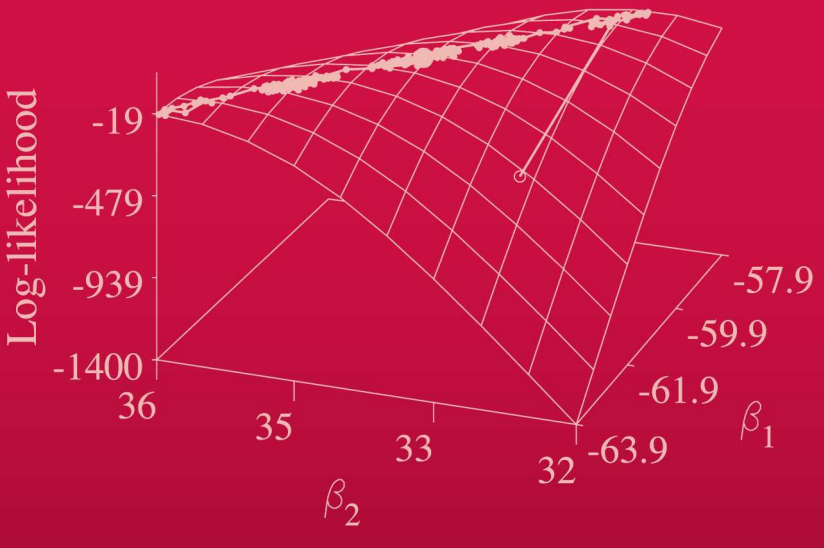
Texts in Statistical Science

An Introduction to Generalized Linear Models

Third Edition



Annette J. Dobson
Adrian G. Barnett

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

An Introduction to Generalized Linear Models

Third Edition

CHAPMAN & HALL/CRC
Texts in Statistical Science Series

Series Editors
Bradley P. Carlin, *University of Minnesota, USA*
Julian J. Faraway, *University of Bath, UK*
Martin Tanner, *Northwestern University, USA*
Jim Zidek, *University of British Columbia, Canada*

Analysis of Failure and Survival Data
P. J. Smith

The Analysis of Time Series — An Introduction, Sixth Edition
C. Chatfield

Applied Bayesian Forecasting and Time Series Analysis
A. Pole, M. West and J. Harrison

Applied Nonparametric Statistical Methods, Fourth Edition
P. Sprent and N.C. Smeeton

Applied Statistics — Handbook of GENSTAT Analysis
E.J. Snell and H. Simpson

Applied Statistics — Principles and Examples
D.R. Cox and E.J. Snell

Bayesian Data Analysis, Second Edition
A. Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin

Bayesian Methods for Data Analysis, Third Edition
B.P. Carlin and T.A. Louis

Beyond ANOVA — Basics of Applied Statistics
R.G. Miller, Jr.

Computer-Aided Multivariate Analysis, Fourth Edition
A.A. Afifi and V.A. Clark

A Course in Categorical Data Analysis
T. Leonard

A Course in Large Sample Theory
T.S. Ferguson

Data Driven Statistical Methods
P. Sprent

Decision Analysis — A Bayesian Approach
J.Q. Smith

Elementary Applications of Probability Theory, Second Edition
H.C. Tuckwell

Elements of Simulation
B.J.T. Morgan

Epidemiology — Study Design and Data Analysis, Second Edition
M. Woodward

Essential Statistics, Fourth Edition
D.A.G. Rees

Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models
J.J. Faraway

A First Course in Linear Model Theory
N. Ravishanker and D.K. Dey

Generalized Additive Models: An Introduction with R
S. Wood

Interpreting Data — A First Course in Statistics
A.J.B. Anderson

An Introduction to Generalized Linear Models, Third Edition
A.J. Dobson and A.G. Barnett

Introduction to Multivariate Analysis
C. Chatfield and A.J. Collins

Introduction to Optimization Methods and Their Applications in Statistics
B.S. Everitt

Introduction to Probability with R
K. Baclawski

Introduction to Randomized Controlled Clinical Trials, Second Edition
J.N.S. Matthews

Introduction to Statistical Methods for Clinical Trials
Thomas D. Cook and David L. DeMets

Large Sample Methods in Statistics
P.K. Sen and J. da Motta Singer

Linear Models with R
J.J. Faraway

Markov Chain Monte Carlo — Stochastic Simulation for Bayesian Inference, Second Edition
D. Gamerman and H.F. Lopes

Mathematical Statistics
K. Knight

Modeling and Analysis of Stochastic Systems
V. Kulkarni

Modelling Binary Data, Second Edition
D. Collett

Modelling Survival Data in Medical Research, Second Edition
D. Collett

Multivariate Analysis of Variance and Repeated Measures — A Practical Approach for Behavioural Scientists
D.J. Hand and C.C. Taylor

