

Predictive Modeling Applications in Actuarial Science

Volume I: Predictive Modeling Techniques

Predictive modeling involves the use of data to forecast future events. It relies on capturing relationships between explanatory variables and the predicted variables from past occurrences and exploiting these relationships to predict future outcomes. Forecasting future financial events is a core actuarial skill – actuaries routinely apply predictive modeling techniques in insurance and other risk management applications.

This book is for actuaries and other financial analysts who are developing their expertise in statistics and wish to become familiar with concrete examples of predictive modeling. The book also addresses the needs of more seasoned practicing analysts who would like an overview of advanced statistical topics that are particularly relevant in actuarial practice.

Predictive Modeling Applications in Actuarial Science emphasizes life-long learning by developing tools in an insurance context, providing the relevant actuarial applications, and introducing advanced statistical techniques that can be used by analysts to gain a competitive advantage in situations with complex data.

Edward W. Frees is the Hickman-Larson Professor of Actuarial Science at the Wisconsin School of Business, University of Wisconsin-Madison.

Richard A. Derrig is the president of Opal Consulting LLC and a visiting professor of Risk, Insurance, and Healthcare Management at Fox School of Business, Temple University.

Glenn Meyers has recently retired as vice president and chief actuary at ISO Innovative Analytics.

INTERNATIONAL SERIES ON ACTUARIAL SCIENCE

Editorial Board

Christopher Daykin (Independent Consultant and Actuary)
Angus Macdonald (Heriot-Watt University)

The *International Series on Actuarial Science*, published by Cambridge University Press in conjunction with the Institute and Faculty of Actuaries, contains 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 is 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.

A complete list of books in the series can be found at www.cambridge.org/statistics. Recent titles include the following:

Computation and Modelling in Insurance and Finance
Erik Bølviken

Solutions Manual for Actuarial Mathematics for Life Contingent Risks (2nd Edition)
David C.M. Dickson, Mary R. Hardy, & Howard R. Waters

Actuarial Mathematics for Life Contingent Risks (2nd Edition)
David C.M. Dickson, Mary R. Hardy, & Howard R. Waters

Risk Modelling in General Insurance
Roger J. Gray & Susan M. Pitts

Financial Enterprise Risk Management
Paul Sweeting

Regression Modeling with Actuarial and Financial Applications
Edward W. Frees

Nonlife Actuarial Models
Yiu-Kuen Tse

Generalized Linear Models for Insurance Data
Piet De Jong & Gillian Z. Heller

PREDICTIVE MODELING APPLICATIONS IN ACTUARIAL SCIENCE

Volume I: Predictive Modeling Techniques

Edited by

EDWARD W. FREES

University of Wisconsin, Madison

RICHARD A. DERRIG

Opal Consulting LLC, Providence, Rhode Island

GLENN MEYERS

ISO Innovative Analytics, Jersey City, New Jersey



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE

UNIVERSITY PRESS

32 Avenue of the Americas, New York, NY 10013-2473, USA

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of
education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781107029873

© Cambridge University Press 2014

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 2014

Printed in the United States of America

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

Library of Congress Cataloging in Publication Data

Predictive modeling applications in actuarial science / [edited by] Edward W. Frees, University
of Wisconsin, Madison, Richard A. Derrig, Opal Consulting LLC, Glenn Meyers,

ISO Innovative Analytics, Jersey City, New Jersey.

volumes cm. – (International series on actuarial science)

Includes bibliographical references and index.

Contents: volume 1. Predictive modeling techniques

ISBN 978-1-107-02987-3 (v. 1: hardback)

1. Actuarial science. 2. Insurance – Mathematical models. 3. Forecasting – Mathematical models.
I. Frees, Edward W. II. Derrig, Richard A. III. Meyers, Glenn.

HG8781.P74 2014

368'.01–dc23 2013049070

ISBN 978-1-107-02987-3 Hardback

Additional resources for this publication at <http://research.bus.wisc.edu/PredModelActuaries>

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>Contributor List</i>	<i>page</i> xiii
<i>Acknowledgments</i>	xix
1 Predictive Modeling in Actuarial Science	1
Edward W. Frees, Richard A. Derrig, and Glenn Meyers	
1.1 Introduction	1
1.2 Predictive Modeling and Insurance Company Operations	3
1.3 A Short History of Predictive Modeling in Actuarial Science	5
1.4 Goals of the Series	8
References	9
I Predictive Modeling Foundations	
2 Overview of Linear Models	13
Marjorie Rosenberg and James Guszcza	
2.1 Introduction	13
2.2 Linear Model Theory with Examples	15
2.3 Case Study	45
2.4 Conclusion	59
2.5 Exercises	60
References	63
3 Regression with Categorical Dependent Variables	65
Montserrat Guillén	
3.1 Coding Categorical Variables	65
3.2 Modeling a Binary Response	66
3.3 Logistic Regression Model	67
3.4 Probit and Other Binary Regression Models	78

3.5	Models for Ordinal Categorical Dependent Variables	79
3.6	Models for Nominal Categorical Dependent Variables	81
3.7	Further Reading	85
	References	86
4	Regression with Count-Dependent Variables	87
	Jean-Philippe Boucher	
4.1	Introduction	87
4.2	Poisson Distribution	87
4.3	Poisson Regression	89
4.4	Heterogeneity in the Distribution	92
4.5	Zero-Inflated Distribution	102
4.6	Conclusion	105
4.7	Further Reading	105
	References	105
5	Generalized Linear Models	107
	Curtis Gary Dean	
5.1	Introduction to Generalized Linear Models	107
5.2	Exponential Family of Distributions	110
5.3	Link Functions	115
5.4	Maximum Likelihood Estimation	118
5.5	Generalized Linear Model Review	121
5.6	Applications	122
5.7	Comparing Models	129
5.8	Conclusion	133
5.9	Appendix A. Binomial and Gamma Distributions in Exponential Family Form	133
5.10	Appendix B. Calculating Mean and Variance from Exponential Family Form	135
	References	136
6	Frequency and Severity Models	138
	Edward W. Frees	
6.1	How Frequency Augments Severity Information	138
6.2	Sampling and the Generalized Linear Model	140
6.3	Frequency-Severity Models	148
6.4	Application: Massachusetts Automobile Claims	152
6.5	Further Reading	160

6.6	Appendix A. Sample Average Distribution in Linear Exponential Families	161
6.7	Appendix B. Over-Sampling Claims	162
	References	164

II Predictive Modeling Methods

7	Longitudinal and Panel Data Models Edward W. Frees	167
7.1	Introduction	167
7.2	Linear Models	172
7.3	Nonlinear Models	176
7.4	Additional Considerations	180
7.5	Further Reading	181
	References	181
8	Linear Mixed Models Katrien Antonio and Yanwei Zhang	182
8.1	Mixed Models in Actuarial Science	182
8.2	Linear Mixed Models	192
8.3	Examples	201
8.4	Further Reading and Illustrations	213
	References	215
9	Credibility and Regression Modeling Vytaras Brazauskas, Harald Dornheim, and Ponnalar Ratnam	217
9.1	Introduction	217
9.2	Credibility and the LMM Framework	220
9.3	Numerical Examples	224
9.4	Theory versus Practice	227
9.5	Further Reading	232
9.6	Appendix	233
	References	234
10	Fat-Tailed Regression Models Peng Shi	236
10.1	Introduction	236
10.2	Transformation	238
10.3	GLM	241

10.4	Regression with Generalized Distributions	243
10.5	Median Regression	250
10.6	Appendix A. Tail Measure	255
10.7	Appendix B. Information Matrix for <i>GB2</i> Regression	256
	References	258
11	Spatial Modeling	260
	Eike Brechmann and Claudia Czado	
11.1	Introduction	260
11.2	Exploratory Analysis of Spatial Data	262
11.3	Spatial Autoregression	265
11.4	Average Claim Size Modeling	269
11.5	Hierarchical Model for Total Loss	273
11.6	Discussion and Conclusion	278
	References	278
12	Unsupervised Learning	280
	Louise Francis	
12.1	Introduction	280
12.2	Datasets	283
12.3	Factor and Principal Components Analysis	285
12.4	Cluster Analysis	294
12.5	Exercises	309
	References	310
III Bayesian and Mixed Modeling		
13	Bayesian Computational Methods	315
	Brian Hartman	
13.1	Why Bayesian?	315
13.2	Personal Automobile Claims Modeling	316
13.3	Basics of Bayesian Statistics	316
13.4	Computational Methods	319
13.5	Prior Distributions	326
13.6	Conclusion	330
13.7	Further Reading	330
	References	331
14	Bayesian Regression Models	334
	Luis E. Nieto-Barajas and Enrique de Alba	
14.1	Introduction	334

14.2 The Bayesian Paradigm	335
14.3 Generalized Linear Models	338
14.4 Mixed and Hierarchical Models	354
14.5 Nonparametric Regression	358
14.6 Appendix. Formal Definition of a Polya Tree	362
References	365
15 Generalized Additive Models and Nonparametric Regression	367
Patrick L. Brockett, Shuo-Li Chuang, and Utai Pitaktong	
15.1 Motivation for Generalized Additive Models and Nonparametric Regression	367
15.2 Additive Models for Nonparametric Regression	370
15.3 The Generalized Additive Model	388
15.4 Conclusion	395
References	396
16 Nonlinear Mixed Models	398
Katrien Antonio and Yanwei Zhang	
16.1 Introduction	398
16.2 Model Families for Multilevel Non-Gaussian Data	399
16.3 Generalized Linear Mixed Models	399
16.4 Nonlinear Mixed Models	411
16.5 Bayesian Approach to (L,GL,NL)MMs	411
16.6 Example: Poisson Regression for Workers' Compensation Insurance Frequencies	412
References	423
IV Longitudinal Modeling	
17 Time Series Analysis	427
Piet de Jong	
17.1 Exploring Time Series Data	427
17.2 Modeling Foundations	431
17.3 Autoregressive, Moving Average (ARMA) Models	434
17.4 Additional Time Series Models	443
17.5 Further Reading	448
References	448
18 Claims Triangles/Loss Reserves	449
Greg Taylor	
18.1 Introduction to Loss Reserving	449
18.2 The Chain Ladder	453

18.3 Models of Aggregate Claims Triangles	455
18.4 Models of Individual Claims	467
References	479
19 Survival Models	481
Jim Robinson	
19.1 Survival Distribution Notation	481
19.2 Survival Data Censoring and Truncation	482
19.3 National Nursing Home Survey	482
19.4 Nonparametric Estimation of the Survival Function	486
19.5 Proportional Hazards Model	497
19.6 Parametric Survival Modeling	507
19.7 Further Reading	511
19.8 Exercises	511
19.9 Appendix. National Nursing Home Survey Data	513
References	514
20 Transition Modeling	515
Bruce Jones and Weijia Wu	
20.1 Multistate Models and Their Actuarial Applications	516
20.2 Describing a Multistate Model	518
20.3 Estimating the Transition Intensity Functions	522
20.4 Estimating the Transition Intensities with Outcomes Observed at Distinct Time Points	534
References	536
<i>Index</i>	539

Contributor List

Katrien Antonio is an assistant professor in actuarial science at the University of Amsterdam and KU Leuven. She holds a PhD from KU Leuven. Her research includes claims reserving, ratemaking, and stochastic mortality models.

Jean-Philippe Boucher is an associate professor of Actuarial Sciences in the Département de mathématiques at the Université du Québec à Montréal (UQAM). He worked for several years in insurance companies and at a consulting firm, primarily on ratemaking and claims reserving. His research topics are risk classification, count data, and credibility theory, as well as stochastic claim reserving in general insurance.

Vytautas Brazauskas, PhD, ASA, is a professor in the Department of Mathematical Sciences at the University of Wisconsin-Milwaukee. He is an associate of the Society of Actuaries; an academic correspondent of the Casualty Actuarial Society; and a member of the American Statistical Association and of the American Risk and Insurance Association and has served as president of the Milwaukee Chapter of the American Statistical Association. His areas of expertise are actuarial science and statistics.

Eike Brechmann completed his PhD in mathematics at Technische Universität München on the topic of dependence modeling with copulas. He is currently working at Allianz in Munich.

Patrick L. Brockett holds the Gus S. Wortham Chair in Risk Management and Insurance at the University of Texas at Austin with a joint appointment in Finance, Mathematics, and Information Risk and Operations Management. He is director of the Risk Management and Insurance Program and former director of the Actuarial Science Program and of the Center for Cybernetic Studies at the University of Texas. He is a Fellow of the Institute of Mathematical Statistics, the American Statistical Association, the American Association for the Advancement of Science, and the Institute of Risk Management and is an elected member of the International Statistics Institute.

Shuo-Li Chuang received her PhD from the Department of Information, Risk, and Operations Managements at the Red McCombs School of Business, The University of Texas at Austin. Her research area is mortality modeling and application of statistical modeling. She was a doctoral student in the school of statistics, University of Minnesota at Twin Cities, and was involved in various academic projects applying statistical modeling in economics, finance, education, and insurance.

Claudia Czado is an associate professor of Applied Mathematical Statistics at the Zentrum Mathematik of the Technische Universität in Munich. Her interests are in regression modeling with space and time effects, as well as general dependence modeling using vine copulas. She has published about 100 research articles on statistical models, methods, and computation, which have been applied to topics in insurance, finance, medicine, and biology.

Enrique de Alba is professor emeritus at Instituto Tecnológico Autónomo de México (ITAM), where he was dean of the Division of Actuarial Science, Statistics and Mathematics for more than 20 years. He is also an adjunct professor at the University of Waterloo in Canada. He is currently co-editor of the *North American Actuarial Journal* and vice president of the National Statistical Institute of Mexico (INEGI).

Curtis Gary Dean is the Lincoln Financial Distinguished Professor of Actuarial Science at Ball State University. He is a Fellow of the Casualty Actuarial Society and member of the American Academy of Actuaries. He worked for many years at American States Insurance and later at SAFECO and Travelers.

Piet de Jong is Professor of Actuarial Studies at Macquarie University in Sydney, Australia. He has held a variety of academic positions throughout the world and has been involved in many quantitative consulting projects. His research interests focus on quantitative tools and techniques, particularly in relation to issues in actuarial science, finance, and economics, including time series analysis and generalized linear modeling.

Richard A. Derrig is president of OPAL Consulting LLC, established in February 2004 to provide research and regulatory support to the property-casualty insurance industry. Before forming OPAL, Dr. Derrig held various positions at the Automobile Insurers Bureau (AIB) of Massachusetts and at the Insurance Fraud Bureau (IFB) of Massachusetts over a 27-year period, retiring in January 2004 as senior vice president at AIB and vice president of research at IFB.

Harald Dornheim, ASA, CERA, is an actuarial and risk management consultant at KPMG in Zurich. He completed his PhD in mathematical statistics at the University of Wisconsin-Milwaukee on the topic of robust efficient fitting of mixed linear models

for prediction and risk pricing in insurance. He is a Fellow of the German Actuarial Association and of the Swiss Actuarial Association.

Louise Francis, FCAS, MAAA, is the consulting principal and founder of Francis Analytics and Actuarial Data Mining, Inc., where she leads reserving, pricing, predictive modeling, simulation, and related actuarial projects and engagements. She is a former Vice President of Research for the Casualty Actuarial Society and has been involved in a number of CAS initiatives, including estimating reserve variability, improving data quality, and reviewing papers for its journal *Variance*. She presents frequently on data-mining-related topics and is a five-time winner of the CAS's Data Management and Information call paper program.

Edward W. (Jed) Frees is the Hickman-Larson Professor of Actuarial Science at the University of Wisconsin-Madison. He is a Fellow of both the Society of Actuaries and the American Statistical Association. He has published extensively (a four-time winner of the Halmstad Prize for best paper published in the actuarial literature) and has written three books; his most recent is *Regression Modeling with Actuarial and Financial Applications* (Cambridge University Press, 2009).

Montserrat Guillén is a full professor of econometrics at the University of Barcelona and director of the research group Riskcenter. She is member of the Reial Acadèmia de Doctors, and in 2011 she served as president of the European Group of Risk and Insurance Economists. Her books include *Quantitative Models for Operational Risk* (Chapman and Hall, 2012).

James Guszcza is the U.S. Predictive Analytics Lead for Deloitte Consulting's Actuarial, Risk, and Advanced Analytics practice. He is a Fellow of the Casualty Actuarial Society and a past faculty member in the Department of Actuarial Science, Risk Management, and Insurance at the University of Wisconsin-Madison.

Brian Hartman is an assistant professor of Actuarial Science at the University of Connecticut and an associate of the Society of Actuaries. He has worked as an actuary for Mercer (retirement consulting) and The Hartford (small commercial auto insurance). His current research interests include Bayesian methods, regime-switching models, predictive modeling, and risk management.

Bruce Jones is a professor in the Department of Statistical and Actuarial Sciences at the University of Western Ontario, where he has been a faculty member since 1996. His research addresses a variety of questions related to modeling for life and health insurance. He is a Fellow of the Society of Actuaries and of the Canadian Institute of Actuaries.

Glenn Meyers, FCAS, MAAA, CERA, PhD, recently retired after a 37-year actuarial career that spanned both industry and academic employment. For his last 23 years of working he was employed by ISO as a research actuary. He has received numerous awards for his publications from the Casualty Actuarial Society, including being the first recipient of the Michaelbacher Significant Achievement Award, which “recognizes a person or persons who have significantly and fundamentally advanced casualty actuarial science.”

Luis E. Nieto-Barajas is a full-time professor at Instituto Tecnológico Autónomo de México. He has published numerous research articles on Bayesian and Bayesian nonparametric statistics, with applications ranging from survival analysis to claims reserving. He has served as consultant for the federal presidential elections in Mexico.

Utai Pitaktong is the Senior Modeling Analyst for Actuarial Department at USAA. His former research positions were in the Actuarial Department of State Farm Insurance and in the Predictive Modeling Research Unit of FICO. One of his publications that he co-authored won the 2004 Robert Mehr Award from the American Risk and Insurance Association. He received his PhD in Management Science from the McCombs School of Business, The University of Texas at Austin, and his M.S. in Applied Mathematical Science from Rice University.

Ponmalar Ratnam is a PhD candidate in the Department of Mathematical Sciences at the University of Wisconsin-Milwaukee working on the topic of robust statistical methods for claim frequency models. She has more than 10 years of actuarial modeling experience in the insurance industry and is currently employed as a data scientist at Systech in Los Angeles.

Jim Robinson, PhD, FSA, is the Director of the Center for Health Systems Research & Analysis at the University of Wisconsin-Madison.

Marjorie Rosenberg is a professor of Actuarial Science, Risk Management, and Insurance in the School of Business at the University of Wisconsin-Madison with a joint appointment in Biostatistics and Medical Informatics in the School of Medicine and Public Health. Her research interests are in the application of statistical methods to health care and applying her actuarial expertise to cost and policy issues in health care.

Peng Shi is an assistant professor of Actuarial Science, Risk Management, and Insurance at the University of Wisconsin-Madison. He is an associate of the Society of Actuaries. His research interests are predictive modeling, multivariate regression and dependence models, longitudinal data, and asymmetric information in insurance.

Greg Taylor holds an honorary professorial position in Risk and Actuarial Studies at the University of New South Wales. He previously spent 44 years in commercial actuarial practice and 8 years as an actuarial academic. Taylor has published two books on loss reserving and numerous articles in mathematics, statistics, and actuarial science. He is an Officer of the Order of Australia and holds a Gold Medal from the Australian Actuaries Institute and a Silver Medal from the United Kingdom Institute and Faculty of Actuaries.

Weijia Wu

Yanwei (Wayne) Zhang is a PhD student of marketing at the University of Southern California. He is a Fellow of the Casualty Actuarial Society. He has published in statistical and actuarial journals, and his current research is on social media and consumer decision making.

Acknowledgments

Many, many people contributed to this project. Our editorial and author team represents academia and industry, as well as seven different countries, reflecting the broad interest in predictive modeling. The editors would like to especially commend Xiaoli Jin, who served as project manager for this book while in the doctoral program at the University of Wisconsin-Madison. Her efforts helped improve this book immensely. We were also fortunate to receive extensive comments from many chapter reviewers. We thank them for their contributions.

Funding for this project was provided by the Casualty Actuarial Society and the Canadian Institute of Actuaries.

Reviewer Acknowledgment

Daniel Alai, University of New South Wales

John Baldan, ISO

Lee Bowron, Kerper and Bowron LLC

Doug Bujakowski, University of Wisconsin-Madison

Alan Chalk, AIG

Wai-Sum Chan, Chinese University of Hong Kong

Arthur Charpentier, University of Rennes

Dave Clark, Munich Reinsurance America, Inc.

Steven Craighead, Pacific Life

Marc-André Desrosiers, Intact Insurance Company

Mario DiCaro, Ultimate Risk Solutions

Robert Erhardt, Wake Forest University

Paul Ferrara, Homesite Group Incorporated

Luyang Fu, Cincinnati Financial Corporation

Wu-Chyuan Gau, Florida Blue

Mordechai Goldburd, ISO

Gillian Heller, Macquarie University
Xiaoli Jin, University of Wisconsin-Madison
Marian Keane, Marian Keane Actuarial Consulting Ltd
Anand Khare, ISO
Mary Louie, ISO
Charles Ng, SCOR
Iqbal Owadally, Cass Business School, City University London
Jeffrey S. Pai, University of Manitoba
Pietro Paradi, Willis Global Solutions (Consulting Group)
Georgios Pitselis, University of Piraeus
Jacques Rioux, SAS Institute
Marjorie Rosenberg, University of Wisconsin-Madison
Frank Schmid, AIG
David Scollnik, University of Calgary
Nariankadu Shyamalkumar, University of Iowa
Jaap Spreeuw, Cass Business School, City University London
Carsten Steinebach, Interthinx
Dan Tevet, ISO
Ranee Thiagarajah, Illinois State University
Benjamin Walker, Aon Benfield
Chun-Shan Wong, The Chinese University of Hong Kong
Mark Wood, Legal & General Group
Jia-Hsing Yeh, The Chinese University of Hong Kong

1

Predictive Modeling in Actuarial Science

Edward W. Frees, Richard A. Derrig, and Glenn Meyers

Chapter Preview. Predictive modeling involves the use of data to forecast future events. It relies on capturing relationships between explanatory variables and the predicted variables from past occurrences and exploiting them to predict future outcomes. The goal of this two-volume set is to build on the training of actuaries by developing the fundamentals of predictive modeling and providing corresponding applications in actuarial science, risk management, and insurance. This introduction sets the stage for these volumes by describing the conditions that led to the need for predictive modeling in the insurance industry. It then traces the evolution of predictive modeling that led to the current statistical methodologies that prevail in actuarial science today.

1.1 Introduction

A classic definition of an actuary is “one who determines the current financial impact of future contingent events.”¹ Actuaries are typically employed by insurance companies whose job is to spread the cost of risk of these future contingent events.

The day-to-day work of an actuary has evolved over time. Initially, the work involved tabulating outcomes for “like” events and calculating the average outcome. For example, an actuary might be called on to estimate the cost of providing a death benefit to each member of a group of 45-year-old men. As a second example, an actuary might be called on to estimate the cost of damages that arise from an automobile accident for a 45-year-old driver living in Chicago. This works well as long as there are large enough “groups” to make reliable estimates of the average outcomes.

Insurance is a business where companies bid for contracts that provide future benefits in return for money (i.e., premiums) now. The viability of an insurance company depends on its ability to accurately estimate the cost of the future benefits it promises to provide. At first glance, one might think that it is necessary to obtain data

¹ Attributed to Frederick W. Kilbourne.

from sufficiently large groups of “like” individuals. But when one begins the effort of obtaining a sufficient volume of “like” data, one is almost immediately tempted to consider using “similar” data. For example, in life insurance one might want to use data from a select group of men with ages between 40 and 50 to estimate the cost of death benefits promised to a 45-year-old man. Or better yet, one may want to use the data from all the men in that age group to estimate the cost of death benefits for all the men in that group. In the automobile insurance example, one may want to use the combined experience of all adult men living in Chicago and Evanston (a suburb north of Chicago) to estimate the cost of damages for each man living in either city arising from an automobile accident.

Making use of “similar” data as opposed to “like” data raises a number of issues. For example, one expects the future lifetime of a 25-year-old male to be longer than that of a 45-year-old male, and an estimate of future lifetime should take this difference into account. In the case of automobile insurance, there is no a priori reason to expect the damage from accidents to a person living in Chicago to be larger (or smaller) than the damage to a person living in Evanston. However, the driving environment and the need to drive are quite different in the two cities and it would not be prudent to make an estimate assuming that the expected damage is the same in each city.

The process of estimating insurance costs is now called “predictive modeling.”² In a very real sense, actuaries had been doing “predictive modeling” long before the term became popular. It is interesting to see how the need for predictive modeling has evolved in the United States.

In 1869, the U.S. Supreme Court ruled in *Paul v. Virginia* that “issuing a policy of insurance is not a transaction of commerce.” This case had the effect of granting antitrust immunity to the business of insurance. As a result, insurance rates were controlled by cartels whose insurance rates were subject to regulation by the individual states. To support this regulation, insurers were required to report detailed policy and claim data to the regulators according to standards set by an approved statistical plan.

The Supreme Court changed the regulatory environment in 1944. In *United States v. Southeast Underwriters Association* it ruled that federal antitrust law did apply under the authority of the Commerce Clause in the U.S. Constitution. But by this time, the states had a long-established tradition of regulating insurance companies. So in response, the U.S. Congress passed the McCarran-Ferguson Act in 1945 that grants insurance companies exemption from federal antitrust laws so long as they are regulated by the states. However, the federal antitrust laws did apply in cases of boycott, coercion, and intimidation.

The effect of the McCarran-Ferguson Act was to eliminate the cartels and free the insurers to file competitive rates. However, state regulators still required the insurers

² The name “predictive modeling” is also used in many enterprises other than insurance.

to compile and report detailed policy and claim data according to approved statistical plans. Industry compilations of these data were available, and insurance companies were able to use the same systems to organize their own data.

Under the cartels, there was no economic incentive for a refined risk classification plan, so there were very few risk classifications. Over the next few decades, insurance companies competed by using these data to identify the more profitable classes of insurance. As time passed, “predictive modeling” led to more refined class plans.

As insurers began refining their class plans, computers were entering the insurance company workplace. In the 60 s and 70 s, mainframe computers would generate thick reports from which actuaries would copy numbers onto a worksheet, and using at first mechanical and then electronic calculators, they would calculate insurance rates. By the late 70 s some actuaries were given access to mainframe computers and the use of statistical software packages such as SAS. By the late 80 s many actuaries had personal computers with spreadsheet software on their desks.

As computers were introduced into the actuarial work environment, a variety of data sources also became available. These data included credit reports, econometric time series, geographic information systems, and census data. Combining these data with the detailed statistical plan data enabled many insurers to continue refining their class plans. The refining process continues to this day.

1.2 Predictive Modeling and Insurance Company Operations

Although actuarial predictive modeling originated in ratemaking, its use has now spread to loss reserving and the more general area of product management. Specifically, actuarial predictive modeling is used in the following areas:

- Initial Underwriting. As described in the previous section, predictive modeling has its actuarial roots in ratemaking, where analysts seek to determine the right price for the right risk and avoid adverse selection.
- Renewal Underwriting. Predictive modeling is also used at the policy renewal stage where the goal is to retain profitable customers.
- Claims Management. Predictive modeling has long been used by actuaries for (1) managing claim costs, including identifying the appropriate support for claims-handling expenses and detecting and preventing claims fraud, and for (2) understanding excess layers for reinsurance and retention.
- Reserving. More recently predictive modeling tools have been used to provide management with an appropriate estimate of future obligations and to quantify the uncertainty of the estimates.

As the environment became favorable for predictive modeling, some insurers seized the opportunity it presented and began to increase market share by refining their risk

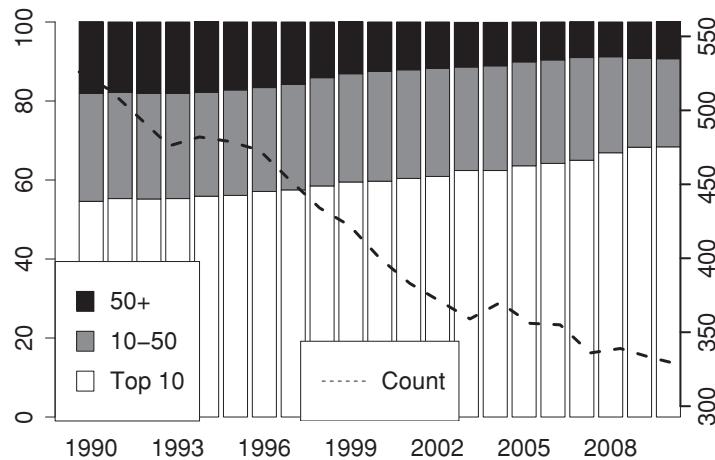


Fig. 1.1. Growth in market share by large insurers. The bar chart shows the percentage, on the left-hand vertical axis, of personal automobile premiums written by the top 10 insurers (in terms of premium volume), the next 40, and other insurers. The right-hand vertical axis shows the decreasing number of insurer groups over time, as indicated by the dashed line. *Source: ISO.*

classification systems and “skimming the cream” underwriting strategies. Figure 1.1 shows how the top American insurers have increased their market share over time. It was this growth in market share that fueled the growth in predictive modeling in insurance.

Actuaries learn and develop modeling tools to solve “actuarial” problems. With these tools, actuaries are well equipped to make contributions to broader company areas and initiatives. This broader scope, sometimes known as “business analytics,” can include the following areas:

- Sales and Marketing – these departments have long used analytics to predict customer behavior and needs, anticipate customer reactions to promotions, and to reduce acquisition costs (direct mail, discount programs).
- Compensation Analysis – predictive modeling tools can be used to incentivize and reward appropriate employee/agent behavior.
- Productivity Analysis – more general than the analysis of compensation, analytic tools can be used to understand production by employees and other units of business, as well as to seek to optimize that production.
- Financial Forecasting – analytic tools have been traditionally used for predicting financial results of firms.

Predictive modeling in the insurance industry is not an exercise that a small group of actuaries can do by themselves. It requires an insurer to make significant investments in their information technology, marketing, underwriting, and actuarial functions. Figure 1.2 gives a schematic representation of the investments an insurer has to make.

	INVESTMENT	REFINED GRANULARITY AND NEW DATA	RETURN
Next Big Thing	Search for new data, method, and new lift 7	Refined Granularity and New Data	Disruptive innovation
	Trusted advisor engaged for ongoing improvement 6	Ensemble Models and Higher Order Interactions	Market leader in data and skilled resources
Predictive Modeling	Multidimensional customer view, including profitability 5	Customization of Third-Party Data	Knowledge of drivers in value creation
	Search for data scientists + mining from federated data warehouses + 360° segmentation data, methods, and new lift 4	Inclusion of Enterprise and Vendor Data	Achieve customer expectations of 'know me'
	Evolution from do-it-yourself to a supply chain of analytical value 3	Customized Third-Party Scores	Efficient deployment of talent and capital
Statistical Analysis	Analytical skills + business acumen + compliance 2	Linear Model-Based Ratemaking	Competitive advantages: smart, fast, and informed
	Add credit score to risk-rating capability 1	Inclusion of Third-Party Scores	Prevents inadvertent omission of the obvious

Fig. 1.2. Investing in predictive modeling. Source: ISO.

1.3 A Short History of Predictive Modeling in Actuarial Science

We would like to nominate the paper “Two Studies in Automobile Insurance Ratemaking” by Bailey and Simon (1960) as the initiator of the modern era of predictive modeling in actuarial science.³ In this paper, they addressed the problem of classification ratemaking. They gave the example of automobile insurance that has five use classes cross-classified with four merit rating classes. At that time, rates for use classes and merit rating classes were estimated independently of each other. They started the discussion by noting that the then current methodology “does not reflect the relative credibility of each subgroup and does not produce the best fit to the actual data. Moreover, it produces differences between the actual data and the fitted values which are far too large to be caused by chance.” They then proposed a set of criteria that an “acceptable set of relativities” should meet:

- (1) “It should reproduce the experience for each class and merit rating class and also the overall experience; i.e., be *balanced* for each class and in total.”
- (2) “It should reflect the relative *credibility* of the various groups involved.”
- (3) “It should provide a minimum amount of *departure* for the maximum number of people.”

³ Mr. Bailey is the son of Arthur L. Bailey, who introduced Bayes’ theorem to the actuarial theory of credibility.

- (4) “It should produce a rate for each subgroup of risk which is close enough to the experience so that the differences could reasonably be caused by *chance*.”

Bailey and LeRoy then considered a number of models to estimate the relative loss ratios, r_{ij} , for use class i and merit rating class j . The two most common models they considered were the additive model $r_{ij} = \alpha_i + \beta_j$ and the multiplicative model $r_{ij} = \alpha_i \cdot \beta_j$. They then described some iterative numerical methods to estimate the coefficients $\{\alpha_i\}$ and $\{\beta_j\}$. The method that most insurers eventually used was described a short time later by Bailey (1963) in his paper titled “Insurance Rates with Minimum Bias.” Here is his solution for the additive model:

- (1) Calculate the initial estimates,

$$\beta_j = \sum_i n_{ij} \cdot r_{ij}$$

for each j . For the additive model, Bailey used the weights, n_{ij} , to represent the exposure measured in car years.

- (2) Given the “balanced” criterion (#1 above) for each i , we have that

$$\sum_j n_{ij} \cdot (r_{ij} - \alpha_i - \beta_j) = 0.$$

One can solve the above equation for an estimate of each α_i ,

$$\alpha_i = \frac{\sum_j n_{ij} (r_{ij} - \beta_j)}{\sum_j n_{ij}}.$$

- (3) Given the estimates $\{\alpha_i\}$, similarly calculate updated estimates of β_j :

$$\beta_j = \frac{\sum_i n_{ij} (r_{ij} - \alpha_i)}{\sum_i n_{ij}}.$$

- (4) Repeat Steps 2 and 3 using the updated estimates of $\{\alpha_i\}$ and $\{\beta_j\}$ until the estimates converge.

This iterative method is generally referred to as the Bailey minimum bias additive model. There is a similarly derived Bailey minimum bias multiplicative model. These models can easily be generalized to more than two dimensions.

As mentioned in the description of the iterative method, the “balance” criterion is met by the design of the method. The method meets the “credibility” criterion because its choice of weights are equal to the exposure. To satisfy the “minimum departure” or “goodness of fit” (in current language) criterion, Bailey and Simon recommended testing multiple models to see which one fit the best. They proposed using the chi-square statistic to test if the differences between actual and fitted were “due to chance.” Although their technology and language were different from now, they were definitely thinking like current actuarial modelers.

The Bailey minimum bias models were easy to program in FORTRAN and BASIC, the predominant computer programming languages of the sixties and seventies, and they were quickly adopted by many insurers. We did a quick poll at a recent Casualty Actuarial Society seminar and found that some insurers are still using these models.

Meanwhile, statisticians in other fields were tackling similar problems and developing general statistical packages. SPSS released its first statistical package in 1968. Minitab was first released in 1972. SAS released the first version of its statistical package in 1976. The early versions of these packages included functions that could fit multivariate models by the method of least squares.

Let's apply the least squares method to the previous example. Define

$$B = \sum_{i,j} n_{ij} \cdot (r_{ij} - \alpha_i - \beta_j)^2.$$

To find the values of the coefficients $\{\alpha_i\}$ and $\{\beta_j\}$ that minimize B we set

$$\frac{\partial B}{\partial \alpha_i} = 2 \sum_j n_{ij} \cdot (r_{ij} - \alpha_i - \beta_j) = 0 \text{ for all } i$$

and

$$\frac{\partial B}{\partial \beta_j} = 2 \sum_i n_{ij} \cdot (r_{ij} - \alpha_i - \beta_j) = 0 \text{ for all } j.$$

These equations are equivalent to the “balance” criterion of Bailey and Simon. The Bailey minimum bias additive model and the least square method differ in how they solve for the coefficients $\{\alpha_i\}$ and $\{\beta_j\}$. Bailey uses an iterative numerical solution, whereas the statistical packages use an analytic solution.

Recognizing that the method of least squares is equivalent to maximum likelihood for a model with errors having a normal distribution, statisticians developed a generalization of the least squares method that fit multivariate models by maximum likelihood for a general class of distributions. The initial paper by Nelder and Wedderburn (1972) was titled “Generalized Linear Models.” The algorithms in this paper were quickly implemented in 1974 with a statistical package called GLIM, which stands for “Generalized Linear Interactive Modeling.” The second edition of McCullagh and Nelder’s book *Generalized Linear Models* (McCullagh and Nelder 1989) became the authoritative source for the algorithm that we now call a GLM.

Brown (1988) was the first to compare the results of the Bailey minimum bias models with GLMs produced by the GLIM package.⁴ The connection between GLMs and Bailey minimum bias models was further explored by Venter (1990),

⁴ Brown attributes his introduction to GLMs to Ben Zehnwirth and Piet DeJong while on a sabbatical leave in Australia.

Zehnwirth (1994), and then a host of others. The connection between GLMs and Bailey minimum bias models reached its best expression in Mildenhall (1999). For any given GLM model there is a corresponding set of weights $\{w_{ij}\}$, (given by Mildenhall's Equation 7.3) for which

$$\sum_i w_{ij} \cdot (r_{ij} - \mu_{ij}) = 0 \text{ for all } i \text{ and } \sum_j w_{ij} \cdot (r_{ij} - \mu_{ij}) = 0 \text{ for all } j.$$

As an example one can set $\mu_{ij} = \alpha_i + \beta_j$ and there is a GLM model for which the corresponding set of weights $\{w_{i,j}\} = \{n_{ij}\}$ yields the Bailey additive model. If we set $\mu_{ij} = \alpha_i \cdot \beta_j$ there is a GLM model for which the corresponding set of weights $\{w_{ij}\}$ yields the Bailey multiplicative model. Other GLM models yield other minimum bias models.

The strong connection between the Bailey minimum bias models and the GLM models has helped make the actuarial community comfortable with GLMs. The actuarial community is rapidly discovering the practical advantages of using statistical packages such as SAS and R. These advantages include the following:

- the ability to include continuous variables (such as a credit score) in the rating plan
- the ability to include interactions among the variables in the rating plan
- the diagnostic tools that allow one to evaluate and compare various models

As a result, the leading edge of actuarial scientists are now using statistical packages for their predictive modeling tasks.

As described in Section 1.2, these tasks go beyond the classification ratemaking applications discussed earlier. Loss reserving is another important task typically performed by actuaries. An early reference to loss reserving that reflects the statistical way of thinking is found in Taylor (1986). Since then there have been several papers describing statistical models in loss reserving. We feel that the papers by Mack (1994), Barnett and Zehnwirth (2000), and England and Verrall (2002) represent the different threads on this topic. These papers have led to a number of specialized loss reserving statistical packages. Although some packages are proprietary, the R `chainladder` package is freely available.

1.4 Goals of the Series

In January 1983, the North American actuarial education societies (the Society of Actuaries and the Casualty Actuarial Society) announced that a course based on regression and time series would be part of their basic educational requirements. Since that announcement, a generation of actuaries has been trained in these fundamental applied statistical tools. This two-set volume builds on this training by developing

the fundamentals of predictive modeling and providing corresponding applications in actuarial science, risk management, and insurance.

Predictive modeling involves the use of data to forecast future events. It relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting those relationships to predict future outcomes. This two-set volume emphasizes life-long learning by developing predictive modeling in an insurance and risk management context, providing actuarial applications, and introducing more advanced statistical techniques that can be used by actuaries to gain a competitive advantage in situations with complex data.

Volume 1 lays out the foundations of predictive modeling. Beginning with reviews of regression and time series methods, this book provides step-by-step introductions to advanced predictive modeling techniques that are particularly useful in actuarial practice. Readers will gain expertise in several statistical topics, including generalized linear modeling and the analysis of longitudinal, two-part (frequency/severity), and fat-tailed data. Thus, although the audience is primarily professional actuaries, we have in mind a “textbook” approach, and so this volume will also be useful for continuing professional development where analytics play a central role.

Volume 2 will examine applications of predictive models, focusing on property and casualty insurance, primarily through the use of case studies. Case studies provide a learning experience that is closer to real-world actuarial work than can be provided by traditional self-study or lecture/work settings. They can integrate several analytic techniques or alternatively can demonstrate that a technique normally used in one practice area could have value in another area. Readers can learn that there is no unique correct answer. Practicing actuaries can be exposed to a variety of techniques in contexts that demonstrate their value. Academic actuaries and students will see that there are multiple applications for the theoretical material presented in Volume 1.

References

- Bailey, R. A. (1963). Insurance rates with minimum bias. *Proceedings of the Casualty Actuarial Society Casualty Actuarial Society L*, 4.
- Bailey, R. A. and L. J. Simon (1960). Two studies in automobile insurance ratemaking. *Proceedings of the Casualty Actuarial Society XLVII*, 192–217.
- Barnett, G. and B. Zehnwirth (2000). Best estimates for reserves. *Proceedings of the Casualty Actuarial Society LXXXVII*, 245–321.
- Brown, R. A. (1988). Minimum bias with generalized linear models. *Proceedings of the Casualty Actuarial Society LXXV*, 187.
- England, P. and R. Verrall (2002). Stochastic claims reserving in general insurance. *Institute of Actuaries and Faculty of Actuaries 28*.
- MacCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). Chapman and Hall/CRC Press.
- Mack, T. (1994). Measuring the variability of chain ladder reserve estimates. In *Casualty Actuarial Society Forum*.

- Mildenhall, S. J. (1999). A systematic relationship between minimum bias and generalized linear models. *Proceedings of the Casualty Actuarial Society LXXXVI*, 393.
- Nelder, J. A. and R. W. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 370–384.
- Taylor, G. C. (1986). *Claims Reserving in Non-Life Insurance*. Elsevier Science Publishing Company, Inc.
- Venter, G. G. (1990). Minimum bias with generalized linear models [discussion]. *Proceedings of the Casualty Actuarial Society LXXVII*, 337.
- Zehnwirth, B. (1994). Ratemaking: From Bailey and Simon [1960] to generalized linear regression models. In *Casualty Actuarial Society Forum*, 615.

Part I

Predictive Modeling Foundations

2

Overview of Linear Models

Marjorie Rosenberg and James Guszczza

Chapter Preview. Linear modeling, also known as *regression analysis*, is a core tool in statistical practice for data analysis, prediction, and decision support. Applied data analysis requires judgment, domain knowledge, and the ability to analyze data. This chapter provides a summary of the linear model and discusses model assumptions, parameter estimation, variable selection, and model validation around a series of examples. These examples are grounded in data to help relate the theory to practice. All of these practical examples and exercises are completed using the open-source R statistical computing package. Particular attention is paid to the role of exploratory data analysis in the iterative process of criticizing, improving, and validating models in a detailed case study. Linear models provide a foundation for many of the more advanced statistical and machine-learning techniques that are explored in the later chapters of this volume.

2.1 Introduction

Linear models are used to analyze relationships among various pieces of information to arrive at insights or to make predictions. These models are referred to by many terms, including linear regression, regression, multiple regression, and ordinary least squares. In this chapter we adopt the term *linear model*.

Linear models provide a vehicle for quantifying relationships between an *outcome* (also referred to as *dependent* or *target*) variable and one or more *explanatory* (also referred to as *independent* or *predictive*) variables. Tradition perhaps accounts for some of the widespread use of linear models, because they are among the oldest data analysis techniques. As discussed in Stigler (1986), Carl Friedrich Gauss and Pierre Simon Laplace developed the method of least squares for estimating planetary orbits from observational data where the functional form of the relationship was provided by physical theory. A second major conceptual advance occurred when the

late 19th-century scientist Francis Galton identified the phenomenon of *regression to the mean*.

Linear models are also ubiquitous because they work well in practice. The core assumption of linearity in the regression parameters offers a structured framework within which to analyze even highly complex data, and it results in models that are intuitive to professionals who do not have extensive training in statistics. The linearity assumption offers a trade-off between simplicity and flexibility that has proven useful in many domains. Because of its relative simplicity, it is often advantageous to start an analysis with a linear model and gradually relax certain assumptions to employ more advanced techniques as the situation demands.

Freedman (2005) states that linear models have at least three different uses: (i) to summarize data, (ii) to predict the future, and (iii) to predict the results of interventions. Regarding the first use, this chapter discusses the relationship between the concepts of regression and correlation. This relationship helps one understand how linear models summarize the associations among various quantities in a dataset of interest. The second use, “predicting the future,” is our primary focus. In predictive modeling applications, one analyzes data to build a model that will be used to estimate an unknown (perhaps future) quantity using one or more variables that are known. Predicting health care utilization or the ultimate loss associated with a particular claim or insurance risk, and estimating the probability that a customer will lapse his or her policy or that a patient will be readmitted to a hospital are all common examples of predictive modeling in insurance. More generally, predictive modeling is increasingly being used in a wide variety of business analytics applications both within and outside the insurance domain. Freedman’s third use of linear models, predicting the results of interventions, involves principles of causal analysis that are outside the scope of this chapter.

Much has been written about linear models, and a short chapter cannot do justice to all of the issues involved (Draper and Smith 1998; Freedman 2005; Frees 2010; Gelman and Hill 2008). The focus of this chapter is on fundamental concepts, assumptions, applications, and the practical issues involved in building regression models. In particular, this chapter focuses on choosing appropriate data samples, specifying an appropriate outcome variable, missing data considerations, selecting explanatory variables, variable transformations and interactions, model validation, and interpreting model output. It provides a number of examples and exercises, all of which are completed in the open-source R statistical computing software¹ (R Core Team 2013). Section 2.2 illustrates the theory of linear models around a series of examples, and then Section 2.3 investigates a detailed case study from beginning to end. This chapter

¹ Source code and data for the examples and exercises are available on the book’s website.

provides a basic foundation for statistical modeling that is expanded on throughout the remainder of this volume.

2.2 Linear Model Theory with Examples

In this section we motivate the use of linear models with targeted examples. We begin with a model that includes no explanatory variables and build to one that includes multiple explanatory variables with transformations and interaction terms. The theory of linear models is interwoven around these examples to allow the reader to put it in context.

The examples and exercises introduced in this chapter are based on a sample of data from the Medical Expenditure Panel Survey (MEPS); (Medical Expenditure Panel Survey 2013). The MEPS study began in 1996 and is a large survey of both families and individuals, focusing on health care utilization and expenditures. The data are collected in panels, in which a household is followed for two calendar years, with each calendar year containing data from two overlapping panels. Our data are from calendar year 2009 based on panels 13 and 14.

We chose to focus on data for individuals who have been diagnosed with diabetes. Diabetes is a “group of diseases characterized by high blood glucose levels that result from defects in the body’s ability to produce and/or use insulin.” (American Diabetes Association 2013a). In 2005–2006, 7.7% of the U.S. population was diagnosed with diabetes, another 5.1% was estimated to be undiagnosed with diabetes, and 29.5% was estimated to be prediabetic, a condition with a strong chance of becoming diabetic. The total annual economic cost of diabetes in 2007 was estimated at \$174 billion, with \$116 billion for medical expenditures (direct diabetes care, chronic diabetes-related complications, and excess general medical costs) and \$58 billion for indirect costs of the disease (increased absenteeism, reduced productivity, disease-related unemployment disability, and loss of productive capacity due to early mortality) (Dall et al. 2008).

In this section we focus on predicting the variable body mass index (BMI) of an individual with diabetes. Studies have shown that those who are overweight have a greater chance of developing diabetes (American Diabetes Association 2013a). The quantity BMI is a number calculated from a person’s weight and height that is used as a screening tool to identify possible weight problems for adults.² Guidelines from the Centers for Disease Control and Prevention indicate that a person with a BMI below 18.5 is underweight, between 18.5 and 24.9 is normal, between 25.0 and 29.9 is overweight, and 30.0 and higher is obese (Centers with Disease Control and Prevention 2013).

² The calculation of $BMI = 703 \times \text{weight in pounds} / (\text{height in inches})^2$.

Example 2.1 (Basic Average Body Mass Index). We analyze a sample of data for 1,101 persons. BMI is the outcome variable of interest and is denoted by $(\text{BMI}_i, i = 1, \dots, 1,101)$. Our statistical model (with $n = 1,101$) is

$$\text{BMI}_i = \mu + \epsilon_i \quad i = 1, \dots, n. \quad (2.1)$$

The BMI population average equals μ , an unknown quantity. Each individual error is represented by ϵ_i , where $E[\epsilon_i] = 0$ and $\text{Var}[\epsilon_i] = \sigma_y^2$. The subscript “y” on σ_y^2 refers to the outcome variable, in this case BMI, and is further discussed in Section 2.2.2. Essentially a person’s BMI is predicted by the mean in this model. If no other information is available, the best predictor of BMI, in the sense of minimizing squared errors, is the sample mean $\bar{\text{BMI}} = \frac{1}{n} \sum_{i=1}^n \text{BMI}_i$. The residual, e_i , is an estimate of ϵ_i and is equal to $e_i = \text{BMI}_i - \bar{\text{BMI}}$. In our sample of $n = 1,101$ observations, the sample average BMI is $\bar{\text{BMI}} = 31.97$, and the sample variance is equal to 55.42. Thus we use 31.97 as our prediction of a person’s BMI and 55.42 as our estimate of σ_y^2 .

Example 2.2 (One-Variable Regression for Body Mass Index). We can improve the basic model by incorporating one or more explanatory (or independent) variables to predict BMI. Suppose we construct a linear model that uses AGE to predict BMI. Here the statistical model can be represented as

$$\text{BMI}_i = \beta_0 + \beta_1 \text{AGE}_i + \epsilon_i \quad i = 1, \dots, n. \quad (2.2)$$

Comparing this equation with Equation 2.1, the outcome variable is BMI, ϵ_i is the error term, and we have modified μ to vary by person ($\mu_i = \beta_0 + \beta_1 \text{AGE}_i$). The person-level mean is a linear function of AGE, where β_0 is the intercept and β_1 is the slope. Note in Example 2.1 that the slope is equal to zero with an intercept equal to 31.97. The empirical relationship between BMI and AGE is suggested by the *scatterplot* in Figure 2.1. In this plot, vertical and horizontal lines have been placed at the sample average values of AGE and BMI, respectively.

The estimated linear model relating BMI to AGE contains an intercept term equal to 36.98 and a slope term on AGE equal to -0.085 . The solid line in Figure 2.1 indicates that as AGE increases by 1, then BMI decreases by -0.085 . As a check, the correlation between BMI and AGE, which measures the linear association between these two variables, is -0.15 . Note that the sign of the slope coefficient agrees with that of the correlation statistic. We see in Example 2.5 how these statistics are related.

2.2.1 Types of Explanatory Variables

Explanatory variables are commonly classified as continuous, discrete, or categorical. A *continuous* variable could equal any value in an interval, whereas a *discrete* variable

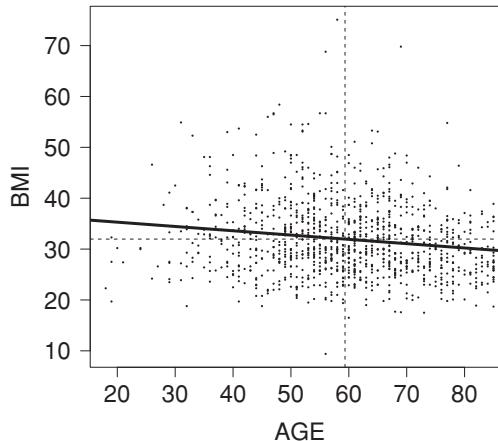


Fig. 2.1. Scatterplot of BMI vs. AGE. Dotted lines represent sample mean of BMI and AGE. Solid line is the least squares line.

takes on a countable, possibly infinite, number of distinct values. Our variable `AGE` in Example 2.2 could be considered either continuous or discrete.

In contrast, a *categorical* variable has a distinct number of levels that can be either ordered or unordered. The estimation of the regression coefficients is not affected by whether a categorical variable is viewed as ordered or unordered. However, the regression coefficients are expected to maintain a certain relationship for ordered categorical variables.

A *binary* (also referred to as *dummy* or *indicator*) variable is a special case of a categorical variable, denoting group membership where the levels are “1” if the observation is a member of the group and “0” otherwise.

Example 2.3 (Linear Model with Sex). Suppose we are interested in the relationship between body mass index (BMI) and sex for individuals who have been diagnosed with diabetes. Using the MEPS data, we regress BMI on the two-level categorical variable `SEX`. The question of interest is whether one’s sex has an effect on the magnitude of BMI. Table 2.1 summarizes the results of three alternate, but equivalent, models using panel 13 of the MEPS data.

The first model, displayed in Table 2.1(a), transforms the categorical variable for sex into a binary indicator variable for males. The estimating equation is $\widehat{BMI} = 32.7923 - 1.9135 \cdot MALE$, with the estimated BMI for males equal to 30.8788 ($= 32.7923 - 1.9135(1)$), whereas the estimated BMI for females is 32.7923 ($= 32.7923 - 1.9135(0)$). In this particular model, the *reference category* is females; that is, the results for males are with reference to the base category of female. We see that the BMI for males is lower than that for females by 1.9135.

Table 2.1. *Linear Model of BMI on SEX*

(a) With reference category = Females	
	Estimate
(Intercept)	32.7923
MALE	-1.9135
(b) With reference category = Males	
	Estimate
(Intercept)	30.8788
FEMALE	1.9135
(c) With no intercept term	
	Estimate
MALE	30.8788
FEMALE	32.7923

If instead we used males as the reference category as in Table 2.1(b), we have an estimated linear model of $\widehat{BMI} = 30.8788 + 1.9135 \cdot FEMALE$. Not surprisingly, the estimated BMI value for males is 30.8788 and for females is 32.7923. Thus changing the reference value does not change the estimated values of the outcome variable. As seen, however, the interpretation of the linear model does change because the comparison is relative to the reference category, males. Here the model indicates that the BMI for females is 1.9135 larger than for males, an equivalent version of that in Table 2.1(a).

Because our model contains one and only one categorical variable, we can remove the intercept term to obtain a different interpretation. In Table 2.1(c), the output of the model is two lines – one for males and one for females – with the values of the estimate of 30.8788 for males and 32.7923 for females, the same as earlier. When there are multiple categorical variables, it is helpful to include an intercept term to facilitate the interpretation of the reference category and for other model summary measures.

Note that including both a male indicator and a female indicator in a model together with an intercept term would result in perfect multicollinearity and is called an *over-specified* model. The male indicator is a linear transformation of the female indicator ($FEMALE = 1 - MALE$), and vice versa. More generally, if a categorical variable contains c categories, we include indicators for at most $(c - 1)$ levels in a model that also contains an intercept term to avoid over-specifying the model.

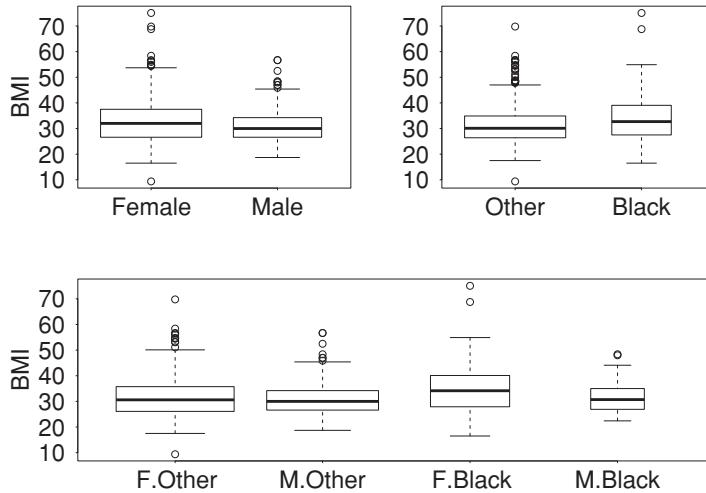


Fig. 2.2. Boxplots of BMI vs. SEX; BMI vs. RACE; and BMI vs. SEX/RACE combination. Solid lines in the boxplots represent the median BMI by category.

Example 2.4 (Multiple Variable Regression for Body Mass Index). We extend Examples 2.2 and 2.3 to include the binary variables MALE and BLACK, along with the continuous variable AGE:

$$BMI_i = \beta_0 + \beta_1 AGE_i + \beta_2 MALE_i + \beta_3 BLACK_i + \epsilon_i \quad i = 1, \dots, n \quad (2.3)$$

The marginal relationships between BMI and male (vs. females) and BMI and BLACK (vs. OTHER RACES) are shown in Figure 2.2.

We see that the median BMI for males is slightly lower than that for females, whereas the median BMI for BLACKS is higher than that for OTHER RACES. When looking at the distribution across the four distinct categories, the FEMALE BLACK median BMI is the highest and the interquartile range is the widest.

The parameter β_1 for AGE is now estimated as -0.083 , similar to the estimate found in Example 2.2. The interpretation of the coefficient is that the estimated BMI for an individual in the population of interest for a given sex and race changes by -0.083 for each additional age.

The estimate for β_2 is -1.60 , indicating that a male of a particular age and race is expected to have a BMI approximately 1.60 lower than a female of the same age and race. Finally, the estimate for β_3 is 2.46 , indicating that being black increases one's BMI for a given age and sex. Thus for a given age, a black female has the highest estimated BMI, a conclusion that concurs with the boxplot.

The estimate of the intercept term β_0 is 36.95 . In general, the intercept of a linear model is the predicted value of the outcome variable given a value of zero for each of the model's explanatory variables. In the current example, the intercept term represents

the BMI for a (diabetic) non-black, female, newborn baby (`BLACK = 0`, `MALE = 0`, `AGE = 0`). Because the minimum age in our data is 18, the interpretation of the intercept term is not useful. Techniques such as *centering the data*, as discussed in Section 2.3.3, adjust the use of the variables and enable a useful interpretation of the intercept term.

Part of the art of statistics involves deciding how to incorporate explanatory variables. `AGE` could be available with integer and fractional components or used as age at last birthday. Using `AGE` in the form as received is certainly one way of including its impact. The assumption then is that there is a linear association between `AGE` and the outcome variable. If the regression coefficient for `AGE` is positive, then an increase in `AGE` by one unit would increase the outcome variable by the amount of the regression coefficient.

However, if the assumed linear relationship is not true over all ages, the outcome variable would not be modeled appropriately. In prediction, using `AGE` as a linear variable would not result in a good fit at certain ages. One possible modification is to include `AGE` as a polynomial, such as quadratic or cubic, to represent the change of the outcome variable with a change in `AGE`. Another modification is to categorize `AGE` into different levels, such as young, adult, and elder. A further approach might be to include `AGE` as a linear term for younger ages and to create a categorical variable for older ages. Such decisions are routinely made in data analyses and exemplify how data analysis is an art requiring sound judgment and domain expertise in addition to the mastery of statistical concepts.

2.2.2 Notation and Assumptions

Using these examples as background, we formally define some key linear model concepts. In general, we have n observations. Our outcome variable is $(y_i, i = 1, \dots, n)$, and for each observation i , there is a set of k explanatory, or independent, variables $\{x_{i1}, x_{i2}, \dots, x_{ik}\}$, each potentially exhibiting some relationship with y_i . The idea is to postulate a model that approximates the relationship between the outcome variable and the explanatory variables.

With linear models, we use an estimator other than the sample mean to better estimate y_i and to reduce the error. Using a set of unknown coefficients $\{\beta_0, \dots, \beta_k\}$ and a random error or disturbance term ϵ_i , the linear model is defined as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i \quad i = 1, \dots, n. \quad (2.4)$$

The error is represented as $\epsilon_i = y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik})$ for each observation i . The assumptions underlying the linear model include the following:

- (i) $\mathbf{y} = (y_i, i = 1, \dots, n)$, a vector of independent outcome random variables.
- (ii) $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$, a vector of given explanatory variables for observation i .

- (iii) $\beta = (\beta_0, \dots, \beta_k)$, a vector of fixed, unknown coefficients, with β_0 representing the intercept term and $(\beta_1, \dots, \beta_k)$ representing k slope parameters.
- (iv) $\epsilon = (\epsilon_i, i = 1, \dots, n)$, a vector of independent random variables, with $E[\epsilon_i] = 0$ and $\text{Var}[\epsilon_i] = \sigma^2$.

From these assumptions, it follows that \mathbf{y} is a vector of independent, continuous random variables with $E[y_i | \mathbf{x}_i, \beta] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}$ and $\text{Var}[y_i | \mathbf{x}_i, \beta] = \sigma^2$. The assumption of constant error variance is called *homoskedasticity*.³

Another way to view the linear model is via the expected value of y_i . This representation is that used with generalized linear models (GLMs), which model the mean of the outcome variable. If the errors are normally distributed, then $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ for a one-variable model. The mean of the normal distribution for a given x_i lies on the regression line. The variance is the spread of the values from the line, conditional on x_i , and is the same σ^2 as defined in the earlier error model.

2.2.3 Estimation

One way to estimate these parameters is by minimizing the sum of squared errors. Starting with Equation 2.4, we solve for ϵ_i and define a function of the β -coefficients to minimize:

$$SS(\beta_0, \dots, \beta_k) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}))^2.$$

To minimize the sum of squared errors with respect to the β -coefficients, we take separate derivatives of $SS(\beta_0, \dots, \beta_k)$ with respect to each β -coefficient and set each equation equal to 0. This results in a system of $(k + 1)$ equations that determine the estimates of the β -coefficients. The vector \mathbf{b} is the solution of this system.⁴

Example 2.5 (One-Variable Linear Model). A one-variable linear model is sometimes referred to as *simple linear regression* or *basic linear regression*. For a set of data, the model is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with two equations to minimize:

$$\frac{\partial SS(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i)) = 0. \quad (2.5)$$

$$\frac{\partial SS(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^n x_i (y_i - (\beta_0 + \beta_1 x_i)) = 0.$$

³ This word is derived from the Greek roots *homo*, meaning “same,” and *skedannoumi*, meaning “to scatter.” Therefore “homoskedasticity” can informally be translated as “uniform scatter.”

⁴ If the outcome data, \mathbf{y} , are normally distributed conditional on the explanatory variables, then \mathbf{b} is also normally distributed. In this case, one can derive that the least squares estimator is equivalent to the maximum likelihood estimator.

Solving for each of the β -coefficients (relabelled as b_0, b_1 for the least squares estimators) yields

$$b_0 = \bar{y} - b_1 \bar{x}. \quad (2.6)$$

$$b_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Defining the standard deviation of x as $s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$, the standard deviation of y as $s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}$, and the sample correlation of (x, y) as $r_{xy} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{(n-1)s_x s_y}$ yields

$$b_1 = r_{xy} \left(\frac{s_y}{s_x} \right). \quad (2.7)$$

This one-variable linear model shows the relationship between regression and correlation. The estimated slope coefficient b_1 is proportional to the sample correlation between x and y .

We define $\hat{y}_i = b_0 + b_1 x_i$, labeling \hat{y}_i as the *fitted value* or *predicted value*. As shown in Section 2.2.6, (b_0, b_1) are unbiased estimators of (β_0, β_1) . This property implies that $E[\hat{y}_i] = \beta_0 + \beta_1 x_i$ or equivalently that the expected value of \hat{y}_i lies on a line determined by x_i and β . The residuals, defined as the set of $(e_i = y_i - \hat{y}_i, i = 1, \dots, n)$, are our estimates of the error terms $(\epsilon_i, i = 1, \dots, n)$. Interestingly, as a consequence of the estimation process, Equation 2.5 implies that the sum, or average, of the residuals is always equal to 0.

Example 2.6 (Output of One-Variable Linear Model). The estimate of the intercept from the regression of BMI on AGE is $b_0 = 36.98$, and the estimate of the slope is $b_1 = -0.085$. The remainder of the output is discussed later in this section.

```
Regression of BMI on AGE

            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 36.98262   1.01031  36.605 < 2e-16 ***
AGE        -0.08455   0.01661  -5.091  4.2e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.361 on 1099 degrees of freedom
Multiple R-squared: 0.02304, Adjusted R-squared: 0.02215 
F-statistic: 25.91 on 1 and 1099 DF,  p-value: 4.196e-07
```

Table 2.2. Mean and Standard Deviation of BMI and AGE

	Mean	Std. Dev.
BMI	31.97	7.44
AGE	59.34	13.36

Using the correlation between BMI and AGE as calculated earlier (-0.1518), along with the mean and standard deviation of BMI and AGE shown in Table 2.2, we can re-create the statistics in the regression output.

With Equation 2.7 we estimate the slope term as $b_1 = -0.1518 \cdot 7.44/13.36 = -0.0845$. Then using Equation 2.6, we find $b_0 = 31.97 - (-0.0845 \cdot 59.34) = 36.98$.

Example 2.7 (Output of Null Linear Model). The mean and standard deviation of BMI shown in Table 2.2 are also outputs of the model shown in Example 2.1, known as the *null model* or an *intercept-only model*.

```
Regression of BMI on No Explanatory Variables

      Estimate Std. Error t value Pr(>|t|)    
(Intercept) 31.9650    0.2244   142.5 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.444 on 1100 degrees of freedom
```

Note that the coefficient estimate for the intercept is same as the mean of BMI and that the residual standard error is the same as the sample standard deviation of BMI. In this case, the standard error of the estimate (0.2244) is the quotient of the residual standard error (7.444) and the square root of the degrees of freedom (1100), or the usual formula to calculate the standard error of the estimate of the mean. Section 2.2.4 discusses how to determine the degrees of freedom, and Section 2.2.6 illustrates the calculation of the standard error of the estimate when adding explanatory variables.

Example 2.8 (Output of Multiple Variable Linear Model). The format of output from the regression of BMI on AGE, MALE, and BLACK is identical to that with Example 2.6, with the addition of two lines for the two additional variables.

```
Regression of BMI on AGE + MALE + BLACK

      Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.95073   1.02348 36.103 < 2e-16 ***
AGE        -0.08301   0.01631 -5.088 4.24e-07 ***
MALE       -1.59767   0.44418 -3.597 0.000336 ***
BLACK       2.45570   0.50358  4.876 1.24e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.23 on 1097 degrees of freedom
Multiple R-squared: 0.05944, Adjusted R-squared: 0.05686
F-statistic: 23.11 on 3 and 1097 DF, p-value: 1.655e-14
```

We present the details of this output for completeness with further discussion of the output later in this section.

2.2.4 Partitioning the Sum of Squares

Linear models provide a method for “explaining the variance” of the outcome variable. To understand this statement, let us analyze two estimators of BMI, namely \bar{y} and \hat{y}_i . The first estimator is what we refer to as a *marginal* estimator, because it depends only on the outcome data. The second estimator is a *conditional* estimator, because \hat{y}_i depends both on the outcome data and also on the explanatory variables in the model being considered.

We decompose the deviation of the estimator, $y_i - \bar{y}$, referred to as *total error*, by defining the unexplained deviation (or error) as $y_i - \hat{y}_i$ and the explained deviation as $\hat{y}_i - \bar{y}$:

$$\underbrace{y_i - \bar{y}}_{\text{total}} = \underbrace{y_i - \hat{y}_i}_{\text{unexplained}} + \underbrace{\hat{y}_i - \bar{y}}_{\text{explained}}. \quad (2.8)$$

Squaring both sides of Equation 2.8, summing over i , and performing a little algebra yields

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{Total SS}} = \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\text{Error SS}} + \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{Regression SS}}. \quad (2.9)$$

In this way, the linear model enables us to decompose the total sum of squares (*Total SS*) into two portions: the error sum of squares (*Error SS*) and the regression

sum of squares (*Regression SS*). Note that the *Total SS* is a property of the outcome variable and is not model dependent. Recall that $\sum_{i=1}^n \frac{(y_i - \bar{y})^2}{n-1}$ is used to calculate the marginal variance of the y_i .

As the *Regression SS* increases, the *Error SS* necessarily decreases by the same amount. Recall that the linear model coefficients are the parameter estimates that minimize $SS(\beta_0, \dots, \beta_k)$. From the above equation, we see that this is tantamount to shifting some of the *Total SS* to *Regression SS*.

A common way to summarize this relationship is through the *coefficient of determination*:

$$R^2 = \frac{\text{Regression SS}}{\text{Total SS}} = 1 - \frac{\text{Error SS}}{\text{Total SS}}. \quad (2.10)$$

R^2 varies between $[0, 1]$ and is a summary statistic that measures the goodness of fit of a model. It can be shown that R^2 is equal to the square of the Pearson correlation coefficient between the outcome variable y and the fitted values \hat{y} . The positive square root of the coefficient of determination is denoted by R and is called the *multiple correlation coefficient*.

An *analysis of variance* or *ANOVA* table is a useful way to summarize some components of a linear model. For a linear model with n observations, k explanatory variables, and an intercept term, the ANOVA table is defined as follows:

Analysis of Variance Table for Linear Model

With k Explanatory Variables + Intercept			
Source	Sum of Squares	Deg. of Freedom	Mean Square
Regression	<i>Regression SS</i>	k	<i>Regression MS</i>
Error	<i>Error SS</i>	$n - (k + 1)$	<i>MSE</i>
Total	<i>Total SS</i>	$n - 1$	s_y^2

The total degrees of freedom is $n - 1$. The sum of the degrees of freedom associated with the *Regression SS* and *Error SS* must equal $n - 1$. The *Regression SS* has k degrees of freedom corresponding to the k explanatory variables. The degrees of freedom associated with the *Error SS* must equal $n - (k + 1)$ by subtraction. This can be seen directly by beginning with the n observations and subtracting the $(k + 1)$ estimated regression parameters, corresponding to the k explanatory variables plus the intercept term.

