit Statistics for Sparse Product Multinomials. *Journal of the Royal Statistical Society B* 48, 48–59.
- DAVIDSON, R. AND MCKINNON, J. G. (1980). On a Simple Procedure for Testing Non-nested Regression Models. *Economic Letters* 5, 45–48.
- DAVIS, C. S. (1991). Semi-parametric and Non-parametric Methods for the Analysis of Repeated Measurements with Applications to Clinical Trials. *Statistics in Medicine* 10, 1959–1980.
- DAVIS, P. J. AND RABINOWITZ, P. (1975). *Numerical Integration*. Waltham, MA: Blaisdell.
- DAVIS, P. J. AND RABINOWITZ, P. (1984). *Methods of Numerical Integration*. Orlando, FL: Academic Press.
- DAVISON, A. C. AND HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.
- DE BOOR, C. (1978). *A Practical Guide to Splines*. New York: Springer-Verlag.
- DE JONG, P. (1989). Smoothing and Interpolation with the State Space Model. *Journal of the American Statistical Association* 84, 1085–1088.
- DE JONG, P. AND SHEPARD, N. (1995). The Simulation Smoother for Time Series Models. *Biometrika* 82, 339–350.
- DEAN, C. B. (1992). Testing for Overdispersion in Poisson and Binomial Regression Models. *Journal of the American Statistical Association* 87, 451–457.
- DECARLI, A., FRANCIS, B. J., GILCHRIST, R., AND SEEGER, G. H. (1989). *Statistical Modelling*. Number 57 in Springer Lecture Notes in Statistics. Berlin: Springer-Verlag.
- DELLAPORTAS, P. AND SMITH, A. F. M. (1993). Bayesian Inference for Generalized Linear and Proportional Hazard Models via Gibbs Sampling. *Applied Statistics* 42, 443–459.
- DELLAPORTAS, P., FORSTER, J., AND NTZOUFRAS, J. (1999). Bayesian Variable Selection Using the Gibbs Sampler. In D. Dey, M. Gosh & B. Mallick

- (Eds.), *Generalized Linear Models: A Bayesian Perspective*. New York: Marcel Dekker.
- DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B* 39, 1–38.
- DENISON, D., MALICK, B., AND SMITH, A. (1998). Automatic Bayesian Curve Fitting. *Journal of the Royal Statistical Society B* 60, 333–350.
- VAN DEUSEN, P. (1989). A Model-Based Approach to Tree Ring Analysis. *Biometrics* 45, 763–779.
- DEVROYE, L. (1986). *Non-uniform Random Variate Generation*. New York: Springer-Verlag.
- DEY, D. K., GELFAND, A., SWARTZ, T., AND VLACHOS, P. (1998). Simulation Based Model Checking for Hierarchical Models. *Test* 7, 325–346.
- DEY, D. K., MÜLLER, P., AND SINHA, D. (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics*. Number 133 in Lecture Notes in Statistics. New York: Springer-Verlag.
- DEY, D. P., GOSH, S., AND MALICK, B. (1999). *Generalized Linear Models: A Bayesian Perspective*. New York: Marcel Dekker.
- DI CICCIO, T., KASS, R., RAFTERY, A., AND WASSERMAN, L. (1997). Computing Bayes Factors by Combining Simulations and Asymptotic Approximations. *Journal of the American Statistical Association* 92, 903–915.
- DIelman, T. E. (1989). *Pooled Cross-Sectional and Time Series Analysis*. New York: Marcel Dekker.
- DIETZ, E. AND BÖNING, D. (1995). Statistical Inference Based on a General Model of Unobserved Heterogeneity. In B. J. Francis, R. Hatzinger, G. U. H. Seeber & G. Steckel-Berger (Eds.), *Statistical Modelling*. New York: Springer-Verlag.
- DIGGLE, P. J., LIANG, K., AND ZEGER, S. (1994). *Analysis of Longitudinal Data*. London: Chapman & Hall.
- DOBSON, A. J. (1989). *Introduction to Statistical Modelling*. London: Chapman & Hall.
- DOKSUM, K. A. AND GASKO, M. (1990). On a Correspondence Between Models in Binary Regression Analysis and in Survival Analysis. *International Statistical Review* 58, 243–252.
- DONOHO, D. AND JOHNSTONE, I. (1995). Adapting to Unknown Smoothness via Wavelet Shrinkage. *Journal of the American Statistical Association* 90, 1200–1224.
- DONOHO, D., JOHNSTONE, I., KERKYACHARIAN, G., AND PICARD, D. (1995). Wavelet Shrinkage: Asymptopia? (with Discussion). *Journal of the Royal Statistical Society B* 57, 301–369.
- DOUCET, A., FREITAS, J., AND GORDON, N. (2000). *Sequential Monte Carlo Methods*. New York: Springer-Verlag.
- DUFFY, D. E. AND SANTNER, T. J. (1989). On the Small Sample Properties of

- Restricted Maximum Likelihood Estimators for Logistic Regression Models. *Communication in Statistics, Theory & Methods* 18, 959–989.
- DURBIN, J. AND KOOPMAN, S. J. (1993). *Filtering, Smoothing and Estimation of Time Series Models When the Observations come from Exponential Family Distributions*. London: London School of Economics.
- EDWARDS, A. AND THURSTONE, L. (1952). An Internal Consistency Check for Scale Values Determined by the Method of Successive Intervals. *Psychometrika* 17, 169–180.
- EDWARDS, D. AND HAVRANEK, T. (1987). A Fast Model Selection Procedure for Large Family of Models. *Journal of the American Statistical Association* 82, 205–213.
- EFRON, B. (1986). Double Exponential Families and Their Use in Generalized Linear Regression. *Journal of the American Statistical Association* 81, 709–721.
- EFRON, B. (1988). Logistic Regression, Survival Analysis, and the Kaplan-Meier-Curve. *Journal of the American Statistical Association* 83, 414–425.
- EFRON, B. AND TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.
- EILERS, P. H. C. AND MARX, B. D. (1996). Flexible Smoothing with B-Splines and Penalties. *Statistical Science* 11, 89–121.
- ENGLE, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of United Kingdom Inflation. *Econometrica* 50, 987–1007.
- ENGLE, R. F., GRANGER, R., RICE, J., AND WEISS, A. (1986). Semiparametric Estimates of the Relation Between Weather and Electricity Sales. *Journal of the American Statistical Association* 81, 310–320.
- EPANECHNIKOV, V. (1969). Nonparametric Estimates of a Multivariate Probability Density. *Theory of Probability and Applications* 14, 153–158.
- ESCOBAR, M. AND WEST, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association* 90, 577–588.
- EUBANK, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker.
- EUBANK, R. L. AND HART, J. D. (1993). Commonality of Cusum, von Neumann and Smoothingbased Goodness-of-Fit Tests. *Biometrika* 80, 89–98.
- FAHRMEIR, L. (1987a). Asymptotic Likelihood Inference for Nonhomogeneous Observations. *Statist. Hefte (N.F.)* 28, 81–116.
- FAHRMEIR, L. (1987b). Asymptotic Testing Theory for Generalized Linear Models. *Math. Operationsforsch. Statist. Ser. Statist.* 18, 65–76.
- FAHRMEIR, L. (1988). A Note on Asymptotic Testing Theory for Nonhomogenous Observation. *Stochastic Processes and Its Applications* 28, 267–273.
- FAHRMEIR, L. (1990). Maximum Likelihood Estimation in Misspecified Generalized Linear Models. *Statistics* 21, 487–502.
- FAHRMEIR, L. (1992a). Posterior Mode Estimation by Extended Kalman Filter-

- ing for Multivariate Dynamic Generalized Linear Models. *Journal of the American Statistical Association* 87, 501–509.
- FAHRMEIR, L. (1992b). State Space Modeling and Conditional Mode Estimation for Categorical Time Series. In D. Brillinger et al. (Eds.), *New Directions in Time Series Analysis*, pp. 87–110. New York: Springer-Verlag.
- FAHRMEIR, L. (1994). Dynamic Modelling and Penalized Likelihood Estimation for Discrete Time Survival Data. *Biometrika* 81, 317–330.
- FAHRMEIR, L., FRANCIS, B., GILCHRIST, R., AND TUTZ, G. (1992). *Advances in GLIM and Statistical Modelling*. Number 78 in Lecture Notes in Statistics. New York: Springer-Verlag.
- FAHRMEIR, L. AND FROST, H. (1992). On Stepwise Variable Selection in Generalized Linear Regression and Time Series Models. *Computational Statistics* 7, 137–154.
- FAHRMEIR, L. AND HAMERLE, A. (1984). *Multivariate statistische Verfahren*. Berlin/New York: de Gruyter.
- FAHRMEIR, L., HAMERLE, A., AND TUTZ, G. (1996). *Multivariate statistische Verfahren* (2nd ed.). Berlin, New York: de Gruyter.
- FAHRMEIR, L., HENNEVOGL, W., AND KLEMME, K. (1992). Smoothing in Dynamic Generalized Linear Models by Gibbs Sampling. In L. Fahrmeir, B. Francis, R. Gilchrist & G. Tutz (Eds.), *Advances in GLIM and Statistical Modelling*. Heidelberg: Springer-Verlag.
- FAHRMEIR, L. AND KAUFMANN, H. (1985). Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models. *The Annals of Statistics* 13, 342–368.
- FAHRMEIR, L. AND KAUFMANN, H. (1986). Asymptotic Inference in Discrete Response Models. *Statistical Papers* 27, 179–205.
- FAHRMEIR, L. AND KAUFMANN, H. (1987). Regression Model for Nonstationary Categorical Time Series. *Journal of Time Series Analysis* 8, 147–160.
- FAHRMEIR, L. AND KAUFMANN, H. (1991). On Kalman Filtering, Posterior Mode Estimation and Fisher Scoring in Dynamic Exponential Family Regression. *Metrika* 38, 37–60.
- FAHRMEIR, L., KAUFMANN, H., AND MORAWITZ, B. (1989). Varying Parameter Models for Panel Data with Applications to Business Test Data. Regensburger Beiträge zur Statistik und Ökonometrie, Universität Regensburg.
- FAHRMEIR, L. AND KNORR-HELD, L. (1997). Dynamic Discrete-Time Duration Models: Estimation via Markov Chain Monte Carlo. *Sociological Methodology* 27, 417–452.
- FAHRMEIR, L. AND KNORR-HELD, L. (2000). Dynamic and Semiparametric Models. In M. Schimek (Ed.), *Smoothing and Regression: Approaches, Computation and Application*, Chapter 18, pp. 505–536. New York: John Wiley & Sons.
- FAHRMEIR, L. AND KREDLER, C. (1984). Verallgemeinerte lineare Modelle. In L. Fahrmeir & A. Hamerle (Eds.), *Multivariate statistische Verfahren*. Berlin: De Gruyter.

- FAHRMEIR, L. AND LANG, S. (1999). Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors. *Applied Statistics* (to appear).
- FAHRMEIR, L. AND LANG, S. (2000). Bayesian Semiparametric Regression Analysis of Multicategorical Time-Space Data. *Annals of the Institute of Statistical Mathematics* (to appear).
- FAHRMEIR, L. AND PRITSCHER, L. (1996). Regression Analysis of Forest Damage by Marginal Models for Correlated Ordinal Responses. *Journal of Environmental and Ecological Statistics* 3, 257–268.
- FAHRMEIR, L. AND WAGENPFEIL, S. (1996). Smoothing Hazard Functions and Time-Varying Effects in Discrete Duration and Competing Risks Models. *Journal of the American Statistical Association* 91, 1584–1594.
- FAHRMEIR, L. AND WAGENPFEIL, S. (1997). Penalized Likelihood Estimation and Iterative Kalman Filtering for non-Gaussian Dynamic Regression Models. *Computational Statistics and Data Analysis* 24, 295–320.
- FAN, J. AND GIBBELS, I. (1994). Censored Regression: Local Linear Approximation and Their Applications. *J. Amer. Statist. Assoc.* 89, 560–570.
- FAN, J. AND GIBBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman & Hall.
- FERREIRA, M. A. R. AND GAMERMAN, D. (1999). Dynamic Generalized Linear Models. In D. Dey, S. Gosh & B. Mallick (Eds.), *Generalized Linear Models: A Bayesian Perspective*, Chapter 4, pp. 57–72. New York: Marcel Dekker.
- FINKELSTEIN, D. (1986). A Proportional Hazard Model for Interval-Censored Failure Time Data. *Biometrika* 73, 845–854.
- FINNEY, D. (1947). The Estimation from Individual Records of the Relationship Between Dose and Quantal Response. *Biometrika* 34, 320–334.
- FIRTH, D. (1991). Generalized Linear Models. In D. V. Hinkley, N. Reid & E. J. Snell (Eds.), *Statistical Theory and Modelling*. London: Chapman & Hall.
- FIRTH, D., GLOSUP, J., AND HINKLEY, D. V. (1991). Model Checking with Nonparametric Curves. *Biometrika* 78, 245–252.
- FITZMAURICE, G. M. (1995). A Caveat Concerning Independence Estimating Equations with Multivariate Binary Data. *Biometrics* 51, 309–317.
- FITZMAURICE, G. M. AND LAIRD, N. M. (1993). A Likelihood-based Method for Analysing Longitudinal Binary Responses. *Biometrika* 80, 141–151.
- FITZMAURICE, G. M., LAIRD, N. M., AND ROTNITZKY, A. G. (1993). Regression Models for Discrete Longitudinal Responses. *Statistical Science* 8, 284–309.
- FITZMAURICE, G. M., LIPSITZ, S. R., AND MOLENBERGHS, G. M. (1995). Regression Models for Longitudinal Binary Responses with Informative Drop-Outs. *Journal of the Royal Statistical Society* 57, 691–704.
- FOKIANOS, K. (2000). Truncated Poisson Regression for Time Series of Counts. *Scandinavian Journal of Statistics* (to appear).
- FOKIANOS, K. AND KEDEM, B. (1998). Prediction and Classification of Non-Stationary Categorical Time Series. *Journal of Multivariate Analysis* 67,

277–296.

- FOLKS, J. L. AND CHHIKARA, R. S. (1978). The Inverse Gaussian Distribution and Its Statistical Application, A Review (with Discussion). *Journal of the Royal Statistical Society B* 40, 263–289.
- FORCINA, A., MARCHETTI, G. M., HATZINGER, R., AND GALMACCI, G. (Eds.) (1996). *Statistical Modelling*, Proceedings of the 11th International Workshop on Statistical Modelling, Orvieto, 1996.
- FORTHOFER, R. N. AND LEHNEN, R. G. (1981). *Public Program Analysis. A New Categorical Data Approach*. Belmont, CA: Lifetime Learning Publications.
- FOUTZ, R. V. AND SRIVASTAVA, R. C. (1977). The Performance of the Likelihood Ratio Test When the Model Is Incorrect. *Annals of Statistics* 5, 1183–1194.
- FRIEDL, H. (1991). *Verallgemeinerte logistische Modelle in der Analyse von Zervix-Karzinomen*. Ph.D. thesis, Universität Graz.
- FRIEDL, H., BERGHOLD, A., AND KAUERMANN, G. (Eds.) (1999). *Statistical Modelling*, Proceedings of the 14th International Workshop on Statistical Modelling, Graz, 1999.
- FRIEDMAN, J. (1991). Multivariate Adaptive Regression Splines (with Discussion). *Ann. Statist.* 19, 1–14.
- FRIEDMAN, J. AND SILVERMAN, B. (1989). Flexible Parsimonious Smoothing and Additive Modelling (with Discussion). *Technometrics* 31, 3–39.
- FRIEDMAN, J. H. AND STÜTZLE, W. (1981). Projection Pursuit Regression. *J. Amer. Statist. Assoc.* 76, 817–823.
- FROST, H. (1991). *Modelltests in generalisierten linearen Modellen*. Ph.D. thesis, Universität Regensburg.
- FRÜHWIRTH-SCHNATTER, S. (1991). Monitoring von ökologischen und biometrischen Prozessen mit statistischen Filtern. In C. Minder & G. Seeber (Eds.), *Multivariate Modelle: Neue Ansätze für biometrische Anwendungen*. Berlin: Springer-Verlag.
- FRÜHWIRTH-SCHNATTER, S. (1994). Data Augmentation and Dynamic Linear Models. *Journal of Time Series Analysis* 15(2), 183–202.
- FURNIVAL, G. M. AND WILSON, R. W. (1974). Regression by Leaps and Bounds. *Technometrics* 16, 499–511.
- GAMERMAN, D. (1997a). Efficient Sampling from the Posterior Distribution in Generalized Linear Mixed Models. *Statistics and Computing* 7, 57–68.
- GAMERMAN, D. (1997b). *Markov Chain Monte Carlo; Stochastic Simulation for Bayesian Inference*. London: Chapman & Hall.
- GAMERMAN, D. (1998). Markov Chain Monte Carlo for Dynamic Generalized Linear Models. *Biometrika* 85, 215–227.
- GARBER, A. M. (1989). A Discrete-Time Model of the Acquisition of Antibiotic-Resistant Infections in Hospitalized Patients. *Biometrics* 45, 797–816.
- GASSER, T. AND MÜLLER, H.-G. (1979). Smoothing Techniques for Curve Estimation. In R. Gasser & H. Rosenblatt (Eds.), *Kernel Estimation of Regression Function*. Heidelberg: Springer-Verlag.

