

Springer Texts in Statistics

Advisors:

George Casella Stephen Fienberg Ingram Olkin

Springer

New York

Berlin

Heidelberg

Barcelona

Hong Kong

London

Milan

Paris

Singapore

Tokyo

- Alfred*: Elements of Statistics for the Life and Social Sciences
Berger: An Introduction to Probability and Stochastic Processes
Bilodeau and Brenner: Theory of Multivariate Statistics
Blom: Probability and Statistics: Theory and Applications
Brockwell and Davis: An Introduction to Times Series and Forecasting
Chow and Teicher: Probability Theory: Independence, Interchangeability, Martingales, Third Edition
Christensen: Plane Answers to Complex Questions: The Theory of Linear Models, Second Edition
Christensen: Linear Models for Multivariate, Time Series, and Spatial Data
Christensen: Log-Linear Models and Logistic Regression, Second Edition
Creighton: A First Course in Probability Models and Statistical Inference
Dean and Voss: Design and Analysis of Experiments
du Toit, Steyn, and Stumpf: Graphical Exploratory Data Analysis
Durrett: Essentials of Stochastic Processes
Edwards: Introduction to Graphical Modelling
Finkelstein and Levin: Statistics for Lawyers
Flury: A First Course in Multivariate Statistics
Jobson: Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design
Jobson: Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods
Kalbfleisch: Probability and Statistical Inference, Volume I: Probability, Second Edition
Kalbfleisch: Probability and Statistical Inference, Volume II: Statistical Inference, Second Edition
Karr: Probability
Keyfitz: Applied Mathematical Demography, Second Edition
Kiefer: Introduction to Statistical Inference
Kokoska and Nevison: Statistical Tables and Formulae
Kulkarni: Modeling, Analysis, Design, and Control of Stochastic Systems
Lehmann: Elements of Large-Sample Theory
Lehmann: Testing Statistical Hypotheses, Second Edition
Lehmann and Casella: Theory of Point Estimation, Second Edition
Lindman: Analysis of Variance in Experimental Design
Lindsey: Applying Generalized Linear Models
Madansky: Prescriptions for Working Statisticians
McPherson: Statistics in Scientific Investigation: Its Basis, Application, and Interpretation
Mueller: Basic Principles of Structural Equation Modeling: An Introduction to LISREL and EQS

(continued after index)

Springer Texts in Statistics *(continued from page ii)*

Nguyen and Rogers: Fundamentals of Mathematical Statistics: Volume I:

Probability for Statistics

Nguyen and Rogers: Fundamentals of Mathematical Statistics: Volume II:

Statistical Inference

Noether: Introduction to Statistics: The Nonparametric Way

Peters: Counting for Something: Statistical Principles and Personalities

Pfeiffer: Probability for Applications

Pitman: Probability

Rawlings, Pantula and Dickey: Applied Regression Analysis

Robert: The Bayesian Choice: A Decision-Theoretic Motivation

Robert and Casella: Monte Carlo Statistical Methods

Santner and Duffy: The Statistical Analysis of Discrete Data

Saville and Wood: Statistical Methods: The Geometric Approach

Sen and Srivastava: Regression Analysis: Theory, Methods, and Applications

Shao: Mathematical Statistics

Shumway and Stoffer: Time Series Analysis and Its Applications

Terrell: Mathematical Statistics: A Unified Introduction

Whittle: Probability via Expectation, Third Edition

Zacks: Introduction to Reliability Analysis: Probability Models and Statistical Methods

James K. Lindsey

Applying Generalized Linear Models

With 35 Illustrations



Springer

James K. Lindsey
Department of Biostatistics
Limburgs Universitair Centrum
3590 Diepenbeek
Belgium

Editorial Board

George Casella
Biometrics Unit
Cornell University
Ithaca, NY 14853
USA

Stephen Fienberg
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
USA

Ingram Olkin
Department of Statistics
Stanford University
Stanford, CA 94305
USA

Library of Congress Cataloging-in-Publication Data

Lindsey, James K.

Applying generalized linear models / J.K. Lindsey

p. cm. — (Springer texts in statistics)

Includes bibliographical references (p. —) and index.

