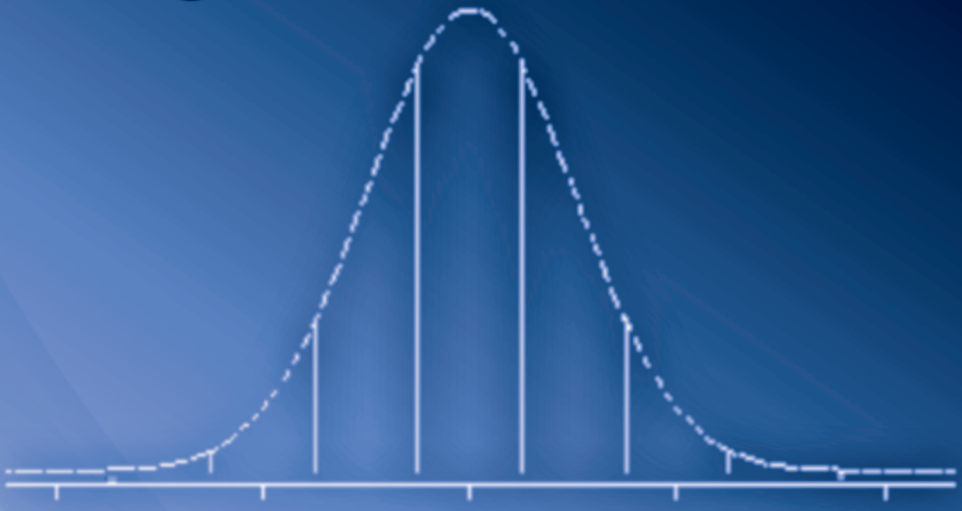
Multivariate Generalized Linear Mixed Models Using R



Damon M. Berridge
Robert Crouchley

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Multivariate Generalized Linear Mixed Models Using R

Damon M. Berridge
Robert Crouchley

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

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

Version Date: 20111012

International Standard Book Number-13: 978-1-4398-1327-0 (eBook - PDF)

Contents

This page intentionally left blank

List of Figures

This page intentionally left blank

List of Tables

This page intentionally left blank

List of Applications

List of Datasets

Preface

The main aims of this book are to provide an introduction to the principles of modelling as applied to longitudinal data from panel and related studies with the necessary statistical theory, and to describe the application of these principles to the analysis of a wide range of examples using the Sabre software (`http://sabre.lancs.ac.uk/`) from within R.

This material on multivariate generalized linear mixed models arises from the activities at the Economic and Social Research Council (ESRC)-funded Colaboratory for Quantitative e-Social Science (CQeSS) at Lancaster University from 2003 to 2008. Sabre is a program for the statistical analysis of multi-process event/response sequences. These responses can take the form of binary, ordinal, count and linear recurrent events. The response sequences can also be of different types, for example, a linear response (wages) and a binary one (trade union membership). Such multi-process data are common in many research areas, for example, in the analysis of work and life histories from the British Household Panel Survey or the German Socio-Economic Panel Study where researchers often want to disentangle state dependence (the effect of previous responses or related outcomes) from any omitted effects that might be present in recurrent behaviour (for example, unemployment). Understanding the need to disentangle these generic substantive issues dates back to the study of accident proneness in the 1950s and has since been discussed in many applied areas, including consumer behaviour and voting behaviour. These issues, and others relating to the analysis of longitudinal or event history data, are discussed in more detail in the following text:

- Shahtahmasebi, S. and Berridge, D. (2010) *Conceptualizing Human Behaviour in Health and Social Research: A Practical Guide to Data Analysis*, New York: Nova

Some key contributions in the References, including a number of Heckman's seminal works, have been reprinted in the following series:

1. Penn, R. and Berridge, D. (2010) *Social Statistics Volume 1: The Fundamentals of Descriptive Social Statistics*, London: Sage
2. Penn, R. and Berridge, D. (2010) *Social Statistics Volume 2: The Development of Statistical Modelling*, London: Sage

3. Penn, R. and Berridge, D. (2010) *Social Statistics Volume 3: Statistical Modelling of Longitudinal Data*, London: Sage

4. Penn, R. and Berridge, D. (2010) *Social Statistics Volume 4: Statistical Modelling of Ordinal Categorical Data*, London: Sage

Those contributions appearing in this series are indicated by asterisks in the References. One asterisk indicates Volume 1, two asterisks indicate Volume 2, and so on.