The mean square column of the table is found by dividing the sum of squares column by the degrees of freedom. Dividing the *Total SS* (the left-side of Equation 2.9) by

Table 2.3. ANOVA Table for Regression of BMI on AGE, MALE, and BLACK

Source	Sum of Squares	Deg. of Freedom	Mean Square
Regression	3,623	3	1,208.67
Error	57,336	1,097	52.27
Total	60,959	1,100	55.42

its degrees of freedom ($n - 1$) results in the sample variance of \mathbf{y} or $s_y^2 = \frac{\text{Total SS}}{n-1}$, which is an unbiased estimator of σ_y^2 . Dividing the *Error SS* by $n - (k + 1)$ results in the *mean squared error (MSE)*, denoted s^2 . The *MSE* is an unbiased estimator of σ^2 , the variance of the linear model error term, and is viewed as a measure of the improvement offered by the model over simply using the sample mean using s_y^2 . The σ^2 parameter is that introduced at the beginning of Section 2.2.2 when defining the $\text{Var}[\epsilon_i]$.

Example 2.9 (Output of Multiple Variable Linear Model). In this example, we connect the output from the regression of BMI on AGE, MALE, and BLACK in Example 2.8 to the ANOVA output shown in Table 2.3.

The degrees of freedom for the *Regression SS* are the number of slope parameters, that is, one for AGE, MALE, and BLACK. The degrees of freedom for *Total SS* are one less than the number of observations. The degrees of freedom for the *Error SS* are the difference between these two quantities.

The square of the *residual standard error* ($= 7.23^2$) is the *MSE*. The *Multiple R-squared* equals $R^2 = 0.05944$. Using R^2 and the *MSE*, we use Equation 2.10 to calculate the *Total SS*. We use that same equation to calculate *Regression SS*. The mean square column in Table 2.3 is the quotient between the sum of squares and the degrees of freedom. Note that the square of the standard deviation of BMI shown in Table 2.2 is the same as that of the mean square total in Table 2.3 (any difference due to rounding).

Of course, computer packages include commands to calculate the ANOVA table, but understanding the quantities in the ANOVA table and how they are displayed in regression output helps remind the analyst of the concept of partitioning the variance.

2.2.5 Variable Selection

Deciding which variables to include in a model and the manner in which they are to be included is one of the most challenging aspects of data analysis. In this section we summarize some of the major criteria used for variable selection and focus on

the basic types of statistics used to choose which variables to include in a linear model. We discuss how to assess the statistical significance of a variable or set of variables. In addition, we investigate whether the explanatory variables are *collinear* with one another, which could affect their statistical significance. We use the output from the three models in Examples 2.6, 2.7, and 2.8 to decide whether the variables are statistically significant.

2.2.5.1 Variable Significance

One way of determining whether to include a particular variable x_j is with a *t*-test. We consider the *null hypothesis* (H_0) that $\beta_j = 0$ against the *alternative hypothesis* (H_1) that $\beta_j \neq 0$.⁵ In a one-variable linear model as in Example 2.6, if $\beta_1 = 0$ then the regression line is a horizontal line with no slope, or as we have called the null model. If $\beta_1 \neq 0$, then there would be some positive or negative relationship between the outcome and explanatory variable. This kind of test is called a *two-tailed test*. We calculate the *t*-statistic as $t = \frac{b_j}{se(b_j)}$, where $se(b_j)$ is the standard error of the regression estimator and is covered formally in Section 2.2.6. The sampling distribution of the *t*-statistic is a *t*-distribution with $n - (k + 1)$ degrees of freedom, assuming that the null hypothesis is true.

We start by selecting a *significance level* α . Formally, $\alpha = Pr [Rejecting H_0 | H_0]$ or the probability of rejecting H_0 given that H_0 is in fact true. The choice of significance level of 0.05, although conventional, is arbitrary. Many other choices of α , such as 1%, 5%, or 10%, are possible and are discipline and contextually dependent.

With a two-sided test, we *reject* H_0 if $|t|$ exceeds the $(1 - \alpha/2)^{th}$ percentile of a *t*-distribution with $n - (k + 1)$ degrees of freedom. Otherwise, we *fail to reject* H_0 . With $\alpha = 0.05$, a rule of thumb is that one rejects H_0 (and concludes that the variable in question is statistically significant) if $|t| > 2$.

As an alternative to examining the *t*-statistic, the *p-value* can be used to decide whether to reject H_0 . Intuitively, a *p-value* is the probability, *under the assumption that the null hypothesis is true*, of observing a *t*-statistic at least as extreme as the *t*-statistic actually observed. More formally, for a two-tailed test, the *p-value* = $2 \cdot Pr [T > |t| | H_0]$, where T is a *t*-distributed random variable with $n - (k + 1)$ degrees of freedom. We reject H_0 if the *p-value* $< \alpha$. See Exercise 2.5 for more intuition for the hypothesis testing approach and its limitations.

Example 2.10 (Significance of Age in a Linear Model). Example 2.6 displays the output of a regression with AGE as the single explanatory variable. The *t*-statistic is equal to -5.091 and the *p-value* is < 0.001 . We therefore reject the null hypothesis

⁵ We could consider additional alternative hypotheses such as $\beta_j > 0$ or other variations.

that the coefficient is equal to zero and say that there is a statistically significant relationship between BMI and AGE.

The F -statistic, located at the bottom of the regression output, tests whether all of the slope parameters are equal to zero against the alternative hypothesis that at least one parameter is not equal to zero.

With one slope coefficient in this model, the F -test and the t -statistic produce identical results. Mathematically, the square of the t -statistic is the F -statistic. In this case, the p -value of the t -statistic for a two-tailed test equals the p -value of the corresponding F -statistic.

Instead of analyzing a single variable, we often wish to compare two models for which the variables of one model are a subset of the variables of the other. Such models are called *nested* models.

Example 2.11 (Nested Linear Model Test of RACE). Suppose we modify our variable BLACK and expand it to RACE, with categories WHITE, BLACK, and OTHER. We might ask whether RACE with its multiple categories is statistically significant to include in the model. We refer to the two models as the *reduced* (without RACE) and *full* (with RACE) models. With three levels for the variable RACE, we essentially add two binary variables to the model as shown in Equation 2.11:

$$\begin{aligned} BMI_i &= \beta_0 + \beta_1 AGE_i + \beta_2 MALE_i + \beta_3 BLACK_i + \beta_4 OTHER_i + \epsilon_i \\ i &= 1, \dots, n. \end{aligned} \quad (2.11)$$

Here we have designated the race category WHITE as the reference category with the inclusion of the binary variables BLACK and OTHER. The ANOVA decomposition concepts previously introduced can be used to address whether the data support the full model over the reduced model. Table 2.4 displays a portion of the ANOVA table for this full model, as well as the residuals portion of the ANOVA table for the reduced model.

We consider the *null hypothesis* (H_0) that $\{\beta_3, \beta_4\}$ in Equation 2.11 are both equal to zero against the *alternative hypothesis* (H_1) that at least one of $\{\beta_3, \beta_4\}$ is not equal to zero. Thus the reduced model is the one in which only MALE and AGE are retained. The data collected are used to test whether there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis.

In this case, we calculate an F -statistic that considers information from the full and the reduced models and use the information from Table 2.4:

$$F = \frac{\text{Error } SS_{\text{reduced}} - \text{Error } SS_{\text{full}}}{q \cdot s_{\text{full}}^2} = \frac{58578.72 - 56351.73}{2 \cdot 51.42} = 21.66 \quad (2.12)$$

Table 2.4. ANOVA Summaries

BMI ~ AGE + MALE + RACE (full model)					
	Df	Sum Sq	Mean Sq	F value	Pr(> F)
RACE	2	2226.99	1113.49	21.66	< 0.0001
Residuals	1096	56351.73	51.42		
BMI ~ AGE + MALE (reduced model)					
	Df	Sum Sq	Mean Sq	F value	Pr(> F)
Residuals	1098	58578.72	53.35		

The q (or 2 in this case) in the denominator refers to the number of coefficients that we are testing as equal to zero, whereas the 51.42 is the s_{full}^2 , or the residual variance from the full regression. Not surprisingly, the F -statistic that we calculated is identical to that shown in the full model ANOVA output in Table 2.4. The corresponding p -value indicates that at least one of the two coefficients (shown by degrees of freedom) for RACE is significantly different from zero.

While in Example 2.11 we focused on inclusion of one variable RACE, multiple variables can be tested for inclusion in the model with the F -test. In general, if our reduced model has k slope parameters and we are testing whether q additional slope parameters are useful, we calculate an F -statistic that considers information from the full and the reduced models:

$$F = \frac{\text{Error SS}_{\text{reduced}} - \text{Error SS}_{\text{full}}}{q \cdot s_{\text{full}}^2} = \frac{\frac{R_{\text{full}}^2 - R_{\text{reduced}}^2}{q}}{\frac{1 - R_{\text{full}}^2}{n - (k + q + 1)}}. \quad (2.13)$$

The first equation is the formula that we used in Example 2.11. The second equation is a mathematically equivalent version. The advantage of the second version is that it can calculate the F -statistic from the usual regression output of the full and reduced models. The disadvantage is that we need to assess the significance of that value.

The sampling distribution of the F -statistic under the null hypothesis is a F -distribution with q degrees of freedom in the numerator and $n - (k + q + 1)$ degrees of freedom in the denominator, because the statistic is a ratio of two χ^2 random variables divided by their degrees of freedom. We compare the F -statistic to the $(1 - \alpha)^{th}$ percentile of the F -distribution, because the support of the F -distribution is the positive real line. Alternatively, the p -value of the F -statistic is calculated and compared against α to determine whether or not to reject H_0 .

Example 2.12 (Null Linear Model). We can view Example 2.10 in this notation of full and reduced models. If the coefficient on AGE is equal to zero, our reduced model contains only an intercept term or what we have called the null model. Because the intercept is the only parameter in the model, the estimator \hat{y} is \bar{y} for every observation. Therefore, the $Error\ SS_{reduced} = Total\ SS$ and $R^2_{reduced} = 0$. Using the R^2 from Example 2.6 and setting $q = 1$, $k = 0$, and $n = 1,101$, the second part of Equation 2.13 becomes

$$F = \frac{R^2_{full}}{\frac{1-R^2_{full}}{n-2}} = \frac{0.02304}{(1 - 0.02304)/1,099} = 25.92. \quad (2.14)$$

which is an F -statistic and ties to the results at the bottom of the regression output in Example 2.6.

2.2.5.2 Collinearity

We can assess the importance of including one or more candidate explanatory variables in the model with the t and F tests. It is also good practice to check for *collinearity* among the explanatory variables themselves. Collinearity occurs when one explanatory variable is a linear combination of other explanatory variables. High correlation among the explanatory variables results in high variances of the corresponding regression parameter estimators. This relationship is discussed further at the end of Section 2.2.6 in Example 2.14.

The variance inflation factor (*VIF*) is a useful statistic in measuring the degree of collinearity present among the explanatory variables.⁶ In this case, we are seeking to determine whether one explanatory variable can be explained by a linear combination of the other explanatory variables.

The *VIF* for the j th explanatory variable is found by regressing this variable against the other explanatory variables in the model. The $VIF_j = (1 - R_j^2)^{-1}$, where R_j^2 is the coefficient of determination for this linear model.

There is not a consensus for a rule of thumb to use to assess whether collinearity is an issue. Brien (2007) Values for *VIF* above 5 or even above 10 are used to suggest high collinearity of the explanatory variables.

Example 2.13 (VIF for AGE, MALE, and BLACK Model). Using the model with explanatory variables as in Example 2.9, we calculate the *VIF* as:

⁶ The term *aliasing* is sometimes used synonymously with collinearity. In the general regression literature, *VIF* is used to identify collinearity among the explanatory variables.

Table 2.5. VIF for AGE, MALE, and BLACK

	AGE	MALE	BLACK
VIF	1.0002	1.0200	1.0202

Because the *VIF* for each of the explanatory variables is below the threshold of either 5 or 10, then we conclude that these variables are not collinear.

2.2.6 Expected Value and Variance of Regression Estimators

We represent the linear model with multiple explanatory variables as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Here \mathbf{y} represents the outcome variable and is a column vector of length n . The explanatory variables are represented by \mathbf{X} , which is sometimes referred to as the *design matrix* and is of dimension $n \times (k + 1)$. The design matrix contains a column of ones corresponding to an intercept term and followed by k explanatory variables. The column vector $\boldsymbol{\beta}$ of length $k + 1$ represents the unknown intercept and k slope parameters. The vector $\boldsymbol{\epsilon}$ contains the errors and is represented by a column vector of length n .

We use results from linear algebra to solve the system of $(k + 1)$ equations. Starting with the sum of squared errors, $\boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, we take derivatives with respect to each of the $(\beta_j, j = 0, 1, \dots, k)$. The solution of the system of linear equations for $\boldsymbol{\beta}$ is $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, and the fitted values for \mathbf{y} are equal to $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$.

The least squares estimators, $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, are functions of the known explanatory variables and the outcome data before they are realized. The least squares estimates result once the data are realized and are substituted in the equation for the estimators. The estimators are random variables, whereas the estimates are specific values determined by the data being modeled.

The expected value of the regression parameter estimators is

$$E[\mathbf{b}] = E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right] = E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\overbrace{(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})}^{\mathbf{y}}\right] = \boldsymbol{\beta}$$

and shows that \mathbf{b} is an unbiased estimator for $\boldsymbol{\beta}$. Using the matrix property that $\text{Var}[AX] = A\text{Var}[X]A'$, where A is a constant matrix and X is a random matrix, the variance of the regression estimators is

$$\text{Var}[\mathbf{b}] = \text{Var}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right] = \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\text{Var}[\mathbf{y}|\mathbf{X}]\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

where $\text{Var}[\mathbf{y}|\mathbf{X}] = \sigma^2$ as discussed in Section 2.2.2. The Gauss-Markov theorem states that, of all unbiased estimators of $\boldsymbol{\beta}$, the least squares estimator has minimum

variance (Stigler 1986, p. 148). Note that the resulting matrix for the $\text{Var}[\mathbf{b}]$ is a $(k + 1) \times (k + 1)$ symmetric matrix.

Technically, we need to estimate σ^2 and call it s^2 , or the residual variance. Thus what we can actually calculate is the estimated variance or $\widehat{\text{Var}}[\mathbf{b}] = s^2 (\mathbf{X}'\mathbf{X})^{-1}$. The standard error of b_j , $se(b_j)$, is the square root of the $(j + 1)^{\text{st}}$ diagonal value in the matrix.

Example 2.14 (Linear Model Output). Referring again to Examples 2.6 to 2.8, the square of the residual standard error is our estimate of s^2 , either 54.18 or 52.27, respectively. Note that, with the addition of explanatory variables, we have reduced the residual variance s^2 from that of the marginal variance s_y^2 (i.e., the null model equal to 55.41).

The column labeled *Std. Error* in the regression output represents the square root of the diagonal elements of $\widehat{\text{Var}}[\mathbf{b}]$. It can be shown that the standard error of b_j is proportional to the square root of the *VIF*:

$$se(b_j) = \frac{s \cdot \sqrt{VIF_j}}{s_{x_j} \cdot \sqrt{n - 1}},$$

where s is the residual standard error and s_{x_j} is the standard deviation of the j th explanatory variable. A higher *VIF* corresponds to a higher $se(b_j)$, which leads to a lower t -statistic and a higher p -value. Thus collinearity, which manifests itself in a higher *VIF*, can result in distorted measures of the statistical importance of particular variables.

Note too that the standard error is a function of the number of observations. If that number is large, then the standard error for all regression estimates will be small, resulting in larger t -statistics and smaller p -values. Thus the larger sample size will lead to more statistically significant slope coefficients. However, greater statistical significance does not imply that the variable is more substantively meaningful to the analysis.

2.2.7 Model Adequacy: R_a^2 and AIC

Section 2.2.5 discussed choosing variables individually or in groups for variable selection in situations where one model is nested in another. Suppose instead that one wishes to compare two non-nested models. One might be tempted to use R^2 to decide between the two models. However, this statistic is ill suited for this purpose because R^2 will never decrease when adding any explanatory variable, even if the new variable has no relationship to the outcome variable.

Table 2.6. Comparison of R^2 , R_a^2 , and AIC

Model	R^2	R_a^2	AIC
MALE	0.01623	0.01534	7531.882
AGE	0.02304	0.02215	7524.238
AGE + MALE	0.03905	0.03730	7508.045
AGE + MALE + BLACK	0.05944	0.05686	7486.433
AGE + MALE + RACE	0.07558	0.07221	7469.372
AGE + MALE + RACE + UNINSURED	0.07629	0.07207	7470.524

Note: All output not shown in chapter.

Because of this limitation of R^2 , a coefficient of determination adjusted for degrees of freedom (R_a^2 , also known as *adjusted-R²*) is commonly used:

$$R_a^2 = 1 - \frac{\frac{\text{Error SS}}{n-(k+1)}}{\frac{\text{Total SS}}{n-1}} = 1 - \frac{s^2}{s_y^2},$$

where the number of regression slope parameters is equal to k . Because s_y^2 is a property of the outcome variable and not of the model, maximizing R_a^2 is equivalent to minimizing the residual variance s^2 .

Example 2.15 (Comparison of R^2 and R_a^2). A summary of the R^2 and R_a^2 for the models discussed and an additional model that includes a binary variable for UNINSURED, are shown in Table 2.6.

Notice that the variable AGE alone explains more of the total variance than does the addition of MALE. Also see that the use of the RACE variable explains more than including only a binary variable on BLACK. In all examples, the R_a^2 statistic is less than R^2 . However, note the last line for the model, which includes the UNINSURED variable. The p -value to determine whether this coefficient is different from zero equals 0.36 (not shown). For this model, the R_a^2 for this model is less than the model without this variable, but its R^2 is larger. This illustrates that selecting a model based on R^2 is not appropriate, and it is better to choose based on R_a^2 .

Another model selection criterion, motivated by concepts from information theory, was introduced by Hirotugu Akaike in 1974. The *Akaike information criterion (AIC)* is a general model selection criterion that applies to both nonlinear and linear models (Akaike 1974). The general definition of AIC is

$$AIC = -2 \log(L) + 2 \times (\text{number of parameters}), \quad (2.15)$$

where $\log(L)$ is the log-likelihood function of the model. If the errors are normally distributed, then Equation 2.15 reduces to

$$AIC = n \cdot \log(2\pi) + n \cdot \log(s^2) + n + 3 + k.$$

The AIC incorporates a penalty for model complexity as measured by the number of parameters. Therefore selecting the model with the lowest AIC is one way of optimizing a trade-off between goodness of fit and model complexity.

Example 2.16 (Comparison of AIC). Table 2.6 includes a column for the AIC statistic. Note that the statistic decreases down the list until the last model with inclusion of the `UNINSURED` variable. Thus in this case, the AIC and R_a^2 measures reach the same conclusion. This does not necessarily occur in all cases.

2.2.8 Model Validation

There is no unique stopping rule indicating when the modeling process is complete. Economic considerations may suggest a trade-off between the costs of additional modeling efforts against the likely marginal improvements in predictive accuracy and analytical insights. In this section we present different ways of analyzing the model results to assess the adequacy of the model. First we define influential points (high leverage and outliers) and discuss how to identify them. We then discuss various types of graphical plots that may be of value. We end the section with a discussion of two metrics useful in model selection and model validation.

2.2.8.1 Analysis of Influential Points

An influential point is an observation with a disproportionate impact on the overall results. Two types of influential points are high leverage and outliers.

Observations that are *high leverage* points are those that are extreme in value relative to one or more explanatory variables. If we are predicting `BMI` for people with diabetes, then as we show in Example 2.17, persons with high income might disproportionately influence the result of the analysis. Intuitively, one can think of a balance beam to represent the regression line or plane, with an observation far out in the range of an explanatory variable tipping the balance beam one way or another, or as having a large impact on the value of the regression coefficient estimates.

In Section 2.2.6, we introduced $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$ for the fitted value of \mathbf{y} and the estimator for the regression coefficients as $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. If we define $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, then we can rewrite the prediction equation as $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$. From this matrix multiplication we see that each predicted observation is written as a linear combination of the elements of \mathbf{H} :

$$\hat{y}_i = h_{i1} \cdot y_1 + h_{i2} \cdot y_2 + \cdots + h_{in} \cdot y_n \quad (2.16)$$

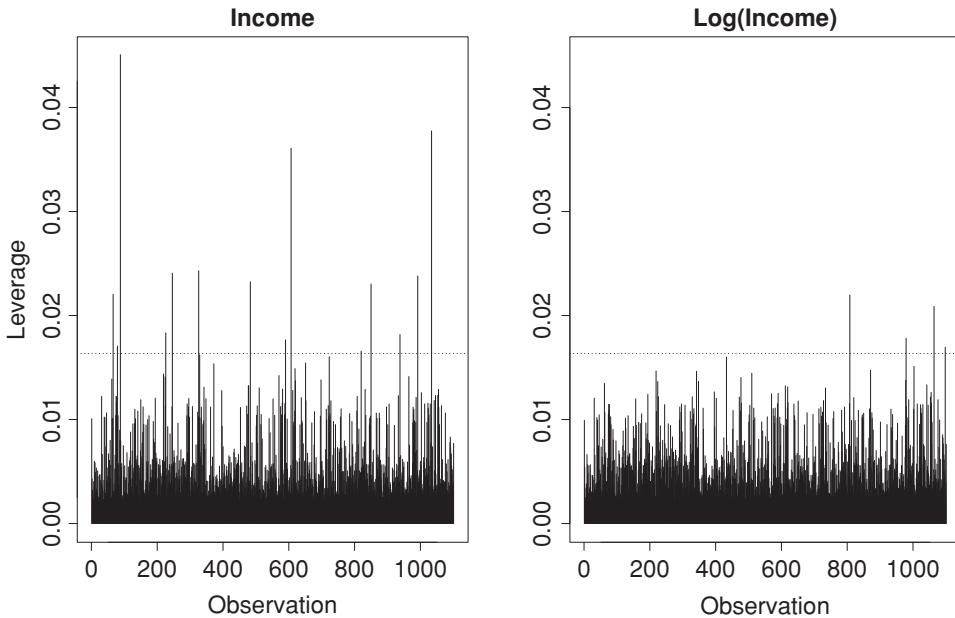


Fig. 2.3. Leverage plot for models including `INCOME` or `Log (INCOME)` in addition to `AGE`, `MALE`, and `RACE`. The dotted line represents three times the average leverage for each model.

The diagonal elements of \mathbf{H} , $\{h_{ii}, i = 1, \dots, n\}$, are called the *leverage* values and are restricted to lie between $\frac{1}{n} \leq h_{ii} \leq 1$. The average leverage is $\frac{k+1}{n}$. There is no standard rule of thumb for a high leverage point, but authors have suggested two or three times the average as cut-off values.

Example 2.17 (High Leverage). Suppose we believe that in addition to `AGE`, `MALE`, and `RACE`, `INCOME` is associated with `BMI`. We include this variable in the model and find that it is significant with a p -value = 0.039 and a negative coefficient. The adjusted- R^2 is 0.075, an improvement on the model without this variable as seen in Table 2.6. Thus conceptually, this relationship makes sense as `INCOME` increases, `BMI` decreases. We create a plot that visually shows the leverage values, $h_{ii}, i = 1, \dots, n$ in Figure 2.3.

For the model including `INCOME` there are 14 observations whose leverage values exceed three times the average. The observed value for `BMI` for these observations will receive greater weight in their predicted value for `BMI` as inferred from Equation 2.16. Investigating these 14 observations reveals that these individuals all have incomes exceeding \$190,000.

One strategy for dealing with high leverage points is to do nothing and disclose the influences of these observations. Another solution is to consider certain variable

transformations to reduce the influence. Here we calculate the log-transform of INCOME. The coefficient on $\log(\text{INCOME})$ is still negative, its p -value is 0.016, and the adjusted- R^2 equals 0.076. This model is an improvement on the model with INCOME, with only four leverage points above three times the average leverage as shown in the right-hand plot of Figure 2.3. Interestingly, in this case, these four observations represent individuals with very low incomes (< \$2,500).

Thus INCOME and its log counterpart are very significant variables, but have some highly influential observations.

Although high-leverage observations are indicators of extremes in the explanatory variables, very high or very low residuals are indicators of outliers. As noted in Example 2.5, the residual is defined as $e_i = y_i - \hat{y}_i$. We can think of a residual as the outcome variable after removing the information contained in the explanatory variables included in the model to help explain the variation of y . A residual must be standardized to determine whether any investigation is required.

The standard error of the residual $se(e_i)$ is equal to $\sigma\sqrt{1 - h_{ii}}$, where h_{ii} is the leverage. The *standardized residual* is therefore estimated as $\frac{e_i}{s\sqrt{1-h_{ii}}}$. A large residual leads to a large outlier. But a large leverage value also increases the standardized residual. The rule of thumb for detecting outliers is similar to that for high leverage points, where 2 or 3 may be used as a threshold.

Example 2.18 (Outliers). Using the same models as in Example 2.17, Figure 2.4 shows graphs of outliers similar to those of the leverage values.

Both graphs indicate a number of observations whose residual exceeds the threshold of ± 2 . In fact, there are 49 such outliers for the model using INCOME and 48 outliers for the model using $\log(\text{INCOME})$, and are a subset of the outliers of the model containing the INCOME variable.

For the model including INCOME, only 3 of the 14 high leverage points are outliers, whereas for the model including $\log(\text{INCOME})$, 2 of the 4 high leverage points are outliers. Figure 2.7 in Example 2.21 shows this pictorially and extends this example.

2.2.8.2 Residual Plots

Plots of residuals, as well as other validation summaries, help ensure that given all of the information available, the model seems reasonable and that further changes are unnecessary. As mentioned previously, residuals measure the leftover variation of the outcome variable from a model containing certain explanatory variables. We create plots of the residuals to show (i) whether the residuals resemble a normal distribution,

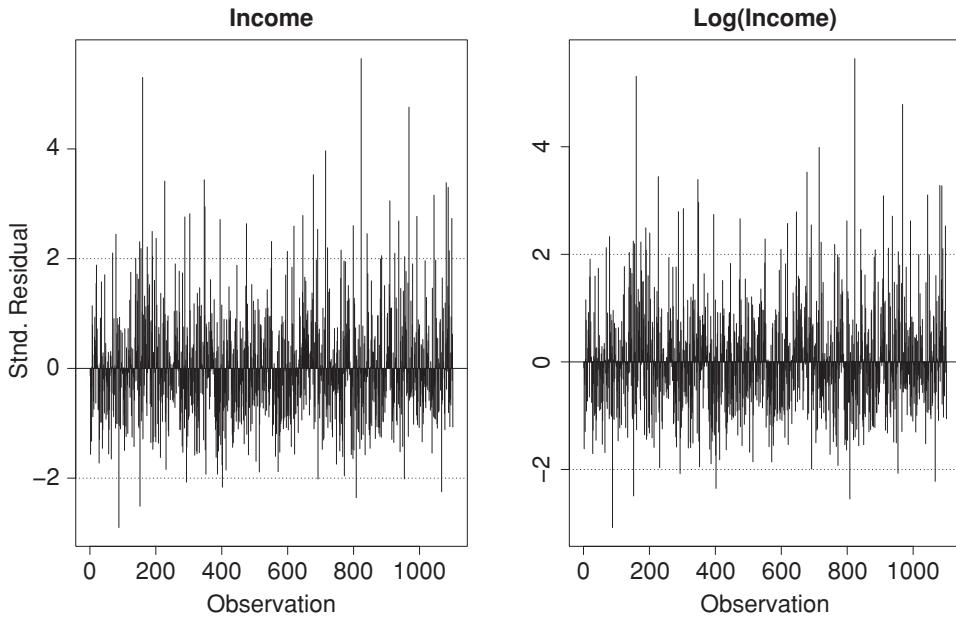


Fig. 2.4. Standardized residual plot for models including INCOME or Log (INCOME) in addition to AGE, MALE, and RACE. The ± 2 levels for outliers for each model are represented as a dotted line.

- (ii) no relationship between the residuals and explanatory variables in the model, (iii) no relationship between residuals and explanatory variables not in the model, and (iv) no relationship between residuals and the fitted values.

When plotted against explanatory variables already in the model, the residuals could display a pattern, such as a quadratic, that would suggest that a squared term of the variable is appropriate to add to the model. A similar plot against the fitted values would show a similar pattern, but would not identify which variable to alter. Plotting the residuals against the fitted values helps check for homoskedasticity of the residuals. If there is a pattern where the spread of the residuals increases or decreases with larger fitted values, then some adjustments must be made or other techniques considered.

If no patterns exist, then plotting the residuals against explanatory variables not in the model would be the next step. If a pattern exists, then additional iterations of model building are required.

Example 2.19 (Residual Plots). Suppose we examine the standardized residuals from the model of BMI on AGE + MALE + RACE + Log (INCOME) with respect

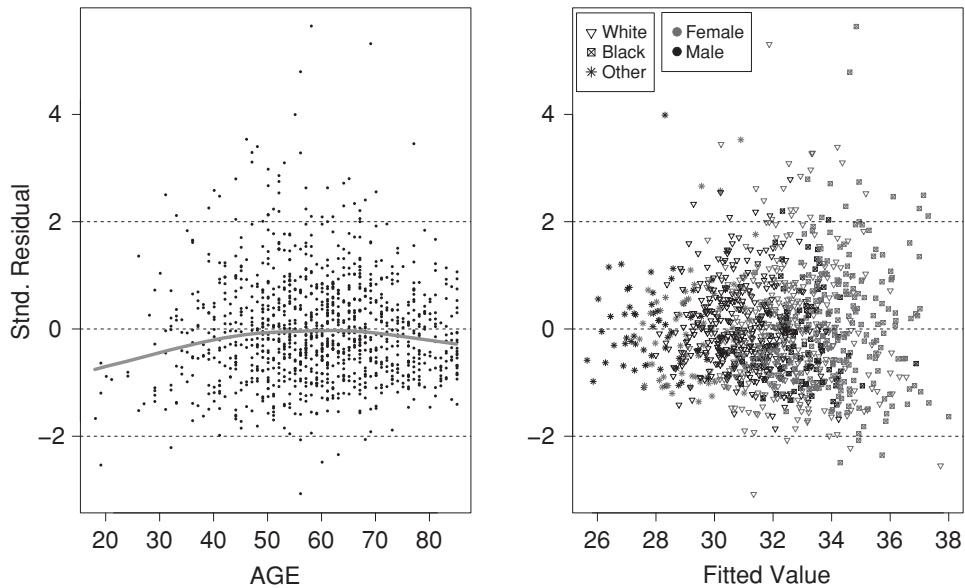


Fig. 2.5. Examination of residuals from `BMI` on `AGE + MALE + RACE + Log(INCOME)`. Left-hand graph shows relationship of standardized residuals and `AGE`. Right-hand graph shows interaction of `MALE` and `RACE` with x-axis being the fitted values.

to `AGE` and another with respect to `RACE` and `SEX`. Example 2.4 examined a similar relationship in Figure 2.2 where we saw that black females had a different median `BMI` than other groups. We had grouped whites and non-blacks into the other category, whereas now we separate them.

Figure 2.5 shows these residual plots, with the left-hand plot showing standardized residuals versus `AGE`, and the right-hand plot showing standardized residuals plotted against the fitted values with the plotting symbols differing by `SEX` and by `RACE`. The `AGE` plot includes a smoothed nonparametric line, known as a *LOESS* curve, that indicates that there exists a modest quadratic-type relationship between the residuals and `AGE`.

The `SEX` and `RACE` plot is more complicated but quite informative. Data points for males are colored black, whereas for females the dots are gray. Whites, blacks and other races are denoted by triangles, boxes, and asterisks, respectively. Interestingly, observations for females are a large proportion of the outliers. In the data, there are approximately equal numbers of males and females for the white and other groups, but twice as many black females as black males. The plot indicates that there might be an interaction effect between `SEX` and `RACE`. Including such a term in the model amounts to the creation of six categories, one for each `SEX/RACE` combination, similar to what was presented in Figure 2.2.

The output from this regression is shown in the box.

```
Regression of BMI on AGE + AGE^2 + MALE + RACE + LOGINCOME +
MALE:RACE

            Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.2034500 3.8338023 8.139 1.08e-15 ***
AGE          0.4262089 0.1092942 3.900 0.000102 ***
AGE^2        -0.0044595 0.0009345 -4.772 2.07e-06 ***
MALE         -0.9482630 0.5424591 -1.748 0.080731 .
RACE black   2.6170878 0.6311280 4.147 3.63e-05 ***
RACE other   -3.1374132 0.9721822 -3.227 0.001287 **
LOGINCOME    -0.7410235 0.2432718 -3.046 0.002374 **
MALE:RACE black -2.5768794 1.0522306 -2.449 0.014483 *
MALE:RACE other -0.1448401 1.4061484 -0.103 0.917978
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.073 on 1092 degrees of freedom
Multiple R-squared: 0.1039, Adjusted R-squared: 0.09737
F-statistic: 15.83 on 8 and 1092 DF, p-value: < 2.2e-16
```

The squared term on AGE is statistically significant as a negative estimate, confirming the quadratic relationship in Figure 2.5. The interaction term for the black males is also significant.

The residuals, or standardized residuals, can be summarized in a histogram, where a symmetric distribution is desired. Alternatively, residuals can be displayed in a quantile-quantile plot (*QQ*-plot) in which the quantiles of the distribution of residuals are plotted against the quantiles of a normal distribution. We look to see whether the points fall on a diagonal line.

Example 2.20 (*QQ*-Plots). Figure 2.6 shows the *QQ*-plot of the null model with no explanatory variables compared to the *QQ*-plot of the model with $AGE + AGE^2 + MALE + RACE + MALE:RACE + \text{Log}(INCOME)$. Conditional on no explanatory variables, BMI is not normally distributed, because the lower and upper tails deviate from the line. In this example, we see that the residuals for the lower quantiles look more normally distributed once covariates have been added. However, it is necessary to explore other variables for inclusion in the model, because the upper quantiles still deviate from the line.

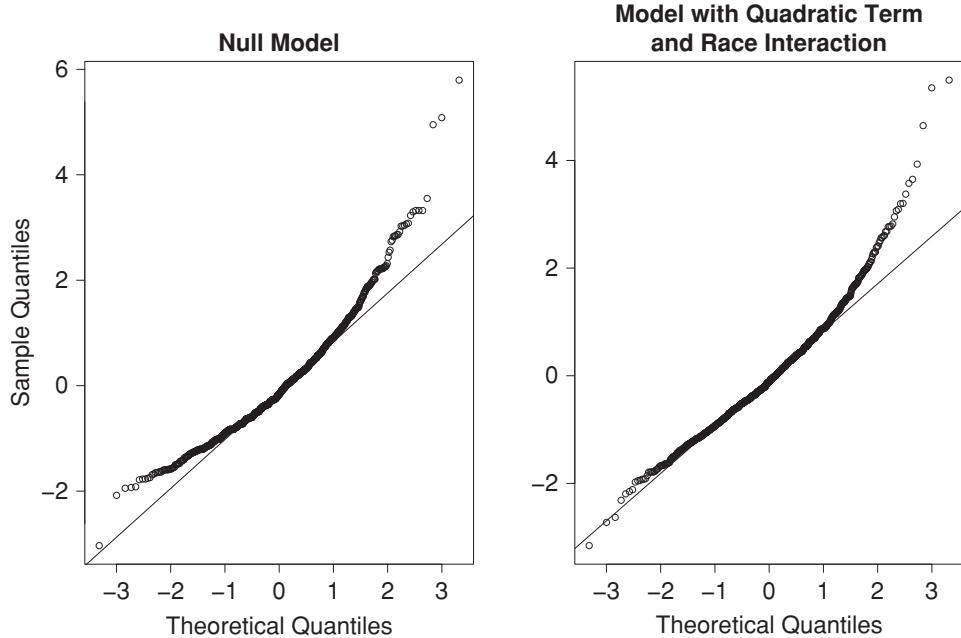


Fig. 2.6. *Q*-*Q*-plots from null model for BMI versus BMI on AGE + AGE^2 + MALE + RACE + MALE:RACE + Log(INCOME).

2.2.8.3 Cook's Distance

Of course an observation can be both an outlier and a high leverage point. A third measure, called *Cook's distance* (D_i), is used to help summarize these effects simultaneously. Define $\hat{y}_{j(i)}$ as the predicted value of y_j from a linear model fit without the i th observation. Cook's distance is defined as

$$D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{(k+1) \cdot s^2} = \left(\frac{e_i}{se(e_i)} \right)^2 \cdot \frac{h_{ii}}{(k+1) \cdot (1-h_{ii})}.$$

In the first equality, D_i summarizes the difference in fitted values from the original model using all n observations versus those in the model after removing the i th observation. The larger the difference, the larger is D_i . The second equality shows a different version of the statistic that separates the outlier component from the leverage component; it is in a form that is easier to compute. Overall the statistic summarizes the influence of the i th observation. Again, there is no standard rule of thumb for assessing Cook's distance, but $D_i > 1$, $D_i > \frac{4}{n-(k+1)}$, or $D_i > \frac{4}{n}$, have appeared in the literature.

Example 2.21 (Cook's Distance). Using the same models as in Examples 2.17 and 2.18, Figure 2.7 shows a combination plot of Cook's distance, outliers, and high

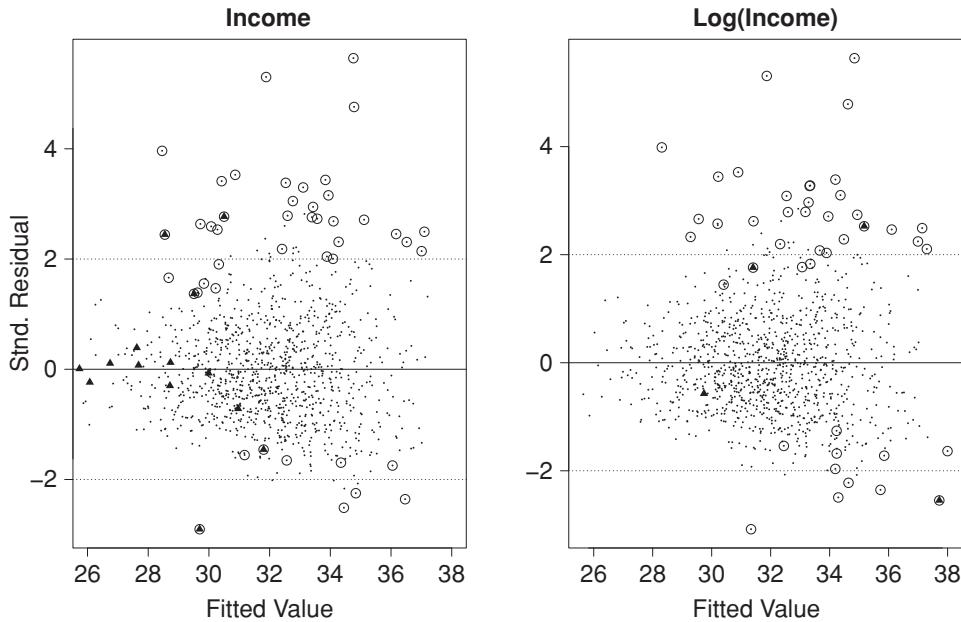


Fig. 2.7. Combination plot of Cook's distance, outliers, and high leverage points for models including `INCOME` or $\text{Log}(\text{INCOME})$ in addition to `AGE`, `MALE`, and `RACE`. The dotted lines represent thresholds for outliers, the triangles represent the high leverage points, the tiny dots represent the other observations, and the open circles represent those observations with high Cook's distance.

leverage points, together with their fitted values. The approach is to plot standardized residuals versus the fitted values. The triangles represent the high leverage observations, the tiny dots are the other observations, and the observations with high Cook's distance have an open circle around them.

The plot is complicated but quite illustrative. First, points that have a high Cook's distance are sometimes outliers, sometimes high leverage, sometimes both outlier and high leverage, and sometimes neither outlier nor high leverage. Dots with circles are those nonhigh leverage points that are above the threshold for a high Cook's distance.

For the left-hand plot using `INCOME`, we see many high leverage observations with residuals close to zero with small fitted values. Examining the data reveals that these observations have the highest values for `INCOME`. Thus these observations do influence the estimation of the regression plane and result in small residual values. Yet other observations with high leverage values do not have this same effect.

In reviewing both plots, the observations with large values of Cook's distance generally seem to be the same observations that are outliers. However, there are observations using `INCOME` or $\text{Log}(\text{INCOME})$ where the threshold for Cook's distance is exceeded but they are not outliers or high leverage points. These observations

bear some investigation because they may be ones whose characteristics give guidance as to the next steps.

As the right-panel indicates, the use of `LOG(INCOME)` helps reduce the number of high leverage points and the influence on the regression parameter estimates. However, many outliers warrant some investigation.

2.2.8.4 SSPE and PRESS

Although the R^2 and AIC statistics are good ways of assessing model adequacy with the data used to estimate the model parameters, the gold standard of model validation is out-of-sample testing. The idea is to first fit the model to a *training* (or referred to as an *in-sample*) dataset. Once the estimates are found, they are applied to evaluate the predictive accuracy of a model on a separate *hold-out* (or referred to as an *out-of-sample*) data set.

The sum of squared prediction errors (*SSPE*) is a statistic that considers the squared residuals of the hold-out sample when applying the parameter estimates from the training sample. Consider a dataset of n observations where n_1 observations are used in the training sample and n_2 observations are used in the hold-out sample. The *SSPE* is

$$SSPE = \sum_{i=n_1+1}^{n_1+n_2} (y_i - \hat{y}_i)^2.$$

As with the *AIC*, the *SSPE* is not meaningful in isolation, but is useful for comparing between models. The lower the *SSPE*, the better the prediction fit.

With small datasets, it may not be feasible to split a sample into the two pieces. Sophisticated techniques called *cross validation* involve an iterative process using one dataset. One such method is called predicted residual sum of squares (*PRESS*). Similar to the notation introduced with Cook's distance, define $\hat{y}_{(i)}$ as the predicted value for the i th observation after removing that observation to fit the model. For each of the n observations in the sample, calculate the squared residual $(y_i - \hat{y}_{(i)})^2$ and sum over the n observations. The *PRESS* statistic is

$$PRESS = \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^n \left(\frac{e_i}{1 - h_{ii}} \right)^2.$$

As with Cook's distance, the *PRESS* statistic is a function of the residual and leverage. A larger residual or larger leverage values lead to a larger *PRESS* statistic. Also as with the *SSPE*, a lower *PRESS* is preferred.

Example 2.22 (Section 2.2 Examples Summary). In this section we have discussed 10 models to illustrate the concepts underlying the main linear models. Table 2.7 summarizes these models by listing their included variables, as well as indicating the

Table 2.7. Model Summary

Name	Explanatory Variables	Examples
m0	Null	2.1, 2.7, 2.12
m1	AGE	2.2, 2.6, 2.10, 2.14, 2.15, 2.16
m2	MALE	2.3, 2.15, 2.16
m3	AGE + MALE + BLACK	2.4, 2.8, 2.9, 2.13, 2.14, 2.15, 2.16
m4	AGE + MALE + RACE	2.11, 2.15, 2.16
m5	AGE + MALE	2.11, 2.15, 2.16
m6	AGE + MALE + RACE + UNINSURED	2.15, 2.16
m7	AGE + MALE + RACE + INCOME	2.17, 2.18, 2.21
m8	AGE + MALE + RACE + log(INCOME)	2.17, 2.18, 2.21, 2.19
m9	AGE + AGE ² + MALE + RACE + MALE:RACE + log(INCOME)	2.19, 2.20

examples in which these models are discussed. The overall summary measures are included in Table 2.8, including the *PRESS* and *SSPE* results.

As discussed at the beginning of Section 2.2, the data consist of one year of data from two different panels (13, 14) of MEPs data that represent different sets of households. The *PRESS* statistics are calculated from panel 13 data that were used to estimate the model parameters. The *SSPE* statistics are calculated from panel 14 data, a separate sample of persons with diabetes. Panel 13 consists of 1,101 observations, whereas panel 14 consists of 1,039 observations.

Table 2.8 shows the model name, as well as the summary measures that we have discussed. The “best” model of a group of candidate models results in the lowest s^2 , *AIC*, *PRESS*, and *SSPE* statistics and the highest R_a^2 statistic. Although this constraint might not be met in most analyses, in this particular case, m9 with the explanatory

Table 2.8. Comparison of Summary Measures

Model	df	s^2	R^2	R_a^2	<i>AIC</i>	<i>PRESS</i>	<i>SSPE</i>
m0	1100	55.42	0.00	0.00	7548	61070	51841
m1	1099	54.19	0.02	0.02	7524	59765	50137
m2	1099	54.57	0.02	0.01	7532	60179	51301
m3	1097	52.27	0.06	0.06	7486	57751	50707
m4	1096	51.42	0.08	0.07	7469	56850	49756
m5	1098	53.35	0.04	0.04	7508	58884	49587
m6	1095	51.42	0.08	0.07	7471	56933	49727
m7	1095	51.26	0.08	0.07	7467	56775	49795
m8	1095	51.19	0.08	0.08	7465	56681	49937
m9	1092	50.02	0.10	0.10	7443	55497	48580

variables AGE + AGE² + MALE + RACE + MALE:RACE + log(INCOME) satisfies all constraints.

We are not advocating the use of this model to predict BMI for those with diabetes, but of the models examined, this one is the best choice.

2.2.9 Prediction

In much of actuarial work, once a linear model is finalized, the next step is to use the model to calculate predictions of some future outcome. As introduced in Section 2.2.6, the linear model is expressed as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with the fitted values expressed as $E[\mathbf{y} | \mathbf{X}] = \mathbf{X}\mathbf{b} = \hat{\mathbf{y}}$. For a given set of explanatory variables the fitted value, or the average value of the outcome variable, is estimated as $\mathbf{X}\mathbf{b}$.

To address the variability of this estimator, the literature presents two types of intervals: the *confidence* interval and the *prediction* interval. The confidence interval addresses the variance from the regression line, whereas the prediction interval addresses the variability of new observations.

We begin with $\text{Var}[\hat{z}]$, the variance of the fitted value for a particular observation with covariate vector of \mathbf{x}_z of dimension $1 \times (k + 1)$. Thus \hat{z} represents a particular value on the regression line. Recall that the design matrix, \mathbf{X} , has dimension $n \times (k + 1)$ and the dimension of $\boldsymbol{\beta}$ is $(k + 1) \times 1$:

$$\begin{aligned}\text{Var}[\hat{z}] &= \text{Var}[\mathbf{x}_z \mathbf{b}] \\ &= \mathbf{x}_z (\text{Var}[\mathbf{b}]) \mathbf{x}'_z \\ &= \sigma^2 \left(\mathbf{x}_z (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}'_z \right). \\ \widehat{\text{Var}}[\hat{z}] &= s^2 \left(\mathbf{x}_z (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}'_z \right).\end{aligned}$$

The third line substitutes $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ for $\text{Var}[\mathbf{b}]$ as in Section 2.2.6 for the variance of the regression parameter estimator. Because σ^2 is not known, we substitute s^2 as an estimator in the last line to indicate that we are estimating the variance.

Note that the dimension of the matrix $\widehat{\text{Var}}[\hat{z}]$ is a scalar of dimension 1×1 . The *confidence* interval for the fitted value of this observation is expressed as $\hat{z} \pm t_{1-\alpha/2} \cdot \sqrt{\widehat{\text{Var}}[\hat{z}]}$, where $t_{1-\alpha/2}$ is the *t*-value that results in an overall significance level of α .

The prediction interval is developed similarly; however, the starting point is $z = \mathbf{x}_z \boldsymbol{\beta} + \epsilon_z$, because we are indicating the variability for the value of the outcome variable itself and not its average. Here the variability of the predicted value is larger, because it includes an extra term for the variance of the individual error term. The *prediction* interval is expressed as $\hat{z} \pm t_{1-\alpha/2} \cdot \sqrt{s^2 + \widehat{\text{Var}}[\hat{z}]}$ or equivalently as $\hat{z} \pm t_{1-\alpha/2} \cdot s \cdot \sqrt{1 + \mathbf{x}_z (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}'_z}$.

2.3 Case Study

Using the foundation of linear models established in Section 2.2, we now address the broader methodology of building regression models in an applied setting. Successful practical applications of regression or other statistical modeling techniques do not begin with data and end with models. Rather, they begin with a question, problem, or strategic initiative and end with a model or models being used to improve decision making.

The following sections illustrate the major steps in (i) designing an analysis and preparing a suitable dataset, (ii) visually exploring the data, (iii) engaging in an iterative modeling process, and (iv) selecting and validating a final model. Although the steps appear to be sequentially completed, the process is truly iterative because patterns in residuals, outliers, or high leverage points might lead to changes in the model.

2.3.1 Analysis Design and Data Preparation

The first step of the modeling process is to design a data analysis strategy that adequately addresses the question or problem motivating the project. The purpose of our case study is to model annual health care expenditures of individuals who have been diagnosed with diabetes.⁷ The MEPS data are from panels 13 and 14 from 2009 and represent individuals who had some health care expenditures for 2009.⁸

The set of possible variables include those discussed in Section 2.2, as well as others. Medical expenditures (EXPEND), income (INCOME), age (AGE), and body mass index (BMI) are all continuous variables. The Mental Health Index (MENTHEALTH) is an ordinal categorical variable, with 1 to 4 denoting the best to worst mental health status. The race variable (RACE) is a categorical variable, with the values white, black, Asian, American Indian, Hawaiian, and multiethnic. For our case study, we collapsed the last four race categories into the single category “other.” Finally our MEPS sample contains 13 binary variables, 7 of which are comorbidity indicators. Finally we include a derived comorbidity count variable (COMORB.CNT) by simply counting the number of binary comorbidity variables that have the value “1.” The range of the COMORB.CNT variable is the set of integer values $\{0, 1, \dots, 7\}$. Deriving new variables is another aspect of the art of data analysis calling on creativity and judgment.

⁷ We continue to use the same MEPS data as in Section 2.2, but switch our analysis from studying BMI to analyzing annual health care expenditures.

⁸ These data are a weighted sample from the United States to allow national estimates. In this case study we do not take into account the weights of each observation, so the conclusions from this study are not representative of the U.S. population. For our study we consider only those with expenditures greater than 0.

The initial dataset contained 2,164 observations. As is often the case in actuarial work, several of the variables have missing values. Although a complete discussion of the treatment of missing data is beyond the scope of this chapter, we mention a few central concepts: for more detail, see Little and Rubin (2002) and Rubin (1976).

Broadly speaking, one can discard observations with missing data, elect not to use variables with missing data, impute missing values, or employ a combination of all three methods. If the outcome variable is missing for a particular observation, then that observation would provide no benefit in the estimation phase. If observations with missing data for explanatory variables are deleted from the dataset, the timing of when these observations are deleted could influence the resulting predictions. If the observations were deleted before the analysis started, then the dataset could be reduced dramatically because some potential variables might have had many missing data points. If those variables were determined not to be needed, then some observations may have been deleted prematurely and results could be biased. We deleted 24 observations that had missing values in either the COMORB.CNT or MENTHEALTH variables. See Exercise 2.6 for a further exploration of the treatment of missing data.

Summary statistics and plots of the reduced sample of 2140 observations are displayed in Table 2.9 and Figures 2.8 to 2.10. The SMOKER variable, an important risk factor influencing health expenditures, is problematic because it is missing in approximately 7% of the total cases. Furthermore, the median health care log-expenditures of smokers and nonsmokers are not materially different, as seen in the boxplot in Figure 2.10. Although not impossible, this finding is perhaps counterintuitive and seemingly at odds with the literature considering the health complications associated with smoking. The median expenditures for the missing data are lower than the medians for the smokers and nonsmokers, with a wider variance. These relationships are true in the aggregate and are not conditional on other variables that could modify the relationship. For the purpose of this case study we elect to discard the SMOKER variable from consideration.

We see the mean of our outcome variable, EXPEND, is \$11,162 with a standard deviation of \$18,331 and maximum of \$247,828. The distribution is quite skewed, as shown in Figure 2.8. Income is also a skewed variable. The mean of AGE is 60 with a range of 18 to 85. Fifty-seven percent of the observations are FEMALE.

Note that the categorical variables RACE and MENTHEALTH are both treated as sets of binary variables, where the sum of their means equals one. RACE is an unordered categorical variable, whereas MENTHEALTH is ordered. Twenty-five percent of the sample is BLACK, indicating the emphasis in the survey design of oversampling minority populations. Eleven percent of the observations represent those who are UNINSURED, and 18% have MEDICAID. Forty percent of the sample is EMPLOYED as measured on their first interview during the calendar year. The proportion of observations by MENTHEALTH category is highest in level 3 (good) and lowest in

Table 2.9. Summary of MEPS Panels 13 & 14 Data

	# Obs	Mean	Median	Std Dev	Min	Max
EXPEND	2,140	11,162	4804	18,331	4	247,828
LOGEXPEND	2,140	8.38	8.48	1.52	1.39	12.42
INCOME	2,140	47,816	34,518	43,511	1,000	359,704
LOGINCOME	2,140	10.4	10.45	0.91	6.91	12.79
AGE	2,140	59.78	60	13.72	18	85
BMI	2,140	31.77	30.7	7.26	9.4	75.1
COMORB . CNT	2,140	2.16	2	1.21	0	7
HIGHBP	2,140	0.76	1	0.42	0	1
ASTHMA	2,140	0.14	0	0.35	0	1
CANCER	2,140	0.15	0	0.36	0	1
CHOLEST	2,140	0.74	1	0.44	0	1
CORONARY	2,140	0.20	0	0.40	0	1
STROKE	2,140	0.12	0	0.32	0	1
EMPHYSEMA	2,140	0.05	0	0.22	0	1
SMOKER	1,981	0.16	0	0.36	0	1
MENTHEALTH.1	2,140	0.25	0	0.43	0	1
MENTHEALTH.2	2,140	0.24	0	0.43	0	1
MENTHEALTH.3	2,140	0.36	0	0.48	0	1
MENTHEALTH.4	2,140	0.15	0	0.36	0	1
MEDICAID	2,140	0.18	0	0.39	0	1
UNINSURED	2,140	0.11	0	0.31	0	1
EMPLOYED	2,140	0.40	0	0.49	0	1
HISPANIC	2,140	0.25	0	0.43	0	1
RACE.white	2,140	0.65	1	0.48	0	1
RACE.black	2,140	0.25	0	0.43	0	1
RACE.other	2,140	0.10	0	0.30	0	1
FEMALE	2,140	0.57	1	0.50	0	1

level 4 (fair or poor). Those with diabetes have from 0 to 7 of the specific comorbidities, with a mean of 2.16. Three-quarters of the sample have high blood pressure or high cholesterol.

Figure 2.8 displays histograms of health care expenditures (EXPEND) as well as its natural logarithm (LOGEXPEND). The former quantity is highly skewed. In contrast, LOGEXPEND has a roughly symmetric, bell-shaped distribution. We model LOGEXPEND as a linear combination of explanatory variables.

In large-scale projects, the choice of which potential explanatory variables to collect and analyze typically involves weighing the marginal cost of gathering and analyzing

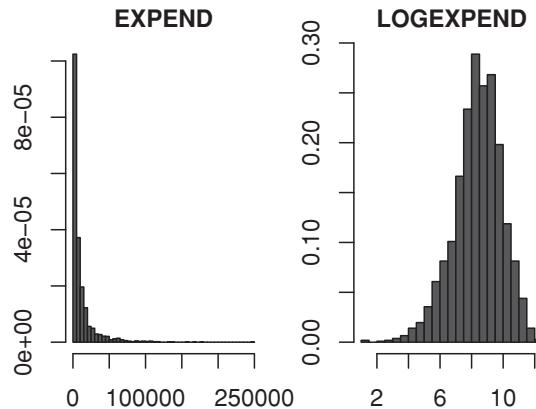


Fig. 2.8. Distribution of health care expenditures.

data elements against the likely benefits in the form of improved predictive accuracy. In many contexts, it is not uncommon to confront databases with hundreds or even thousands of candidate explanatory variables. According to the American Diabetes Association website (American Diabetes Association 2013b), complications of having diabetes include weight issues, eye troubles, hearing difficulties, foot and skin complications, heart troubles, risk of stroke, kidney disease, neuropathy, and mental health issues. For this case study, we used our judgment to preselect a manageable number of variables for further analysis.

A final step in the analysis design phase is to specify a model validation strategy. The gold standard of model validation is to evaluate the predictive accuracy of a model on a hold-out sample of data that were not used to train the model. We use panel 13 (1,101 observations) to train candidate models and set aside panel 14 data (1,039 observations) to validate the final selected model.

2.3.1.1 Exploratory Data Analysis

Once the analysis design is in place and the data have been prepared, the analysis phase begins. It is best to visually explore data before launching into a formal model-building process. The statistician John Tukey pioneered the discipline of graphical exploratory data analysis (EDA) and commented that “the greatest value of a picture is when it forces us to notice what we never expected to see” (Tukey, 1977, p. vi). At a minimum, we want to inspect variable distributions, as well as the relationships of the various explanatory variables with the outcome variable. For this case study, much of this relevant information can be conveyed in two images.

Figure 2.9 displays a scatterplot matrix of the outcome variable `LOGEXPEND` and the available continuous explanatory variables. The lower cells of this graphical matrix plot one variable versus one other variable. Note that `INCOME` is included on both the

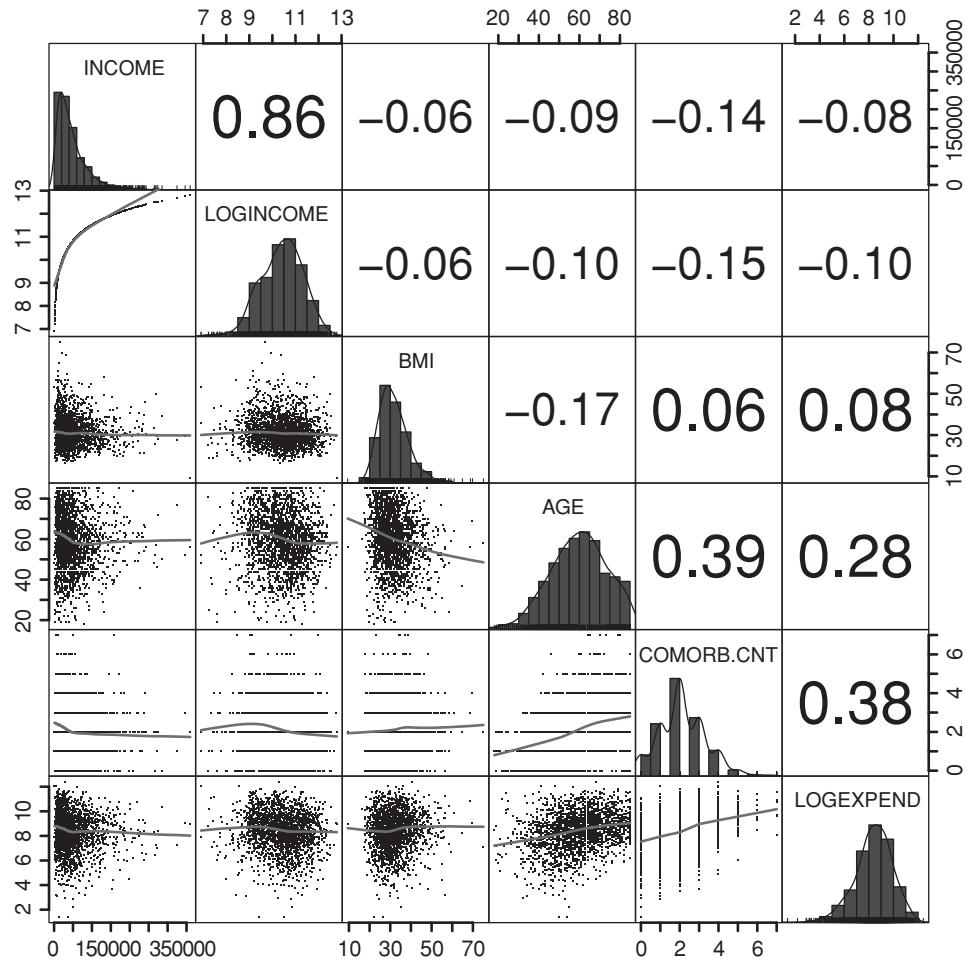


Fig. 2.9. Scatterplot matrix of outcome and continuous explanatory variables.

raw and logarithmic scale. The upper cells of the matrix display the sample Pearson correlation coefficient between variables. Finally, histograms of the variables are included along the diagonal.

The plots of the relationship of the variables with LOGEXPEND are shown in the last row, whereas the pairwise correlations with LOGEXPEND are displayed in the last column. These relationships are all roughly linear, the strongest being with age and comorbidity count. The scatterplot matrix also enables a check for pairwise collinearity among the explanatory variables using the correlation statistic. The highest explanatory variable correlation (between age and comorbidity count) is 0.39. All of the correlation coefficients displayed are significant at the 0.01 level, meaning that all correlations are significantly different from zero.

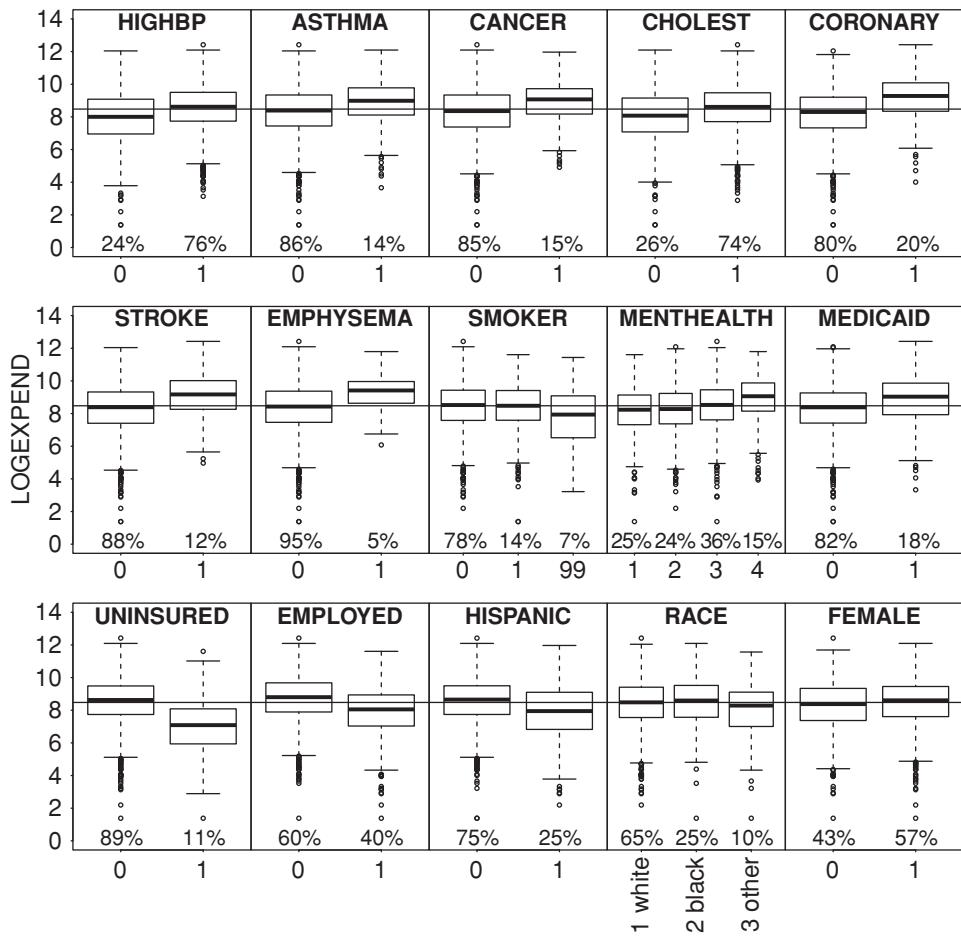


Fig. 2.10. Boxplots of LOGEXPEND versus categorical explanatory variables.

Figure 2.10 displays boxplots summarizing the distribution of LOGEXPEND (on the y-axis) by the levels for the various categorical variables. The outer edges of the boxes denote the 25th and 75th percentiles of LOGEXPEND within the particular category. The thick black line inside each box denotes the median value of LOGEXPEND within each category. Under each boxplot is the percent of observations contained in that level. These boxplots indicate that most of the categorical variables (with the possible exceptions of FEMALE and SMOKER) are somewhat correlated with LOGEXPEND. The SMOKER variable has three levels: smoker, nonsmoker, and missing (coded as '99').

2.3.1.2 Iterative Modeling Process

Given the small number of available explanatory variables, we begin by fitting a model containing all of the explanatory variables other than comorbidity count. If

Table 2.10. Case Study Model CS1

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	6.605280	0.662784	9.966	< 2e-16	***
LOGINCOME	0.031166	0.052052	0.599	0.549472	
FEMALE	0.033370	0.085051	0.392	0.694879	
AGE	0.009847	0.003780	2.605	0.009307	**
BMI	0.016560	0.005760	2.875	0.004119	**
RACE black	-0.073056	0.106146	-0.688	0.491442	
RACE other	-0.367215	0.138907	-2.644	0.008322	**
HISPANIC	-0.436667	0.104586	-4.175	3.22e-05	***
EMPLOYED	-0.318890	0.105312	-3.028	0.002520	**
UNINSURED	-0.944318	0.137622	-6.862	1.14e-11	***
MEDICAID	0.073254	0.123566	0.593	0.553414	
MENTHEALTH2	-0.022231	0.117598	-0.189	0.850095	
MENTHEALTH3	0.176169	0.109892	1.603	0.109203	
MENTHEALTH4	0.473380	0.140091	3.379	0.000753	***
EMPHYSEMA	0.290178	0.192247	1.509	0.131489	
STROKE	0.223260	0.141316	1.580	0.114432	
CORONARY	0.598302	0.110794	5.400	8.19e-08	***
CHOLEST	0.208910	0.096664	2.161	0.030899	*
CANCER	0.265052	0.121303	2.185	0.029099	*
ASTHMA	0.240329	0.124027	1.938	0.052919	.
HIGHBP	0.260372	0.100555	2.589	0.009745	**

Signif. codes:	0 '****'	0.001 '***'	0.01 '**'	0.05 '*'	0.1 '.'
					1
Residual standard error:	1.325	on 1080 degrees of freedom			
Multiple R-squared:	0.2706	, Adjusted R-squared:	0.2571		
F-statistic:	20.03	on 20 and 1080 DF,	p-value:	< 2.2e-16	

all of the comorbidity indicators were included together with COMORB.CNT, perfect multicollinearity would result, because COMORB.CNT is a linear combination of the disease indicators. The correlations shown in the last column of Figure 2.9 show the pairwise correlation of LOGEXPEND and motivate the inclusion of INCOME on a log scale rather than the original scale. This decision has the added benefit of preventing possible high leverage points later in the analysis. The resulting model, denoted *CS1*, is displayed in Table 2.10.

The very low *t*-statistics for the model coefficients corresponding to LOGINCOME, FEMALE, and MEDICAID suggest that these variables have fairly low marginal explanatory power in the presence of the other variables.⁹ Excluding these variables yields Model CS2 displayed in Table 2.11.

⁹ Although the coefficient for FEMALE is not significant, it is usually prudent to keep this variable in the model for communication with other stakeholders because it is an important variable of interest.

Table 2.11. Case Study Model CS2

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	6.967229	0.342968	20.315	< 2e-16	***
AGE	0.009575	0.003751	2.553	0.010829	*
BMI	0.016878	0.005706	2.958	0.003167	**
RACE black	-0.071886	0.103793	-0.693	0.488714	
RACE other	-0.353627	0.136984	-2.582	0.009967	**
HISPANIC	-0.427723	0.102873	-4.158	3.47e-05	***
EMPLOYED	-0.320100	0.097041	-3.299	0.001003	**
UNINSURED	-0.972722	0.131579	-7.393	2.87e-13	***
MENTHEALTH2	-0.023457	0.117416	-0.200	0.841696	
MENTHEALTH3	0.175822	0.109604	1.604	0.108972	
MENTHEALTH4	0.477781	0.138733	3.444	0.000595	***
EMPHYSEMA	0.284333	0.191251	1.487	0.137385	
STROKE	0.226059	0.140961	1.604	0.109071	
CORONARY	0.596957	0.110296	5.412	7.66e-08	***
CHOLEST	0.208344	0.096469	2.160	0.031015	*
CANCER	0.266285	0.121077	2.199	0.028067	*
ASTHMA	0.241906	0.121815	1.986	0.047302	*
HIGHBP	0.261191	0.100370	2.602	0.009387	**

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '
Residual standard error:	1.324	on 1083 degrees of freedom			
Multiple R-squared:	0.2701	, Adjusted R-squared:	0.2587		
F-statistic:	23.58	on 17 and 1083 DF,	p-value:	< 2.2e-16	

Because Models CS1 and CS2 are nested, an *F*-test can be performed to ascertain the joint significance of the three omitted variables: LOGINCOME, FEMALE, and MEDICAID. The resulting *F*-statistic is 0.2258 (*p*-value = 0.8785); that is, does not reject the null hypothesis that all coefficients are equal to zero. Excluding these three variables decreases the *AIC* from 3767 to 3762. This is consistent evidence that the three variables in question offer little explanatory or predictive power in the presence of the remaining explanatory variables. Note that Model CS1 has a higher R^2 but a lower R_a^2 than Model CS2. This illustrates why it is unwise to decide between models using R^2 .