- GASSER, T. AND MÜLLER, H.-G. (1984). Estimating Regression Functions and Their Derivatives by the Kernel Method. *Scand. J. Statist.* 11, 171–185.
- GASSER, T., MÜLLER, H.-G., AND MAMMITZSCH, V. (1985). Kernels for Nonparametric Curve Estimation. *Journal of the Royal Statistical Society B* 47, 238–252.
- GAY, D. M. AND WELSCH, R. E. (1988). Maximum Likelihood and Quasi-Likelihood for Nonlinear Exponential Family Regression Models. *Journal of the American Statistical Association* 83, 990–998.
- GELFAND, A. AND GHOSH, M. (1999). Generalized Linear Models: A Bayesian View. In D. Dey, M. Gosh & B. Mallick (Eds.), *Generalized Linear Models: A Bayesian Perspective*. New York: Marcel Dekker.
- GELFAND, A. E. (1998). Approaches for Semiparametric Bayesian Regression. In S. Ghosh (Ed.), *Asymptotics, Nonparametrics and Time Series*. New York: Marcel Dekker (to appear).
- GELFAND, A. E., DEY, D. K., AND CHANG, H. (1992). Model Determination Using Predictive Distributions with Implementation via Sampling-based Methods (with Discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith (Eds.), *Bayesian Statistics 4*. Oxford: University Press.
- GELFAND, A. E. AND GHOSH, S. K. (1998). Model Choice: A Minimum Posterior Predictive Loss Approach. *Biometrika* 85, 1–11.
- GELFAND, A. E. AND SMITH, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* 85, 398–409.
- GELFAND, E. E. (1996). Model Determination Using Sampling-based Methods. In W. Gilks, S. Richardson & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 145–161. London: Chapman & Hall.
- GELMAN, A., CARLIN, J. B., STERN, H. S., AND RUBIN, D. B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.
- GELMAN, A. AND MENG, X.-L. (1996). Model Checking and Model Improvement. In W. Gilks et al. (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 189–202. London: Chapman & Hall.
- GELMAN, A., MENG, X.-L., AND STERN, H. (1995). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies (with Discussion). *Statistica Sinica* 6, 733–807.
- GEMAN, S. AND GEMAN, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 721–741.
- GENTER, F. C. AND FAREWELL, V. T. (1985). Goodness-of-Link Testing in Ordinal Regression Models. *The Canadian Journal of Statistics* 13, 37–44.
- GERSCH, W. AND KITAGAWA, G. (1988). Smoothness Priors in Time Series. In J. C. Spall (Ed.), *Bayesian Analysis of Time Series and Dynamic Models*. New York: Marcel Dekker.
- GIEGER, C. (1998). *Marginale Regressionsmodelle mit varierenden Koeffizienten für kategoriale Zielvariablen*. Aachen: Shaker Verlag.

- GIEGER, C. (1999). Marginal Regression Models with Varying Coefficients for Correlated Ordinal Data. Sfb 386 discussion paper 177, Dept. of Statistics, University of Munich.
- GIESBRECHT, F. G. AND BURNS, J. C. (1985). Two-Stage Analysis Based on a Mixed-Model: Large-Sample Asymptotic Theory and Small-Sample Simulation Results. *Biometrics* 41, 477–486.
- GIGLI, A. (1992). Bootstrap Importance Sampling in Regression. In P. Heyden, W. Jansen, B. Francis & G. Seeber (Eds.), *Statistical Modelling*, pp. 95–104. Amsterdam: North-Holland.
- GILCHRIST, R., FRANCIS, B., AND WHITTAKER, J. (1985). *Generalized Linear Models, Proceedings, Lancaster 1985. Springer Lecture notes*. New York: Springer-Verlag.
- GILKS, W., RICHARDSON, S., AND SPIEGELHALTER, D. J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.
- GILKS, W. R. AND WILD, P. (1992). Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics* 4, 337–348.
- GLONEK, G. V. F. (1996). A Class of Regression Models for Multivariate Categorical Responses. *Biometrika* 83, 15–28.
- GLONEK, G. V. F. AND McCULLAGH, P. (1996). Multivariate Logistic Models. *Journal of the Royal Statistical Society B* 57, 533–546.
- GOLDSTEIN, H. (1986). Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares. *Biometrika* 73, 43–56.
- GOLDSTEIN, H. (1987). *Multilevel Models in Educational and Social Research*. New York: London and Oxford University Press.
- GOLDSTEIN, H. (1989). Restricted Unbiased Iterative Generalized Least Squares Estimation. *Biometrika* 76, 622–623.
- GOODMAN, L. A. (1979). Simple Models for the Analysis of Association in Cross-Classification Having Ordered Categories. *Journal of the American Statistical Society* 74, 537–552.
- GOODMAN, L. A. (1981a). Association Models and Canonical Correlation in the Analysis of Cross-Classification Having Ordered Categories. *Journal of the American Statistical Association* 76, 320–334.
- GOODMAN, L. A. (1981b). Association Models and the Bivariate Normal for Contingency Tables with Ordered Categories. *Biometrika* 68, 347–355.
- GORDON, K. (1986). The Multi State Kalman Filter in Medical Monitoring. *Computer Methods and Programs in Biomedicine* 23, 147–154.
- GORDON, N., SALMOND, D., AND SMITH, A. (1993). Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation. *IEE Proceedings, Part F* 140(2), 107–113.
- Goss, M. (1990). Schätzung und Identifikation von Struktur- und Hyperstrukturparametern in dynamischen generalisierten linearen Modellen. Master's thesis, Universität Regensburg.
- GÖTTLIN, A. AND PRUSCHA, H. (1996). Der Einfluß von Bestandskenngrößen,

- Topographie, Standort und Witterung auf die Entwicklung des Kronenzustandes im Bereich des Forstamtes Rothenbuch. *Forstwirtschaftliches Centralblatt* 114, 146–162.
- GOULD, A. AND LAWLESS, J. F. (1988). Estimation Efficiency in Lifetime Regression Models When Responses Are Censored or Grouped. *Comm. Statist. Simul.* 17, 689–712.
- GORILOUX, C. (1985). Asymptotic Comparison of Tests for Non-nested Hypotheses by Bahadur's A.R.E. In J.-P. Florens (Ed.), *Model Choice, Proceedings of the 4th Franco-Belgian Meeting of Statisticians*. Facultés Univ.: Saint-Louis, Bruxelles.
- GORILOUX, C., HOLLY, A., AND MONFORT, A. (1982). Likelihood Ratio Test, Wald Test and Kuhn-Tucker Test in Linear Models with Inequality Constraints on the Regression Parameters. *Econometrica* 50, 63–80.
- GORILOUX, C. AND MONFORT, A. (1989). Simulation Based Inference in Models with Heterogeneity. Document de Travail INSEE/ENSAE. 8902.
- GORILOUX, C., MONFORT, A., AND TROGNON, A. (1983a). Estimation and Test in Probit Models with Serial Correlation. Discussion Paper 8220, CEPREMAP.
- GORILOUX, C., MONFORT, A., AND TROGNON, A. (1983b). Testing Nested or Non-nested Hypotheses. *Journal of Econometrics* 21, 83–115.
- GORILOUX, C., MONFORT, A., AND TROGNON, A. (1984). Pseudo Maximum Likelihood Methods: Theory. *Econometrica* 52, 681–700.
- GREEN, D. J. AND SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman & Hall.
- GREEN, P. J. (1984). Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives. *Journal of the Royal Statistical Society B*, 149–192.
- GREEN, P. J. (1987). Penalized Likelihood for General Semi-Parametric Regression Models. *International Statistical Review* 55, 245–259.
- GREEN, P. J. (1989). Generalized Linear Models and Some Extensions. Geometry and Algorithms. In A. Decarli, B. J. Francis, R. Gilchrist & G. Seeber (Eds.), *Statistical Modelling*, pp. 26–36. Heidelberg: Springer-Verlag.
- GREEN, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika* 82, 711–732.
- GREEN, P. J. AND YANDELL, B. S. (1985). Semi-parametric Generalized Linear Models. In R. Gilchrist, B. J. Francis & J. Whittaker (Eds.), *Generalized Linear Models, Lecture Notes in Statistics*, Volume 32, pp. 44–55. Berlin: Springer-Verlag.
- GREENLAND, S. (1994). Alternative Models for Ordinal Logistic Regression. *Statistics in Medicine* 13, 1665–1677.
- GREENWOOD, M. (1926). The Natural Duration of Cancer. Reports of Public Health and Medical Subjects 33, His Majesty's Stationery Office, London.
- GRETHER, D. M. AND MADDALA, G. S. (1982). A Time Series Model with Qual-

- itative Variables. In M. Deistler et al. (Eds.), *Games, Economic Dynamics and Time Series Analysis*, pp. 291–305. Wien: Physica.
- GU, C. (1990). Adaptive Spline Smoothing in Non-Gaussian Regression Models. *Journal of the American Statistical Association* 85, 801–807.
- GUAN, D. AND YUAN, L. (1991). The Integer-Valued Autoregressive (INAR(p)) Model. *Journal of Time Series Analysis* 12, 129–142.
- HABERMAN, S. J. (1974). Loglinear Models for Frequency Tables with Ordered Classifications. *Biometrics* 30, 589–600.
- HABERMAN, S. J. (1977). Maximum Likelihood Estimates in Exponential Response Models. *Annals of Statistics* 5, 815–841.
- HABERMAN, S. J. (1978). *Analysis of Qualitative Data, 1: Introductory Topics*. New York: Academic Press.
- HALL, P. AND HEYDE, C. C. (1980). *Martingale Limit Theory and Its Applications*. New York: Academic Press.
- HALL, P. AND JOHNSTONE, I. (1992). Empirical Functionals and Efficient Smoothing Parameter Selection (with Discussion). *Journal of the Royal Statistical Society B* 54, 475–530.
- HAMERLE, A. AND NAGL, W. (1988). Misspecification in Models for Discrete Panel Data: Applications and Comparisons of Some Estimators. Diskussionsbeitrag Nr. 105. Universität Konstanz.
- HAMERLE, A. AND RONNING, G. (1992). Panel Analysis for Qualitative Variables. In G. Arminger et al. (Eds.), *A Handbook for Statistical Modelling in the Social and Behavioral Sciences*. New York: Plenum.
- HAMERLE, A. AND TUTZ, G. (1989). *Diskrete Modelle zur Analyse von Verweildauern und Lebenszeiten*. Frankfurt/New York: Campus Verlag.
- HAN, C. AND CARLIN, B. (2000). MCMC Methods for Computing Bayes Factors: A Comparative Review. Technical Report, University of Minnesota.
- HANEFELD, U. (1987). *Das sozio-ökonomische Panel*. Frankfurt: Campus Verlag.
- HÄRDLE, W. (1990a). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.
- HÄRDLE, W. (1990b). *Smoothing Techniques. With Implementations in S*. New York: Springer-Verlag.
- HÄRDLE, W., HALL, P., AND MARRON, J. S. (1988). How Far Are Automatically Chosen Regression Smoothing Parameters from Their Optimum? *Journal of the American Statistical Association* 83, 86–101.
- HARRISON, P. J. AND STEVENS, C. (1976). Bayesian Forecasting (with Discussion). *Journal of the Royal Statistical Society B* 38, 205–247.
- HART, J. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. New York: Springer-Verlag.
- HART, J. D. AND YI, S. (1996). One-sided Cross-validation. Preprint.
- HARTLEY, H. (1958). Maximum Likelihood Estimation from Incomplete Data. *Biometrics* 14, 174–194.

- HARTZEL, J., LIU, I., AND AGRESTI, A. (2000). Describing Heterogenous Effects in Stratified Ordinal Contingency Tables, with Applications to Multi-Center Clinical Trials. *Computational Statistics & Data Analysis* (to appear).
- HARVEY, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.
- HARVEY, A. C. AND FERNANDES, C. (1988). Time Series Models for Count or Qualitative Observations. *Journal of Business and Economic Statistics* 7, 407–422.
- HARVILLE, D. A. (1976). Extension of the Gauss-Markov Theorem to Include the Estimation of Random Effects. *Annals of Statistics* 4, 384–395.
- HARVILLE, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association* 72, 320–338.
- HARVILLE, D. A. AND MEE, R. W. (1984). A Mixed-model Procedure for Analyzing Ordered Categorical Data. *Biometrics* 40, 393–408.
- HASTIE, T. AND LOADER, C. (1993). Local Regression: Automatic Kernel Carapentry. *Statist. Sci.* 8, 120–143.
- HASTIE, T. AND TIBSHIRANI, R. (1990). *Generalized Additive Models*. London: Chapman & Hall.
- HASTIE, T. AND TIBSHIRANI, R. (1993). Varying-coefficient Models. *Journal of the Royal Statistical Society B* 55, 757–796.
- HASTIE, T. AND TIBSHIRANI, R. (2000). Bayesian Backfitting. *Statistical Science* 15(4).
- HAUSMAN, J., HALL, B. H., AND GRILICHES, Z. (1984). Econometric Models for Count Data with an Application to the Patents-R&D Relationship. *Econometrica* 52, 909–938.
- HAUSMAN, J. A. (1978). Specification Tests in Econometrics. *Econometrica* 46, 1251–1271.
- HAUSMAN, J. A. AND TAYLOR, W. E. (1981). A General Specification Test. *Economics Letters* 8, 239–245.
- HAUSMAN, J. A. AND WISE, D. A. (1978). A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preference. *Econometrica* 46, 403–426.
- HEAGERTY, P. AND ZEGER, S. (1998). Lorelogram: A Regression Approach to Exploring Dependence in Longitudinal Categorical Responses. *Journal of the American Statistical Association* 93(441), 150–162.
- HEAGERTY, P. J. AND ZEGER, S. L. (1996). Marginal Regression Models for Clustered Ordinal Measurements. *Journal of the American Statistical Association* 91, 1024–1036.
- HEBBEL, H. AND HEILER, S. (1987). Trend and Seasonal Decomposition in Discrete Time. *Statistical Papers* 28, 133–158.
- HECKMAN, J. J. (1981). *Dynamic Discrete Probability Models*, pp. 114–195.