Sabre can also be used to model collections of single sequences such as may occur in medical trials on the number of headaches experienced over a sequence of weeks, or in single-equation descriptions of cross-sectional clustered data such as the educational attainment of children in schools.

Sabre is available in three forms: (1) stand-alone (as discussed in Shahtahmasebi and Berridge, 2010), (2) the R plugin (as discussed in the current text), and (3) the Stata plugin (as discussed on the Sabre web page — see above).

The class of models that can be estimated by Sabre may be termed Multivariate Generalized Linear Mixed Models (MGLMMs). These models have special features to help them disentangle state dependence from the incidental parameters (omitted or unobserved effects). The incidental parameters can be treated as random or fixed. The random effects models can be estimated with standard Gaussian quadrature or adaptive Gaussian quadrature. Quadrature methods (and particularly adaptive Gaussian quadrature) are the most reliable way of handling random effects in MGLMMs, as the adequacy of the numerical integration can be improved by adding more quadrature points. The number of quadrature points required will depend on the model being estimated. If additional quadrature points fail to improve the log likelihood, then we have found an accurate evaluation of the integral. Even though the linear model integral has a closed form solution, we do not use it as it cannot easily be used in multivariate models when some of the joint sequences do not have interval level responses. Also current computational facilities on many desktop computers often make the delay involved in using numerical integration for the linear model negligible for many small to medium-sized data sets. 'End effects' can also be added to the models to accommodate 'stayers' or 'non-susceptibles'. The fixed effects algorithm we have developed uses code for large sparse matrices from the Harwell Subroutine Library; see `http://www.cse.scitech.ac.uk/nag/hsl/`.

Also included in Sabre is the option to undertake all the calculations using increased accuracy. Numerical underflow and overflow often occur in the estimation process for models with incidental parameters. We suppose that many of the alternative software systems truncate their

calculations without informing the user when this happens as there is little discussion of this in their respective user manuals.

This book is written in a way that we have found appropriate for some of our short courses. The book starts by discussing members of the family of generalized linear models and gradually adds complexity to the modelling framework by incorporating random effects. We then review the generalized linear model notation before illustrating a range of more substantively appropriate random effects models, for example, the three-level model, multivariate (in particular, bivariate and trivariate) models, endpoint, event history and state dependence models. The MGLMMs are estimated using either standard Gaussian quadrature or adaptive Gaussian quadrature. The book compares two-level fixed and random effects linear models. Additional information on quadrature, model estimation and endogenous variables is included in Appendix A. Appendix B contains an introduction to R and some examples of using R to pre-process the data for Sabre.

There are two other related SabreR booklets available from the Sabre web page:

- Exercises for SabreR
- Solutions Manual for SabreR Exercises

These booklets contain the exercises and solutions on small data sets that have been written to accompany this book. These exercises will run quickly on a desktop PC.

Drafts of the chapters of this book were developed and revised in the process of preparing and delivering short courses in 'Statistical Modelling using Sabre', 'Multilevel Modelling' and 'Event History Analysis' given at CQeSS and the Department of Mathematics and Statistics at Lancaster University and elsewhere. We are grateful to many of the students of these courses who are from a range of backgrounds (for example, computational science and the social sciences) and whose comments and criticisms improved these early drafts. We think that the book should serve as a training manual for postgraduate Masters and research students, and as a self-teaching manual for data analysts.

If you have any suggestions as to how this book could be improved—for instance by the addition of other material—please let us know via the Sabre mailing list, `sabre@lancaster.ac.uk`.

We accept no liability for anything that might happen as a consequence of your use of Sabre, though we are happy to accept recognition of its successful use.

Dr. Damon M. Berridge and Professor Robert Crouchley
Lancaster University
February 2011

This page intentionally left blank

Acknowledgments

Many thanks to Dr. Iraj Kazemi for helping to draft the material in the first ten chapters of this book. Thanks to Professor Richard B. Davies for inspiring the early development of Sabre (Poisson and logit models with endpoints).

Many thanks to Daniel Grose for writing the R side of the SabreR library. Dan also wrote much of the introductory material on R in Appendix B. David Stott and John Pritchard undertook all the recent development work on Sabre. Dave wrote the standard Gaussian and adaptive Gaussian quadrature algorithms. John wrote the algorithm for manipulating the large sparse matrices used by the fixed effects estimator.