Although both prior intuition and the EDA suggest a positive relationship between EMPHYSEMA and LOGEXPEND, as well as between STROKE and LOGEXPEND, an examination of the *t*-statistics from Model CS2 suggests that these comorbidity indicators are of marginal importance to the model. This prompts us to investigate the effect of excluding these variables. The *AIC* increases from 3762 to 3763, and the *F*-statistic is 2.54 (*p*-value = 0.08). These statistics are borderline as to whether to exclude or include these variables. For purposes of this case study, we elect to retain these variables in the model.

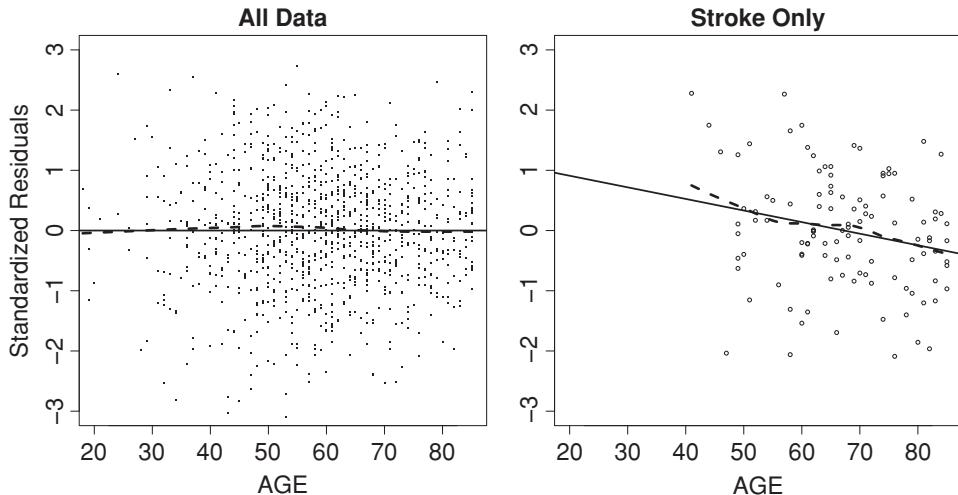


Fig. 2.11. Residual analysis related to age.

If we examine the signs of the coefficients, the directions are intuitive. As people age or increase in BMI, their expenditures tend to increase. Relative to whites, blacks and others have smaller expenditures. If one is employed, expenditures are less. If one is uninsured, there is less spending. As one's mental health declines, expenditures increase. Finally, the presence of a comorbidity increases expenditures.

2.3.2 Explanatory Variable Transformations

Model CS2 contains two continuous variables (AGE and BMI), each of which is assumed to have a linear relationship with LOGEXPEND. At this stage in the modeling process, it is good practice to check the residuals for reasonableness. Failures of linearity motivate the use of variable transformations such as creating categorical variables, taking logarithms, or including higher order polynomial terms.

Figure 2.11 illustrates one analysis relating to the AGE variable. The left figure displays standardized residuals from all of the data plotted against AGE. No pattern is apparent, indicating that there is no relationship between AGE and the residuals from the regression on LOGEXPEND. As a further check, one can verify that, when adding a quadratic term for AGE, the coefficient is not significant. Similar diagnostics can be completed for BMI.

A second assumption implicit in our model is that the linear relationship between AGE and LOGEXPEND is the same across the different levels of the various categorical explanatory variables (and similarly for BMI). The coefficient for AGE is 0.0096, indicating that, for two otherwise identical people, the health care expenditures are

Table 2.12. Case Study Model CS3

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.818068	0.348021	19.591	< 2e-16 ***
AGE	0.012126	0.003896	3.113	0.001903 **
BMI	0.017153	0.005695	3.012	0.002659 **
RACE black	-0.073454	0.103577	-0.709	0.478371
RACE other	-0.359568	0.136718	-2.630	0.008660 **
HISPANIC	-0.416785	0.102760	-4.056	5.35e-05 ***
EMPLOYED	-0.301692	0.097150	-3.105	0.001949 **
UNINSURED	-0.974983	0.131305	-7.425	2.27e-13 ***
MENTHEALTH2	-0.032085	0.117226	-0.274	0.784363
MENTHEALTH3	0.160967	0.109554	1.469	0.142043
MENTHEALTH4	0.476044	0.138442	3.439	0.000607 ***
EMPHYSEMA	0.262535	0.191071	1.374	0.169720
STROKE	2.102759	0.806800	2.606	0.009278 **
CORONARY	0.594447	0.110068	5.401	8.16e-08 ***
CHOLEST	0.206563	0.096268	2.146	0.032119 *
CANCER	0.265431	0.120822	2.197	0.028241 *
ASTHMA	0.243989	0.121561	2.007	0.044985 *
HIGHBP	0.249521	0.100280	2.488	0.012987 *
AGE:STROKE	-0.028056	0.011877	-2.362	0.018339 *
<hr/>				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
<hr/>				
Residual standard error: 1.321 on 1082 degrees of freedom				
Multiple R-squared: 0.2739, Adjusted R-squared: 0.2618				
F-statistic: 22.67 on 18 and 1082 DF, p-value: < 2.2e-16				

on average 0.96% ($= e^{0.0096} - 1$) higher for each additional year of age. This interpretation follows from the fact that the model estimates expenditures on the log scale.

However, it is possible that this rate of increase might differ for different subpopulations. The right-side of Figure 2.11 indicates that this is likely the case for STROKE. The plot shows the standardized residuals versus AGE only for individuals diagnosed with having had a stroke.¹⁰ The downward pattern in the residuals indicates that our model fails to adequately capture the relationship between age and expenditures for the population of stroke victims. This motivates us to add an interaction between AGE and STROKE to the model. The resulting model is displayed in Table 2.12.

The AGE : STROKE interaction is significant, and lowers the AIC from 3762 to 3758. The negative coefficient for this interaction term might seem counterintuitive, because it suggests that expenditures of people who have *not* had a stroke increase more with

¹⁰ Individuals diagnosed with having had a stroke are all more than 40 years old, which is why the data points appear bunched in the right-hand side of the plot.

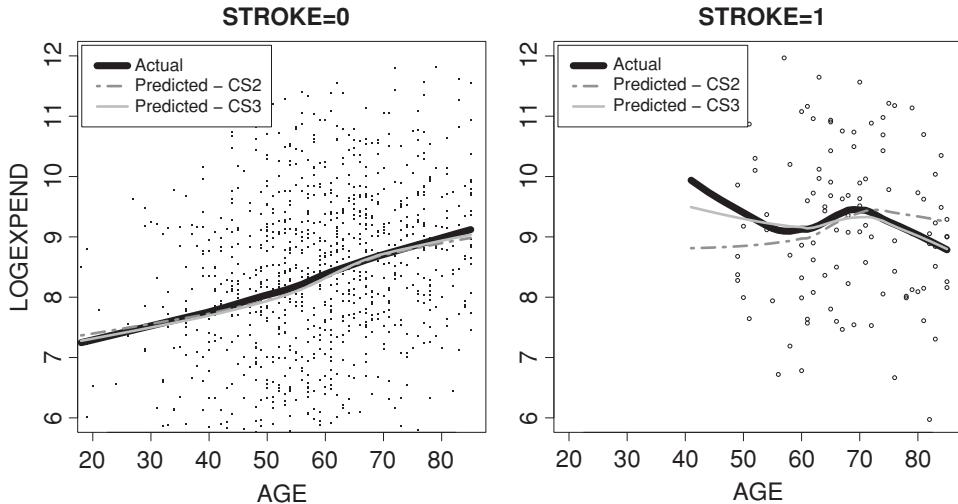


Fig. 2.12. Interaction between age and stroke on expenditures.

age than the expenditures of people who *have* had a stroke. When an interaction term is present, all of the related model parameters must be interpreted jointly. In this case study, this means that one needs to consider the coefficients for AGE and STROKE, and the interaction term for both. Data visualization provides helpful insight for this situation.

Figure 2.12 plots LOGEXPEND versus AGE separately for the subpopulations without and with a stroke. The thick black line in each plot is an empirical LOESS curve of the data, whereas the gray lines are LOESS-smoothed predicted values for Models CS2 and CS3, respectively. The scatterplot on the left suggests that for the nonstroke subpopulation, Models CS2 and CS3 capture the relationship between AGE and LOGEXPEND equally well, because the lines are indistinguishable.

The thick line on the right plot suggests that having a stroke is associated with higher expenditures earlier in life that do not increase with age, but generally decrease slightly. This is consistent with the coefficients of Model CS3. However, Model CS2 predicts an upward relationship that does not exist in the data. The coefficient for AGE in Model CS3 shows that each additional year of age for otherwise identical nonstroke victims is on average associated with 1.2% ($= e^{0.012} - 1$) higher health care expenditures. This, together with the coefficient for the AGE : STROKE interaction, implies that each additional year of age for otherwise identical stroke victims is on average associated with 1.6% ($= 1 - e^{(0.012-0.028)}$) lower health care expenditures. It is also noted from the right-hand graph that the level of expenditures is higher for those having had a stroke versus those who have not, which concurs with our prior knowledge.

As a caveat, what the included data do not indicate is the age at which the person had a stroke or the time since they had a stroke. This extra information may be informative for the analysis. Our purpose here is to demonstrate the use of graphical techniques and ways to interpret regression output with interaction terms.

2.3.3 Interpreting Model Coefficients

The overall intercept term is equal to 6.818. This is the average LOGEXPEND for unemployed, insured white, non-Hispanic individuals in excellent mental health and with no comorbidities, and for whom AGE = 0 and BMI = 0. Because the last two conditions are never met, the intercept term has no interpretation in isolation from the other terms in the model.

When the AGE : STROKE interaction term is introduced, the coefficient for STROKE jumps from 0.226 in Model CS2 to 2.103 in Model CS3. This seemingly large increase in the STROKE coefficient cannot be analyzed in isolation from the AGE : STROKE interaction term. Because there are no observations with AGE = 0, the STROKE coefficient has no meaning in isolation. In Model CS3, an amount on a log scale, ranging from $-0.504 (= -0.028 \times 18)$ to $-2.38 (= -0.028 \times 85)$, is reflected in the fitted equation for the interaction term AGE : STROKE depending on the age of the person.

More interpretable model parameters result if we *center* the continuous explanatory variables. Specifically, we replace AGE and BMI with $AGE.C = AGE - 60$ and $BMI.C = BMI - 32$, where the values subtracted are the appropriate means of the respective variables. All of the resulting model coefficients and goodness of fit statistics remain unchanged except for the intercept term, which increases from 6.818 to 8.095, and the STROKE coefficient, which decreases from 2.10 to 0.419. For otherwise identical subpopulations of 60-year-old individuals, having had a stroke diagnosed is associated with 52% higher ($= e^{0.419} - 1$) health care expenditures. Similarly, the intercept term corresponds to the expenditures of individuals with average age and BMI, and in the reference category for the other variables in the model.

2.3.4 Model Validation

A number of residual plots from Model CS3 are displayed in Figures 2.13 and 2.14. The upper left image in Figure 2.13 is a *QQ*-plot indicating that the standardized model residuals are, to a reasonable approximation, normally distributed. The upper right image displays a scatterplot of the standardized residuals versus the predicted values. This is the most commonly displayed residual plot. No relationship between the residuals and predicted values is discernible, and the homoskedasticity assumption does not appear to be violated. Further, horizontal lines are drawn at the empirical 2.5 and 97.5 percentiles of the standardized residual distribution. These lines are close to

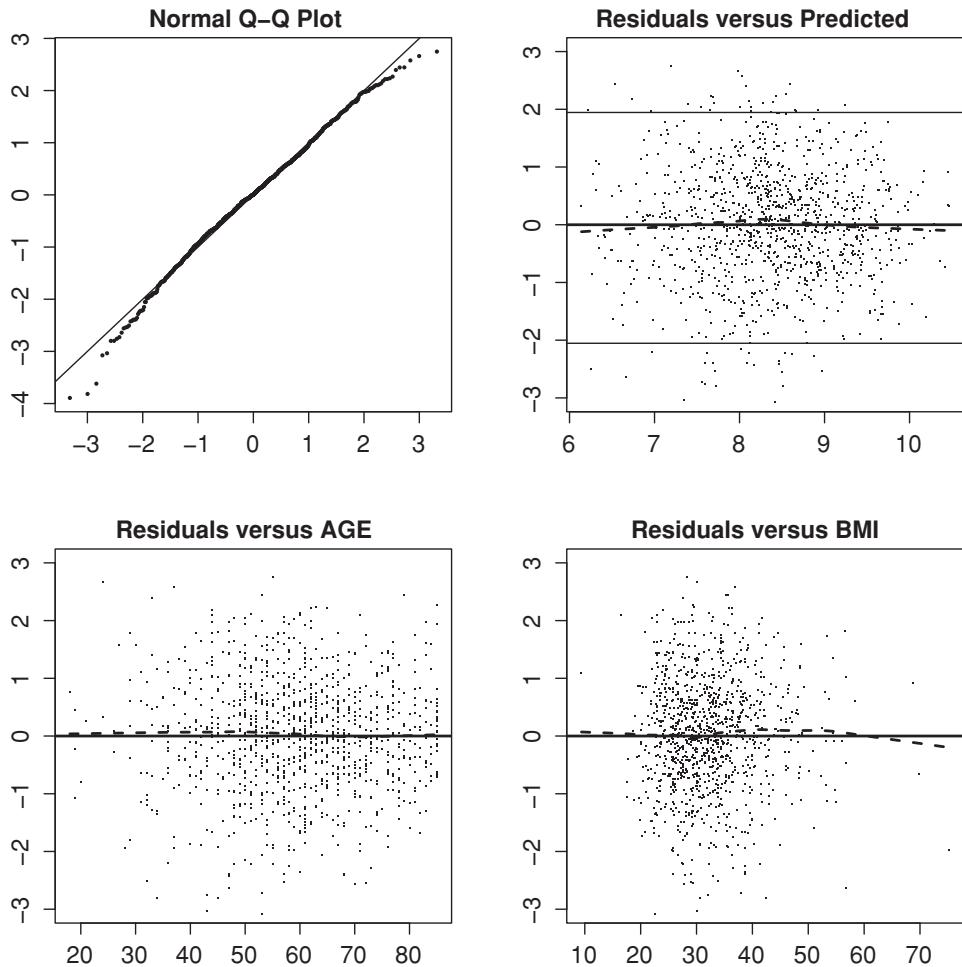


Fig. 2.13. Final residual analysis – model CS3.

the corresponding theoretical standard normal percentile values of -1.96 and 1.96 , respectively.

The bottom images in Figure 2.13 display the standardized residuals plotted against AGE and BMI, respectively. Again, no suspicious patterns are apparent. The boxplots in Figure 2.14 indicate that the standardized residuals have essentially the same distribution within each level of the various categorical explanatory variables. In short, at a high level everything seems to be in order. Further checks for interactions could be completed, but as discussed previously, the benefits of spending more time must be weighed against the costs of further effort.

Once we are satisfied that our model adequately captures the salient features of the training data, it is appropriate to evaluate its predictive accuracy on the hold-out

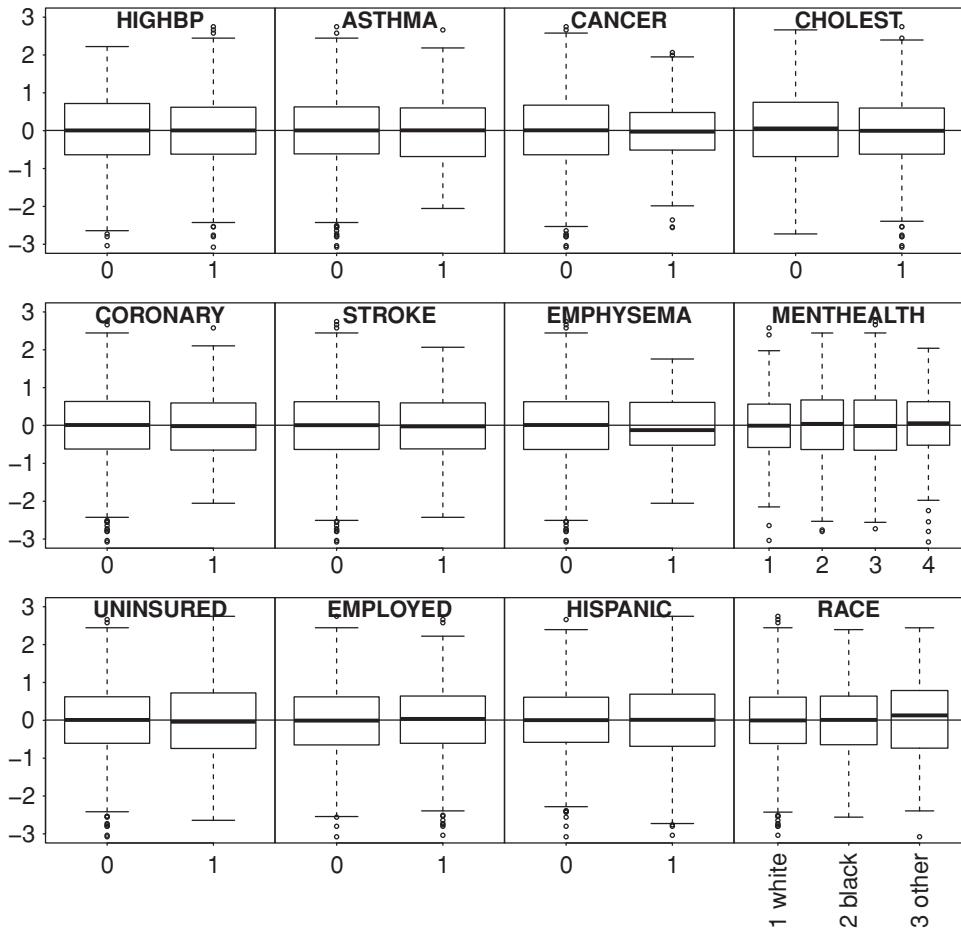


Fig. 2.14. Standardized residuals versus categorical variables.

data. Recall that the hold-out data have not been used in the model estimation process. Figure 2.15 displays two predictive validation analyses. On the left is a scatterplot of predicted versus actual values of LOGEXPEND. The sample correlation of these quantities is 0.49 (p -value < 0.0001), an improvement over the one-way correlations observed in the EDA scatterplot.

On the right is a *gains chart*, a model validation tool that facilitates an assessment of the economic significance of the predictive accuracy of the model. Here, the hold-out data are sorted from highest to lowest predicted LOGEXPEND. Next, the data are grouped into 10 equal-sized deciles. For example, decile 1 contains the 10% of the observations with the *highest* predicted expenditures, and decile 10 contains the 10% of the observations with the *lowest* predicted expenditures. The total dollars of expenditures are calculated within each decile. Finally, the cumulative percent of these expenditures are plotted against the cumulative percent of individuals. The fourth

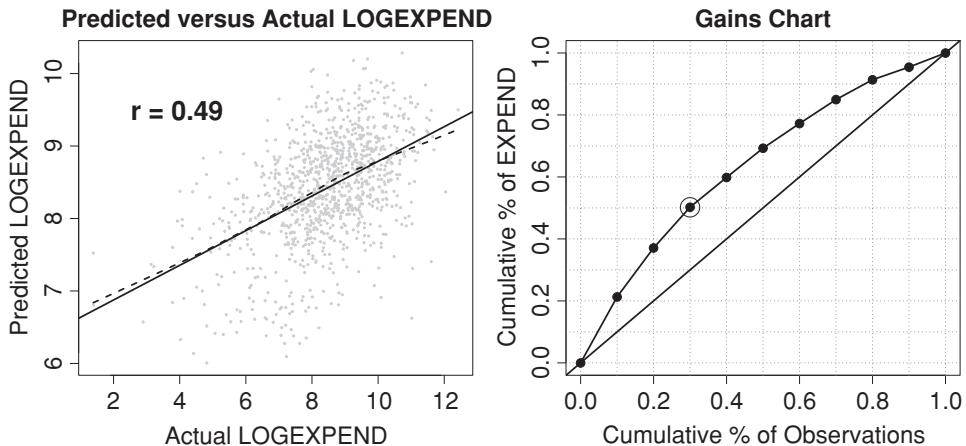


Fig. 2.15. Predictive accuracy of hold-out data.

point from the left of the gains chart (circled) indicates that the 30% of the individuals with the highest *predicted* expenditures account for approximately 50% of the total *actual* expenditures in the hold-out data. In contrast a random selection of 30% of the individuals in the hold-out data would on average account for approximately 30% of the expenditures. A gains chart conveys the degree to which using the model to rank individuals in terms of predicted expenditures improves on randomly selecting individuals.

Another model exercise, suggested in Exercise 2.3, is to reestimate the model parameters on the hold-out data and apply that resulting model to the training data. This enables a comparison of the two sets of model coefficients for consistency. It provides further evidence that our model primarily reflects aspects of the processes that generated the data, rather than accidental features of the data arising from random variation. Therefore it is not unreasonable to expect the model to make comparably good predictions when applied to future observations. Finally, applying this reestimated model to the training data to calculate the predicted values yields a correlation coefficient of 0.52 with LOGEXPEND.

In short, although the model could be improved, no major problems have emerged from our various validation checks, and it is reasonable to bring the analysis to a close.

2.4 Conclusion

This chapter has provided a practical overview of linear models. First, it outlined fundamental assumptions and other statistical concepts. Second, these concepts were woven into a series of examples to help reinforce the material. This chapter serves as a foundation for the rest of the volume in that many of the subsequent chapters relate to linear models in some way.

Chapter 5, “Generalized Linear Models,” relaxes two of the linear model assumptions. First, rather than assuming that the expected value of the outcome variable equals a linear combination of explanatory variables, the GLM framework assumes linearity on a monotonic scale (such as logarithmic, inverse, or log-odds) provided by a link function. Second, the normality assumption for the dependent variable is relaxed to one from an exponential family distribution. The *Error SS* concept is generalized to the concept of deviance. Many of the variable selection, EDA, and model validation concepts discussed in this chapter carry over to the broader GLM framework.

Chapters 8 and 16 on mixed models, also known as multilevel/hierarchical models (Gelman and Hill 2008), can be viewed as regression models for which certain model parameters are given probability submodels. Such models are often useful when the data are naturally grouped along one or more dimensions. These models incorporate parameter shrinkage, an important phenomenon that is closely related to credibility weighting. Furthermore, Bayesian versions of (generalized) linear and multilevel/hierarchical models can be specified by providing prior probability distributions for each of the model parameters.

Chapter 15 on generalized additive models generalize the GLM and linear modeling frameworks by replacement of a linear combination of predictors with a linear combination of semi-parametric transformations of predictors.

Finally, a number of machine learning techniques can be viewed as extensions of the linear model. For example the multivariate adaptive regression spline (MARS) algorithm is a computationally intensive technique that searches through a high-dimensional space of transformed variables and variable interactions. (Hastie, Tibshirani, and Friedman 2009). Other popular machine-learning techniques include classification and regression trees (CART), random forests, and support vector machines, (Hastie et al. 2009). It can be helpful to use machine learning techniques in the course of (generalized) linear modeling projects for at least two reasons. First, such techniques can be used as heuristic tools for suggesting certain variables, variable transformations, and variable interactions for further investigation. Second, highly complex yet accurate “black box” models provided by machine-learning algorithms can serve as baselines against which to compare the predictive accuracy of more traditional models.

Many of the topics covered in later chapters in this volume, such as credibility, survival models, and time series models, can be viewed within the linear model framework.

2.5 Exercises

The first four exercises use the MEPS data (MEPS.Diabetes.csv) that were analyzed in Sections 2.2 and 2.3 of this chapter.

Exercise 2.1. This exercise uses the training data to explore aspects of correlation, addition of variables, and interpretation of coefficients in a linear model.

- Regress LOGEXPEND on LOGINCOME. From this output, calculate the standard deviation of LOGEXPEND on LOGINCOME. Using the regression output and the sample correlation shown in Figure 2.9, verify the formula introduced in Example 2.5 between the regression coefficient estimate and the sample correlation. How do you interpret this regression coefficient? What is the relationship between the R^2 of this model and the sample correlation coefficient between LOGEXPEND and LOGINCOME?
- Regress LOGEXPEND on LOGINCOME and MENTHEALTH and perform an F -test to evaluate the significance of MENTHEALTH. Does the result indicate whether or not to include MENTHEALTH in the model?
- How would you justify the difference in the value of the regression coefficient for LOGINCOME between models in 2.1(a) and 2.1(b)?

Exercise 2.2. Use the model developed in Exercise 2.1(b) to explore the residuals.

- To gain insight into whether AGE might be a significant addition to the model, create a scatterplot with the standardized residuals plotted on the y -axis and AGE plotted on the x -axis. Add (i) a horizontal line where the residuals are equal to zero and (ii) a regression line or a loess curve to help visualize the patterns. What does this residual analysis suggest?
- Create a new model with AGE as the dependent variable and LOGINCOME and MENTHEALTH as the explanatory variables. Calculate the residuals from this model. Create a scatterplot of the residuals from the model created in Exercise 2.1(b) on the y -axis and the residuals from this new model on the x -axis. The resulting graph is called an *added variable plot*.
- Calculate the sample correlation between the residuals from the models used in the scatterplot in 2.2(b). This quantity is known as the *partial correlation* between LOGEXPEND and AGE, controlling for LOGINCOME and MENTHEALTH.
- Regress LOGEXPEND on LOGINCOME, MENTHEALTH, and AGE. Using the t -statistic for AGE and degrees of freedom for the residual standard error, verify that the partial correlation computed in 2.2(c) is directly computed by:

$$r(y, x_j | x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_k) = \frac{t(b_j)}{\sqrt{t(b_j)^2 + n - (k + 1)}}$$

Exercise 2.3. To perform further validation of the final case study Model CS3, we reverse the roles of the training and hold-out data samples to provide insight into the stability of the estimated model coefficients.

- Refit the final case study Model CS3 on the hold-out data and compare the resulting model coefficients with the coefficients in Model CS3.

- (b) Use the resulting regression model estimates to compute the predictions on the training data, and compute the sample correlation coefficient between these predictions and LOGEXPEND.

Exercise 2.4. As an alternative to the final case study Model CS3, this exercise examines the use of the derived variable, COMORB.CNT, defined as the sum of the various comorbidity indicators in the model.

- (a) The output of Model CS3 indicates that the coefficients for EMPHYSEMA, CHOLEST, CANCER, ASTHMA, HIGHBP are of similar magnitude. Modify the model in the following way: (i) drop the comorbidity indicators EMPHYSEMA, CHOLEST, CANCER, ASTHMA, and HIGHBP and (ii) add COMORB.CNT as an explanatory variable. Label the resulting Model CS4. Justify the use of these covariates and explain the results.
- (b) Is it meaningful to perform an F -test to compare Models CS3 and CS4? Why or why not?
- (c) Compare the AIC of both models. What conclusion can be drawn as to the choice of model?
- (d) Create a scatterplot matrix to compare the predicted values of the two models on the hold-out data.

Exercise 2.5. This exercises highlights the dangers of *data snooping* using simulation.

- (a) Generate samples of length 1,000 from the independent standard normal random variables Y, X_1, \dots, X_{100} . Next, regress Y on X_1, \dots, X_{100} .
- (b) Given that Y is generated independently from the X_i , how many explanatory variables would you expect to find significant at the 5% level? Compare your answer to the output of the model.

Exercise 2.6. This exercise uses simulation to investigate various strategies for handling missing data.

- (a) Simulate 1,000 draws for x_1 and x_2 from a bivariate normal distribution with $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$, and $\rho = 0.7$. Hint: In R, install the `mvtnorm` package and use the following line of code:

```
x <- rmvnorm(1000, mean=c(0, 0), sigma=matrix(c(1, .7, .7, 1),
ncol=2))
```

See `help(package = mvtnorm)` for assistance.

- (b) Simulate 1,000 draws from a standard normal distribution and label the resulting vector ϵ . Finally set $y = 3 + 5 * x_1 + 7 * x_2 + \epsilon$ and create a data frame containing the fields y, x_1, x_2 , and ϵ .

- (c) Conduct audit checks on the data to verify that the simulated data conform to your specifications. For example, for the vector of errors, see that the mean equals 0, the variance equals 1, and the histogram looks normally distributed. For the multivariate distribution, you can assess the marginals and check on the correlation between x_1 and x_2 .
- (d) Create a scatterplot matrix of the resulting dataset and examine the relationship between the outcome variable, the explanatory variables, and the error term.
- (e) Regress y on x_1 and x_2 and compare the resulting coefficients with the “true” parameter values.
- (f) Replace a random 25% of the values of x_2 with the missing value NA. Rerun the regression in 2.6(e) and compare the resulting coefficients with the “true” parameter values. Note the number of observations used in estimating the model.
- (g) Replace the missing values of x_2 with the mean value of the nonmissing values. Label the resulting variable $x2.imp$. Regress y on x_1 and $x2.imp$. Note that all 1,000 observations are used to fit this model. Note that the parameter estimates are biased (i.e., differ from truth) and that the standard errors are larger than in the previous models.
- (h) Now create a binary variable that takes on the value 1 if an observation was mean-imputed, and 0 otherwise. Add this binary variable to the model created in 2.6(g) and create a new model.
- (i) Compare the resulting coefficients of all models to the “true” parameter values. Create a scatterplot matrix of the modified dataset to help understand the changes in model coefficients. Note the large number of observations (25% in this case) that are zero.
- (j) Add an interaction term between x_1 and the mean-imputation indicator variable created above to create a new model. Compare the resulting coefficients with the “true” parameter values.
- (k) Finally, change $\rho = 0.7$ to $\rho = 0$ and repeat steps 2.6(a) to 2.6(j) of this exercise. Note the effects of the mean-imputation strategy in the scenario when x_1 and x_2 are uncorrelated.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6), 716–723.
- American Diabetes Association (2013a). Diabetes information. <http://www.diabetes.org/diabetes-basics/?loc=GlobalNavDB>.
- American Diabetes Association (2013b). Diabetes information. <http://www.diabetes.org/living-with-diabetes/complications/>.
- Brien, R. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity* 41, 673–690.
- Centers with Disease Control and Prevention (2013). Body mass index information. http://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html.

- Dall, T. M., S. E. Mann, Y. Zhang, J. Martin, Y. Chen, and P. Hogan (2008). Economic costs of diabetes in the U.S. in 2007. *Diabetes Care* 31(3), 1–20.
- Draper, N. R. and H. S. Smith (1998). *Applied Regression Analysis*. Wiley, New York.
- Freedman, D. A. (2005). *Statistical Models: Theory and Practice*. Cambridge University Press, New York.
- Frees, E. W. (2010). *Regression Modeling with Actuarial and Financial Applications*. Cambridge University Press, New York.
- Gelman, A. and J. Hill (2008). *Data Analysis Using Regression and Multilevel/Hierarchical Modeling*. Cambridge University Press, New York.
- Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (2nd edition). Springer-Verlag, New York.
- Little, R. and D. Rubin (2002). *Statistical Analysis with Missing Data, Second Edition*. Wiley, New Jersey.
- Medical Expenditure Panel Survey (2013). Meps website. www.meps.ahrq.gov/.
- Rubin, D. (1976). Inference and missing data. *Biometrika* 63(3), 581–592.
- Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press, Cambridge, MA.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>.
- Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts.

3

Regression with Categorical Dependent Variables

Montserrat Guillén

Chapter Preview. This chapter presents regression models where the dependent variable is categorical, whereas covariates can either be categorical or continuous. In the first part binary dependent variable models are presented, and the second part is aimed at covering general categorical dependent variable models, where the dependent variable has more than two outcomes. This chapter is illustrated with datasets, inspired by real-life situations. It also provides the corresponding R programs for estimation, which are based on R packages `glm` and `mlogit`. The same output can be obtained when using SAS or similar software programs for estimating the models presented in this chapter.

3.1 Coding Categorical Variables

Categorical variables measure qualitative traits; in other words, they evaluate concepts that can be expressed in words. Table 3.1 presents examples of variables that are measured in a categorical scale and are often found in insurance companies databases. These variables are also called *risk factors* when they denote characteristics that are associated with losses.

Categorical variables must have mutually exclusive outcomes. The number of categories is the number of possible response levels. For example, if we focus on insurance policies, we can have a variable such as `TYPE OF POLICY CHOSEN` with as many categories as the number of possible choices for the contracts offered to the customer. This categorical variable is valid if the customer can choose one and only one option. An introduction to categorical variables was given in Chapter 2.

A categorical variable with many categories can be understood as a set of binary variables that capture the presence or absence of each possible choice. For example, we can work with the categorical variable `TYPE OF VEHICLE`. In Table 3.1 this variable has four possible categories (motorbike, car, van, or truck). Equivalently, we can define four binary variables. The first one indicates whether or not the vehicle is

Table 3.1. Examples of Categorical Variables in Insurance

Variable	Categories
Sex	Man or woman
Bodily injury coverage	Yes or no
Deductible	Present or absent
Policy renewal status	Renewed or canceled
Opinion	Against, neutral, or favorable
Size	Small, medium, or large
Type of vehicle	Motorbike, car, van, or truck
Number of policies owned by the same customer	1, 2, at least 3

a motorbike, the second one indicates whether or not the vehicle is a car, and so on. For simplification, when using a binary variable we usually denote the presence by “1” and the absence by “0.”

When a categorical variable measures choices, then we say that each category is a possible response. Choices can be unordered, as with sex, or ordered, such as a small, medium, or large type of answer. Ordered choices can be modeled using special purpose models, but they can also be treated as unordered responses, if we just ignore the order.

It is common to call the dependent categorical variable itself the response. If the response has two possible outcomes, then one outcome is called the *event* and the other one is called the *baseline*, the *reference*, or the *non-event*. When using statistical software, one has to be careful about the way the categories are labeled by internal procedures, which fix the way models are specified by default. Before starting the model estimation, one has to be sure about the way categorical responses are handled, and particularly which is the response that corresponds to the *baseline* choice.

The baseline category is often assigned by practitioners to the one that is most frequent in the database or in the population, or to the one that indicates the non-event. This is done to facilitate the interpretation of coefficients. It seems more natural to speak about an event occurrence, rather than having to talk about it not occurring having too many negative words in a sentence makes it much more complex, and communicating ideas becomes cumbersome.

3.2 Modeling a Binary Response

Databases contain information on n individuals, who are often also called units, cases, or observations and who correspond, in the insurance context, to policyholders, firms, business units, agents, and so on. Let us assume that y_i is the categorical response for individual i . In the simplest case, y_i is binary and can only take two possible values, which we code as 1 for the event, and 0 for the non-event, respectively. We assume y_i follows a Bernoulli distribution with probability π_i .

As in linear regression techniques, we are interested in using characteristics of policyholders, such as age, sex, education, driving experience, and type of vehicle, to help explain the dependent variable y_i . The probability of the event response, π_i , depends on a set of individual characteristics, which we denote by column vector \mathbf{x}_i . In generalized linear models, π_i is expressed as a general function of a linear combination of the characteristics, $\mathbf{x}'_i \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a vector of unknown parameters to be estimated.

The simplest model for a binary response just mimics the classical linear regression model described in Chapter 2. Here it is called the *linear probability model*, which specifies that $y_i = \mathbf{x}'_i \boldsymbol{\beta} + \epsilon_i$, where ϵ_i is an error term. This model has several drawbacks. First, the fitted response, \hat{y}_i , which is the predicted expectation of y_i , does not necessarily lie between 0 and 1 and so cannot be interpreted as a probability. Second if ϵ_i is normally distributed, then y_i should also be normally distributed because it depends linearly on ϵ_i . As y_i is binary, then a normal law is wrong. And finally, the linear probability model is heteroskedastic because the variance of y_i equals $\pi_i(1 - \pi_i)$, which is not constant in the dataset because π_i depends on \mathbf{x}_i . Note that $E(y_i) = 0 \times \Pr(y_i = 0|\mathbf{x}_i) + 1 \times \Pr(y_i = 1|\mathbf{x}_i) = \pi_i$ and then $\pi_i = \mathbf{x}'_i \boldsymbol{\beta}$ if we assume that the expectation of the error term is zero.

For a binary dependent variable, given the predictors, the probabilities of each binary outcome are

$$\Pr(y_i = 1|\mathbf{x}_i) = \pi_i$$

and

$$\Pr(y_i = 0|\mathbf{x}_i) = 1 - \Pr(y_i = 1|\mathbf{x}_i) = 1 - \pi_i \quad i = 1, \dots, n.$$

3.3 Logistic Regression Model

In the logistic regression model the dependent variable is binary. This model is the most popular for binary dependent variables. It is highly recommended to start from this model setting before carrying out more sophisticated categorical modeling. Dependent variable y_i can only take two possible outcomes. We assume y_i follows a Bernoulli distribution with probability π_i . The probability of the event response π_i depends on a set of individual characteristics \mathbf{x}_i .

The logistic regression model is an extension of the linear probability model.

3.3.1 Specification and Interpretation

The *logistic regression model* specifies that

$$\Pr(y_1 = 1|\mathbf{x}_i) = \pi_i = \frac{1}{1 + \exp(-\mathbf{x}'_i \boldsymbol{\beta})} = \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta})}{1 + \exp(\mathbf{x}'_i \boldsymbol{\beta})}, \quad (3.1)$$

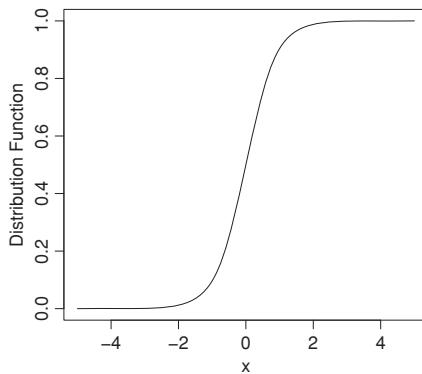


Fig. 3.1. Shape of the logistic function.

and the inverse of this relationship, called the *link function* in generalized linear models, expresses $\mathbf{x}'\boldsymbol{\beta}$ as a function of π_i as

$$\mathbf{x}'\boldsymbol{\beta} = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \text{logit}(\pi_i). \quad (3.2)$$

We call $\pi_i/(1 - \pi_i)$, the *odds* and its natural logarithm is the *log-odds* or the *logit*. So, because it clearly follows from (3.2) in the logistic regression model, we say the log-odds is a linear combination of the predictors.

In the logistic regression model, parameters are interpreted in terms of log-odds rather than directly on the response. Moreover, the predicted response is the probability of the event. To put it in simple words, the logistic regression is a model to predict a probability.

The logistic regression model is popular because the fitted responses lie between 0 and 1, and therefore, the fitted values can always be interpreted easily as the estimated probability for the modeled event, namely $y_i = 1$. The logistic function is defined as $f(x) = 1/(1 + \exp(-x))$, and it is represented in Figure 3.1. The S-shaped curve that results from the combination of risk factors, or covariates, is appealing because in the extremes the effect of the linear combination of risk factors on the estimated probability smooths down.

The logistic regression model is widely used in insurance, as well as in many other disciplines. In medicine or pharmaceutics, this model is the basis for analysis of the cause effects of risk factors or conditions that influence the positive or negative response of a patient to a treatment.

Parameters are interpreted in the binary logistic regression model setting in a more complicated way than in the linear regression model context. A bit of practice is needed to get used to talking about fitted probabilities instead of fitted values.

When interpreting the estimation results of a logistic regression model, a first, useful step is to analyze the sign of the parameters and to check if the estimated signs are those that prior intuition or theory was indicating. A positive parameter means that an increase in the covariate that is associated with this parameter implies an increase in the probability of the event response. Conversely, if a parameter is negative, then when the predictor increases, the probability of the modeled event decreases.

Logistic regression models used in practice should always contain a constant term. The value of the constant is not directly interpretable. It is used as a level, and it corresponds to the natural logarithm of the event probability when all regressors are equal to zero.

The *odds-ratio* is the name given to the exponential of a parameter. Let us assume that individual i changes its k -th predictor, x_{ik} , in one unit. For instance, let us assume that initially that characteristic x_{ik} equals c and then it equals $c + 1$. Then it is easily seen that

$$\begin{aligned}\exp(\beta_k) &= \frac{\Pr(y_i = 1 | \mathbf{x}_i, x_{ik} = c + 1) / (1 - \Pr(y_i = 1 | \mathbf{x}_i, x_{ik} = c + 1))}{\Pr(y_i = 1 | \mathbf{x}_i, x_{ik} = c) / (1 - \Pr(y_i = 1 | \mathbf{x}_i, x_{ik} = c))} \\ &= \frac{e^{\beta_j(c+1)}}{e^{\beta_j(c)}}.\end{aligned}$$

So,

$$\beta_k = \ln \left(\frac{\Pr(y_i = 1 | \mathbf{x}_i, x_{ik} = c + 1)}{1 - \Pr(y_i = 1 | \mathbf{x}_i, x_{ik} = c + 1)} \right) - \ln \left(\frac{\Pr(y_i = 1 | \mathbf{x}_i, x_{ik} = c)}{1 - \Pr(y_i = 1 | \mathbf{x}_i, x_{ik} = c)} \right).$$

If K is the total number of regressors in the model, including the constant term, then odds-ratio OR_k equals $\exp(\beta_k)$, for $k = 1, \dots, K$.

We can say that the odds for $x_{ik} = c + 1$ are $\exp(\beta_k)$ times the odds when $x_{ik} = c$. If the k -th explanatory variable is continuous, we can also derive that parameter β_k is the proportional change in the logarithm of the odds-ratio. This is connected to the notion of the economic *elasticity*.

A connected concept to the *odds-ratio* is the notion of *risk-ratio*. The *risk-ratio* is the ratio of event probabilities when the predictor indicator changes by one unit. It is especially important in indicating the relative change in the probability of an event when a risk indicator is present rather than absent. The definition is

$$RR_k = \frac{\Pr(y_i = 1 | \mathbf{x}_i, x_{ik} = c + 1)}{\Pr(y_i = 1 | \mathbf{x}_i, x_{ik} = c)} \quad k = 1, \dots, K.$$

Risk-ratios depend on the \mathbf{x}_i and cannot be expressed as a function of a single parameter.

Example 3.1 (Fullcoverage.csv data file). Table 3.2 shows the distribution by gender of 4,000 policyholders of motor insurance selected at random. The response

Table 3.2. Number of Policyholders who Choose Other Coverage or Full Coverage as a Function of Predictors

		Other Coverage $y = 0$	Full Coverage $y = 1$	Total
SEX	woman	498(50.30%)	492(49.70%)	990
	man	2115(70.27%)	895(29.73%)	3010
DRIVING AREA	rural	1906(72.83%)	711(27.17%)	2617
	urban	707(51.12%)	676(48.88%)	1383
VEHICLE USE	commercial	33(84.62%)	6(15.38%)	39
	private	2580(65.14%)	1381(34.86%)	3961
MARITAL STATUS	single	467(54.24%)	394(45.76%)	861
	married	2047(68.85%)	926(31.15%)	2973
	other	99(59.64%)	67(40.36%)	166
AGE (YEARS)		48.27	43.09	46.47
SENIORITY IN COMPANY (YEARS)		9.93	10.88	10.88

Notes: Row percents indicate the proportion of policyholders. For quantitative regressors, the mean is shown.

variable is whether they bought full coverage (event coded as 1) or not (non-event coded as 0). For this example, full coverage usually refers to a type of contract that compensates for any physical damage to the insured vehicle, including collision damage as a result of an accident; it also covers stolen or vandalized vehicles as well as storm or flood damage or an impact with an inanimate object.

Predictors for coverage choice are SEX, DRIVING AREA, VEHICLE USE, MARITAL STATUS of the policyholder, AGE, and SENIORITY in the company.

Table 3.3 presents the parameter estimates for the logistic regression model to predict the probability that a policy holder chooses FULL COVERAGE. Parameter estimates are obtained by the maximum likelihood method. The logistic regression is a special case of the generalized linear model, and maximum likelihood estimation in this context is covered in Chapter 5.

We can see that the parameter associated with variable AGE is negative (-0.058) and significantly different from zero (small p -value). This result means that the older the policyholder is, the lower are the chances he or she buys full coverage. Likewise, we expect that men would be less prone to buying full coverage than women. The parameter for MEN is negative (-0.961) and significantly different from zero.

Two additional interesting conclusions can be found in our dataset from the model estimation results. First, customers driving in an urban area have a higher probability

Table 3.3. Logistic Regression Model Results for FullCoverage.csv Dataset

Variable	Parameter Estimate	Standard Error	p-Value	Odds-Ratio
(INTERCEPT)	-0.257	0.486	0.5959	0.773
MEN	-0.961	0.086	<0.001	0.382
URBAN	1.173	0.078	<0.001	3.230
PRIVATE	1.065	0.469	0.0232	2.901
MARITAL (MARRIED)	-0.083	0.096	0.3889	0.921
MARITAL (OTHER)	0.161	0.200	0.4212	1.175
AGE	-0.058	0.004	<0.001	0.944
SENIORITY	0.133	0.007	<0.001	1.143
<hr/>				
-2Log-Likelihood	$-2 \times (-2143.905)$			
Likelihood ratio test	875.483 (df = 7, p-value <0.001)			
Pseudo- R^2	16.96%			

Notes: Dependent variable is the choice of FULL COVERAGE ($y = 1$) versus OTHER COVERAGE ($y = 0$).

of buying full coverage compared to those who drive in rural areas. This is because the parameter for URBAN is positive (1.173) and significant (p -value lower than 1%). Second, the SENIORITY in the company has a positive influence. Its parameter is positive (0.133), and the corresponding p -value is below 1%. This means that those customers who have been in the company for a longer period of time have a higher probability of buying full coverage than recent or new customers. Marital status of the policyholder, which separates single, married, and other individuals, does not seem to have a clear effect on the coverage choice. If we look at the significance in the p -value column, we see that the p -value for the coefficient of variables corresponding to marital status equal to MARRIED or OTHER as opposed to SINGLES is large. So, those coefficients are not significantly different from zero. A zero parameter implies that the predictor has no influence on the response. Likewise, the p -value for variable PRIVATE is above 2%. Vehicles for private use are compared to those that have commercial uses, such as those used for transportation or for business. This factor has no significant effect on the choice of coverage if we consider significance at 1%, but it has a significant effect if we fix significance at the 5% level. At this level, the effect is significant and policyholders who own a private use vehicle have a larger probability of buying FULL COVERAGE than insureds with a vehicle for commercial use.

In Table 3.2, we can look at the odds-ratio column and say that when the predictor increases by one unit, the odds that the insured chooses FULL COVERAGE are multiplied by the odds-ratio value. So, when the odds-ratio is larger than one the odds

of FULL COVERAGE versus OTHER COVERAGE increase and they diminish if the odds-ratio is smaller than one.

3.3.2 Parameter Estimation

The maximum likelihood method is the procedure used for parameter estimation and standard error estimation in logistic regression models. This method is based on the fact that responses from the observed units are independent and that the likelihood function can then be obtained from the product of the likelihood of single observations.

In real-life examples there are occasions when the independence between units assumption may not hold. For instance, in periods of economic depression an insurer may perceive a correlated behavior as if there was a contagion between policyholders. As a result of significant economic burdens, policyholders may tend not to buy full coverage in order to reduce their expenses, so all responses to the type of coverage chosen may be somehow connected by exogenous causes. Similarly, risks that are geographically located close to each other are affected by the same meteorological and other external phenomena. So their claiming behavior is strongly correlated.

In this section, we assume that responses are independent. The likelihood of a single observation in a logistic regression model is simply the probability of the event that is observed, so it can be expressed as

$$\Pr(y_i = 1|\mathbf{x}_i)^{y_i} (1 - \Pr(y_i = 1|\mathbf{x}_i))^{(1-y_i)} = \pi_i^{y_i} (1 - \pi_i)^{(1-y_i)}.$$

Note that when the response y_i equals 1, then the previous expression equals $\Pr(y_i = 1|\mathbf{x}_i)$, because the second term is equal to 1 as $(1 - y_i)$ is 0. Conversely, when the response y_i equals 0 then the expression equals $1 - \Pr(y_i = 1|\mathbf{x}_i)$ simply because the first term is equal to one.

The log-likelihood function for a dataset is a function of the vector of unknown parameters β . The observed values of y_i and \mathbf{x}_i are given by the information in the dataset on n individuals. Then, when observations are independent, we can write the log-likelihood function as

$$\begin{aligned} L(\beta) &= \ln \left[\prod_{i=1}^n (\Pr(y_i = 1|\mathbf{x}_i))^{y_i} (1 - \Pr(y_i = 1|\mathbf{x}_i))^{(1-y_i)} \right] \\ &= \sum_{i=1}^n [y_i \ln \Pr(y_i = 1|\mathbf{x}_i) - (1 - y_i) \ln(1 - \Pr(y_i = 1|\mathbf{x}_i))]. \end{aligned}$$

Conventional software can do the job of maximizing the log-likelihood function and providing the parameter estimates and their standard errors. Unless covariates are perfectly correlated the parameter estimates exist and are unique. Moreover, they can be obtained quickly. The algorithm that is used to obtain the maximum of the likelihood function is an iterative method based on a Newton optimization approach.

The method is comparable to the way we would climb a mountain. The likelihood function would play the role of the mountain to be climbed. A walker's steps lead us to the summit. The latitude and longitude of terrestrial coordinates would play the role of the parameter values that are updated at each step until the summit is reached. Similarly, in the likelihood maximization, we modify the values of the parameters to locate the summit of the likelihood function. Starting at convenient values for the parameter vector, this method calculates a new vector of parameter estimates in such a way that the likelihood function is larger than its initial value obtained from the initial parameter values. The procedure is repeated until the likelihood evaluated at new parameter values stabilizes and does not increase when updating the parameter values. Estimation of generalized linear models is further developed in Chapter 5.

3.3.3 Model Fitting

The parameter estimates are denoted by $\hat{\beta}$. Once they are obtained, we can predict the probability of the response modeled event. Given the individual characteristics, the fitted probability is based on the following expression:

$$\hat{\pi}_i = \frac{\exp(\mathbf{x}'_i \hat{\beta})}{1 + \exp(\mathbf{x}'_i \hat{\beta})}.$$

So, every individual in the dataset has an observed response event, y_i , and we can predict the fitted probability of the modeled event, $\hat{\pi}_i$. Let us compare those two values. For instance, an individual in the dataset may have y_i equal to 1, and the estimated probability $\hat{\pi}_i$ may be $0.9 = 90\%$. That is quite plausible. But the response can be $y_i = 1$, while the prediction can be much lower, say 20%. In this case the model prediction would be surprising. In our example, consider a policyholder who has bought the full coverage, but has an estimated probability of buying it of only 20%. We would not have expected this policyholder to buy this coverage, according to the fitted probability. We can say that the odds that he buys full coverage are 1 in 4.

In most datasets when predictions and observations are compared, we can find an unexpected low fitted probability for the event observed responses. The opposite situation can also occur, when the response is the non-event, whereas the probability of the event is large.

In linear models, it is sensible to discuss the correlation between an observation and its fitted value. As pointed out in Chapter 2, squaring this correlation gives the statistic R^2 , the coefficient of determination. However, in nonlinear models, correlations are less useful.

In practice, we will be happy if our logistic regression model is a good predictive tool for the binary response outcome. This means that we will prefer a model in which

the predicted probability that is assigned to a case in the original database is coherent with the observed response. So, when the observed response for individual i is the event, then we would like our model to predict a large $\hat{\pi}_i$ for individual i . Likewise, when the observed response for a particular individual is a non-event, then we want that $\hat{\pi}_i$ for individual i to be small, which means that the probability of a non-event, namely $(1 - \hat{\pi}_i)$, is high.

When looking at $\hat{\pi}_i$, we do not really know what we should consider large or small. In general we fix the *discrimination threshold* at 50%. That corresponds to odds equal to one. So, when the estimated probability is larger than 50%, then we would expect that the response y_i is the event, and similarly, when the estimated probability of the event is smaller than 50%, then we would expect the response to be the non-event.

The discrimination threshold can vary and should not necessarily be equal to 50%. In fact, it is recommended that the threshold be set equal to the proportion of events in the dataset to keep the proportion of predicted event cases in line with the observations. Using the proportion of events in the dataset as the discrimination threshold generally maximizes the overall prediction performance.

The classification table is the simplest way to present the prediction performance of a logistic regression model.

Example 3.1 – (continued): An example of a classification table is presented in Table 3.4. We set the threshold at 35%, which is the proportion of full coverage choice in the sample in the example. Columns represent the predicted responses, and rows are observed behavior. Each cell in the table counts the number of observations from the dataset that are classified within each group after the model and threshold have been used to predict the response behavior. Note that a classification can be done using an out-of-sample new dataset that is not used in the estimation phase. In our example, we reclassified the same observations that were used for estimation.

We see that 2,982 cases are correctly classified, as the model predicts the correct response for 1,858 policyholders who do not have FULL COVERAGE and 1,074 who have FULL COVERAGE. This means that the overall classification rate is equal to $2,982/4,000=74.55\%$, which is excellent. However, a number of cases do not have the expected response. On the one hand, there are 755 cases that do not choose FULL COVERAGE, but the fitted probability is high and the model predicts that they are likely to have made this choice. On the other hand, there are 313 cases that have FULL COVERAGE in the initial database, but the model predicts the opposite response for them.

There are typical traits from the classification table that we would like to see in practice when using a logistic regression model for predictive purposes. The number of cases that are correctly classified has to be high, whereas the number of cases that are incorrectly classified should be low.

Table 3.4. Classification Table for the Logistic Regression Model
in the FullCoverage.csv Dataset

	Predicted $y = 0$	Predicted $y = 1$
Observed $y = 0$	1,858	755
Observed $y = 1$	313	1,074

Notes: Dependent variable is the choice of FULL COVERAGE ($y = 1$) versus OTHER COVERAGE ($y = 0$).

The number of *false-positive* outcomes corresponds to cases where the model predicts a high probability of the response event, but the observed response has been a non-event. False positives are not good in practice. In our example, they mean that the model predicts that the customer is likely to buy a full coverage, but the customer does not buy it. We have 755 such cases in Table 3.4. False positives in this example indicate the failure of commercial efforts. If we use the model to predict the behavior of new customers and try to sell them the FULL COVERAGE with no success, we would incur costs that could be saved.

The number of *false-negative* responses corresponds to the number of cases where the model predicts that the customer will choose a non-event, whereas the true response has been the opposite. In our example 313 customers bought full coverage, although the prediction for doing so is low. If our example model has only a few false-negative cases and it is used for predictive purposes, it means that only a few customers will be predicted not to buy, but decide to do so.

Ideally a perfect model in terms of prediction performance would have no false-positive and no false-negative cases in the classification table. The proportions of false positives or false negatives are calculated column-wise.

Successful logistic regression models usually maximize two additional characteristics: *sensitivity* and *specificity*. *Sensitivity* is the proportion of correctly classified true event responses. *Specificity* is the proportion of correctly classified true non-event responses. In Table 3.4 sensitivity is $1,074/(313 + 1,074) = 77.43\%$, and specificity is $1,858/(1,858 + 755) = 71.11\%$. Both need to be as high as possible for a good prediction performance.

3.3.4 Goodness of Fit

The standard way to evaluate model fit in the binary logistic regression model is the likelihood ratio test for the significance of one or more parameters. The likelihood ratio test requires the identification of two models to be compared, one of which is a special case of the other. In the reduced model, some covariates have been eliminated,

so there are less parameters than in the original model. The number of degrees of freedom for the test statistic is the number of parameters that have been eliminated in the reduced model compared to the initial one. In the null hypotheses, the parameters that have been eliminated are zero. In the alternative hypothesis at least one of those parameters differs from zero. When the sample is large, the likelihood ratio test statistic follows a χ^2 distribution.

The expression for the likelihood ratio test statistic is

$$LRT = 2 \times (L_A - L_R),$$

where L_A is the maximized log-likelihood function in the large, initial model and L_R is the maximized log-likelihood function in the reduced model.

In practice an abuse of goodness-of-fit statistics may lead to selecting a logistic regression model that is difficult to understand. As practitioners, we may be tempted to start from a model with all possible explanatory variables that can be predictors of the dependent variable in the dataset and to eliminate variables step by step, using the likelihood ratio test. This is called a *backward elimination procedure*. Each time a variable or a group of variables is eliminated. The likelihood ratio test indicates that those variables can be deleted from the model if the alternative hypothesis is rejected.

It is highly recommended to be aware of the variables that are being used in the model specification and their corresponding meaning. Backward elimination may lead to selecting variables no matter what they mean or whether they are precise. For instance, an insurer may be confident about the information in the dataset regarding the driving experience of policyholders. A copy of the driving license can always be checked. However, the policyholder's driving area is a predictor that may be kept in the model before dropping the driving experience variable or even replace it. Note that unless a device is installed in the car, it can be difficult to determine what is the true driving area for a customer. One would prefer driving experience rather than driving area as a predictor for models in insurance databases, but automatic backward elimination does not consider the practitioner's preferences for predictors or their information quality.

Pseudo-R squared is a popular goodness-of-fit measure for logistic regression models, but there are many expressions available in the literature. The simplest version can take values between zero and one and is easy to interpret. The larger it is, the better the model fit. The pseudo- R^2 statistic is defined as

$$Pseudo - R^2 = 1 - \left(\frac{L_A}{L_0} \right),$$

where L_A is the maximized log-likelihood function in the initial model and L_0 is the maximized log-likelihood function in the model with only a constant term. Other

Table 3.5. Hosmer-Lemeshow Test for the
FullCoverage.csv Dataset

Group	Observed Number of $y = 1$	Expected
1	22	20.83
2	16	44.45
3	30	67.28
4	31	90.20
5	128	114.98
6	225	140.23
7	219	166.60
8	245	198.92
9	237	241.17
10	234	302.33
χ^2	Df	P(>Chi)
289.4653	8	<0.001

Notes: Dependent variable is the choice of FULL COVERAGE ($y = 1$) versus OTHER COVERAGE ($y = 0$).

commonly used statistics are variations of pseudo- R^2 that are defined between 0 and 1.

The *Hosmer-Lemeshow test* (see Table 3.5) is aimed at evaluating the coherence between the observed responses and the predicted probabilities. The fitted probabilities obtained from the model estimation are ordered from smallest to highest, and individuals are grouped in decile groups. Then 10 groups with the same number of individuals in each group are available.

The Hosmer-Lemeshow test can be calculated as

$$H = \sum_{g=1}^G \frac{(O_g - E_g)^2}{n_g \bar{\pi}_g (1 - \bar{\pi}_g)},$$

where O_g is the number of cases with the event in group g , E_g is the expected number of choices in that group, n_g is the number of cases in group g , and $\bar{\pi}_g$ is the average probability predicted by the model in that group.

There is a graphical tool to assess goodness of fit in logistic regression modeling: the *receiver operating characteristic*, also known as the ROC curve. The ROC curve is a graphical plot of the sensitivity, or true positive rate, versus the false-positive rate (one minus the specificity or true negative rate), for the estimated logistic regression model. Each point in the curve corresponds to a discrimination threshold level. The best model fit would be for a model that has a ROC curve as close as possible to the upper left edge. A nondiscriminant model would have a flat ROC curve, close to the diagonal.

ROC analysis provides a way to select possibly optimal models and to discard suboptimal ones based on classification performance at various threshold levels. For an objective rule to compare ROC curves, the area under the curve, simply referred to as AUROC, can be used. A model with the largest AUROC should be preferred to other possible models.

Example 3.1 – (continued): To compute the Hosmer-Lemeshow test in our sample of 4,000 individuals, we can define groups of 400 individuals. The first group is formed by the individuals in the dataset who have the lowest estimated probabilities of the modeled response, as predicted by the model. The second group is formed by the next bunch of individuals whose estimated probabilities are low, but not as low as the first group, and so forth. Inside each group the number of event responses is calculated. So, for example in our case study, we would expect that in the first group, where the estimated probabilities of buying full coverage are the lowest in the dataset, the proportion of individuals who bought full coverage in this group is indeed low. Likewise, in the highest group we would expect the contrary – that most customers would buy full coverage – so there the observed proportion would be high. The Hosmer-Lemeshow test evaluates how close the proportions of observed event responses within each group are compared to the expected probability for the modeled event in that group. A model fits perfectly if the observed proportions and the predicted proportions are equal. The test establishes a distance between observed and expected proportions. The closer they are, the better is the model fit.

The results for the Hosmer-Lemeshow test are presented in Table 3.5. The test indicates that the model has a poor fit because the p -value is too low. In large samples, this test frequently rejects the null hypothesis.

An example is shown in Figure 3.2 of the ROC curve. In this plot, the curve is distant from the diagonal, which is a good sign. The area under the curve (AUROC) equals 78.78%, which also indicates a good fit.

3.4 Probit and Other Binary Regression Models

The logistic regression model is closely related to other models such as the probit regression. The only difference between logit and probit regression is the way in which the relationship between π_i as a nonlinear function of $\mathbf{x}'_i\boldsymbol{\beta}$ is established. The distribution for y_i is Bernoulli, because y_i is a binary variable and there is no other possible choice. The probit regression model specifies that π_i equals the distribution function of a standard normal random variable at $\mathbf{x}'_i\boldsymbol{\beta}$; in other words, it is the probability that a standard normal random variable is lower than $\mathbf{x}'_i\boldsymbol{\beta}$. This specification is quite natural if we assume that there is an unobserved random utility that drives the decision of each individual choice. That random utility can be specified

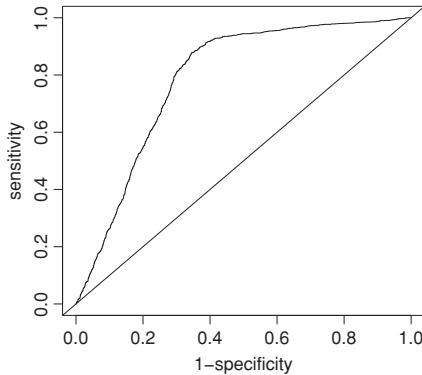


Fig. 3.2. The ROC curve for the logistic regression model in the `FullCoverage.csv` dataset. Dependent variable is the choice of FULL COVERAGE ($y = 1$) versus OTHER COVERAGE ($y = 0$).

as following a normal distribution and be set equal to $\mathbf{x}_i'\boldsymbol{\beta}$ plus an error term. In that case the probit model follows directly as a result of assuming that an individual makes the decision to chose the event response when the random unobserved utility exceeds a threshold that depends on \mathbf{x}_i . So, the probit model arises naturally from a linear regression model for the underlying unobserved utility.

Other binary regression models are possible, such as the complementary log-log model. In terms of prediction performance, these other models do not seem to be more flexible or easily interpretable as the logistic regression model.

3.5 Models for Ordinal Categorical Dependent Variables

This section describes the cumulative logit regression model aimed at ordinal dependent variables. It focuses on model specification and examples.

In ordinal categorical dependent variable models the responses have a natural ordering. These models are quite common in insurance; an example is to model possible claiming outcomes as ordered categorical responses. For instance, we can have policyholders who had no claims during one period of time, policyholders who had one claim in the same period, and others who had two or more claims.

Let us assume that an ordinal categorical variable has J possible choices. The most straightforward model in this case is the *cumulative logit model*, also known as *ordered logit*. Let us denote by y_i the choice of individual i for a categorical ordered response variable. Let us assume that π_{ij} is the probability that i chooses j , $j = 1, \dots, J$. So, $\pi_{i1} + \pi_{i2} + \dots + \pi_{iJ} = 1$. Response probabilities depend on the individual predictors; again, we assume they depend on $\mathbf{x}_i'\boldsymbol{\beta}$. It is important to bear in mind that the ordered logit model concentrates on the cumulative probabilities

Table 3.6. *Ordered Logistic Regression Model Results for TypeofCoverage.csv Dataset*

Variable	Parameter Estimate	Standard Error	p-Value	Odds-Ratio
INTERCEPT (COMPREHENSIVE)	-1.754	0.578	0.002	0.173
INTERCEPT (FULL COVERAGE)	-3.094	0.580	<0.001	0.045
MEN	0.067	0.097	0.492	1.069
URBAN	0.151	0.086	0.079	1.163
PRIVATE	1.478	0.562	0.009	4.380
MARITAL (MARRIED)	-0.083	0.098	0.213	0.885
MARITAL (OTHER)	0.161	0.705	0.207	2.433
AGE	-0.001	0.004	0.917	1.000
SENIORITY	0.018	0.007	0.008	1.018
–2Log-Likelihood	4330.819			
Likelihood ratio test	25.156 (df = 7, p-value = 0.001)			

Notes: Dependent variable is the coverage choice: full coverage, comprehensive, or third-party liability (baseline).

$\Pr(y_i \leq j | \mathbf{x}_i)$. Then, for $j = 1, \dots, J - 1$

$$\text{logit}(\Pr(y_i \leq j | \mathbf{x}_i)) = \alpha_j + \mathbf{x}_i' \boldsymbol{\beta}.$$

Note that

$$\text{logit}(\Pr(y_i \leq j | \mathbf{x}_i)) = \ln \left(\frac{\Pr(y_i \leq j | \mathbf{x}_i)}{1 - \Pr(y_i \leq j | \mathbf{x}_i)} \right).$$

Parameters are interpreted similar to the logistic regression model case. A positive parameter means that if the predictor increases then the cumulated probability increases. There are as many constant terms as possible choices minus one, because $\Pr(y_i \leq J | \mathbf{x}_i) = 1$. Those constant term parameters (α_j) are increasing with j because the cumulative probabilities also increase when more choices are accumulated, but they can be negative.

Example 3.2 (TypeofCoverage.csv data file). Table 3.6 presents the results of a sample dataset in which are 2,148 policy holders who are offered three ordered types of coverage: THIRD PARTY LIABILITY, COMPREHENSIVE, and FULL COVERAGE. Third-party liability is the minimum compulsory insurance in the market for those customers. Comprehensive insurance includes additional coverage for stolen or vandalized vehicles. Full coverage also includes damage to the insured vehicle. We interpret the parameters and the odds-ratio as before. Here we see that the effect of predictors is not as neat as in the logistic regression model presented in Example 3.1.

Age does not influence the propensity to increase insurance coverage, nor does sex or marital status; p -values are too large. Private use vehicles do seem to be associated with higher levels of coverage. Negative signs in the intercept coefficients mean a resistance to increase coverage, which is certainly explained by the associated price increase, but price is not included in this model. Only larger seniority in the company and private as opposed to commercial vehicles indicate a tendency to select extended (cumulated) coverages.

3.6 Models for Nominal Categorical Dependent Variables

The last section is aimed at specifying the multinomial logit and general categorical response models, with an emphasis on parameter interpretation.

In nominal categorical dependent variable models, the responses correspond to choices that are identified with labels and do not respond to a natural ordering. A classical example is the color chosen when buying a car. Another one is the type of policy chosen by policyholders in an insurance company. An insurer may offer several possibilities to the customer, and although we could say that the price of each possible contract may induce an ordering in the choice, we can ignore price, and study the choice of products as purely nominal alternatives. A marketing firm could also study customer behavior by offering the customer a choice of possible insurers to model the preference for a particular insurance company over some others.

The multinomial logit model (Section 3.6.1) and the conditional logit model (Section 3.6.2) deal with data where the dependent variable contains more than just two alternatives. The former handles only explanatory variables that are *individual specific*, whereas the latter handles explanatory variables that are *individual and alternative specific*. Many applications of both types exist in the insurance and risk management context. For example, if we have three possible insurance policy coverages to choose from, a model that would just consider the consumer's characteristics would only include individual-specific explanatory variables, whereas a model that would consider, in addition, the differences in quality or price between each coverage would include individual- and alternative-specific characteristics.

Let us assume that a nominal categorical variable has J possible choices. We recommend using the models in this section with a moderate number of possible choices. The larger the number of options, the more complex is parameter interpretation.

3.6.1 Specification of the Generalized Logit Model

Let us start with the *generalized logit model*. This model is often called the *multinomial logit model*, which we present in the next section and which is a bit more general. However, the generalized logit model is so widely used that it is often called the multinomial logit model. The generalized multinomial logit model is used to predict

the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables that measure individual risk factors.

Let us denote by y_i the choice of individual i for a nominal categorical response variable. Let us assume that π_{ij} is the probability that i chooses j , $j = 1, \dots, J$ and $i = 1, \dots, n$. So, $\pi_{i1} + \pi_{i2} + \dots + \pi_{iJ} = 1$. The probabilities depend on the individual predictors, and again, we assume these choice probabilities depend on $\mathbf{x}'_i \boldsymbol{\beta}$.

We assume that the J -th alternative is the *baseline* choice. Then, the generalized logit regression model is specified as

$$\Pr(y_i = j | \mathbf{x}_i) = \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta}_j)}{1 + \sum_{k=1}^{J-1} \exp(\mathbf{x}'_i \boldsymbol{\beta}_k)}, \quad j = 1, \dots, J-1$$

$$\Pr(y_i = J | \mathbf{x}_i) = \frac{1}{1 + \sum_{k=1}^{J-1} \exp(\mathbf{x}'_i \boldsymbol{\beta}_k)}.$$

So there are $J - 1$ vectors of parameters to be estimated, namely $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_{J-1}$. We set vector $\boldsymbol{\beta}_J$ to zero for identification purposes.

In the generalized logit regression model, the linear predictor $(\mathbf{x}'_i \boldsymbol{\beta}_j)$ equals the log-odds of the corresponding response (j) to the baseline response (J), i.e.

$$\mathbf{x}'_i \boldsymbol{\beta}_j = \ln \left(\frac{\pi_{ij}}{\pi_{iJ}} \right).$$

Parameters can easily be interpreted. A positive parameter in vector $\boldsymbol{\beta}_j$ increases the odds of choosing j with respect to the baseline (J th choice), when the predictor increases. Conversely, a negative parameter diminishes the odds.