Multivariate Statistics — A Practical Approach
B. Flury and H. Riedwyl

Practical Data Analysis for Designed Experiments
B.S. Yandell

Practical Longitudinal Data Analysis
D.J. Hand and M. Crowder

Practical Statistics for Medical Research
D.G. Altman

A Primer on Linear Models
J.F. Monahan

Probability — Methods and Measurement
A. O'Hagan

Problem Solving — A Statistician's Guide, Second Edition
C. Chatfield

Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition
B.F.J. Manly

Readings in Decision Analysis
S. French

Sampling Methodologies with Applications
P.S.R.S. Rao

Statistical Analysis of Reliability Data
M.J. Crowder, A.C. Kimber, T.J. Sweeting, and R.L. Smith

Statistical Methods for Spatial Data Analysis
O. Schabenberger and C.A. Gotway

Statistical Methods for SPC and TQM
D. Bissell

Statistical Methods in Agriculture and Experimental Biology, Second Edition
R. Mead, R.N. Curnow, and A.M. Hasted

Statistical Process Control — Theory and Practice, Third Edition
G.B. Wetherill and D.W. Brown

Statistical Theory, Fourth Edition
B.W. Lindgren

Statistics for Accountants
S. Letchford

Statistics for Epidemiology
N.P. Jewell

Statistics for Technology — A Course in Applied Statistics, Third Edition
C. Chatfield

Statistics in Engineering — A Practical Approach
A.V. Metcalfe

Statistics in Research and Development, Second Edition
R. Caulcutt

Survival Analysis Using S — Analysis of Time-to-Event Data
M. Tableman and J.S. Kim

The Theory of Linear Models
B. Jørgensen

Time Series Analysis
H. Madsen

Texts in Statistical Science

An Introduction to Generalized Linear Models

Third Edition

Annette J. Dobson
University of Queensland
Herston, Australia

Adrian G. Barnett
Queensland University of Technology
Kelvin Grove, Australia

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2008 by Taylor & Francis Group, LLC
Chapman & Hall/CRC is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-58488-950-2 (Softcover)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Dobson, Annette J., 1945-
An introduction to generalized linear models / Annette J. Dobson and Adrian G. Barnett. -- 3rd ed.
p. cm. -- (Chapman & Hall/CRC texts in statistical science series ; 77)
Includes bibliographical references and index.
ISBN 978-1-58488-950-2 (alk. paper)
1. Linear models (Statistics) I. Barnett, Adrian G. II. Title. III. Series.

QA276.D589 2008
519.5--dc22 2008013034

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface

The original purpose of the book was to present a unified theoretical and conceptual framework for statistical modelling in a way that was accessible to undergraduate students and researchers in other fields.

The second edition was expanded to include nominal and ordinal logistic regression, survival analysis and analysis of longitudinal and clustered data. It relied more on numerical methods, visualizing numerical optimization and graphical methods for exploratory data analysis and checking model fit. These features have been extended further in this new edition.

The third edition contains three new chapters on Bayesian analysis. The fundamentals of Bayesian theory were written long before the development of classical theory but practical Bayesian analysis has only recently become available. This availability is mostly thanks to Markov chain Monte Carlo methods which are introduced in Chapter 13. The increased availability of Bayesian analysis means that more people with a classical knowledge of statistics are trying Bayesian methods for generalized linear models. Bayesian analysis offers significant advantages over classical methods because of the ability formally to incorporate prior information, greater flexibility and an ability to solve complex problems.

This edition has also been updated with Stata and R code, which should help the practical application of generalized linear models. The chapters on Bayesian analyses contain R and WinBUGS code.

The data sets and outline solutions of the exercises are available on the publisher's website: http://www.crcpress.com/e_products/downloads/.

We are grateful to colleagues and students at the Universities of Queensland and Newcastle, Australia, and those taking postgraduate courses through the Biostatistics Collaboration of Australia for their helpful suggestions and comments about the material.

Annette Dobson and Adrian Barnett
Brisbane

CHAPTER 1

Introduction

1.1 Background

This book is designed to introduce the reader to generalized linear models; these provide a unifying framework for many commonly used statistical techniques. They also illustrate the ideas of statistical modelling.

The reader is assumed to have some familiarity with classical statistical principles and methods. In particular, understanding the concepts of estimation, sampling distributions and hypothesis testing is necessary. Experience in the use of t-tests, analysis of variance, simple linear regression and chi-squared tests of independence for two-dimensional contingency tables is assumed. In addition, some knowledge of matrix algebra and calculus is required.