This work was supported by the following ESRC research grants:

- RES-149-28-1003: The Colaboratory for Quantitative e-Social Science (E-Social Science Centre Lancaster Node), principal investigator: Professor Robert Crouchley
- RES-149-25-0010: An OGSA Component-Based Approach to Middleware for Statistical Modelling, principal investigator: Professor Robert Crouchley
- RES-576-25-0019: The Lancaster-Warwick-Stirling Node: Developing Statistical Modelling in the Social Sciences (National Centre for Research Methods (NCRM) Phase 2), principal investigator: Professor Brian Francis

The NCRM Phase 2 grant was particularly important for the development of the bivariate ordered response model reported in Chapter 11.

Finally, we wish to express our gratitude to Professor Roger Penn for his assistance in proofreading the final draft of this book.

This page intentionally left blank

1

Introduction

A major objective of this book is to provide data analysts with the tools to analyze large and complex datasets using methodologically sound models, thereby enabling them to answer increasingly complex research questions. The statistical software used in this book is SabreR. This is a version of the package Sabre, for the statistical analysis of multi-process event/response sequences, which has been implemented within the R environment.

These responses can take the form of binary, ordinal, count and linear recurrent events. The response sequences can also be of different types, for example, a linear response (wages) and a binary response (trade union membership). Such multi-process data are common in many research areas, for example, in the analysis of work and life histories from the British Household Panel Survey or the German Socio-Economic Panel Study where researchers often want to disentangle state dependence (the effect of previous responses or related outcomes) from any omitted effects that might be present in recurrent behaviour (unemployment).

Understanding of the need to disentangle these generic substantive issues dates back to the study of accident proneness [14] and has been discussed in many applied areas, including consumer behaviour [75] and voting behaviour [34].

SabreR can also be used to model collections of single sequences such as those that may occur in medical trials, for example, headaches and epileptic seizures [29,30], or in single-equation descriptions of cross-sectional clustered data such as the educational attainment of children in schools.

The class of models that can be estimated by SabreR may be called multivariate generalized linear mixed models. These models have special features added to standard models to help us disentangle state dependence from the incidental parameters (omitted or unobserved effects). The incidental parameters can be treated as random or fixed, the random effects models being estimated using standard Gaussian quadrature or adaptive Gaussian quadrature. 'End effects' can also be added to the models to accommodate 'stayers' or 'non-susceptibles', resulting in a more parsimonious model which provides a better fit to the data with

fewer parameters than a non-parametric specification of the random effects. The fixed effects algorithm we have developed uses code for large sparse matrices from the Harwell Subroutine Library [49].

SabreR also includes the option to undertake all of the calculations using increased accuracy. This is important because numerical underflow and overflow often occur in the estimation process for models with incidental parameters.

Chapters 2 and 3 cover the analysis of single-level data of various types: continuous, binary, ordinal and count data using univariate generalized linear models. The material covered in these chapters is summarized in Chapter 4. Chapters 5 to 8 extend these models to handle multi-level, specifically two-level, data of various types: continuous, binary, ordinal and count data, using univariate generalized linear mixed models. The models considered in Chapters 5 to 8 are summarized in Chapter 9, and are generalized to handle three-level data in Chapter 10.

A key feature of this book is the emphasis on the application of statistical models to real-life examples. At the heart of each chapter will be a fully worked example. In addition, readers will have the opportunity to apply these statistical models and to interpret the resulting output through a large number of exercises spanning a wide variety of areas of application. The exercises illustrating the use of models for continuous/interval scale data in Chapters 2 and 5 are based on the following examples:

Example 1.1. Psychological distress

Twelve students completed the twelve-item version of Goldberg's General Health Questionnaire (GHQ) [42]. The questionnaire was completed by each student on two different occasions, separated by three days. A psychological distress score was computed, on the basis of the twelve GHQ items, for each student on each of the two occasions [39]. These student-occasion-specific scores are saved in the file `ghq2.tab`.

Example 1.2. Essay grading (continuous response)

Johnson and Albert [66] analyzed data on the grading of essays by several experts. Essays were graded on a scale between 1 and 10, with a score of 10 corresponding to 'excellent'. In this example, we consider a subset of the data limited to the grades given to 198 essays by markers 1 and 4. This subset of data is stored in the data file `grader1.tab` which may be found on the Sabre web page. The grades given by markers 1 and 4 are stacked in a single column `grade` in the file `grader2.tab`. This file also includes an identifier which distinguishes between the two graders, in other words, the variable `dg4` which takes value 1 if the grader is number 4, and value 0 otherwise. Alternative treatments of the response are considered in *Examples 1.7* and *1.11*.