ISBN 0-387-98218-3 (hardcover : alk. paper)

1. Linear models (Statistics) I. Title. II. Series

QA279.L594 1997

97-6926

519.5'3—dc21

© 1997 Springer-Verlag New York, Inc.

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

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

Preface

Generalized linear models provide a unified approach to many of the most common statistical procedures used in applied statistics. They have applications in disciplines as widely varied as agriculture, demography, ecology, economics, education, engineering, environmental studies and pollution, geography, geology, history, medicine, political science, psychology, and sociology, all of which are represented in this text.

In the years since the term was first introduced by Nelder and Wedderburn in 1972, generalized linear models have slowly become well known and widely used. Nevertheless, introductory statistics textbooks, and courses, still most often concentrate on the normal linear model, just as they did in the 1950s, as if nothing had happened in statistics in between. For students who will only receive one statistics course in their career, this is especially disastrous, because they will have a very restricted view of the possible utility of statistics in their chosen field of work. The present text, being fairly advanced, is not meant to fill that gap; see, rather, Lindsey (1995a).

Thus, throughout much of the history of statistics, statistical modelling centred around this normal linear model. Books on this subject abound. More recently, log linear and logistic models for discrete, categorical data have become common under the impetus of applications in the social sciences and medicine. A third area, models for survival data, also became a growth industry, although not always so closely related to generalized linear models. In contrast, relatively few books on generalized linear models, as such, are available. Perhaps the explanation is that normal and discrete, as well as survival, data continue to be the major fields of application. Thus, many students, even in relatively advanced statistics courses, do not have

an overview whereby they can see that these three areas, linear normal, categorical, and survival models, have much in common. Filling this gap is one goal of this book.

The introduction of the idea of generalized linear models in the early 1970s had a major impact on the way applied statistics is carried out. In the beginning, their use was primarily restricted to fairly advanced statisticians because the only explanatory material and software available were addressed to them. Anyone who used the first versions of GLIM will never forget the manual which began with pages of statistical formulae, before actually showing what the program was meant to do or how to use it.

One had to wait up to twenty years for generalized linear modelling procedures to be made more widely available in computer packages such as Genstat, Lisp-Stat, R, S-Plus, or SAS. Ironically, this is at a time when such an approach is decidedly outdated, not in the sense that it is no longer useful, but in its limiting restrictions as compared to what statistical models are needed and possible with modern computing power. What are now required, and feasible, are nonlinear models with dependence structures among observations. However, a unified approach to such models is only slowly developing and the accompanying software has yet to be put forth. The reader will find some hints in the last chapter of this book.

One of the most important accomplishments of generalized linear models has been to promote the central role of the likelihood function in inference. Many statistical techniques are proposed in the journals every year without the user being able to judge which are really suitable for a given data set. Most *ad hoc* measures, such as mean squared error, distinctly favour the symmetry and constant variance of the normal distribution. However, statistical models, which by definition provide a means of calculating the probability of the observed data, *can* be directly compared and judged: a model is preferable, or more likely, if it makes the observed data more probable (Lindsey, 1996b). This direct likelihood inference approach will be used throughout, although some aspects of competing methods are outlined in an appendix.

A number of central themes run through the book:

- the vast majority of statistical problems can be formulated, in a unified way, as regression models;
- any statistical models, for the same data, can be compared (whether nested or not) directly through the likelihood function, perhaps, with the aid of some model selection criterion such as the AIC;
- almost all phenomena are dynamic (stochastic) processes and, with modern computing power, appropriate models should be constructed;
- many so called “semi-” and “nonparametric” models (although not nonparametric inference procedures) are ordinary (often saturated)

generalized linear models involving factor variables; for inferences, one must condition on the observed data, as with the likelihood function.

Several important and well-known books on generalized linear models are available (Aitkin *et al.*, 1989; McCullagh and Nelder, 1989; Dobson, 1990; Fahrmeir and Tutz, 1994); the present book is intended to be complementary to them.

For this text, the reader is assumed to have knowledge of basic statistical principles, whether from a Bayesian, frequentist, or direct likelihood point of view, being familiar at least with the analysis of the simpler normal linear models, regression and ANOVA. The last chapter requires a considerably higher level of sophistication than the others.