- Cambridge, MA: MIT Press.
- HECKMAN, J. J. AND SINGER, B. (1984). A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models of Duration. *Econometrica* 52, 271–320.
- HEDEKER, D. AND GIBBONS, R. B. (1994). A Random-Effects Ordinal Regression Model for Multilevel Analysis. *Biometrics* 50, 933–944.
- VAN DER HEIJDEN, P., JANSEN, W., FRANCIS, B., AND SEEGER, G. U. H. (1992). *Statistical Modelling*. Amsterdam: North-Holland.
- HENNEVOGL, W. (1991). *Schätzung generalisierter Regressions- und Zeitreihenmodelle mit variierenden Parametern*. Ph.D. thesis, Universität Regensburg.
- HENNEVOGL, W. AND KRANERT, T. (1988). Residual and Influence Analysis for Multi Categorical Response Models. Regensburger Beiträge zur Statistik und Ökonometrie 5, Universität Regensburg.
- HEUMANN, C. (1997). *Likelihoodbasierte marginale Regressionsmodelle für korrelierte kategoriale Daten*. Frankfurt: Peter Lang.
- HINDE, J. (1982). Compound Poisson Regression Models. In R. Gilchrist (Ed.), *GLIM 1982 Internat. Conf. Generalized Linear Models*, New York, pp. 109–121. Springer-Verlag.
- HINDE, J. (1992). *Choosing Between Non-Nested Models: A Simulation Approach*, Volume 78, pp. 119–124. New York: Springer Lecture Notes in Statistics.
- HINDE, J. AND DÉMETRIO, C. (1998). Overdispersion: Models and Estimation. *Comp. Stat. & Data Analysis* 27, 151–170.
- HOAGLIN, D. AND WELSCH, R. (1978). The Hat Matrix in Regression and ANOVA. *American Statistician* 32, 17–22.
- HOBERT, J. P. AND CASELLA, G. (1996). The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models. *Journal of the American Statistical Association* 91, 1461–1473.
- HOCKING, R. R. (1976). The Analysis and Selection of Variables in Linear Regression. *Biometrics* 32, 1–49.
- HODGES, P. AND HALE, D. (1993). A Computational Method for Estimating Densities of Non-Gaussian Non-Stationary Univariate Time Series. *Journal of Time Series Analysis* 14, 163–178.
- HOERL, A. E. AND KENNARD, R. W. (1970). Ridge Regression: Bias Estimation for Nonorthogonal Problems. *Technometrics* 12, 55–67.
- HOLLAND, P. W. AND WELSCH, R. E. (1977). Robust Regression Using Iteratively Reweighted Least Squares. *Communications in Statistics A, Theory and Methods* 6, 813–827.
- HOLLY, A. (1982). A Remark on Hausman's Specification Test. *Econometrica* 50, 749–759.
- HOLMES, M. C. AND WILLIAMS, R. (1954). The Distribution of Carriers of Streptococcus Pyogenes among 2413 Healthy Children. *J. Hyg. Camb.* 52, 165–179.

- HOLTBRÜGGE, W. AND SCHUHMACHER, M. (1991). A Comparison of Regression Models for the Analysis of Ordered Categorical Data. *Applied Statistics* 40, 249–259.
- HOPPER, J. L. AND YOUNG, G. P. (1989). A Random Walk Model for Evaluating Clinical Trials Involving Serial Observations. *Statistics in Medicine* 7, 581–590.
- HOROWITZ, J. AND HÄRDLE, W. (1994). Testing a Parametric Model Against a Semiparametric Alternative. *Econometric Theory* 10, 821–848.
- HOROWITZ, J. L. (1998). *Semiparametric Methods in Econometrics*. New York: Springer-Verlag.
- HSIAO, C. (1986). *Analysis of Panel Data*. Cambridge: Cambridge University Press.
- HUBER, P. (1981). *Robust Statistics*. New York: John Wiley.
- HUFFER, F. W. AND MCKEAGUE, I. W. (1991). Weighted Least Squares Estimation for Aalen's Risk Model. *Journal of the American Statistical Association* 86, 114–129.
- HUNSDERGER, S. (1994). Semiparametric Regression in Likelihood Models. *Journal of the American Statistical Association* 89, 1354–1365.
- HURVICH, C. M., SIMONOFF, J. S., AND TSAI, C. (1998). Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion. *Journal of the Royal Statistical Society B* 60, 271–293.
- HÜRZELER, M. (1998). *Statistical Methods for General State Space Models*. Ph.D. thesis, ETH Zürich. Nr. 12674.
- HÜRZELER, M. AND KÜNSCH, H. (1998). Monte Carlo Approximations for General State Space Models. *J. Comp. and Graph. Statist.* 7, 175–193.
- HUTCHINSON, C. E. (1984). The Kalman Filter Applied to Aerospace and Electronic Systems. *IEEE Transactions Aero. Elect. Systems AES-20*, 500–504.
- IBRAHIM, J. AND KLEINMAN, K. (1998). Semiparametric Bayesian Methods for Random Effects Models. In D. Dey, P. Müller & D. Sinha (Eds.), *Practical Nonparametric and Semiparametric Bayesian Statistics*, Number 133 in Lecture Notes in Statistics, Chapter 5, pp. 89–114. New York: Springer-Verlag.
- IM, S. AND GIANOLA, D. (1988). Mixed Models for Binomial Data with an Application to Lamb Mortality. *Applied Statistics* 37, 196–204.
- JACOBS, P. A. AND LEWIS, P. A. W. (1983). Stationary Discrete Autoregressive Moving-Average Time Series Generated by Mixtures. *J. Time Series Analysis* 4, 19–36.
- JANSEN, J. (1990). On the Statistical Analysis of Ordinal Data when Extravariation is Present. *Applied Statistics* 39, 74–85.
- JAZWINSKI, A. (1970). *Stochastic Processes and Filtering*. New York: Academic Press.
- JOHNSON, V. AND ALBERT, J. (1999). *Ordinal Data Modelling*. New York: Springer-Verlag.

- JONES, R. H. (1993). *Longitudinal Data with Serial Correlation: A State-Space Approach*. London: Chapman & Hall.
- JORGENSEN, B. (1982). *Statistical Properties of the Generalized Inverse Gaussian Distribution*. Heidelberg: Springer Lecture Notes.
- JORGENSEN, B. (1992). Exponential Dispersion Models and Extension: A Review. *International Statistical Review* 60, 5–20.
- JORGENSEN, B. (1997). *The Theory of Dispersion Models*. London: Chapman & Hall.
- KALBFLEISCH, J. AND PRENTICE, R. (1973). Marginal Likelihoods Based on Cox's Regression and Life Model. *Biometrika* 60, 256–278.
- KALBFLEISCH, J. AND PRENTICE, R. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.
- KASHIWAGI, N. AND YANAGIMOTO, T. (1992). Smoothing Serial Count Data through a State Space Model. *Biometrics* 48, 1187–1194.
- KAUERMANN, G. (2000). Modelling Longitudinal Data with Ordinal Response by Varying Coefficients. *Biometrics* 56, 692–698.
- KAUERMANN, G. AND TUTZ, G. (1999). On Model Diagnostics and Bootstrapping in Varying Coefficient Models. *Biometrika* 86, 119–128.
- KAUERMANN, G. AND TUTZ, G. (2000a). Local Likelihood Estimation in Varying Coefficient Models Including Additive Bias Correction. *Journal of Nonparametric Statistics* 12, 343–371.
- KAUERMANN, G. AND TUTZ, G. (2000b). Testing Generalized and Semiparametric Models against Smooth Alternatives. *J. Roy. Stat. Soc. B* (to appear).
- KAUFMANN, H. (1987). Regression Models for Nonstationary Categorical Time Series: Asymptotic Estimation Theory. *The Annals of Statistics* 15, 79–98.
- KAUFMANN, H. (1988). On Existence and Uniqueness of Maximum Likelihood Estimates in Quantal and Ordinal Response Models. *Metrika* 35, 291–313.
- KAUFMANN, H. (1989). On Likelihood Ratio Test for and against Convex Cones in the Linear Model. Regensburger Beiträge zu Statistik und Ökonometrie 17, Universität Regensburg.
- KIRCHEN, A. (1988). Schätzung zeitveränderlicher Strukturparameter in ökonometrischen Prognosemodellen. Mathematical Systems in Economics.
- KITAGAWA, G. (1987). Non-Gaussian State-Space Modelling of Nonstationary Time Series (with Comments). *Journal of the American Statistical Association* 82, 1032–1063.
- KITAGAWA, G. (1996). Monte Carlo Filter and Smoother for Non-Gaussian Non-linear State Space Models. *J. Comp. and Graph. Statistics* 5, 1–25.
- KITAGAWA, G. AND GERSCH, W. (1984). Smoothness Priors, State Space Modelling of Time Series with Trend and Seasonality. *Journal of the American Statistical Association* 79, 378–389.
- KITAGAWA, G. AND GERSCH, W. (1996). *Smoothness Priors Analysis of Time Series*. Lecture Notes in Statistics 116. New York: Springer-Verlag.

- KLINGER, A. (1998). *Hochdimensionale Generalisierte Lineare Modelle*. Ph.D. thesis, LMU München. Shaker Verlag, Aachen.
- KLINGER, A., DANNEGGER, F., AND ULM, K. (2000). Identifying and Modelling Prognostic Factors with Censored Data. *Statistics in Medicine* 19, 601–615.
- KNORR HELD, L. (1997). *Hierarchical Modelling of Discrete Longitudinal Data; Applications of Markov Chain Monte Carlo*. München: Utz.
- KNORR HELD, L. (1999). Conditional Prior Proposals in Dynamic Models. *Scandinavian Journal of Statistics* 26, 129–144.
- KNORR HELD, L. (2000). Bayesian Modelling of Inseparable Space-Time Variation in Disease Risk. *Statistics in Medicine* (to appear).
- KNORR HELD, L. AND BESAG, J. (1998). Modelling Risk from a Disease in Time and Space. *Statistics in Medicine* 17, 2045–2060.
- KOEHLER, J. AND OWEN, A. (1996). Computer Experiments. In S. Gosh & C. R. Rao (Eds.), *Handbook of Statistics* 13, pp. 261–308. Amsterdam: Elsevier Science.
- KOHN, R. AND ANSLEY, C. (1987). A New Algorithm for Spline Smoothing Based on Smoothing a Stochastic Process. *SIAM Journal Sci. Stat. Comp.* 8, 33–48.
- KOHN, R. AND ANSLEY, C. (1989). A Fast Algorithm for Signal Extraction, Influence and Cross-Validation in State-Space Models. *Biometrika* 76, 65–79.
- KÖNIG, H., NERLOVE, M., AND OUDIZ, G. (1981). On the Formation of Price Expectations, an Analysis of Business Test Data by Log-Linear Probability Models. *European Economic Review* 16, 421–433.
- KOOPMAN, S. J. (1993). Disturbance Smoother for State Space Models. *Biometrika* 80, 117–126.
- KRÄMER, W. (1986). Bemerkungen zum Hausman-Test. *Allgemeines Statistisches Archiv* 70, 170–179.
- KRANTZ, D. H., LUCE, R. D., SUPPES, P., AND TVERSKY, A. (1971). *Foundations of Measurement*, Volume 1. New York: Academic Press.
- KREDLER, C. (1984). Selection of Variables in Certain Nonlinear Regression Models. *Comp. Stat. Quarterly* 1, 13–27.
- KÜCHENHOFF, H. AND ULM, K. (1997). Comparison of Statistical Methods for Assessing Threshold Limiting Values in Occupational Epidemiology. *Comput. Statist.* 12, 249–264.
- KULLBACK, S. (1959). *Information Theory and Statistics*. New York: Wiley.
- KULLBACK, S. (1985). Minimum Discrimination Information (MDI) Estimation. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*, pp. 527–529. New York: Wiley.
- KÜNSCH, H. R. (1987). Intrinsic Autoregressions and Related Models on the Two-Dimensional Lattice. *Biometrika* 74, 517–524.
- KÜNSCH, H. R. (2000). State Space and Hidden Markov Models. In O. Barndorff-Nielsen, D. Cox & C. Klüppelberg (Eds.), *Complex Stochastic Systems*.

- London: Chapman & Hall.
- KÜSTERS, U. (1987). *Hierarchische Mittelwert- und Kovarianzstrukturmodelle mit nichtmetrischen endogenen Variablen*. Ph.D. thesis, Bergische Universität Wuppertal.
- LÄÄRÄ, E. AND MATTHEWS, J. N. (1985). The Equivalence of Two Models for Ordinal Data. *Biometrika* 72, 206–207.
- LAIRD, N. M. (1978). Empirical Bayes Methods for Two-Way Contingency Tables. *Biometrika* 65, 581–590.
- LAIRD, N. M., BECK, G. J., AND WARE, J. H. (1984). Mixed Models for Serial Categorical Response. Quoted in A. Eckholm (1991). Maximum Likelihood for Many Short Binary Time Series (preprint).
- LAIRD, N. M. AND WARE, J. H. (1982). Random Effects Models for Longitudinal Data. *Biometrics* 38, 963–974.
- LANCASTER, T. (1990). *The Econometric Analysis of Transition Data*. Cambridge: Cambridge University Press.
- LANDWEHR, J. M., PREGIBON, D., AND SHOEMAKER, A. C. (1984). Graphical Methods for Assessing Logistic Regression Models. *Journal of the American Statistical Association* 79, 61–71.
- LANG, J. B. (1996). Maximum Likelihood Methods for a Generalized Class of Log-Linear Models. *The Annals of Statistics* 24, 726–752.
- LANG, J. B. AND AGRESTI, A. (1994). Simultaneous Modelling Joint and Marginal Distributions of Multivariate Categorical Responses. *Journal of the American Statistical Assoc.* 89, 625–632.
- LANG, S. AND BREZGER, A. (2000). Bayesian P-Splines. Technical Report, Proc. International Workshop on Statistical Modelling, Bilbao.
- LAURITZEN, S. (1981). Time Series Analysis in 1880: A Discussion of Contributions Made by T.N. Thiele. *Int. Statist. Review* 49, 319–331.
- LAURITZEN, S. (1998). *Graphical Models*. New York: Oxford University.
- LAURITZEN, S. L. AND WERMUTH, N. (1989). Graphical Models for Associations Between Variables, Some of Which Are Qualitative and Some Quantitative. *Annals of Statistics* 17, 31–57. (Corr: V.17 p. 1916).
- LAWLESS, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. New York: Wiley.
- LAWLESS, J. F. (1987). Negative Binomial and Mixed Poisson Regression. *The Canadian Journal of Statistics* 15, 209–225.
- LAWLESS, J. F. AND SINGHAL, K. (1978). Efficient Screening of Nonnormal Regression Models. *Biometrics* 34, 318–327.
- LAWLESS, J. F. AND SINGHAL, K. (1987). ISMOD: An All-Subsets Regression Program for Generalized Linear Models. *Computer Methods and Programs in Biomedicine* 24, 117–134.
- LEE, A. H. (1988). Assessing Partial Influence in Generalized Linear Models. *Biometrics* 44, 71–77.