The generalized logit model relies on the assumption of independence of irrelevant alternatives (IIA), which is not always desirable. This assumption states that the odds of preferring one class over another do not depend on the presence or absence of other alternatives. Let us give an example. Assume a consumer is asked about two brands for a soda drink, and she says that she prefers brand A twice as much as brand B. So the odds are 2 to 1. Then a marketing expert offers two cans, one for brand A and one for brand B. The probabilities of choice are $2/3$ versus $1/3$, respectively. But what happens if a third can from brand A is added to the set of choices? Numerous studies in experimental economics show that individuals assign equal probabilities to the three cans. This means $1/3$ to each can, but then the odds for the initial two cans would change from 2:1 to 1:1. This means that the third can was not irrelevant, and its presence has violated the hypothesis of independence of irrelevant alternatives. The presence of the third can has changed the odds between the initial two choices.

In practice, the preferences between any two choices in the generalized logit model must be independent of the other possible choices.

Table 3.7. Generalized Logistic Regression Model Results for
VehOwned.csv Dataset

Variable	Parameter Estimate	Standard Error	p-Value
Choice of FOUR WHEEL DRIVE			
Intercept	-2.774	0.412	<0.001
MEN	0.468	0.272	0.086
URBAN	-0.804	0.267	0.003
AGE	-0.006	0.008	0.470
Choice of MOTORBIKE			
Intercept	-1.380	0.272	<0.001
MEN	0.644	0.178	<0.001
URBAN	0.071	0.145	0.625
AGE	-0.024	0.005	<0.001
-2 Log-Likelihood	$-2 \times (-1104.4)$		
Likelihood ratio test	44.051(df = 6, p-value < 0.001)		

Notes: Dependent variable is type of vehicle chosen: MOTORBIKE, FOUR WHEEL DRIVE, and CAR (baseline).

Example 3.3 (VehOwned.csv data file). An example of a generalized logit model estimation is shown in Table 3.7. This sample corresponds to 2,067 customers of an insurance firm. We study the type of vehicle that they choose when given three unordered options: MOTORBIKE, CAR, or FOUR WHEEL DRIVE. The baseline is CAR. We model the choice of MOTORBIKE versus CAR and the choice of FOUR WHEEL DRIVE versus CAR. Predictors are SEX, DRIVING AREA, and AGE. We can see from the results that MEN have more preference for motorbikes over cars compared to women. The parameter value is positive 0.644 and it is significant (i.e., small *p*-value). We can see that the preference for motorbikes versus cars diminishes with age, because the parameter is significant and negative (-0.024). There is no clear effect of the driving area on the preferences for motorbikes and cars, because the *p*-value for the coefficient of variable urban is too large. So, it seems that motorbikes are as appreciated in rural as they are in urban areas.

When we compare preference for FOUR WHEEL DRIVE vehicles over CAR vehicles, we see that the parameter for the driving area is negative and significant, which means that FOUR WHEEL DRIVE vehicles are much less preferred than cars in urban areas rather than in rural areas. Age and sex have no significant effect on this preference.

3.6.2 Multinomial Logistic Regression Model

In the *multinomial logistic regression model* individual characteristics can be different for different choices. For instance, when choosing a transportation mode to commute,

each choice is associated with a time and cost. This model is also known as the *conditional logit model* because individual characteristics depend on the chosen alternative. For instance, when underwriting a policy, the price depends on the policyholder's risk factors and on the type of policy or brand that is chosen.

The multinomial logistic regression model specification is

$$\Pr(y_1 = j | \mathbf{x}_{ij}) = \frac{\exp(\mathbf{x}'_{ij} \boldsymbol{\beta})}{\sum_{k=1}^J \exp(\mathbf{x}'_{ik} \boldsymbol{\beta})}, \quad j = 1, \dots, J. \quad (3.3)$$

There is only one vector of unknown parameters $\boldsymbol{\beta}$, but we have J vectors of known characteristics $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iJ}$. For instance, if we offer three insurance policies to a customer, we would have three different prices, even if the customer finally chooses one of them. We can then use that price information in the model as a factor that changes with choice. The regressors can also contain individual characteristics that do not change with the choice, such as age or driving experience.

The generalized logit model can be understood as a special case of the multinomial logistic model. If we define $\mathbf{x}_{ij} = (\mathbf{0}', \dots, \mathbf{0}', \mathbf{x}'_i, \mathbf{0}', \dots, \mathbf{0}')$, then it is a vector that contains zeros, Except for the components corresponding to the j -th alternative. Then we define $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_J)$. Because of this special case definition, many authors refer to the multinomial logit model when they mean to refer to the generalized logit model.

Example 3.4 (VehChoicePrice.csv data file). This example corresponds to a similar situation to Example 3.3. Customers report the type of vehicle that they choose; however additional information is now available on the price of each possible alternative. Price is an alternative specific explanatory variable.

Estimation results for a multinomial logistic regression model are shown in Table 3.8. Not surprisingly, the variable PRICE has a positive (0.056) and significant effect, which means that customers tend to choose the most expensive alternative. Indeed, given the nature of the alternatives, higher price vehicles correspond to those that have more capabilities.

Noticeably, those driving in urban areas tend to find the choice of a four-wheel drive vehicle less attractive than a regular car, compared to those living in rural areas (parameter equal to -0.767 and significant). Conclusions similar to those reported for the generalized logit model, where the alternative specific price was not included in the model, are also found here. This model found a larger preference for motorbikes in men compared to women and in younger compared to older respondents.

3.6.3 Advanced Categorical Response Models

The *nested logit model* is a hierarchical model where decision, are taken as in a tree structure setting. First, we model one decision, and then, conditional on the choice

Table 3.8. *Multinomial Logistic Regression Model Results for VehChoicePrice.csv Dataset*

Variable	Parameter Estimate	Standard Error	p-Value
Intercept (FOUR WHEEL DRIVE)	-0.411	0.288	0.153
Intercept (MOTORBIKE)	-3.493	0.434	<0.001
PRICE	0.056	0.006	<0.001
MEN (FOUR WHEEL DRIVE)	0.331	0.276	0.229
MEN (MOTORBIKE)	0.713	0.180	<0.001
URBAN (FOUR WHEEL DRIVE)	-0.767	0.270	0.004
URBAN (MOTORBIKE)	0.082	0.146	0.574
AGE (FOUR WHEEL DRIVE)	-0.001	0.008	0.884
AGE (MOTORBIKE)	-0.033	0.006	<0.001
-2Log-Likelihood		$-2 \times (-1055.8)$	
Likelihood ratio test		141.110 (df = 7, p-value <0.001)	

Notes: Dependent variable is type of vehicle chosen: ‘Motorbike’, ‘Four wheel drive’ and ‘Car’ (baseline).

made for the first step, a second choice is possible. This is a flexible setting because the number and type of possible choices do not have to be the same in each branch. Moreover, the IIA hypothesis is eliminated in this model.

It is often the case that we have repeated observations for the same individuals. The mixed logit model can take this panel dimension into account. Panel data models involving categorical dependent variables are aimed at modeling a sequence of categorical responses. They can be presented as nested logit models. Mixed logit models assume that parameter vectors are subject specific. This means that the parameters change from one individual to another. Estimation for this type of models is done using simulation methods.

3.7 Further Reading

The conditional logit model was introduced by McFadden (1974) in the context of econometrics. Daniel McFadden and James Heckman were awarded the Nobel Prize in 2000 for their contributions to transportation economics. Hosmer and Lemeshow (2000) wrote a book on applied logistic regression in which they present their test. Long (1997) is a great resource for categorical and limited dependent variables. Hilbe (2009) is full of guided examples in *Stata*, and Frees (2010) presents several actuarial and financial applications of categorical dependent variable models and includes an excellent summary of the theoretical background. Additional reference material can be found in Greene (2011), Cameron and Trivedi (2005), and Artis, Ayuso and Guillén (1999).

References

- Artis, M., Ayuso, M., and Guillén, M. (1999). *Modelling Different Types of Automobile Insurance Fraud Behaviour in the Spanish Market*, 24(1), 67–81.
- Cameron, A. C. and P. K. Trivedi (2005). *Microeometrics: Methods and Applications*. Cambridge University Press, New York.
- Frees, E. W. (2010). *Regression Modeling with Actuarial and Financial Applications*. Cambridge University Press, New York.
- Greene, W. H. (2011). *Econometric Analysis* (7th ed.). Prentice Hall, New York.
- Hilbe, J. M. (2009). *Logistic Regression Models*. CRC Press, Chapman & Hall. Boca Raton, FL.
- Hosmer, D. W. and S. Lemeshow (2000). *Applied Logistic Regression* (2nd ed.). John Wiley & Sons, New York.
- Long, J. S. (1997). *Regression Models of Categorical and Limited Dependent Variables*. Sage, Thousand Oaks, CA.
- McFadden, D. (1974). *The Measurement of Urban Travel Demand*. *Journal of Public Economics*, 3(4), 303–328.

4

Regression with Count-Dependent Variables

Jean-Philippe Boucher

Chapter Preview. This chapter presents regression models where the random variable is a count and compares different risk classification models for the annual number of claims reported to the insurer. Count regression analysis allows identification of risk factors and prediction of the expected frequency given characteristics of the risk. This chapter details some of the most popular models for the annual number of claims reported to the insurer, the way the actuary should use these models for inference, and how the models should be compared.

4.1 Introduction

In the early 20th century, before the theoretical advances in statistical sciences, a method called *the minimum bias technique* was used to find the premiums that should be offered to insureds with different risk characteristics. This technique's aim was to find the parameters of the premiums that minimize their bias by using iterative algorithms.

Instead of relying on these techniques that lack theoretical support, the actuarial community now bases its methods on probability and statistical theories. Using specific probability distributions for the count and the costs of claims, the premium is typically calculated by obtaining the conditional expectation of the number of claims given the risk characteristics and combining it with the expected claim amount. In this chapter, we focus on the number of claims.

4.2 Poisson Distribution

The starting point for modeling the number of claims, a random variable noted Y , is the Poisson distribution. Introduced by Siméon-Denis Poisson (1781–1840) in his 1837 work titled *Recherches sur la probabilité des jugements en matière criminelle et matière civile*, the Poisson distribution is indeed the basis of almost all analysis of

count data. The probability function of the Poisson distribution can be expressed as

$$\Pr(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!},$$

where λ is the mean parameter. The Poisson distribution has an equidispersion property, meaning that the mean $E(Y)$ of the distribution is equal to its variance $\text{Var}(Y)$. Because the Poisson distribution is a member of the exponential family, it has some useful statistical properties (see Chapter 5 for details).

As we see in this chapter, this distribution can be used to model the number of claims in nonlife insurance. However, it can also be used in other popular areas in actuarial sciences. For example, count regression can be used with run-off triangles (see Chapter 18), to model mortality (see Chapter 19) and natality, or to estimate the severity scores in third-party liability.

4.2.1 Law of Small Numbers

There are many ways to construct the Poisson distribution, some of which lead to important insurance interpretations. It can be shown that the Poisson distribution is the limit of a binomial distribution with a success probability going to zero, and a number of trials going to infinity.

Indeed, expressing $Y_{n,\pi}$ as the total number of success with n independent tries, with π representing the success probability, we have the following probability function:

$$\Pr[Y_{n,\pi} = k] = \binom{n}{k} \pi^k (1 - \pi)^{n-k}, \quad k = 0, 1, 2, \dots$$

When $n \rightarrow \infty$, $\pi \rightarrow 0$, and $n\pi = \lambda > 0$, meaning that the mean λ is fixed when $n \rightarrow \infty$, $\pi \rightarrow 0$, we have

$$\lim_{n \rightarrow \infty} \pi \rightarrow 0 \left[\binom{n}{k} \left(\frac{\lambda}{n} \right)^k \left(1 - \frac{\lambda}{n} \right)^{n-k} \right] = \frac{\lambda^k e^{-\lambda}}{k!},$$

which is the probability function of a Poisson distribution. Thus, it is natural to base the claim count analysis on such distributions when modeling the accident process by the Poisson distribution, because its interpretation is direct, with n measuring the use of the car (in kilometers, for example) and p the probability of having an accident for each use of the car. So, even if the probability of having an accident on a specific day is extremely small, because the car is used a large number of times during the year, the probability of having an accident becomes less marginal.

Because the Poisson distribution is the limit of a binomial when the probability of success is small compared with the number of tries, the Poisson distribution is often called the *law of small numbers*.

4.2.2 Exponential Waiting Times

Another interesting property of the Poisson distribution refers to the time between two claims. If we suppose that τ_i is the time between the $i - 1^{th}$ and the i^{th} claim, we can define the time arrival of the j^{th} claim as

$$\nu(j) = \sum_{i=1}^j \tau_i.$$

Consequently, we can link the time arrival $\nu(j)$ of the j^{th} claim to a claim count process $Y(t)$:

$$\nu(j) \leq t \Leftrightarrow Y(t) \geq j.$$

Consequently, when the time occurrence of the j^{th} claim is less than t , the number of claims during the time period t is greater or equal to j . Formally, we have

$$\begin{aligned} \Pr(Y(t) = j) &= \Pr(Y(t) < j + 1) - \Pr(Y(t) < j) \\ &= \Pr(\nu(j + 1) > t) - \Pr(\nu(j) > t) \\ &= F_j(t) - F_{j+1}(t), \end{aligned}$$

where $F_j(t)$ is the cumulative function of $\nu(j)$.

If we suppose that the waiting time between two claims is exponentially distributed with mean $1/\lambda$, which indicates that $\nu(j)$ is gamma distributed, we can show that

$$\Pr(Y(t) = y) = \frac{e^{-\lambda t} (\lambda t)^y}{y!},$$

which is the probability function of a Poisson distribution of mean λt . Note that because the hazard function does not depend on t (memoryless property of the exponential distribution), the Poisson does not imply duration dependence, meaning that a claim does not modify the expected waiting time to the next claim. Another important property of the Poisson distribution that follows the waiting time interpretation is that the mean parameter of the model is proportional to the observed time length. Normally, t is considered as the number of years of coverage. For example, insureds covered for 6 months will have a premium half as high as if they were insured for 1 year because the mean parameter of the Poisson distribution will be 0.5λ , compared to λ .

4.3 Poisson Regression

The characteristics of insureds that should influence their premiums, such as age, sex, or marital status, are included as regressors in the parameter of the count distribution.

For classic regression models (see Chapter 2), the exogenous information can be coded with binary variables. For example, we can model the sex with a variable x that takes the value 1 if the insured is a man and 0 otherwise.

In statistical modeling, we use the a link function $h(\mathbf{x}'\boldsymbol{\beta})$, where $\boldsymbol{\beta}' = (\beta_0, \beta_1, \dots, \beta_k)$ is a vector of regression parameters for the binary explanatory variables $\mathbf{x}'_i = (x_{i,0}, x_{i,1}, \dots, x_{i,k})$. Usually in insurance, the link function $h()$ is an exponential function. Consequently, we have $\lambda_i = t_i \exp(\mathbf{x}'_i \boldsymbol{\beta}) = \exp(\mathbf{x}'_i \boldsymbol{\beta} + \ln(t_i))$ where t_i represents the risk exposure of insured i , as explained in the previous section.

There is a big advantage of using an exponential function for the mean parameter. It allows the insurer to construct a premium based on multiplicative relativities:

$$\begin{aligned}\lambda_i &= t_i \exp(\beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k}) \\ &= t_i \exp(\beta_0) \exp(\beta_1 x_{i,1}) \dots \exp(\beta_k x_{i,k}) \\ &= t_i p_0 \times r_{i,1} \times \dots \times r_{i,k},\end{aligned}$$

where p_0 can be viewed as the base premium, and $r_{i,j}$, $j = 1, \dots, k$ the relativities applied to insureds having the property j (i.e $x_{i,j} = 1$).

4.3.1 Maximum Likelihood Estimator

With covariates in the mean parameter of the Poisson, the maximum likelihood estimator (MLE) of the parameters of the Poisson distribution can be obtained with the log-likelihood function:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^n \left[y_i \ln(\exp(\mathbf{x}'_i \boldsymbol{\beta})) - \exp(\mathbf{x}'_i \boldsymbol{\beta}) - \ln(y_i !) \right],$$

where n is the total number of observations and y_i the number of claims of insured i . We then obtain the first-order condition:

$$\frac{\delta l(\boldsymbol{\beta})}{\delta \boldsymbol{\beta}} = \sum_{i=1}^n (y_i - \lambda_i) \mathbf{x}_i = \mathbf{0}. \quad (4.1)$$

The Newton-Raphson algorithm can be used to obtain the $\hat{\boldsymbol{\beta}}$ using the equation (4.1). However, classic statistical software such as R or SAS can be used with pre-programmed routines. The properties of the MLE are well known and allow us to compute the variance of the estimators. The Hessian and the outer-product estimates

of the variance are

$$\text{Var}_{Hess}[\hat{\beta}] = \left[\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i \lambda_i \right]^{-1}.$$

$$\text{Var}_{OP}[\hat{\beta}] = \left[\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i (y_i - \lambda_i)^2 \right]^{-1}.$$

It is well known that the Poisson distribution has some severe drawbacks, such as its equidispersion property, that limit its use. Because the estimated parameters of a Poisson distribution are consistent even if the *real* distribution is not Poisson, it is very advantageous to use these estimates. So, we can still use $\hat{\beta}$ obtained by the MLE of a Poisson distribution and *correct* the overdispersion of the Poisson distribution by adding a multiplicative factor ϕ to the variance of the Poisson to obtain $\text{Var}_{Over.}[Y_i] = \phi \lambda_i$. We can estimate the ϕ parameter by several techniques including

$$\hat{\phi} = \frac{\sum_{i=1}^n (y_i - \hat{\lambda}_i)^2}{\sum_{i=1}^n \hat{\lambda}_i},$$

where the estimator ϕ is computed using $\hat{\lambda}_i$, which corresponds to the estimate of the mean parameter λ_i . In this case, it can be shown that

$$\text{Var}_{Over.}[\hat{\beta}] = \phi \text{Var}_{Hess}[\hat{\beta}].$$

Based on equation (4.1), other estimation techniques can be used such as the estimating equation approach, in which the link between the mean and the variance is broken.

Parameters' statistical significance can be tested using classic Wald or likelihood ratio tests. Deviance, as explained in Chapter 5, can also be used to verify the fit of the Poisson because this distribution is a member of the linear exponential family.

4.3.2 Empirical Illustration

We use the Singapore Automobile Claims Database [singapore.csv] to apply the models. The following two covariates have been used in the application:

- (1) The No Claim Discount (NCDClass). Based on the accident record, we used three categories: 0, 10–30, and 40–50. A high discount means a better accident record than a lower one.
- (2) The age (in years) of the vehicle (Vagecat1). We used five categories: 0–2, 3–5, 6–10, 11–15, and >15.

Table 4.1. MLE for the Poisson Distribution

Covariates	Parameter	Estimate	Std. Err.		
			Hess	OP	Over.
Intercept	β_0	-3.7748	0.5118	0.5151	0.5233
NCDClass	β_1	0.7218	0.1246	0.1227	0.1274
	(10, 20, 30)	0.3584	0.1264	0.1240	0.1292
	(0–2)	1.5896	0.5035	0.5106	0.5149
	(3–5)	1.5329	0.5148	0.5222	0.5265
	(6–10)	1.1051	0.5179	0.5244	0.5296
	(11–15)	0.3489	0.5304	0.5380	0.5424
Log-likelihood					$\hat{\phi} = 1.0475$
					-1803.841

Results of the application of the Poisson distribution are in Table 4.1. R programs that computed these results are available on the book's website.

Large differences between the Hessian and the outer-product variances indicate a misspecification of the model, meaning that the Poisson distribution is not indicated for the data. In our case, the variances of the parameter estimates of β are quite close. However, obtaining a sound theoretical answer to requires construction of a more formal test. The inclusion of a dispersion parameter ϕ increases the Hessian variance by a factor of 1.0475.

4.4 Heterogeneity in the Distribution

Instead of adding a dispersion parameter ϕ on the variance of the estimators to generalize the Poisson distribution, we can construct other distributions that allow for overdispersion. Indeed, we can suppose that the overdispersion of the insurance data is caused by the omission of some important classification variables (swiftness of reflexes, aggressiveness behind the wheel, consumption of drugs, and so forth). Consequently, by supposing that the insurance portfolio is heterogeneous, we can generalize the Poisson distribution by adding a random heterogeneity term:

$$\Pr[Y = y] = \int_0^\infty \Pr[Y = y|\theta]g(\theta)d\theta,$$

where $\Pr[Y = y|\theta]$ is the conditional distribution of Y and $g(\theta)$ is the density of Θ . The introduction of an heterogeneity term means that the mean parameter is also a random variable.

4.4.1 Negative Binomial

When the random variable Θ follows a gamma distribution of mean 1 (to ensure that the heterogeneity mean is equal to 1, both parameters of the gamma distribution are chosen to be identical and equal to $a = 1/\alpha$), such as

$$\Pr[Y = y|\theta] = \frac{(\lambda\theta)^y e^{-\lambda\theta}}{y!},$$

$$f(\theta) = \frac{\theta^a}{\Gamma(a)} \theta^{a-1} \exp(-\theta a),$$

the mixed model yields a negative binomial distribution:

$$\Pr[Y = y] = \frac{\Gamma(y+a)}{\Gamma(y+1)\Gamma(a)} \left(\frac{\lambda}{a+\lambda}\right)^y \left(\frac{a}{a+\lambda}\right)^a.$$

From this, it can be proved that $E[Y] = \lambda$. Because $\text{Var}[Y] = \lambda + \frac{\lambda^2}{a} = \lambda + \alpha\lambda^2$, with a λ^2 , this negative binomial is often called a negative binomial 2 (NB2).

We can estimate the parameters of the distribution by maximum likelihood. For the β parameters, the NB2 model leads to a first-order condition of

$$\sum_{i=1}^n x_i \left(\frac{y_i - \lambda_i}{1 + \lambda_i/a} \right) = 0.$$

For the a parameter, it can be shown that the first-order condition can be expressed as

$$\sum_{i=1}^n \left(\left[\sum_{j=0}^{y_i-1} \frac{1}{j+a} \right] - \left[\ln \left(1 + \frac{\lambda_i}{a} \right) - \frac{y_i - \lambda_i}{a + \lambda_i} \right] \right) = 0.$$

Additionally, we can show that the β and the a estimators are independent, meaning that

$$\text{Cov}_{MLE}[\hat{\beta}, \hat{a}] = 0.$$

By changing the parameter of the heterogeneity distribution, it is possible to obtain other forms of the negative binomial distribution. Indeed, using

$$f(\theta) = \frac{\tau^\tau}{\Gamma(\tau)} \theta^{\tau-1} \exp(-\theta\tau),$$

with $\tau = a\lambda^{(2-p)}$, this general form of the negative distribution has the same mean λ , but because it has a variance of the form $\lambda + \alpha\lambda^p$, it is called a *negative distribution p* (NBp).

For $p \neq 2$, note that the variance of Θ now depends on the individual characteristics of the policyholders. When $p = 2$, the NB2 distribution coincides with the negative

binomial 2 distribution seen earlier. When p is set to 1, we get the classic *NB1* distribution that has the following probability function:

$$\Pr(Y = y) = \frac{\Gamma(y + a\lambda)}{\Gamma(y + 1)\Gamma(a\lambda)}(1 + 1/a)^{-a\lambda}(1 + a)^{-y},$$

with $a = 1/\alpha$. The *NB1* model is interesting because the variance $\text{Var}[Y] = \lambda + \alpha\lambda = \phi\lambda$ is the one used previously to correct the overdispersion of the Poisson (noted $\text{Var}_{\text{Over.}}[Y]$).

We can even treat the variable p as an unknown parameter to be estimated. This can be seen as an hyper-model that can allow use of a test to distinguish between the *NB2* and the *NB1* distributions. For example, if the confidence interval of the variable p includes the value 2 but not the value 1, it means that the distribution *NB2* should be preferred to the *NB1* distribution.

Finally, note that the *NB2* distribution is clearly different from the *NB1* or, more generally, from distributions *NB p*, with $p \neq 2$, in the following three ways:

- (1) The *NB2* supposes an independence between β and $a(\alpha)$ for the estimation by maximum likelihood (the variance-covariance matrix is 0 outside the diagonal), which is not the case for other forms of the *NB* distributions.
- (2) For a parameter $a(\alpha)$ that is known, the *NB2* is a member of the exponential linear family; consequently the MLE of β is consistent if the true distribution is not *NB2*.
- (3) An *NB2* distribution with $a = 1$ ($\alpha = 1$) corresponds to a classic geometric distribution.

4.4.2 Poisson-Inverse Gaussian Heterogeneity

There are other choices of distribution for the heterogeneity parameter Θ . For example, if Θ follows an inverse Gaussian distribution with unit mean and variance α , we obtain the Poisson-inverse Gaussian (PIG) distribution. The probability distribution of the *PIG2* distribution is

$$\Pr(Y = y) = \frac{\lambda^y}{y!} \left(\frac{2}{\pi\alpha} \right)^{0.5} e^{1/\alpha} (1 + 2\alpha\lambda)^{-s/2} K_s(z),$$

where $s = y - 0.5$ and $z = (1 + 2\alpha\lambda)^{0.5}/\alpha$. The function $K_j(.)$ is the modified Bessel function of the second kind that satisfies the following recursive properties:

$$\begin{aligned} K_{-1/2}(u) &= \left(\frac{\pi}{2u} \right)^{0.5} e^{-u} \\ K_{1/2}(u) &= K_{-1/2}(u) \\ K_{s+1}(u) &= K_{s-1}(u) + \frac{2s}{u} K_s(u). \end{aligned}$$

As for the NB distributions, the PIG distributions can be extended to cover many forms of variance. If Θ has variance $\alpha\lambda^{2-p}$, we get the PIG_{*p*} distribution with probability mass function:

$$\Pr(Y = y) = \frac{\lambda^y}{y!} \left(\frac{2}{\pi\alpha\lambda^{2-p}} \right)^{0.5} e^{1/\alpha\lambda^{2-p}} (1 + 2\alpha\lambda^{3-p})^{-s/2} K_s(z), \quad (4.2)$$

where $s = y - 0.5$ and $z = (1 + 2\alpha\lambda^{2-p})^{0.5}/\alpha\lambda^{k-1}$. In this case, $\text{Var}[Y] = \lambda + \alpha\lambda^{2-p}$. Similar to the NB distribution, the parameters of the model can be estimated by maximum likelihood, and the *p* parameter can also be estimated to use as a tool to choose between the PIG2 and the PIG1 distributions.

4.4.3 Poisson-Lognormal Heterogeneity

Another possibility is to suppose that Θ is a lognormal distribution with parameters $\mu = -\alpha/2$ and $\sigma^2 = \alpha$. In this case, we get the Poisson-lognormal (PLN) distribution. For insured *i*, the probability distribution of the PLN distribution is:

$$\Pr(Y_i = y_i) = \int_{-\infty}^{\infty} \frac{\exp(-\gamma_i)\gamma_i^{y_i}}{y_i!} \frac{1}{\sqrt{2\pi\alpha}} \exp\left(-\frac{\epsilon_i^2}{2\alpha}\right) d\epsilon_i,$$

where $\gamma_i = \exp(\mathbf{x}'_i \boldsymbol{\beta} + \epsilon_i)$. As opposed to the NB and the PIG distributions, we cannot express this probability distribution in closed form. A closed-form log-likelihood function is needed to estimate the parameters of the distribution. In SAS, using the random effects option, the NLMIXED procedure can be used to evaluate the parameters.

Here again, other forms of variance having the same form as the other heterogeneous models (i.e., $\text{Var}[Y_i] = \lambda_i + \alpha\lambda_i^{2-p}$) can be considered.

4.4.4 Specification Test for Heterogeneity

There is a link between the Poisson and the models constructed with a heterogeneity parameter added to the mean of a Poisson distribution. Indeed, when the NB, the PIG, or the PLN is used, the Poisson distribution is obtained when the extra parameter α converges to 0, which corresponds to the border of the parameter space of the three heterogeneity distributions studied in this chapter. To test the possibility that $\alpha = 0$, classical hypothesis tests can be used. The three standard tests are the log-likelihood ratio (LR), the Wald, and the score tests. Asymptotically, all three tests are equivalent.

However, a problem with standard specification tests (the Wald or the log-likelihood ratio tests) happens when the null hypothesis is on the boundary of the parameter space. Indeed, when a parameter is bounded by the H_0 hypothesis, the estimate is also bounded, and the asymptotic normality of the MLE no longer holds under H_0 . In the negative binomial case, it was shown that, under the null hypothesis, the distribution

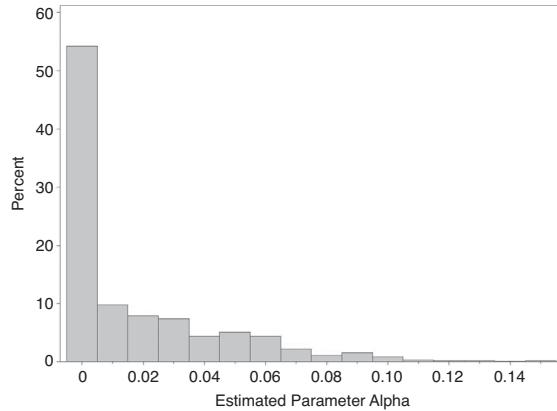


Fig. 4.1. Histogram of estimated values of NB2* α under Poisson simulations.

of the LR statistic is a mixture of a probability mass of $\frac{1}{2}$ on the boundary and a half- $\chi^2(1)$ distribution above zero. The non-normality of the estimators of α can be observed in Figure 4.1, where a Poisson distribution was simulated 1,000 times and estimated as an NB2 distribution. Figure 4.1 illustrates the estimated value of the α parameter under such simulations.

When testing at a level δ , one must reject the H_0 hypothesis if the test statistic exceeds $\frac{1}{2} \chi^2_{1-2\delta}(1)$, rather than $\chi^2_{1-\delta}(1)$, meaning that in this situation a one-sided test must be used. Analogous results stand for the Wald test, given that parameter distribution consists of a mass of one-half at zero and a normal distribution for the positive values. Again in this case, the usual one-sided test critical value of $z_{1-\delta}$ should be used.

Nevertheless, when the hypothesized parameter lies on the boundary of the parameter space, other tests can be used without. Indeed, the asymptotic properties of the score test are not altered when testing on the boundary of the parameter space.

The Poisson distribution can be compared to the heterogeneous models by testing the variance function of the heterogeneous model:

$$\text{Var}[Y_i] = \lambda_i + \alpha g(\lambda_i). \quad (4.3)$$

The function $g(\lambda_i)$ has to be replaced by the form of variance to be tested, such as $g(\lambda_i) = \lambda_i^2$ for an NB2 variance form or $g(\lambda_i) = \lambda_i$ for an NB1 variance form. By construction, the NB2, the PIG2, and the PLN2 are all tested by this score test when $g(\lambda_i) = \lambda_i^2$. Similarly, the NB1, the PIG1, and the PLN1 are all tested simultaneously against the Poisson distribution for $g(\lambda_i) = \lambda_i$.

Then, the null hypothesis $H_0 : \alpha = 0$ is tested against $H_a : \alpha > 0$, yielding the following score statistic test for the Poisson distribution against heterogeneous models

having a variance function of the form (4.3):

$$T_{LM} = \left[\sum_{i=1}^n \frac{1}{2} \hat{\lambda}_i^{-2} g^2(\hat{\lambda}_i) \right]^{-\frac{1}{2}} \sum_{i=1}^n \frac{1}{2} \hat{\lambda}_i^{-2} g(\hat{\lambda}_i) ((y_i - \hat{\lambda}_i)^2 - y_i).$$

For various forms $g(\lambda_i)$, the statistics T_{LM} are normally distributed with mean 0 and variance 1.

4.4.5 True and Apparent Occurrence Contagion

Strong assumptions are needed when we suppose a Poisson distribution for the number of claims. One of these assumptions is the absence of occurrence dependence. A positive (negative) occurrence means that the event of a claim increases (decreases) the probability of another claim. It has been shown that positive occurrence dependence can lead to a binomial negative distribution. The negative binomial distribution has also been shown to be the limit distribution of apparent occurrence dependence, where the modification of the probability of another claim arises from the recognition of the accident proneness of an individual. Consequently, we cannot deduce the kind of dependence implied by the data when we suppose a negative binomial distribution. This impossibility of distinguishing between true and apparent contagion has been called the *impossibility theorem*.

In the actuarial literature, the apparent occurrence dependence explanation seems to be the most accepted thesis. This indicates that although past events do not truly influence the probability of reporting a claim, they provide some information about the true nature of the driver. The heterogeneity term of the models can be updated according to the insured's accident history (see Chapter 7 for heterogeneity analysis for panel data or Chapter 14 for the a posteriori analysis of heterogeneity).

4.4.6 Numerical Application

Negative binomial distributions that have been applied to the Singapore dataset are shown in Table 4.2. The mean parameters $\hat{\beta}$ are approximately the same for the NB2, the NB1, and the Poisson distributions. Fit of the distributions can be verified by the log-likelihood, which shows that the introduction of another parameter α by the NB2 and the NB1 improves the adjustment slightly. However, because another parameter (α) has been added to the distributions NB2 and NB1, compared with the Poisson distribution, we cannot compare the log-likelihood directly. To compare the fit, information criteria that penalize models with a large number of parameters should be used. The classical criteria include Akaike information criteria ($AIC = -2 \log(L) + 2k$, where L is the likelihood and k is the number of parameters of the model) and the Bayesian information criteria ($BIC = -2 \log(L) + \log(n)k$, where k

Table 4.2. MLE for the NB2 and the NB1 Distributions

Parameter	NB2		NB1	
	Estimate	(Std.Err.)	Estimate	(Std.Err.)
β_0	-3.7735	(0.5146)	-3.7509	(0.5122)
β_1	0.7191	(0.1272)	0.7112	(0.1259)
β_2	0.3572	(0.1286)	0.3470	(0.1277)
β_3	1.5906	(0.5061)	1.5729	(0.5037)
β_4	1.5304	(0.5180)	1.5247	(0.5151)
β_5	1.1048	(0.5209)	1.0895	(0.5184)
β_6	0.3481	(0.5332)	0.3374	(0.5310)
α	0.4137	(0.2328)	0.0273	(0.0227)
Log-likelihood	-1801.7336		-1802.9360	

and n represent, respectively, the number of parameters of the model and the number of observations).

Another way of comparing the NB distributions with the Poisson distribution is to use the score test defined earlier. Applications of these tests lead to values of 2.25 and 1.23, where these statistics are normally distributed with mean 0 and variance 1 under H_0 . Consequently, the Poisson distribution is rejected against the NB2 distribution, but not against the NB1 distribution.

The difference between the Poisson, the NB1, and the NB2 distribution can be visualized by looking at the variance function. Figure 4.2 illustrates the variance of

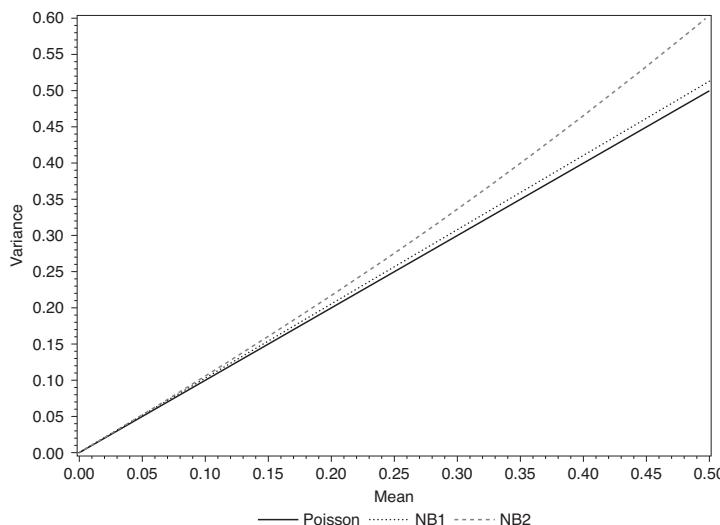


Fig. 4.2. Link between the mean and the variance.

Table 4.3. MLE for the PIG2 and the PIG1 Distributions

Parameter	PIG2		PIG1	
	Estimate	(Std.Err.)	Estimate	(Std.Err.)
β_0	-3.7731	(0.5146)	-3.7506	(0.5123)
β_1	0.7187	(0.1272)	0.7109	(0.1260)
β_2	0.3569	(0.1286)	0.3467	(0.1278)
β_3	1.5903	(0.5062)	1.5727	(0.5037)
β_4	1.5308	(0.5180)	1.5248	(0.5151)
β_5	1.1048	(0.5209)	1.0897	(0.5184)
β_6	0.3481	(0.5332)	0.3376	(0.5310)
τ	0.4170	(0.2382)	0.0277	(0.0232)
Log-likelihood	-1801.7263		-1802.9296	

each distribution compared with its mean. We can see that the differences between the models concern the riskiest drivers, where the NB2 distribution, having an exponential variance function, supposes a bigger variance for these clients.

The PIG and the PLN distributions have also been applied to the Singapore dataset, with results shown in Tables 4.3 and 4.4, respectively. Covariates are again approximately equal to the ones obtained with the Poisson and the NB distributions. As explained earlier, the same score test used to test the NB2 and the NB1 distributions must be applied to the models PIG2-PLN2/PIG1-PLN1. Consequently, we can again say that the Poisson distribution is rejected against the PIG2 and against the PLN2, whereas it is not rejected against the PIG1 and the PLN1.

Differences between models can be observed through the mean and the variance of the annual number of claims for some insured profiles. Several insured profiles have

Table 4.4. MLE for the PLN2 and the PLN1 Distributions

Parameter	PLN2		PLN1	
	Estimate	(Std.Err.)	Estimate	(Std.Err.)
β_0	-3.7732	(0.5146)	-3.7506	(0.5124)
β_1	0.7186	(0.1272)	0.7108	(0.1260)
β_2	0.3567	(0.1286)	0.3466	(0.1278)
β_3	1.5904	(0.5062)	1.5728	(0.5039)
β_4	1.5312	(0.5180)	1.5251	(0.5153)
β_5	1.1050	(0.5209)	1.0901	(0.5186)
β_6	0.3483	(0.5332)	0.3379	(0.5311)
σ^2	0.3462	(0.1693)	0.0277	(0.0232)
Log-likelihood	-1801.7411		-1802.9236	

Table 4.5. *The Four Types of Policyholders to be Compared*

Profile Number	x1	x2	x3	x4	x5	x6	x7
1	0	0	0	0	0	0	0
2	0	0	0	0	1	0	0
3	0	0	0	1	0	0	0
4	1	0	1	0	0	0	0

been selected and are described in Table 4.5. Profile 1 is classified as a good driver, whereas profile 4 usually exhibits bad loss experience. Other profiles are medium risk. The premium and the variance for each profile are given in Table 4.6, which shows that the differences between premiums are smaller than the differences between the variances. So, although we cannot see differences in premium corresponding to the expected value ($E[Y]$), in the case where the premiums are calculated using a premium principle based on the standard error ($E[Y] + \theta\sqrt{\text{Var}[Y]}$) or on the variance ($E[Y] + \theta\text{Var}[Y]$) or, when using other transformations of the distribution (Esscher premium, for example = $E[Se^{hY}]/E[e^{hY}]$, for a given h), the choice of the distribution has an impact.

Finally, using the estimated parameters $\hat{\alpha}$ for the NB2, the PIG2, and the PLN2 (that do not depend on the covariates), we can illustrate the heterogeneity distribution of the underlying gamma, inverse-Gaussian, and lognormal. This is shown in Figure 4.3, where we see a close similarity between the distributions. Figure 4.4 illustrates the heterogeneity distribution of profiles 1, 2, 3, and 4, which depends on λ_i . For all distributions, the inverse-Gaussian shows a fatter tail than the gamma and the lognormal distributions. This property of the inverse-Gaussian does not have a big impact on the count analysis for cross-section data, but may lead to the highest penalties for bad drivers (see Chapter 9 on credibility for details).

Table 4.6. *Comparison of a Priori Claim Frequencies*

Dist.	1 st Profile		2 nd Profile		3 rd Profile		4 th Profile	
	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.
Poisson	0.0229	0.0229	0.0693	0.0693	0.1063	0.1063	0.2314	0.2314
NB2	0.0230	0.0232	0.0693	0.0713	0.1061	0.1108	0.2313	0.2535
NB1	0.0235	0.0241	0.0699	0.0718	0.1079	0.1109	0.2306	0.2369
PIG2	0.0230	0.0232	0.0694	0.0714	0.1062	0.1109	0.2313	0.2536
PIG1	0.0235	0.0242	0.0699	0.0718	0.1080	0.1110	0.2306	0.2370
PLN2	0.0230	0.0232	0.0693	0.0710	0.1061	0.1100	0.2314	0.2499
PLN1	0.0235	0.0242	0.0699	0.0718	0.1080	0.1110	0.2306	0.2370

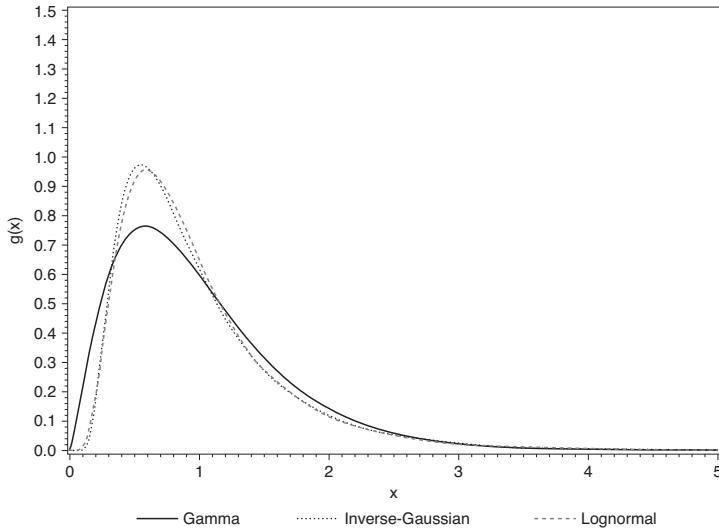


Fig. 4.3. Densities of the heterogeneity distributions.

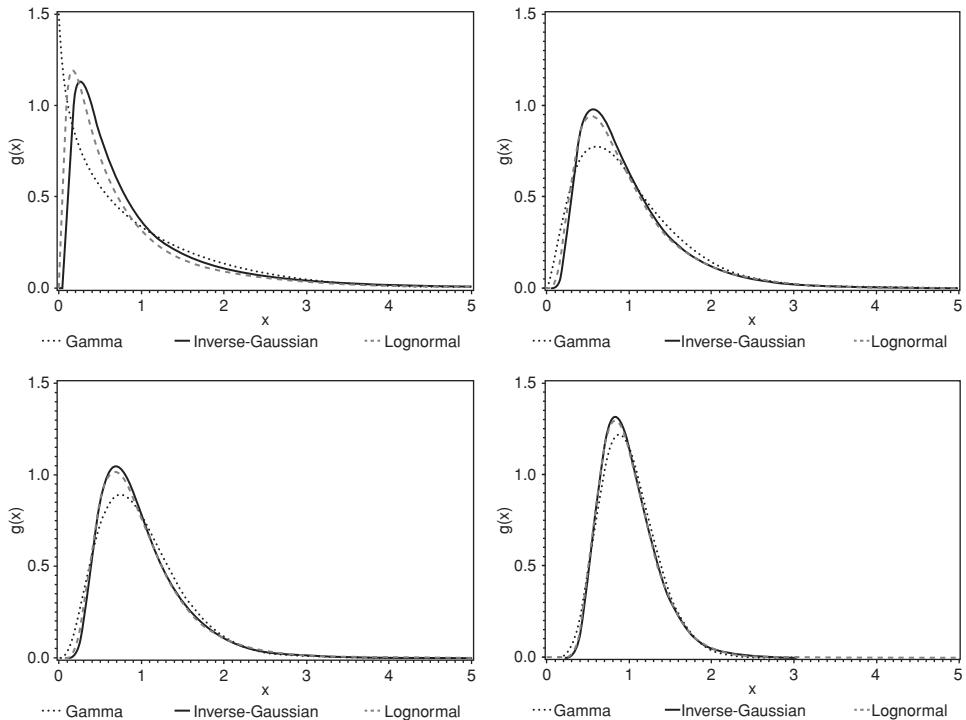


Fig. 4.4. Densities of the heterogeneity distributions for profiles 1 to 4.

4.5 Zero-Inflated Distribution

A high number of zero values is often observed in the fitting of claim counts. An increasingly popular solution for this kind of data is a distribution with excess zeros, often called zero-inflated distribution. A zero-inflated distribution is a finite mixture model of two distributions combining an indicator distribution for the zero case and a standard count distribution. The probability distribution of this zero-inflated distribution can be expressed as

$$\Pr[Y = y] = \begin{cases} \phi + (1 - \phi) \Pr[V = 0] & \text{for } n = 0 \\ (1 - \phi) \Pr[V = y] & \text{for } n = 1, 2, \dots \end{cases}, \quad (4.4)$$

where the random variable V follows a standard count distribution to be modified by an additional excess zero function. In the limiting case, where $\phi \rightarrow 0$, the zero-inflated model corresponds to the distribution of V .

Many distributions may be used with the zero-inflated models. Obviously, the classic distribution is the zero-inflated Poisson (ZIP) distribution. With the use of equation (4.4), the density of the ZIP model is

$$\Pr[Y = y] = \begin{cases} \phi + (1 - \phi)e^{-\lambda} & \text{for } y = 0 \\ (1 - \phi) \frac{e^{-\lambda} \lambda^y}{y!} & \text{for } y = 1, 2, \dots \end{cases}.$$

From this, we can determine the first two moments of the ZIP distribution: $E[Y] = (1 - \phi)\lambda$ and $\text{Var}[Y] = E[Y] + E[Y](\lambda - E[Y])$. Covariates can be included in the λ parameter with a log link function, as done with the Poisson and the other models with heterogeneity. Moreover, covariates can also be included in the ϕ parameter, with a logit, a probit, or other appropriate link functions.

The ZIP model could be useful for modeling purposes because it accounts for overdispersion. Since the only difference between the ZIP model and the Poisson distribution is found when $Y = 0$, it is easy to adjust the MLE equations of the Poisson distribution to find the parameters of the ZIP model. Other count distributions can also be used with the zero-inflated distributions, such as the NB, the PIG, or the PLN, leading to distributions called ZI-NB, ZI-PIG, or ZI-PLN, respectively.

4.5.1 Interpretation and Risk Exposure

Recently, it has been shown that applying the zero-inflated distributions to model the number of claims can be a tool to describe the behavior of insureds. More precisely, it has been shown that fitting zero-inflated distributions to claims data can model the probability of having an accident. Indeed, as is well known in the insurance industry, not all accidents are reported and the insurer is only aware of the accident claims.

There are mainly two ways to explain how the zero-inflated distribution can differentiate between the claim and the accident:

- (1) A first interpretation to justify the use of a zero-inflated distribution is the assumption that each year, a number of insureds will not claim at all, whatever the case. In this situation, one might question why these insureds procure insurance. Some explanations refer to their fear of insurance, their having minimal protection (mandatory insurance), or their being insured only for major risk (with probability close to 0).
- (2) Another way of interpreting the zero-inflated model assumes that the number of accidents is Poisson distributed. In addition, it considers the probability of each accident being reported. The model assumes that the first accident of each insured year indicates the way the insured will act for the rest of the year. Accordingly, if the first accident is reported, the succeeding ones will also be reported. If the first accident is not reported, the following accidents will not be reported. Those who will not claim their first accident, because they made an effort to financially support their decision, tend to defend the way they act and consequently will not claim other accidents.

When all the insureds analyzed in the portfolio have a risk exposure of 1, meaning that we observe all the clients for complete one-year contracts, the two interpretations of the ZIP models generate the same model. Both models are equivalent.

However, when some insureds in the portfolio do not have complete coverage, the model differs depending on which interpretation we choose. In the first interpretation, because the insurance portfolio supposes that a portion of the insureds will not report at all, the parameter ϕ that corresponds to the extra probability of having no claim does not need to be modeled using risk exposure. However, in the second case, the ϕ parameter should include the risk exposure because it is linked to the probability of reporting the first accident. As seen in Section 4.2.2, the risk exposure of a Poisson distribution is proportional to the observed time length. It is also important to understand that the Poisson in the ZIP model is also modeled with risk exposure, meaning that mean parameter of the Poisson is equal to λt .

4.5.2 Numerical Applications

Zero-inflated models have been applied to the exposure dataset [exposure.csv]. The data contain the number of claims and the risk exposure for 10,000 automobile insurance holders. No covariates are included in the data; we focus only on the interpretation of the zero-inflated models. The Poisson distribution and the two forms of the zero-inflated Poisson distribution have been used with the data. Results are shown in Table 4.7.

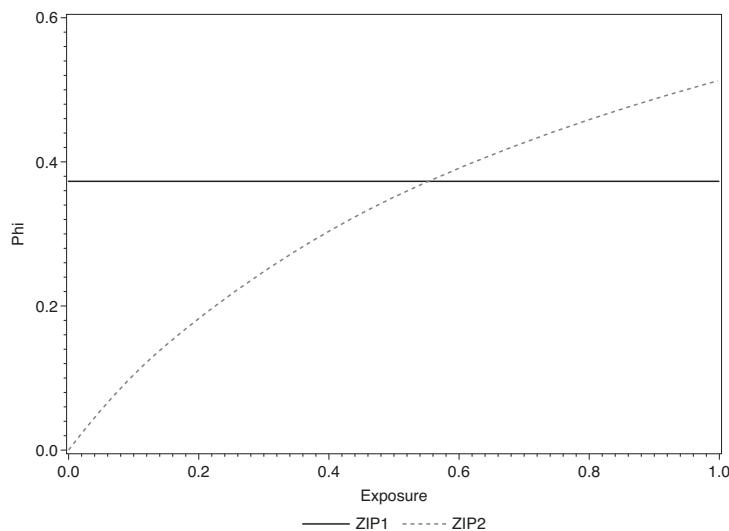
The first zero-inflated Poisson (ZIP1) supposes that the probability that an insured will not report at all is independent of the risk exposure. Consequently, the extra

Table 4.7. MLE for the Poisson and the Zero-Inflated Poisson Distribution

Par.	Poisson		ZIP1		ZIP2	
	Est.	(Std.Err.)	Est.	(Std.Err.)	Est.	(Std.Err.)
β_0	-2.2147	(0.0381)	-1.7444	(0.1553)	-1.6262	(0.1199)
ϕ	.	.	0.3731	(0.0937)	.	.
ψ_0	0.0505	(0.2400)
ψ_1	0.9578	(0.2633)
Log-likelihood	-2533.8862		-2529.7992		-2516.2043	

probability assigned to the possibility of having $Y = 0$ is modeled as ϕ for the whole portfolio. In contrast, the second zero-inflated Poisson (ZIP2) supposes that the additional mass-at-zero is modeled as $\phi = \text{logit}(\psi_0 + \psi_1 \ln(t))$. The ZIP2 model generalizes the ZIP1 model because a value of ψ_1 equal to 0 means that the ZIP2 model collapses to the ZIP1 model. Looking at the log-likelihood (and the corresponding AIC or BIC), we can see that the ZIP2 seems to offer a better fit to the data. Moreover, a classic Wald test shows that the value of ψ_1 is statistically different from 0, meaning that we should prefer the second interpretation of the ZIP models.

To help us understand the difference between the two ZIP models in more detail, Figure 4.5 illustrates the values of the ϕ parameters depending on the risk exposure. As the number of days of coverage increases, the extra probability-at-zero becomes higher.

Fig. 4.5. Values of ϕ for both ZIP models, depending on the risk exposure.

As mentioned, the ZIP model allows us to distinguish underreporting from the driving behavior. Consequently, using zero-inflated distributions on the number of claims, we can *uncensor* these zero-inflated distributions to obtain an approximation of the accident frequency distribution. By removing all the effects of reporting that we modeled by the censorship parameter ϕ , the accident process is Poisson distributed, which is simple and easily understood. Indeed, as mentioned in Section 4.2.1, the Poisson process as a limit of a binomial distribution is an intuitive way to model the number of accidents.

If the uncensored version of the zero-inflated distribution is used, all the ϕ parameters must be set to zero. In other words, the zero-inflated distributions are fitted to claims data to find the $\lambda_i = \exp(x'_i\beta)$ of the Poisson distribution of the number of accidents. In our numerical example, according to the ZIP2 model, this means that the number of accidents follows a Poisson distribution, with mean $\lambda = 0.1966 (= \exp(-1.6262))$.

4.6 Conclusion

A wide selection of models can be used to model count data and, more precisely, to model the number of claims by an insured. We chose to focus on the Poisson distribution, on its generalization by the introduction of a random heterogeneity term, and on zero-inflated models. It is also possible to use other advanced models to model claim count. Different interpretations of the parameters can be done depending on the chosen model. The choice of the best distribution should be supported by specification tests and a goodness-of-fit test. Such count distributions can serve as a basis for the modeling of time series of count, for panel data, or for hierarchical models.

4.7 Further Reading

For a general overview of count data, particularly for specification tests, or for general tests of fitting for count data, we refer the reader to Cameron and Trivedi (1998). The distinction between apparent and true contagion in claim counts has been widely studied by Pinquet (2000). Zero-inflated models with an analysis of the dichotomy between the number of claims and the number of accidents have been analyzed in Boucher, Denuit, and Guillén (2009), in Lemaire (1995), and in Denuit et al. (2007).

References

- Boucher, J.-P., M. Denuit, and M. Guillén (2009). Number of accidents or number of claims? An approach with zero-inflated Poisson models for panel data. *Journal of Risk and Insurance* 76(4), 821–846.

- Cameron, A. C. and P. Trivedi (1998). *Regression Analysis of Count Data*. Cambridge University Press, New York.
- Denuit, M., X. Maréchal, S. Pitrebois, and J.-F. Walhin (2007). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. Wiley, New York.
- Lemaire, J. (1995). *Bonus-Malus Systems in Automobile Insurance*. Kluwer Academic Publishers, Boston.
- Pinquet, J. (2000). Experience rating through heterogeneous models. In G. Dionne (Ed.), *Handbook of Insurance*, pp. 459–500. Kluwer Academic Publishers, Boston.

5

Generalized Linear Models

Curtis Gary Dean

Chapter Preview. Generalized linear models (GLMs) generalize linear regression in two important ways: (1) the response variable y can be linked to a linear function of predictor variables x_j with a nonlinear link function, and (2) the variance in the response variable y is not required to be constant across observations but can be a function of y 's expected value. For example, if y represents the number of claims, then the variance in y may depend on the expected value of y as in a Poisson distribution. In linear regression the normal distribution plays a key role, but with GLMs the response variable y can have a distribution in a linear exponential family, which includes distributions important to actuaries: Poisson, binomial, normal, gamma, inverse-Gaussian, and compound Poisson-gamma. Actuaries can model frequency, severity, and loss ratios with GLMs, as well as probabilities of events such as customers renewing policies.

The likelihood function has a key role in GLMs. Maximum likelihood estimation replaces least squares in the estimation of model coefficients. The log-likelihood function is used to perform statistical tests.

5.1 Introduction to Generalized Linear Models

5.1.1 Assumptions of Linear Models

Multiple linear regression (Chapter 2) models response variables y_i , with $i = 1, \dots, n$, as a linear function of predictor variables x_{ij} , often called explanatory variables, plus a constant β_0 :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i. \quad (5.1)$$

The k predictor variables are given, nonrandom variables whose values can change with i . The error terms ε_i are the differences between the response variables and their

predicted values:

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}).$$

Two key assumptions are that error terms ε_i have expected value 0, $E[\varepsilon_i] = 0$, and that the variance of ε_i is constant and does not change across observations i , a property referred to as homoskedasticity: $\text{Var}[\varepsilon_i] = \sigma^2$. The errors ε_i are usually assumed to be independent and normally distributed.

Taking expected values in (5.1) yields

$$E[y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}. \quad (5.2)$$

Taking the variance of both sides of (5.1), $\text{Var}[y_i] = \text{Var}[\varepsilon_i] = \sigma^2$. The normality of error terms implies that response variables y_i are normally distributed about their respective means $E[y_i]$.

The coefficients in the linear model are estimated by the method of least squares.¹ Normality is not a requirement to construct a linear model using least squares, but is important for hypothesis testing and constructing confidence intervals. Linear models have shown their value in modeling, but in many situations linear models need to be generalized as demonstrated in the next section.

Problems with Predicting Number of Claims

Suppose that y_i represents the number of claims for risk i in a portfolio of n risks. The actuary may want to predict the expected number of claims for each risk i , $E[y_i]$, based on k risk characteristics $x_{i1}, x_{i2}, \dots, x_{ik}$. Multiple linear regression may not be the best tool for this job. (For an in-depth discussion of modeling counts, see Chapter 4.) Here are three problems with applying the standard linear model:

- (1) The Poisson is commonly used to model the number of claims. If y_i is Poisson distributed, then $\text{Var}[y_i] = E[y_i]$: the variance is not constant across observations, but depends on the expected value of the response variable. The assumption of constancy of the variance across risks, $\text{Var}[y_i] = \text{Var}[\varepsilon_i] = \sigma^2$, does not hold.
- (2) When modeling the expected number of claims, the left-hand side of equation (5.2) needs to be non-negative, but this cannot be guaranteed in the linear model. It is quite possible that some combination of the predictors x_{ij} could result in a negative value for $b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik}$.²
- (3) Rather than building an additive model where the contributions of risk characteristics $x_{i1}, x_{i2}, \dots, x_{ik}$ are added, perhaps a multiplicative model is more appropriate. For example, a bad driver may drive predominantly in a territory with a high rate of

¹ The true coefficients $\beta_0, \beta_1, \dots, \beta_k$ are unknowable, but are estimated by quantities b_0, b_1, \dots, b_k determined from the data.

² The calculated coefficient values b_0, b_1, \dots, b_k are used to compute quantities in a model.

accidents. Should the contributions from poor driving ability be added to territory effects, or should these contributions be multiplied?

The linear model needs to be adjusted. Letting $\ln(E[y_i])$ equal the linear combination of predictor variables addresses items 2 and 3. Equation (5.2) becomes $\ln(E[y_i]) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$ which gives

$$E[y_i] = e^{\beta_0} e^{\beta_1 x_{i1}} \cdots e^{\beta_k x_{ik}}.$$

With this model the expected number of claims for risk i , $E[y_i]$, will not be negative, and the predictive model is multiplicative. This adjustment to the linear model raises another issue: how to determine coefficients β_j in this nonlinear equation.

Linear models can be extended to account for heteroskedasticity (i.e., $\text{Var}[y_i]$ is not constant across risks as in item 1). One technique is to use weighted least squares with weights $w_i = 1/\text{Var}[y_i]$ that are inversely proportional to variances to estimate coefficients $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. Less weight would be given when the variance is greater. This approach needs a model for how $\text{Var}[y_i]$ changes across risks i . If we assume that y_i is Poisson distributed for each risk i , then it follows that $\text{Var}[y_i] = E[y_i]$ but we have the problem that $\text{Var}[y_i]$ is unknown until $E[y_i]$ is calculated.

Some of the complications arising from predicting the number of claims with linear models can be addressed by moving on to generalized linear models. Predicting the number of claims is only one of many applications where linear models need to be extended. Generalized linear models were developed to unify a variety of statistical models.

5.1.2 Generalized Linear Model Assumptions

Generalized linear models (GLMs) generalize linear regression in two important ways:

- (1) The independent response variables y_i can be linked to a linear function of predictor variables x_{ij} with a nonlinear link function.
- (2) The variance in the response variables y_i is not required to be constant across risks, but can be a function of y_i 's expected value.

The GLM predictive equation for response random variables y_i is

$$g(E[y_i]) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.$$

The link function $g()$ can be a nonlinear function. In classic linear regression, $g()$ is just the identity function $g(x) = x$. There is a restriction on link function $g()$ that it be differentiable and strictly monotonic.³ Because it is strictly monotonic, its inverse

³ A function is strictly monotonic if it is strictly increasing or strictly decreasing. If $f(x)$ is strictly increasing then $x_1 < x_2$ implies $f(x_1) < f(x_2)$. Strictly decreasing is defined similarly.

function exists, and the previous equation can be rewritten as

$$E[y_i] = g^{-1}(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}). \quad (5.3)$$

The predictor variables x_{ij} are still combined into a linear function, but the response variable $E[y_i]$ can be a nonlinear function of this linear combination. The linear function of predictor variables is often assigned the symbol η :

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}. \quad (5.4)$$

Letting $\mu_i = E[y_i]$ yields a shorthand equation for (5.3):

$$\mu_i = g^{-1}(\eta_i).$$

If the link function is $g(x) = \ln(x)$ with the corresponding inverse function $g^{-1}(x) = e^x$, then a multiplicative model will result.

The other important GLM assumption is that random variables y_i can be members of a linear exponential family of distributions. The modeler can choose a distribution from this family that is appropriate for the application; for example, a Poisson distribution to model the number of claims. The relationship between variance $\text{Var}[y_i]$ and expected value $E[y_i]$ for risks will depend on the chosen distribution.

An advantage of choosing a particular distribution for the model is that maximum likelihood estimation can be used to calculate the coefficients, and there are algorithms for computing the coefficients that work for all distributions in the exponential family.

5.2 Exponential Family of Distributions

Many distributions used by actuaries share a common structure and can be grouped into an exponential family. This has made it possible to construct the common framework for analysis referred to as generalized linear models.

For GLMs, response variable y is assumed to have a probability distribution function⁴ that can be written as⁵

$$f(y; \theta, \phi) = \exp \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right]. \quad (5.5)$$

Note that function $c(y, \phi)$ does not include parameter θ . Parameter θ is often referred to as the canonical parameter, natural parameter, or parameter of interest. Parameter

⁴ The function can be a probability density function (e.g., normal), probability mass function (e.g., binomial), or a mixture (e.g., Tweedie).

⁵ Formula (5.5) is a special case of the more general formula $f(y; \theta) = \exp[(r(\theta)h(y) + s(y) + q(\theta))]$ that defines the exponential family. Note that $h(y)$ has been replaced by a linear term y in (5.5). Distributions whose density functions can be written as (5.5) are members of the linear exponential family.

ϕ is called the dispersion parameter or, sometimes, nuisance parameter because the mean of the distribution does not depend directly on ϕ . The functions $b(\theta)$, $a(\phi)$, and $c(y, \phi)$ determine the type of distribution; for example, normal, Poisson, and so on. The Poisson distribution, with one parameter, also fits into this family as shown in Example 5.1.

The mean and variance of the distribution are simply

$$E[y] = b'(\theta), \quad (5.6)$$

$$\text{Var}[y] = a(\phi)b''(\theta), \quad (5.7)$$

where $b'(\theta)$ is the first derivative with respect to θ and $b''(\theta)$ is the second derivative. Derivations of the formulas for the mean and variance are shown in the appendices.

Distributions in this exponential family include the normal, binomial, Poisson, exponential, gamma, inverse-Gaussian, and the compound Poisson-gamma. With a little algebra the common forms of these distributions can be rewritten in exponential family form (5.5).

Example 5.1. The Poisson distribution is a member of the exponential family with probability mass function

$$\begin{aligned} f(y; \lambda) &= \frac{\lambda^y e^{-\lambda}}{y!} \\ &= \exp \left[\ln \left(\frac{\lambda^y e^{-\lambda}}{y!} \right) \right] \\ &= \exp \left[\frac{y \ln \lambda - \lambda}{1} - \ln y! \right]. \end{aligned}$$

Making the substitution $\theta = \ln \lambda$ or $e^\theta = \lambda$ produces

$$f(y; \theta) = \exp \left[\frac{y\theta - e^\theta}{1} - \ln y! \right].$$

Note that $b(\theta) = e^\theta$ and $c(y, \phi) = -\ln y!$. We can let $a(\phi) = \phi = 1$. Calculating the mean and variance of the distribution,

$$E[y] = b'(\theta) = e^\theta = \lambda$$

$$\text{Var}[y] = a(\phi)b''(\theta) = e^\theta = \lambda.$$

For the Poisson random variable y , $\text{Var}[y] = \lambda = E[y]$. As the expected value of y varies so does its variance.

Example 5.2. Suppose that $y \sim N(\mu, \sigma^2)$. The normal distribution is a member of the exponential family:

$$\begin{aligned} f(y; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \\ &= \exp\left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\right] \exp\left[-\frac{y^2 - 2y\mu + \mu^2}{2\sigma^2}\right] \\ &= \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \left(\frac{y^2}{2\sigma^2} + \ln(\sqrt{2\pi\sigma^2})\right)\right]. \end{aligned}$$

The parameter μ corresponds to θ in (5.5) and $\phi = \sigma^2$. Rewriting the pdf with parameters θ and ϕ ,

$$f(y; \theta, \phi) = \exp\left[\frac{y\theta - \theta^2/2}{\phi} - \left(\frac{y^2}{2\phi} + \ln(\sqrt{2\pi\phi})\right)\right].$$

So, $b(\theta) = \theta^2/2$, $a(\phi) = \phi$, and $c(y, \phi) = -\left(\frac{y^2}{2\phi} + \ln(\sqrt{2\pi\phi})\right)$. The mean and variance are

$$\begin{aligned} E[y] &= b'(\theta) = \frac{d(\theta^2/2)}{d\theta} = \theta = \mu \\ \text{Var}[y] &= a(\phi)b''(\theta) = \phi \frac{d^2(\theta^2/2)}{d\theta^2} = \phi = \sigma^2. \end{aligned}$$

The most complicated term, $c(y, \phi)$, is not relevant in the computation of the mean and the variance.

Table 5.1 displays five common distributions in the exponential family. The pdfs for distributions are shown on the left side of the table in commonly displayed forms. The exponential family parameters for each distribution can be derived from the pdfs on the left.

Suppose that response variable y_i represents the loss severity for the i^{th} risk where *loss severity* = *losses / number of losses*. A loss severity computed from several observed losses will be a better predictor of the expected loss severity than would a single loss. If the variance of a single loss is $\text{Var}[L]$, then the variance in loss severity based on m losses would be $\text{Var}[L]/m$. Greater weights should be attached to observed severities based on more losses. This same argument would apply to other types of response variables whose value may depend on an average.

Example 5.3. Let v_j be independent random variables with $v_j \sim N(\mu, \sigma^2)$ for $j = 1, \dots, w$. If $y = (v_1 + \dots + v_w)/w$ is the average, then $E[y] = \mu$ and $\text{Var}[y] = \sigma^2/w$.

Table 5.1. Exponential Family Form

$$f(y; \theta, \phi) = \exp \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right]$$

$$E[y] = b'(\theta) \quad , \quad \text{Var}[y] = a(\phi)b''(\theta)$$

Common form of pdf	θ	$b(\theta)$	ϕ	$a(\phi)$	$c(y, \phi)$
Normal: $\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(y-\mu)^2}{2\sigma^2} \right]$	μ	$\theta^2/2$	σ^2	ϕ	$-\frac{1}{2} \left[\frac{y^2}{\phi} + \ln(2\pi\phi) \right]$
Poisson: $\lambda^y e^{-\lambda} / y!$	$\ln \lambda$	e^θ	1	1	$-\ln(y!)$
Binomial ^a : $\binom{n}{y} p^y (1-p)^{n-y}$	$\ln[p/(1-p)]$	$n \ln(1+e^\theta)$	1	1	$\ln \binom{n}{y}$
Gamma ^b : $\beta^\alpha y^{\alpha-1} e^{-\beta y} / \Gamma(\alpha)$	$-\frac{\beta}{\alpha}$	$-\ln(-\theta)$	$\frac{1}{\alpha}$	ϕ	$\frac{1}{\phi} \ln \frac{y}{\phi} - \ln y - \Gamma \left(\frac{1}{\phi} \right)$
Inverse-Gaussian: $\sqrt{\frac{\lambda}{2\pi y^3}} \exp \left[\frac{-\lambda(y-\mu)^2}{2y\mu^2} \right]$	$\frac{-1}{2\mu^2}$	$-\sqrt{-2\theta}$	$\frac{1}{\lambda}$	ϕ	$-\frac{1}{2} \left[\ln(2\pi\phi x^3) + \frac{1}{\phi y} \right]$

Notes: Other sources may show different parameterizations. Common forms of pdfs are displayed so that readers can make their own calculations and reconcile formulas.

^a n is assumed to be a known, fixed quantity.

^b With this parameterization the mean and variance are $E[y] = \alpha/\beta$ and $\text{Var}[y] = \alpha/\beta^2$.

Using the results of Example 5.2, the probability density function for y is

$$f(y; \theta, \phi, w) = \exp \left[\frac{y\theta - \theta^2/2}{\phi/w} - \left(\frac{y^2}{2\phi/w} + \ln(\sqrt{2\pi\phi/w}) \right) \right].$$

where $\phi = \sigma^2$ again. Note that the pdf for random variable y depends on the weight w . As the weight w increases, the variance of y decreases.

For response variables y_i it is standard practice, as shown in Examples 5.2 and 5.3, to define parameter ϕ and function $a()$ so that $a(\phi)$ can be replaced by ϕ/w_i in equation (5.5). The pdf for y_i becomes

$$f(y_i; \theta_i, \phi, w_i) = \exp \left[\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i, \phi/w_i) \right]. \quad (5.8)$$

Parameter ϕ is called the scale parameter or dispersion and is usually assumed to be constant for all response variables in the sample. The weights can be identically one or can be inputs into the GLM. In insurance applications the number of claims or exposures are common measures for weights w_i .