Example 1.3. Educational attainment

Garner and Raudenbush [41] and Raudenbush and Bryk [93] studied the role of school and/or neighbourhood effects on the educational attainment of young people, from one Scottish Local Education Authority, who left school between 1984 and 1986. The primary outcome of interest is a young person's combined end-of-school educational attainment as measured by his/her grades.

Explanatory variables are available at two levels: (i) the individual young person level and (ii) the school and/or neighbourhood level. Most explanatory variables present in the dataset are specific to each young person. These variables include: young person's gender; verbal reasoning quotient and reading ability as measured by tests in primary school at age 11–12; father's occupation and education. The one school/neighbourhood-specific explanatory variable is an index of social deprivation for the local community within which the young person lived. The data are stored in the file `neighbourhood.tab` on the Sabre web page.

Example 1.4. Unemployment claims

Indiana's enterprise zone programme provided tax credits for cities with high poverty and unemployment levels. In a bid to establish whether those cities targetted by the programme had significantly lower unemployment claims than those cities lying outside enterprise zones, Papke [85] analyzed annual data from 1980 to 1988. The dataset (`ezunem2.tab`) comprises the number of unemployment claims in 22 cities, and whether each city was located within an enterprise zone, in each of the nine years 1980 to 1988.

Example 1.5. Wage determinants

Vella and Verbeek [103] analyzed annual data on 545 males from the Youth Sample of the US National Longitudinal Survey for the period 1980 to 1987. The version of the data used in this book (`wagepan.tab`) was obtained from Wooldridge [106]. We wish to relate the outcome of primary interest, log hourly wage (in US dollars), to a time-invariant factor (ethnicity) and a variety of time-dependent explanatory variables. Those variables allowed to vary over time include respondent demographics (marital status, region of US lived in, rural/urban area lived in), education (years of schooling), labour market experience and trade union membership. These data are re-considered in *Example 1.8*, where trade union membership is regarded as the binary response of interest.

Having analyzed these data in Chapters 2 and 5, we will return to this dataset on further occasions in this book. In Chapter 11, in the context of bivariate models, we will estimate a joint model for wages and

trade union membership. We will allow trade union membership to be endogenous in the wage equation. In Chapter 14, we will use the data on trade union membership to illustrate Wooldridge's [107] treatment of the initial conditions problem in first-order Markov models. In Chapter 15, we compare and contrast the inferences made when we first assume fixed effects and then proceed under the assumption of random effects. We will use these data to relate log wages to time-varying explanatory variables such as number of years of labour market experience, marital status and trade union membership, and to time-invariant factors including race and education.

Example 1.6. Pupil rating of school managers

856 pupils in 94 schools were asked to rate the performance of their school managers/directors on the basis of six questions, each response recorded on a four-point scale [64]. The response to each item given by each pupil is presented in the dataset `manager.tab`. Pupil-specific explanatory variables are gender and school year. School-specific factors are gender of the school manager/director and type of school which is classified into the following three categories: 'general (AVO)', 'professional (MBO&T)' and 'day/evening'.

The exercises illustrating the use of models for binary data in Chapters 3 and 6 are based on the following examples:

Example 1.7. Essay grading (binary response)

In an extension to *Example 1.1*, we use data on the grades given to 198 essays by markers 1 to 5. Essays were graded on a scale from 1 to 10, with 10 classified as 'excellent'. For the purposes of the current example, the original essay grading variable is converted into a binary response variable, labelled as `pass` in the dataset `essays2.tab`. The variable `pass` takes the value 1 for grades 5 to 10, and value 0 for grades 1 to 4. The primary objective in this example is to test for significant differences in this binary response between markers, whilst adjusting for six explanatory variables which characterize the 198 essays.

Four of these factors are lexical in nature: average word length (`wordlength`), square root of the number of words (`sqrtwords`), average sentence length (`sentlength`) and proportion of words in the essay which are prepositions (`prepos`). A fifth explanatory variable is related to punctuation: number of commas, multiplied by 100 and divided by the total number of words in the essay (`commas`). The sixth factor is the percentage of words in the essay which are spelt incorrectly (`errors`).

Example 1.8. Trade union membership

In *Example 1.5*, we related data from the Youth Sample of the US