- LEE, Y. AND NELDER, J. A. (1996). Hierarchical Generalized Linear Models. *Journal of the Royal Statistical Society B* 58, 619–678.
- LEONARD, T. (1972). Bayesian Methods for Binomial Data. *Biometrika* 59, 869–874.
- LEONARD, T. AND NOVICK, M. R. (1986). Bayesian Full Rank Marginalization for Two-Way Contingency Tables. *Journal of Educational Statistics* 11, 33–56.
- LERMAN, S. AND MANSKI, C. (1981). On the Use of Simulated Frequencies to Approximate Choice Probabilities. In C. Manski & D. Fadden (Eds.), *Structural Analysis of Discrete Data*. Cambridge, MA: MIT Press.
- LESAFFRE, E. AND ALBERT, A. (1989). Multiple-group Logistic Regression Diagnostics. *Applied Statistics* 38, 425–440.
- LESAFFRE, E., MOLENBERGS, G., AND DEWULF, L. (1996). Effect of Dropouts in a Longitudinal Study: An Application of a Repeated Ordinal Model. *Statistics in Medicine* 15, 1123–1141.
- LEVINE, D. (1983). A Remark on Serial Correlation in Maximum Likelihood. *Journal of Econometrics* 23, 337–342.
- LI, G. AND DOSS, H. (1995). An Approach to Nonparametric Regression for Life History Data Using Local Linear Fitting. *Annals of Statistics* 23, 787–823.
- LI, K. C. AND DUAN, N. (1989). Regression Analysis under Link Violation. *The Annals of Statistics* 17, 1009–1052.
- LIANG, K.-Y. AND McCULLAGH, P. (1993). Case Studies in Binary Dispersion. *Biometrics* 49, 623–630.
- LIANG, K.-Y. AND ZEGER, S. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* 73, 13–22.
- LIANG, K.-Y., ZEGER, S. L., AND QAQISH, B. (1992). Multivariate Regression Analysis for Categorical Data (with Discussion). *Journal of the Royal Statistical Society B* 54, 3–40.
- LIN, X. AND BRESLOW, N. E. (1996). Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion. *Journal of the American Statistical Association* 91, 1007–1016.
- LIN, X. AND ZHANG, D. (1999). Inference in Generalized Additive Mixed Models by Using Smoothing Splines. *Journal of the Royal Statistical Society B* 61, 381–400.
- LINDSEY, J. J. (1993). *Models for Repeated Measurements*. Oxford: Oxford University Press.
- LINDSTROM, M. AND BATES, D. (1990). Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics* 46, 673–687.
- LINTON, O. B. AND HÄRDLE, W. (1996). Estimation of Additive Regression Models with Known Links. *Biometrika* 83, 529–540.
- LIPSITZ, S., LAIRD, N., AND HARRINGTON, D. (1991). Generalized Estimation Equations for Correlated Binary Data: Using the Odds Ratio as a Measure of Association in Unbalanced Mixed Models with Nested Random Effects.

- Biometrika* 78, 153–160.
- LIU, Q. AND PIERCE, D. A. (1994). A Note on Gauss-Hermite Quadrature. *Biometrika* 81, 624–629.
- LOADER, C. (1999). *Local Regression and Likelihood*. New York: Springer-Verlag.
- LOADER, C. R. (1995). Old Faithful Erupts: Bandwidth Selection Reviewed. Preprint.
- LONGFORD, N. L. (1993). *Random Effect Models*. New York: Oxford University Press.
- LONGFORD, N. T. (1987). A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models with Nested Random Effects. *Biometrika* 74, 817–827.
- LOS, C. (1984). *Econometrics of Models with Evolutionary Parameter Structures*. Ph.D. thesis, Columbia University, NY.
- LOUIS, T. A. (1982). Finding the Observed Information Matrix When Using the EM Algorithm. *Journal of the Royal Statistical Society B* 44, 226–233.
- MACDONALD, I. AND ZUCCHINI, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*. London: Chapman & Hall.
- MACK, Y. P. (1981). Local Properties of k-NN Regression Estimates. *SIAM J. Alg. Disc. Meth.* 2, 311–323.
- MADDALA, G. S. (1983). *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.
- MAGNUS, J. R. AND NEUDECKER, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. London: John Wiley.
- MALLAT, S. G. (1989). Multiresolution Approximations and Wavelet Orthonormal Basis of  $L^2(\mathbb{R})$ . *Transaction of the American Mathematical Society* 315, 69–88.
- MALLICK, B., DENISON, D., AND SMITH, A. (1999). Semiparametric Generalized Linear Models: Bayesian Approaches. In D. Dey, S. Ghosh & B. Mallick (Eds.), *Generalized Linear Models: A Bayesian Perspective*. New York: Marcel Dekker.
- MAMMEN, E. AND VAN DE GEER, S. (1997). Locally Adaptive Regression Splines. *Annals of Statistics* 25, 387–413.
- MANTEL, N. AND HANKEY, B. F. (1978). A Logistic Regression Analysis of Response Time Data Where the Hazard Function Is Time Dependent. *Communication in Statistics, Theory & Methods A* 7, 333–347.
- MARTIN, R. (1979). Approximate Conditional-Mean Type Smoothers and Interpolators. In T. Gasser & M. Rosenblatt (Eds.), *Smoothing Techniques for Curve Estimation*. Berlin: Springer-Verlag.
- MARTIN, R. AND RAFTERY, A. (1987). Robustness, Computation and Non-Euclidian Models (Comment). *Journal of the American Statistical Association* 82, 1044–1050.
- MARX, B., EILERS, P., AND SMITH, E. (1992). Ridge Likelihood Estimation for Generalized Linear Regression. In R. van der Heijden, W. Jansen, B. Fran-

- cis & G. Seeber (Eds.), *Statistical Modelling*, pp. 227–238. Amsterdam: North-Holland.
- MARX, B. AND FRIEDL, H. (Eds.) (1998). *Statistical Modelling*, Proceedings of the 13th International Workshop on Statistical Modelling, New Orleans, 1998.
- MARX, B. AND SMITH, E. (1990). Principal Component Estimation for Generalized Linear Regression. *Biometrika* 77, 23–31.
- MARX, D. B. AND EILERS, P. (1998). Direct Generalized Additive Modelling with Penalized Likelihood. *Comp. Stat. & Data Analysis* 28, 193–209.
- MASTERS, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika* 47, 149–174.
- MCCULLAGH, P. (1980). Regression Model for Ordinal Data (with Discussion). *Journal of the Royal Statistical Society B* 42, 109–127.
- MCCULLAGH, P. (1983). Quasi-Likelihood Functions. *Annals of Statistics* 11, 59–67.
- MCCULLAGH, P. AND NELDER, J. A. (1989). *Generalized Linear Models* (2nd ed.). New York: Chapman & Hall.
- MCCULLOCH, C. E. (1994). Maximum Likelihood Variance Components Estimation for Binary Data. *J. Am. Statist. Assoc.* 89, 330–335.
- MCCULLOCH, C. E. (1997). Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association* 92, 162–170.
- McDONALD, B. W. (1993). Estimating Logistic Regression Parameters for Bivariate Binary Data. *Journal of the Royal Statistical Society B* 55, 391–397.
- MFADDEN, D. (1973). Conditional Logit Analysis of Qualitative Choice Behaviour. In P. Zarembka (Ed.), *Frontiers in Econometrics*. New York: Academic Press.
- MFADDEN, D. (1981). Econometric Models of Probabilistic Choice. In C. F. Manski & D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, pp. 198–272. Cambridge, MA: MIT Press.
- MFADDEN, D. (1984). Qualitative Response Models. In Z. Grilicher & M. D. Intrilligator (Eds.), *Handbook of Econometrics*. Cambridge, MA: MIT Press.
- MCKINNON, J. G. (1983). Model Specification Tests Against Non-nested Alternatives. *Econometric Reviews* 2, 85–110.
- MEHTA, C. R., PATEL, N. R., AND TSIATIS, A. A. (1984). Exact Significance Testing to Establish Treatment Equivalence with Ordered Categorical Data. *Biometrics* 40, 819–825.
- MEILIJSSEN, I. (1989). A Fast Improvement to the EM-Algorithm on Its Own Terms. *Journal of the Royal Statistical Society B* 51, 127–138.
- MENGERSEN, K., ROBERT, C., AND GUIHENNEUC-JOUYAU, C. (1998). MCMC Convergence Diagnostics: A Review. In J. Bernardo, O. Berger, A. Dawid & A. F. M. Smith (Eds.), *Bayesian Statistics 6*, pp. 415–440. Oxford: Oxford University Press.

- MILLER, A. J. (1984). Selection of Subsets and Regression Variables (with Discussion). *Journal of the Royal Statistical Society A* 147, 389–429.
- MILLER, A. J. (1989). *Subset Selection in Regression*. London: Chapman & Hall.
- MILLER, M. E., DAVIS, C. S., AND LANDIS, R. J. (1993). The Analysis of Longitudinal Polytomous Data: Generalized Estimated Equations and Connections with Weighted Least Squares. *Biometrics* 49, 1033–1044.
- MINDER, C. E. AND FRIEDL, H. (1997). Good Statistical Practice. In *Proceedings of the 12th International Workshop on Statistical Modelling*, Volume 5 of *Schrifreihe der Österreichischen Statistischen Gesellschaft*, Wien.
- MOLENBERGHS, G. AND LESAFFRE, E. (1992). Marginal Modelling of Correlated Ordinal Data Using an n-way Plackett Distribution. *Advances in GLIM and Statistical Modelling. Springer Lecture Notes in Statistics* 78, 139–144.
- MOLENBERGHS, G. AND LESAFFRE, E. (1994). Marginal Modelling of Correlated Ordinal Data Using a Multivariate Plackett Distribution. *Journal of the American Statistical Association* 89, 633–644.
- MOLENBERGHS, G. AND LESAFFRE, E. (1999). Marginal Modeling of Multivariate Categorical Data. *Statistics in Medicine* 18, 2237–2255.
- MOORE, D. F. (1987). Modelling the Extraneous Variance in the Presence of Extra-binomial Variation. *Appl. Statist.* 36, 8–14.
- MORAWITZ, G. AND TUTZ, G. (1990). Alternative Parametrizations in Business Tendency Surveys. *ZOR-Methods and Models of Operations Research* 34, 143–156.
- MORGAN, B. J. T. (1985). The Cubic Logistic Model for Quantal Assay Data. *Applied Statistics* 34, 105–113.
- MORRIS, C. N. (1982). Natural Exponential Families with Quadratic Variance Functions. *Annals of Statistics* 10, 65–80.
- MOULTON, L. AND ZEGER, S. (1989). Analysing Repeated Measures in Generalized Linear Models via the Bootstrap. *Biometrics* 45, 381–394.
- MOULTON, L. AND ZEGER, S. (1991). Bootstrapping Generalized Linear Models. *Computational Statistics and Data Analysis* 11, 53–63.
- MUKHOPADHYAY, S. AND GELFAND, A. E. (1997). Dirichlet Process Mixed Generalized Linear Models. *Journal of the American Statistical Association* 92, 633–639.
- MÜLLER, H.-G. (1984). Smooth Optimum Kernel Estimators of Densities, Regression Curves and Modes. *The Annals of Statistics* 12, 766–774.
- MÜLLER, H.-G. (1991). Smooth Optimum Kernel Estimators near Endpoints. *Biometrika* 78(3), 521–530.
- MÜLLER, H.-G. AND STADTMÜLLER, U. (1987). Estimation of Heteroscedasticity in Regression Analysis. *Annals of Statistics* 15, 221–232.
- MÜLLER, P., ERKANLI, A., AND WEST, M. (1996). Bayesian Curve Fitting Using Multivariate Normal Mixtures. *Biometrika* 83, 67–79.
- MÜLLER, P. AND PARMIGIANI, G. (1995). Numerical Evaluation of Information Theoretic Measures. In D. A. Berry, K. M. Chaloner & J. F. Geweke (Eds.),

- Bayesian Statistics and Econometrics: Essays in Honor of A. Zellner.* New York: John Wiley.
- MUTHEN, B. (1984). A General Structural Equation Model with Dichotomous Categorical and Continuous Latent Variable Indicators. *Psychometrika* 49, 115–132.
- NADARAYA, E. A. (1964). On Estimating Regression. *Theory Prob. Appl.* 10, 186–190.
- NASON, G. AND SILVERMAN, B. (2000). Wavelets for Regression and Other Statistical Problems. In M. G. Schimek (Ed.), *Smoothing and Regression: Approaches, Computation and Application*. New York: Wiley.
- NAYLOR, J. C. AND SMITH, A. F. M. (1982). Applications of a Method for the Efficient Computation of Posterior Distributions. *Applied Statistics* 31, 214–225.
- NELDER, J. A. (1992). Joint Modelling of Mean and Dispersion. In P. van der Heijden, W. Jansen, B. Francis & G. Seeber (Eds.), *Statistical Modelling*. Amsterdam: North-Holland.
- NELDER, J. A. AND PREGIBON, D. (1987). An Extended Quasi-Likelihood Function. *Biometrika* 74, 221–232.
- NELDER, J. A. AND WEDDERBURN, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society A* 135, 370–384.
- NERLOVE, M. (1983). Expectations, Plans and Realizations in Theory and Practice. *Econometrica* 51, 1251–1279.
- NEUHAUS, J., HAUCK, W., AND KALBFLEISCH, J. (1991). The Effects of Mixture Distribution. Misspecification When Fitting Mixed Effect Logistic Models. Technical Report 16, Dept. of Epidemiology and Biostatistics University of California, San Francisco.
- NEUHAUS, J. M., KALBFLEISCH, J. D., AND HAUCK, W. W. (1991). A Comparison of Cluster-Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data. *International Statistical Review* 59, 25–35.
- NEWTON, M. A., CZADO, C., AND CHAPPELL, R. (1996). Semiparametric Bayesian Inference for Binary Regression. *Journal of the American Statistical Association* 91, 142–153.
- NEYMAN, J. (1949). Contributions to the Theory of the Chi-square-test. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley.
- NYQUIST, H. (1990). Restricted Estimation of Generalized Linear Models. *Applied Statistics* 40, 133–141.
- OGDEN, R. T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*. Boston: Birkhäuser.
- OPSOMER, J. D. (2000). Asymptotic Properties of Backfitting Estimators. *Journal of Multivariate Analysis* 73, 166–179.
- OPSOMER, J. D. AND RUPPERT, D. (1997). Fitting a Bivariate Additive Model by Local Polynomial Regression. *Annals of Statistics* 25, 186–211.

- Osius, G. and Rojek, D. (1992). Normal Goodness-of-Fit Tests for Parametric Multinomial Models with Large Degrees of Freedom. *Journal of the American Statistical Association* 87, 1145–1152.
- O'Sullivan, F., Yandell, B. S., and Raynor, W. J. (1986). Automatic Smoothing of Regression Functions in Generalized Linear Models. *Journal of the American Statistical Association* 81, 96–103.
- Parr, W. C. (1981). Minimum Distance Estimation. A Bibliography. *Communications in Statistics, Theory and Methods* 10, 1205–1224.
- Pauly, R. (1989). A General Structural Model for Decomposing Time Series and Its Analysis as a Generalized Regression Model. *Statistical Papers* 30, 245–261.
- Pendergast, J. F., Gange, S., Newton, M., and Lindstrom, M. (1996). A Survey of Methods for Analyzing Clustered Binary Response Data. *International Statistical Review* 64, 89–118.
- Pepe, M. S. and Anderson, G. L. (1994). A Cautionary Note on Inference for Marginal Regression Models with Longitudinal Data and General Correlated Response Data. *Communications in Statistics, Part B-Simulation and Computation* 23, 939–951.
- Piegorsch, W. (1992). Complementary Log Regression for Generalized Linear Models. *The American Statistician* 46, 94–99.
- Piegorsch, W. W., Weinberg, C. R., and Margolin, B. H. (1988). Exploring Simple Independent Action in Multifactor Tables of Proportions. *Biometrics* 44, 595–603.
- Pierce, D. A. and Schafer, D. W. (1986). Residuals in Generalized Linear Models. *Journal of the American Statistical Association* 81, 977–986.
- Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the Log-likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics* 4, 12–35.
- Pitt, M. and Shephard, N. (1999). Filtering via Simulation: Auxiliary Particle Filters. *Journal of the American Statistical Association* 94, 590–599.
- Poortema, K. L. (1999). On Modelling Overdispersion of Counts. *Statistica Neerlandica* 53, 5–20.
- PRAKASA RAO, B. L. S. (1987). *Asymptotic Theory of Statistical Inference*. New York: Wiley.
- PREGIBON, D. (1980). Goodness of Link Tests for Generalized Linear Models. *Applied Statistics* 29, 15–24.
- PREGIBON, D. (1981). Logistic Regression Diagnostics. *The Annals of Statistics* 9, 705–724.
- PREGIBON, D. (1982). Resistant Fits for Some Commonly Used Logistic Models with Medical Applications. *Biometrics* 38, 485–498.
- PREGIBON, D. (1984). Review of Generalized Linear Models by McCullagh and Nelder. *American Statistician* 12, 1589–1596.
- PREISLER, H. K. (1989). Analysis of a Toxicological Experiment Using a Gener-

