

Cambridge University Press

978-0-521-87914-9 - Generalized Linear Models for Insurance Data

Piet de Jong and Gillian Z. Heller

Frontmatter

[More information](#)

## **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 book 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**

Mark Davis, Imperial College London

John Hylands, Standard Life

John McCutcheon, Heriot-Watt University

Ragnar Norberg, London School of Economics

H. Panjer, Waterloo University

Andrew Wilson, Watson Wyatt

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.

Cambridge University Press

978-0-521-87914-9 - Generalized Linear Models for Insurance Data

Piet de Jong and Gillian Z. Heller

Frontmatter

[More information](#)

# 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  
978-0-521-87914-9 - Generalized Linear Models for Insurance Data  
Piet de Jong and Gillian Z. Heller  
Frontmatter  
[More information](#)

---

C A M B R I D G E U N I V E R S I T Y P R E S S  
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](http://www.cambridge.org)

© P. de Jong and G. Z. Heller 2008

This publication is in copyright. Subject to statutory exception  
and to the provisions of relevant collective licensing agreements,  
no reproduction of any part may take place without  
the written permission of Cambridge University Press.

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

Cambridge University Press has no responsibility for the persistence or  
accuracy of URLs for external or third-party internet websites referred to  
in this publication, and does not guarantee that any content on such  
websites is, or will remain, accurate or appropriate.

## Contents

---

<i>Preface</i>	<i>page</i>	<i>ix</i>
<b>1 Insurance data</b>		<b>1</b>
1.1 Introduction		2
1.2 Types of variables		3
1.3 Data transformations		4
1.4 Data exploration		6
1.5 Grouping and runoff triangles		10
1.6 Assessing distributions		12
1.7 Data issues and biases		13
1.8 Data sets used		14
1.9 Outline of rest of book		19
<b>2 Response distributions</b>		<b>20</b>
2.1 Discrete and continuous random variables		20
2.2 Bernoulli		21
2.3 Binomial		22
2.4 Poisson		23
2.5 Negative binomial		24
2.6 Normal		26
2.7 Chi-square and gamma		27
2.8 Inverse Gaussian		29
2.9 Overdispersion		30
Exercises		33
<b>3 Exponential family responses and estimation</b>		<b>35</b>
3.1 Exponential family		35
3.2 The variance function		36
3.3 Proof of the mean and variance expressions		37
3.4 Standard distributions in the exponential family form		37
3.5 Fitting probability functions to data		39
Exercises		41

Cambridge University Press

978-0-521-87914-9 - Generalized Linear Models for Insurance Data

Piet de Jong and Gillian Z. Heller

Frontmatter

[More information](#)

vi

*Contents*

<b>4</b>	<b>Linear modeling</b>	42
4.1	History and terminology of linear modeling	42
4.2	What does “linear” in linear model mean?	43
4.3	Simple linear modeling	43
4.4	Multiple linear modeling	44
4.5	The classical linear model	46
4.6	Least squares properties under the classical linear model	47
4.7	Weighted least squares	47
4.8	Grouped and ungrouped data	48
4.9	Transformations to normality and linearity	49
4.10	Categorical explanatory variables	51
4.11	Polynomial regression	53
4.12	Banding continuous explanatory variables	54
4.13	Interaction	55
4.14	Collinearity	55
4.15	Hypothesis testing	56
4.16	Checks using the residuals	58
4.17	Checking explanatory variable specifications	60
4.18	Outliers	61
4.19	Model selection	62
<b>5</b>	<b>Generalized linear models</b>	64
5.1	The generalized linear model	64
5.2	Steps in generalized linear modeling	65
5.3	Links and canonical links	66
5.4	Offsets	66
5.5	Maximum likelihood estimation	67
5.6	Confidence intervals and prediction	70
5.7	Assessing fits and the deviance	71
5.8	Testing the significance of explanatory variables	74
5.9	Residuals	77
5.10	Further diagnostic tools	79
5.11	Model selection	80
	Exercises	80
<b>6</b>	<b>Models for count data</b>	81
6.1	Poisson regression	81
6.2	Poisson overdispersion and negative binomial regression	89
6.3	Quasi-likelihood	94
6.4	Counts and frequencies	96
	Exercises	96
<b>7</b>	<b>Categorical responses</b>	97
7.1	Binary responses	97
7.2	Logistic regression	98

Cambridge University Press

978-0-521-87914-9 - Generalized Linear Models for Insurance Data

Piet de Jong and Gillian Z. Heller

Frontmatter

[More information](#)

<i>Contents</i>		vii
7.3	Application of logistic regression to vehicle insurance	99
7.4	Correcting for exposure	102
7.5	Grouped binary data	105
7.6	Goodness of fit for logistic regression	107
7.7	Categorical responses with more than two categories	110
7.8	Ordinal responses	111
7.9	Nominal responses	116
	Exercises	119
<b>8</b>	<b>Continuous responses</b>	120
8.1	Gamma regression	120
8.2	Inverse Gaussian regression	125
8.3	Tweedie regression	127
	Exercises	128
<b>9</b>	<b>Correlated data</b>	129
9.1	Random effects	131
9.2	Specification of within-cluster correlation	136
9.3	Generalized estimating equations	137
	Exercise	140
<b>10</b>	<b>Extensions to the generalized linear model</b>	141
10.1	Generalized additive models	141
10.2	Double generalized linear models	143
10.3	Generalized additive models for location, scale and shape	143
10.4	Zero-adjusted inverse Gaussian regression	145
10.5	A mean and dispersion model for total claim size	148
	Exercises	149
<b>Appendix 1</b>	<b>Computer code and output</b>	150
A1.1	Poisson regression	150
A1.2	Negative binomial regression	156
A1.3	Quasi-likelihood regression	159
A1.4	Logistic regression	160
A1.5	Ordinal regression	169
A1.6	Nominal regression	175
A1.7	Gamma regression	178
A1.8	Inverse Gaussian regression	181
A1.9	Logistic regression GLMM	183
A1.10	Logistic regression GEE	185
A1.11	Logistic regression GAM	187
A1.12	GAMLSS	189
A1.13	Zero-adjusted inverse Gaussian regression	190
	<i>Bibliography</i>	192
	<i>Index</i>	195

Cambridge University Press

978-0-521-87914-9 - Generalized Linear Models for Insurance Data

Piet de Jong and Gillian Z. Heller

Frontmatter

[More information](#)

---

## 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.

Cambridge University Press

978-0-521-87914-9 - Generalized Linear Models for Insurance Data

Piet de Jong and Gillian Z. Heller

Frontmatter

[More information](#)

x

*Preface*

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  $y_i = \beta_0 + \beta_1x_{i1} + \dots + \beta_px_{ip}$ . Generally such laboring is avoided. Usually  $x$  denotes the vector  $(1, x_1, \dots, x_p)'$  and  $\beta$  denotes  $(\beta_0, \dots, \beta_p)'$ . 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