The Appendix shows how to put the gamma and binomial distributions into the exponential family form shown in (5.5).

5.2.1 The Variance Function and the Relationship between Variances and Means

The variance formula from (5.7) with ϕ/w_i replacing $a(\phi)$ is

$$\text{Var}[y_i] = \frac{\phi}{w_i} b''(\theta_i). \quad (5.9)$$

The parameter θ_i can be replaced by a function of $\mu_i = E[y_i]$ as shown later. (The substitution of μ_i for $E[y_i]$ will make what follows more readable.) Inverting formula (5.6) for the mean

$$\mu_i = b'(\theta_i),$$

$$\theta_i = b'^{-1}(\mu_i),$$

and replacing θ_i in (5.9) by the right-hand side above gives

$$\text{Var}[y_i] = \frac{\phi}{w_i} V(\mu_i), \quad (5.10)$$

where $V(\mu_i) = b''(b'^{-1}(\mu_i))$. Function $V(\mu_i)$ is called the *variance function*. The variance function defines the relationship between the variance and the mean for a distribution in the exponential family (see Table 5.2).

Table 5.2. Variance Functions $V(\mu)$

Distribution	$V(\mu)$	Distribution	$V(\mu)$
normal	$\mu^0 = 1$	Tweedie	$\mu^p, 1 < p < 2$
binomial	$\mu(1 - \mu)$	gamma	μ^2
Poisson	μ	inverse-Gaussian	μ^3

The Tweedie⁶ distribution is the name commonly used for the compound Poisson-gamma distribution. Note that its exponent p lies between that of the Poisson and gamma. Intuitively that makes sense because the number of losses is Poisson distributed and the size of each loss is gamma distributed.

The variance of response variable y_i controls how much weight to put on observation y_i when fitting the GLM. As is demonstrated in Section 5.4, when $\text{Var}[y_i]$ is larger, less weight will be given to observation y_i in calculating GLM coefficients b_0, b_1, \dots, b_k than when $\text{Var}[y_i]$ is smaller:

$$\text{Weight for observation } i \propto \frac{1}{\text{Var}[y_i]} = \frac{w_i}{V(\mu_i)}.$$

For example, the inverse-Gaussian distribution with variance function $V(\mu) = \mu^3$ will assign much less weight to observations that have larger expected means. Note that we have assumed that ϕ is constant for all i .

5.3 Link Functions

The link function must be differentiable and strictly monotonic – either strictly increasing or strictly decreasing – so that its inverse exists:

$$\begin{aligned} g(\mu_i) &= \eta_i, \\ \mu_i &= g^{-1}(\eta_i). \end{aligned}$$

Recall $\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$ from (5.4).

Common choices for the link function are shown in Table 5.3. The modeler has a choice of link functions, but some links may be more appropriate than others for a model. For example, an important consideration is selecting the link function is the range of $\mu_i = E[y_i]$.

Response Variable y_i Is Number of Claims. The expected number of claims μ_i has range $(0, \infty)$. The linear predictor $\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$ may have range

⁶ Maurice Charles Kenneth Tweedie (1919–1996) was a medical statistician. Tweedie distributions that bear his name are a particular subclass of an exponential dispersion model. The compound Poisson-gamma is included in the subclass.

Table 5.3. Common Link Functions

	$g(\mu)$	$g^{-1}(\eta)$	Range of $g^{-1}(\eta)$
identity	μ	η	$(-\infty, \infty)$
log	$\ln(\mu)$	e^η	$(0, \infty)$
logit	$\ln[\mu/(1 - \mu)]$	$e^\eta/(1 + e^\eta)$	$(0, 1)$
probit	$\Phi^{-1}(\mu)$	$\Phi(\eta)$	$(0, 1)$
complementary log-log	$\ln(-\ln(1 - \mu))$	$1 - e^{-e^\eta}$	$(0, 1)$
inverse	$1/\mu$	$1/\eta$	$(-\infty, 0) \cup (0, \infty)$
inverse squared	$1/\mu^2$	$1/\sqrt{\eta}$	$(0, \infty)$

$(-\infty, \infty)$. It may be that $\eta_i < 0$ for possible combinations of predictors x_{ij} . A solution to this contradiction is a log-link function. The log link is $g(\mu) = \ln(\mu)$:

$$\ln(\mu) : (0, \infty) \rightarrow (-\infty, \infty).$$

The inverse of log link $g^{-1}(\eta) = e^\eta$ maps $(-\infty, \infty)$ onto $(0, \infty)$.

μ_i is Probability of an Event. The GLM may model probability of events such as customers renewing policies or claims being fraudulent. The response variable y_i will take on values of 1 or 0 depending on whether the event happens or not and μ_i will be the probability of the event. Probabilities μ_i have range $[0, 1]$. As discussed earlier, η_i has possible range $(-\infty, \infty)$. An appropriate link function can be constructed in two steps. If p is the probability of an event, then the odds-ratio is $p/(1 - p)$:

$$p/(1 - p) : (0, 1) \rightarrow (0, \infty).$$

Next take the natural log of the odds-ratio:

$$\ln(p/(1 - p)) : (0, 1) \rightarrow (-\infty, \infty).$$

Link $g(\mu) = \ln(\mu/(1 - \mu))$ is called the logit link. Two other common links for this mapping are

- (1) probit: $g(\mu) = \Phi^{-1}(\mu)$ where $\Phi^{-1}()$ is the inverse standard cumulative normal distribution,
- (2) complementary log-log: $g(\mu) = \ln(-\ln(1 - \mu))$

Another consideration in the selection of the link function is the relationship between response variables and predictors. If the effects of the predictive variables are additive, then the identity link may be appropriate. If the effects of predictors are multiplicative, then log link is better because its inverse is exponential:

identity link: $\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$.

log link: $\mu_i = e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}} = e^{\beta_0} e^{\beta_1 x_{i1}} \cdots e^{\beta_k x_{ik}}$.

5.3.1 Canonical Links

The canonical parameter θ in the pdf for random variable y is related to $\mu = E[y]$ by $\mu = b'(\theta)$. Inverting the function gives canonical parameter θ as a function of μ : $\theta = b'^{-1}(\mu)$.

The GLM predictive equation is

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.$$

If link function $g()$ is chosen such that $\theta_i = g(\mu_i)$ then

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}. \quad (5.11)$$

Canonical parameters θ_i are equal to a linear function of the predictors. The chosen link function $g()$ is called a *canonical link function*.

Canonical links generate linear equations (5.11) for unknown parameters θ_i .⁷ A canonical link may be a good choice for modeling a particular problem, but the modeler should not feel compelled to select a canonical link. The overall fit of the model and other considerations such as intuitive appeal may be more important.

Example 5.4. Parameters and functions for the gamma distribution are shown in Table 5.1. For a gamma distribution $b(\theta) = -\ln(-\theta)$ so $b'(\theta) = -1/\theta$, which implies $\mu = -1/\theta$. Solving for θ gives $\theta = -1/\mu$.

If link function $g(\mu) = -1/\mu$ is used with a gamma distribution, then the predictive equations can be written as

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.$$

Link function $g(\mu) = -1/\mu$ is a canonical link for the gamma. If one uses the link $g(\mu) = 1/\mu$ instead, then

$$-\theta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.$$

But multiplying through by -1 will reverse the sign of the coefficients on the right, showing that the θ_i can still be calculated as linear functions of the predictors. The same goes for any constant times θ on the left. Multiply through by the reciprocal. Most sources will get rid of the negative sign of $-1/\mu$ and display $g(\mu) = 1/\mu$ as the canonical link for the gamma distribution.

⁷ On page 32 of McCullagh and Nelder (1997) they explain that canonical links result in the existence of a sufficient statistic for the model.

5.4 Maximum Likelihood Estimation

The coefficients $\beta_0, \beta_1, \dots, \beta_k$ for the GLM are estimated from the data using maximum likelihood estimation (MLE). Choosing a distribution to model random variables y_i allows one to apply MLE.

The likelihood function is

$$L(\mathbf{y}; \boldsymbol{\beta}) = \prod_{i=1}^n \exp \left[\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c_i(y_i, \phi) \right].$$

using (5.5).⁸ The left-hand side shows that the likelihood is a function of the n observations y_1, \dots, y_n represented by vector \mathbf{y} and parameters β_0, \dots, β_k represented by vector $\boldsymbol{\beta}$. It is easier to maximize the log-likelihood:

$$l(\mathbf{y}; \boldsymbol{\beta}) = \sum_{i=1}^n \left[\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c_i(y_i, \phi) \right]. \quad (5.12)$$

The log-likelihood can be maximized by calculating partial derivatives with respect to the β_j 's and setting them equal to zero. The partial derivative with respect to β_j is

$$\begin{aligned} \frac{\partial l(\mathbf{y}; \boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^n \frac{\partial}{\partial \beta_j} \left[\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c_i(y_i, \phi) \right] \\ &= \sum_{i=1}^n \frac{1}{a_i(\phi)} \left[y_i \frac{\partial \theta_i}{\partial \beta_j} - \frac{\partial b(\theta_i)}{\partial \beta_j} \right]. \end{aligned}$$

Note that only the θ_i 's are functions of the β_j 's. The y_i 's and ϕ do not depend on the β_j 's.

The connection between the θ_i 's and β_j 's is a little complicated. The following three equations are the thread:

$$\mu_i = b'(\theta_i),$$

$$g(\mu_i) = \eta_i,$$

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}.$$

The chain rule for differentiation says

$$\frac{\partial}{\partial \beta_j} = \frac{\partial}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}.$$

⁸ The subscript i is applied to $a_i(\phi)$ and $c_i(y_i, \phi)$ to recognize that a weight w_i can be included in the model as shown in Example 5.3 and equation (5.8).

Applying the chain rule, the result is

$$\frac{\partial l(\mathbf{y}; \boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^n \frac{(y_i - \mu_i)x_{ij}}{a_i(\phi)b''(\theta_i)g'(\mu_i)}.$$

Equation (5.7) says that the first two terms in the denominator are the variance of y_i : $\text{Var}[y_i] = a_i(\phi)b''(\theta_i)$. This can be rewritten using the variance function and weight, as shown in equation (5.10), as $\text{Var}[y_i] = (\phi/w_i)V(\mu_i)$, giving

$$\frac{\partial l(\mathbf{y}; \boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^n \frac{w_i(y_i - \mu_i)x_{ij}}{\phi V(\mu_i)g'(\mu_i)} = 0. \quad (5.13)$$

for $j = 0, 1, \dots, k$. Note that $x_{i0} = 1$ because β_0 is the intercept. The β_j 's are included in (5.13) through

$$\mu_i = g^{-1}(\beta_0 + \beta_1x_{i1} + \dots + \beta_kx_{ik}).$$

There are $k+1$ equations with $k+1$ unknowns.

What is going on inside (5.13)? The weighted sum of the differences $(y_i - \mu_i)$ should equal 0. This is the MLE solution for the best coefficients b_0, b_1, \dots, b_k to predict $E[y_i]$.

The weights applied to $(y_i - \mu_i)$ are $w_i x_{ij}/(\phi V(\mu_i)g'(\mu_i))$. Values for w_i are specified by the modeler; for example, w_i = number of claims if y_i represents claim severity. A larger w_i gives more weight to the difference $(y_i - \mu_i)$. A larger variance at data point i as captured by variance function $V(\mu_i)$ reduces the weight placed on the difference $(y_i - \mu_i)$. Note that ϕ has no effect on the solution if it is constant across risks. The $g'(\mu_i)$ in the denominator makes an adjustment for the effect of the link function.

Note that (5.13) works for any distribution in the exponential family defined by (5.5). The variance function $V(\mu_i)$ is the only characteristic of the distribution that is required to make the fit.

Statistical packages have numerical methods to maximize the log-likelihood function (5.12). A technique commonly used is referred to as *iteratively reweighted least squares*. One also sees the name *Fisher Scoring algorithm*. For those interested in the details of these numerical techniques, see Dobson and Barnett (2008); Gill (2000); McCullagh and Nelder (1997); and Nelder and Wedderburn (1972).

Example 5.5. Suppose that a GLM model were built assuming a Poisson distribution, log link $g(\mu_i) = \ln(\mu_i)$, and constant weights $w_i = 1$. The Poisson assumption implies $V(\mu_i) = \mu_i$ and $\phi = 1$. The log link gives $g'(\mu_i) = 1/\mu_i$. Equations (5.13) become

$$\sum_{i=1}^n (y_i - \mu_i)x_{ij} = 0$$

for $j = 0, 1, \dots, k$. The log link means that $\ln \mu_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$, and exponentiating both sides yields $\mu_i = e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}}$. The $k + 1$ nonlinear equations to calculate coefficients b_j are

$$\sum_{i=1}^n (y_i - e^{b_0 + b_1 x_{i1} + \dots + b_k x_{ik}}) x_{ij} = 0.$$

As explained earlier, statistical packages have algorithms to calculate coefficients b_j .

Example 5.6. Linear regression uses least squares to find estimators for coefficients β_0, \dots, β_k . Letting $\mu_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$, then the coefficients are estimated by minimizing

$$G(\mathbf{y}; \boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \mu_i)^2.$$

Taking partial derivatives with respect to β_j gives

$$\frac{\partial G(\mathbf{y}; \boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^n 2(y_i - \mu_i)x_{ij}$$

where $x_{i0} = 1$. Setting the partial derivatives equal to zero yields a system of $k + 1$ equations to solve for the b_j 's:

$$\sum_{i=1}^n (y_i - \mu_i)x_{ij} = 0.$$

These are the same as equation (5.13) if one assumes that (1) weights $w_i = 1$, (2) the link function is the identity implying $g'(x) = 1$, and (3) the distribution is normal and homoskedasticity holds. Requirement (3) means that $V(\mu_i) = 1$ (see Table 5.2) and $\phi = \sigma^2$ is a constant and can be canceled. So, classic linear regression is a special case of a GLM. Weighted least squares is also a special case of (5.13).

5.4.1 Quasi-Likelihood

Here we take a quick look at a generalization of the results of the prior section. Inside the sum in equation (5.13) is the fraction $\frac{(y-\mu)}{\phi V(\mu)}$ where we have dropped the subscripts. Integrating with respect to μ and assigning limits gives

$$Q(\mu, y) = \int_y^\mu \frac{y-t}{\phi V(t)} dt. \quad (5.14)$$

Note that $\partial Q/\partial\mu = \frac{(y-\mu)}{\phi V(\mu)}$. Function $Q(\mu, y)$ is called the *log quasi-likelihood*, though “log” is usually dropped from the name. The quasi-likelihood for dataset y_1, \dots, y_n is

$$Q(\boldsymbol{\mu}, \mathbf{y}) = \sum_{i=1}^n Q_i(\mu_i, y_i). \quad (5.15)$$

A quasi-likelihood function can be created by specifying the variance function $V[\mu]$, whereas a log-likelihood function is based on a specific distribution. The quasi-likelihood function has similar properties to a log-likelihood.

As in a GLM a predictive equation is specified, $g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$, and the quasi-likelihood is maximized to determine the coefficients. Quasi-likelihood models can be fit using algorithms similar to those in GLMs. See McCullagh and Nelder (1997) for further discussion.

Some mean-variance relationships will produce results equivalent to the selection of a distribution. For example, if a constant variance is specified, then the quasi-likelihood model will be equivalent to a GLM with a normal distribution. The reader is encouraged to try it! Let $\phi = \sigma^2$ and $V(t) = 1$ in (5.14), do the integration, and put your answer into (5.15).

Specifying a dispersion parameter $\phi = 1$ and variance function $V[\mu] = \mu$ will generate a Poisson-like model, but with a Poisson model the response variables should be integers, whereas this is not a requirement in the quasi-likelihood model. Another application is the overdispersed Poisson model where the dispersion parameter ϕ is greater than 1. The quasi-likelihood function will specify the mean-variance relationship, $V[\mu] = \mu$, but dispersion parameter ϕ will be estimated from the data.

5.5 Generalized Linear Model Review

Here is a quick summary of a generalized linear model:

- Response variables y_i have a distribution from the exponential family and are independently distributed.
- Predictor variables x_{ij} are combined into linear predictors plus a constant

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}.$$

- Link function $g(x)$ is strictly monotonic and differentiable with inverse function $g^{-1}(x)$.
- The expected values of y_i , $\mu_i = E[y_i]$, are predicted by the equations

$$g(\mu_i) = \eta_i \text{ or } \mu_i = g^{-1}(\eta_i) \text{ for } i = 1, \dots, n.$$

- Coefficients $\beta_0, \beta_1, \dots, \beta_k$ are estimated from data using maximum likelihood estimation.
- The modeler must choose the distribution and link function appropriate for the model.

Table 5.4. *Selling a Homeowners Policy to an Auto Policyholder*

Policy	Insured Age	Gender	Marital Status	Territory	Collision Coverage	Purchased Homeowners
1	30	F	M	T01	Y	1
2	51	M	M	T03	Y	0
3	23	F	S	T06	N	0
4	42	F	M	T11	Y	1
5	60	M	S	T03	N	0
:	:	:	:	:	:	:

5.6 Applications

This section presents three stylized examples that demonstrate basic GLM modeling in insurance.

5.6.1 Modeling Probability of Cross Selling with Logit Link

The probability of selling a homeowners policy to a customer who already has an auto policy with the company is modeled in this section. Five data records from a larger file are shown in Table 5.4. The response variable y_i is the Purchased Homeowners column. If the customer purchased a Homeowners policy then a value of 1 appears, and if he or she did not, a 0 is recorded. There are five predictive variables in this example: insured age, gender, marital status, territory, and collision coverage.

The outcome for response variable y_i is simply 1 or 0. This is a Bernoulli process, with a special case of a binomial when $n = 1$. The goal is to predict $E[y_i]$, which is the probability of renewal. As explained in Section 5.3, logit link is appropriate for this model. With $g(x) = \ln[x/(1 - x)]$ the link maps the interval $(0, 1)$ onto $(-\infty, +\infty)$. The inverse function $g^{-1}(x) = e^y/(1 + e^y)$ maps $(-\infty, +\infty)$ onto $(0, 1)$.

The R code to run the GLM is

```
BuyHomeowners <- glm(Homeowners ~ Age + Gender + Marital
+ Territory + Collision,
family = binomial(link=logit), data = CrossSell)
```

Even without an understanding of R, the code should be decipherable. The estimated coefficients for the five predictor variables as well as goodness-of-fit measures from the R `glm` function were put into an R object that we named `BuyHomeowners`. Table 5.4 was loaded into R under the data table name `CrossSell`. The response

Table 5.5. *Claim Count Data*

Policy	Policy Years	Gender	Territory	Claims
1	5	M	East	0
2	5	F	East	0
3	4	M	East	1
4	3	F	West	1
5	4	F	East	0
6	3	F	West	1
7	5	M	West	0
8	5	M	West	2
9	3	M	East	1
10	2	F	East	1
11	4	M	West	1
12	5	F	West	0

variable `Homeowners` takes on values 0 or 1 as shown in the table, and it is modeled with a binomial distribution and logit link.

5.6.2 Modeling Frequency with an Offset

In this example, a GLM is constructed using only the data in Table 5.5. The goal is to produce a multiplicative model to estimate the expected *annual* claims frequency for any combination of `GENDER` and `TERRITORY`.

The observed number of claims for a policy depends on both the annual claims frequency and the number of years of observation, labeled “policy years” in Table 5.5. For two similar policies, the one with more policy years would be expected to have more claims in total. If a policy has more claims, is that because of a higher annual claims frequency or because there are more years of experience? Because we are trying to estimate annual claims frequency, an offset term is put into the model to account for the varying number of `policy years`. This is explained in detail later.

We select the Poisson distribution to model the number of claims. The log link is the right choice for the link function for two reasons. First, the log link is a natural choice for a Poisson distribution as discussed in the section on links. Its inverse maps $(-\infty, \infty)$ the range of the linear predictor η_i , onto $(0, \infty)$, a range for claim frequency. Second, the log link generates a multiplicative model. The model is then

$$\begin{aligned} \ln(f_i) &= \beta_0 + \beta_1 x_{\text{Gender}(i)} + \beta_2 x_{\text{Territory}(i)}, \quad \text{or} \\ f_i &= e^{\beta_0 + \beta_1 x_{\text{Gender}(i)} + \beta_2 x_{\text{Territory}(i)}} \end{aligned} \tag{5.16}$$

where f_i is the expected annual claims frequency for the i^{th} risk.

GENDER and TERRITORY are categorical variables, and in this simple example, each has two levels or categories. The predictors $x_{Gender(i)}$ and $x_{Territory(i)}$ are binary indicator variables that take on values of either 0 or 1. Choosing the base frequency to be FEMALE and EAST, the predictors take on values

$$x_{Gender(i)} = \begin{cases} 0, & \text{Female}, \\ 1, & \text{Male}, \end{cases}$$

$$x_{Territory(i)} = \begin{cases} 0, & \text{East}, \\ 1, & \text{West}. \end{cases}$$

The expected annual claims frequency for risk i can be written as

$$f_i = e^{\beta_0}(e^{\beta_1})^{x_{Gender(i)}}(e^{\beta_2})^{x_{Territory(i)}}.$$

A FEMALE in EAST territory has expected frequency e^{β_0} . The multiplying factor for MALE is e^{β_1} . The factor for WEST territory is e^{β_2} .

5.6.2.1 Offset

The number of claims is a function of the number of policy years, which varies in Table 5.5. The observed annual claims frequencies are (claims/years). Letting y_i be a random variable representing the total number of claims for risk i and m_i be the measure of exposure – in this case the number of policy years – then $f_i = E[y_i/m_i]$. Substituting into equation (5.16),

$$\ln(f_i) = \ln(E[y_i/m_i]) = \beta_0 + \beta_1 x_{Gender(i)} + \beta_2 x_{Territory(i)}.$$

Moving $\ln(m_i)$ to the right-hand side, the log of exposure becomes part of the linear predictor,

$$\ln(E[y_i]) = \beta_0 + \beta_1 x_{Gender(i)} + \beta_2 x_{Territory(i)} + \ln(m_i).$$

The offset term $\ln(m_i)$ is a known effect and must be included because the number of claims depends on the number of years of observations. If the offset is left out of the model then risks with more years of experience will be predicted to have higher annual claims frequencies f_i . See Chapter 6 and Yan et al. (2009) for further discussion of offsets.

5.6.2.2 Offset or Weight?

What is wrong with the following argument? “A risk with more policy years of experience will have more credibility and should be given more weight in the prior model. Policy years should go into the GLM as weights rather than offsets.” The fallacy of this reasoning is that response variable y_i is the *total number of claims* for risk i observed during its whole experience period. If the number of policy years m_i

increases, then both $E[y_i]$ and $\text{Var}[y_i]$ will increase. Policy year offsets $\ln(m_i)$ capture these effects in the log-link Poisson model.

The measure of variability, $\text{Var}[y_i]$, in claim counts y_i increases as policy years increase. Putting policy years m_i into the model as weights has the opposite and wrong effect. A bigger weight m_i means that more weight will be given to observation y_i because you have told the model that y_i is less variable. You also neglected to tell the model that $E[y_i]$ will be bigger for risks with larger m_i . The model will identify risks with more experience as poorer risks. So, a risk with many policy years of experience will tend to be classified as a poor risk, and the model will assign a lot of weight to this observation.

Now the `glm` function can be run with Table 5.5 as input. Although statistical software can handle categorical variables, `GENDER` (`FEMALE=0` and `MALE=1`) and `TERRITORY` (`EAST=0`, `WEST=1`) are switched to numeric for this example. Here is the one line of code to run the GLM in R.

```
freqmodel <- glm(Claims ~ Gender + Territory,
                   family = poisson(link=log), data = ClaimCountData,
                   offset=log(Years))
```

The R code is similar to that in the prior section except that we are modeling with a Poisson distribution and log link. We also include an offset `log(Years)` to account for the varying number of experience years across policies. The program will take the natural log of `YEARS` from Table 5.5 and use that as an offset.

The coefficients are $b_0 = -2.2214$, $b_1 = 0.3282$, and $b_2 = 0.4152$.⁹ The expected frequencies are

$$\mu_i = e^{-2.2214}(e^{0.3282})^{x_{Gender(i)}}(e^{0.4152})^{x_{Territory(i)}}.$$

The four possibilities for annual claim frequencies are $\mu(\text{FEMALE}, \text{EAST}) = 0.1085$, $\mu(\text{FEMALE}, \text{WEST}) = 0.1643$, $\mu(\text{MALE}, \text{EAST}) = 0.1506$, and $\mu(\text{MALE}, \text{WEST}) = 0.2281$.

A quick check for reasonability can be performed. For each policy in Table 5.5 multiply the number of years by the appropriate μ_i from the GLM model and add up the 12 results. The total is 8.00, which exactly matches the eight claims shown in the `Claims` column of the table.

For more complete coverage of modeling claim counts see Chapter 4, which discusses regression models that model count data. In addition to the Poisson distribution, it discusses more general distributions including the negative binomial and the zero-inflated Poisson.

⁹ The unknown coefficients β_0 , β_1 , and β_2 have been relabeled as b_0 , b_1 , and b_2 now that specific values have been estimated.

Table 5.6. *Claim Severity Data*

Policy	Industry	Territory	Number of Claims	Total \$ Claims	Average Claim
1	Retail	B	1	850	850
2	Retail	B	1	2,070	2,070
3	Service	C	3	5,430	1,810
4	Office	A	2	400	200
5	Restaurant	D	1	1,100	1,100
6	Contract	C	2	18,560	9,280
:	:	:	:	:	:

5.6.3 Modeling Severity with Weights

This section presents a simple example of a claim severity model. It uses claim severity and average claim cost interchangeably. The goal is to predict the expected cost of claims based on risk characteristics. Table 5.6 shows a small extract from a file of claim data. We want to build a multiplicative model so a log-link function is appropriate. Because average claim cost is positive, a log link is reasonable.

We chose a gamma distribution to model claim severity. Although it may not be a perfect model for claims severity, it is not necessary to exactly model the severity distribution because we are trying to estimate only the expected cost of a claim – the mean of the distribution. The key feature of the gamma distribution function that is used to fit the model is its variance function: $V(\mu_i) = \mu^2$. The variance in the claim size is proportional to the square of its expected value as shown in Table 5.2.

5.6.3.1 Weights

If random variable y_i is the average claim for risk i , then $\text{Var}[y_i]$ depends on the number of claims used to estimate the average. Let u_{ij} be the amount of the j^{th} claim for risk i . If risk i has w_i claims then $y_i = (u_{i1} + u_{i2} + \dots + u_{iw_i})/w_i$. Assuming that individual claims are i.i.d. for given risk i , then $\text{Var}[y_i] = \text{Var}[u_{ij}]/w_i$ for any j . In GLMs the weight given to a data point is inversely proportional to the variance in the response variable. An observed claim severity computed from w_i claims should be given w_i times the weight of a claim severity that is estimated using just one claim if the expected cost is the same.

Table 5.6 contains a portion of the data file called “SeverityData” shown in the following code. The response variable is AVERAGE CLAIM. Predictors are INDUSTRY and TERRITORY. Weights are NUMBER OF CLAIMS. The R code to run the GLM is

Note the similarities and differences of this R code with that in the prior section where we modeled frequency. In this example we use a gamma distribution rather than Poisson for *family* and use *weights* rather than *offset*. Another difference is that this example has categorical variables `INDUSTRY` and `TERRITORY`, whereas in the frequency example `GENDER` and `TERRITORY` are binary variables.

5.6.4 Modeling Pure Premiums or Loss Ratios

A pure premium is expressed as *pure premium* = *losses/exposure*. A loss ratio is similar, but with premium in the denominator: *loss ratio* = *losses/premium*. One can use similar techniques to model both pure premiums and loss ratios, with the caveat that the modeler must take care with loss ratios to ensure that the premiums are all on the same rate level, usually the current rate level. There are two approaches to modeling pure premiums and loss ratios. One approach is to model frequency and severity separately as done in prior sections and then combine the results. A second approach is to model losses directly.

5.6.4.1 Model Frequency and Severity Separately

Both pure premiums and loss ratios can be split into frequency and severity components:

$$\text{pure premium} = \text{freq} \times \text{severity} = \left(\frac{\text{number of losses}}{\text{exposure}} \right) \times \left(\frac{\$ \text{losses}}{\text{number of losses}} \right).$$

The frequency and severity components can be modeled separately, and then the predicted frequencies and severities can be multiplied together to produce pure premiums. For loss ratios the denominator for frequency is premium, so frequency would be number of losses per unit of premium. Note that if the loss ratio is modeled, the offset in the frequency model as in Section 5.6.2 would be premium, the measure of exposure to loss.

Frequency and severity may be affected differently by risk characteristics so coefficients and predictors in frequency and severity models may be different. Modeling frequency and severity separately may provide more insight into the loss process than modeling the losses directly.

5.6.4.2 Model Losses Directly

A compound Poisson-gamma distribution is often used to model aggregate losses. If the number of losses n is Poisson distributed and loss amounts u_j are i.i.d. gamma distributed, then aggregate loss y can be written as

$$y = u_1 + u_2 + \cdots + u_n.$$

Frequency and severity are combined into a single distribution and model.

In GLM modeling the compound Poisson-gamma is often referred to as a Tweedie distribution. Tweedie family distributions are members of the exponential family with variance functions $V(\mu) = \mu^p$ where p can have values in $(-\infty, 0] \cup [1, \infty)$; that is, p cannot be in the interval $(0,1)$ but all other values work. The Poisson distribution with $V(\mu) = \mu^1$ and gamma distribution with $V(\mu) = \mu^2$ are in the Tweedie family, as are the normal and inverse-Gaussian.

The compound Poisson-gamma has been shown to be a Tweedie distribution with variance function

$$V[\mu] = \mu^p \text{ with } 1 < p < 2,$$

where exponent p is determined by shape parameter α of the gamma according to the formula: $p = \frac{\alpha+2}{\alpha+1}$ (see Jørgensen 1987; Smyth and Jørgensen 2002). The distribution has a point mass at $y = 0$ and continuous density for $y > 0$.

GLM modeling with the Tweedie is complicated by the fact that relative frequency and severity components may vary across risks. Some risks may have high frequency and low severity, whereas other risks have low frequency and high severity such that variance $\text{Var}[y]$ has a more complicated relationship to mean μ .

One approach is to model the variance of response variables y_i as

$$\text{Var}[y_i] = \phi_i \mu^p,$$

allowing dispersion parameter ϕ_i to vary across risks but keeping p constant. A fixed p means that shape parameter α for the gamma severity distributions is constant across risks i . The coefficient of variation for the gamma is $1/\sqrt{\alpha}$, so a constant shape parameter is equivalent to a constant c.v. for severity. This model is described in Smyth and Jørgensen (2002). Because of the complications mentioned earlier, not all GLM software packages are capable of modeling a Tweedie distribution.

5.6.5 Generalized Linear Models and Classification Ratemaking

GLMs allow modelers to include many predictors in one model and consider all of their effects simultaneously. Traditionally, actuaries might have looked at rating variables one at a time: they would (1) build a class plan, (2) compute territorial relativities, (3) compute construction relativities, and so on. Using these ad hoc, piecemeal approaches the actuary had to be careful that one rating factor was not being influenced by a heterogeneous distribution of other rating factors. For example, data might imply that a particular territorial relativity should be high, but a closer look could reveal that this finding was being caused by an abundance of some other high-risk characteristics in the territory, such as a lack of fire protection. A properly constructed GLM gives each factor its due.

5.7 Comparing Models

With a variety of distributions unified into one framework under the label GLM, analyzing goodness of fit and comparing models can be challenging. The log-likelihood is used to fit models to data and is also useful in making model comparisons.

5.7.1 Deviance

A statistical measure called *deviance* is commonly used to evaluate and compare GLMs and it is based on log-likelihoods. The log-likelihood function for linear exponential family distributions is

$$l(\mathbf{y}; \boldsymbol{\theta}) = \sum_{i=1}^n \left[\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c_i(y_i, \phi) \right],$$

where vector $\boldsymbol{\theta}$ represents the n canonical parameters $\theta_1, \theta_2, \dots, \theta_n$. When a particular GLM is constructed – let's call it model M – coefficients b_j are calculated to maximize the log-likelihood. As explained earlier in the chapter, canonical parameters can be computed using these coefficients and predictive variables $x_{i1}, x_{i2}, \dots, x_{ik}$ for the chosen distribution and link function. If a canonical link is chosen for Model M then the canonical parameters have the simple linear form $\hat{\theta}_i^M = b_0 + b_1 x_{i1} + \dots + b_k x_{ik}$ for $i = 1, \dots, n$; otherwise the function is nonlinear. Let $l(\mathbf{y}; \boldsymbol{\theta}^M)$ denote the value of the log-likelihood for these parameters $\hat{\theta}_i^M$.

If the n parameters θ_i were free to take on any values, what values would maximize the log-likelihood? The log-likelihood will be maximized if each term of the sum is maximized. Take the partial derivative of term i with respect to θ_i and set it equal to zero:

$$\frac{\partial}{\partial \theta_i} \left[\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c_i(y_i, \phi) \right] = y_i - b'(\theta_i) = 0.$$

The canonical parameters will satisfy $y_i = b'(\hat{\theta}_i^S)$ where the superscript S indicates that we have a *saturated model* with a separate parameter $\hat{\theta}_i^S$ for each observation y_i . The mean of a distribution is $\mu = b'(\theta)$ (see Appendix B) so in the saturated model the estimated mean $\hat{\mu}_i^S$ of distribution i equals the value of observation i : $\hat{\mu}_i^S = b'(\hat{\theta}_i^S) = y_i$.¹⁰

¹⁰ The conclusion that the log-likelihood will be maximized by $\hat{\mu}_i^S = y_i$ is also apparent from equation (5.13). Note that all of the $(y_i - \mu_i)$ terms in the sum are zero when $\mu_i = y_i$. The log-likelihood can also be written as a function of the distribution means in place of the canonical parameters: $l(\mathbf{y}; \boldsymbol{\mu})$. In this form the saturated model log-likelihood $l(\mathbf{y}; \hat{\boldsymbol{\mu}}^S)$ can be written simply as $l(\mathbf{y}; \mathbf{y})$.

The difference between the log-likelihoods of the saturated model S and model M is

$$l(\mathbf{y}; \hat{\boldsymbol{\theta}}^S) - l(\mathbf{y}; \hat{\boldsymbol{\theta}}^M) = \sum_{i=1}^n \left[\frac{y_i(\hat{\theta}_i^S - \hat{\theta}_i^M) - (b(\hat{\theta}_i^S) - b(\hat{\theta}_i^M))}{a_i(\phi)} \right].$$

This quantity is non-negative because $l(\mathbf{y}; \hat{\boldsymbol{\theta}}^S) \geq l(\mathbf{y}; \hat{\boldsymbol{\theta}}^M)$. If $a_i(\phi) = \phi/w_i$ then *scaled deviance* $D^*(\mathbf{y}; \hat{\boldsymbol{\theta}}^M)$ for model M is defined as

$$\begin{aligned} D^*(\mathbf{y}; \hat{\boldsymbol{\theta}}^M) &= 2[l(\mathbf{y}; \hat{\boldsymbol{\theta}}^S) - l(\mathbf{y}; \hat{\boldsymbol{\theta}}^M)] \\ &= \frac{1}{\phi} \sum_{i=1}^n 2w_i [y_i(\hat{\theta}_i^S - \hat{\theta}_i^M) - (b(\hat{\theta}_i^S) - b(\hat{\theta}_i^M))]. \end{aligned}$$

Deviance $D(\mathbf{y}; \hat{\boldsymbol{\theta}}^M)$ drops the dispersion parameter ϕ : $D(\mathbf{y}; \hat{\boldsymbol{\theta}}^M) = \phi D^*(\mathbf{y}; \hat{\boldsymbol{\theta}}^M)$. The deviance for the saturated model is 0: $D(\mathbf{y}; \hat{\boldsymbol{\theta}}^S) = 0$.

Example 5.7. Model M , was constructed with n response variables y_i and corresponding $n \times k$ predictive variables $x_{i1}, x_{i2}, \dots, x_{ik}$. Responses y_i were assumed to be normally distributed: $y_i \sim N(\mu_i, \sigma^2)$. An identity link was selected and no weights were applied. The pdf's for the response variables can be written as

$$f(y_i; \theta_i, \phi) = \exp \left[\frac{y_i \theta_i - \theta_i^2/2}{\phi} - \frac{1}{2} \left(\frac{y_i^2}{\phi} + \ln(2\pi\phi) \right) \right]$$

with $\theta_i = \mu_i$ and $\phi = \sigma^2$.

With the identity link the predicted means for the n response variables in Model M are $\hat{\mu}_i^M = b_0 + b_1x_{i1} + \dots + b_kx_{ik}$. With a normal distribution the identity link is the canonical link so $\hat{\theta}_i^M = \hat{\mu}_i^M$. In saturated model S the fitted means are the actual observations: $\hat{\theta}_i^S = \hat{\mu}_i^S = y_i$. The deviance for model M is

$$\begin{aligned} D(\mathbf{y}; \hat{\boldsymbol{\theta}}^M) &= \sum_{i=1}^n 2[y_i(y_i - \hat{\mu}_i^M) - (y_i^2/2 - (\hat{\mu}_i^M)^2/2)] \\ &= \sum_{i=1}^n (y_i - \hat{\mu}_i^M)^2. \end{aligned}$$

The deviance is the residual sum of squares, a goodness of fit measure used in linear regression.

Example 5.8. Generalized linear model Q assumed Poisson distributions $f(y_i) = \mu_i^{y_i} e^{-\mu_i} / y_i!$ for response variables y_i , log links, and no weights. The distribution for

response variables y_i can be written as

$$f(y_i; \theta_i) = \exp \left[\frac{y_i \theta_i - e^{\theta_i}}{1} - \ln y_i! \right]$$

with $\theta_i = \ln \mu_i$ and distribution means μ_i . In the saturated model $\hat{\mu}_i^S = y_i$ so $\hat{\theta}_i^S = \ln y_i$ in the deviance formula. The fitted parameters are $\hat{\theta}_i^Q = \ln \hat{\mu}_i^Q$. The deviance for model Q is

$$D(\mathbf{y}; \hat{\theta}^Q) = \sum_{i=1}^n 2[y_i(\ln y_i - \ln \hat{\mu}_i^Q) - (y_i - \hat{\mu}_i^Q)].$$

with $\ln \hat{\mu}_i^Q = b_0 + b_1 x_{i1} + \dots + b_k x_{ik}$ for $i = 1, \dots, n$. Note that the deviance goes to zero as fitted means $\hat{\mu}_i^Q$ approach observations y_i .

The output from the `glm` function in R displays null deviance and residual deviance. Null deviance is the residual deviance for a GLM with only a constant term. The null deviance provides an upper bound on residual deviance; 0, the deviance of the saturated model, is the lower bound: $0 \leq D^*(\mathbf{y}; \hat{\theta}^M) \leq D^*(\mathbf{y}; \hat{\theta}^{Null})$.

The residual deviance can be used to compare two nested models. Suppose model P has p explanatory variables, x_{i1}, \dots, x_{ip} .¹¹ Suppose that model Q has q explanatory variables that include all of the explanatory variables of model P plus $q - p > 0$ additional explanatory variables. The difference in the residual deviances is equivalent to a likelihood-ratio statistic:

$$D^*(\mathbf{y}; \hat{\theta}^P) - D^*(\mathbf{y}; \hat{\theta}^Q) = 2[l(\mathbf{y}; \hat{\theta}^Q) - l(\mathbf{y}; \hat{\theta}^P)] = 2 \ln \frac{L(\mathbf{y}; \hat{\theta}^Q)}{L(\mathbf{y}; \hat{\theta}^P)}.$$

This statistic has an approximate χ^2 distribution with $q - p$ degrees of freedom.

Example 5.9. In Section 5.6.1 a model was created to predict the probability that an auto insurance policyholder would also buy a homeowners policy. Suppose that the actuary has built two models and wants to compare:

$$\text{Model } P: g(\mu_i) = \beta_0 + \beta_1 \text{AGE}_i + \beta_2 \text{GENDER}_i$$

$$\begin{aligned} \text{Model } Q: g(\mu_i) = \beta_0 + \beta_1 \text{AGE}_i + \beta_2 \text{GENDER}_i + \beta_3 \text{MARITALSTATUS}_i \\ + \beta_4 \text{COLLCOVERAGE}_i. \end{aligned}$$

Model P is more restrictive than model Q in that model P says β_3 and β_4 must be 0, whereas coefficients β_3 and β_4 can have nonzero values in model Q . Setting this up as a hypothesis test, the null hypothesis is $H_0: \beta_3 = \beta_4 = 0$. The alternative hypothesis is H_1 : At least one of β_3 or β_4 is nonzero.

¹¹ The model will have $p + 1$ parameters b_0, b_1, \dots, b_p .

The test statistic is $T = D^*(\mathbf{y}; \hat{\boldsymbol{\theta}}^P) - D^*(\mathbf{y}; \hat{\boldsymbol{\theta}}^Q)$. The critical value for a χ^2 test with two degrees of freedom with an $\alpha = 5\%$ significance level is $c = 5.99$. The null hypothesis will be rejected if $T > 5.99$. If one chooses model Q instead of model P based on this χ^2 test, then the change in the deviance has to be large enough to justify the choice, in this case, $T > 5.99$.

5.7.2 Log-Likelihood, AIC, AICC, and BIC

Deviance is a useful measure for comparing nested models as described in the prior section, but if models are not nested then the χ^2 test may not apply. This section discusses several statistics for comparing non-nested models that may be displayed by GLM software. In the previous section we defined $l(\mathbf{y}; \boldsymbol{\theta}^M)$ as the maximum value of the log-likelihood function for model M with its set of predictor variables.

By itself, the value of $l(\mathbf{y}; \boldsymbol{\theta}^M)$ may not reveal much about model fit, but generally, bigger is better. However, there is a problem with a direct comparison of log-likelihood measures between models. If one model includes more predictive variables than another, then there is a good chance that the model with more predictors will have a larger value for $l(\mathbf{y}; \boldsymbol{\theta}^M)$. To address this problem, three measures are often used with GLMs to adjust for the number of estimated parameters in a model.

AIC is an abbreviation for Akaike information criterion. It is the simplest of the three with the formula

$$\text{AIC} = 2[-l(\mathbf{y}; \boldsymbol{\theta}^M) + r],$$

where r is the number of fitted parameters in the model. Smaller values for *AIC* indicate a better fit.

AICC (or *AICc*) adjusts *AIC* for finite sample size n and can be written as

$$\text{AICC} = \text{AIC} + \frac{2r(r+1)}{n-r-1}.$$

As the sample size becomes large *AICC* approaches *AIC*. In many insurance applications that include individual policies or claims, the *AIC* and *AICC* may be essentially the same because n is large.

BIC is short for Bayesian information criterion. It is also referred to as SBC, the Bayesian criterion. Like *AIC* and *AICC* it makes an adjustment to the log-likelihood:

$$\text{BIC} = 2[-l(\mathbf{y}; \boldsymbol{\theta}^M) + r \ln(n)].$$

In addition to these three measures, there are other measures that make adjustments to the log-likelihood for number of parameters and sample size.

5.8 Conclusion

The goal of this chapter is to provide an introduction to generalized linear models. Many useful topics were not covered, and for those topics that were addressed much was left out. This chapter is only a starting point for those who will use GLMs.

GLMs greatly extend the power of regression models. Data that actuaries deal with constantly such as claim frequencies, loss amounts, pure premiums, and loss ratios can be response variables in GLMs. The confining assumptions of classic linear regression such as normality, homoskedasticity, and linearity are eliminated.

GLMs allow modelers to combine many predictors into one model and consider all of their effects simultaneously, thereby making it more efficient to construct actuarial models while improving the accuracy of models. Though this chapter has focused on ratemaking, GLMs can be applied in many other areas of interest to actuaries, including reserving, marketing, underwriting, and claims analysis.

5.9 Appendix A. Binomial and Gamma Distributions in Exponential Family Form

Appendix A Preview. The binomial and gamma distributions are put into exponential family form.

Binomial. Let $y \sim B(n, p)$, then its pdf is

$$\begin{aligned} f(y; n, p) &= \binom{n}{y} p^y (1-p)^{(n-y)} \\ &= \exp \left[\ln \left[\binom{n}{y} p^y (1-p)^{(n-y)} \right] \right] \\ &= \exp \left[\ln \left(\binom{n}{y} \right) + y \ln(p) + (n-y) \ln(1-p) \right] \\ &= \exp \left[\frac{y \ln \left(\frac{p}{1-p} \right) - [-n \ln(1-p)]}{1} + \ln \left(\binom{n}{y} \right) \right]. \end{aligned}$$

This is the exponential family form with $\theta = \ln \left(\frac{p}{1-p} \right)$, $b(\theta) = -n \ln(1-p)$, $a(\phi) = 1$, and $c(y) = \ln \left(\binom{n}{y} \right)$. It assumed that n is a known and fixed quantity. Inverting $\theta = \ln \left(\frac{p}{1-p} \right)$ gives $p = e^\theta / (1 + e^\theta)$. So, $b(\theta) = n \ln(1 + e^\theta)$.

If y is assumed to have a binomial distribution in a GLM, then p is the value that the model is trying to predict. Or, equivalently, the model is predicting $E[y] = np$ where n is known. If the link function is the logit link, $g(p) = \ln[p/(1-p)]$, then the predictive equations will have the very simple form $\theta_i = \eta_i$, where η_i is the

linear predictor. The logit link is referred to as the canonical link for the binomial distribution.

The mean for any distribution in exponential family form is $b'(\theta)$. Tying this back to familiar parameters n and p ,

$$\begin{aligned} b'(\theta) &= \frac{d}{d\theta}[n \ln(1 + e^\theta)] \\ &= n \frac{e^\theta}{1 + e^\theta} \\ &= np. \end{aligned}$$

The variance is $a(\phi)b''(\theta) = np(1 - p)$.

Gamma. Let y have a gamma distribution; then its pdf is

$$f(y; \alpha, \beta) = \frac{\beta^\alpha y^{\alpha-1} e^{-\beta y}}{\Gamma(\alpha)}.$$

The mean of the distribution is $E[y] = \alpha/\beta$ and the variance is $\text{Var}[y] = \alpha/\beta^2$. Changing parameters in the distribution $\mu = \alpha/\beta$ and $v = \alpha$ gives

$$\begin{aligned} f(y; \mu, v) &= \frac{1}{\Gamma(v)} \left(\frac{v}{\mu}\right)^v y^{v-1} e^{-vy/\mu} \\ &= \exp[-\ln \Gamma(v) + v \ln v - v \ln \mu + (v-1) \ln y - vy/\mu] \\ &= \exp \left[\frac{y \left(\frac{-1}{\mu}\right) - \ln \mu}{1/v} + (v-1) \ln y - \ln \Gamma(v) + v \ln v \right]. \end{aligned}$$

This is exponential family form with $\theta = -1/\mu$, $b(\theta) = \ln \mu$, $a(\phi) = 1/v$, and $c(y, v) = (v-1) \ln y - \ln \Gamma(v) + v \ln v$.

Because $\theta = -1/\mu$ and $\mu = \alpha/\beta$ is positive, it follows that $\theta < 0$. Substituting $-1/\theta$ in for μ , then $b(\theta) = \ln(-1/\theta) = -\ln(-\theta)$. Calculating the mean for the distribution,

$$b'(\theta) = \frac{d}{d\theta}(-\ln(-\theta)) = \frac{-1}{\theta} = \mu.$$

The variance is $a(\phi)b''(\theta)$, which is $(1/v)(1/\theta^2)$. Converting back to original parameters this is $\alpha/(\beta^2)$, which is the variance $\text{Var}[y]$.

5.10 Appendix B. Calculating Mean and Variance from Exponential Family Form

Appendix B Preview. After a brief discussion of the score function, the mean and variance are derived for distributions that can be written in the exponential family form assumed for GLM.

The mean and variance for distributions whose pdfs can be written as

$$f(y; \theta, \phi) = \exp \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right]$$

are

$$E[y] = b'(\theta) \text{ and } \text{Var}[y] = a(\phi)b''(\theta).$$

These formulas can be derived several ways. We follow a common derivation that is shown in McCullagh and Nelder (1997).

Maximum likelihood estimation finds parameters that maximize the log of the likelihood function. The log-likelihood function for one observation of random variable y with pdf $f(y; \theta, \phi)$ is $l(\theta, \phi; y) = \ln(f(y; \theta, \phi))$; the log-likelihood is a function of θ and ϕ given observation y . The log-likelihoods for functions in the exponential family assumed in GLMs have the convenient form

$$l(\theta, \phi; y) = \ln \left(\exp \left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right] \right) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi).$$

The score function, sometimes called score, is the derivative of the log-likelihood function with respect to a parameter of interest. In this case the parameter of interest is θ and the score function is

$$\frac{\partial l(\theta, \phi; y)}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}.$$

The score function is used to compute maximum likelihood estimates and also in approximating the error in maximum likelihood estimates. Two properties of the score function that we use without proof are

$$E_y \left[\frac{\partial l(\theta, \phi; y)}{\partial \theta} \right] = 0, \quad \text{Identity 1.}$$

$$E_y \left[\frac{\partial^2 l(\theta, \phi; y)}{\partial \theta^2} \right] + E_y \left[\left(\frac{\partial l(\theta, \phi; y)}{\partial \theta} \right)^2 \right] = 0. \quad \text{Identity 2.}$$

Using identity 1 with the score function for the exponential family pdf given earlier,

$$\begin{aligned} 0 &= E_y \left[\frac{\partial l(\theta, \phi; y)}{\partial \theta} \right] \\ &= E_y \left[\frac{y - b'(\theta)}{a(\phi)} \right] \\ &= \frac{E[y] - b'(\theta)}{a(\phi)}. \end{aligned}$$

The expected value of y follows immediately: $E[y] = b'(\theta)$.

Substituting the score function above into Identity 2,

$$\begin{aligned} E_y \left[\left(\frac{\partial l(\theta, \phi; y)}{\partial \theta} \right)^2 \right] &= -E_y \left[\frac{\partial^2 l(\theta, \phi; y)}{\partial \theta^2} \right] \\ E_y \left[\left(\frac{y - b'(\theta)}{a(\phi)} \right)^2 \right] &= -E_y \left[\frac{\partial}{\partial \theta} \left(\frac{y - b'(\theta)}{a(\phi)} \right) \right] \\ E_y \left[\frac{y^2 - 2yb'(\theta) + (b'(\theta))^2}{(a(\phi))^2} \right] &= -E_y \left[\frac{-b''(\theta)}{a(\phi)} \right] \\ E[y^2] - 2E[y]b'(\theta) + (b'(\theta))^2 &= a(\phi)b''(\theta) \\ E[y^2] - (E[y])^2 &= a(\phi)b''(\theta) \\ \text{Var}[y] &= a(\phi)b''(\theta). \end{aligned}$$

This is the variance for distributions that can be put into the linear exponential family form.

References

- Anderson, D., S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher, and N. Thandi (2005). *A Practitioner's Guide to Generalized Linear Models: A Foundation for Theory, Interpretation, and Application*. Chapman & Hall/CRC.
- Dobson, A. J. and A. G. Barnett (2008). *An Introduction to Generalized Linear Models* (Third ed.). Chapman & Hall/CRC.
- Gill, J. (2000). *Generalized Linear Models: A Unified Approach*. Sage University Papers Series on Quantitative Applications in the Social Sciences, series no. 07-134. Sage Publications.
- Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)* 49(2), 127–162.
- Klugman, S. A., H. H. Panger, and G. E. Willmot (2008). *Loss Models: From Data to Decisions* (Third ed.). Wiley.
- McCullagh, P. and J. Nelder (1997). *Generalized Linear Models* (Second ed.). Chapman & Hall/CRC.

- Mildenhall, S. J. (1999). A systematic relationship between minimum bias and generalized linear models. *Proceedings of the Casualty Actuarial Society LXXXVI*, 393–487.
- Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)* 135(3), 370–384.
- Smyth, G. K. and B. Jørgensen (2002, May). Fitting Tweedie’s compound Poisson model to insurance claims data: Dispersion modelling. *ASTIN Bulletin* 32(2), 143–157.
- Yan, J., J. Guszcza, M. Flynn, and C.-S. P. Wu (Winter 2009). Applications of the offset in property-casualty predictive modeling. *Casualty Actuarial Society E-Forum*, 366–385.

6

Frequency and Severity Models

Edward W. Frees

Chapter Preview. Many insurance datasets feature information about frequency, how often claims arise, in addition to severity, the claim size. This chapter introduces tools for handling the joint distribution of frequency and severity. Frequency-severity modeling is important in insurance applications because of features of contracts, policyholder behavior, databases that insurers maintain, and regulatory requirements. Model selection depends on the data form. For some data, we observe the claim amount and think about a zero claim as meaning no claim during that period. For other data, we observe individual claim amounts. Model selection also depends on the purpose of the inference; this chapter highlights the Tweedie generalized linear model as a desirable option. To emphasize practical applications, this chapter features a case study of Massachusetts automobile claims, using out-of-sample validation for model comparisons.

6.1 How Frequency Augments Severity Information

At a fundamental level, insurance companies accept premiums in exchange for promises to indemnify a policyholder on the uncertain occurrence of an insured event. This indemnification is known as a *claim*. A positive amount, also known as the *severity*, of the claim, is a key financial expenditure for an insurer. One can also think about a zero claim as equivalent to the insured event not occurring. So, knowing only the claim amount summarizes the reimbursement to the policyholder. Ignoring expenses, an insurer that examines only amounts paid would be indifferent to two claims of 100 when compared to one claim of 200, even though the number of claims differs.

Nonetheless, it is common for insurers to study how often claims arise, known as the frequency of claims. Let us think about reasons why an insurance analyst should be concerned with models of frequency as well as severity.

Contractual. It is common for insurance contracts to impose deductibles and policy limits on a per occurrence basis. For example, if the policy has a deductible of 100 per occurrence, then two losses of 100 would result in a payout (or claim) of zero from the insurer, whereas a single loss of 200 would result in a payout of 100. Models of total insured losses need to account for deductibles and policy limits for each insured event.

Behaviorial. Models of insurance losses implicitly or explicitly account for decisions and behavior of people and firms that can affect losses; these decision makers can include not only the policyholder but also the insurance adjuster, repair specialist, medical provider, and so forth. Behavioral explanatory (rating) variables can have different effects on models of how often an event occurs in contrast to the size of the event.

For example, in homeowners insurance, consider a very careful policyholder who lives in an expensive neighborhood. We might look to characteristics of the homeowner as an indication of the introduction of loss prevention measures (e.g., sprinklers) as determinants that suggest low frequency. In contrast, we might look to the overall income level of the geographic area where the house is located as a proxy for the level of repair costs in the event of an accident, suggesting high severity.

In health care, the decision to utilize health care by individuals is related primarily to personal characteristics, whereas the cost per user may be more related to characteristics of the health care provider (such as the physician).

In automobile insurance, we might think of population density as positively correlated with the frequency of accidents and negatively associated with severity. For example, in a densely populated urban area, the traffic congestion is high, meaning that drivers are likely to have frequent, but relatively low-cost, accidents. This is in contrast to a more sparsely populated rural area where there is an opportunity to drive speedily. Less congestion may mean less frequent accidents, but greater speeds mean greater severity.

Prior claims history is another variable that provides information about a policyholder's risk appetite. Especially in personal lines, it is common to use an indicator of whether or not a claim has occurred in, for example, the last three years, rather than the claim amount. (Claim amounts are commonly used in commercial lines through credibility calculations). In many countries, automobile premiums are adjusted by a so-called bonus-malus system where prior claim frequency is used to dynamically adjust premiums.

Databases. Many insurers keep separate data files that suggest the development of separate frequency and severity models. For example, insurers maintain a "policyholder" file that is established when a policy is written. This file records much underwriting information about the insured(s), such as age, gender, and prior claims experience; policy information such as coverage, deductibles, and limitations; and information

about the insurance claims event. A separate file, often known as the “claims” file, records details of the claim against the insurer, including the amount. (There may also be a “payments” file that records the timing of the payments, although we do not deal with that here.) This recording process makes it natural for insurers to model the frequency and severity as separate processes.

Regulatory and Administrative. Insurance is a closely monitored industry sector. Regulators routinely require the reporting of both claims numbers and amounts. This may be due to the fact that there can be alternative definitions of an “amount” (e.g., paid versus incurred), and there is less potential error when reporting claim numbers.

At a broad level, it is clear that insurers need very different administrative systems for handling small, frequently occurring, reimbursable losses (e.g., prescription drugs) versus rare occurrence, high-impact events, such as inland marine. Every insurance claim means that the insurer incurs additional expenses, suggesting that claims frequency is an important determinant of expenses.

There are considerable differences of opinion concerning the importance of frequency models for allocated loss adjustment expenses (ALAE), costs that can be associated with a specific claim (e.g., legal defense fees and claims adjuster costs). According to Werner and Modlin (2010), it is common to assume that ALAE vary by the amount of the claim rather than by frequency.

6.2 Sampling and the Generalized Linear Model

6.2.1 Sampling

For a sampling basis, begin by thinking about the policyholder and claims databases that an insurer maintains. An insurer enters new contracts with insureds and administers claims continuously over time. For some purposes, it is helpful to consider a continuous-time stochastic process as a sampling model; this is the perspective of the loss reserving described in Chapter 18 and the survival modeling presented in Chapter 19 (where we examine processes that influence policy retention). In contrast, this chapter focuses on collections of policies without an emphasis on the exact calendar start date of the policy. In particular, it only considers closed claims, leaving issues of claim development to those chapters. Ratemaking and reinsurance are the primary purposes of this chapter rather than reserving and policy retention.

To establish notation, for each policy $\{i\}$, the potentially observable responses are

- N_i – the number of claims (events),
- y_{ij} , $j = 1, \dots, N_i$ – the amount of each claim (loss), and
- $S_i = y_{i1} + \dots + y_{iN_i}$, the aggregate claim amount.

By convention, the set $\{y_{ij}\}$ is empty when $N_i = 0$.

For a specific accounting period (such as a year), the sample of observable responses may consist of the following:

- (1) S_i , so that only aggregate losses are available. For example, when examining losses for commercial insurance, it is common that only aggregate losses are available.
- (2) (N_i, S_i) , so that the number and amount of aggregate losses are available.
- (3) $(N_i, y_{i1}, \dots, y_{i,N_i})$, so that detailed information about each claim is available. For example, when examining personal automobile claims, losses for each claim are available. Let $\mathbf{y}_i = (y_{i1}, \dots, y_{i,N_i})'$ be the vector of individual losses.

We can use ideas from conditional probability to decompose the distribution into frequency and severity components. To be specific, consider the third data form. Suppressing the $\{i\}$ subscript, we decompose the distribution of the dependent variables as

$$\begin{aligned} f(N, \mathbf{y}) &= f(N) \times f(\mathbf{y}|N) \\ \text{joint} &= \text{frequency} \times \text{conditional severity}, \end{aligned}$$

where $f(N, \mathbf{y})$ denotes the joint distribution of (N, \mathbf{y}) . This joint distribution equals the product of the two components:

1. claims frequency: $f(N)$ denotes the probability of having N claims, and
2. conditional severity: $f(\mathbf{y}|N)$ denotes the conditional density of the claim vector \mathbf{y} given N .

The second data form follows similarly, replacing the vector of individual losses \mathbf{y} with the aggregate loss S . We can even decompose the first data form by breaking off the zero event through the indicator notation $r_i = I(S_i > 0)$ for the frequency component and conditioning on $r_i = 1$ for the severity component. We examine this data form using two-part models in Section 6.3.1.

Through this decomposition, we do *not* require independence of the frequency and severity components as is traditional in the actuarial science literature. There are many ways to model dependence when considering the joint distribution $f(N, \mathbf{y})$. For example, one may use a latent variable that affects both frequency N and loss amounts \mathbf{y} , thus inducing a positive association. Copulas are another tool used regularly by actuaries to model nonlinear associations. The conditional probability framework is a natural method of allowing for potential dependencies and provides a good starting platform for empirical work.

6.2.2 Generalized Linear Model

A natural starting platform for empirical modeling of both frequency and severity components is the generalized linear model (GLM) introduced in Chapter 5. Indeed,

one reason for the popularity of this modeling framework is that it has the flexibility to address both frequency and severity models.

Thinking of a generic dependent variable y_i (without regard to whether it represents frequency or severity), we focus on logarithmic links so that the mean function is $E y_i = \exp(\mathbf{x}_i' \boldsymbol{\beta})$.

In some instances, the mean is known to vary proportionally with a variable that we label as E_i , for “exposure;” see the next subsection for more discussion of exposures. To incorporate exposures, one can always specify one of the explanatory variables to be $\ln E_i$ and restrict the corresponding regression coefficient to be 1; this term is known as an *offset*. With this convention, the link function is

$$\ln \mu_i = \ln E_i + \mathbf{x}_i' \boldsymbol{\beta}.$$

Example 6.1 (Relativities). In this example, we consider a small fictitious dataset that appears in Werner and Modlin (2010). The data consist of loss and loss adjustment expenses (`LossLAE`), decomposed by three levels of an amount of insurance (`AOI`) and three territories (`Terr`). For each combination of `AOI` and `Terr`, we have available the number of policies issued, given as the exposure.

AOI	Terr	Exposure	LossLAE
Low	1	7	210.93
Medium	1	108	4458.05
High	1	179	10565.98
Low	2	130	6206.12
Medium	2	126	8239.95
High	2	129	12063.68
Low	3	143	8441.25
Medium	3	126	10188.70
High	3	40	4625.34

Source: Werner and Modlin, 2010.

Our objective is to fit a generalized linear model (GLM) to the data using `LossLAE` as the dependent variable. We would like to understand the influence of the amount of insurance and territory on `LossLAE`.

We now specify two factors and estimate a generalized linear model using a gamma distribution with a logarithmic link function. In the R output that follows, the “`relevel`” command allows us to specify the reference level. For this example, a medium amount of insurance (`AOI = medium`) and the second territory (`Terr = 2`) are chosen as the reference levels. Logarithmic exposure is used as an offset variable

so that cells (combinations of the two categorical variables) with larger numbers of exposures/policies will have larger expected losses.

Selected R Output

```

> Sampdata$AOI = relevel(Sampdata$AOI, ref = "Medium")
> Sampdata$Terr = factor(Sampdata$Terr)
> Sampdata$Terr = relevel(Sampdata$Terr, ref = "2")
> summary(glm(LossLAE ~ AOI + Terr, offset = log(Exposure),
  data = Sampdata, + family = Gamma(link = "log")))

Call:
glm(formula = LossLAE ~ AOI + Terr, family = Gamma(link = "log"),
  data = Sampdata, offset = log(Exposure))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.180e+00 1.975e-06 2116446   <2e-16 ***
AOIHigh     3.577e-01 2.164e-06   165302   <2e-16 ***
AOILow      -3.147e-01 2.164e-06  -145448   <2e-16 ***
Terr1       -4.601e-01 2.164e-06  -212656   <2e-16 ***
Terr3        2.123e-01 2.164e-06    98109   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1    1

(Dispersion parameter for Gamma family taken to be 7.022767e-12)

Null deviance: 1.3528e+00  on 8  degrees of freedom
Residual deviance: 2.8091e-11  on 4  degrees of freedom
AIC: -47.141

```

Parameter estimates can be readily converted to relativities by exponentiation, as follows:

Variable	Parameter Estimate	Relativity (exponential parameter estimate)
Intercept	4.18	65.366
AOILow	-0.3147	0.730
AOIMedium	0	1
AoIHigh	0.3577	1.430
Terr1	-0.4601	0.631
Terr2	0	1
Terr3	0.2123	1.237

With the relativities and exposures, it is straightforward to compute predictions. For example, for a high amount of insurance in territory 1, the exposure is 179, so the fitted value is $179 \times 65.366 \times 1.430 \times 0.631 = 10,558$. This is close to the actual value 10,565.98.

By comparing all actual to fitted values, or the null to the residual deviance, or examining the t -values or p -values, we see that we have done a pretty amazing job of fitting this data. In fact, these data are artificially constructed by Werner and Modlin to prove that various univariate methods of identifying relativities can do poorly. A multivariate method such as GLM is usually preferred in practice. Recall that the purpose of linear, as well as generalized linear, modeling is to simultaneously fit several factors to a set of data, not each in isolation of the others. As discussed in the following subsection, we should pay attention to the variability when introducing exposures. However, weighting for changing variability is not needed for this artificial example.

6.2.3 Exposures

As illustrated in the prior example, actuaries commonly use the idea of an “exposure” to calibrate the size of a potential loss. This subsection discusses exposure from a statistical perspective. To begin, an *exposure* is a variable that can be used to explain the distribution of losses; that is, it is a rating variable. It is typically the most important rating variable; it is so important that both premiums and losses are quoted on a “per exposure” basis. Here are some examples:

Typical Exposure Bases for Several Lines of Business

Line of Business	Exposure Basis
Personal Automobile	Earned Car Year
Homeowners	Earned House Year
Workers' Compensation	Payroll
Commercial General Liability	Sales Revenue, Payroll, Square Footage, Number of Units
Commercial Business Property	Amount of Insurance Coverage
Physician's Professional Liability	Number of Physician Years
Professional Liability	Number of Professionals (e.g., Lawyers or Accountants)
Personal Articles Floater	Value of Item

Source: Werner and Modlin, 2010.

Naturally, selection of a good exposure base goes beyond statistics. An exposure basis should

- be an accurate measure of the quantitative exposure to loss,
- be easy for the insurer to determine (at the time the policy is calculated) and not subject to manipulation by the insured,
- be easy to understand by the insured and to calculate by the insurer,

- consider any preexisting exposure base established within the industry, and
- for some lines of business, be proportional to inflation. In this way, rates are not sensitive to the changing value of money over time because these changes are captured in exposure base.

To illustrate, consider personal automobile coverage. Instead of the exposure basis “earned car year,” a more accurate measure of the quantitative exposure to loss might be the number of miles driven. However, this measure is difficult to determine at the time the policy is issued and is subject to potential manipulation by the insured.

For frequency and severity modeling, it is customary to think about the frequency aspect as proportional to exposure and the severity aspect in terms of loss per claim (not dependent on exposure). However, this does not cover the entire story. For many lines of business, it is convenient for exposures to be proportional to inflation. Inflation is typically viewed as unrelated to frequency, but as proportional to severity.

6.2.3.1 Small Exposures

We begin by considering instances where the units of exposure may be fractions. To illustrate, for our automobile data, E_i will represent the fraction of the year that a policyholder had insurance coverage. The logic behind this is that the expected number of accidents is directly proportional to the length of coverage. This can also be motivated by a probabilistic framework based on collections of Poisson-distributed random variables known as *Poisson processes* (see, for example, Klugman, Panjer, and Willmot, 2008).

For binary outcomes, this situation is less clear. One way to handle exposures is to let the logit link function depend on exposures. To this end, the basic logit link function is $\pi(z) = \frac{e^z}{1+e^z}$. Define an exposure-weighted logit link function to be $\pi_i(z) = E_i \frac{e^z}{1+e^z}$. With this definition, the probability of a claim is

$$\Pr(r_i = 1) = \pi_i = \pi_i(\mathbf{x}'_i \boldsymbol{\beta}) = E_i \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta})}{1 + \exp(\mathbf{x}'_i \boldsymbol{\beta})}. \quad (6.1)$$

For more discussion, see de Jong and Heller (2008, p. 102); for illustrative SAS code, see de Jong and Heller (2008, p. 162).

Variations. There are alternative ways of incorporating partial year exposures, none clearly superior to the others. Equation (6.1) is based on a uniform distribution of failures within a year. Two other measures are as follows:

1. A constant hazard rate within the year assumption, resulting in $\pi_{i,H}(z) = 1 - (1 - \frac{e^z}{1+e^z})^{E_i}$.
2. A hyperbolic assumption (known as the “Balducci” assumption for an Italian actuary), resulting in $\pi_{i,B}(z) = \frac{E_i \frac{e^z}{1+e^z}}{1 - (1 - E_i) \frac{e^z}{1+e^z}}$.

See Bowers et al. (1997) for a discussion of these variations.

For some applications, the event of a claim is a relatively infrequent event, and the analyst would like to use all the information available in a claims database. One may wish to “over-sample” policyholders with claims; the idea is to draw a larger proportion of a subset of the population that is of interest in the study. Appendix Section 6.7 provides details of this type of sampling scheme.

6.2.4 Grouped versus Individual Data

A discussion of large exposures leads naturally into a discussion of grouped versus individual data.

6.2.4.1 Using Offsets to Handle Large Exposures

To begin, recall that sums of independent Poisson random variables also have a Poisson distribution. So, when summing random variables from independent policies, it is sensible to think of exposures as large positive numbers. Thus, it is common to model the number of accidents per thousand vehicles or the number of homicides per million population.

For a Poisson distribution, we can use the (logarithmic) number of policies in a group as an offset variable. Mathematically, if we are thinking about E_i independent Poisson variables in group i , each with mean μ_i , then the sum will also be Poisson distributed with mean $E_i\mu_i$. For the Poisson distribution, the variance equals the mean, so both the mean and the variance grow proportionally to the exposure E_i . When using a Poisson distribution with a logarithmic link function, one only needs to specify an offset variable $\ln E_i$ to automatically account for the growing variability.