- alized Linear Model with Nested Random Effects. *International Statistical Review* 57, 145–159.
- PRENTICE, R. L. (1976). A Generalization of the Probit and Logit Methods for Close Response Curves. *Biometrics* 32, 761–768.
- PRENTICE, R. L. (1988). Correlated Binary Regression with Covariates Specific to Each Binary Observation. *Biometrics* 44, 1033–1084.
- PRENTICE, R. L. AND SELF, S. G. (1983). Asymptotic Distribution Theory for Cox-Type Regression Models with General Relative Risk Form. *Annals of Statistics* 11, 804–813.
- PRIESTLEY, M. B. AND CHAO, M. T. (1972). Nonparametric Function Fitting. *Journal of the Royal Statistical Society B* 34, 385–392.
- PRUSCHA, H. (1993). Categorical Time Series with a Recursive Scheme and with Covariates. *Statistics* 24, 43–57.
- PUDNEY, S. (1989). *Modelling Individual Choice. The Econometrics of Corners, Kinks and Holes*. London: Basil Blackwell.
- QU, Y., WILLIAMS, G. W., BECK, G. J., AND GOORMASTIC, M. (1987). A Generalized Model of Logistic Regression for Clustered Data. *Communications in Statistics, A Theory and Methods* 16, 3447–3476.
- QUINTANA, R., LIA, J., AND DEL PINO, G. (1999). Monte Carlo EM with Importance Reweighting and Its Application in Random Effect Models. *Computational Statistics & Data Analysis* 29, 429–444.
- RAFTERY, A. E. (1996). Hypothesis Testing and Model Selection. In W. G. et al. (Ed.), *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.
- RANDALL, J. H. (1989). The Analysis of Sensory Data by Generalized Linear Models. *Biom. Journal* 31, 781–793.
- RAO, C. AND KLEFFE, J. (1988). *Estimation of Variance Components and Applications*. Amsterdam: North-Holland.
- RASCH, G. (1961). On General Laws and the Meaning of Measurement in Psychology. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley.
- READ, I. AND CRESSIE, N. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.
- REDNER, R. A. AND WALKER, H. F. (1984). Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review* 26, 195–239.
- REINSCH, C. (1967). Smoothing by Spline Functions. *Numerische Mathematik* 10, 177–183.
- RICE, J. A. (1984). Bandwidth Choice for Nonparametric Regression. *Annals of Statistics* 12, 1215–1230.
- RIPLEY, B. (1987). *Stochastic Simulation*. New York: Wiley.
- ROBERTS, F. (1979). *Measurement Theory*. Reading, MA: Addison-Wesley.
- ROBIN, J. M., ROTNITZKY, A. G., AND ZHAO, L. P. (1995). Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of

- Missing Data. *Journal of the American Statistical Association* 90, 106–120.
- ROJEK, D. (1989). *Asympototik für Anpassungstests in Produkt- Multinomial-Modellen bei wachsendem Freiheitsgrad*. Ph.D. thesis, Universität Bremen.
- RONNING, G. (1980). Logit, Tobit and Markov Chains. Three Different Approaches to the Analysis of Aggregate Tendency Data. In W. H. Strigel (Ed.), *Business Cycle Analysis*. Westmead: Farnborough.
- RONNING, G. (1987). The Informational Content of Responses from Business Surveys. Diskussionsbeitrag 961s, Universität Konstanz.
- RONNING, G. (1991). *MikroÖkonometrie*. Berlin: Springer-Verlag.
- RONNING, G. AND JUNG, R. (1992). Estimation of a First Order Autoregressive Process with Poisson Marginals for Count Data. In L. Fahrmeir et al. (Eds.), *Advances in GLIM and Statistical Modelling, Lecture Notes in Statistics*. Berlin: Springer-Verlag.
- ROSENBERG, B. (1973). Random Coefficient Models. *Annals of Economic and Social Measurement* 4, 399–418.
- ROSNER, B. (1984). Multivariate Methods in Ophthalmology with Applications to Other Paired-Data Situations. *Biometrics* 40, 1025–1035.
- RUE, H. (2000). Fast Sampling of Gaussian Markov Random Fields with Applications. Available under <http://www.math.ntnu.no/preprint/statistics/2000/s1-2000.ps>.
- SAGE, A. AND MELSA, J. (1971). *Estimation Theory, with Applications to Communications and Control*. New York: McGraw-Hill.
- SANTNER, T. J. AND DUFFY, D. E. (1989). *The Statistical Analysis of Discrete Data*. New York: Springer-Verlag.
- SCALLAN, A., GILCHRIST, R., AND GREEN, M. (1984). Fitting Parametric Link Functions in Generalised Linear Models. *Comp. Statistics and Data Analysis* 2, 37–49.
- SCHALL, R. (1991). Estimation in Generalised Linear Models with Random Effects. *Biometrika* 78, 719–727.
- SCHLICHT, E. (1981). A Seasonal Adjustment Principle and Seasonal Adjustment Method Derived from this Principle. *Journal of the American Statistical Association* 76, 374–378.
- SCHNATTER, S. (1992). Integration-Based Kalman-Filtering for a Dynamic Generalized Linear Trend Model. *Comp. Statistics and Data Analysis* 13, 447–459.
- SCHNEIDER, W. (1986). *Der Kalmanfilter als Instrument zur Diagnose und Schätzung variabler Parameterstrukturen in Ökonometrischen Modellen*. Heidelberg: Physica.
- SCHUMAKER, L. (1993). *Spline Functions: Basic Theory*. Malabar, FL: Krieger Publishing Co.
- SCHWARZ, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics* 6, 461–464.
- SEARLE, S., CASELLA, G., AND McCULLOCH, C. (1992). *Variance Components*.

New York: Wiley.

- SEEBER, G. (1977). *Linear Regression Analysis*. New York: Wiley.
- SEEBER, G. U. H. (1989). *Statistisches Modellieren in Exponentialfamilien*. Innsbruck: Unpublished.
- SEEBER, G. U. H., FRANCIS, B. J., HATZINGER, R., AND STECKEL-BERGER, G. (1995). *Statistical Modelling*. New York: Springer-Verlag.
- SEGERSTEDT, B. (1992). On Ordinary Ridge Regression in Generalized Linear Models. *Commun. Statist. – Theory Meth.* 21, 2227–2246.
- SEIFERT, B. AND GASSER, T. (1996). Variance Properties of Local Polynomials and Ensuring Modifications. In W. Härdle & M. Schimek (Eds.), *Statistical Theory and Computational Aspects of Smoothing*. Heidelberg: Physica-Verlag.
- SEVERINI, T. A. AND STANISWALIS, J. G. (1994). Quasi-likelihood Estimation in Semiparametric Models. *Journal of the American Statistical Association* 89, 501–511.
- SHAO, J. AND TU, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.
- SHAW, J. E. H. (1988). Aspects of Numerical Integration and Summarisation. *Bayesian Statistics 3*, 411–428.
- SHEPHARD, N. (1994). Partial Non-Gaussian State Space. *Biometrika* 81, 115–131.
- SHEPHARD, N. AND PITT, M. K. (1997). Likelihood Analysis of Non-Gaussian Measurement Timeseries. *Biometrika* 84, 653–657.
- SILVAPULLE, M. J. (1981). On the Existence of Maximum Likelihood Estimates for the Binomial Response Models. *Journal of the Royal Statistical Society B* 43, 310–313.
- SILVERMAN, B. W. (1984). Spline Smoothing: the Equivalent Variable Kernel Method. *Annals of Statistics* 12, 898–916.
- SIMONOFF, J. S. (1996). *Smoothing Methods in Statistics*. New York: Springer-Verlag.
- SIMONOFF, J. S. AND TSAI, C. (1991). Assessing the Influence of Individual Observation on a Goodness-of-Fit Test Based on Nonparametric Regression. *Statistics and Probability Letters* 12, 9–17.
- SIMONOFF, J. S. AND TUTZ, G. (2000). Smoothing Methods for Discrete Data. In M. Schimek (Ed.), *Smoothing and Regression. Approaches, Computation and Application*. New York: Wiley.
- SINGH, A. C. AND ROBERTS, G. R. (1992). State Space Modelling of Cross-Classified Time Series of Counts. *International Statistical Review* 60, 321–336.
- SMALL, K. A. (1987). A Discrete Choice Model for Ordered Alternatives. *Econometrica* 55, 409–424.
- SMITH, A. AND WEST, M. (1983). Monitoring Renal Transplants: An Application of the Multi-Process Kalman Filter. *Biometrics* 39, 867–878.

- SMITH, A. F. M., SKENE, A. M., SHAW, J. E. H., NAYLOR, J. C., AND DRANSFIELD, M. (1985). The Implementation of the Bayesian Paradigm. *Communication in Statistics, Theory & Methods* 14, 1079–1102.
- SMITH, M. AND KOHN, R. (1996). Nonparametric Regression Using Bayesian Variable Selection. *Journal of Econometrics* 75, 317–343.
- SMITH, R. AND MILLER, J. (1986). A Non-Gaussian State Space Model and Application to Prediction of Records. *Journal of the Royal Statistical Society B* 48, 79–88.
- SONG, P. (2000). Multivariate Dispersion Models Generated from Gaussian Copula. *Scandinavian Journal of Statistics* 27, 305–320.
- SPECKMAN, P. (1988). Kernel Smoothing in Partial Linear Models. *Journal of the Royal Statistical Society B* 50, 413–436.
- SPIEGELHALTER, D. J., BEST, N., AND CARLIN, B. (1998). Bayesian Deviance, the Effective Number of Parameters and the Comparison of Arbitrarily Complex Models. Technical Report 98-009, Div. Biostatistics, University of Minnesota.
- STANISWALIS, J. G. (1989). Local Bandwidth Selection for Kernel Estimates. *Journal of the American Statistical Association* 84, 284–288.
- STANISWALIS, J. G. AND SEVERINI, T. A. (1991). Diagnostics for Assessing Regression Models. *Journal of the American Statistical Association* 86, 684–692.
- STIRATELLI, R., LAIRD, N., AND WARE, J. H. (1984). Random-Effects Models for Serial Observation with Binary Response. *Biometrics* 40, 961–971.
- STONE, C., HANSEN, M., KOOPERBERG, C., AND TRUONG, Y. (1997). Polynomial Splines and Their Tensor Products in Extended Linear Modeling. *The Annals of Statistics* 25, 1371–1470.
- STONE, C. J. (1977). Consistent Nonparametric Regression (with Discussion). *Annals of Statistics* 5, 595–645.
- STRAM, D. O. AND WEI, L. J. (1988). Analyzing Repeated Measurements with Possibly Missing Observations by Modelling Marginal Distributions. *Statistics in Medicine* 7, 139–148.
- STRAM, D. O., WEI, L. J., AND WARE, J. H. (1988). Analysis of Repeated Categorical Outcomes with Possibly Missing Observations and Time-Dependent Covariates. *Journal of the American Statistical Association* 83, 631–637.
- STRANG, G. (1993). Wavelet Transform versus Fourier Transforms. *Bulletin of the American Mathematical Society* 28, 288–305.
- STROUD, A. (1971). *Approximate Calculation of Multiple Integrals*. Englewood Cliffs, NJ: Prentice-Hall.
- STROUD, A. H. AND SECREST, D. (1966). *Gaussian Quadrature Formulas*. Englewood Cliffs, NJ: Prentice-Hall.
- STUKEL, T. A. (1988). Generalized Logistic Models. *Journal of the American Statistical Association* 83(402), 426–431.

- STUTE, W. (1984). Asymptotic Normality of Nearest Neighborhood Regression Function Estimates. *Annals of Statistics* 12, 917–926.
- STUTE, W. (1997). Nonparametric Model Checks for Regression. *The Annals of Statistics* 23, 461–471.
- STUTE, W., GONZÁLEZ-MONTEIGA, W., AND PRESEDO-QUINDIMIL, M. (1998). Bootstrap Approximations in Model Checks for Regression. *J. Am. Stat. Ass.* 93, 141–149.
- TANIZAKI, H. (1993). *Nonlinear Filters; Estimation and Applications. Lecture Notes in Economics and Mathematical Systems 400*. Berlin: Springer-Verlag.
- TERZA, J. V. (1985). Ordinal Probit: A Generalization. *Commun. Statist. Theor. Meth.* 14, 1–11.
- THALL, P. F. AND VAIL, S. C. (1990). Some Covariance Models for Longitudinal Count Data with Overdispersion. *Biometrics* 46, 657–671.
- THIELE, T. (1980). *Sur la Compensation de Quelques Erreurs Quasi-Systematiques par la Méthode des Moindres Carrés*. Copenhagen: Reitzel.
- THOMPSON, R. AND BAKER, R. J. (1981). Composite Link Function in Generalized Linear Models. *Applied Statistics* 30, 125–131.
- THOMPSON, W. A. (1977). On the Treatment of Grouped Observations in Life Studies. *Biometrics* 33, 463–470.
- TIBSHIRANI, R. AND HASTIE, T. (1987). Local Likelihood Estimation. *Journal of the American Statistical Association* 82, 559–568.
- TIELSCH, J. M., SOMMER, A., KATZ, J., AND EZRENE, S. (1989). *Sociodemographic Risk Factors for Blindness and Visual Impairment*. Archives of Ophthalmology: The Baltimore Eye Survey.
- TIERNEY, L. (1994). Markov Chains for Exploring Posterior Distributions. *Annals of Statistics* 22, 1701–1762.
- TIERNEY, L. AND KADANE, J. B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association* 81, 82–86.
- TITTERINGTON, D. M. AND BOWMAN, A. W. (1985). A Comparative Study of Smoothing Procedures for Ordered Categorical Data. *J. Statist. Comput. Simul.* 21, 291–312.
- TSIATIS, A. A. (1981). A Large Sample Study of Cox's Regressions Model. *Annals of Statistics* 9, 93–108.
- TSUTAKAWA, R. K. (1988). Mixed Model for Analyzing Geographic Variability in Mortality Rates. *Journal of the American Statistical Association* 83, 37–42.
- TUTZ, G. (1989). Compound Regression Models for Categorical Ordinal Data. *Biometrical Journal* 31, 259–272.
- TUTZ, G. (1990). Sequential Item Response Models with an Ordered Response. *British Journal of Statistical and Mathematical Psychology* 43, 39–55.
- TUTZ, G. (1991a). Choice of Smoothing Parameters for Direct Kernels in Dis-

- crimination. *Biometrical Journal* 33, 519–527.
- TUTZ, G. (1991b). Consistency of Cross-validatory Choice of Smoothing Parameters for Direct Kernel Estimates. *Computational Statistics Quarterly* 4, 295–314.
- TUTZ, G. (1991c). Sequential Models in Ordinal Regression. *Computational Statistics & Data Analysis* 11, 275–295.
- TUTZ, G. (1993). Invariance Principles and Scale Information in Regression Models. *Methodika VII*, 112–119.
- TUTZ, G. (1995a). Competing Risks Models in Discrete Time with Nominal or Ordinal Categories of Response. *Quality & Quantity* 29, 405–420.
- TUTZ, G. (1995b). Smoothing for Categorical Data: Discrete Kernel Regression and Local Likelihood Approaches. In H. H. Bock & W. Polasek (Eds.), *Data Analysis and Information Systems*, pp. 261–271. New York: Springer-Verlag.
- TUTZ, G. (1997). Sequential Models for Ordered Responses. In W. van der Linden & R. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, pp. 139–152. New York: Springer-Verlag.
- TUTZ, G. (2000). *Die Analyse kategorialer Daten – eine anwendungsorientierte Einführung in Logit-Modellierung und kategoriale Regression*. München: Oldenbourg Verlag.
- TUTZ, G. AND HENNEVOGL, W. (1996). Random Effects in Ordinal Regression Models. *Computational Statistics and Data Analysis* 22, 537–557.
- TUTZ, G. AND KAUERMANN, G. (1997). Local Estimators in Multivariate Generalized Linear Models with Varying Coefficients. *Computational Statistics* 12, 193–208.
- TUTZ, G. AND KAUERMANN, G. (1998). Locally Weighted Least Squares in Categorical Varying-coefficient Models. In R. Galata & H. Küchenhoff (Eds.), *Econometrics in Theory and Practice. Festschrift für Hans Scheeweß*, pp. 119–130. Heidelberg: Physika Verlag.
- TUTZ, G. AND PRITSCHER, L. (1996). Nonparametric Estimation of Discrete Hazard Functions. *Lifetime Data Analysis* 2, 291–308.
- VIDAKOVIC (1999). *Statistical Modelling by Wavelets*. Wiley Series in Probability and Statistics. New York: Wiley.
- WACLAWIW, M. AND LIANG, K. Y. (1993). Prediction of Random Effects in the Generalized Linear Model. *Journal of the American Statistical Association* 88, 171–178.
- WACLAWIW, M. AND LIANG, K. Y. (1994). Empirical Bayes Estimation and Inference for the Random Effects Model with Binary Response. *Statistics in Medicine* 13, 541–551.
- WAGENPFEIL, S. (1996). *Dynamische Modelle zur Ereignisanalyse*. München: Utz Verlag.
- WAHABA, G. (1978). Improper Prior, Spline Smoothing and the Problem of Guarding against Model Errors in Regression. *Journal of the Royal Statistical Society B* 44, 364–372.