The reader will find it necessary to have access to statistical computing facilities. Many statistical programs, languages or packages can now perform the analyses discussed in this book. Often, however, they do so with a different program or procedure for each type of analysis so that the unifying structure is not apparent.

Some programs or languages which have procedures consistent with the approach used in this book are **Stata**, **R**, **S-PLUS**, **SAS** and **Genstat**. For Chapters 13 to 14 programs to conduct Markov chain Monte Carlo methods are needed and WinBUGS has been used here. This list is not comprehensive as appropriate modules are continually being added to other programs.

In addition, anyone working through this book may find it helpful to be able to use mathematical software that can perform matrix algebra, differentiation and iterative calculations.

1.2 Scope

The statistical methods considered in this book all involve the analysis of relationships between measurements made on groups of subjects or objects. For example, the measurements might be the heights or weights and the ages of boys and girls, or the yield of plants under various growing conditions. We use the terms **response**, **outcome** or **dependent variable** for measurements that are free to vary in response to other variables called **explanatory variables** or **predictor variables** or **independent variables**—although this last term can sometimes be misleading. Responses are regarded as random variables. Explanatory variables are usually treated as though they are non-

random measurements or observations; for example, they may be fixed by the experimental design.

Responses and explanatory variables are measured on one of the following scales.

1. **Nominal** classifications: e.g., red, green, blue; yes, no, do not know, not applicable. In particular, for **binary**, **dichotomous** or **binomial** variables there are only two categories: male, female; dead, alive; smooth leaves, serrated leaves. If there are more than two categories the variable is called **polychotomous, polytomous** or **multinomial**.

2. **Ordinal** classifications in which there is some natural order or ranking between the categories: e.g., young, middle aged, old; diastolic blood pressures grouped as ≤ 70, 71–90, 91–110, 111–130, ≥ 131 mmHg.

3. **Continuous** measurements where observations may, at least in theory, fall anywhere on a continuum: e.g., weight, length or time. This scale includes both **interval scale** and **ratio scale** measurements—the latter have a well-defined zero. A particular example of a continuous measurement is the time until a specific event occurs, such as the failure of an electronic component; the length of time from a known starting point is called the **failure time**.

Nominal and ordinal data are sometimes called **categorical** or **discrete variables** and the numbers of observations, **counts** or **frequencies** in each category are usually recorded. For continuous data the individual measurements are recorded. The term **quantitative** is often used for a variable measured on a continuous scale and the term **qualitative** for nominal and sometimes for ordinal measurements. A qualitative, explanatory variable is called a **factor** and its categories are called the **levels** for the factor. A quantitative explanatory variable is sometimes called a **covariate**.

Methods of statistical analysis depend on the measurement scales of the response and explanatory variables.

This book is mainly concerned with those statistical methods which are relevant when there is just *one response variable* although there will usually be several explanatory variables. The responses measured on different subjects are usually assumed to be statistically independent random variables although this requirement is dropped in Chapter 11, which is about correlated data, and in subsequent chapters. Table 1.1 shows the main methods of statistical analysis for various combinations of response and explanatory variables and the chapters in which these are described.The last three chapters are devoted to Bayesian methods which substantially extend these analyses.

The present chapter summarizes some of the statistical theory used throughout the book. Chapters 2 through 5 cover the theoretical framework that is common to the subsequent chapters. Later chapters focus on methods for analyzing particular kinds of data.

Chapter 2 develops the main ideas of classical or frequentist statistical modelling. The modelling process involves four steps:

Table 1.1 *Major methods of statistical analysis for response and explanatory variables measured on various scales and chapter references for this book. Extensions of these methods from a Bayesian perspective are illustrated in Chapters 12–14.*

Response (chapter)	Explanatory variables	Methods
Continuous (Chapter 6)	Binary	t-test
	Nominal, >2 categories	Analysis of variance
	Ordinal	Analysis of variance
	Continuous	Multiple regression
	Nominal & some continuous	Analysis of covariance
	Categorical & continuous	Multiple regression
Binary (Chapter 7)	Categorical	Contingency tables Logistic regression
	Continuous	Logistic, probit & other dose-response models
	Categorical & continuous	Logistic regression
Nominal with >2 categories (Chapters 8 & 9)	Nominal	Contingency tables
	Categorical & continuous	Nominal logistic regression
Ordinal (Chapter 8)	Categorical & continuous	Ordinal logistic regression
Counts (Chapter 9)	Categorical	Log-linear models
	Categorical & continuous	Poisson regression
Failure times (Chapter 10)	Categorical & continuous	Survival analysis (parametric)
Correlated responses (Chapter 11)	Categorical & continuous	Generalized estimating equations Multilevel models

1. Specifying models in two parts: equations linking the response and explanatory variables, and the probability distribution of the response variable.
2. Estimating fixed but unknown parameters used in the models.
3. Checking how well the models fit the actual data.
4. Making inferences; for example, calculating confidence intervals and testing hypotheses about the parameters.