However, for other distributions, this need not be the case. In the GLM linear exponential family, we saw that the variance can be expressed as a function of the mean, $v(\mu)$. To be specific, consider the gamma distribution where $v(\mu) = \mu^2/\phi$ and ϕ is a dispersion parameter. If we are thinking about E_i independent gamma random variables in group i , each with mean μ and variance θ , then the sum will also be gamma distributed with mean $E_i\mu$ and variance $E_i\theta$. When using a gamma distribution with a logarithmic link function and offset variable $\ln E_i$, the mean will grow proportionally to the exposure E_i , but the variability will grow proportionally to E_i^2 , not E_i . So, an offset by itself cannot handle large exposures.

6.2.4.2 Using Variable Scale Parameters to Handle Exposures

For a general distribution in the linear exponential family, suppose that we have m independent variables from the same distribution with location parameter θ and scale parameter ϕ . Then, basic arguments given in Section 6.6 show that the sample average comes from the same distributional family with location parameter θ and scale parameter ϕ/m . To apply this result, let us consider a problem that analysts regularly face: the use of grouped versus individual data.

To be specific, think about a sample $i = 1, \dots, n$ categories, or *groups*. For example, each group could be formed by the intersection of the amount of insurance, territory, and so forth. For the i th group, we have E_i independent observations with the same distribution from the linear exponential family. It has, for example, location parameter θ_i , mean μ_i , and scale parameter ϕ (that may or may not depend on i). One could run an analysis with a dataset based on individual observations $j = 1, \dots, E_i$, $i = 1, \dots, n$.

However, with these assumptions, then the average outcome from the i th group comes from the same exponential family with the same mean μ_i (or location parameter θ_i) but with a scale parameter ϕ/E_i . An alternative method of analysis would be to use the smaller grouped data sample consisting of only n observations, using the reciprocal of exposure as the weight. The Section 6.6 result guarantees the following:

- Estimates of location/mean parameters (e.g., the regression coefficients) will be the same. For ratemaking purposes, analysts typically focus on location parameter estimates.
- Only in the case when the scale parameter is known (e.g., binomial, Poisson) would other inferential aspects (standard errors, t -statistics, p -values, and so forth) be the same. Individual data analysis provides more accurate estimates of scale parameters than the corresponding grouped data analysis.

The book's website provides a demonstration of this comparison using the statistical package R.

6.2.4.3 Large Exposures for Frequency and Severity Models

As noted earlier, for frequency and severity modeling, it is customary to think about the frequency aspect as proportional to exposure and the severity aspect in terms of loss per claim. Let us make this advice a bit more specific in the context of an individual versus grouped analysis.

Suppose that individual data consist of a sample $i = 1, \dots, n$ groups, with $j = 1, \dots, E_i$ independent observations within each group i . For observation $\{ij\}$, the dependent variables consist of (N_{ij}, S_{ij}) : the frequency and total amount of claims.

If explanatory/rating variables are available at the individual observation level, then aggregating information up to the group level is problematic because one loses the information in individual level variables.

Instead, assume that explanatory/rating variables are available only at the group level and we wish to model aggregate frequency and severity variables $\{N_i = \sum_{j=1}^n N_{ij}, S_i = \sum_{j=1}^n S_{ij}\}$.

- For claims frequency, one alternative is to use a Poisson model with the response N_i and offset $\ln E_i$.

- For claims frequency, another alternative is to use a count member from the exponential distribution family (e.g., binomial) with the response N_i/E_i and scale parameter ϕ/E_i .
- For claims severity, use a severity member from the exponential distribution family (e.g., gamma) with the response S_i/N_i and scale parameter ϕ/N_i .

As noted earlier, these modeling strategies provide reliable estimates of location (mean) parameters but not scale parameters. This is a comparative advantage of analysis with individual level analysis; other advantages include the following:

- Group-level analysis was important before modern day computing and databases became available. Currently, however, the computing requirements for an individual-level analysis rarely present a substantial barrier.
- Group-level analysis precludes the examination of individual observations. Often, a highly unusual observation (“outlier”) can provide important information to the analyst.
- The equivalence between the two procedures relies on a number of unverifiable assumptions, including the independence of observations within a group. In some instances, we can think of reasons for positive associations among observations from the same category. In this case, the variance of the sum grows faster than linearly, and so specifying the scale parameter as inversely proportional to the exposure may give too large a weight to categories with large exposure.

6.3 Frequency-Severity Models

6.3.1 Two-Part Models

In Section 6.2.1, we introduced three forms of dependent variables. We now focus on the first type where the only dependent variable of interest is the total claims from a policy. However, let us think about datasets where there is a large proportion of zeros, corresponding to no claims. For example, in homeowners, it is not uncommon to consider data where 93% of policies do not have a claim.

To address this large proportion of zeros, we consider a two-part model that is a special type of frequency-severity model. In a two-part model, one part indicates whether an event (claim) occurs, and the second part indicates the size of the event. Specifically, let r_i be a binary variable indicating whether or not the i th subject has an insurance claim and y_i describe the amount of the claim.

To estimate a two-part model, the analyst first considers the frequency and then the severity, conditional on the frequency.

- (1) Use a binary regression model with r_i as the dependent variable and \mathbf{x}_{1i} as the set of explanatory variables. Denote the corresponding set of regression coefficients as $\boldsymbol{\beta}_1$. The logit is a typical binary regression model.

- (2) Conditional on $r_i = 1$, specify a regression model with y_i as the dependent variable and \mathbf{x}_{2i} as the set of explanatory variables. Denote the corresponding set of regression coefficients as $\boldsymbol{\beta}_2$. The gamma with a logarithmic link is a typical severity model.

There is usually overlap in the sets of explanatory variables, where variables are members of both \mathbf{x}_1 and \mathbf{x}_2 . Typically, one assumes that $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are not related so that the joint likelihood of the data can be separated into two components and run separately.

6.3.1.1 Tobit Models

Another way of modeling a large proportion of zeros is to assume that the dependent variable is (left) censored at zero. Chapter 19 on survival models provides a more complete introduction to censored regression. This section emphasizes the application to two-part data.

With censored regression models, we use an unobserved, or latent, variable y^* that is assumed to follow a linear regression model of the form

$$y_i^* = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i. \quad (6.2)$$

The responses are censored or “limited” in the sense that we observe $y_i = \max(y_i^*, 0)$. Estimation of this model is typically done by assuming normally distributed disturbances ε_i and using maximum likelihood estimation.

It is straightforward to extend this model to allow for limiting values that vary by policy. In actuarial applications, we think about d_i as representing a (known) deductible that varies by policyholder.

One drawback of the tobit model is its reliance on the normality assumption of the latent response. A second, and more important, drawback is that a single latent variable dictates both the magnitude of the response and the censoring. There are many instances where the limiting amount represents a choice or activity that is separate from the magnitude. For example, in a population of smokers, zero cigarettes consumed during a week may simply represent a lower bound (or limit) and may be influenced by available time and money. However, in a general population, zero cigarettes consumed during a week can indicate that a person is a nonsmoker, a choice that could be influenced by other lifestyle decisions (for which time and money may or may not be relevant).

6.3.2 Other Frequency-Severity Models

We now focus on the second and third types of dependent variables introduced in Section 6.2.1.

For the second form, we have aggregate counts and severities (N_i, S_i) (or use the notation y_i instead of S_i). Then, the two-step frequency-severity model procedure is as follows:

- 1. Use a count regression model with N_i as the dependent variable and \mathbf{x}_{1i} as the set of explanatory variables. Denote the corresponding set of regression coefficients as $\boldsymbol{\beta}_1$. Typical models include the Poisson and negative binomial models.
- 2. Conditional on $N_i > 0$, use a GLM with S_i/N_i as the dependent variable and \mathbf{x}_{2i} as the set of explanatory variables. Denote the corresponding set of regression coefficients as $\boldsymbol{\beta}_2$. Typical models include the gamma regression with a logarithmic link and a dispersion parameter proportional to $1/N_i$.

For the third form of dependent variables, we have individual claims $\mathbf{y}_i = (y_{i1}, \dots, y_{i,N_i})'$ available. In this case, the first step for the count model is the same. The second step for severity modeling becomes

- 2*. Conditional on $N_i > 0$, use a regression model with y_{ij} as the dependent variable and \mathbf{x}_{2i} as the set of explanatory variables. Denote the corresponding set of regression coefficients as $\boldsymbol{\beta}_2$. Typical models include the linear regression (with logarithmic claims as the dependent variable), gamma regression, and mixed linear models. For the mixed linear models, one uses a subject-specific intercept to account for the heterogeneity among policyholders.

6.3.3 Tweedie GLMs

The natural exponential family includes continuous distributions, such as the normal and gamma, as well as discrete distributions, such as the binomial and Poisson. It also includes distributions that are mixtures of discrete and continuous components. In insurance claims modeling, a widely used mixture is the Tweedie (1984) distribution. It has a positive mass at zero representing no claims and a continuous component for positive values representing the total amount for one or more claims.

The Tweedie distribution is defined as a Poisson sum of gamma random variables. Specifically, suppose that N has a Poisson distribution with mean λ , representing the number of claims. Let y_j be an i.i.d. sequence, independent of N , with each y_j having a gamma distribution with parameters α and γ , representing the amount of a claim. Then, $S_N = y_1 + \dots + y_N$ is a Poisson sum of gammas.

To understand the mixture aspect of the Tweedie distribution, first note that it is straightforward to compute the probability of zero claims as

$$\Pr(S_N = 0) = \Pr(N = 0) = e^{-\lambda}.$$

The distribution function can be computed using conditional expectations,

$$\Pr(S_N \leq y) = e^{-\lambda} + \sum_{n=1}^{\infty} \Pr(N = n) \Pr(S_n \leq y), \quad y \geq 0.$$

Because the sum of i.i.d. gammas is a gamma, S_n (not S_N) has a gamma distribution with parameters $n\alpha$ and γ . Thus, for $y > 0$, the density of the Tweedie distribution is

$$f_S(y) = \sum_{n=1}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} \frac{\gamma^{n\alpha}}{\Gamma(n\alpha)} y^{n\alpha-1} e^{-y\gamma}. \quad (6.3)$$

At first glance, this density does not appear to be a member of the linear exponential family. To see the relationship, we first calculate the moments using iterated expectations as

$$\mathbb{E} S_N = \lambda \frac{\alpha}{\gamma} \quad \text{and} \quad \text{Var } S_N = \frac{\lambda\alpha}{\gamma^2} (1 + \alpha).$$

Now, define three parameters μ, ϕ, p through the relations

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}, \quad \alpha = \frac{2-p}{p-1}, \quad \text{and} \quad \frac{1}{\gamma} = \phi(p-1)\mu^{p-1}. \quad (6.4)$$

Inserting these new parameters in equation (6.3) yields

$$f_S(y) = \exp \left[\frac{-1}{\phi} \left(\frac{\mu^{2-p}}{2-p} + \frac{y}{(p-1)\mu^{p-1}} \right) + S(y, p, \phi) \right],$$

where

$$\exp S(y, p, \phi) = \frac{1}{y} \sum_{n=1}^{\infty} \frac{\left(\frac{y^\alpha}{\phi^{1/(p-1)}(2-p)(p-1)^\alpha} \right)^n}{n! \Gamma(n\alpha)}.$$

Thus, the Tweedie distribution is a member of the linear exponential family. Easy calculations show that

$$\mathbb{E} S_N = \mu \quad \text{and} \quad \text{Var } S_N = \phi\mu^p, \quad (6.5)$$

where $1 < p < 2$. The Tweedie distribution can also be viewed as a choice that is intermediate between the Poisson and the gamma distributions.

For the Tweedie GLM, we might use $\mathbf{x}_{i,T}$ as a set of covariates and $\boldsymbol{\beta}_T$ as the corresponding set of regression coefficients. With a logarithmic link, $\mu_i = \exp(\mathbf{x}'_{i,T} \boldsymbol{\beta}_T)$. For the distribution function, there is no closed-form expression, but we could compute this directly, for example, using the R function `ptweedie`.

6.3.4 Comparing the Tweedie to a Frequency-Severity Model

As an alternative, consider a model composed of frequency and severity components. Then, we might use a Poisson regression model for the frequency, thinking of the number of claims for the i th person as

$$N_i \sim \text{Poisson}(\lambda_i), \quad \lambda_i = \exp(\mathbf{x}'_{i,F} \boldsymbol{\beta}_F),$$

using a logarithmic link function. Here, $\mathbf{x}_{i,F}$ is a set of covariates to be used in the frequency modeling, and $\boldsymbol{\beta}_F$ is the corresponding set of regression coefficients.

For the severity, we might use a gamma regression also with a logarithmic link function. Thus, we would model loss amounts as

$$y_{ij} \sim \text{gamma}(\alpha, \gamma_i), \quad \text{where } \frac{\alpha}{\gamma_i} = \text{E } y_{ij} = \exp(\mathbf{x}'_{i,S} \boldsymbol{\beta}_S),$$

for $j = 1, \dots, N_i$. Similar to frequency, $\mathbf{x}_{i,S}$ is a set of covariates to be used in the severity modeling, and $\boldsymbol{\beta}_S$ is the corresponding set of regression coefficients. Thus, the frequency and severity models need not employ the same set of covariates.

Putting the frequency and severity components together yields the aggregate loss

$$S_{N,i} = y_{ij} + \dots + y_{i,N_i}.$$

This has mean

$$\text{E } S_{N,i} = \text{E } N_i \times \text{E } y_{ij} = \exp(\mathbf{x}'_{i,F} \boldsymbol{\beta}_F + \mathbf{x}'_{i,S} \boldsymbol{\beta}_S) \quad (6.6)$$

and variance

$$\text{Var } S_{N,i} = \lambda_i \frac{\alpha}{\gamma_i^2} (1 + \alpha) = \exp(\mathbf{x}'_{i,F} \boldsymbol{\beta}_F + 2\mathbf{x}'_{i,S} \boldsymbol{\beta}_S + \ln(1 + 1/\alpha)). \quad (6.7)$$

Note that for frequency-severity modeling, two parameters, λ_i and γ_i , vary with i . To compute the distribution function, one could use the Tweedie for S_N with the R function `ptweedie`. This would be done by reversing the relations in (6.4) to get

$$p = \frac{\alpha + 2}{\alpha + 1}, \quad \mu_i = \lambda_i \frac{\alpha}{\gamma_i}, \quad \text{and} \quad \phi_i \mu_i^p = \lambda_i \frac{\alpha}{\gamma_i^2} (1 + \alpha). \quad (6.8)$$

Note that if one begins with the frequency-severity model formulation, the scale parameter ϕ depends on i .

6.4 Application: Massachusetts Automobile Claims

We investigate frequency-severity modeling using an insurance automobile claims dataset studied in Ferreira and Minikel (2010; 2012). These data, made public by the Massachusetts Executive Office of Energy and Environmental Affairs (EOEEA), summarize automobile insurance experience from the state of Massachusetts in year

Table 6.1. *Number of Policies by Rating Group and Territory*

Rating Group	Territory						Total
	1	2	3	4	5	6	
A – Adult	13,905	14,603	8,600	15,609	14,722	9,177	76,616
B – Business	293	268	153	276	183	96	1,269
I – Youthful with less than 3 years Experience	706	685	415	627	549	471	3,453
M – Youthful with 3–6 years Experience	700	700	433	830	814	713	4,190
S – Senior Citizens	2,806	3,104	1,644	2,958	2,653	1,307	14,472
Totals	18,410	19,360	11,245	20,300	18,921	11,764	100,000

2006. The dataset consists of approximately 3.25 million policies representing more than a half-billion dollars of claims.

Because the dataset represents experience from several insurance carriers, it is not surprising that the amount of policyholder information is less than typically used by large carriers that employ advanced analytic techniques. Nonetheless, we do have basic ratemaking information that is common to all carriers, including primary driver characteristics and territory groupings. At the vehicle level, we also have mileage driven in a year, the focus of the Ferreira and Minikel study.

6.4.1 Data and Summary Statistics

From the Ferreira and Minikel (2010) data, we drew a random sample of 100,000 policyholders for our analysis. Table 6.1 shows the distribution of number of policies by rating group and territory. The distribution of policies is reasonably level across territories. In contrast, the distribution by rating group is more uneven; for example, more than three-quarters of the policies are from the “Adult” group. The sparsest cell is business drivers in territory 6; the most heavily populated cell is territory 4 adult drivers.

For this study, an insurance claim is from only bodily injury, property damage liability, and personal injury protection coverages. These are the compulsory, and thus fairly uniform, types of insurance coverages in Massachusetts; it is critical to have uniformity in reporting standards in an intercompany study such as in Ferreira and Minikel (2010; 2012). In Table 6.2, the averages of the loss might appear to be lower than in other studies because the “total” is over the three compulsory coverages and does not represent, for example, losses from the commonly available (and costly) comprehensive coverage. The average total loss in Table 6.2 is 127.48. We also see important differences by rating group, where average losses for inexperienced

Table 6.2. *Averages by Rating Group*

Rating Group	Total Loss	Claim Number	Earned Exposure	Annual Mileage	Number of Policies	
					Total	With Valid Annual Miles
A	115.95	0.040	0.871	12,527	76,616	69,201
B	159.67	0.055	0.894	14,406	1,269	1,149
I	354.68	0.099	0.764	12,770	3,453	2,786
M	187.27	0.065	0.800	13,478	4,190	3,474
S	114.14	0.038	0.914	7,611	14,472	13,521
Total	127.48	0.043	0.870	11,858	100,000	90,131

youthful drivers are more than three times greater than for adult drivers. We can think of this total loss as a “pure premium.”

Table 6.2 shows that the average claim frequency is 4.3%. Specifically, for the 100,000 policies, 95,875 had zero claims, 3,942 had one claim, 176 had two claims, and 7 had three claims. The table also reports important differences by rating group, where the average number of losses for inexperienced youthful drivers are about 2.5 times greater than for adult drivers.

Table 6.2 also summarizes information on the earned exposure, defined here as the amount of time that the policy was in force in the study, and annual mileage. Annual mileage was estimated by Ferreira and Minikel (2010; 2012) based on Massachusetts’ Commonwealth’s Registry of Motor Vehicles mandatory annual safety checks, combined with automobile information from the vehicle identification number (commonly known using the acronym VIN). Interestingly, Table 6.2 shows that only about 90% of our data possess valid information about the number of miles driven, so that about 10% are missing this information.

Table 6.3 provides similar information but by territory. Here, we see that the average total loss and number of claims for territory 6 are about twice that for territory 1.

Table 6.3. *Averages by Territory*

Territory	Total Loss	Claim Number	Earned Exposure	Annual Mileage	Number of Policies	
					Total	With Valid Annual Miles
1	98.24	0.032	0.882	12,489	18,410	16,903
2	94.02	0.036	0.876	12,324	19,360	17,635
3	112.21	0.037	0.870	12,400	11,245	10,092
4	126.70	0.044	0.875	11,962	20,300	18,331
5	155.62	0.051	0.866	10,956	18,921	16,944
6	198.95	0.066	0.842	10,783	11,764	10,226
Total	127.48	0.043	0.870	11,858	100,000	90,131

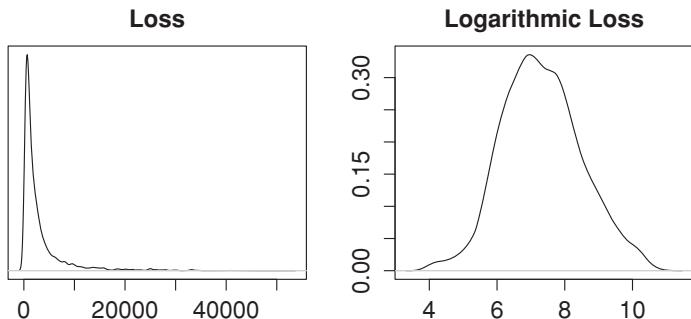


Fig. 6.1. Loss distribution. The left-hand panel shows the distribution of loss; the right-hand panel shows the same distribution but on a (natural) logarithmic scale.

There are 4,125 ($= 100,000 - 95,875$) policies with losses. To get a better handle on claim sizes, Figure 6.1 provides smooth histograms of the loss distribution. The left-hand panel is in the original (dollar) units, indicating a distribution that is right-skewed. The right-hand panel shows the same distribution on a logarithmic scale where we see a more symmetric behavior.

Do our rating factors affect claim size? To get some insights into this question, Figure 6.2 shows the logarithmic loss distribution by each factor. The left-hand panel shows the distribution by rating group; the right-hand panel shows the distribution by territory. Neither figure suggests that the rating factors have a strong influence on the size distribution.

6.4.2 Model Fitting

We report three types of fitted models here: (1) frequency models, (2) a severity model, and (3) a pure premium model.

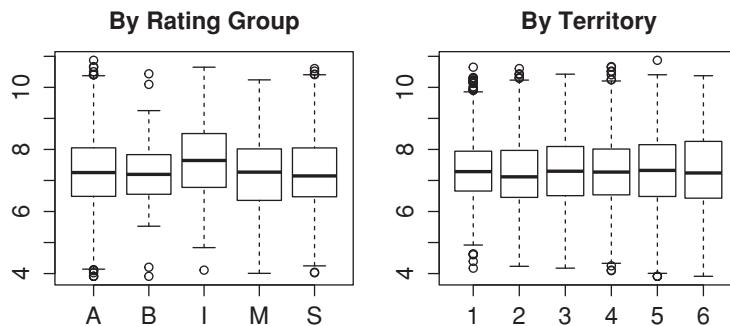


Fig. 6.2. Logarithmic loss distribution by factor. The left-hand panel shows the distribution by rating group; the right-hand panel shows the distribution by territory.

Table 6.4. Comparison of Poisson and Negative Binomial Models

Effect	Poisson		Negative Binomial		Relativity (Poisson)
	Estimate	t-Statistic	Estimate	t-Statistic	
(Intercept)	-2.636	-70.92	-2.639	-69.67	
Rating Group					
B	0.344	2.85	0.343	2.79	1.41
I	1.043	18.27	1.038	17.64	2.84
M	0.541	8.58	0.539	8.38	1.72
S	-0.069	-1.49	-0.069	-1.48	0.93
Territory					
1	-0.768	-14.02	-0.766	-13.79	0.46
2	-0.641	-12.24	-0.640	-12.04	0.53
3	-0.600	-9.87	-0.598	-9.70	0.55
4	-0.433	-8.81	-0.432	-8.64	0.65
5	-0.265	-5.49	-0.264	-5.37	0.77

Notes: Both models use logarithmic exposure as an offset. Estimated negative binomial dispersion parameter is 2.128. Reference levels are “A” for Rating Group and “6” for Territory.

Table 6.4 summarizes the results from two frequency models: Poisson and negative binomial regression models. For both models, we used a logarithmic link with logarithmic exposure as an offset variable. Focussing on the Poisson fit, we see that the *t*-statistics indicate strong statistical significance for several levels of each factor: rating group and territory. Additional tests confirm that they are statistically significant factors. Although not reported in Table 6.4, we also ran a model that included interactions among terms. The interaction terms were statistically insignificant with a *p-value* = 0.303 level. Hence, we report on the model without interactions, in part because of our desire for simplicity.

We also ran an analysis including annual mileage. This variable turned out to be strongly statistically significant, with a *t*-statistic equal to 12.08. However, by including this variable, we also lost 9,869 observations due to missing values in annual mileage. From the perspective taken in the Ferreira and Minikel (2010; 2012) study, mileage is the key variable of interest, and so the analyst would wish to retain this variable. From another perspective, including the mileage variable might result in analyzing a biased sample; that is, the roughly 10% population without mileage might differ dramatically from the 90% with mileage. Because of the biased sample concern, we treat the potential inclusion of the mileage variable as an interesting follow-up study.

For some audiences, analysts may wish to present the more flexible negative binomial regression model. Table 6.4 shows that there is little differences in the estimated

Table 6.5. Gamma Regression Models

Effect	Without Number		With Number	
	Estimate	t-Statistic	Estimate	t-Statistic
(Intercept)	7.986	137.33	7.909	76.87
Rating Group				
B	0.014	0.08	0.020	0.11
I	0.222	2.49	0.218	2.43
M	-0.013	-0.13	-0.015	-0.15
S	0.036	0.50	0.038	0.52
Territory				
1	0.026	0.31	0.027	0.31
2	-0.137	-1.67	-0.138	-1.68
3	0.004	0.04	0.005	0.05
4	-0.029	-0.38	-0.029	-0.38
5	0.019	0.26	0.018	0.23
Claim Number	-	-	0.071	0.90
Estimated Dispersion Parameter	2.432		2.445	

Notes: Reference levels are “A” for Rating Group and “6” for Territory.

coefficients for this dataset, indicating that the simpler Poisson model is acceptable for some purposes. We use the Poisson distribution in our out-of-sample analysis in Section 6.4.3, primarily because it provides an analytic expression for the predictive distribution (using the fact that a Poisson sum of independent gammas has a Tweedie distribution).

Table 6.5 summarizes the fit of a gamma regression severity model. As described in Section 6.3.2, we use total losses divided by the number of claims as the dependent variable and the number of claims as the weight. We fit a gamma distribution with a logarithmic link and the two factors, rating group and territory. Table 6.5 shows small *t*-statistics associated with the levels of rating group and territory – only “inexperienced” drivers are statistically significant. Additional tests indicate that the territory factor is not statistically significant and the rating group factor is marginally statistically significant with a *p-value* = 0.042. This is an interesting finding.

Table 6.5 also shows the result of using claim number as an explanatory variable in the severity model. For our data, the variable was not statistically significant and so was not included in subsequent modeling. Had the variable been statistically significant, a proxy would need to be developed for out-of-sample prediction. That is, although we can condition on claim number and it may be a sensible explanatory variable of (average) severity, it is not available a priori and so cannot be used directly for out-of-sample prediction.

Table 6.6. Tweedie Regression Model

Effect	Estimate	t-Statistic	Relativity
(Intercept)	5.356	63.47	
Rating Group			
B	0.340	1.28	1.41
I	1.283	9.39	3.61
M	0.474	3.22	1.61
S	-0.033	-0.36	0.97
Territory			
1	-0.743	-6.53	0.48
2	-0.782	-6.92	0.46
3	-0.552	-4.37	0.58
4	-0.480	-4.44	0.62
5	-0.269	-2.50	0.76

Notes: This model uses logarithmic exposure as an offset. Estimated dispersion parameter is 2.371. Reference levels are “A” for Rating Group and “6” for Territory.

As an alternative to the frequency-severity approach, we also fit a model using “pure premiums,” total losses, as the dependent variable. Similar to the frequency and severity models, we used a logarithmic link function with the factors of rating group and territory. The Tweedie distribution was used. We approximated the Tweedie shape parameter p using profile likelihood and found that the value $p = 1.5$ was acceptable. This was the value used in the final estimation.

Table 6.6 reports the fitted Tweedie regression model. The t -statistics associated with several levels of rating group and territory are statistically significant. This suggests, as was confirmed through additional testing, that both factors are statistically significant determinants of total loss. The table also reports the relativities (computed as the exponentiated parameter estimates). Interestingly, these relativities turn out to be close to those of the frequency model; this is not surprising given the lack of statistical significance associated with the factors in the severity model.

6.4.3 Out-of-Sample Model Comparisons

To compare the frequency-severity and pure premium approaches, we examined a “held-out” validation sample. Specifically, from our original database, we drew a random sample of 100,000 policies and developed the models reported in Section 6.4.2. Then, we drew an (independent) sample of 50,000 policies. For the frequency-severity model, our predictions are based on equation (6.6), using Poisson frequency coefficients in Table 6.4 to estimate β_F , severity coefficients in Table 6.5 to estimate β_S ,

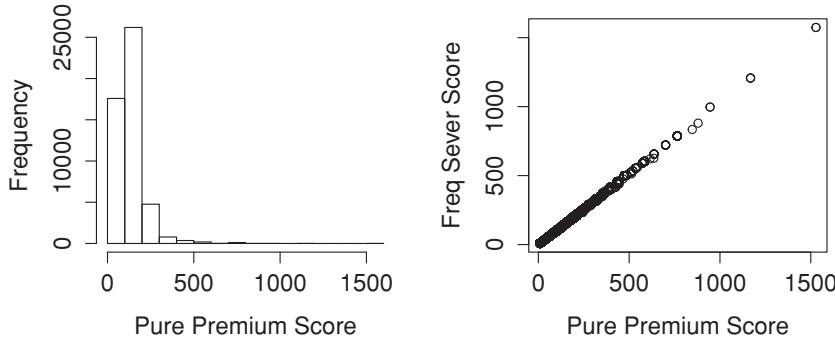


Fig. 6.3. Out-of-sample mean performance. The left-hand panel shows the distribution of the out-of-sample predictions calculated using the pure premium, or Tweedie, model. The right-hand panel shows the strong relationship between the scores from the frequency-severity and the pure premium models.

and with values of the independent variables from the held-out validation sample. The predictions for the Tweedie model followed similarly using the coefficients reported in Table 6.6.

Figure 6.3 compares the predictions for frequency-severity and the pure premium models. The left-hand panel shows the distribution of our pure premium predictions. The right-hand panel shows the strong relationship between the two predictions; it turns out that the correlation is approximately 0.999. For the purposes of predicting the mean, which is typically the primary focus of ratemaking, these two models yield virtually indistinguishable predictions. Both models provided some ability to predict total losses; the (Spearman) correlation between held-out losses and (either) predictor turned out to be 8.2%.

Because the mean predictor did not provide a way of discriminating between the pure premium and frequency-severity models, we also looked to tail percentiles. Specifically, in the Tweedie regression model, in Section 6.4.2 we cited $p = 1.5$ and $\hat{\phi} = 2.371$ and described how to estimate $\hat{\mu}_i$ for each observation i . Then, in Section 6.3.3 we noted that one could use the `pTweedie` function in R to get the distribution function. We did this for *each* held-out observation and evaluated it using the actual realized value. Recall the “probability integral transform,” a result in probability theory that says that when a continuous random variable is evaluated using its distribution function, the resulting transformed random variable should have a uniform (on $[0,1]$) distribution. Thus, if our distribution function calculation is approximately correct, then we can expect the held-out transformed random variables to have an approximate uniform distribution.

The procedure for Poisson frequency is similar to that for gamma severity models, but is more complex. In Section 6.3.3, we noted that a Poisson sum of gamma random

Table 6.7. *Out-of-Sample Quantile Performance*

Percentile	Pure Premium	Frequency-Severity
0.960	0.50912	0.42648
0.970	0.85888	0.79766
0.980	0.93774	0.86602
0.985	0.97092	0.90700
0.990	0.99294	0.93948
0.995	0.99528	0.97722
0.999	0.99784	0.99870

variables has a Tweedie distribution. So, even though we estimate the frequency and severity parameters separately, they can still be combined when we look at the loss distribution. In display (6.8), we show explicitly how to get Tweedie parameters from the Poisson frequency and gamma severity models. Then, as with the Tweedie GLM, we can calculate the transformed (using the distribution function) actual realized value.

Table 6.7 provides the comparisons for selected percentiles. Both models provide disappointing results below the 98th percentile; perhaps this is to be expected for a distribution with approximately 96% zeros. For the 99th percentile and above, the Tweedie does a good job tracking the actual held-out losses. In comparison, the frequency-severity approach is only competitive at the 99.9th percentile. On the one hand, this table suggests that fitting the tails of the distribution is a more complex problem that requires more refined data and sophisticated models. On the other hand, the similarity of results in Figure 6.3 when predicting the mean suggests a robustness of the GLM procedures that gives the analyst confidence when providing recommendations.

6.5 Further Reading

There is a rich literature on modeling the joint frequency and severity distribution of automobile insurance claims. There has been substantial interest in statistical modeling of claims frequency, yet the literature on modeling claims severity, especially in conjunction with claims frequency, is less extensive. One possible explanation, noted by Coutts (1984), is that most of the variation in overall claims experience may be attributed to claim frequency. Coutts (1984) also notes that the first paper to analyze claim frequency and severity separately seems to be Kahane and Levy (1975); see also Weisberg and Tomberlin (1982).

In the econometrics literature, Cragg (1971) provides an introduction to different frequency and severity covariates in two-part models, citing an example from fire

insurance. Mullahy (1998) provides an overview of two-part models and discusses health care applications.

Brockman and Wright (1992) provide an early overview of how statistical modeling of claims and severity can be helpful for pricing automobile coverage. Renshaw (1994) shows how generalized linear models can be used to analyze both the frequency and severity portions based on individual policyholder-level data. At the individual policyholder level, Frangos and Vrontos (2001) examined a claim frequency and severity model, using negative binomial and Pareto distributions, respectively. They use their statistical model to develop experience-rated (bonus-malus) premiums.

A trend in recent research has been to explore multivariate frequency-severity models, examining different lines of business or different perils simultaneously. The first papers in this area seem to be those of Pinquet (1997; 1998), fitting not only cross-sectional data but also following policyholders over time. Pinquet was interested in two lines of business: claims at fault and not at fault with respect to a third party. Frees et al. (2012) examine multivariate two-part models for different perils in homeowners insurance. Frees et al. (2013) review multivariate two-part models, examining several types of medical care expenditures jointly.

6.6 Appendix A. Sample Average Distribution in Linear Exponential Families

The distribution of the linear exponential family with parameters θ and ϕ is

$$f(y; \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + S(y, \phi)\right).$$

With this notation, it can be readily shown (e.g., Frees, 2010, chapter 13) that the moment-generation function can be expressed as

$$M(s; \theta, \phi) = E e^{sy} = \exp\left(\frac{b(\theta + s\phi) - b(\theta)}{\phi}\right).$$

Suppose that y_1, \dots, y_m are independently distributed with this moment-generating function. Then, the moment-generating function of the sample average is

$$\begin{aligned} E \exp\left(s \frac{y_1 + \dots + y_m}{m}\right) &= \prod_{i=1}^m E \exp\left(\frac{s}{m} y_i\right) \\ &= \prod_{i=1}^m \exp\left(\frac{b(\theta + \frac{s}{m}\phi) - b(\theta)}{\phi}\right) \\ &= \exp\left(\frac{b(\theta + s\frac{\phi}{m}) - b(\theta)}{\phi/m}\right) = M(s; \theta, \phi/m). \end{aligned}$$

Thus, the sample average is from the same linear exponential family with parameters θ and ϕ/m .

6.7 Appendix B. Over-Sampling Claims

If you work with government surveys such as the Medical Expenditure Survey (MEPS) in Chapter 2 or the Survey of Consumer Finances (SCF) in Frees (2010), you will see that it is common for such surveys to use unequal probabilities when drawing samples from larger populations. For example, the MEPS data over-sample poor and minority individuals; the SCF over-samples the wealthy. The idea is to draw a larger proportion of a subset of the population that is of interest in the study. In insurance, it is common to “over-sample” policyholders with claims.

Specifically, consider the two-part model introduced in Section 6.2.1 and presented in more detail in Section 6.3.1. Suppose that we have a very large database consisting of $\{r_i, y_i, \mathbf{x}_i\}$, $i = 1, \dots, N$ observations. We want to make sure to get plenty of $r_i = 1$ (corresponding to claims or “cases”) in our sample, plus a sample of $r_i = 0$ (corresponding to nonclaims or “controls”). Thus, we split the dataset into two pieces. For the first piece consisting of observations with $r_i = 1$, take a random sample with probability τ_1 . Similarly, for the second piece consisting of observations with $r_i = 0$, take a random sample with probability τ_0 . For example, we might use $\tau_1 = 1$ and $\tau_0 = 0.2$, corresponding to taking all of the claims and a 20% sample of nonclaims. Thus, the “sampling weights” τ_0 and τ_1 are considered known to the analyst. This over-sampling procedure is sometimes known as the “case-control” method.

How does this sampling procedure affect the inference in a two-part model? Think about this question from a likelihood perspective. To develop the likelihood, let $\{s_i = 1\}$ denote the event that the observation is selected to be included in the sample and $\{s_i = 0\}$ means that it is not included. Suppressing the $\{i\}$ subscript, we decompose the likelihood of the dependent variables that can be observed as

$$f(r, y|s = 1) = f(r|s = 1) \times f(y|s = 1, r)$$

“observable” likelihood = conditional frequency \times conditional severity.

For the conditional severity, it is common to assume that $f(y|s = 1, r) = f(y|r)$ – given the absence or presence of a claim, the selection mechanism has no effect on the amount. This is an assumption that may need to be verified, but does seem to commonly hold.

For the conditional frequency, here are some basic probability calculations to show how the conditional (on selection) claim frequency relates to the (population) claim frequency. Conditional on $\{r_i = 1\}$, we have that $\Pr(s_i = 1|r_i = 1) = \tau_1$, a Bernoulli

distribution. Similarly, $\Pr(s_i = 1|r_i = 0) = \tau_0$. From this, we have

$$\begin{aligned}\Pr(r_i = 1, s_i = 1) &= \Pr(s_i = 1|r_i = 1)\Pr(r_i = 1) = \tau_1\pi_i \\ \Pr(r_i = 0, s_i = 1) &= \Pr(s_i = 1|r_i = 0)\Pr(r_i = 1) = \tau_0(1 - \pi_i).\end{aligned}$$

Thus, the probability of the observation being selected into the sample is

$$\Pr(s_i = 1) = \tau_1\pi_i + \tau_0(1 - \pi_i).$$

Further, the probability of observing a claim in the sample is

$$\begin{aligned}\Pr(r_i = 1|s_i = 1) &= \frac{\Pr(r_i = 1, s_i = 1)}{\Pr(s_i = 1)} = \frac{\tau_1\pi_i}{\tau_1\pi_i + \tau_0(1 - \pi_i)} \\ &= \frac{\tau_1\pi_i/(1 - \pi_i)}{\tau_1\pi_i/(1 - \pi_i) + \tau_0}.\end{aligned}$$

Now, using the logit form in equation (6.1), we can express the odds-ratio as

$$\frac{\pi_i}{1 - \pi_i} = \frac{\frac{E_i}{1 + \exp(-\mathbf{x}'_i \boldsymbol{\beta})}}{1 - \frac{E_i}{1 + \exp(-\mathbf{x}'_i \boldsymbol{\beta})}} = \frac{E_i}{1 + \exp(-\mathbf{x}'_i \boldsymbol{\beta}) - E_i}. \quad (6.9)$$

Thus,

$$\begin{aligned}\Pr(r_i = 1|s_i = 1) &= \frac{\tau_1 \frac{E_i}{1 + \exp(-\mathbf{x}'_i \boldsymbol{\beta}) - E_i}}{\tau_1 \frac{E_i}{1 + \exp(-\mathbf{x}'_i \boldsymbol{\beta}) - E_i} + \tau_0} = \frac{\tau_1 E_i}{\tau_1 E_i + \tau_0(1 + \exp(-\mathbf{x}'_i \boldsymbol{\beta}) - E_i)} \\ &= \frac{\tau_1 E_i}{c_i + \tau_0 \exp(-\mathbf{x}'_i \boldsymbol{\beta})},\end{aligned}$$

where $c_i = \tau_1 E_i + \tau_0(1 - E_i)$. From this, we can express the probability of observing a claim in the sample as

$$\Pr(r_i = 1|s_i = 1) = \frac{E_i^*}{1 + \gamma_i \exp(-\mathbf{x}'_i \boldsymbol{\beta})} \quad (6.10)$$

where $E_i^* = \tau_1 E_i / c_i = \frac{\tau_1 E_i}{\tau_1 E_i + \tau_0(1 - E_i)}$ and $\gamma_i = \tau_0 / c_i = \frac{\tau_0}{\tau_1 E_i + \tau_0(1 - E_i)}$.

In summary, equation (6.10) has the same form as equation (6.1) with a new definition of exposure and the introduction of an offset term, $-\ln \gamma_i$, assuming that you use logistic regression (not probit) for your claim frequency modeling. If all of your exposures are identically equal to 1 ($E_i \equiv 1$), then γ_i is a constant and you simply reinterpret the constant in the systematic component $\mathbf{x}'_i \boldsymbol{\beta}$ (which we typically ignore). If exposures are not constant, then equation (6.10) gives a straightforward method of adjusting the exposure and introducing an offset term, allowing you to run the usual logistic regression software without the need for specialized software routines.

References

- Bowers, N. L., H. U. Gerber, J. C. Hickman, D. A. Jones, and C. J. Nesbitt (1997). *Actuarial Mathematics*. Society of Actuaries.
- Brockman, M. J. and T. S. Wright (1992). Statistical motor rating: making effective use of your data. *Journal of the Institute of Actuaries* 119, 457–543.
- Coutts, S. M. (1984). Motor insurance rating, an actuarial approach. *Journal of the Institute of Actuaries* 111, 87–148.
- Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* 39(5), 829–844.
- de Jong, P. and G. Z. Heller (2008). *Generalized Linear Models for Insurance Data*. Cambridge University Press, Cambridge, UK.
- Ferreira, J. and E. Minikel (2010). Pay-as-you-drive auto insurance in Massachusetts: A risk assessment and report on consumer, industry and environmental benefits. In *Conservation Law Foundation, Boston*. http://www.clf.org/wp-content/uploads/2010/12/CLF-PAYD-Study_November-2010.pdf.
- Ferreira, J. and E. Minikel (2012). Measuring per mile risk for pay-as-you-drive automobile insurance. *Transportation Research Record: Journal of the Transportation Research Board* 2297, 97–103.
- Frangos, N. E. and S. D. Vrontos (2001). Design of optimal bonus-malus systems with a frequency and a severity component on an individual basis in automobile insurance. *ASTIN Bulletin* 31(1), 1–22.
- Frees, E., X. Jin, and X. Lin (2013). Actuarial applications of multivariate two-part regression models. *Annals of Actuarial Science* 7(2), 258–287.
- Frees, E., G. Meyers, and A. D. Cummings (2012). Predictive modeling of multi-peril homeowners insurance. *Variance* 6(1), 11–31.
- Kahane, Y. and H. Levy (1975). Regulation in the insurance industry: Determination of premiums in automobile insurance. *Journal of Risk and Insurance* 42, 117–132.
- Klugman, S. A., H. H. Panjer, and G. E. Willmot (2008). *Loss Models: From Data to Decisions*. John Wiley & Sons, Hoboken, NJ.
- Mullahy, J. (1998). Much ado about two: Reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics* 17, 247–281.
- Pinquet, J. (1997). Allowance for cost of claims in bonus-malus systems. *ASTIN Bulletin* 27(1), 33–57.
- Pinquet, J. (1998). Designing optimal bonus-malus systems from different types of claims. *ASTIN Bulletin* 28(2), 205–229.
- Renshaw, A. E. (1994). Modeling the claims process in the presence of covariates. *ASTIN Bulletin* 24(2), 265–285.
- Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. In *Statistics: Applications and New Directions*. Proceedings of the Indian Statistical Golden Jubilee International Conference (Editors J. K. Ghosh and J. Roy), pp. 579–604. Indian Statistical Institute, Calcutta.
- Weisberg, H. I. and T. J. Tomberlin (1982). A statistical perspective on actuarial methods for estimating pure premiums from cross-classified data. *Journal of Risk and Insurance* 49, 539–563.
- Werner, G. and C. Modlin (2010). *Basic Ratemaking* (4th ed.). Casualty Actuarial Society.

Part II

Predictive Modeling Methods

7

Longitudinal and Panel Data Models

Edward W. Frees

Chapter Preview. This chapter considers regression methods where the analyst has the ability to follow a unit of analysis, such as a policyholder, over time. In the biomedical literature, this type of information is known as *longitudinal* data and, in the econometric literature, as *panel* data.

7.1 Introduction

7.1.1 What Are Longitudinal and Panel Data?

Regression is a statistical technique that serves to explain the distribution of an outcome of interest (y) in terms of other variables, often called “explanatory” or “predictor” (x ’s) variables. For example, in a typical ratemaking exercise, the analyst gathers a cross-section of policyholders and uses various rating variables (x ’s) to explain losses (y ’s).

In this chapter, we assume that the analyst has the ability to follow each policyholder *over time*. In many situations, a policyholder’s past loss experience can provide an important supplement to the information gleaned from available rating variables. We use the notation i to represent the unit of analysis (e.g., policyholder) that we will follow over time t . Using double subscripts, the notation y_{it} refers to a dependent variable y for policyholder i at time t , and similarly for explanatory variables \mathbf{x}_{it} . Consider four applications that an actuary might face that fall into this framework:

- (1) *Personal lines insurance such as automobile and homeowners:* Here, i represents the policyholder that we follow over time t , y represents the policy loss or claim, and the vector \mathbf{x} represents a set of rating variables. One could use $y_{i1} > 0$ to mean that a claim has occurred in period 1 – this often signals that we are likely to observe a claim in period 2.
- (2) *Commercial lines insurance such as commercial auto:* Here, i represents the commercial customer (policyholder) and y represents claims per premium (i.e., the loss

- ratio). Without intervening loss reduction measures, a high loss ratio in period 1 might signal a high loss ratio in period 2.
- (3) *Insurance sales:* Here, i represents a sales agent and y represents annual sales. Although agent information (e.g., years of experience) and sales territory information can be useful, often prior sales history is the most important variable for predicting sales.
 - (4) *Customer retention:* In any line of business, the analyst may follow a customer i over time. We might use $y_{i1} = 1$ to indicate that customer i bought a policy in period 1, and we wish to predict $y_{i2} = 1$ or 0, whether or not a customer buys a policy in period 2.

Of course, sometimes actuaries and other analysts attack these and related problems without a formal background in statistics, using well-established benchmarks and procedures that have stood the test of time. For example, in personal lines auto insurance, it is common to use the presence of a claim to indicate a rate increase by omitting the “safe driver discount.” In commercial lines, actuaries have a long history of incorporating past claims experience using credibility techniques. In contrast, statistical methods provide a disciplined way to reexamine and reevaluate old methods and can also be brought to bear on emerging problems.

7.1.2 Why Longitudinal and Panel Data?

7.1.2.1 Dynamic versus Cross-Sectional Effects

Analysts use standard cross-sectional regression analysis to make inferences about how changes in explanatory variables affect the dependent variable. Because there is no time element in cross-sectional data, we refer to these anticipated changes as *static*. In contrast, the actuary is typically interested in changes over time, known as *temporal* or *dynamic* changes. To underscore how these concepts can differ, consider the following example.

Example 7.1 (Effects of a Rating Variable on a Loss Outcome). Consider a (cross-sectional) sample of $n = 50$ policyholders. A synthetic dataset of a loss (y) and a rating variable (x) is displayed in the left-hand panel of Figure 7.1. This graph shows, for a single year (1), that the rating variable is an effective, although not perfect, explanatory variable for the loss. From this graph, an analyst would typically conclude that, as the rating variable increases, the expected loss increases.

For year 2, suppose that a similar relationship holds between the loss and rating variables. However, the rating variable has a natural increase associated with it (e.g., inflation). A plot of the rating variable and the loss for the combined years 1 and 2 (not pictured here) would provide the same overall conclusion as in the left-hand panel of Figure 7.1: as the rating variable increases, the expected loss increases.

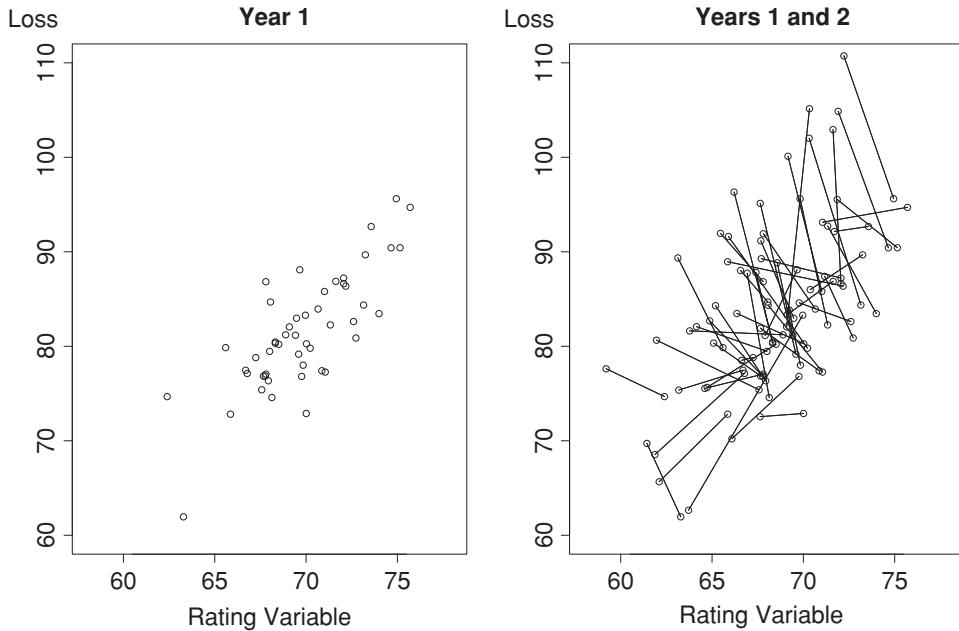


Fig. 7.1. Loss and rating variable. The left-hand panel shows the *positive* period 1 relationship between the loss and a rating variable. The right-hand panel shows the loss and rating variables for both periods 1 and 2, with lines connecting the periods for each policyholder. Most of the lines have a *negative* slope, indicating that increases in the rating variable result in a decrease in the loss variable.

In contrast, the right-hand panel of Figure 7.1 shows the plot of the rating variable and the loss for the combined years 1 and 2, but with a line connecting year 1 and year 2 results for each policyholder. The line emphasizes the dynamic effect, moving from the year 1 rating variable to the year 2. To interpret this plot, the data were constructed so that the year 1 rating variable is generally smaller than the year 2 result, so you can think of the change over time as moving from left to right. Most of the lines have a negative slope, indicating that increases in the rating variable result in a decrease in the loss variable. That is, the overall dynamic effect is *negative*, in contrast to the *positive* static effect.

How can these data happen? We can think of one scenario in which the rating variable naturally increases from year 1 to year 2 due to inflation. Further, suppose that an unobserved loss reduction measure has been introduced that serves to reduce expected losses for *all* policyholders. This would mean that each policyholder could expect to have an increase in the horizontal x axis and a decrease in the vertical y axis, resulting in a negative slope.

For another scenario, suppose that x represents a loss prevention measure such as a burglar alarm, and y represents theft losses in a home or commercial building.

We might interpret the left-hand panel as both x and y being positively related to an unobserved variable such as the “safety” of the home/building’s neighborhood. That is, in very safe neighborhoods, theft losses y tend to be low, and expenditures on burglar alarms x tend to be low (why pay a lot for a burglar alarm in such a safe neighborhood?) and, conversely for unsafe neighborhoods, resulting in the overall positive slope. However, for *each* home or building, the introduction of a more extensive burglar alarms means that expenditures x increase while expected losses y tend to decrease.

Regardless of the scenario or model generating the data, the important point is that static analysis without paying attention to temporal effects gives a grossly biased inference about the effects of the rating variable on the losses.

7.1.2.2 Efficiency and Sharing of Information

Example 7.1 illustrates a potential type of bias that one can encounter if the analyst does not think carefully about the longitudinal or panel nature of the data. Bias is a serious concern – as we saw in Example 7.1, the static relationship summarizing the rating variable’s effect on losses (positive) displayed the opposite tendency to that of the dynamic relationship (negative).

Analysts also use longitudinal and panel data techniques because they wish to optimize the use of information by promoting efficient inference and sharing of information among different quantities of interest.

Suppose that observations from different years are independent of one another. Even in this case, it is clear that an analyst would prefer two years of information about, for example, the relationship between a rating and a loss variable. Having two years of information, as compared to one, enables the analyst to provide more precise (efficient) estimates of parameters that summarize this relationship, with correspondingly smaller standard errors. Having multiple years of information allows the analyst to make efficient estimates about loss prediction for each policyholder, which is the goal of credibility theory.

However, it is more common for longitudinal and panel data to exhibit features of *clustering*, where observations from the same unit of analysis tend to be similar or “close” to one another in some sense. Referring to Example 7.1, it is commonly observed in insurance claims that y_{i2} tends to be related to y_{i1} ; that is, year 2 claims are related to year 1 claims for the same policyholder. By recognizing and incorporating this relationship into our models, we can (i) develop more efficient estimators and (ii) use the information in prior years (e.g., year 1) to predict current year (e.g., year 2) losses.

Example 7.2 (Bivariate Normal Example). It is helpful to assume normality to reaffirm basic concepts. Suppose that we have a sample of size n of logarithmic losses

from two years $\mathbf{y}_i = (y_{i1}, y_{i2})'$ that we assume are bivariate normal. That is, y_{it} is normally distributed with mean $\mu_{it} = \beta_0 + \beta_1 x_{it}$ and variance σ_t^2 , for years $t = 1$ and $t = 2$, and ρ is the correlation between y_{i1} and y_{i2} . Here, x_{i1} and x_{i2} are known rating variables. Assume that we have a sufficiently large sample size n so that estimation errors of the parameters β_0 , β_1 , σ_1^2 , σ_2^2 , and ρ are not a significant issue.

If we want to predict year 2 losses given the information in year 1, standard probability theory tells us that the conditional distribution is normal. Specifically, we have

$$y_{i2}|y_{i1} \sim N\left(\mu_{i2} + \rho \frac{\sigma_2}{\sigma_1} (y_{i1} - \mu_{i1}), \sigma_2^2(1 - \rho^2)\right).$$

Without information about y_{i1} , the optimal predictor of y_{i2} is its mean, $\mu_{i2} = \beta_0 + \beta_1 x_{i2}$. However, with information about prior year losses, y_{i1} , we can do better. The optimal predictor of y_{i2} given y_{i1} is its conditional mean, $\mu_{i2} + \rho \frac{\sigma_2}{\sigma_1} (y_{i1} - \mu_{i1})$. For the conditional predictor, the stronger the relationship between the two years, the larger is the value of ρ , and the smaller is the variance of the conditional distribution.

The conditional predictor outperforms the original (marginal) predictor because we are “sharing information” over the two years through the correlation parameter ρ . Of course, this improved performance relies on sufficient knowledge of the year 2 parameters, which is an empirical question. Nonetheless, this example opens the gateway so that you can see how this type of information might improve standard cross-sectional-based prediction.

7.1.3 Notation and Names

Models of longitudinal data are sometimes differentiated from regression and time series through their “double subscripts.” We use the subscript i to denote the unit of analysis, or *subject*, and t to denote time. To this end, define y_{it} to be the dependent variable for the i th subject during the t th time period. A longitudinal dataset consists of observations of the i th subject over $t = 1, \dots, T_i$ time periods, for each of $i = 1, \dots, n$ subjects. Because the number of repeated observations (need not equal 2 and) may vary by subject, we remind ourselves of the potential “imbalance” through the notation T_i for the number of observations per subject. Thus, we observe

$$\begin{aligned} &\text{first subject } \{y_{11}, \dots, y_{1T_1}\} \\ &\text{second subject } \{y_{21}, \dots, y_{2T_2}\} \\ &\vdots \qquad \qquad \vdots \\ &\text{nth subject } \{y_{n1}, \dots, y_{nT_n}\}. \end{aligned}$$

Looking at the data this way, we see that longitudinal data analysis can be thought of a special case of *multivariate* regression, where each subject features a multivariate response $\mathbf{y}_i = (y_{i1}, \dots, y_{iT_i})'$.

The term “panel study” was coined in a marketing context when Lazarsfeld and Fiske (1938) considered the effect of radio advertising on product sales. Traditionally, hearing radio advertisements had been thought to increase the likelihood of purchasing a product. Lazarsfeld and Fiske considered whether those who bought the product would have been more likely to hear the advertisement, thus positing a reverse in the direction of causality. They proposed repeatedly interviewing a set of people (the “panel”) to clarify the issue.

Baltes and Nesselroade (1979) trace the history of longitudinal data and methods with an emphasis on childhood development and psychology. They describe longitudinal research as consisting of “a variety of methods connected by the idea that the entity under investigation is observed repeatedly as it exists and evolves over time.” Moreover, they trace the need for longitudinal research to at least as early as the 19th century.

7.2 Linear Models

Longitudinal and panel data consist of repeated observations of a subject. To see the variety of ways to represent this type of data, we begin by working in a traditional linear model framework, where mean effects are linear in the parameters. Naturally, many actuarial applications fall into the nonlinear model context, but the linear model framework provides a convenient pedagogic framework on which we can base our discussions.

For period 1, think about a sample of the form $\{y_{i1}, \mathbf{x}_{i1}\}$ where we have a dependent variable of interest (y) and a collection of variables (\mathbf{x}) that could be used to explain y . This is the usual regression set-up. For our extension to longitudinal and panel data, we have a similar sample for the other periods of the form $\{y_{it}, \mathbf{x}_{it}\}$. Unlike some surveys (known as repeated cross-sections), we can link the “ i ” across periods.

Linear Model 1. Cross-Sectional Model. Here, we assume that all observations are independent and have a common variance $\text{Var } y_{it} = \sigma^2$ and regression function:

$$\begin{aligned} E y_{it} &= \alpha + \beta_1 x_{it,1} + \beta_2 x_{it,2} + \cdots + \beta_k x_{it,k} \\ &= \alpha + \mathbf{x}'_{it} \boldsymbol{\beta}. \end{aligned}$$

This model is helpful because it is regularly used as the “strawman” in any panel data analysis, one that can be easily defeated by using more sophisticated techniques. However, it is useful as a benchmark because many consumers are familiar and comfortable with cross-sectional regression analysis.

The model is useful in some isolated cases. A good illustration of this point is the capital asset pricing model, known by the acronym CAPM, of stock returns. In this application, i represents a firm whose stock price return, y_{it} , is followed over time t . (Sometimes it is the return in excess of the risk-free rate.) The only explanatory variable is the return based on a market index. Essentially, the argument from financial economics is that any patterns in the errors would be discovered, taken advantage of, and disappear in a liquid market. This model, without clustering or other relationships over time, is a good representation in this application.

Linear Model 2. Fixed Effects Model. As with Linear Model 1, assume independence among all observations and a common variance $\text{Var } y_{it} = \sigma^2$. The difference is in the regression function

$$\mathbb{E} y_{it} = \alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta},$$

where the intercept parameter α has a subscript i attached to it – this means that each subject has its own parameter. Intuitively, the parameter α_i is meant to capture all of the effects for the i th subject that are not observable in the other explanatory variables $\mathbf{x}_{it} = (x_{it,1}, \dots, x_{it,k})'$. With at least one parameter for each subject, you can see where repeated observations are necessary to estimate parameters of this model; in technical terms, this model is not identifiable with cross-sectional data where $T_i \equiv 1$.

The model is popular in part because it is easy to estimate with commonly available software. Think of the subject as a categorical variable, a “factor,” and so the varying intercepts α_i correspond to different levels of the factor. One can replace categorical variables with an appropriate set of binary variables; for this reason, panel data estimators are sometimes known as “least squares dummy variable model” estimators.

This is known as a “fixed effects” model because the quantities α_i are fixed, although unknown to the analyst, parameters to be estimated. This is contrast to the next model where we interpret the quantities to be random variables.

Linear Model 3. Random Effects Model. To allow intercepts to be random, instead of writing down only the regression function, consider the model equation

$$y_{it} = \alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta} + \varepsilon_{it},$$

where ε_{it} is an identically and independently distributed error, or disturbance, term with mean zero and variance σ_ε^2 . In this “random effects” formulation, the intercept α_i is a random variable, independent of $\{\varepsilon_{it}\}$.

The random effects model is commonly used in actuarial practice (although not by this name). As seen in Chapter 9, it is the basis of widely used credibility formulas for prediction.

Both the fixed and random effects models account for the clustering through common intercept parameters. For example, both y_{i1} and y_{i2} are based in part on the

intercept α_i , so are related through this quantity. Because of this, we interpret α_i to capture attributes of the dependent variable that are constant over time.

Linear Model 4. Model with Lagged Dependent Variables. A more direct way of relating y_{i1} to y_{i2} is through the model equation

$$y_{it} = \alpha + \gamma y_{i,t-1} + \mathbf{x}'_{it} \boldsymbol{\beta} + \varepsilon_{it}.$$

In this model equation, the dependent variable lagged by one period, $y_{i,t-1}$, is used as an explanatory variable to predict y_{it} . The parameter γ controls the strength of this relationship.

A strength of this model is that it is easy to interpret and to explain. It is similar in appearance to the popular autoregressive model of order one, *AR1*, that appears in Chapter 17. As with the *AR1* models, one loses the time $t = 1$ observations because there is no lagged version for the first period. However, for panel data, this means losing n observations. For example, if you are using $T = 4$ years of data, then with this model, you have lost your entire time 1 data for estimation, or approximately 25% of the data. For longer panels, this is not an issue. However, for shorter panels such as $T = 4$, this is a serious limitation of this model.

Linear Model 5. Model with Correlated Errors. As an alternative, instead of relating the dependent variables through a (conditional) regression function, one could relate the disturbance terms ε_{it} . For example, we might use the regression function

$$y_{it} = \alpha + \mathbf{x}'_{it} \boldsymbol{\beta} + \varepsilon_{it},$$

with disturbance terms

$$\varepsilon_{it} = \gamma_\varepsilon \varepsilon_{i,t-1} + \eta_{it}.$$

Here, η_{it} are i.i.d. error terms, and the parameter γ_ε controls the strength of association between ε_{it} and its lagged version, $\varepsilon_{i,t-1}$. The idea is that the disturbance terms follow an *AR1* structure, not the dependent variables. If you are familiar with time series modeling (as introduced in Chapter 17), this is a “moving average” formulation, a counterpart to the autoregressive model.

The limitation of this model is that it is more difficult to explain and interpret than a model with lagged dependent variables. A strength is that, with suitable conditions on the disturbance terms, the analyst does not lose year 1 observations in the estimation. This means that, in many situations, parameter estimates are more precise than competing models.

Table 7.1. Summary Statistics for Group Term Life Data

	Mean	Median	Standard Deviation	Minimum	Maximum
Coverage (000's)	30,277	11,545	54,901	25	427,727
Claims (000's)	14.724	5.744	32.517	0	290.206
Logarithmic Coverage	16.272	16.262	1.426	10.145	19.874
Logarithmic Claims	8.029	8.656	2.710	0	12.578

7.2.1 Group Term Life Example

This example shows how to model claims and exposure information for 88 Florida credit unions. These claims are “life savings” claims from term life contracts between the credit union and their members that provide a death benefit based on the member’s savings deposited in the credit union. For this example, the unit of analysis is the credit union, each of which is observed over four years, 1993–1996 inclusive. The interest is in modeling group term claims. For simplicity, we consider only one rating variable, the contract coverage.

Table 7.1 provides summary statistics for the coverages and claims. These variables are the sum over all contracts for each credit union; it shows summaries in terms of thousands (000’s) of U.S. dollars. Based on a preliminary analysis, it turns out that both variables are highly skewed and so we analyzed their (natural) logarithmic versions, also given in Table 7.1. Specifically, we used the transformation $\text{LnClaims} = \ln(1+\text{Claims})$, so that credit unions with zero claims remain at zero when on a logarithmic scale (and similarly for coverages).

To visualize the claim development over time, Figure 7.2 provides a *trellis plot* of logarithmic claims. This plot shows logarithmic claims versus year, with a panel for each credit union, arranged roughly in order of increasing size of claims. A trellis plot provides extensive information about a dataset and so is unlikely to be of interest to management, although it can be critical to the analyst developing a model. For our data, Figure 7.2 shows that claims are increasing over time and that this pattern of increase largely holds for *each* credit union. We see that credit union number 26, in the upper right-hand corner, has substantially larger claims than even the next largest credit union. We also see that, as the typical size of the claims decreases, the (downward) variability increases. Much can be learned by the analyst from close inspection of trellis plots.

Table 7.2 summarizes parameter estimates of the five model fits. Interestingly, the coefficient estimates across the five models are relatively consistent, indicating that the parameter estimates are robust to model specification. For the lagged dependent variable model, when we adjust for the mean effect of the lagged dependent variable,

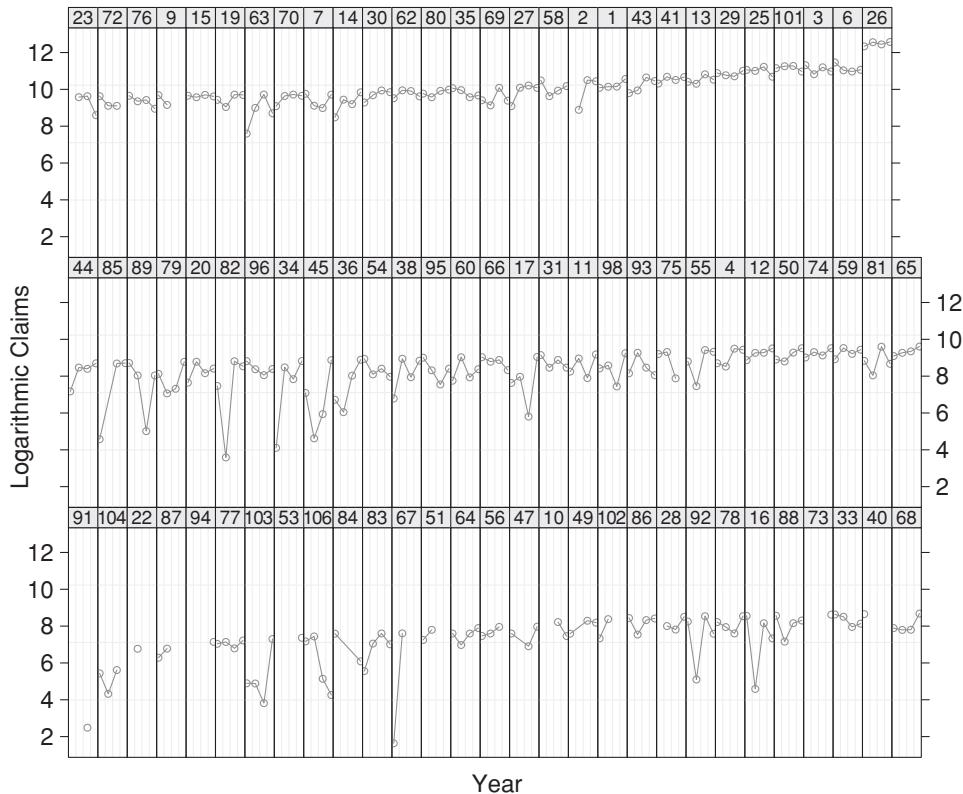


Fig. 7.2. Trellis plot of logarithmic group term life claims from 1993–1996.

the intercept is $\frac{-8.8657}{1-0.303} = -12.420$, and the slope for logarithmic coverage is $\frac{0.884}{1-0.303} = 1.268$, consistent with the other models. The t -statistic for the lagged dependent variable model is 5.35, indicating strong statistical evidence of serial autocorrelation. For the correlated errors model, the estimate of auto-covariance turns out to be $\hat{\gamma}_\varepsilon = 0.351$ which is qualitatively similar to findings from the lagged dependent variable model. For the model fits, the cross-sectional model turns out to have an $R^2 = 0.434$ goodness-of-fit statistics. The corresponding measure for the fixed effects model is $R^2 = 0.711$, a statistically significant improvement. Additional examination of model diagnostics (not displayed here) shows that all four longitudinal data models are superior to our “strawman,” the ordinary cross-sectional regression model.

7.3 Nonlinear Models

Although the linear model framework provides a convenient pedagogic framework on which to base our discussions, many actuarial applications fall into the nonlinear

Table 7.2. Summaries of Five Model Fits to the Group Term Life Data

	Cross-Sectional	Fixed Effects	Models Random Effects	Lagged Dependent	Correlated Errors
Intercept	-12.337	-12.286	-12.882	-8.657	-12.567
<i>t</i> -statistic	-9.49	-1.65	-7.48	-5.80	-7.73
Logarithmic Coverage	1.252	1.264	1.282	0.884	1.265
<i>t</i> -statistic	15.73	3.04	12.14	8.36	12.69
Lagged Claims				0.303	
<i>t</i> -statistic				5.35	

model context. This section provides a roadmap of how to think about modeling choices when your data cannot be reasonably be represented using a linear model.

7.3.1 Binary Outcomes

Many actuarial applications involve analyses of datasets where the outcome of interest is binary, often using $y = 1$ to signal the presence of an attribute and $y = 0$ to signal its absence. For a policy, this outcome can be a check function to see whether or not there was a claim during the period. For a customer, it can be whether or not a customer from one period is retained from one period to the next.

7.3.1.1 Random Effects

It is common to use a random effects model to reflect clustering of binary longitudinal data. To see how to incorporate random effects, as in Chapter 3, we use a logistic function $\pi(z) = \ln \frac{1}{1+e^{-z}}$. Now, conditional on α_i , define the probability

$$\pi_{it} = \Pr(y_{it} = 1|\alpha_i) = \pi(\alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta}).$$

As with linear random effects models, the quantity α_i can capture effects for the i th subject that are not observable in the other explanatory variables.

Estimation of binary outcomes random effects models is generally conducted using maximum likelihood. This is not as straightforward in the cross-sectional case because, to calculate the likelihood of observable data, one must take the expectation over unobservable random intercepts α_i . However, computational concerns about this aspect of estimation have been addressed using modern-day statistical software and high-speed computing, for all except the largest of datasets.

It is tempting to implement a fixed effects version of this model, thinking of $\{\alpha_i\}$ as fixed, yet unknown, parameters to be estimated. Unfortunately, in this nonlinear context, the behavior of the resulting estimates has been shown to be unreliable.

Intuitively, our ability to estimate global parameters β that are common to all subjects is corrupted by the inability to estimate the subject-specific effects α_i . In contrast, in the linear model context, the estimation procedure “sweeps out” the intercept terms when producing estimates of β , producing reliable estimates even when the estimates of α_i are not.