This is a book about statistical *modelling*, not statistical inference. The idea is to show the unity of many of the commonly used models. In such a text, space is not available to provide complete detailed coverage of each specific area, whether categorical data, survival, or classical linear models. The reader will not become an expert in time series or spatial analysis by reading this book! The intention is rather to provide a taste of these different areas, and of their unity. Some of the most important specialized books available in each of these fields are indicated at the end of each chapter.

For the examples, every effort has been made to provide as much background information as possible. However, because they come from such a wide variety of fields, it is not feasible in most cases to develop prior theoretical models to which confirmatory methods, such as testing, could be applied. Instead, analyses primarily concern exploratory inference involving model selection, as is typical of practice in most areas of applied statistics. In this way, the reader will be able to discover many direct comparisons of the application of the various members of the generalized linear model family.

Chapter 1 introduces the generalized linear model in some detail. The necessary background in inference procedures is relegated to Appendices A and B, which are oriented towards the unifying role of the likelihood function and include details on the appropriate diagnostics for model checking. Simple log linear and logistic models are used, in Chapter 2, to introduce the first major application of generalized linear models. These log linear models are shown, in turn, in Chapter 3, to encompass generalized linear models as a special case, so that we come full circle. More general regression techniques are developed, through applications to growth curves, in Chapter 4. In Chapter 5, some methods of handling dependent data are described through the application of conditional regression models to longitudinal data. Another major area of application of generalized linear models is to survival, and duration, data, covered in Chapters 6 and 7, followed by spatial models in Chapter 8. Normal linear models are briefly reviewed in Chapter 9, with special reference to model checking by comparing them to

nonlinear and non-normal models. (Experienced statisticians may consider this chapter to be simpler than the others; in fact, this only reflects their greater familiarity with the subject.) Finally, the unifying methods of dynamic generalized linear models for dependent data are presented in Chapter 10, the most difficult in the text.

The two-dimensional plots were drawn with MultiPlot, for which I thank Alan Baxter, and the three-dimensional ones with Maple. I would also like to thank all of the contributors of data sets; they are individually cited with each table.

Students in the masters program in biostatistics at Limburgs University have provided many comments and suggestions throughout the years that I have taught this course there. Special thanks go to all the members of the Department of Statistics and Measurement Theory at Groningen University who created the environment for an enjoyable and profitable stay as Visiting Professor while I prepared the first draft of this text. Philippe Lambert, Patrick Lindsey, and four referees provided useful comments that helped to improve the text.

Diepenbeek
December, 1996

J.K.L.

Contents

Preface	v
1 Generalized Linear Modelling	1
1.1 Statistical Modelling	1
1.1.1 A Motivating Example	1
1.1.2 History	4
1.1.3 Data Generating Mechanisms and Models	6
1.1.4 Distributions	6
1.1.5 Regression Models	8
1.2 Exponential Dispersion Models	9
1.2.1 Exponential Family	10
1.2.2 Exponential Dispersion Family	11
1.2.3 Mean and Variance	11
1.3 Linear Structure	13
1.3.1 Possible Models	14
1.3.2 Notation for Model Formulae	15
1.3.3 Aliasing	16
1.4 Three Components of a GLM	18
1.4.1 Response Distribution or “Error Structure”	18
1.4.2 Linear Predictor	18
1.4.3 Link Function	18
1.5 Possible Models	20
1.5.1 Standard Models	20
1.5.2 Extensions	21