- WAHBA, G. (1990). *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.
- WANG, P. C. (1985). Adding a Variable in Generalized Linear Models. *Technometrics* 27, 273–276.
- WANG, P. C. (1987). Residual Plots for Detecting Non-Linearity in Generalized Linear Models. *Technometrics* 29, 435–438.
- WARE, J. H., LIPSITZ, S., AND SPEIZER, F. E. (1988). Issues in the Analysis of Repeated Categorical Outcomes. *Statistics in Medicine* 7, 95–107.
- WATSON, G. S. (1964). Smooth Regression Analysis. *Sankhyā, Series A*, 26, 359–372.
- WECKER, W. AND ANSLEY, C. (1983). The Signal Extraction Approach to Nonlinear Regression and Spline Smoothing. *Journal of the American Statistical Association* 78, 81–89.
- WEDDERBURN, R. W. M. (1974). Quasilikelihood Functions, Generalized Linear Models and the Gauss-Newton Method. *Biometrika* 61, 439–447.
- WEDDERBURN, R. W. M. (1976). On the Existence and Uniqueness of the Maximum Likelihood Estimates for Certain Generalized Linear Models. *Biometrika* 63, 27–32.
- WERMUTH, N. AND LAURITZEN, S. L. (1990). On Substantive Research Hypotheses, Conditional Independence Graphs and Graphical Chain Models. *Journal of the Royal Statistical Society B* 52, 21–50.
- WEST, M. (1981). Robust Sequential Approximate Bayesian Estimation. *Journal of the Royal Statistical Society B* 43, 157–166.
- WEST, M. AND HARRISON, P. J. (1989). *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag.
- WEST, M. AND HARRISON, P. J. (1997). *Bayesian Forecasting and Dynamic Models* (2nd ed.). New York: Springer-Verlag.
- WEST, M., HARRISON, P. J., AND MIGON, M. (1985). Dynamic Generalized Linear Models and Bayesian Forecasting. *Journal of the American Statistical Association* 80, 73–97.
- WHITE, H. (1981). Consequences and Detection of Misspecified Nonlinear Regression Models. *Journal of the American Statistical Association* 76, 419–433.
- WHITE, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica* 50, 1–25.
- WHITE, H. (1984). Maximum Likelihood Estimation of Misspecified Dynamic Models. In T. Dijlestra (Ed.), *Misspecification Analysis*. Berlin: Springer-Verlag.
- WHITTAKER, E. T. (1923). On a New Method of Graduation. *Proc. Edinborough Math. Assoc.* 78, 81–89.
- WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.
- WILD, C. J. AND YEE, T. W. (1996). Additive Extensions to Generalized Estimating Equation Methods. *Journal of the Royal Statistical Society B* 58,

- 711–725.
- WILLIAMS, D. A. (1982). Extra Binomial Variation in Logistic Linear Models. *Applied Statistics* 31, 144–148.
- WILLIAMS, D. A. (1987). Generalized Linear Model Diagnostics Using the Deviance and Single Ease Deletions. *Applied Statistics* 36, 181–191.
- WILLIAMS, O. D. AND GRIZZLE, J. E. (1972). Analysis of Contingency Tables Having Ordered Response Categories. *Journal of the American Statistical Association* 67, 55–63.
- WILLIAMSON, J. M., KIM, K., AND LIPSITZ, S. R. (1995). Analyzing Bivariate Ordinal Data Using a Global Odds Ratio. *Journal of the American Statistical Association* 90, 1432–1437.
- WILSON, J. AND KOEHLER, K. (1991). Hierarchical Models for Cross-Classified Overdispersed Multinomial Data. *J. Bus. & Econ. St.* 9, 103–110.
- WINKELMANN, R. (1997). *Econometric Analysis of Count Data* (2nd ed.). Berlin: Springer-Verlag.
- WINKLER, G. (1995). *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*. New York: Springer-Verlag.
- WOLAK, F. (1987). An Exact Test for Multiple Inequality and Equality Constraints in the Linear Regression Model. *Journal of the American Statistical Association* 82, 782–793.
- WOLAK, F. (1989). Local and Global Testing of Linear and Nonlinear Inequality Constraints in Nonlinear Econometric Models. *Econometric Theory* 5, 1–35.
- WOLFINGER, R. W. (1994). Laplace's Approximation for Nonlinear Mixed Models. *Biometrika* 80, 791–795.
- WONG, C. AND KOHN, R. (1996). A Bayesian Approach to Estimating and Forecasting Additive Nonparametric Autoregressive Models. *Journal of Time Series Analysis* 17, 203–220.
- WONG, G. Y. AND MASON, W. M. (1985). The Hierarchical Logistic Regression Model for Multilevel Analysis. *Journal of the American Statistical Association* 80, 513–524.
- WONG, W. H. (1986). Theory of Partial Likelihood. *Annals of Statistics* 14, 88–123.
- WU, J. C. F. (1983). On the Covergence Properties of the EM-Algorithm. *The Annals of Statistics* 11, 95–103.
- YANG, S. (1981). Linear Function of Concomitants of Order Statistics with Application to Nonparametric Estimation of a Regression Function. *Journal of the American Statistical Association* 76, 658–662.
- YAU, R., KOHN, R., AND WOOD, S. (2000). Bayesian Variable Selection and Model Averaging in High Dimensional Multinomial Nonparametric Regression. Technical Report, University of NSW, Sydney.
- YELLOTT, J. I. (1977). The Relationship Between Luce's Choice Axiom, Thurstone's Theory of Comparative Judgement, and the Double Exponential

- Distribution. *Journal of Mathematical Psychology* 15, 109–144.
- ZEGER, S. AND LIANG, K.-Y. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics* 42, 121–130.
- ZEGER, S. AND LIANG, K.-Y. (1989). A Class of Logistic Regression Models for Multivariate Binary Time Series. *Journal of the American Statistical Association* 84, 447–451.
- ZEGER, S., LIANG, K.-Y., AND ALBERT, P. S. (1988). Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics* 44, 1049–1060.
- ZEGER, S. L. (1988a). Commentary. *Statistics in Medicine* 7, 161–168.
- ZEGER, S. L. (1988b). A Regression Model for Time Series of Counts. *Biometrika* 75, 621–629.
- ZEGER, S. L., DIGGLE, P. J., AND YASUI, Y. (1990). Marginal Regression Models for Time Series. Inst. for Mathematics and Its Application, Time Series Workshop, Preprint.
- ZEGER, S. L. AND KARIM, M. R. (1991). Generalized Linear Models with Random Effects; A Gibbs' Sampling Approach. *Journal of the American Statistical Association* 86, 79–95.
- ZEGER, S. L. AND QAQISH, B. (1988). Markov Regression Models for Time Series: A Quasi-Likelihood Approach. *Biometrics* 44, 1019–1031.
- ZELLNER, A. AND ROSSI, P. E. (1984). Bayesian Analysis of Dichotomous Quantal Response Models. *Journal of Econometrics* 25, 365–393.
- ZHAO, L. P. AND PRENTICE, R. L. (1990). Correlated Binary Regression Using a Quadratic Exponential Model. *Biometrika* 77, 642–648.
- ZHAO, L. P., PRENTICE, R. L., AND SELF, S. (1992). Multivariate Mean Parameter Estimation by Using a Partly Exponential Model. *Journal of the Royal Statistical Society B* 54, 805–811.
- ZIEGLER, A., KASTNER, C., AND BLETTNER, M. (1998). The Generalized Estimating Equations: An Annotated Bibliography. *Biometrical Journal* 40, 115–139.

# Author Index

- Aalen, 422  
Abrahamowicz, 422  
Abramowitz, 445  
Agresti, 14, 36, 81, 94, 104, 113, 120,  
    130, 265, 267, 284, 307, 328  
Aitchison, 429  
Aitken, 201, 429  
Aitkin, 15, 285, 288, 295, 306, 308,  
    310, 312, 315, 393–395, 460  
Albert, 64, 65, 137, 160, 171, 236, 271,  
    284, 285, 326–328  
Andersen, 264, 265, 385, 395, 429  
Anderson, 15, 34, 82, 103, 104, 239,  
    274, 285, 295, 306, 312, 338,  
    339, 345, 350, 394, 460  
Andrews, 166, 167  
Ansley, 337, 339, 341  
Aoki, 343  
Aranda-Ordaz, 87, 402  
Arjas, 395  
Arminger, 113, 165, 166  
Armstrong, 95, 99  
Artes, 278  
Ashby, 113  
Atkinson, 145, 156  
Azzalini, 162, 172, 239, 252
- Bacchetti, 113  
Baker, 65  
Barnhart, 95  
Bartholomew, 113  
Bates, 285, 307  
Beck, 11, 118, 271  
Belsley, 145  
Ben-Akiva, 80  
Benedetti, 187  
Benjamin, 247, 248, 252, 255  
Berghold, 14  
Berhane, 275
- Besag, 118, 321, 322, 378–380, 382,  
    453  
Best, 172  
Bhapkar, 110  
Bickel, 180  
Biller, 228, 233, 369, 373, 424  
Birch, 111  
Bishop, 14, 36  
Blettner, 267  
Blossfeld, 385  
Bock, 306  
Böhning, 308  
Börsch-Supan, 78  
Bonney, 115, 116  
De Boor, 176, 177, 181, 182  
Booth, 314, 328  
Borgan, 385, 395, 429  
Bowman, 162, 172, 201, 239  
Boyles, 443  
Breiman etc, 210  
Breslow, 36, 62, 64, 284, 298–300, 323,  
    328, 373, 380  
Brezger, 228, 234  
Brillinger, 295, 312, 314  
Brown, 328  
Buja, 217  
Burns, 291  
Buse, 50
- Cameron, 35, 37  
Carlin, 171, 172, 343, 362, 452, 453  
Carroll, 67, 172, 209, 274  
Carter, 227, 343  
Casella, 285, 321  
Le Cessie, 172, 239  
Chambers, 461  
Chang, 171  
Chao, 187  
Chappell, 221  
Chen, 137, 171, 236, 278

- Chhikara, 24  
 Chib, 137, 171, 236, 278  
 Christensen, 14  
 Di Ciccio, 171  
 Clayton, 62, 63, 284, 298–300, 322, 323, 328, 373, 380, 382, 393, 395  
 Cleveland, 189  
 Clogg, 115  
 Collett, 15, 35, 212, 463  
 Conaway, 265, 328  
 Congdon, 421  
 Conolly, 118  
 Cook, 145, 156, 159, 160  
 Copas, 172, 421  
 Corbeil, 291  
 Cowles, 453  
 Cox, 64, 95, 114, 167, 242, 243, 389, 403, 440  
 Craven, 192  
 Cressie, 105, 109–112, 378, 383  
 Czado, 65, 87, 162, 221
- Daganzo, 78  
 Dale, 112, 130  
 Dannegger, 429  
 Davidson, 168  
 Davis, 129, 273, 443, 444  
 Davison, 67  
 Dean, 35  
 Decarli, 14  
 De Jong, 339, 343  
 Del Pino, 328  
 Dellaportas, 63, 171  
 Démétrio, 35  
 Dempster, 442, 443  
 Denison, 227, 233  
 van Deusen, 334  
 Devroye, 451  
 Dewulf, 274  
 Dey, 60, 137, 170, 171, 236, 240, 278  
 Dielman, 260, 285  
 Dietz, 308  
 Diggle, 14, 113, 120, 241, 256, 260, 267, 285  
 Dobson, 15  
 Doksum, 391  
 Donoho, 180, 194  
 Doss, 422  
 Doucet, 367  
 Dransfield, 61, 445, 446, 449  
 Duan, 164  
 Duffy, 14, 61, 63
- Durbin, 352, 367  
 Edwards, 83, 140  
 Efron, 59, 67, 403, 429, 430  
 Eilers, 62, 67, 176, 180, 194, 209, 217  
 Engle, 190, 246  
 Epanechnikov, 185  
 Erkanli, 240  
 Escobar, 221  
 Esdaile, 422  
 Eubank, 172, 173, 239, 275  
 Ezrene, 117
- Fahrmeir, 3, 14, 15, 32, 36, 43, 46, 49, 57, 129–131, 140, 143, 144, 162, 163, 223, 233, 235, 236, 238, 245, 251, 323, 340, 346, 352, 355, 356, 362, 367, 369, 372, 373, 424, 425, 440, 441  
 Fan, 173, 188, 189, 201, 209, 422  
 Farewell, 87  
 Fernandes, 346  
 Ferreira, 346, 351  
 Fienberg, 14, 36  
 Finkelstein, 396  
 Finney, 148  
 Firth, 15, 172, 239  
 Fitzmaurice, 120, 123, 135, 136, 274  
 Fokianos, 245, 251, 255  
 Folks, 24  
 Forcina, 14  
 Forster, 171  
 Forthofer, 6, 81  
 Foulley, 52  
 Foutz, 58  
 Francis, 14, 15, 308, 394, 460  
 Freitas, 367  
 Friedl, 14, 87  
 Friedman, 178, 209, 210, 213  
 Frost, 140, 143, 144, 166, 169  
 Frühwirth-Schnatter, 227, 334, 343, 366  
 Furnival, 139, 142  
 Fusaro, 113
- Galmacci, 14  
 Gamerman, 63, 233, 322, 346, 351, 362, 452  
 Gange, 113  
 Garber, 244, 264  
 Gasko, 391  
 Gasser, 187, 189  
 Gay, 59, 66  
 van der Geer, 238