The next three chapters provide the theoretical background. Chapter 3 is about the **exponential family of distributions**, which includes the Normal, Poisson and Binomial distributions. It also covers **generalized linear models** (as defined by Nelder and Wedderburn 1972). Linear regression and many other models are special cases of generalized linear models. In Chapter 4 methods of classical estimation and model fitting are described.

Chapter 5 outlines frequentist methods of statistical inference for generalized linear models. Most of these methods are based on how well a model describes the set of data. For example, **hypothesis testing** is carried out by first specifying alternative models (one corresponding to the null hypothesis and the other to a more general hypothesis). Then test statistics are calculated which measure the "goodness of fit" of each model and these are compared. Typically the model corresponding to the null hypothesis is simpler, so if it fits the data about as well as a more complex model it is usually preferred on the grounds of parsimony (i.e., we retain the null hypothesis).

Chapter 6 is about **multiple linear regression** and **analysis of variance** (ANOVA). Regression is the standard method for relating a continuous response variable to several continuous explanatory (or predictor) variables. ANOVA is used for a continuous response variable and categorical or qualitative explanatory variables (factors). **Analysis of covariance** (ANCOVA) is used when at least one of the explanatory variables is continuous. Nowadays it is common to use the same computational tools for all such situations. The terms **multiple regression** or **general linear model** are used to cover the range of methods for analyzing one continuous response variable and multiple explanatory variables.

Chapter 7 is about methods for analyzing binary response data. The most common one is **logistic regression** which is used to model relationships between the response variable and several explanatory variables which may be categorical or continuous. Methods for relating the response to a single continuous variable, the dose, are also considered; these include **probit analysis** which was originally developed for analyzing dose-response data from bioassays. Logistic regression has been generalized to include responses with more than two nominal categories (**nominal**, **multinomial**, **polytomous** or **polychotomous logistic regression**) or ordinal categories (**ordinal logistic regression**). These methods are discussed in Chapter 8.

Chapter 9 concerns **count** data. The counts may be frequencies displayed in a **contingency table** or numbers of events, such as traffic accidents, which need to be analyzed in relation to some "exposure" variable such as the num-

ber of motor vehicles registered or the distances travelled by the drivers. Modelling methods are based on assuming that the distribution of counts can be described by the Poisson distribution, at least approximately. These methods include **Poisson regression** and **log-linear models**.

Survival analysis is the usual term for methods of analyzing failure time data. The parametric methods described in Chapter 10 fit into the framework of generalized linear models although the probability distribution assumed for the failure times may not belong to the exponential family.

Generalized linear models have been extended to situations where the responses are correlated rather than independent random variables. This may occur, for instance, if they are **repeated measurements** on the same subject or measurements on a group of related subjects obtained, for example, from **clustered sampling**. The method of **generalized estimating equations** (GEEs) has been developed for analyzing such data using techniques analogous to those for generalized linear models. This method is outlined in Chapter 11 together with a different approach to correlated data, namely **multilevel modelling** in which some parameters are treated as random variables rather than fixed but unknown constants. Multilevel modelling involves both fixed and random effects (mixed models) and relates more closely to the Bayesian approach to statistical analysis.

The main concepts and methods of Bayesian analysis are introduced in Chapter 12. In this chapter the relationships between classical or frequentist methods and Bayesian methods are outlined. In addition the software WinBUGS which is used to fit Bayesian models is introduced.

Bayesian models are usually fitted using computer-intensive methods based on Markov chains simulated using techniques based on random numbers. These methods are described in Chapter 13. This chapter uses some examples from earlier chapters to illustrate the mechanics of Markov chain Monte Carlo (MCMC) calculations and to demonstrate how the results allow much richer statistical inferences than are possible using classical methods.

Chapter 14 comprises several examples, introduced in earlier chapters, which are reworked using Bayesian analysis. These examples are used to illustrate both conceptual issues and practical approaches to estimation, model fitting and model comparisons using WinBUGS.

Further examples of generalized linear models are discussed in the books by McCullagh and Nelder (1989), Aitkin et al. (2005) and Myers et al. (2001). Also there are many books about specific generalized linear models such as Hosmer and Lemeshow (2000), Agresti (1990, 1996), Collett (2003a, 2003b), Diggle et al. (2002) and Goldstein (2003).

1.3 Notation

Generally we follow the convention of denoting random variables by upper case italic letters and observed values by the corresponding lower case letters. For example, the observations $y_1, y_2, ..., y_n$ are regarded as realizations of the