1.6	Inference	23
1.7	Exercises	25
2	Discrete Data	27
2.1	Log Linear Models	27
2.1.1	Simple Models	28
2.1.2	Poisson Representation	30
2.2	Models of Change	31
2.2.1	Mover–Stayer Model	32
2.2.2	Symmetry	33
2.2.3	Diagonal Symmetry	35
2.2.4	Long-term Dependence	36
2.2.5	Explanatory Variables	36
2.3	Overdispersion	37
2.3.1	Heterogeneity Factor	38
2.3.2	Random Effects	38
2.3.3	Rasch Model	39
2.4	Exercises	44
3	Fitting and Comparing Probability Distributions	49
3.1	Fitting Distributions	49
3.1.1	Poisson Regression Models	49
3.1.2	Exponential Family	52
3.2	Setting Up the Model	54
3.2.1	Likelihood Function for Grouped Data	54
3.2.2	Comparing Models	55
3.3	Special Cases	57
3.3.1	Truncated Distributions	57
3.3.2	Overdispersion	58
3.3.3	Mixture Distributions	60
3.3.4	Multivariate Distributions	63
3.4	Exercises	64
4	Growth Curves	69
4.1	Exponential Growth Curves	70
4.1.1	Continuous Response	70
4.1.2	Count Data	71
4.2	Logistic Growth Curve	72
4.3	Gomperz Growth Curve	74
4.4	More Complex Models	76
4.5	Exercises	82
5	Time Series	87
5.1	Poisson Processes	88
5.1.1	Point Processes	88

5.1.2	Homogeneous Processes	88
5.1.3	Nonhomogeneous Processes	88
5.1.4	Birth Processes	90
5.2	Markov Processes	91
5.2.1	Autoregression	93
5.2.2	Other Distributions	96
5.2.3	Markov Chains	101
5.3	Repeated Measurements	102
5.4	Exercises	103
6	Survival Data	109
6.1	General Concepts	109
6.1.1	Skewed Distributions	109
6.1.2	Censoring	109
6.1.3	Probability Functions	111
6.2	“Nonparametric” Estimation	111
6.3	Parametric Models	113
6.3.1	Proportional Hazards Models	113
6.3.2	Poisson Representation	113
6.3.3	Exponential Distribution	114
6.3.4	Weibull Distribution	115
6.4	“Semiparametric” Models	116
6.4.1	Piecewise Exponential Distribution	116
6.4.2	Cox Model	116
6.5	Exercises	117
7	Event Histories	121
7.1	Event Histories and Survival Distributions	122
7.2	Counting Processes	123
7.3	Modelling Event Histories	123
7.3.1	Censoring	124
7.3.2	Time Dependence	124
7.4	Generalizations	127
7.4.1	Geometric Process	128
7.4.2	Gamma Process	132
7.5	Exercises	136
8	Spatial Data	141
8.1	Spatial Interaction	141
8.1.1	Directional Dependence	141
8.1.2	Clustering	145
8.1.3	One Cluster Centre	147
8.1.4	Association	147
8.2	Spatial Patterns	149
8.2.1	Response Contours	149

8.2.2	Distribution About a Point	152
8.3	Exercises	154
9	Normal Models	159
9.1	Linear Regression	160
9.2	Analysis of Variance	161
9.3	Nonlinear Regression	164
9.3.1	Empirical Models	164
9.3.2	Theoretical Models	165
9.4	Exercises	167
10	Dynamic Models	173
10.1	Dynamic Generalized Linear Models	173
10.1.1	Components of the Model	173
10.1.2	Special Cases	174
10.1.3	Filtering and Prediction	174
10.2	Normal Models	175
10.2.1	Linear Models	176
10.2.2	Nonlinear Curves	181
10.3	Count Data	186
10.4	Positive Response Data	189
10.5	Continuous Time Nonlinear Models	191
Appendices		
A	Inference	197
A.1	Direct Likelihood Inference	197
A.1.1	Likelihood Function	197
A.1.2	Maximum Likelihood Estimate	199
A.1.3	Parameter Precision	202
A.1.4	Model Selection	205
A.1.5	Goodness of Fit	210
A.2	Frequentist Decision-making	212
A.2.1	Distribution of the Deviance Statistic	212
A.2.2	Analysis of Deviance	214
A.2.3	Estimation of the Scale Parameter	215
A.3	Bayesian Decision-making	215
A.3.1	Bayes' Formula	216
A.3.2	Conjugate Distributions	216
B	Diagnostics	221
B.1	Model Checking	221
B.2	Residuals	222
B.2.1	Hat Matrix	222
B.2.2	Kinds of Residuals	223

B.2.3	Residual Plots	225
B.3	Isolated Departures	226
B.3.1	Outliers	227
B.3.2	Influence and Leverage	227
B.4	Systematic Departures	228
References		231
Index		243