Generalized Linear Models for Insurance Data

Actuaries should have the tools they need. Generalized linear models are used in the insurance industry to support critical decisions. Yet no text introduces GLMs in this context and addresses problems specific to insurance data. Until now.

Practical and rigorous, this books treats GLMs, covers all standard exponential family distributions, extends the methodology to correlated data structures, and discusses other techniques of interest and how they contrast with GLMs. The focus is on issues which are specific to insurance data and all techniques are illustrated on data sets relevant to insurance.

Exercises and data-based practicals help readers to consolidate their skills, with solutions and data sets given on the companion website. Although the book is package-independent, SAS code and output examples feature in an appendix and on the website. In addition, R code and output for all examples are provided on the website.

International Series on Actuarial Science

The International Series on Actuarial Science, published by Cambridge University Press in conjunction with the Institute of Actuaries and the Faculty of Actuaries, will contain textbooks for students taking courses in or related to actuarial science, as well as more advanced works designed for continuing professional development or for describing and synthesizing research. The series will be a vehicle for publishing books that reflect changes and developments in the curriculum, that encourage the introduction of courses on actuarial science in universities, and that show how actuarial science can be used in all areas where there is long-term financial risk.

GENERALIZED LINEAR MODELS FOR INSURANCE DATA

PIET DE JONG
Department of Actuarial Studies, Macquarie University, Sydney

GILLIAN Z. HELLER
Department of Statistics, Macquarie University, Sydney

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org

First published 2008

Printed in the United Kingdom at the University Press, Cambridge

A catalog record for this publication is available from the British Library

ISBN 978-0-521-87914-9 hardback

Contents

Preface

The motivation for this book arose out of our many years of teaching actuarial students and analyzing insurance data. Generalized linear models are ideally suited to the analysis of non-normal data which insurance analysts typically encounter. However the acceptance, uptake and understanding of this methodology has been slow in insurance compared to other disciplines. Part of the reason may be the lack of a suitable textbook geared towards an actuarial audience. This book seeks to address that need.

We have tried to make the book as practical as possible. Analyses are based on real data. All but one of the data sets are available on the companion website to this book:

`http://www.acst.mq.edu.au/GLMsforInsuranceData/`.

Computer code and output for all examples is given in Appendix 1.

The SAS software is widely used in the insurance industry. Hence computations in this text are illustrated using SAS. The statistical language R is used where computations are not conveniently performed in SAS. In addition, R code and output for all the examples is provided on the companion website. Exercises are given at the end of chapters, and fully worked solutions are available on the website.

The body of the text is independent of software or software "runs." In most cases, fitting results are displayed in tabular form. Remarks on computer implementation are confined to paragraphs headed "SAS notes" and "Implementation" and these notes can be skipped without loss of continuity.

Readers are assumed to be familiar with the following statistical concepts: discrete and continuous random variables, probability distributions, estimation, hypothesis testing, and linear regression (the normal model). Relevant basics of probability and estimation are covered in Chapters 2 and 3, but familiarity with these concepts is assumed. Normal linear regression is covered in Chapter 4: again it is expected readers have previously encountered the material. This chapter sets the scene for the rest of the book and discuss concepts that are applicable to regression models in general.

Excessive notation is avoided. The meanings of symbols will be clear from the context. For example a response variable is denoted by y, and there is no notational distinction between the random variable and its realization. The vector of outcomes is also denoted by y. Derivatives are denoted using the dot notation: $\dot{f}(y)$ and double dots denote second derivatives. This avoids confusion with the notation for matrix transposition X', frequently required in the same mathematical expressions. Tedious and generally uninformative subscripting is avoided. For example, the expression $y = x'\beta$ used in this text can be written as $y_i = x_i'\beta$, or even more explicitly and laboriously as $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$. Generally such laboring is avoided. Usually x denotes the vector $(1, x_1, \ldots, x_p)'$ and β denotes $(\beta_0, \ldots, \beta_p)'$. The equivalence symbol "$\equiv$" is used when a quantity is defined. The symbol "$\sim$" denotes "distributed as," either exactly or approximately.

Both authors contributed equally to this book, and authorship order was determined by the alphabetical convention. Much of the book was written while GH was on sabbatical leave at CSIRO Mathematical and Information Sciences, Sydney, whom she thanks for their hospitality. We thank Christine Lu for her assistance. And to our families Dana, Doryon, Michelle and Dean, and Steven, Ilana and Monique, our heartfelt thanks for putting up with the many hours that we spent on this text.

Piet de Jong
Gillian Heller

Sydney, 2007