7.3.1.2 Markov Transition Models

In the biomedical and econometrics literatures, the random effects model is the most common method of analyzing longitudinal binary outcomes. However, in actuarial applications, it is helpful to think about Markov transition modeling. With these models, actuaries can account for *persistency* by tracing the development of a dependent variable over time and representing the distribution of its current value as a function of its history. In this context, think about the events $\{y = 1\}$ and $\{y = 0\}$ as representing two “states.” “Persistency” connotes the tendency to remain in a state over time. It is the same idea as clustering, yet applied to a state space context.

To be more precise about the development over time, define H_{it} to be the history of the i th subject up to time t . For example, if the explanatory variables are assumed to be nonstochastic, then we might use $H_{it} = \{y_{i1}, \dots, y_{i,t-1}\}$. With this information set, we may partition the likelihood for the i th subject as

$$f(y_{it}) \prod_{t=2}^{T_i} f(y_{it}|H_{it}),$$

where $f(y_{it}|H_{it})$ is the conditional distribution of y_{it} given its history and $f(y_{i1})$ is the marginal (unconditional) distribution of y_{i1} .

For a Markov model (of order 1), one assumes that the entire history is captured by the most recent outcome, so that

$$f(y_{it}|H_{it}) = f(y_{it}|y_{i,t-1}).$$

For binary outcomes, we can write the conditional probability of a 1, given $y_{i,t-1} = 0$, as

$$\pi_{it,0} = \Pr(y_{it} = 1|y_{i,t-1} = 0) = \pi(\alpha_0 + \mathbf{x}'_{it}\boldsymbol{\beta}_0)$$

and, given $y_{i,t-1} = 1$, as

$$\pi_{it,1} = \Pr(y_{it} = 1|y_{i,t-1} = 1) = \pi(\alpha_1 + \mathbf{x}'_{it}\boldsymbol{\beta}_1).$$

In this context, $\pi_{it,0}$ and $\pi_{it,1}$ are examples of *transition probabilities*, quantifying the probability of moving or transiting from one state to another. Here, we have used one set of parameters $\{\alpha_0, \boldsymbol{\beta}_0\}$ when the “origin” (determined by $y_{i,t-1}$) state is 0 and another set, $\{\alpha_1, \boldsymbol{\beta}_1\}$, when the origin state is 1. Whether this is appropriate

is an empirical question, although many datasets are rich enough to support this parameterization.

With this notation, the conditional distribution is given as

$$f(y_{it}|y_{i,t-1}) = \begin{cases} \pi_{it,1} & \text{if } y_{i,t-1} = 1, y_{it} = 1 \\ 1 - \pi_{it,1} & \text{if } y_{i,t-1} = 1, y_{it} = 0 \\ \pi_{it,0} & \text{if } y_{i,t-1} = 0, y_{it} = 1 \\ 1 - \pi_{it,0} & \text{if } y_{i,t-1} = 0, y_{it} = 0 \end{cases}.$$

Then, it is customary to estimate model parameters by maximizing a *partial log-likelihood*, given as

$$L_P = \sum_i \sum_{t=2}^{T_i} \ln f(y_{it}|y_{i,t-1}).$$

As with the lagged dependent variable linear model, this modeling choice does lose the period 1 observations; in this sense, it is a “partial” likelihood. For many problems of interest, the interesting part is estimating the transition probabilities (e.g., what is the probability of not retaining a customer?), and so one loses little by focusing on the partial likelihood.

7.3.2 Other Outcomes

7.3.2.1 Random Effects and Generalized Linear Model Outcomes

As we have seen, many “non-normal” outcomes can be handled using a generalized linear model (GLM). To handle clustering in a panel data context, the random effects formulation is commonly used.

For this formulation, we follow the usual three-stage GLM set-up. Specifically, we do the following:

1. Specify a distribution from the linear exponential family of distributions.
2. Introduce a systematic component. With random effects, this is conditional on α_i so that $\eta_{it} = \alpha_i + \mathbf{x}'_{it}\beta$.
3. Relate the systematic component to the (conditional) mean of the distribution through a specified link function

$$\eta_{it} = g(\mu_{it}) = g(E(y_{it}|\alpha_i)).$$

This modeling framework is sufficiently flexible to handle many practical applications. From a user’s viewpoint, it is convenient to use a single statistical software program, regardless of whether one wants to model a count (e.g., Poisson) or a medium-tailed (e.g., gamma) distribution.

The general random effects GLM model has the same limitations discussed in the special case of binary outcomes:

- This is a computationally intensive formulation that is tractable only with modern-day software and hardware.
- A fixed effects version is often not a reliable alternative.

See Chapter 16 for further discussion.

7.3.2.2 Categorical Outcomes and Markov Transition Models

Instead of only two (binary) outcomes, actuaries may want to model transitions to more than two states. For example, one might want to follow an automobile driver over time and assess whether the driver belongs to one of the three states: “preferred,” “standard,” or “substandard.” As another application, one might want to follow a bond over time and model how it moves among ratings “AAA,” “AA,” “A,” and so forth. On the one hand, it is possible to extend the generalized linear model framework to handle categorical outcomes using multinomial logits. On the other hand, Markov transition modeling provides greater flexibility and is easier to explain and to implement. See Chapter 20 for further discussion.

7.4 Additional Considerations

7.4.1 Unbalanced and Missing Data

Modern-day statistical software can readily handle “unbalanced” data situations where the number of repeated observations from a subject differs among subjects ($T_i \neq T$). In actuarial applications, balanced data are rarely encountered. As with cross-sectional regression, with longitudinal and panel data, the analyst must be especially wary of the reasons for data nonavailability, or “missingness.” It is common for the cause of missingness to be related to the outcome of interest, thus violating a basic sample selection assumption. For example, think about studying the financial strength of a cross-section of firms over time. If your sample includes financially weak firms, over time, some of these firms will no longer report financial information due to mergers or insolvencies. The outcome of interest, financial strength, is directly related to the availability of data, meaning that your analysis is based on a biased sample; in the latter years, it consists of only those companies that are financially strong enough to survive.

7.4.2 Clustered Data (Nontemporal)

Longitudinal and panel data techniques have been developed based on the assumption that repeated observations from a subject are observed over time. However, many of

these techniques can be applied to other sampling schemes that do not have a time element. The key is the replication of observations, not the observation over time. For example, you may organize your unit of analysis i to represent different geographic regions of a country (e.g., counties). Then, there may be several policies within each county, a type of replication without a time element.

Extending this line of thought, it is possible to think about several layers of a hierarchy. For example, in commercial automobile lines, one might consider a customer as the unit of analysis that we follow over time. For each customer, there may be several vehicles types insured (e.g., trucks, automobiles, motorcycles), and within each type, several vehicles. This type of analysis is known as *multilevel* modeling.

7.5 Further Reading

The ideas in this chapter are expanded on in the book-length treatment of Frees (2004). Readers may also wish to refer to Diggle et al. (2002) and Hsiao (2002), both of which provide excellent introductions to longitudinal and panel data analysis.

References

- Baltes, P. B. and J. R. Nesselroade (1979). History and rational of longitudinal research. In P. B. Baltes and J. R. Nesselroade (Eds.), *Longitudinal Research in the Study of Behavior and Development*. Academic Press, New York.
- Diggle, P. J., P. Heagarty, K. Liang, and S. L. Zeger (2002). *Analysis of Longitudinal Data* (2nd ed.). Oxford University Press, Oxford.
- Frees, E. W. (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge University Press, Cambridge.
- Hsiao, C. (2002). *Analysis of Panel Data* (2nd ed.). Cambridge University Press, Cambridge.
- Lazarsfeld, P. F. and M. Fiske (1938). The panel as a new tool for measuring opinion. *Public Opinion Quarterly* 2, 596–612.

8

Linear Mixed Models

Katrien Antonio and Yanwei Zhang

Chapter Preview. We give a general discussion of linear mixed models and continue by illustrating specific actuarial applications of this type of model. Technical details on linear mixed models follow: model assumptions, specifications, estimation techniques, and methods of inference. We include three worked-out examples with the R `lme4` package and use `ggplot2` for the graphs. Full code is available on the book's website.

8.1 Mixed Models in Actuarial Science

8.1.1 What Are Linear Mixed Models?

A First Example of a Linear Mixed Model. As explained in Chapter 7, a panel dataset follows a group of subjects (e.g., policyholders in an insurance portfolio) over time. We therefore denote variables (e.g., y_{it} , x_{it}) in a panel dataset with double subscripts, indicating the subject (say, i) and the time period (say, t). As motivated in Section 1.2 of Chapter 7, the analysis of panel data has several advantages. Panel data allow one to study the effect of certain covariates on the response of interest (as in usual regression models for cross-sectional data), while accounting appropriately for the dynamics in these relations. For actuarial ratemaking the availability of panel data is of particular interest in a posteriori ratemaking. An a posteriori tariff predicts the current year loss for a particular policyholder, using (among other factors) the dependence between the current year's loss and losses reported by this policyholder in previous years. Credibility theory, being a cornerstone of actuarial mathematics, is an example of such an a posteriori rating system. Section 2 in Chapter 7 presents a sequence of models suitable for the analysis of panel data in the context of linear models. Recall in particular the well-known linear regression model with common intercept or the cross-sectional model; see Linear Model 1 in Section 7.2:

$$\mathrm{E} y_{it} = \alpha + \mathbf{x}'_{it} \boldsymbol{\beta}. \quad (8.1)$$

This model *completely pools* the data, ignores the panel structure, and produces identical estimates for all subjects i (for a given x_{it}). The linear fixed effects model (Linear Model 2 in Section 7.2) specifies

$$\text{E } y_{it} = \alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta}, \quad (8.2)$$

where each subject i has its own unknown – but *fixed* – intercept α_i . Hence, the name *fixed effects* model. Independence among all observations is assumed, and $\text{Var}(y_{it}) = \sigma^2$. This regression model *does not pool* information and estimates each α_i separately using least squares or maximum likelihood. This approach often results in overfitting and unreasonable $\hat{\alpha}_i$'s (see Gelman 2006). The linear random effects model (see Linear Model 3 in Section 7.2) is an alternative approach, balancing between *no pooling* and *complete pooling* of data. It allows for *random* intercepts, with model equation

$$y_{it} = \alpha_i + \mathbf{x}'_{it} \boldsymbol{\beta} + \varepsilon_{it}, \quad (8.3)$$

where $\varepsilon_{it} \sim (0, \sigma_\varepsilon^2)$.¹ The subject specific intercept α_i is now a random variable with zero mean and variance σ_α^2 . Hence, the name *random effects* model. Moreover, the model in (8.3) is a first example of a *linear mixed model* (LMM), with a combination ('mix') of *fixed* and *random* effects in the linear predictor. The errors ε_{it} with variance σ_ε^2 structure represent variability within subject i , whereas the random intercepts with variance σ_α^2 represent variation between subjects. Compared with the *no pooling* and *complete pooling* examples, the linear mixed model has many interesting features, as explained later.

Mixed or Multilevel Models for Clustered Data. Panel data are a first example of so-called clustered data. As mentioned in Section 7.4, predictive modeling in actuarial science (and in many other statistical disciplines) is based on data structures going beyond the cross-sectional as well as panel data design. Section 8.3 in this chapter includes multiple motivating examples. Mixed (or multilevel) models are statistical models suitable for the analysis of data structured in nested (i.e., *hierarchical*) or non-nested (i.e., cross-classified, *next to* each other instead of hierachically nested) *clusters or levels*. In this chapter we explain the use of linear mixed models for multilevel data. A discussion of nonlinear mixed models is found in Chapter 16. Chapters 7 (on longitudinal and panel data), 9 (on credibility), and 11 (on spatial statistics) include additional examples of clustered data and their analysis with mixed models.

Textbook Examples. A standard textbook example of multilevel data is the students in schools data structure. Extended versions are the students in classes in schools or students followed repeatedly over time, in classes in schools examples, where

¹ The notation $\varepsilon_{it} \sim (0, \sigma_\varepsilon^2)$ implies $E[\varepsilon_{it}] = 0$ and $\text{Var}[\varepsilon_{it}] = \sigma_\varepsilon^2$.

each example adds an extra level of observations to the data hierarchy. Reflecting the actuarial audience of this book, we consider the example of a collection of vehicles j (with $j = 1, \dots, n_i$) insured under fleets i (with $i = 1, \dots, m$). Let y_{ij} be the loss observed for vehicle j in fleet i (in a well defined period of exposure). Denote with $x_{1,ij}$ covariate information at vehicle-level (our **level 1**): $x_{1,ij}$ is, for example, the cubic capacity or vehicle age of car j in fleet i . $x_{2,i}$ is a predictor at fleet-level (our **level 2**) and could, for example, refer to the size of the fleet or the business in which the fleet is operating. The so-called varying intercepts model is a basic example of a multilevel model. It combines a linear model at **vehicle-level** (i.e., level 1),

$$y_{ij} = \beta_i + \beta_{1,0} + x_{1,ij}\beta_{1,1} + \varepsilon_{1,ij}, \quad j = 1, \dots, n_i, \quad (8.4)$$

with a linear model at **fleet-level** (i.e., level 2),

$$\beta_i = \varepsilon_{2,i}, \quad i = 1, \dots, m, \quad (8.5)$$

or, when fleet-specific information is available,

$$\beta_i = x_{2,i}\beta_2 + \varepsilon_{2,i}, \quad i = 1, \dots, m. \quad (8.6)$$

Here $\varepsilon_{2,i} \sim (0, \sigma_2^2)$ and $\varepsilon_{1,ij} \sim (0, \sigma_1^2)$ are mean zero, independent error terms, representing variability (or heterogeneity) at both levels in the data. Written as a *single model equation*, the combination of (8.4) and, for example, (8.5), is

$$y_{ij} = \beta_{1,0} + \varepsilon_{2,i} + x_{1,ij}\beta_{1,1} + \varepsilon_{1,ij}. \quad (8.7)$$

This regression model uses an overall intercept, $\beta_{1,0}$; a fleet-specific intercept, $\varepsilon_{2,i}$; a vehicle-level predictor $x_{1,ij}$ with corresponding regression parameter, $\beta_{1,1}$; and an error term $\varepsilon_{1,ij}$. We model the fleet-specific intercepts, $\varepsilon_{2,i}$, as random variables, which reflects *heterogeneity* among fleets in an efficient way, even for a large number of fleets. Indeed, by assigning a distribution to these error terms, we basically only need an estimate for the unknown parameters (i.e., the variance component σ_2^2) in their distribution. The other regression parameters, $\beta_{1,0}$ and $\beta_{1,1}$, are considered *fixed* (in frequentist terminology); we do not specify a distribution for them. The model in (8.7) is – again – an example of a linear mixed model (LMM). *Mixed* refers to the combination of fixed and random effects, combined in a model specification that is linear in the random ($\varepsilon_{2,i}$) and in the fixed effects ($\beta_{1,0}$ and $\beta_{1,1}$).

Allowing for varying slopes and intercepts results in the following model equations:

$$y_{ij} = \beta_{i,0} + x_{1,ij}\beta_{i,1} + \beta_{1,0} + x_{1,ij}\beta_{1,1} + \varepsilon_{1,ij}, \quad (8.8)$$

$i = 1, \dots, m$, $j = 1, \dots, n_i$, with

$$\begin{aligned} \beta_{i,0} &= \varepsilon_{2,i,0}, \\ \beta_{i,1} &= \varepsilon_{2,i,1}. \end{aligned} \quad (8.9)$$

Written as a single model equation, this multilevel model becomes

$$y_{ij} = \beta_{1,0} + \varepsilon_{2,i,0} + x_{1,ij}\beta_{1,1} + x_{1,ij}\varepsilon_{2,i,1} + \varepsilon_{1,ij}. \quad (8.10)$$

In addition to having random intercepts ($\varepsilon_{2,i,0}$), the model also allows the effect of predictor $x_{1,ij}$ on the response to vary by fleet. This is modeled here by the random slopes $\varepsilon_{2,i,1}$.

Main Characteristics and Motivations. The examples of varying intercepts and varying slopes reveal the essential characteristics of a multilevel model: (1) varying coefficients and (2) a regression model for these varying coefficients (possibly using group-level predictors). Motivations for using multilevel modeling are numerous (see Gelman and Hill 2007); we illustrate many throughout this chapter. Because data often are clustered (e.g., students in schools, students in classes in schools, cars in fleets, policyholder data over time, policies within counties), statistical methodology should reflect the structure in the data and use it as relevant information when building statistical models. Using traditional (say linear or generalized linear models, as in Chapters 2 and 5) regression techniques, the clustering in groups is either ignored (*complete pooling*) or groups are analyzed separately (*no pooling*), resulting in overfitting because even small clusters will get their own regression model. The multilevel model enhances both extremes: for example, in the varying intercepts model from (8.7), complete pooling corresponds with $\sigma_2^2 \rightarrow 0$, and no pooling corresponds with $\sigma_2^2 \rightarrow \infty$. Multilevel modeling is a compromise between these two extremes, known as *partial pooling*. In this case, we impose a distributional assumption on $\varepsilon_{2,i}$ (with variance σ_2^2) and estimate σ_2^2 from the data. This takes heterogeneity among clusters into account, making appropriate cluster-specific predictions and structuring the dependence between observations belonging to the same cluster. Moreover, predictions related to new clusters become readily available. Whereas in classical regression cluster-specific indicators cannot be included along with cluster-specific predictors, multilevel models allow doing this in a convenient way (see (8.6)). When specifying regression models at different levels in the data, interactions between explanatory variables at different levels (so-called cross-level effects) may appear. This feature is often mentioned as another advantage of multilevel models.

What's in a Name?: Labels and Notation. Multilevel models carry many labels in statistical literature. They are sometimes called *hierarchical*, because data are often hierarchically structured (see the students in schools example) and because of the hierarchy in the model specifications. However, non-nested models, with levels structured next to each other instead of being hierarchically nested, can also be analyzed with the multilevel methodology. Multilevel models are also known as *random effects* or *mixed* models, because they combine (a mix of) fixed and random effects. This distinction is only applicable when using frequentist methodology and

terminology. A Bayesian analysis treats all regression parameters as random variables, specifying an appropriate prior distribution for each parameter.

In addition to terminology, mathematical notation can vary greatly among statistical sources. This should not be a surprise, because multilevel models can be formulated for any number of levels, involving nested and non-nested group (or cluster) effects, predictor information at different levels, and so on. For instance, Gelman and Hill (2007) denote the varying coefficients and varying slopes models in (8.4) + (8.6) and (8.10), respectively, in a more intuitive way:

$$\begin{aligned} y_i &= \alpha_{j[i]} + \beta x_i + \varepsilon_i, \quad i = 1, \dots, N \\ \alpha_j &= a + b u_j + \eta_j, \quad j = 1, \dots, m, \end{aligned} \tag{8.11}$$

and

$$\begin{aligned} y_i &= \alpha_{j[i]} + \beta_{j[i]} x_i + \varepsilon_i, \quad i = 1, \dots, N \\ \alpha_j &= a_0 + b_0 u_j + \eta_{j1}, \quad j = 1, \dots, m \\ \beta_j &= \eta_{j2}. \end{aligned} \tag{8.12}$$

Observations in the dataset are indexed with i , where N is the total number of observations j denotes the fleets in the dataset, and $j[i]$ is the fleet to which observation i belongs; x_i refers to covariate information available at vehicle-level (i.e., level 1 in (8.4)); and u_j refers to covariate information available at fleet-level (i.e., level 2 in (8.6)).

The notation used from Section 8.2 on is motivated by generality and is inspired by Frees (2004). This notation allows model equations to be written in a structured way, with clear reference to the particular level in the data to which the parameter/predictor is attached. Moreover, this notation can be used for any number of levels in a concise way. Section 8.2 explains the connection between this particular notation and the matrix notation (and corresponding manipulations) that is often developed in statistical literature on mixed models. When discussing examples, we replace this general notation with a more intuitive one, explicitly referring to the structure of the data under consideration.

8.1.2 Why? Motivating Examples from Actuarial Science

Research on mixed models originated in bio- and agricultural statistics. For example, the topic of variance components models, a particular example of models with random effects (see Searle, Casella, and McCulloch 2008), was studied extensively in the context of animal breeding experiments. The following (nonexhaustive) list of examples should convince the reader of the usefulness of mixed models as a modeling

tool in actuarial science, with applications ranging from ratemaking to reserving and smoothing. We deploy some of these examples within the framework of linear mixed models, whereas others are more appropriate for analysis with generalized linear mixed models.

Example 8.1 (Credibility Models). Credibility theory is an a posteriori ratemaking technique. Credibility models are designed to predict an insured's risk premium by weighting the insured's own loss experience and the experience in the overall portfolio. An extensive discussion of credibility models is available in Chapter 9. Credibility models have a natural and explicit interpretation as special examples of mixed models. Frees, Young, and Luo (1999) demonstrate this connection by reinterpreting credibility models using mixed model parlance. This mapping greatly increases the accessibility and usefulness of such models. Indeed, the complete machinery (including computational methods and software) of mixed models becomes available for the analysis of these actuarial models. The famous Hachemeister dataset (see Hachemeister 1975) has often been used in credibility literature. This dataset considers 12 periods, from the third quarter of 1970 to the second quarter of 1973, of bodily injury losses covered by a private passenger auto insurance; it registers the total loss and corresponding number of claims for five states. Figure 8.1 shows a trellis plot of the average loss per claim, followed over time, per state. The plot also shows a linear regression line and corresponding confidence intervals (in gray). In Section 8.3 we use linear mixed models to analyze this dataset and predict the next year's average claim per state. Further analysis – with focus on credibility theory – follows in Chapter 9.

Example 8.2 (Workers' Compensation Insurance: Losses). The dataset is from the National Council on Compensation Insurance (United States) and contains losses due to permanent partial disability (see Klugman 1992); 121 occupation or risk classes are observed over a period of seven years. The variable of interest is the `Loss` paid out (on a yearly basis) per risk class. Possible explanatory variables are `Year` and `Payroll`. Frees, Young, and Luo (2001) and Antonio and Beirlant (2007) present mixed models for the pure premium, $PP = Loss/Payroll$. For a random subsample of 10 risk classes, Figure 8.2 shows the time series plot of `Loss` (left) and corresponding `Payroll` (right).

Example 8.3 (Workers' Compensation Insurance: Frequencies). The data are from Klugman (1992); see Scollnik (1996), Makov et al. (1996), and Antonio and Beirlant (2007) for further discussion. Frequency counts in workers' compensation insurance are observed on a yearly basis for 133 occupation classes followed over 7 years. `Count` is the response variable of interest. Possible explanatory variables are `Year` and `Payroll`, a measure of exposure denoting scaled payroll totals adjusted for

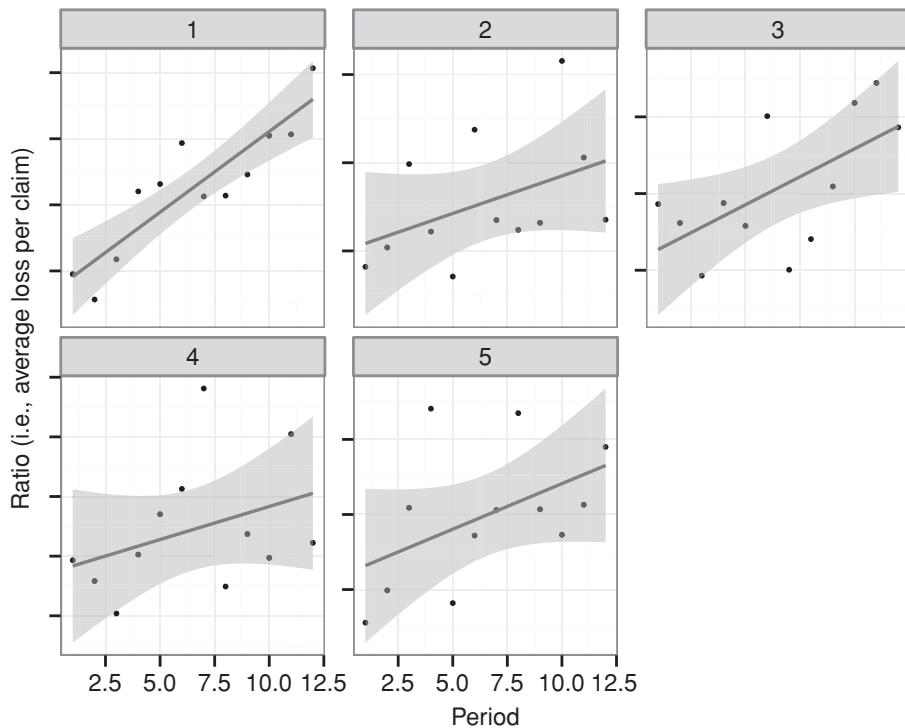


Fig. 8.1. Trellis plot of average losses per period and a linear regression line with corresponding confidence intervals (in gray); each panel represents one state. *Source:* Hachemeister (1975).

inflation. Figure 8.3 shows exploratory plots for a random subsample of 10 occupation classes. Statistical modeling should take into account the dependence between observations on the same occupation class and reflect the heterogeneity among different classes. In ratemaking (or tariffication) an obvious question for this example would be: *What is the expected number of claims for a risk class in the next observation period, given the observed claims history of this particular risk class and the whole portfolio?*

Example 8.4 (Hierarchical Data Structures). With panel data a group of subjects is followed over time, as in Examples 8.2 and 8.3. This is a basic and widely studied example of hierarchical data. Obviously, more complex structures may occur. Insurance data often come with some kind of *inherent hierarchy*. Motor insurance policies grouped in zip codes within counties within states are one example. Workers' compensation or fire insurance policies operating in similar industries or branches is another one. Consider the manufacturing versus education branch, with employees in manufacturing firms indicating larger claims frequencies, and restaurants versus stores,

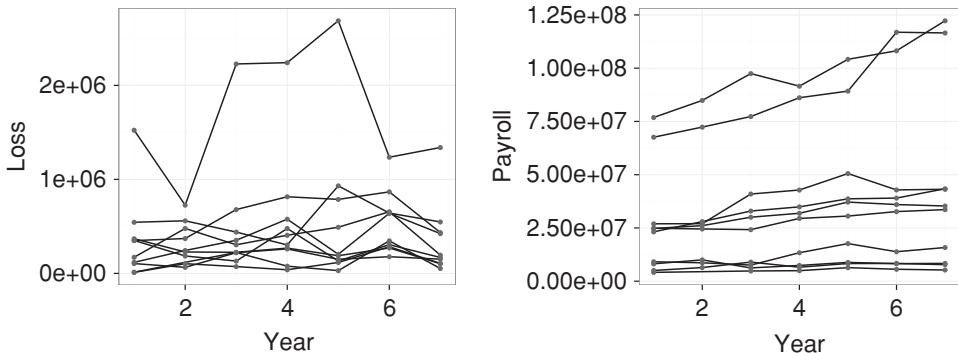


Fig. 8.2. Time series plot of losses (left) and payroll (right) for a random sample of 10 risk classes: workers' compensation data (losses).

with restaurants having a higher frequency of fire incidents than stores, and so on. A policyholder holding multiple policies (e.g., for theft, motor, flooding), followed over time, within the same company, is an example of a hierarchical data structure studied in the context of *multidimensional credibility* (see Bühlmann and Gisler 2005).

Another detailed multilevel analysis (going beyond the panel data structure) is Antonio, Frees, and Valdez (2010). These authors model claim count statistics for vehicles insured under a *fleet policy*. *Fleet policies* are umbrella-type policies issued to customers whose insurance covers more than a single vehicle. The hierarchical or multilevel structure of the data is as follows: vehicles (v) observed over time (t), nested within fleets (f), with policies issued by insurance companies (c). Multilevel models allow for incorporation of the hierarchical structure of the data by specifying random effects at vehicle, fleet, and company levels. These random effects represent unobservable characteristics at each level. At the vehicle level, the missions assigned to a

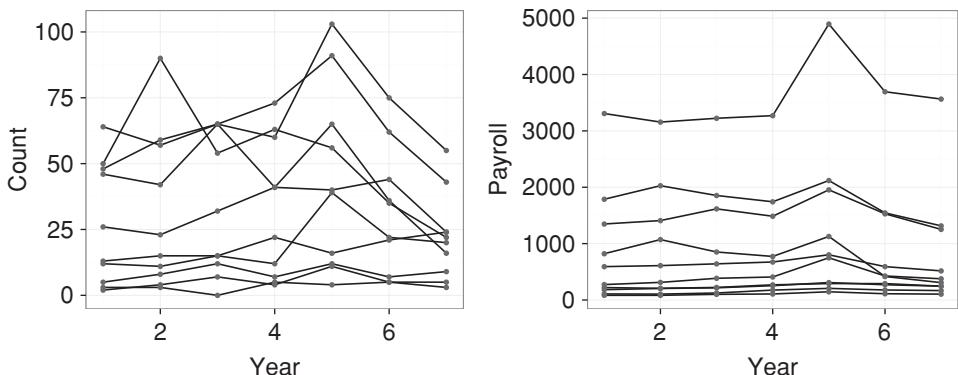


Fig. 8.3. Time series plot of counts (left) and payroll (right) for a random sample of 10 risk classes: workers' compensation data (counts).

Table 8.1. *Number of Payments (left) and Average Loss Per Combination of Status and Experience Risk Class: Credit Insurance Data*

Status	Experience			status	Experience		
	1	2	3		1	2	3
1	40	43	41	1	180.39	246.71	261.58
2	54	53	48	2	172.05	232.67	253.22
3	39	39	44	3	212.30	269.56	366.61

vehicle or unobserved driver behavior may influence the riskiness of a vehicle. At the fleet level, guidelines on driving hours, mechanical check-ups, loading instructions, and so on, may influence the number of accidents reported. At the insurance company level, underwriting and claim settlement practices may affect claims. Moreover, random effects allow a posteriori updating of an a priori tariff, by taking into account the past performance of vehicle, fleet, and company. As such, these models are relevant for a posteriori or experience rating with clustered data. See Antonio et al. (2010) and Antonio and Valdez (2012) for further discussion.

Example 8.5 (Non-Nested or Cross-Classified Data Structures). Data may also be structured in levels that are not nested or hierarchically structured, but instead act next to each other. An example is the dataset from Dannenburg, Kaas, and Goovaerts (1996) on private loans from a credit insurer. The data are payments of the credit insurer to several banks to cover losses caused by clients who were no longer able to pay off their loans. These payments are categorized by civil status of the debtors and their work experience. The civil status is single (1), divorced (2), or other (3), and the work experience is less than 2 years (<2 , category 1), from 2 up to 10 years (≥ 2 and < 10 , category 2), and more than 10 years (≥ 10 , category 3). Table 8.1 shows the number of clients and the average loss paid per risk class. Boxplots of the observed payments per risk class are in Figure 8.4. Using linear mixed models we estimate the expected loss per risk category and compare our results with the credibility premiums derived by Dannenburg et al. (1996).

Example 8.6 (Loss Reserving). Zhang, Dukic, and Guszczca (2012) analyze data from the workers' compensation line of business of 10 large insurers, as reported to the National Association of Insurance Commissioners.² Common accident years available are from 1988 to 1997. Losses are evaluated at 12-month intervals, with the highest available development age being 120 months. The data have a multilevel

² NAIC is a consortium of state-level insurance regulators in the United States.

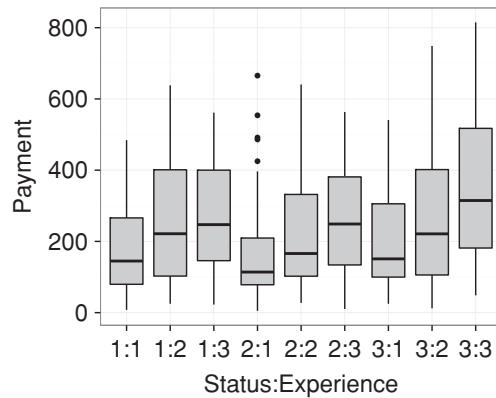


Fig. 8.4. Boxplots of payments versus combination of status and experience: credit insurance data.

structure with losses measured repeatedly over time, among companies and accident years. A plot of the cumulative loss over time for each company clearly shows a nonlinear growth pattern (see Figure 8.5). Predicting the development of these losses beyond the range of the available data is the major challenge in loss reserving. Figure 8.5 reveals that the use of a nonlinear growth curve model is an interesting path to explore. Random effects are included to structure heterogeneity among companies and accident years.

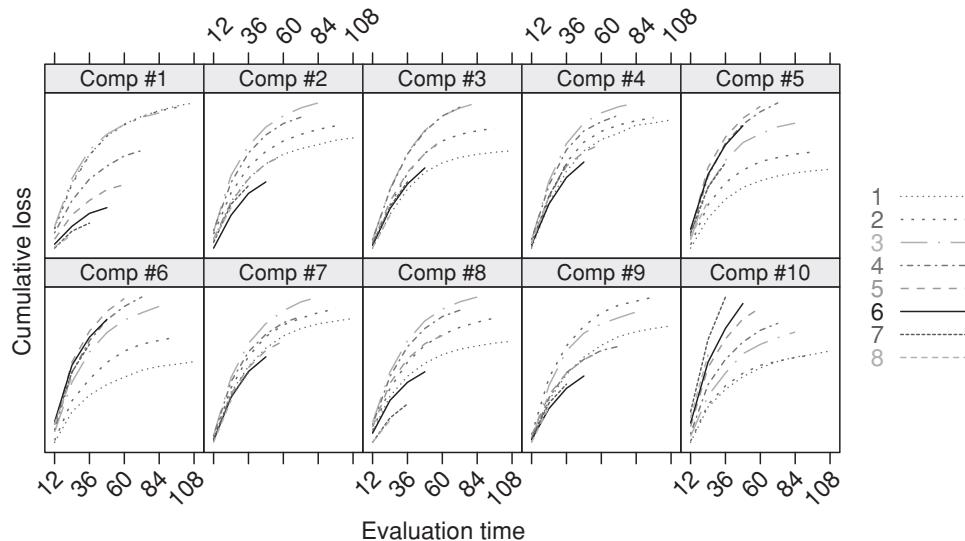


Fig. 8.5. Observed growth of cumulative losses for the 10 companies in the study. The colored lines represent accident years.

8.2 Linear Mixed Models

This section is based on Verbeke and Molenberghs (2000), McCulloch and Searle (2001), Ruppert, Wand, and Carroll (2003), Czado (2004), and Frees (2004).

8.2.1 Model Assumptions and Notation

The basic linear model specifies $E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$, with \mathbf{y} the response vector, $\boldsymbol{\beta}$ the vector of regression parameters, and \mathbf{X} the model design matrix. In traditional statistical parlance, all parameters in $\boldsymbol{\beta}$ are fixed (i.e., no distribution is assigned to them). They are unknown, but are fixed constants that should be estimated. In a linear mixed model we start from $\mathbf{X}\boldsymbol{\beta}$, but add $\mathbf{Z}\mathbf{u}$ to it, where \mathbf{Z} is a model matrix corresponding with a vector of random effects \mathbf{u} . A distribution is specified for this random effects vector \mathbf{u} with mean zero and covariance matrix \mathbf{D} . As discussed in Section 16.1 and illustrated later these random effects structure between-cluster heterogeneity and within-cluster dependence. All together, textbook notation for linear mixed models is as follows³:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \\ \mathbf{u} &\sim (\mathbf{0}, \mathbf{D}) \\ \boldsymbol{\varepsilon} &\sim (\mathbf{0}, \Sigma), \end{aligned} \tag{8.13}$$

with $\boldsymbol{\varepsilon}$ an $N \times 1$ vector of error terms with covariance matrix Σ (see later discussion for examples), which is independent of \mathbf{u} . This is the hierarchical specification of a linear mixed model. For a given \mathbf{u} , the conditional mean and variance are

$$\begin{aligned} E[\mathbf{y}|\mathbf{u}] &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \\ \text{Var}[\mathbf{y}|\mathbf{u}] &= \Sigma. \end{aligned} \tag{8.14}$$

The combined, unconditional, or *marginal* model states

$$\mathbf{y} \sim (\mathbf{X}\boldsymbol{\beta}, \mathbf{V} := \mathbf{Z}\mathbf{D}\mathbf{Z}' + \Sigma), \tag{8.15}$$

showing that fixed effects enter the (implied) mean of \mathbf{Y} and random effects structure the (implied) covariance matrix of \mathbf{y} .

Usually, normality is assumed for \mathbf{u} and $\boldsymbol{\varepsilon}$, thus,

$$\begin{pmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim N \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \Sigma \end{pmatrix} \right). \tag{8.16}$$

³ The notation $\mathbf{u} \sim (\mathbf{0}, \mathbf{D})$ implies $E[\mathbf{u}] = \mathbf{0}$ and $\text{Var}[\mathbf{u}] = \mathbf{D}$.

With these distributional assumptions the hierarchical LMM becomes

$$\begin{aligned} \mathbf{y}|\mathbf{u} &\sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \boldsymbol{\Sigma}) \\ \mathbf{u} &\sim N(\mathbf{0}, \mathbf{D}). \end{aligned} \quad (8.17)$$

This implies the marginal model $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, but not vice versa. When interest is only in the fixed effects parameters $\boldsymbol{\beta}$, the marginal model can be used. When there is explicit interest in $\boldsymbol{\beta}$ and \mathbf{u} , the specification in (8.13) and (8.17) should be used.

Examples 8.7 and 8.8 focus on particular examples of two and three level data and explain in detail the structure of vectors and matrices in (8.13) and (8.15).

Example 8.7 (A Two-Level Model for Longitudinal Data). Y_{ij} represents the j th measurement on a subject i (with $i = 1, \dots, m$ and $j = 1, \dots, n_i$); m is the number of subjects under consideration; and n_i the number of observations registered on subject i . \mathbf{x}_{ij} ($p \times 1$) is a column vector with fixed effects' covariate information from observation j on subject i . Correspondingly, \mathbf{z}_{ij} ($q \times 1$) is a column vector with covariate information corresponding with random effects. $\boldsymbol{\beta}$ ($p \times 1$) is a column vector with fixed effects parameters, and \mathbf{u}_i ($q \times 1$) is a column vector with random effects regression parameters. These are subject-specific and allow one to model heterogeneity between subjects. The combined model is

$$y_{ij} = \underbrace{\mathbf{x}'_{ij}\boldsymbol{\beta}}_{\text{fixed}} + \underbrace{\mathbf{z}'_{ij}\mathbf{u}_i}_{\text{random}} + \underbrace{\varepsilon_{ij}}_{\text{random}}. \quad (8.18)$$

The distributional assumptions for the random parts in (8.18) are

$$\begin{aligned} \mathbf{u}_i &\sim (\mathbf{0}, \mathbf{G}) \quad \mathbf{G} \in \mathbb{R}^{q \times q} \\ \varepsilon_i &\sim (\mathbf{0}, \boldsymbol{\Sigma}_i) \quad \boldsymbol{\Sigma}_i \in \mathbb{R}^{n_i \times n_i}. \end{aligned} \quad (8.19)$$

The covariance matrix \mathbf{G} is left unspecified (i.e., no particular structure is implied). Various structures are available for $\boldsymbol{\Sigma}_i$. Very often just a simple diagonal matrix is used: $\boldsymbol{\Sigma}_i := \sigma^2 I_{n_i}$. However, when the inclusion of random effects is not enough to capture the dependence between measurements on the same subject, we can add serial correlation to the model and specify $\boldsymbol{\Sigma}_i$ as nondiagonal (e.g., unstructured, Toeplitz, or autoregressive structure; see Verbeke and Molenberghs 2000, for more discussion). $\mathbf{u}_1, \dots, \mathbf{u}_m, \varepsilon_1, \dots, \varepsilon_m$ are independent. Typically, normality is assumed for both vectors, as in (8.17). In vector notation we specify

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \varepsilon_i, \quad i = 1, \dots, m, \\ \mathbf{u}_i &\sim (\mathbf{0}, \mathbf{G}) \\ \varepsilon_i &\sim (\mathbf{0}, \boldsymbol{\Sigma}_i), \end{aligned} \quad (8.20)$$

where

$$\mathbf{X}_i := \begin{pmatrix} \mathbf{x}'_{i1} \\ \vdots \\ \mathbf{x}'_{in_i} \end{pmatrix} \in \mathbb{R}^{n_i \times p}, \quad \mathbf{Z}_i = \begin{pmatrix} \mathbf{z}'_{i1} \\ \vdots \\ \mathbf{z}'_{in_i} \end{pmatrix} \in \mathbb{R}^{n_i \times q}, \quad \mathbf{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{pmatrix} \in \mathbb{R}^{n_i \times 1}. \quad (8.21)$$

All subjects or clusters $i = 1, \dots, m$, (8.13) are combined in the matrix formulation of this LMM for longitudinal data (with $N = \sum_{i=1}^m n_i$ being the total number of observations):

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_m \end{pmatrix} \in \mathbb{R}^{N \times 1}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{pmatrix} \in \mathbb{R}^{N \times p}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{pmatrix} \in \mathbb{R}^{N \times 1},$$

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0}_{n_1 \times q} & \dots & \mathbf{0}_{n_1 \times q} \\ \mathbf{0}_{n_2 \times q} & \mathbf{Z}_2 & & \\ \vdots & & \ddots & \\ \mathbf{0}_{n_m \times q} & & & \mathbf{Z}_m \end{pmatrix} \in \mathbb{R}^{N \times (m \cdot q)}, \quad \mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_m \end{pmatrix} \in \mathbb{R}^{(m \cdot q) \times 1}. \quad (8.22)$$

The covariance matrix of the combined random effects vector \mathbf{u} , on the one hand, and the combined residual vector $\boldsymbol{\varepsilon}$, on the other hand, are specified as

$$\mathbf{D} = \begin{pmatrix} \mathbf{G} & & \\ & \ddots & \\ & & \mathbf{G} \end{pmatrix} \in \mathbb{R}^{m \cdot q \times m \cdot q}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & & \\ & \ddots & \\ & & \boldsymbol{\Sigma}_m \end{pmatrix} \in \mathbb{R}^{N \times N}. \quad (8.23)$$

Covariance matrix \mathbf{V} in this particular example is block diagonal and is given by

$$\begin{aligned} \mathbf{V} &= \mathbf{Z} \mathbf{D} \mathbf{Z}' + \boldsymbol{\Sigma} \\ &= \begin{pmatrix} \mathbf{Z}_1 \mathbf{G} \mathbf{Z}'_1 + \boldsymbol{\Sigma}_1 & \dots & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{Z}_m \mathbf{G} \mathbf{Z}'_m + \boldsymbol{\Sigma}_m \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{V}_1 & & \\ & \ddots & \\ & & \mathbf{V}_m \end{pmatrix}, \end{aligned} \quad (8.24)$$

with $\mathbf{V}_i = \mathbf{Z}_i \mathbf{G} \mathbf{Z}'_i + \boldsymbol{\Sigma}_i$.

Example 8.8 (A Three-Level Example.). y_{ijk} is the response variable of interest, as observed for, say, vehicle k , insured in fleet j by insurance company i . At vehicle-level

(or level 1), we model this response as

$$y_{ijk} = z'_{1,ijk} \beta_{ij} + x'_{1,ijk} \beta_1 + \varepsilon_{1,ijk}. \quad (8.25)$$

Hereby, predictors $z_{1,ijk}$ and $x_{1,ijk}$ may depend on the insurance company, fleet, or vehicle. β_1 is a vector of regression parameters that do not vary by company or fleet; they are fixed effects regression parameters. Parameters β_{ij} vary by company and fleet. We model them in a level-two equation:

$$\beta_{ij} = Z_{2,ij} \gamma_i + X_{2,ij} \beta_2 + \varepsilon_{2,ij}. \quad (8.26)$$

$X_{2,ij}$ and $Z_{2,ij}$ may depend on the company or fleet, but not on the insured vehicle. The regression parameters in γ_i are company-specific and modeled in (8.27):

$$\gamma_i = X_{3i} \beta_3 + \varepsilon_{3i}, \quad (8.27)$$

where the predictors in X_{3i} may depend on the company, but not on the fleet or vehicle. The combined level 1, 2, and 3 models lead to the following model specification:

$$\begin{aligned} Y_{ijk} &= z'_{1,ijk} (Z_{2,ij} (X_{3i} \beta_3 + \varepsilon_{3i}) + X_{2,ij} \beta_2 + \varepsilon_{2,ij}) + x'_{1,ijk} \beta_1 + \varepsilon_{1,ijk} \\ &= x'_{ijk} \beta + z'_{ijk} u_{ij} + \varepsilon_{1,ijk}, \end{aligned} \quad (8.28)$$

where $x'_{ijk} = (x'_{1,ijk} \quad z'_{1,ijk} X_{2,ij} \quad z'_{1,ijk} Z_{2,ij} X_{3i})$, $\beta = (\beta'_1 \quad \beta'_2 \quad \beta'_3)'$, $z'_{i,j,k} = (z'_{1,i,j,k} \quad z'_{1,i,j,k} Z_{2,ij})$ and $u_{ij} = (\varepsilon'_{2,ij} \quad \varepsilon'_{3i})'$. Formulating this three-level model in matrix notation follows from stacking all observations Y_{ijk} .

More examples of LMM specifications are in McCulloch and Searle (2001). A standard notation for a k -level model is in Frees (2004; see Chapter 5, Appendix A).

8.2.2 The Structure of Random Effects

Since the random effects u often correspond to factor predictors, the design matrix Z is often highly sparse, with a high proportion of elements being exactly zero. Moreover, the covariance matrix D is highly structured and depends on some parameter vector θ that is to be estimated. We consider the following structures for the random effects:

- **Single random effect per level:** This is the simplest yet most common case where the random effect corresponds to a certain level of a single grouping factor. For example, we may have the state indicator in the model, and each state has its own intercept (i.e., $y \sim (1 | \text{state})$ (in R parlance)). We illustrate this structure in Section 8.3 with the workers' compensation losses data.
- **Multiple random effects per level:** Another common case is that the model has both random intercepts and random slopes that vary by some grouping factor. For example, each state in the model has its own intercept and also its own slope with respect to some predictor (i.e., $y \sim (1 + \text{time} | \text{state})$). In general, the multiple random effects are

correlated, and so the matrix \mathbf{D} is not diagonal. We illustrate this structure in Section 8.3 with the workers' compensation losses data.

- **Nested random effects:** In the nested classification, some levels of one factor occur only within certain levels of a first factor. For example, we may have observations within each county and then of the counties within each state. The county from state A never occurs for state B , so counties are nested within states, forming a hierarchical structure (i.e., $y \sim (1 | \text{county}/\text{state})$). Antonio et al. (2010) provide an example of this type of structuring.
- **Crossed random effects:** This happens when each level of each factor may occur with each level of each other factor. For example, we may have both state and car make in the model, and cars of different makes can occur within each state (i.e., $y \sim (1 | \text{state}) + (1 | \text{make})$). The credit insurance example in Section 8.3 is an example of crossed random effects.

8.2.3 Parameter Estimation, Inference, and Prediction

Mixed models use a combination of fixed effects regression parameters, random effects, and covariance matrix parameters (also called *variance components*). For example, in the varying intercepts example from (8.4) and (8.5), $\beta_{1,0}$ and $\beta_{1,1}$ are regression parameters corresponding with fixed effects, σ_1^2 and σ_2^2 are variance components, and $\varepsilon_{2,i}$ ($i = 1, \dots, m$) are the random effects. We use standard statistical methodology, such as maximum likelihood, to estimate parameters in a LMM. For the random effects we apply statistical knowledge concerning *prediction* problems (for an overview, see McCulloch and Searle 2001). The difference in terminology stems from the nonrandomness of the parameters versus the randomness of the random effects.

We first derive an estimator for the fixed effects parameters in $\boldsymbol{\beta}$ and a predictor for the random effects in \mathbf{u} , under the assumption of known covariance parameters in \mathbf{V} (see (8.15)).

Estimating $\boldsymbol{\beta}$. The generalized least squares (GLS) estimator – which coincides with the maximum likelihood estimator (MLE) under normality (as in (8.17)) – of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{y}. \quad (8.29)$$

See Frees (2004) or Czado (2004) for a formal derivation of this result.

Predicting \mathbf{u} . In the sense of the minimal mean squared error of prediction (MSEP), the best predictor (BP) of \mathbf{u} is the conditional mean $E[\mathbf{u}|Y]$. This predictor obviously requires knowledge of the conditional distribution $\mathbf{u}|Y$. The BP is often simplified by restricting the predictor to be a linear function of Y : the best linear predictor (BLP).

The BLP of a random vector \mathbf{u} is

$$\text{BLP}[\mathbf{u}] = \hat{\mathbf{u}} = \mathbb{E}[\mathbf{u}] + \mathbf{C}\mathbf{V}^{-1}(\mathbf{y} - \mathbb{E}[\mathbf{y}]), \quad (8.30)$$

where $\mathbf{V} = \text{Var}(\mathbf{y})$ and $\mathbf{C} = \text{Cov}(\mathbf{u}, \mathbf{y}')$. $\text{BP}(\mathbf{u})$ and $\text{BLP}(\mathbf{u})$ are unbiased, in the sense that their expected value equals $\mathbb{E}[\mathbf{u}]$. Normality is not required in BP or BLP, but with (\mathbf{y}, \mathbf{u}) being multivariate normally distributed, the BP and BLP coincide. See McCulloch and Searle (2001) and Chapter 9 for more details.

In the context of the LMM sketched in (8.17) the predictor of \mathbf{u} is usually called the best linear unbiased predictor (BLUP). Robinson (1991) describes several ways to derive this BLUP. For instance, under normality assumptions,

$$\begin{aligned} \text{Cov}(\mathbf{y}, \mathbf{u}') &= \text{Cov}(X\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \mathbf{u}') \\ &= \text{Cov}(X\boldsymbol{\beta}, \mathbf{u}') + \mathbf{Z}\text{Var}(\mathbf{u}, \mathbf{u}') + \text{Cov}(\boldsymbol{\epsilon}, \mathbf{u}') \\ &= \mathbf{Z}\mathbf{D}, \end{aligned}$$

which leads to the multivariate normal distribution

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{u} \end{pmatrix} \sim N \left(\begin{pmatrix} X\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{V} & \mathbf{Z}\mathbf{D} \\ \mathbf{Z}'\mathbf{D} & \mathbf{D} \end{pmatrix} \right). \quad (8.31)$$

Using either properties of this distribution⁴ or the result in (8.30), the BLUP of \mathbf{u} follows:

$$\text{BLUP}(\mathbf{u}) := \hat{\mathbf{u}} = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - X\boldsymbol{\beta}). \quad (8.32)$$

Of course, (8.32) relies on the (unknown) vector of fixed effects $\boldsymbol{\beta}$, as well as on unknown covariance parameters in \mathbf{V} . When we replace both with their estimates, we call the BLUP an empirical or estimated BLUP. Estimated BLUPs are confronted with multiple sources of variability, including from the estimation of $(\boldsymbol{\beta}, \mathbf{u})$ of \mathbf{V} . Histograms and scatterplots of components of $\hat{\mathbf{u}}$ are often used to detect outlying clusters or to visualize between-cluster heterogeneity.

A Unified Approach: Henderson's Justification. Maximizing the joint log likelihood of $(\mathbf{y}', \mathbf{u}')$ (see assumptions (8.17)) with respect to $(\boldsymbol{\beta}, \mathbf{u})$ leads to Henderson's mixed model equations:

$$\begin{aligned} f(\mathbf{y}, \mathbf{u}) &= f(\mathbf{y}|\mathbf{u}) \cdot f(\mathbf{u}) \\ &\propto \exp \left(-\frac{1}{2}(\mathbf{y} - X\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - X\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) \right) \cdot \exp \left(-\frac{1}{2}\mathbf{u}'\mathbf{D}^{-1}\mathbf{u} \right). \end{aligned} \quad (8.33)$$

⁴ Namely, with $\mathbf{X} = \begin{pmatrix} \mathbf{Y} \\ \mathbf{Z} \end{pmatrix} \sim N \left(\begin{pmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_Z \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma}_{YZ} \\ \boldsymbol{\Sigma}_{ZY} & \boldsymbol{\Sigma}_Z \end{pmatrix} \right)$ we know $\mathbf{Z}|Y \sim N(\boldsymbol{\mu}_{Z|Y}, \boldsymbol{\Sigma}_{Z|Y})$ where $\boldsymbol{\mu}_{Z|Y} = \boldsymbol{\mu}_Z + \boldsymbol{\Sigma}_{ZY}\boldsymbol{\Sigma}_Y^{-1}(\mathbf{Y} - \boldsymbol{\mu}_Y)$ and $\boldsymbol{\Sigma}_{Z|Y} = \boldsymbol{\Sigma}_Z - \boldsymbol{\Sigma}_{ZY}\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\Sigma}_{YZ}$.

It is therefore enough to minimize

$$Q(\boldsymbol{\beta}, \mathbf{u}) := (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \mathbf{u}' \mathbf{D}\mathbf{u}, \quad (8.34)$$

which corresponds to solving the set of equations

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} Q(\boldsymbol{\beta}, \mathbf{u}) &= \mathbf{0} \text{ and } \frac{\partial}{\partial \mathbf{u}} Q(\boldsymbol{\beta}, \mathbf{u}) = \mathbf{0} \\ \Leftrightarrow \begin{pmatrix} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X} & \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{Z} \\ \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{X} & \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{Z} + \mathbf{D}^{-1} \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{pmatrix} &= \begin{pmatrix} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{y} \\ \mathbf{Z}' \boldsymbol{\Sigma}^{-1} \mathbf{y} \end{pmatrix}. \end{aligned} \quad (8.35)$$

(8.29) and (8.32) solve this system of equations.

More on Prediction. With $\hat{\boldsymbol{\beta}}$ from (8.29) and $\hat{\mathbf{u}}$ from (8.32), the profile of cluster i is predicted by

$$\begin{aligned} \hat{\mathbf{y}}_i &:= \mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{Z}_i \hat{\mathbf{u}}_i \\ &= \mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}'_i \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{\Sigma}_i \mathbf{V}_i^{-1} \mathbf{X}_i \hat{\boldsymbol{\beta}} + (\mathbf{I}_{n_i} - \boldsymbol{\Sigma}_i \mathbf{V}_i^{-1}) \mathbf{Y}_i, \end{aligned} \quad (8.36)$$

using $\mathbf{V}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}'_i + \boldsymbol{\Sigma}_i$ and n_i the cluster size. $\hat{\mathbf{y}}_i$ is a weighted mean of the global profile $\mathbf{X}_i \hat{\boldsymbol{\beta}}$ and the data observed on cluster i , \mathbf{y}_i . $\hat{\mathbf{y}}_i$ is a so-called *shrinkage estimator*. Actuaries will recognize a credibility type formula in (8.36).

The prediction of a future observation is discussed in detail in Frees (2004 and in Section 4.4). The case of nondiagonal residual covariance matrices $\boldsymbol{\Sigma}_i$ requires special attention. For instance, with panel data the BLUP for y_{i,T_i+1} is $\mathbf{x}'_{i,T_i+1} \boldsymbol{\beta} + \mathbf{z}'_{i,T_i+1} \hat{\mathbf{u}}_i + \text{BLUP}(\varepsilon_{i,T_i+1})$. From (8.30) we understand that the last term in this expression is zero when $\text{Cov}(\varepsilon_{i,T_i+1}, \boldsymbol{\varepsilon}_i) = \mathbf{0}$. This is not the case when serial correlation is taken into account. Chapter 9 carefully explains this kind of prediction problems.

Estimating Variance Parameters. The parameters or variance components used in \mathbf{V} are in general unknown and should be estimated from the data. With $\boldsymbol{\theta}$ the vector of unknown parameters used in $\mathbf{V} = \mathbf{Z} \mathbf{D}(\boldsymbol{\theta}) \mathbf{Z}' + \mathbf{D}(\boldsymbol{\theta})$, the log-likelihood for $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is (with c a constant)

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \log \{L(\boldsymbol{\beta}, \boldsymbol{\theta})\} \\ &= -\frac{1}{2} \left(\ln |\mathbf{V}(\boldsymbol{\theta})| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}(\boldsymbol{\theta})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right) + c. \end{aligned} \quad (8.37)$$

Maximizing (8.37) with respect to $\boldsymbol{\beta}$ and with $\boldsymbol{\theta}$ fixed, we get

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\mathbf{X}' \mathbf{V}(\boldsymbol{\theta})^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}(\boldsymbol{\theta})^{-1} \mathbf{y}. \quad (8.38)$$

We obtain the so-called *profile log-likelihood* by replacing $\boldsymbol{\beta}$ in (8.37) with $\hat{\boldsymbol{\beta}}$ from (8.38):

$$\begin{aligned}\ell_p(\boldsymbol{\theta}) &:= \ell(\hat{\boldsymbol{\beta}}, \boldsymbol{\theta}) \\ &= -\frac{1}{2} \left\{ \ln |\mathbf{V}(\boldsymbol{\theta})| + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}))' \mathbf{V}(\boldsymbol{\theta})^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})) \right\}. \quad (8.39)\end{aligned}$$

Maximizing this profile log-likelihood with respect to $\boldsymbol{\theta}$ gives the maximum likelihood estimates $\hat{\boldsymbol{\theta}}_{MLE}$ of the variance components in $\boldsymbol{\theta}$.

With LMMs, restricted (or residual) maximum likelihood (REML) is a popular alternative to estimate $\boldsymbol{\theta}$. REML accounts for the degrees of freedom used for fixed effects estimation. McCulloch and Searle (2001), (section 6.10) is an overview of important arguments in the discussion, *ML versus REML*? For example, estimates with REML (for balanced data) are minimal variance unbiased under normality⁵ and are invariant to the value of $\boldsymbol{\beta}$. The REML estimation of $\boldsymbol{\theta}$ is based on the marginal log-likelihood obtained by integrating out the fixed effects in $\boldsymbol{\beta}$:

$$\ell_r(\boldsymbol{\theta}) := \ln \left(\int L(\boldsymbol{\beta}, \boldsymbol{\theta}) d\boldsymbol{\beta} \right), \quad (8.41)$$

where (see Czado 2004)

$$\begin{aligned}\int L(\boldsymbol{\beta}, \boldsymbol{\theta}) d\boldsymbol{\beta} &= \int \frac{1}{(2\pi)^{N/2}} |\mathbf{V}(\boldsymbol{\theta})|^{-1/2} \exp \left(-\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}(\boldsymbol{\theta})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right) d\boldsymbol{\beta} \\ &\vdots \\ &= \ell_p(\boldsymbol{\theta}) - \frac{1}{2} \ln |\mathbf{X}' \mathbf{V}(\boldsymbol{\theta})^{-1} \mathbf{X}| + \text{constants}. \quad (8.42)\end{aligned}$$

8.2.3.1 Standard Errors and Inference

Estimation of Standard Errors. In the marginal model $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\theta}))$, the covariance of $\hat{\boldsymbol{\beta}}$ in (8.29) is

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}' \mathbf{V}^{-1}(\boldsymbol{\theta}) \mathbf{X})^{-1}, \quad (8.43)$$

where $\text{Cov}(\mathbf{y}) = \mathbf{V}(\boldsymbol{\theta})$ is used. Replacing the unknown $\boldsymbol{\theta}$ with its ML or REML estimate $\hat{\boldsymbol{\theta}}$ and using $\hat{\mathbf{V}} := \mathbf{V}(\hat{\boldsymbol{\theta}})$, a natural estimate for $\text{Cov}(\hat{\boldsymbol{\beta}})$ is $(\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1}$.

⁵ A well-known example of *REML versus ML* considers the case of a random sample $X_1, \dots, X_N \sim N(\mu, \sigma^2)$. The resulting estimators for the unknown variance σ^2 are

$$\hat{\sigma}_{ML}^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2, \quad \hat{\sigma}_{REML}^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2, \quad (8.40)$$

with \bar{X} the sample mean. The REML estimator is unbiased for σ^2 . The $(N-1)$ in $\hat{\sigma}_{REML}^2$ accounts for the estimation of μ by \bar{X} .

However, this estimate ignores the extra variability originating from the estimation of $\boldsymbol{\theta}$. Kacker and Harville (1984) among others discuss attempts to quantify this extra variability through approximation, but only a fully Bayesian analysis accounts for all sources of variability (see Chapter 16 for a demonstration of a Bayesian analysis of a generalized linear mixed model).

The covariance of the empirical BLUP in (8.32) is equal to

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{u}}) &= \text{Cov}(\mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})) \\ &= \mathbf{D}\mathbf{Z}' \left\{ \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} \right\} \mathbf{Z}\mathbf{D}.\end{aligned}\quad (8.44)$$

However, the estimator in (8.44) ignores the variability in the random vector \boldsymbol{u} . Therefore, as suggested by Laird and Ware (1982), inference for \boldsymbol{u} is usually based on $\text{Cov}(\hat{\boldsymbol{u}} - \boldsymbol{u})$. Estimates of the precision of other predictors involving $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$ are based on

$$\text{Cov} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} - \boldsymbol{u} \end{bmatrix}, \quad (8.45)$$

and are available in McCulloch and Searle (2001, section 9.4 (c)). Accounting for the variability induced by estimating the variance components $\boldsymbol{\theta}$ would require – once again – a fully Bayesian analysis. Using Bayesian statistics, posterior credible intervals of cluster-specific effects follow immediately. They are useful in understanding the between-cluster heterogeneity present in the data.

Inference. We consider testing a set of s ($s \leq p$) hypotheses concerning the fixed effects parameters in $\boldsymbol{\beta}$:

$$\begin{aligned}H_0 : \mathbf{C}\boldsymbol{\beta} &= \boldsymbol{\zeta} \\ \text{versus } H_1 : \mathbf{C}\boldsymbol{\beta} &\neq \boldsymbol{\zeta}.\end{aligned}\quad (8.46)$$

The Wald test statistic

$$[\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\zeta}]'[\mathbf{CVar}(\hat{\boldsymbol{\beta}})\mathbf{C}'][\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\zeta}] \quad (8.47)$$

is approximately χ_s^2 distributed. With $\ell(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Sigma}})$ the log-likelihood obtained with ML in the restricted model (i.e., under H_0) and $\ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}})$ the log-likelihood with ML in the unrestricted model, the likelihood ratio test (LRT) statistic for nested models

$$-2[\ell(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Sigma}}) - \ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}})], \quad (8.48)$$

is approximately χ_s^2 distributed. Estimation should be done with ML instead of REML, because REML maximizes the likelihood of linear combinations of \mathbf{Y} that do not depend on $\boldsymbol{\beta}$.

Testing the necessity of random effects requires a hypothesis test involving the variance components. For example, in the varying intercepts model from (8.7), we want to investigate whether the intercepts of different subjects are significantly different. This corresponds with

$$H_0 : \sigma_2^2 = 0 \quad \text{versus} \quad H_1 : \sigma_2^2 > 0. \quad (8.49)$$

However, because 0 is on the boundary of the allowed parameter space for σ_2^2 , the likelihood ratio test statistic should not be compared with a χ^2_1 distribution, but with a mixture $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$. When testing a hypothesis involving s fixed effects parameters and one variance component, the reference distribution is $\frac{1}{2}\chi^2_s + \frac{1}{2}\chi^2_{s+1}$. When more variance components are involved, the complexity of this problem increases; see Ruppert et al. (2003) and related work from these authors.

8.3 Examples

8.3.1 Workers' Compensation Insurance Losses

We analyze the data from Example 8.2 on losses observed for workers' compensation insurance risk classes. The variable of interest is Loss_{ij} observed per risk class i and year j . The distribution of the losses is right skewed, which motivates the use of $\log(\text{Loss}_{ij})$ as the response variable. To enable out-of-sample predictions, we split the dataset into a training (without Loss_{i7}) versus test set (the Loss_{i7} observations). We remove observations corresponding with zero payroll from the dataset. Models are estimated on the training set, and centering of covariate `Year` is applied. Throughout our analysis we include $\log(\text{Payroll})_{ij}$ as an offset in the regression models, since losses should be interpreted relative to the size of the risk class.

Complete Pooling. We start with the complete pooling model introduced in (8.1). The model ignores the clustering of data in risk classes and fits an overall intercept (β_0) and an overall slope (β_1) for the effect of `Year`:

$$\log(\text{Loss}_{ij}) = \log(\text{Payroll}_{ij}) + \beta_0 + \beta_1 \text{Year}_{ij} + \varepsilon_{ij} \quad (8.50)$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \quad i.i.d. \quad (8.51)$$

We fit the model with `lm` in R.

```
>fitglm.CP <- lm(log(loss) ~ yearcentr, offset=log(payroll) ,
  data=wclassFit)
>summary(fitglm.CP)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -4.34023    0.04105 -105.733   <2e-16 ***
yearcentr    0.03559    0.02410     1.477     0.14
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 1.062 on 667 degrees of freedom
Multiple R-squared:  0.7282,    Adjusted R-squared:  0.7278
F-statistic: 1787 on 1 and 667 DF,  p-value: < 2.2e-16
```

According to this R output, $\hat{\beta}_0 = -4.34$ (with s.e. 0.041), $\hat{\beta}_1 = 0.036$ (with s.e. 0.024), and $\hat{\sigma}_\epsilon = 1.062$.

No Pooling. The fixed effects linear regression model in (8.2) estimates an intercept for each of the 118 risk classes in the dataset. According to model equation (8.52), the intercepts $\beta_{0,i}$ are unknown, but fixed, whereas the error terms ε_{ij} are stochastic:

$$\begin{aligned} \log(\text{Loss}_{ij}) &= \log(\text{Payroll}_{ij}) + \beta_{0,i} + \beta_1 \text{Year}_{ij} + \varepsilon_{ij} \\ \varepsilon_{ij} &\sim N(0, \sigma_\varepsilon^2) \quad i.i.d. \end{aligned} \quad (8.52)$$

We fit this model in R by identifying the risk class variable as a factor variable.

```
>fitglm.NP <- lm(log(loss) ~ 0 + yearcentr + factor(riskclass),
  offset = log(payroll),
  data = wclossFit)
>summary(fitglm.NP)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
yearcentr      0.03843   0.01253   3.067  0.00227 **
factor(riskclass)1 -3.49671   0.22393 -15.615 < 2e-16 ***
factor(riskclass)2 -3.92231   0.22393 -17.516 < 2e-16 ***
factor(riskclass)3 -4.48135   0.22393 -20.012 < 2e-16 ***
factor(riskclass)4 -4.70981   0.22393 -21.032 < 2e-16 ***
...
Residual standard error: 0.5485 on 550 degrees of freedom
Multiple R-squared:  0.9986,    Adjusted R-squared:  0.9983
F-statistic: 3297 on 119 and 550 DF,  p-value: < 2.2e-16
```

The null hypothesis of equal intercepts, $H_0 : \beta_{0,1} = \beta_{0,2} = \dots = \beta_{0,118} = \beta_0$, is rejected (with p -value < 0.05). Therefore, the no pooling model significantly improves on the complete pooling model.

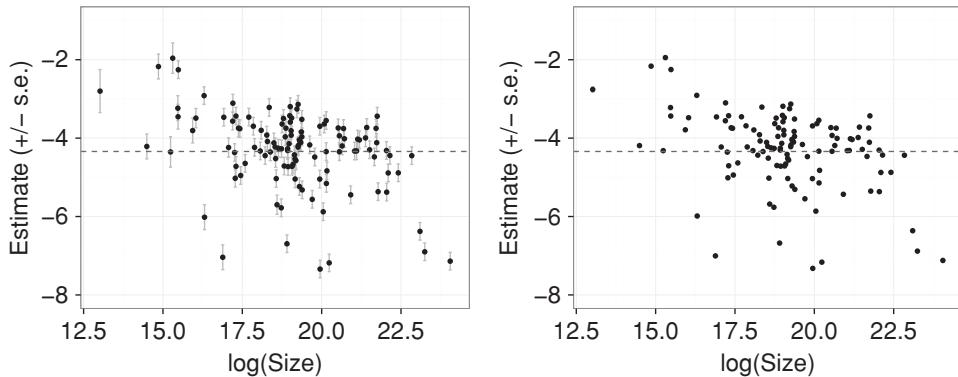


Fig. 8.6. Point estimates for risk class specific intercepts, plus/minus one standard error. Results from no pooling approach (left) and linear mixed model (right). The dashed line is $y = -4.34$ (i.e., the overall intercept from the complete pooling model).

```
> anova(fitglm.CP, fitglm.NP)
Analysis of Variance Table

Model 1: log(loss) ~ yearcentr
Model 2: log(loss) ~ 0 + yearcentr + factor(riskclass)
  Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1     667 751.90
2     550 165.48 117      586.42 16.658 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

Figure 8.6 (left) shows the estimates $\hat{\beta}_{0,i}$, plus/minus one standard error, against the size (on log-scale) of the risk class. The size of a risk class is here defined as $\sum_{j=1}^6 \text{Payroll}_{ij}$. The no pooling model estimates risk class specific intercepts with reasonable precision.

Linear Mixed Models: Random Intercepts. A linear mixed model with random risk class specific intercepts is a meaningful alternative to the no pooling model in (8.52). The regression equation is

$$\begin{aligned} \log(\text{Loss}_{ij}) &= \log(\text{Payroll}_{ij}) + \beta_0 + u_{0,i} + \beta_1 \text{Year}_{ij} + \varepsilon_{ij} \\ u_{0,i} &\sim N(0, \sigma_u^2) \quad i.i.d. \\ \varepsilon_{ij} &\sim N(0, \sigma_\varepsilon^2) \quad i.i.d. \end{aligned} \tag{8.53}$$

Random intercepts $u_{0,i}$ are independent across risk classes and independent of the error terms ε_{ij} . In R we use the `lme4` package to fit this linear mixed model. The

package uses REML by default. Results with ML follow by adding `REML=FALSE` in the `lmer(...)` statement.