- Gelfand, 60, 65, 170–172, 221, 452  
 Gelman, 170, 171, 452  
 Geman, 451, 452  
 Genter, 87  
 Gersch, 334, 336  
 Gianola, 312  
 Gibbons, 308  
 Gieger, 130, 136, 274–276  
 Giesbrecht, 291  
 Gigli, 67  
 Gijbels, 173, 188, 189, 201, 209, 422  
 Gilchrist, 14, 65  
 Gilks, 63, 452  
 Gill, 385, 395, 429  
 Glonek, 136  
 Glosup, 172, 239  
 Göttlein, 275  
 Goldstein, 285, 289  
 González-Manteiga, 172  
 Goodman, 104  
 Goormastic, 118  
 Gordon, 334, 367  
 Gosh, 60, 65, 170, 172  
 Goss, 356  
 Gould, 402  
 Gourieroux, 50, 56, 163, 168, 169, 249,  
     308, 318  
 Granger, 190  
 Green, 14, 65, 66, 171, 173, 182, 195,  
     220, 228, 321, 322, 352, 380,  
     453  
 Greenland, 95, 104  
 Greenwood, 399  
 Grether, 249  
 Griliches, 328  
 Grizzle, 104  
 Gu, 206  
 Guan, 249  
 Guienneuc-Jouyaux, 453  
 Haberman, 43, 46, 212, 421  
 Härdle, 14, 162, 172–174, 191, 193,  
     217, 239  
 Hale, 366  
 Hall, 191, 193, 251, 328  
 Hamerle, 3, 15, 32, 36, 144, 265, 268,  
     271, 328, 385, 414, 416  
 Han, 171  
 Hanefeld, 13, 399  
 Hankey, 403  
 Hansen, 210  
 Harrington, 122  
 Harrison, 61, 331, 332, 334, 335, 346,  
     358, 360  
 Hart, 172, 193, 239  
 Hartley, 442  
 Hartzel, 307  
 Harvey, 331, 332, 334–336, 338, 343,  
     344, 346  
 Harville, 284, 289, 290, 297, 298  
 Hastie, 14, 173, 189, 193, 197, 201,  
     209, 217, 218, 220, 226–228,  
     232, 323, 342, 461  
 Hatzinger, 14  
 Hauck, 113, 284, 295, 327  
 Hausman, 78, 165, 328  
 Havranek, 140  
 Heagerty, 129, 131, 274  
 Hebbel, 341  
 Heckman, 249, 262, 308  
 Hedeker, 308  
 van der Heijden, 14  
 Heilbron, 113  
 Heiler, 341  
 Hennevogl, 158, 312, 321, 362, 366,  
     449  
 Heumann, 131, 136  
 Heyde, 251  
 Higdon, 321, 322, 380, 453  
 Hinde, 15, 35, 67, 285, 295, 306, 312,  
     314, 394, 460  
 Hinkley, 64, 67, 172, 239, 440  
 Hoaglin, 146  
 Hobart, 328  
 Hobert, 314, 321  
 Hocking, 139  
 Hodges, 366  
 Hoerl, 180  
 Holland, 14, 36, 66  
 Holly, 50, 165  
 Holmes, 93  
 Holtbrügge, 104  
 Hopper, 264  
 Horowitz, 162, 209  
 van Houwelingen, 172, 239  
 Hsiao, 260, 285, 287, 288, 290, 369  
 Huber, 66  
 Hürzeler, 366, 367  
 Huffner, 422  
 Hunsberger, 220  
 Hurvich, 193  
 Hutchinson, 334  
 Ibrahim, 329  
 Im, 312  
 Jacobs, 249

- Jansen, 14, 285, 297, 312, 314  
 Jazwinski, 345  
 Jewell, 113  
 Johnstone, 180, 193, 194  
 Jones, 14, 285, 286, 291, 331, 369  
 Jorgensen, 24, 66, 278, 433  
 Jung, 249
- Kadane, 62  
 Kalbfleisch, 86, 113, 284, 295, 327, 385, 389, 391, 401, 405, 410, 414, 416  
 Karim, 317, 321, 322  
 Kashiwagi, 368, 369  
 Kass, 171  
 Kastner, 267  
 Katz, 117  
 Kauermann, 14, 172, 201, 202, 209, 213, 239, 274  
 Kaufmann, 43, 46, 50, 245, 251, 346, 352, 355, 372, 440, 441  
 Kedem, 251  
 Keiding, 385, 395, 429  
 Kennard, 180  
 Kerkyacharian, 180  
 Kim, 129  
 Kirchen, 344  
 Kitagawa, 8, 334, 336, 345, 356, 358, 366, 367, 369  
 Kleffe, 285, 289–291  
 Kleinman, 329  
 Klemme, 362  
 Klinger, 180, 194, 429  
 Knorr-Held, 223, 227, 233, 322, 340, 363, 365, 367, 373, 382, 383, 425  
 Koehler, 240, 328  
 König, 265, 373  
 Kohn, 227, 236, 337, 339, 341, 343  
 Kooperberg, 210, 378  
 Koopman, 339, 352, 367  
 Krämer, 166  
 Kranert, 158  
 Krantz, 81  
 Kredler, 43, 144  
 Küchenhoff, 67, 211  
 Kuensch, 331  
 Künsch, 366, 367, 379  
 Küsters, 113  
 Kuh, 145  
 Kullback, 111
- Laird, 11, 61, 120, 122, 135, 136, 271, 274, 284, 285, 290, 291, 295, 298, 442, 443  
 Lancaster, 385  
 Landis, 129  
 Landwehr, 145  
 Lang, 228, 233–236, 265, 323, 328, 362, 425  
 Lauritzen, 14, 114, 340  
 Lawless, 64, 139, 142, 385, 386, 389, 392, 399, 402  
 Lee, 160, 328  
 Lehnert, 6, 81  
 Leonard, 61, 65  
 Lerman, 78, 80  
 Lesaffre, 113, 120, 136, 160, 268, 274  
 Levine, 252  
 Lewis, 249  
 Li, 164, 422  
 Lia, 328  
 Liang, 7, 8, 11, 14, 35, 113, 117, 118, 120–122, 124, 127, 135, 241, 260, 261, 267, 268, 270, 271, 284, 285, 326–328  
 Lin, 298, 300, 323, 328, 373, 380  
 Lindsey, 14, 285, 346  
 Lindstrom, 113, 285  
 Linton, 217  
 Lipsitz, 122, 129, 264, 274  
 Liu, 307  
 Loader, 14, 173, 189, 193  
 Lombard, 67  
 Longford, 14, 285, 288, 289  
 Los, 344  
 Louis, 443  
 Luce, 81
- MacDonald, 331  
 Mack, 190  
 MacKenzie, 422  
 Maddala, 80, 249  
 Magnus, 302, 304  
 Mallat, 178  
 Mallick, 60, 170, 227, 233  
 Mammen, 238  
 Mammitzsch, 187  
 Manski, 78  
 Mantel, 403  
 Marchetti, 14  
 Margolin, 4, 37  
 Marron, 191, 193  
 Martin, 331, 345  
 Marx, 14, 62, 67, 176, 180, 194, 209, 217
- Läärä, 95

- Mason, 295, 298  
 Masters, 104  
 Matthews, 95  
 Mayer, 385  
 McCullagh, 14, 15, 35, 36, 51, 55, 56,  
     59, 84, 93, 99, 136, 139, 153,  
     162, 439  
 McCulloch, 285, 314, 328  
 McDonald, 123, 271  
 McFadden, 78  
 McKeague, 422  
 McKinnon, 167, 168  
 Mee, 284, 297, 298  
 Mehta, 100  
 Meilija, 443  
 Melsa, 332, 345, 350, 352  
 Meng, 170, 171  
 Mengersen, 322, 453  
 Mengerson, 321, 380, 453  
 Migon, 332, 346, 358, 360  
 Miller, 129, 139, 346  
 Minder, 14  
 Molenberghs, 113, 120, 136, 268, 274  
 Mollie, 379, 380  
 Montfort, 50, 56, 163, 168, 169, 249,  
     308, 318  
 Moore, 328, 332, 338, 339, 345, 350  
 Morawitz, 102, 265, 372  
 Morgan, 87  
 Morris, 56  
 Moulton, 67, 273  
 Müller, 171, 187, 190, 240, 367  
 Mukhopadhyay, 221  
 Muthén, 113  
 Nadaraya, 186, 190, 429  
 Nagl, 268, 271  
 Nason, 178  
 Naylor, 61, 445, 446, 449  
 Nelder, 1, 14, 15, 36, 51, 55, 56, 58,  
     59, 139, 153, 162, 328, 434  
 Nerlove, 265, 373  
 Neudecker, 302, 304  
 Neuhaus, 113, 284, 295, 327  
 Newton, 113, 221  
 Neyman, 111  
 Novick, 65  
 Ntzoufras, 171  
 Nyquist, 194  
 Ogden, 178  
 Opsomer, 217  
 Osius, 112  
 O'Sullivan, 197, 206, 352  
 Oudiz, 265, 373  
 Owen, 240  
 Parmigiani, 171  
 Parr, 111  
 Patel, 100  
 Pauly, 341, 343  
 Payne, 328  
 Pederson, 172  
 Pendergast, 113  
 Pepe, 274  
 Picard, 180  
 Piegorsch, 4, 27, 37  
 Pierce, 153, 154, 307  
 Pinheiro, 307  
 Pitt, 363, 367  
 Polson, 343, 362  
 Poortema, 35  
 Prakasa Rao, 437  
 Pregibon, 56, 58, 59, 65, 87, 145, 147,  
     148, 160, 162, 172, 205  
 Preisler, 295, 312, 314  
 Prentice, 86, 87, 118, 121, 125, 135,  
     136, 385, 389, 391, 401, 405,  
     410, 414, 416, 433  
 Presedo-Quindimil, 172  
 Priestley, 187  
 Pritscher, 129–131, 421  
 Pruscha, 249, 275  
 Pudney, 80  
 Qaqish, 7, 8, 117, 120, 122, 127, 135,  
     245–247, 252  
 Qu, 118  
 Quintana, 328  
 Rabash, 328  
 Rabinowitz, 443, 444  
 Raftery, 170, 171, 331, 345  
 Randall, 292, 293  
 Rao, 285, 289–291  
 Rasch, 264  
 Raynor, 197, 206, 352  
 Read, 105, 109–112  
 Redner, 443  
 Reinsch, 181, 182, 341  
 Riani, 145, 156  
 Rice, 190, 193  
 Richardson, 452  
 Rigby, 247, 248, 252, 255  
 Ripley, 322, 450, 451  
 Robert, 453  
 Roberts, 81, 352

- Robins, 274  
 Rojek, 112  
 Ronning, 80, 249, 265, 328  
 Rosenberg, 369  
 Rosner, 118  
 Rossi, 61, 316, 317  
 Rotnitzky, 120, 135, 274  
 Rubin, 442, 443, 452  
 Rue, 227  
 Ruppert, 67, 217, 274
- Sage, 332, 345, 350, 352  
 Salmon, 367  
 Sampson, 95  
 Santner, 14, 61, 63  
 Scallan, 65  
 Schafer, 153, 154  
 Schall, 300, 328  
 Schlicht, 341  
 Schnatter, 366  
 Schneider, 332, 339, 344  
 Schoenberg, 165, 166  
 Schuhmacher, 104  
 Schumaker, 176  
 Schwarz, 142  
 Searle, 285, 291  
 Secrest, 445  
 Seeber, 14, 66, 139  
 Segal, 113  
 Segerstedt, 194  
 Seifert, 189  
 Self, 136, 389, 433  
 Severini, 172, 220, 239  
 Shao, 67  
 Shaw, 61, 445, 446, 449  
 Shephard, 343, 363, 367  
 Shoemaker, 145  
 Silvapulle, 43  
 Silverman, 14, 173, 178, 182, 190, 220  
 Simonoff, 14, 154, 173, 188, 193, 211, 239  
 Singer, 308  
 Singh, 352  
 Singhal, 139, 142  
 Sinha, 240  
 Skene, 61, 445, 446, 449  
 Sloan, 95, 99  
 Small, 78  
 Smith, 61–63, 67, 227, 233, 334, 346, 367, 445, 446, 449, 452  
 Sobel, 113  
 Sommer, 117  
 Song, 136, 278
- Speckman, 220  
 Speizer, 264  
 Spiegelhalter, 172, 452  
 Srivastava, 58  
 Stadtmüller, 190  
 Staniswalis, 172, 190, 220, 239  
 Stasinopoulos, 247, 248, 252, 255  
 Steckel-Berger, 14  
 Stefanski, 67  
 Stegun, 445  
 Stern, 171, 452  
 Stevens, 335  
 Stiratelli, 61, 284, 295, 298  
 Stoffer, 343, 362  
 Stone, 189, 190, 210  
 Stram, 262, 273, 374  
 Strang, 178  
 Stroud, 444, 445  
 Stützle, 209, 213  
 Stukel, 87, 162  
 Stute, 172, 190  
 Suppes, 81  
 Swartz, 171
- Tanizaki, 366  
 Taylor, 165  
 Terza, 87  
 Thall, 268  
 Thiele, 340  
 Thompson, 65, 403  
 Thurstone, 83  
 Tibshirani, 14, 67, 173, 193, 197, 201, 209, 217, 218, 220, 226–228, 232, 275, 323, 342  
 Tielsch, 117  
 Tierney, 62, 452, 453  
 Titterington, 201  
 Trivedi, 35, 37  
 Trognon, 56, 163, 168, 169, 249  
 Truong, 210  
 Tsai, 154, 193  
 Tsiatis, 100, 389  
 Tsutakawa, 328  
 Tu, 67  
 Tutz, 14, 15, 36, 81, 95, 99, 102, 104, 172, 201, 202, 209, 211, 213, 239, 265, 312, 321, 414, 416, 421  
 Tversky, 81  
 Ulm, 211, 429  
 Vail, 268  
 Vidakovic, 178

- Vlachos, 171  
Waclawiw, 328  
Wagenpfeil, 238, 352, 356, 372, 424,  
    425  
Wahba, 192, 217, 225, 337, 341  
Wald, 270  
Walker, 443  
Wand, 209  
Wang, 145  
Ware, 11, 61, 262, 264, 271, 273, 284,  
    285, 290, 291, 295, 298, 374  
Wasserman, 171  
Watson, 186, 190, 429  
Wecker, 337  
Wedderburn, 1, 43, 55, 56, 434  
Wei, 262, 273, 374  
Weinberg, 4, 37  
Weisberg, 145  
Weiss, 190  
Weissenbacher, 2  
Welsch, 59, 66, 145, 146  
Welsh, 274  
Wermuth, 114  
West, 61, 221, 240, 331, 332, 334, 345,  
    346, 358, 360  
White, 58, 161–163, 166, 252, 264  
Whittaker, 14, 114, 224, 341  
Wild, 63, 274, 452  
Williams, 64, 93, 104, 118, 160, 328  
Williamson, 129  
Wilson, 139, 142, 328  
Winkelmann, 37  
Wise, 78  
Wolak, 50  
Wolfinger, 300, 328  
Wong, 227, 244, 295, 298  
Wood, 236  
Wu, 443  
Yanagimoto, 368, 369  
Yandell, 195, 197, 206, 352  
Yang, 190  
Yasui, 256  
Yau, 236  
Yee, 274  
Yellott, 78  
Yi, 193  
York, 379, 380  
Young, 264  
Yuan, 249  
Zeger, 7, 8, 10, 11, 14, 67, 113, 117,  
    118, 120–122, 124, 127, 129,  
    131, 135, 241, 245–247, 252,  
    256, 257, 259–261, 267, 268,  
    270, 271, 273, 274, 284, 285,  
    317, 321, 322, 326–328  
Zellner, 61, 316, 317  
Zhang, 323, 380  
Zhao, 135, 136, 274, 433  
Ziegler, 267  
Zucchini, 331

# Subject Index

- AIC, 142
- Akaikes information criterion, 142
- Alternative-specific, 80
- Anderson, 332
- ARCH models, 246
- Asymptotics, 49, 51, 111, 112, 263, 270, 437
  - consistency, 44, 270
  - existence, 438
  - existence and uniqueness, 44
  - normality, 44, 62, 438, 439
- Auto-covariance function, 257
- Autocorrelation function, 259
- Autoregressive models, 242
- Average predictive squared error, 190
- Average squared error, 190
- Backfitting, 215–217, 219
- Bandwidth, 186
- Basis function, 174, 227
- Basis function approach, 193, 210
- Bayesian inference, 63
- Bayesian mixed models, 321
- Bayesian models, 60
- BayesX, 455, 465
- Binary response, 24, 122
- Binary time series, 243, 257
- Binomial response, 24
- Binomial time series, 243
- BMDP, 455, 458
- BUGS, 455, 464
- CART, 210
- Case deletion, 156
- Categorical time series, 245, 347
- Censoring, 391
  - random, 391
  - Type I, 392
- Cholesky square root, 448
- Cluster, 120
- Cluster-specific, 285, 293
- Coding
  - dummy, 16
  - effect, 16
  - of covariates, 16
- Conditional covariance, 250
- Conditional expectation, 250
- Conditional gamma models, 247
- Conditional independence, 261
- Conditional information, 250, 263
- Conditional likelihood, 264
- Conditional means, 261
- Conditional model, 242, 261
  - for logit data, 261
  - statistical inference, 249
- Conjugate prior-posterior, 351
- Consistency, 44, 270, 438
- Cook distance, 159
- Correction step, 339
- Correlated categorical responses, 129
- Correlated univariate responses, 120
- Correlation matrix
  - working, 121
- Count data, 244
- Counts, 246
- Covariance matrix
  - working, 121
- Covariate
  - categorical, 20
  - external, 410
  - internal, 410
  - metrical, 20
  - time-varying, 408
- Cross-validation, 191
- DARMA processes, 249
- Data
  - count, 36, 244, 246

- grouped, 17, 22, 72
- longitudinal, 241, 260, 369
- panel, 369
- ungrouped, 17, 22
- Data-driven, 114
- Demmler-Reinsch, 275
- Density
  - marginal posterior, 61
  - posterior, 60
  - prior, 60
- Design matrix, 72, 76, 79, 88, 102, 332
- Design vector, 19, 20
- Deviance, 50, 108
- Discrete choice, 80
- Dispersion parameter, 19, 269, 433
- Distribution
  - $\chi^2$ , 49
  - binomial, 20, 24
  - exponential, 23, 386
  - extreme maximal-value, 86
  - extreme minimal-value, 26, 86
  - extreme-value, 78
  - gamma, 20, 23
  - inverse Gaussian, 20, 24
  - mixing, 286
  - multinomial, 70
  - normal, 20, 22
  - Poisson, 20, 36
  - scaled binomial, 25
  - scaled multinomial, 71
  - Weibull, 387
- Dynamic models, 332, 345, 348
- EGRET, 455, 463
- EM algorithm, 290, 291, 343, 356, 442
- EM-type algorithm, 63, 302, 356
- Empirical Bayes, 310
- Epanechnikov kernel, 185
- Equicorrelation, 122
- Estimating equation, 258
- Estimation, 303
  - by integration techniques, 303
  - existence, 43
  - generalized estimating equation, 56, 124, 269
  - hierarchical Bayes, 65
  - hyperparameter, 356
  - marginal, 325
  - maximum likelihood, 38, 105, 303, 404, 417, 437, 442
  - MINQUE, 290
  - MIVQUE, 290
  - of marginal models, 258
- posterior mean, 315, 351
- posterior mode, 61, 298, 340, 351
- quasi-maximum likelihood, 56, 64, 121, 252, 258
- RMLE, 290
- under misspecification, 162
- uniqueness, 43
- Estimator
  - kernel
    - Nadaraya-Watson, 186
    - loess, 189
    - lowess, 189
  - Exponential family, 19, 433
    - simple, 433
- Filter, 337
  - Fisher-scoring, 354
  - Gauss-Newton, 354
  - non-normal, 350
- Fisher matrix
  - expected, 40, 106, 435
  - Fisher scoring, 42, 107, 300
    - for generalized spline smoothing, 196
- Fixed-parameter models, 241
- Gasser-Müller weight, 187, 189
- GAUSS, 455, 462
- Gauss-Hermite integration, 306
  - multivariate, 447
  - univariate, 444
- Gauss-Hermite quadrature, 366
- Gauss-Seidel algorithm, 219
- Gaussian kernel, 185
- Generalized additive models, 207, 274
- Generalized autoregressive models, 261
- Generalized cross-validation, 192, 197, 220
- Generalized estimating equation, 124, 258, 269
- Generalized linear model
  - dynamic, 345
  - multivariate, 75, 433
  - software, 455
  - univariate, 18, 433
- Generalized partially single index model, 209
- Generalized sequential model, 95
- GENSTAT, 455, 458
- Gibbs sampler, 228
- Gibbs sampling, 317, 451
- GLIM, 455, 460
- Global, 79

- Goodness-of-fit, 47, 50, 107, 151
- Hat matrix, 146  
generalized, 147
- Hazard  
baseline, 403
- Hazard function, 386  
cause-specific, 414  
discrete, 396  
overall, 414
- Heterogeneity  
unobserved, 35
- Hidden Markov, 366
- Hyperparameter, 63, 333, 342, 343, 356
- Importance sampling, 316, 449
- Independence  
working, 122
- Inference, 262, 268  
statistical, 289
- Information matrix, 39, 166  
expected, 40, 106, 435, 438  
observed, 40, 435, 436, 438
- Integrated squared error, 190
- Integration techniques, 303
- Integration-based approaches, 365
- Iteratively weighted least-squares, 42, 107
- Jeffreys prior, 321
- $k$ -nearest neighborhood, 183
- Kalman filter, 332, 333, 338  
generalized extended, 352
- Kalman gain, 339, 353
- Kernel  
Epanechnikov, 185  
Gaussian, 185
- Landau symbol, 437
- Latent variable models, 234
- Leaving-one-out estimate, 191
- Life table, 397
- Likelihood  
marginal, 299, 325  
maximum likelihood estimation, 38, 105, 303, 404, 417, 437, 442  
penalized, 62, 195, 424  
quasi-maximum, 56, 64, 121, 252, 258
- Likelihood ratio statistic, 48, 251
- LIMDEP, 455, 464
- Linear predictor, 434
- Link function, 19, 20, 73, 88, 102, 434  
log-link, 23  
natural, 20, 434  
violation, 161
- Local average, 183
- Local estimators, 183
- Local likelihood estimation, 198
- Local linear trend, 334
- Log-likelihood, 39, 105, 435, 437
- Logit model  
dynamic, 346
- Longitudinal data, 241, 260
- Marginal cumulative response model, 273
- Marginal estimation, 325
- Marginal likelihood, 299
- Marginal logit model, 257, 267
- Marginal model, 119, 255, 257, 258, 267
- Markov chain Monte Carlo, 451
- Markov chains, 233
- Markov model, 243, 261
- Markov random field, 378, 379, 428
- Markovian process, 333
- MARS, 210
- Maximum likelihood  
nonparametric, 308
- Maximum random utility, 77
- MCMC, 60, 63, 227, 228, 233, 362, 365, 373, 381, 425
- Mean average squared error, 190
- MINQUE, 290
- MIVQUE, 290
- Mixed models  
Bayesian, 321
- Mixing distribution, 286
- ML equations, 437
- Model  
Aranda-Ordaz, 402  
ARCH, 246  
asymmetric, 114  
autoregressive, 242  
Bayesian, 60  
binary regression, 390  
competing risks, 414  
complementary log-, 26  
complementary log-log, 26  
compound cumulative, 101  
conditional, 114, 242, 261  
conditional gamma, 247  
conditional Gaussian, 333

- cumulative, 83, 87
- discrete-time, 396
- dispersion, 433
- dynamic cumulative models, 348
- dynamic generalized linear, 345
- dynamic logit model, 346
- dynamic multivariate logistic, 347
- dynamic Poisson model, 346
- exponential, 394
- for categorical time series, 241, 245
- for correlated responses, 112
- for longitudinal data, 241
- for marginal cumulative response, 273
- for nominal responses, 77
- for non-normal time series, 242
- for nonexponential family time series, 248
- for nonlinear time series, 248
- for ordinal responses, 81
- general state space, 350
- generalized autoregressive, 261
- generalized sequential, 95
- grouped Cox, 86
- grouped proportional hazards, 400
- linear Poisson, 36
- linear probability, 25
- linear transformation, 390
- location-scale, 388
- log-linear Poisson, 36, 244
- logistic, 29, 403
- logit, 26
- marginal, 119, 255, 257, 267
- Markov, 243, 261
- multicategorical logit, 72
- multicategorical response, 70
- multilevel, 288
- multinomial logit, 78
- multiple modes of failure, 414
- multivariate dynamic, 347
- non-normal, 345
- non-normal state space, 345
- nonexponential family, 349
- nonexponential family regression, 65
- nonhomogeneous for transition probabilities, 244
- nonlinear, 345
- nonlinear exponential family, 436
- nonlinear family, 349
- nonlinear regression, 65
- piecewise exponential, 388, 395
- probit, 26
- proportional hazards, 86, 389
- proportional odds, 83, 390
- quasi-likelihood, 55, 246, 261, 440
- quasi-likelihood Markov, 246
- random effects, 285, 292
- sequential, 92, 403
- state space, 332
- stereotype regression, 103
- survival, 396
- symmetric, 116
- threshold, 29, 83
- transition, 114, 264
- two-step, 100
- two-step cumulative, 101
- variance components, 287
- Weibull, 394
- Models with multiple covariates, 228
- Monte Carlo methods, 304, 449
- Multilevel models, 288
- Multiple covariates, 202
- Multivariate adaptive regression splines, 210
- Nadaraya-Watson estimate, 201
- Nadaraya-Watson kernel estimator, 186, 188
- Nearest neighborhood, 189
- Non-normal time series, 242
- Nonexponential family regression models, 65
- Nonexponential family time series, 248
- Nonlinear time series, 248
- Nonstationary Markov chains, 244
- Numerical integration, 443
- Observation equation, 332
- Odds ratio, 29, 73, 122, 274
- One-step estimate, 156
- Overdispersion, 35
- Panel waves, 261, 369
- Parameter
  - category-specific, 79
  - choice of smoothing parameter, 197
  - dispersion, 19, 269, 433
  - global, 80
  - natural, 19, 433

- nonconstant dispersion, 59
- overdispersion, 47
- scale, 19
- smoothness, 190, 341
- Partial likelihood, 249
- Partial linear models, 220
- Partially linear models, 208
- Particle filters, 367
- Pearson residuals, 130
- Pearson statistic, 50, 152
- Penalized, 352
- Penalized least-squares, 181, 340
- Penalized log-likelihood, 195, 424
- Penalty matrix, 341
- Plots
  - added variable, 145
  - partial residual, 145
- Poisson distribution, 36
- Poisson model, 36, 244
  - dynamic, 346
- Population-specific, 285
- Posterior covariance matrix, 61, 311
- Posterior curvatures, 299, 302
- Posterior distribution, 338
- Posterior mean, 61, 310, 338, 342
- Posterior mean estimation, 315
- Posterior mode, 61, 298, 299, 302, 340, 342
- Posterior mode approach, 372
- Power-divergence family, 109
- Prediction, 338
- Prediction step, 339
- Projection matrix, 146
- Projection pursuit regression, 209
- Quasi-information, 441
- Quasi-likelihood, 56, 118
- Quasi-likelihood Markov models, 246
- Quasi-likelihood model, 55, 246, 252, 258, 440
- Quasi-likelihood models, 261
- Quasi-score function, 56, 118, 258, 440
- Random effects, 285, 288, 292
  - two-stage, 285
- Random intercepts, 286
- Random slopes, 287
- Random walks, 334
- Regression
  - local, 184
- Rejection sampling, 316, 317, 450
- Reliability function, 386
- Repeated observations, 260
- Residual
  - Anscombe, 153
  - deviance, 153
  - Pearson, 130, 152
  - studentized, 152
- Response
  - binary, 24, 29, 122
  - multicategorical, 347
  - multinomial, 347
  - smoothing techniques for continuous response, 174
- Response function, 19, 20, 76, 434
  - exponential, 23
  - violation, 161
- Ridge estimator, 62
- RMLE, 290
- S-PLUS, 455, 461
- SAS, 455, 456
- SC, 142
- Scatterplot smoother, 174
- Schwarz' criterion, 142
- Score function, 39, 40, 105, 106, 435, 438
- Score statistic, 48, 58, 141, 251, 439
- Seasonal component, 334, 335
- Selection
  - all-subsets, 142
  - criteria, 140
  - stepwise, 143
  - variable, 139
- Semiparametric, 337
- Semiparametric Bayesian inference, 221
- Simple exponential family, 433
- Single index model, 209
- Smoothing, 174, 333, 338, 340, 420
  - choice of smoothing parameter, 197
  - Fisher-scoring, 354
  - Gauss-Newton, 352, 354
  - generalized extended, 352
  - kernel, 420
  - non-normal, 350
  - of a trend component, 341
  - penalized least squares criterion, 341
  - posterior mode, 352
  - running-line smoother, 183
  - simple neighborhood, 183
  - spline, 181
- Smoothing for non-Gaussian data, 193

- Smoothness priors, 221
- Spatio-temporal data, 376
- Splines, 181, 189, 403
  - cubic smoothing splines, 217
- SPSS, 457
- SPSS/PC+, 455
- STATA, 455, 459
- State space model, 332, 334, 369
  - Bayesian interpretation, 341
- State space models, 209
- Statistic
  - (log-)likelihood ratio, 439
  - Freeman-Tukey, 110
  - Kullback, 110
  - likelihood ratio, 251
  - Neyman, 110
  - Pearson, 107
  - Pearson goodness-of-fit, 152
  - power-divergence, 109
  - score, 251, 439
  - Wald, 251, 439
- Statistical inference, 289
- Strict stochastic ordering, 86, 99
- Structural time series, 334
- Subject-specific approaches, 264
- Survival function, 386
  - discrete, 396
- Test
  - for nonnested hypotheses, 167
  - generalized score, 168
  - generalized Wald, 168
  - goodness-of-fit, 47, 50, 107, 151
- Hausman, 165
- information matrix, 166
- likelihood ratio, 48, 50, 108, 141
- misspecification, 161
- modified likelihood ratio, 141
- modified Wald, 58
- quasi-likelihood ratio, 58
- score, 48, 58, 141, 251
- Wald, 48, 50, 141, 251
- Time series, 241, 260
- Time-varying covariate, 337
- Transition equation, 333
  - nonlinear, 350
- Transition model, 264
- Transition probabilities, 244
- Trend component, 334
- Truncated power series basis, 176
- Two-stage random effects, 285
- Variance components, 287, 300
- Variance function, 20, 22, 58
  - working, 56
- Varying-coefficient models, 208
- Wald, 251
- Wald statistic, 251, 439
- Wiener process, 225
- Wishart prior, 321, 373
- Working correlation, 274
- Working covariance, 268
- XPLORER, 455, 463

- Mielke/Berry*: Permutation Methods: A Distance Function Approach.
- Miller, Jr.*: Simultaneous Statistical Inference, 2nd edition.
- Mosteller/Wallace*: Applied Bayesian and Classical Inference: The Case of the Federalist Papers.
- Parzen/Tanabe/Kitagawa*: Selected Papers of Hirotugu Akaike.
- Politis/Romano/Wolf*: Subsampling.
- Ramsay/Silverman*: Functional Data Analysis.
- Rao/Toutenburg*: Linear Models: Least Squares and Alternatives.
- Read/Cressie*: Goodness-of-Fit Statistics for Discrete Multivariate Data.
- Reinsel*: Elements of Multivariate Time Series Analysis, 2nd edition.
- Reiss*: A Course on Point Processes.
- Reiss*: Approximate Distributions of Order Statistics: With Applications to Non-parametric Statistics.
- Rieder*: Robust Asymptotic Statistics.
- Rosenbaum*: Observational Studies.
- Rosenblatt*: Gaussian and Non-Gaussian Linear Time Series and Random Fields.
- Särndal/Swensson/Wretman*: Model Assisted Survey Sampling.
- Schervish*: Theory of Statistics.
- Shao/Tu*: The Jackknife and Bootstrap.
- Siegmund*: Sequential Analysis: Tests and Confidence Intervals.
- Simonoff*: Smoothing Methods in Statistics.
- Singpurwalla and Wilson*: Statistical Methods in Software Engineering:  
Reliability and Risk.
- Small*: The Statistical Theory of Shape.
- Sprott*: Statistical Inference in Science.
- Stein*: Interpolation of Spatial Data: Some Theory for Kriging.
- Taniguchi/Kakizawa*: Asymptotic Theory of Statistical Inference for Time Series.
- Tanner*: Tools for Statistical Inference: Methods for the Exploration of Posterior  
Distributions and Likelihood Functions, 3rd edition.
- Tong*: The Multivariate Normal Distribution.
- van der Vaart/Wellner*: Weak Convergence and Empirical Processes: With  
Applications to Statistics.
- Verbeke/Molenberghs*: Linear Mixed Models for Longitudinal Data.
- Weerahandi*: Exact Statistical Methods for Data Analysis.
- West/Harrison*: Bayesian Forecasting and Dynamic Models, 2nd edition.