```
> lmm1 <- lmer(log(loss) ~ (1|riskclass)+yearcentr+offset
  (log(payroll)),
  data=wcrossFit)
> print(lmm1)
Linear mixed model fit by REML
Formula: log(loss) ~ (1 | riskclass) + yearcentr + offset
  (log(payroll))
Data: wclossFit
AIC  BIC logLik deviance REMLdev
1448 1466 -720.2     1431    1440
Random effects:
Groups      Name        Variance Std.Dev.
riskclass (Intercept) 0.88589  0.94122
Residual            0.30145  0.54904
Number of obs: 669, groups: riskclass, 118

Fixed effects:
          Estimate Std. Error t value
(Intercept) -4.31959   0.08938 -48.33
yearcentr    0.03784   0.01253   3.02

Correlation of Fixed Effects:
  (Intr)
yearcentr  0.001
```

The R output shows the following parameter estimates: $\hat{\beta}_0 = -4.32$ (s.e. 0.089), $\hat{\beta}_1 = 0.037$ (s.e. 0.013), $\hat{\sigma}_u = 0.94$, and $\hat{\sigma}_\epsilon = 0.55$. In Figure 8.6 (right) we plot the point predictions for the $u_{i,0}$'s, and their corresponding standard errors, against size of the risk class. To create this plot we refit the linear mixed model and do not include an intercept.

The point estimates of the random intercepts obtained with the no pooling model in (8.52) and the linear mixed model in (8.53) are similar in this example. For the standard errors of the random intercepts in the LMM, we use the following instructions

```
str(rr1 <- ranef(lmm0, condVar = TRUE))
my.se.risk = sqrt(as.numeric(attributes(rr1$riskclass)$postVar)),
```

which calculate the variance of $u|y$, conditional on the maximum likelihood estimates for β and θ . Thus, these standard errors are different from the approach outlined in

(8.44). We are aware of the fact that they do not account for all sources of variability involved.

Linear Mixed Models: Random Intercepts and Slopes. We now extend the LMM in (8.53) and allow for random slopes as well as random intercepts. This is an example of the *multiple random effects per level* setting from Section 8.2.2. The model equation is

$$\begin{aligned} \log(\text{Loss}_{ij}) &= \log(\text{Payroll}_{ij}) + \beta_0 + u_{0,i} + \beta_1 \text{Year}_{ij} + u_{1,i} \text{Year}_{ij} + \varepsilon_{ij}, \\ \boldsymbol{u}_i &\sim N(\boldsymbol{0}, \boldsymbol{D}(\boldsymbol{\theta})) \quad i.i.d. \\ \varepsilon_{ij} &\sim N(0, \sigma_\varepsilon^2) \quad i.i.d. \end{aligned} \tag{8.54}$$

The random effects vector \boldsymbol{u}_i is now bivariate, say with $\text{Var}(u_{i,0}) = \theta_0$, $\text{Var}(u_{i,1}) = \theta_1$, and $\text{Cov}(u_{i,0}, u_{i,1}) = \theta_{01}$. Random effects are independent across risk classes and independent of the error terms ε_{ij} . We fit this model with `lmer` as follows.

```
> lmm2 <- lmer(log(loss) ~ (1+yearcentr|riskclass)+yearcentr
+offset(log(payroll)),
  data=wclossFit)
> print(lmm2)
Linear mixed model fit by REML
Formula: log(loss) ~ (1 + yearcentr | riskclass) + yearcentr
+ offset(log(payroll))
Data: wclossFit
AIC  BIC logLik deviance REMLdev
1451 1478 -719.4     1429     1439
Random effects:
Groups      Name        Variance Std.Dev. Corr
riskclass (Intercept) 0.885937 0.941242
            yearcentr   0.003171 0.056312 -0.195
Residual                 0.290719 0.539184
Number of obs: 669, groups: riskclass, 118

Fixed effects:
            Estimate Std. Error t value
(Intercept) -4.32030   0.08929 -48.38
yearcentr    0.03715   0.01340    2.77

Correlation of Fixed Effects:
          (Intr)
yearcentr -0.072
```

In this output $\hat{\theta}_0 = 0.89$, $\hat{\theta}_1 = 0.0032$, and $\hat{\theta}_{01} = -0.010$. We test whether the structure of random effects should be reduced; that is, $H_0 : \theta_1 = 0$ (with θ_1 the variance of random slopes), using an anova test comparing models (8.53) and (8.54).

```
> anova(lmm1,lmm2)
Data: wclossFit
Models:
lmm1: log(loss) ~ (1 | riskclass) + yearcentr + offset
      (log(payroll))
lmm2: log(loss) ~ (1 + yearcentr | riskclass) + yearcentr +
      offset(log(payroll))
      Df     AIC     BIC   logLik   Chisq Chi Df Pr(>Chisq)
lmm1    4 1438.5 1456.6 -715.27
lmm2    6 1440.9 1468.0 -714.46 1.6313        2       0.4423
```

When performing the corresponding LRT the software automatically refits lmm1 and lmm2 with ML (instead of REML), as required (see our discussion in Section 8.2.3.1). This explains why the AIC, BIC, and logLik values differ from those printed in the box. The observed Chisq test statistic and reported p -value indicate that $H_0 : \sigma_1^2 = 0$ cannot be rejected. The model with only random intercepts is our preferred specification.

Out-of-Sample Predictions. We compare out-of-sample predictions of Loss_{i7} , for given Payroll_{i7} , as obtained with models (8.50), (8.52), and (8.53). Figure 8.7 plots observed versus fitted losses (on log scale) for (from left to right) the complete pooling, the random intercepts, and the no pooling linear regression model.

8.3.2 Hachemeister Data

We present an analysis of the Hachemeister data using three simple linear mixed models. Chapter 9 presents an in-depth discussion of credibility models for this dataset (namely the Bühlmann, Bühlmann–Straub, and Hachemeister credibility models). By combining the R scripts prepared for our example with the scripts from Chapter 9, readers obtain relevant illustrations of credibility models in R and their analog interpretation as LMMs.

Random Intercepts, No Weights. The response variable is the average loss per claim (i.e., Ratio_{ij}), per state i ($i = 1, \dots, 5$) and quarter j ($j = 1, \dots, 12$). A basic

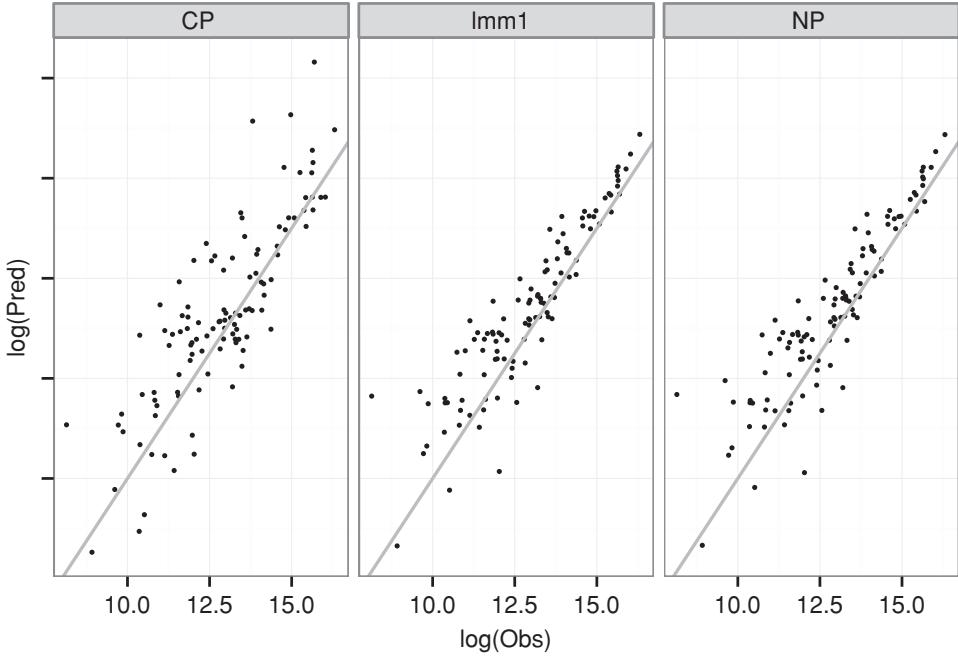


Fig. 8.7. Out-of-sample predictions for Loss_{i7} versus observed losses, as obtained with model (8.50) (CP, complete pooling), (8.53) (lmm1, random intercepts), and (8.52) (NP, no pooling): losses on workers' insurance compensation data.

random state intercept model for Ratio_{ij} is

$$\begin{aligned} \text{Ratio}_{ij} &= \beta_0 + u_{i,0} + \varepsilon_{ij} \\ u_{i,0} &\sim N(0, \sigma_u^2) \quad i.i.d. \\ \varepsilon_{ij} &\sim N(0, \sigma_\varepsilon^2) \quad i.i.d. \end{aligned} \quad (8.55)$$

Apart from the normality assumption, actuaries recognize this as the so-called Bühlmann credibility model, as Chapter 9 explains.

Random Intercepts, Including Weights. Our response variable is average loss per claim, constructed as total loss (per state and quarter) divided by the corresponding number of claims. This average loss is more precise when more claims have been observed. We therefore include the number of observed claims as weights (w_{ij}) in our LMM:

$$\begin{aligned} \text{Ratio}_{ij} &= \beta_0 + u_{i,0} + \varepsilon_{ij} \\ u_{i,0} &\sim N(0, \sigma_u^2) \quad i.i.d. \\ \varepsilon_{ij} &\sim N(0, \sigma_\varepsilon^2 / w_{ij}) \quad i.i.d. \end{aligned} \quad (8.56)$$

The model equation and variance assumptions (apart from normality) correspond with the Bühlmann–Straub credibility model. Including weights goes as follows in R lme4.

```
> lmmBS <- lmer(ratio ~ (1|state), weights=weight, data=hach)
> print(lmmBS)
Linear mixed model fit by REML
Formula: ratio ~ (1 | state)
Data: hach
AIC  BIC logLik deviance REMLdev
1301 1307 -647.5     1306    1295
Random effects:
Groups   Name        Variance Std.Dev.
state    (Intercept) 22.326   4.725
Residual           47928.954 218.927
Number of obs: 60, groups: state, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept) 1688.934    2.265    745.6
```

The risk (or: credibility) premium for state i is $\hat{\beta}_0 + \hat{u}_{i,0}$, and is available in R as follows.

```
## get fixed effects
fe <- fixef(lmmBS)
## get random intercepts
re <- ranef(lmmBS)
## calculate credibility premiums in this lmm
pred.lmmBS <- fe[1]+re$state
> t(pred.lmmBS)
          1         2         3         4         5
(Intercept) 2053.18 1528.509 1790.053 1468.113 1604.815
```

Chapter 9 illustrates how traditional actuarial credibility calculations are available in the `actuar` package in R. The credibility premiums obtained with Bühlmann–Straub are close to – but not exactly the same as – the premiums obtained with (8.56). Note that the actuarial credibility calculations use method of moments for parameter estimation, whereas our LMMs use (RE)ML.

```
> ## BS model (Buhlmann-Straub credibility model)
> ## use actuar package, and hachemeister data as available
  in this package
> fitBS <- cm(~state, hachemeister,ratios = ratio.1:ratio.12,
  weights = weight.1:weight.12)
> pred.BS <- predict(fitBS) # credibility premiums
> pred.BS
[1] 2055.165 1523.706 1793.444 1442.967 1603.285
```

Random Intercepts and Slopes, including Weights. We extend the random intercepts model to a random intercepts and slopes model, using the period of observation as regressor:

$$\begin{aligned} \text{Ratio}_{ij} &= \beta_0 + u_{i,0} + \beta_1 \text{period}_{ij} + u_{i,1} \text{period}_{ij} + \varepsilon_{ij} \\ \boldsymbol{u}_i &\sim N(\mathbf{0}, \mathbf{D}(\boldsymbol{\theta})) \quad i.i.d. \\ \varepsilon_{ij} &\sim N(0, \sigma_\varepsilon^2 / w_{ij}) \quad i.i.d. \end{aligned} \tag{8.57}$$

Our analysis uses period_{ij} as the quarter ($j = 1, \dots, 12$) of observation. The use of a centered version of period is discussed in Chapter 2. In R the `(1+period|state)` instruction specifies random intercepts and slopes per state.

```
> lmmHach <- lmer(ratio ~ period+(1+period|state),weights=weight,
  data=hach)
> lmmHach
Linear mixed model fit by REML
Formula: ratio ~ period + (1 + period | state)
Data: hach
AIC  BIC logLik deviance REMLdev
1242 1255 -615.1      1247     1230
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 state    (Intercept) 4.1153e+00  2.02863
           period       1.9092e-01  0.43695 1.000
 Residual            1.6401e+04 128.06735
Number of obs: 60, groups: state, 5

Fixed effects:
              Estimate Std. Error t value
(Intercept) 1501.5452     1.1265 1333.0
```

period	27.7333	0.2172	127.7
Correlation of Fixed Effects:			
(Intr)			
period	0.540		

Using LMM (8.57) the state specific risk premium for the next time period is

$$E[\widehat{\text{Ratio}_{i,13}}|\boldsymbol{u}_i] = \hat{\beta}_0 + \hat{u}_{i,0} + \hat{\beta}_1 \cdot 13 + \hat{u}_{i,1} \cdot 13. \quad (8.58)$$

```
> t(pred.lmmHach)
     [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 2464.032 1605.676 2067.279 1453.923 1719.48.
```

These premiums correspond with the results (obtained with SAS) reported in Frees et al. (1999, table 3, columns “Prediction and standard errors”). These authors also investigate linear mixed models as a user-friendly and computationally attractive alternative for actuarial credibility models. The traditional Hachemeister credibility premiums are available in R as follows (see also the “Base” results in table 3 from Frees et al. (1999)).

```
fitHach <- cm(~state, hachemeister, regformula = ~time, regdata =
  data.frame(time = 1:12), ratios = ratio.1:ratio.12,
  weights = weight.1:weight.12)
pred.Hach <- predict(fitHach, newdata = data.frame(time = 13))
# cred.premium
# > pred.Hach
# [1] 2436.752 1650.533 2073.296 1507.070 1759.403
```

Once again, with linear mixed models we obtain premiums that are close to, but do not replicate, the traditional actuarial credibility results. Differences in parameter estimation techniques explain why these results are not identical.

Using an LRT we verify whether model (8.57) should be reduced to the model with random intercepts only. The *p*-value indicates that this is not case.

```
lmmHach2 <- lmer(ratio ~ period+(1|state), weights=weight,data=hach)
anova(lmmHach,lmmHach2)
#Data: hach
#Models:
#lmmHach2: ratio ~ period + (1 | state)
#lmmHach: ratio ~ period + (1 + period | state)
```

```

#           Df      AIC      BIC  logLik   Chisq Chi Df Pr(>Chisq)
#lmmHach2  4 1272.2 1280.5 -632.08
#lmmHach    6 1258.7 1271.2 -623.32 17.521        2  0.0001568 ***
#---
#Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

```

Figure 8.8 illustrates the fit of a complete pooling (dark gray, dashed line), a no pooling (black, dashed line), and an LMM with random intercepts and slopes (black, solid line). The regression equations for the complete and no pooling model are

$$\text{Ratio}_{ij} = \beta_0 + \beta_1 \text{period}_{ij} + \varepsilon_{ij} \\ \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2 / w_{ij}), \quad (8.59)$$

and

$$\text{Ratio}_{ij} = \beta_{0,i} + \beta_{1,i} \text{period}_{ij} + \varepsilon_{ij} \\ \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2 / w_{ij}), \quad (8.60)$$

respectively.

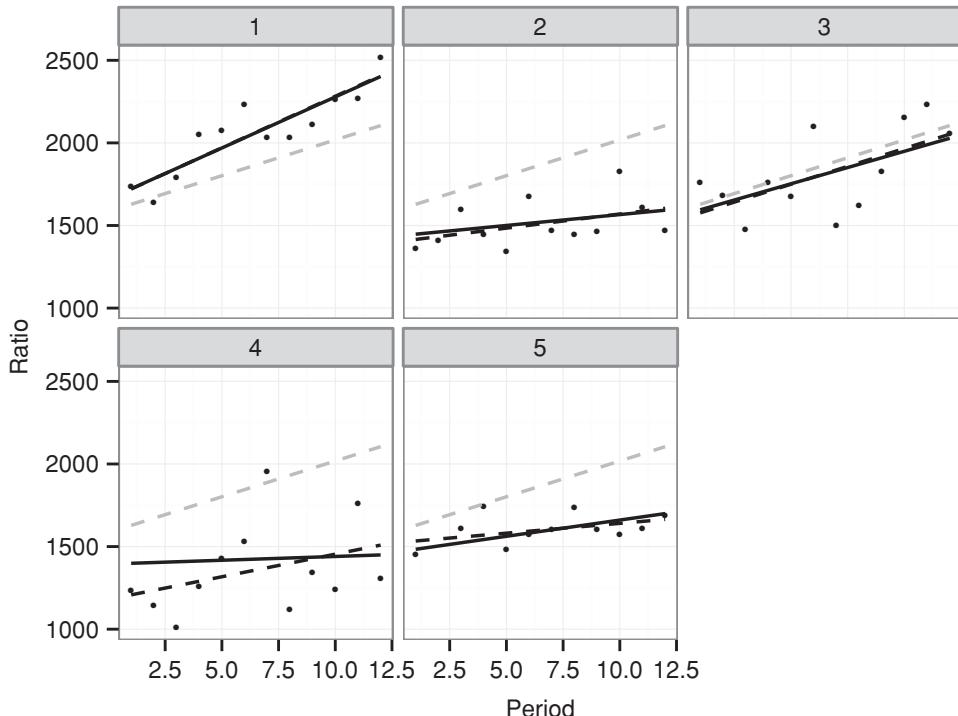


Fig. 8.8. Fit of a complete pooling (dark gray, dashed line), a no pooling (black, dashed line), and an LMM with random intercepts and slopes (black, solid line): Hachemeister data (no centering of period).

Table 8.2. Credibility Premiums Obtained with Crossed-Classification Credibility Model Per Combination of Status and Experience Risk Class: Credit Insurance Data

Status	Experience		
	1	2	3
1	181.05	238.18	277.77
2	172.11	229.16	268.8
3	225.29	282.24	323.68

8.3.3 Credit Insurance Data

We analyze the data from Example 8.5 and demonstrate the use of crossed random effects (see Section 8.2.2) with `lme4`. The response variable of interest is Payment_{ijt} , where $i = 1, 2, 3$ denotes status and $j = 1, 2, 3$ is for working experience of the insured; t is an index going over all observations in cell (i, j) . Dannenburg et al. (1996) use these data to demonstrate the principles of a so-called cross-classification credibility model, with the following model equation (in typical actuarial credibility notation):

$$\text{Payment}_{ijt} = m + \Xi_i^{(1)} + \Xi_j^{(2)} + \Xi_{ij}^{(12)} + \Xi_{ijt}^{(123)}. \quad (8.61)$$

Hereby, m is an overall intercept, $\Xi_i^{(1)}$ is a random effect for level i in factor (1) (i.e., status), $\Xi_j^{(2)}$ “a random intercept for level j in factor (2) (i.e., experience) and $\Xi_{ij}^{(12)}$ is a random effect for the interaction of level i and j . $\Xi_{ijt}^{(123)}$ is an error term for observation t from the combined level i and j . Dannenburg et al. (1996) obtain the credibility premiums in Table 8.2.

The analysis of this data by means of a linear mixed model with crossed random effects (i.e., `(1 | status:experience)`), is directly available in R.

$$\begin{aligned} \text{Payment}_{ijt} &= m + u_i^{(1)} + u_j^{(2)} + u_{ij}^{(12)} + \varepsilon_{ijt} \\ u_i^{(1)} &\sim N(0, \sigma_1^2) \\ u_j^{(2)} &\sim N(0, \sigma_2^2) \\ u_{ij}^{(12)} &\sim N(0, \sigma_{12}^2) \\ \varepsilon_{ijt} &\sim N(0, \sigma_\varepsilon^2), \end{aligned} \quad (8.62)$$

where i and j run over all levels in factors 1 (`status`) and 2 (`experience`) and we assume all random variables to be independent.

```

> lmm2 <- lmer(payment ~ 1+(1|experience)+(1|status)+(1|status:
   experience)
   ,data=credit)
> print(lmm2)
Linear mixed model fit by REML
Formula: payment ~ 1 + (1 | experience) + (1 | status) +
(1 | status:experience)
Data: credit
AIC  BIC logLik deviance REMLdev
5241 5261 -2616      5240      5231
Random effects:
Groups           Name        Variance Std.Dev.
status:experience (Intercept) 14.611   3.8224
status            (Intercept) 992.791  31.5086
experience        (Intercept) 2569.330 50.6886
Residual          26990.398 164.2875
Number of obs: 401, groups: status:experience, 9; status, 3;
               experience, 3

Fixed effects:
            Estimate Std. Error t value
(Intercept) 244.25     35.44   6.892

```

The resulting risk premiums as obtained with `lme4` are very close to the credibility premiums in Table 8.2.

	experience		
#status	1	2	3
1	181.0253	238.1813	277.7692
2	172.1086	229.1551	268.7954
3	225.2921	282.2424	323.6784

Our analysis directly uses `Payment` as response variable to facilitate the comparison between the credibility and linear mixed model calculations. However, the positivity and right skewness of `Payment` suggest the use of a lognormal or gamma distribution for this response.

8.4 Further Reading and Illustrations

We recommend Czado (2004), Gelman and Hill (2007), Frees (2004), McCulloch and Searle (2001), Ruppert et al. (2003), and Verbeke and Molenberghs (2000) as

further readings on linear mixed models. The use of LMMs for smoothing purposes is not discussed in this chapter, but interested readers can find in Example 8.9 a brief introduction and useful references.

Example 8.9 (Smoothing with Mixed Models). A semi-parametric regression model incorporates both parametric and nonparametric functional relationships between a response and a set of covariates. These models are particularly useful when a globally linear pattern is inappropriate or parametric nonlinear curves are difficult to determine. Such nonlinear effects frequently occur when time-related covariates are present, such as driver's age, development lag, or years in business of the insured company. For example, in an LM the effect of age of the insured on the number of claims reported is often expressed with a categorical Age covariate. The analyst splits Age into several categories and estimates a regression parameter for each one. In a nonparametric analysis we model the effect of Age on the response with an unknown, smooth function, in comparison with the piecewise constant assumption in linear models.

Penalized splines (also called P-splines) are popular nonparametric tools that specify the smoothing function as a linear combination of basis functions, in which some coefficients associated with the basis functions are constrained to avoid overfitting. That is, they are penalized, or shrunk toward zero, reducing the effective number of coefficients to be estimated. The widespread popularity of P-splines is largely because they can be written in the form of mixed models (Ruppert et al. 2003; Wood 2006) so that we can rely on software, diagnostic and inferential tools designed for mixed models directly in fitting P-splines, or can use a Bayesian implementation of the model to make inference of the full posterior distribution. Of course, hierarchical components can be included in addition to smoothing terms, thus often leading to models that are both intuitively appealing and structurally flexible when studying practical problems in predictive modeling.

For example, Figure 8.9 shows an application of the P-splines in estimating insurance loss reserves. In this example, the incremental paid insurance losses, represented by the dots in the plot, exhibit a nonlinear dependence on the report lag (the x -axis). Standard loss reserving methods specify a model with these lags as categorical covariates. In contrast, P-splines allow us to estimate a smooth functional relationship between paid losses and report lags. One advantage over the reserving model with dummy variables is the reduced number of model parameters, because generally a small number of knots can capture the observed pattern sufficiently well. The example shown here is based on a four-knot penalized spline, and Zhang and Dukic (2012) find that the resulting model has significantly better predictive performance than a dummy-variable-based reserving model. Another benefit is that estimates at any time point can be produced based on interpolation or extrapolation of the estimated

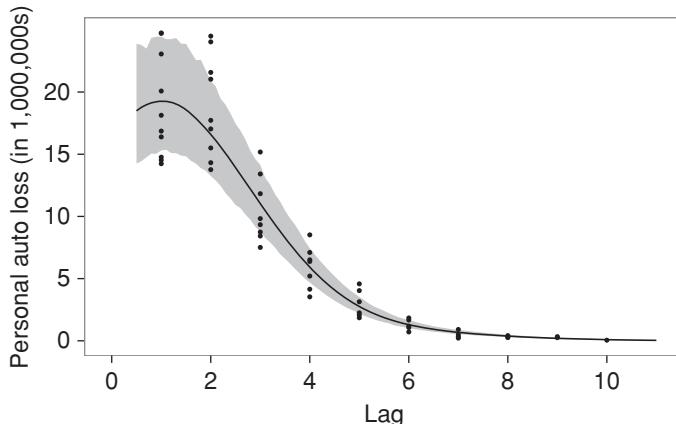


Fig. 8.9. The plot of the company-level smoother (incremental losses) along with the 50% prediction interval for a loss triangle.

functional form. This can be very helpful when the goal of a reserving study is to make forecasts for a short period ahead, say one month or a quarter.

More examples of semi-parametric models in insurance loss reserving can be found in Antonio and Beirlant (2008) and Zhang and Dukic (2012). Multivariate extensions of penalized splines are available for spatial regression (e.g., in postcode rating).

References

- Antonio, K. and J. Beirlant (2007). Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics* 40(1), 58–76.
- Antonio, K. and J. Beirlant (2008). Issues in claims reserving and credibility: A semiparametric approach with mixed models. *Journal of Risk and Insurance* 75(3), 643–676.
- Antonio, K., E. Frees, and E. Valdez (2010). A multilevel analysis of intercompany claim counts. *ASTIN Bulletin* 40(1), 151–177.
- Antonio, K. and E. Valdez (2012). Statistical aspects of *a priori* and *a posteriori* risk classification in insurance. *Advances in Statistical Analysis* 96(2), 187–224.
- Bühlmann, H. and A. Gisler (2005). *A Course in Credibility Theory and Its Applications*. Springer Verlag, Berlin.
- Czado, C. (2004). *Linear Mixed Models*. Lecture slides on GLM, TU Munchen.
- Dannenburg, D., R. Kaas, and M. Goovaerts (1996). *Practical Actuarial Credibility Models*. Institute of Actuarial Science and Econometrics, University of Amsterdam.
- Frees, E. (2004). *Longitudinal and Panel Data. Analysis and Applications in the Social Sciences*. Cambridge University Press, Cambridge.
- Frees, E., V. Young, and Y. Luo (1999). A longitudinal data analysis interpretation of credibility models. *Insurance: Mathematics and Economics* 24(3), 229–247.
- Frees, E., V. Young, and Y. Luo (2001). Case studies using panel data models. *North American Actuarial Journal* 5(4), 24–42.
- Gelman, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics* 48(3), 432–435.

- Gelman, A. and J. Hill (2007). *Applied Regression and Multilevel (Hierarchical) Models*. Cambridge University Press, Cambridge.
- Hachemeister, C. (1975). In Credibility: Theory and Applications, Credibility for regression models with application to trend, pp. 129–163. Academic Press, New York.
- Kacker, R. and D. Harville (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association* 79, 853–862.
- Klugman, S. (1992). *Bayesian Statistics in Actuarial Science with Emphasis on Credibility*. Kluwer, Boston.
- Laird, N. and J. Ware (1982). Random-effects models for longitudinal data. *Biometrics* 38(4), 963–974.
- Makov, U., A. Smith, and Y. Liu (1996). Bayesian methods in actuarial science. *The Statistician* 45(4), 503–515.
- McCulloch, C. and S. Searle (2001). *Generalized, Linear and Mixed Models*. Wiley Series in Probability and Statistics, Wiley, New York.
- Robinson, G. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science* 6, 15–51.
- Ruppert, D., M. Wand, and R. Carroll (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.
- Scollnik, D. (1996). An introduction to Markov Chain Monte Carlo methods and their actuarial applications. *Proceedings of the Casualty Actuarial Society Forum LXXXIII*, 114–165.
- Searle, S., G. Casella, and C. McCulloch (2008). *Variance Components*. Wiley, New York.
- Verbeke, G. and G. Molenberghs (2000). *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics, New York.
- Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall, CRC Texts in Statistical Science.
- Zhang, Y. and V. Dukic (2012). Predicting multivariate insurance loss payments under the Bayesian copula framework. *Journal of Risk and Insurance*.
- Zhang, Y., V. Dukic, and J. Guszczza (2012). A Bayesian nonlinear model for forecasting insurance loss payments. *Journal of the Royal Statistical Society, Series A* 175, 637–656.

9

Credibility and Regression Modeling

Vytautas Brazauskas, Harald Dornheim, and Ponmalar Ratnam

Chapter Preview. This chapter introduces the reader to credibility and related regression modeling. The first section provides a brief overview of credibility theory and regression-type credibility, and it discusses historical developments. The next section shows how some well-known credibility models can be embedded within the linear mixed model framework. Specific procedures on how such models can be used for prediction and standard ratemaking are given as well. Further, in Section 9.3, a step-by-step numerical example, based on the widely studied Hachemeister's data, is developed to illustrate the methodology. All computations are done using the statistical software package R. The fourth section identifies some practical issues with the standard methodology, in particular, its lack of robustness against various types of outliers. It also discusses possible solutions that have been proposed in the statistical and actuarial literatures. Performance of the most effective proposals is illustrated on the Hachemeister's dataset and compared to that of the standard methods. Suggestions for further reading are made in Section 9.5.

9.1 Introduction

9.1.1 Early Developments

Credibility theory is one of the oldest but still most common premium ratemaking techniques in insurance industry. The earliest works in credibility theory date back to the beginning of the 20th century, when Mowbray (1914) and Whitney (1918) laid the foundation for *limited fluctuation credibility theory*. It is a stability-oriented form of credibility, the main objective of which is to incorporate into the premium as much individual experience as possible while keeping the premium sufficiently stable. Despite numerous attempts, this approach never arrived at a unifying principle that covered all special cases and that opened new venues for generalization. Its range of applications is quite limited, and thus it never became a full-fledged theory.

Instead of focusing solely on the stability of the premium, the modern and more flexible approach to credibility theory concentrates on finding the most accurate estimate of an insured's pure risk premium. This is accomplished by striking a balance between the individual's risk experience and the average claim over all risk classes. Although initial contributions to this area can be traced back to the 1920s (see Keffer 1929), it is generally agreed that the systematic development of the field of *greatest accuracy credibility* started in the late 1960s with the seminal paper of Bühlmann (1967). A few years later, Bühlmann and Straub (1970) introduced a credibility model as a means to rate reinsurance treaties, which generalized previous results and became the cornerstone of greatest accuracy credibility theory. The model is one of the most frequently applied credibility models in insurance practice, and it enjoys some desirable optimality properties. For more historical facts and further discussion about credibility, see the classic textbook by Klugman, Panjer, and Willmot (2012, chapters 17–18).

9.1.2 Regression-Type Credibility

The first credibility model linked to regression was introduced by Hachemeister (1975) who employed it to model U.S. automobile bodily injury claims classified by state and with different inflation trends. Specifically, Hachemeister considered 12 periods, from the third quarter of 1970 to the second quarter of 1973, of claim data for bodily injury that was covered by a private passenger auto insurance. The response variable of interest to the actuary is the severity *average loss per claim*, denoted by y_{it} . It is followed over the periods $t = 1, \dots, n_i$ for each state $i = 1, \dots, m$. Average losses were reported for $n_1 = \dots = n_m = 12$ periods and from $m = 5$ different states (see Appendix, Table 9.3).

A multiple time series plot of the observed variable average loss per claim, y_{it} , and of the average loss per claim in logarithmic units, $\ln(y_{it})$, is provided in Figure 9.1 (log-claim modeling is considered in Section 9.4). The plots indicate that states differ with respect to variability and severity. For instance, State 1 reports the highest average losses per claim, whereas State 4 seems to have larger variability compared to other states. For all five states we observe a small increase of severity over time. Since the response variable y_{it} grows over time t and varies from one state to another, this provides a hint about the possible structure of explanatory variables. Therefore, Hachemeister originally suggested the use of the linear trend model – a regression model – which can be viewed as a special case of the linear mixed models of Chapter 8:

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{u}_i + \varepsilon_{it},$$

where $\mathbf{x}'_{it} = \mathbf{z}'_{it} = (1, t)'$ are known designs for the fixed effects (parameter $\boldsymbol{\beta}$) and the subject-specific random effects (parameter \mathbf{u}_i), respectively, and ε_{it} denotes within-subject residuals.

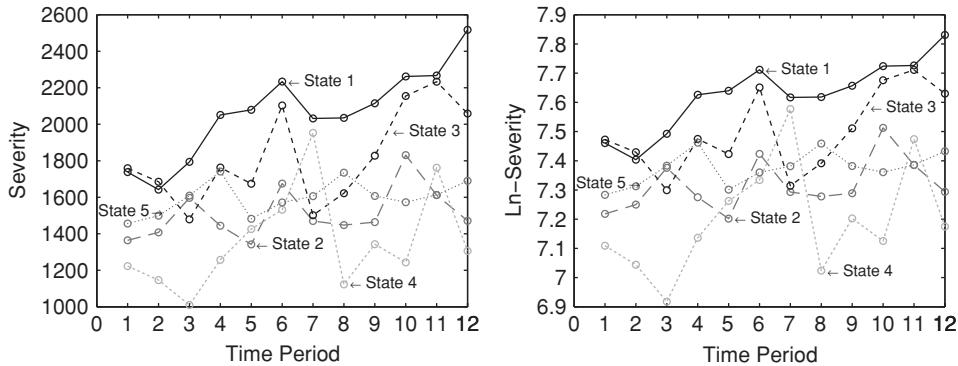


Fig. 9.1. Multiple time series plot of the variable average loss per claim, y_{it} , and the logarithmic average loss per claim, $\ln(y_{it})$.

Later in this chapter, more details and examples are provided about the link between the linear mixed models and the most popular credibility models (see Section 9.2.1). Also, as is shown in Section 9.2.2, the general linear prediction problem for linear mixed models is closely related to credibility ratemaking. It turns out that generalized least squares and best linear unbiased predictors correspond to the well-known pricing formulas of credibility theory.

9.1.3 Recent Developments

Frees, Young, and Luo (1999, ch. 7) provided a longitudinal data analysis interpretation for the aforementioned and other additive credibility ratemaking procedures, which also remains valid in the framework of linear mixed models. The flexibility of linear mixed models for handling, simultaneously, within-risk variation and heterogeneity among risks makes them a powerful tool for credibility (see Chapter 8 for details and generalizations of linear mixed models).

As is the case with many mathematical models, credibility models contain unknown structural parameters (or, in the language of linear mixed models, fixed effects and variance components) that have to be estimated from the data. For statistical inference about fixed effects and variance components, likelihood-based methods such as (restricted) maximum likelihood estimators, (RE)ML, are commonly pursued. However, it is also known that, although these methods offer most flexibility and full efficiency at the assumed model, they are extremely sensitive to small deviations from hypothesized normality of random components, as well as to the occurrence of outliers. To obtain more reliable estimators for premium calculation and prediction of future claims, various robust methods have been successfully adapted to credibility theory in the actuarial literature; see, for example, Pitselis (2008; 2012) and Dornheim and Brazauskas (2007; 2011a).

In the remainder of the chapter, we first present the standard likelihood-based procedures for ratemaking, then provide a step-by-step numerical example, and conclude with a brief review of robust techniques and a comparison of their performance to that of the standard methods. All computations are done using the statistical software package R and are based on Hachemeister's data.

9.2 Credibility and the LMM Framework

In this section, we start by briefly describing how some popular (linear) credibility models are expressed as linear mixed models, or LMM for short. The problem of prediction in linear mixed models and its application to standard credibility ratemaking are discussed in Section 9.2.2.

9.2.1 Credibility Models

Here we demonstrate that some well-known additive credibility models can be interpreted as linear mixed models that enjoy many desirable features. For instance, LMMs allow the modeling of claims across risk classes and time, as well as the incorporation of categorical and continuous explanatory characteristics for prediction of claims. The following descriptions are taken, with some modifications, from Bühlmann and Gisler (2005), Frees (2004), and Frees et al. (1999). The basic credibility models such as Bühlmann and Bühlmann-Straub can also be found in chapter 18 of Klugman et al. (2012). The notation we use is similar to that of other chapters in this book (especially, Chapter 8), but may differ from the notation used elsewhere in the literature.

9.2.1.1 The Bühlmann Model

Let us consider a portfolio of different insureds or risks i , $i = 1, \dots, m$. For each risk i we have a vector of observations $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})'$, where y_{it} represents the *observed* claim amount (or loss ratio) of risk i at time t , $t = 1, \dots, n_i$, where n_i 's are allowed to be unequal. Then, by choosing $p = q = 1$ and $\mathbf{X}_i = \mathbf{Z}_i = \mathbf{1}_{n_i}$ in equation (8.18) of Chapter 8, we arrive at

$$\mathbf{y}_i = \mathbf{1}_{n_i} \beta + \mathbf{1}_{n_i} u_i + \boldsymbol{\varepsilon}_i,$$

where $\beta = E(y_{it}) = E(E(y_{it}|u_i))$ is the overall mean or *collective premium* charged for the whole portfolio, u_i denotes the *unobservable* risk parameter characterizing the subject-specific deviation from the collective premium β , and $\mathbf{1}_{n_i}$ represents the n_i -variate vector of ones. From the hierarchical formulation of linear mixed models (see Section 8.2.1), the risk premium $\mu_i = E(y_{it}|u_i) = \beta + u_i$ is the *true premium* for an insured i if its risk parameter u_i were known. In addition, we obtain $G = \text{Var}(u_i) = \sigma_u^2$

and the variance-covariance matrices

$$\Sigma_i = \mathbf{Var}(\mathbf{y}_i | u_i) = \mathbf{Var}(\boldsymbol{\varepsilon}_i) = \sigma_\varepsilon^2 \mathbf{I}_{n_i \times n_i}.$$

Note that, in general, the structural parameters β , σ_u^2 , and σ_ε^2 are unknown and must be estimated from the data. Also, viewing the Bühlmann model from this broader perspective provides insight about the explanatory variables for claims (or loss ratios) and possible generalizations.

Note 1 (The Balanced Bühlmann Model): When the number of observation periods is the same for all risks, i.e., $n_1 = \dots = n_m$, the basic credibility model becomes the balanced Bühlmann model. \square

9.2.1.2 The Bühlmann-Straub Model

The credibility model of Section 9.2.1.1 can be easily extended to the heteroskedastic model of Bühlmann and Straub (1970) by choosing the variance-covariance matrix as follows:

$$\Sigma_i = \mathbf{Var}(\mathbf{y}_i | u_i) = \mathbf{Var}(\boldsymbol{\varepsilon}_i) = \sigma_\varepsilon^2 \operatorname{diag}(v_{i1}^{-1}, \dots, v_{in_i}^{-1}),$$

where $v_{it} > 0$ are known volume measures. These weights represent varying exposures toward risk for insured i over the period n_i . Practical examples of exposure weights include the number of years at risk in motor insurance, sum insured in fire insurance, and annual turnover in commercial liability insurance (see Bühlmann and Gisler 2005).

9.2.1.3 The Hachemeister Regression Model

Hachemeister's simple linear regression model is a generalization of the Bühlmann-Straub model, which includes the time (as linear trend) in the covariates. To obtain the linear trend model, in equation (8.18), we choose $p = q = 2$ and set $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i})'$ and $\mathbf{Z}_i = (\mathbf{z}_{i1}, \dots, \mathbf{z}_{in_i})'$, where $\mathbf{x}_{it} = \mathbf{z}_{it} = (1, t)'$. This results in the random coefficients model of the form

$$\mathbf{y}_i = \mathbf{X}_i (\boldsymbol{\beta} + \mathbf{u}_i) + \boldsymbol{\varepsilon}_i,$$

with the diagonal matrix Σ_i defined as in Section 9.2.1.2. It is common to assume that (unobservable) risk factors u_1 and u_2 are independent with the variance-covariance matrix $\mathbf{G} = \operatorname{diag}(\sigma_{u_1}^2, \sigma_{u_2}^2)$.

9.2.1.4 The Revised Hachemeister Regression Model

Application of the Hachemeister's model to bodily injury data (see Section 9.1.2) results in unsatisfying model fits that are due to systematic underestimation of the

credibility regression line. To overcome this drawback, Bühlmann and Gisler (1997) suggested taking the intercept of the regression line at the “center of gravity” of the time variable, instead of at the origin of the time axis. That is, choose design matrices $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i})'$ and $\mathbf{Z}_i = (\mathbf{z}_{i1}, \dots, \mathbf{z}_{in_i})'$ with $\mathbf{x}_{it} = \mathbf{z}_{it} = (1, t - C_{i\bullet})'$, where

$$C_{i\bullet} = v_{i\bullet}^{-1} \sum_{t=1}^{n_i} t v_{it}$$

is the center of gravity of the time range in risk i , and $v_{i\bullet} = \sum_{t=1}^{n_i} v_{it}$. This modification ensures that the regression line stays between the individual and collective regression lines; the model is called the revised Hachemeister regression model.

From a practical point of view, volumes are often equal enough across periods for a single risk to be considered constant in time, which yields similar centers of gravity between risks. Then, it is reasonable to use the center of gravity of the collective, which is defined by $C_{\bullet\bullet} = v_{\bullet\bullet}^{-1} \sum_{i=1}^m \sum_{t=1}^{n_i} t v_{it}$, where $v_{\bullet\bullet} = \sum_{i=1}^m \sum_{t=1}^{n_i} v_{it}$ (see Bühlmann and Gisler 2005, Section 8.3).

9.2.2 Prediction and Ratemaking

In the linear mixed model defined by equation (8.18), let $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ and $\widehat{\boldsymbol{\theta}}$ be the likelihood-based estimates of the grand mean $\boldsymbol{\beta}$ and the variance component vector $\boldsymbol{\theta} = (\sigma_{u_1}^2, \dots, \sigma_{u_q}^2, \sigma_\varepsilon^2)$, respectively. Then, the minimum mean square error predictor of the random variable

$$W_i = E(y_{i,n_i+1} | \mathbf{u}_i) = \mathbf{x}'_{i,n_i+1} \boldsymbol{\beta} + \mathbf{z}'_{i,n_i+1} \mathbf{u}_i$$

is given by the best linear unbiased predictor

$$\widehat{W}_{\text{BLUP},i} = \mathbf{x}'_{i,n_i+1} \widehat{\boldsymbol{\beta}}_{\text{GLS}} + \mathbf{z}'_{i,n_i+1} \widehat{\mathbf{u}}_{\text{BLUP},i}, \quad i = 1, \dots, m, \quad (9.1)$$

where \mathbf{x}'_{i,n_i+1} and \mathbf{z}'_{i,n_i+1} are known covariates of risk i in time period $n_i + 1$, and $\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ and $\widehat{\mathbf{u}}_{\text{BLUP},i}$ are computed using equations (8.29) and (8.30), respectively (see also the discussion following equation (8.30)).

In the actuarial literature, $\widehat{W}_{\text{BLUP},i}$ is called a *homogeneous* estimator of W_i (Dannenburg, Kaas, and Goovaerts 1996), and it is used to predict the expected claim size $\mu_{i,n_i+1} = E(y_{i,n_i+1} | \mathbf{u}_i)$ of risk i for time $n_i + 1$. This estimator is even optimal for non-normally distributed claims (Norberg 1980).

Recall that the central objective of credibility is to price fairly heterogeneous risks based on the overall portfolio mean, M , and the risk's individual experience, M_i . This relation can be expressed by the general credibility pricing formula:

$$P_i = \zeta_i M_i + (1 - \zeta_i) M = M + \zeta_i (M_i - M), \quad i = 1, \dots, m, \quad (9.2)$$

where P_i is the credibility premium of risk i , and $0 \leq \zeta_i \leq 1$ is known as the credibility factor. Note, a comparison of equation (9.1) with (9.2) implies that $\mathbf{x}'_{i,n_i+1} \hat{\boldsymbol{\beta}}_{\text{GLS}}$ can be interpreted as an estimate of M , and $\mathbf{z}'_{i,n_i+1} \hat{\mathbf{u}}_{\text{BLUP},i}$ as a predictor of the weighted, risk-specific deviation $\zeta_i (M_i - M)$. This relationship is illustrated next for the Bühlmann-Straub model and the revised Hachemeister regression model.

9.2.2.1 Example 1: The Bühlmann-Straub Model

In Section 9.2.1.2, we saw that the Bühlmann-Straub model can be formulated as a random coefficients model of the form $\mathbf{E}(\mathbf{y}_i | u_i) = \mathbf{1}_{n_i} \boldsymbol{\beta} + \mathbf{1}_{n_i} u_i$. Then, for future expected claims $\mu_i = \mathbf{E}(y_{i,n_i+1} | u_i)$ of risk i , Frees (2004) finds the best linear unbiased predictor $\hat{\mu}_i = \hat{\boldsymbol{\beta}}_{\text{GLS}} + \hat{\mathbf{u}}_{\text{BLUP},i}$ with

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \bar{y}_\zeta \quad \text{and} \quad \hat{\mathbf{u}}_{\text{BLUP},i} = \zeta_i (\bar{y}_i - \hat{\boldsymbol{\beta}}_{\text{GLS}}), \quad (9.3)$$

where $\bar{y}_\zeta = (\sum_{i=1}^m \zeta_i)^{-1} \sum_{i=1}^m \zeta_i \bar{y}_i$, $\bar{y}_i = v_{i\bullet}^{-1} \sum_{t=1}^{n_i} v_{it} y_{it}$, and $\zeta_i = (1 + \sigma_\varepsilon^2 / (v_{i\bullet} \sigma_u^2))^{-1}$. To compute formulas (9.3), one needs to estimate the structural parameters σ_u^2 and σ_ε^2 . The estimators $\hat{\sigma}_u^2$ and $\hat{\sigma}_\varepsilon^2$ are obtained from (RE)ML (i.e., as a byproduct from Henderson's mixed model equations) and coincide, when assuming normality, with the following nonparametric estimators:

$$\begin{aligned} \hat{\sigma}_\varepsilon^2 &= \frac{\sum_{i=1}^m \sum_{t=1}^{n_i} v_{it} (y_{it} - \bar{y}_i)^2}{\sum_{i=1}^m (n_i - 1)} \quad \text{and} \\ \hat{\sigma}_u^2 &= \frac{v_{\bullet\bullet}}{v_{\bullet\bullet}^2 - \sum_{i=1}^m v_{i\bullet}^2} \left(\sum_{i=1}^m v_{i\bullet} (\bar{y}_i - \bar{y})^2 - \hat{\sigma}_\varepsilon^2 (m - 1) \right), \end{aligned}$$

where $\bar{y} = v_{\bullet\bullet}^{-1} \sum_{t=1}^m v_{i\bullet} \bar{y}_i$ (see also Klugman et al. 2012, section 19.2).

9.2.2.2 Example 2: The Revised Hachemeister Regression Model

Here, we provide the necessary details for estimators in the revised Hachemeister regression model. For risk i one can estimate the expected claim amount $\mu_{i,n_i+1} = \mathbf{E}(y_{i,n_i+1} | \mathbf{u}_i)$ by the credibility estimator $\hat{\mu}_{i,n_i+1} = (1, n_i + 1) (\hat{\boldsymbol{\beta}}_{\text{GLS}} + \hat{\mathbf{u}}_{\text{BLUP},i}) = (1, n_i + 1) ((\mathbf{I}_{2 \times 2} - \boldsymbol{\xi}_i) \hat{\boldsymbol{\beta}}_{\text{GLS}} + \boldsymbol{\xi}_i \mathbf{b}_i)$, with

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \left(\sum_{i=1}^m \boldsymbol{\xi}_i \right)^{-1} \sum_{i=1}^m \boldsymbol{\xi}_i \mathbf{b}_i \quad \text{and} \quad \hat{\mathbf{u}}_{\text{BLUP},i} = \boldsymbol{\xi}_i (\mathbf{b}_i - \hat{\boldsymbol{\beta}}_{\text{GLS}}),$$

where

$$\mathbf{b}_i = \mathbf{A}_i^{-1} \begin{bmatrix} \sum_{t=1}^{n_i} v_{it} y_{it} \\ \sum_{t=1}^{n_i} v_{it} y_{it} (t - C_{i\bullet}) \end{bmatrix}$$

is the estimated individual claim experience of risk i ,

$$\boldsymbol{\xi}_i = \text{diag} \begin{bmatrix} \left(1 + \sigma_{\varepsilon}^2 / (\sigma_{u_1}^2 a_{i1})\right)^{-1} \\ \left(1 + \sigma_{\varepsilon}^2 / (\sigma_{u_2}^2 a_{i2})\right)^{-1} \end{bmatrix}$$

is the credibility factor for risk i , $\mathbf{A}_i = \text{diag}(a_{i1}, a_{i2})$ with $a_{i1} = v_{i\bullet}$, $a_{i2} = \tilde{v}_{i\bullet} = \sum_{t=1}^{n_i} v_{it}(t - C_{i\bullet})^2$, and $C_{i\bullet} = v_{i\bullet}^{-1} \sum_{t=1}^{n_i} t v_{it}$ is the center of gravity. We still have to estimate the process variance σ_{ε}^2 and variances of hypothetical means $\sigma_{u_1}^2$ and $\sigma_{u_2}^2$. It is reasonable to estimate σ_{ε}^2 by the natural variance estimator $\hat{\sigma}_{\varepsilon}^2 = m^{-1} \sum_{i=1}^m \hat{\sigma}_{\varepsilon,i}^2$, where $\hat{\sigma}_{\varepsilon,i}^2 = (n_i - 2)^{-1} \sum_{t=1}^{n_i} v_{it}(y_{it} - \hat{\mu}_{it})^2$ is a (conditionally) unbiased estimator of the within-risk variance $\sigma_{\varepsilon,i}^2$, and $\hat{\mu}_{it}$ is the fitted value of the i th regression line in time t . The structural parameters $\sigma_{u_1}^2$ and $\sigma_{u_2}^2$ are estimated by

$$\begin{aligned} \hat{\sigma}_{u_1}^2 &= c_1 \left[\frac{m}{m-1} \sum_{i=1}^m \frac{v_{i\bullet}}{v_{\bullet\bullet}} (b_{i,1} - \bar{b}_1)^2 - \frac{m \hat{\sigma}_{\varepsilon}^2}{v_{\bullet\bullet}} \right] \quad \text{and} \\ \hat{\sigma}_{u_2}^2 &= c_2 \left[\frac{m}{m-1} \sum_{i=1}^m \frac{\tilde{v}_{i\bullet}}{\tilde{v}_{\bullet\bullet}} (b_{i,2} - \bar{b}_2)^2 - \frac{m \hat{\sigma}_{\varepsilon}^2}{\tilde{v}_{\bullet\bullet}} \right], \end{aligned}$$

where

$$\begin{aligned} c_1 &= \frac{m-1}{m} \left\{ \sum_{i=1}^m \frac{v_{i\bullet}}{v_{\bullet\bullet}} \left(1 - \frac{v_{i\bullet}}{v_{\bullet\bullet}}\right) \right\}^{-1}, \quad \bar{b}_1 = v_{\bullet\bullet}^{-1} \sum_{i=1}^m v_{i\bullet} b_{i,1}, \\ c_2 &= \frac{m-1}{m} \left\{ \sum_{i=1}^m \frac{\tilde{v}_{i\bullet}}{\tilde{v}_{\bullet\bullet}} \left(1 - \frac{\tilde{v}_{i\bullet}}{\tilde{v}_{\bullet\bullet}}\right) \right\}^{-1}, \quad \text{and} \quad \bar{b}_2 = \tilde{v}_{\bullet\bullet}^{-1} \sum_{i=1}^m \tilde{v}_{i\bullet} b_{i,2}. \end{aligned}$$

9.3 Numerical Examples

In this section, we revisit Section 9.1.2 and model the Hachemeister's dataset, which, over the years, has been extensively analyzed by a number of authors in the actuarial literature. For example, Dannenburg et al. (1996), Bühlmann and Gisler (1997), Frees et al. (1999), Pitselis (2008), and Dornheim and Brazauskas (2011a) used this dataset to illustrate the effectiveness of various regression-type credibility ratemaking techniques.

To get a feel for how things work, let us study a detailed example that shows how to fit and make predictions based on the Hachemeister model and the revised Hachemeister model. All computations are done using the statistical software package R. Parts of the computer code are available in `actuar`, an R package for actuarial science that is described in Dutang et al. (2008).

To fit the linear trend regression model of Section 9.2.1.3 to the Hachemeister's dataset, we employ the following R-code:

```
> fit <- cm(~state, hachemeister, regformula = ~time,
+ regdata = data.frame(time = 1:12), ratios = ratio.1:ratio.12,
+ weights = weight.1:weight.12)
> fit
> summary(fit, newdata = data.frame(time = 13))
```

The label `hachemeister` in the first line of the code reads the dataset that is available in the `actuar` package. The last line produces predictions that are based on formula (9.1). The R-code yields the following credibility-adjusted parameter estimates and predictions for the five states:

State <i>i</i>	Parameter Estimates		Prediction $\hat{\mu}_{i,12+1}$
	$\hat{\beta}_0 + \hat{u}_{i,0}$	$\hat{\beta}_1 + \hat{u}_{i,1}$	
1	1693.52	57.17	2436.75
2	1373.03	21.35	1650.53
3	1545.36	40.61	2073.30
4	1314.55	14.81	1507.07
5	1417.41	26.31	1759.40

In addition, the grand parameters are found by taking the average across all states; they are $\hat{\beta}_0 = 1468.77$ and $\hat{\beta}_1 = 32.05$. Also, within-state variance is $\hat{\sigma}_e^2 = 49,870,187$, and the other estimates of variance components are $\hat{\sigma}_{u_1}^2 = 24,154.18$ and $\hat{\sigma}_{u_2}^2 = 301.81$. Note that the numerical values in the display table differ from the ones reported by Dutang et al. (2012), who used the same R-code but applied the reversed time variable; that is, `time = 12:1` instead of `time = 1:12`.

Fitting and predictions based on the revised Hachemeister model (see Sections 9.2.1.4 and 9.2.2.2) are much more involved. The complete program for running these tasks is presented in the code on the book website. Comments explaining the formulas or a block of the program are listed between the signs `#` and `#`. To run the program, download the R-code to your computer and use the following command lines:

```
> source("HachemRevised.R")
> HachemRevised(5,12, hachemeister[,2:13], hachemeister[,14:25],
  1:12,13)
```

Let us now go through the main blocks of the program and review the numerical output of each block. The first group of commands (which is labeled “Program Initialization”) defines the variables used in computations and sets their initial values, if necessary. The second group of commands, labeled “Center of Gravity & Recentered X ,” computes the collective center of gravity (the outcome is $C_{\bullet\bullet} = 6.4749$) and recenters the design matrix X , which results in

$$\begin{bmatrix} 1 & -5.4749 \\ 1 & -4.4749 \\ 1 & -3.4749 \\ 1 & -2.4749 \\ 1 & -1.4749 \\ 1 & -0.4749 \\ 1 & 0.5251 \\ 1 & 1.5251 \\ 1 & 2.5251 \\ 1 & 3.5251 \\ 1 & 4.5251 \\ 1 & 5.5251 \end{bmatrix}$$

The third group of commands, labeled “Volume Measures & Other Constants,” computes the constants required for further calculations. The results of this step are $\tilde{v}_{\bullet\bullet} = 2,104,688$; $c_1 = 1.3202$; $c_2 = 1.3120$; $\bar{b}_1 = 1865.4040$; and $\bar{b}_2 = 44.1535$. The fourth and fifth groups of commands, labeled “Variance Components” and “Parameters & Prediction,” respectively, yield estimates of the structural parameters and deliver next-period predictions. The summarized results of these two program blocks are

State i	Estimates of Structural Parameters					Prediction $\hat{\mu}_{i,12+1}$	Std. Error $\hat{\sigma}_{\hat{\mu}_{i,12+1}}$
	$\hat{u}_{i,0}$	$\hat{u}_{i,1}$	$\hat{\beta}_0 + \hat{u}_{i,0}$	$\hat{\beta}_1 + \hat{u}_{i,1}$	$\hat{\sigma}_{\varepsilon,i}^2$		
1	385.71	27.06	2058.85	60.70	121,484,314	2847.94	110.13
2	-157.67	-12.60	1515.48	21.04	30,175.637	1789.01	123.15
3	127.71	6.58	1800.85	40.22	52,560,076	2323.70	195.49
4	-283.51	-2.40	1389.63	31.25	24,362,730	1795.83	242.22
5	-72.24	-18.63	1600.90	15.01	21,075,078	1796.00	76.39

Note that the grand parameters are found by taking the average across all states for $\hat{\beta}_0 + \hat{u}_{i,0}$, $\hat{\beta}_1 + \hat{u}_{i,1}$, and $\hat{\sigma}_{\varepsilon,i}^2$; they are $\hat{\beta}_0 = 1673.14$, $\hat{\beta}_1 = 33.64$, and $\hat{\sigma}_{\varepsilon}^2 = 49,931,567$. In addition, the estimates of variance components are $\hat{\sigma}_{u_1}^2 = 93,021.43$ and $\hat{\sigma}_{u_2}^2 = 665.48$. The last line of the code screen-prints the results of the program.

Note 2 (Potential Outliers): A careful examination of Figure 9.1 reveals that the quarterly observations #6 in State 1, #10 in State 2, #7 in State 4, and perhaps #6 in State 5 differ somewhat from their state-specific inflation trends. This suggests the topic for the next section. That is, we would like to know how to identify outliers and, if they are present, what to do about them. \square

9.4 Theory versus Practice

The modeling approach of Section 9.3 is well understood and widely used, but it is not very realistic in practice. In particular, insurance loss data are often highly skewed and, heavy-tailed, and they contain outliers. Although complete treatment of these issues is beyond the scope of this chapter, in the following we present some insights on how to modify and improve the standard methodology. First, in Section 9.4.1 we formulate heavy-tailed linear mixed models. Note that “heavy-tailed” models are also known as “long-tailed” or “fat-tailed” (see Chapter 10). Then, Section 9.4.2 introduces a three-step procedure for robust-efficient fitting of such models. Robust credibility ratemaking based on heavy-tailed linear mixed models (which are calibrated using the robust-efficient fitting procedure of Section 9.4.2) is described in Section 9.4.3. Finally, in Section 9.4.4, we revisit the earlier examples and illustrate performance of the robust methods using Hachemeister’s data.

9.4.1 Heavy-Tailed Linear Mixed Models

Suppose we are given a random sample $(\mathbf{x}_{i1}, \mathbf{z}_{i1}, y_{i1}, v_{i1}), \dots, (\mathbf{x}_{in_i}, \mathbf{z}_{in_i}, y_{in_i}, v_{in_i})$, where \mathbf{x}_{it} and \mathbf{z}_{it} are known p - and q -dimensional row-vectors of explanatory variables and $v_{it} > 0$ is some known volume measure. Assume the claims y_{it} follow a log-location-scale distribution with cdf of the form:

$$G(y_{it}) = F_0\left(\frac{\log(y_{it}) - \lambda_{it}}{\sigma_\varepsilon v_{it}^{-1/2}}\right), \quad y_{it} > 0, \quad i = 1, \dots, m, \quad t = 1, \dots, n_i,$$

defined for $-\infty < \lambda_{it} < \infty$, $\sigma_\varepsilon > 0$, and where F_0 is the standard (i.e., $\lambda_{it} = 0$, $\sigma_\varepsilon = 1$, $v_{it} = 1$) cdf of the underlying location-scale family $F(\lambda_{it}, \sigma_\varepsilon^2/v_{it})$. Following regression analysis with location-scale models, we include the covariates \mathbf{x}_{it} and \mathbf{z}_{it} only through the location parameter λ_{it} . Then, the following linear mixed model may be formulated:

$$\log(\mathbf{y}_i) = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i + \boldsymbol{\varepsilon}_i = \boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, m,$$

where $\log(\mathbf{y}_i) = (\log(y_{i1}), \dots, \log(y_{in_i}))'$ and $\boldsymbol{\lambda}_i$ is the n_i -dimensional vector of the within-subject locations λ_{it} that consist of the *population location* $\boldsymbol{\beta}$ and the subject-specific *location deviation* \mathbf{u}_i . While normality of \mathbf{u}_i is still assumed, the error term

$\boldsymbol{\varepsilon}_i$ now follows the n_i -dimensional multivariate cdf with location-scale distributions $F(0, \sigma_\varepsilon^2/v_{it})$ as margins. Examples of such marginal log-location-scale families F include lognormal, log-logistic, log- t , log-Cauchy, and Weibull, which after the logarithmic transformation become normal, logistic, t , Cauchy, and Gumbel (extreme-value), respectively. Special cases of the n_i -dimensional distributions $F_{n_i}(\lambda_i, \Sigma_i)$ are the well-known elliptical distributions such as multivariate normal and the heavy-tailed multivariate t .

9.4.2 Robust-Efficient Fitting

For robust-efficient fitting of the linear mixed model with normal random components, Dornheim (2009) and Dornheim and Brazauskas (2011b) developed adaptively truncated likelihood methods. Those methods were further generalized to log-location-scale models with symmetric or asymmetric errors and labeled *corrected adaptively truncated likelihood* methods, CATL (see Dornheim 2009; Dornheim and Brazauskas 2011a). More specifically, the CATL estimators for location λ_i and variance components $\sigma_{u_1}^2, \dots, \sigma_{u_q}^2, \sigma_\varepsilon^2$ can be found by the following three-step procedure:

(1) *Detection of Within-Risk Outliers*

Consider the random sample

$$\left(\mathbf{x}_{i1}, \mathbf{z}_{i1}, \log(y_{i1}), v_{i1} \right), \dots, \left(\mathbf{x}_{in_i}, \mathbf{z}_{in_i}, \log(y_{in_i}), v_{in_i} \right), \quad i = 1, \dots, m.$$

In the first step, the corrected reweighting mechanism automatically detects and removes outlying events *within* risks whose standardized residuals computed from initial high breakdown-point estimators exceed some adaptive cut-off value. This threshold value is obtained by comparison of an empirical distribution with a theoretical one. Let us denote the resulting “pre-cleaned” random sample as

$$\left(\mathbf{x}_{i1}^*, \mathbf{z}_{i1}^*, \log(y_{i1}^*), v_{i1}^* \right), \dots, \left(\mathbf{x}_{in_i^*}^*, \mathbf{z}_{in_i^*}^*, \log(y_{in_i^*}^*), v_{in_i^*}^* \right), \quad i = 1, \dots, m.$$

Note that for each risk i , the new sample size is n_i^* ($n_i^* \leq n_i$).

(2) *Detection of Between-Risk Outliers*

In the second step, the procedure searches the pre-cleaned sample (marked with $*$) and discards entire risks whose risk-specific profile expressed by the random effect significantly deviates from the overall portfolio profile. These risks are identified when their robustified Mahalanobis distance exceeds some adaptive cut-off point. The process results in

$$\left(\mathbf{x}_{i1}^{**}, \mathbf{z}_{i1}^{**}, \log(y_{i1}^{**}), v_{i1}^{**} \right), \dots, \left(\mathbf{x}_{in_i^*}^{**}, \mathbf{z}_{in_i^*}^{**}, \log(y_{in_i^*}^{**}), v_{in_i^*}^{**} \right), \quad i = 1, \dots, i^*,$$

a “cleaned” sample of risks. Note that the number of remaining risks is i^* ($i^* \leq m$).

(3) *CATL Estimators*

In the final step, the CATL procedure employs traditional likelihood-based methods, such as (restricted) maximum likelihood, on the cleaned sample and computes reweighted parameter estimates $\widehat{\boldsymbol{\beta}}_{\text{CATL}}$ and $\widehat{\boldsymbol{\theta}}_{\text{CATL}} = (\widehat{\sigma}_{u_1}^2, \dots, \widehat{\sigma}_{u_q}^2, \widehat{\sigma}_\varepsilon^2)$. Here, the subscript CATL emphasizes that the maximum likelihood type estimators are not computed on the original sample (i.e., the starting point of Step 1), but rather on the cleaned sample, which is the end result of Step 2.

Using the described procedure, we find the shifted *robust best linear unbiased predictor* for location:

$$\widehat{\lambda}_i = \mathbf{X}_i^{**} \widehat{\boldsymbol{\beta}}_{\text{CATL}} + \mathbf{Z}_i^{**} \widehat{\mathbf{u}}_{\text{rBLUP}, i} + \widehat{\mathbf{E}}_{F_0}(\boldsymbol{\varepsilon}_i), \quad i = 1, \dots, m,$$

where $\widehat{\boldsymbol{\beta}}_{\text{CATL}}$ and $\widehat{\mathbf{u}}_{\text{rBLUP}, i}$ are standard likelihood-based estimators but computed on the clean sample from Step 2. Also, $\widehat{\mathbf{E}}_{F_0}(\boldsymbol{\varepsilon}_i)$ is the expectation vector of the n_i^* -variate cdf $F_{n_i^*}(\mathbf{0}, \widehat{\Sigma}_i)$. For symmetric error distributions we obtain the special case $\widehat{\mathbf{E}}_{F_0}(\boldsymbol{\varepsilon}_i) = \mathbf{0}$.

9.4.3 Robust Credibility Ratemaking

The reweighted estimates for location, $\widehat{\lambda}_i$, and structural parameters, $\widehat{\boldsymbol{\theta}}_{\text{CATL}} = (\widehat{\sigma}_{u_1}^2, \dots, \widehat{\sigma}_{u_q}^2, \widehat{\sigma}_\varepsilon^2)$, are used to calculate robust credibility premiums for the ordinary but heavy-tailed claims part of the original data. The robust ordinary net premiums

$$\widehat{\mu}_{it}^{\text{ordinary}} = \widehat{\mu}_{it}^{\text{ordinary}}(\widehat{\mathbf{u}}_{\text{rBLUP}, i}), \quad t = 1, \dots, n_i + 1, \quad i = 1, \dots, m$$

are found by computing the empirical *limited expected value* (LEV) of the fitted log-location distribution of claims. The percentile levels of the lower bound q_l and the upper bound q_g used in LEV computations are usually chosen to be extreme (e.g., 0.1% for q_l and 99.9% for q_g).

Then, robust regression is used to price separately identified excess claims. The risk-specific excess claim amount of insured i at time t is defined by

$$\widehat{O}_{it} = \begin{cases} -\widehat{\mu}_{it}^{\text{ordinary}}, & \text{for } y_{it} < q_l. \\ (y_{it} - q_l) - \widehat{\mu}_{it}^{\text{ordinary}}, & \text{for } q_l \leq y_{it} < q_g. \\ (q_g - q_l) - \widehat{\mu}_{it}^{\text{ordinary}}, & \text{for } y_{it} \geq q_g. \end{cases}$$

Further, let m_t denote the number of insureds in the portfolio at time t and let $N = \max_{1 \leq i \leq m} n_i$, the maximum horizon among all risks. For each period $t = 1, \dots, N$, we find the mean cross-sectional overshoot of excess claims $\widehat{O}_{\bullet t} = m_t^{-1} \sum_{i=1}^{m_t} \widehat{O}_{it}$, and fit robustly the random effects model

$$\widehat{O}_{\bullet t} = \mathbf{o}_t \boldsymbol{\xi} + \tilde{\varepsilon}_t, \quad t = 1, \dots, N,$$

where \mathbf{o}_t is the row-vector of covariates for the hypothetical mean of overshots $\boldsymbol{\xi}$. Here we choose $\mathbf{o}_t = 1$, and let $\widehat{\boldsymbol{\xi}}$ denote a robust estimate of $\boldsymbol{\xi}$. Then, the premium

for extraordinary claims, which is common to all risks i , is given by

$$\mu_{it}^{\text{extra}} = \mathbf{o}_t \hat{\boldsymbol{\xi}}.$$

Finally, the portfolio-unbiased robust regression credibility estimator is defined by

$$\widehat{\mu}_{i,n_i+1}^{\text{CATL}}(\widehat{\boldsymbol{u}}_{\text{rBLUP},i}) = \widehat{\mu}_{i,n_i+1}^{\text{ordinary}}(\widehat{\boldsymbol{u}}_{\text{rBLUP},i}) + \mu_{i,n_i+1}^{\text{extra}}, \quad i = 1, \dots, m.$$

From the actuarial point of view, premiums assigned to the insured have to be positive. Therefore, we determine pure premiums by $\max \{0, \widehat{\mu}_{i,n_i+1}^{\text{CATL}}(\widehat{\boldsymbol{u}}_{\text{rBLUP},i})\}$.

9.4.4 Numerical Examples Revisited

For Hachemeister's regression credibility model and its revised version, we use $\log(y_{it})$ as the response variable and fit it using the CATL method. In Table 9.1, we report loss predictions for individual states, which were computed using CATL and compared to those obtained by other authors: the BASE method, which is the linear trend model used by Goovaerts and Hoogstad (1987), and the M-RC, MM-RC, GM-RC procedures, which were studied by Pitselis (2008, 2012).

As discussed by Kaas et al. (1997) and Frees, Young, and Luo (2001), in practice it is fairly common to observe situations where risks with larger exposure measure exhibit lower variability. As one can infer from Table 9.3 (see Appendix), the number of claims per period, denoted by v_{it} , significantly affects the within-risk variability. Also, State 4 reports high average losses per claim, which in turn yields increased within-state variability (see Figure 9.1). Thus, to obtain homoskedastic error terms, we fit models using v_{it} as subject-specific weights. This model can be written as

$$\log(y_{it}) = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{z}_{it}\boldsymbol{u}_i + \varepsilon_{it} v_{it}^{1/2},$$

where ε_{it} is a sequence of independent normally distributed noise terms.

To assess the quality of credibility predictions, $\widehat{\mu}_{i,12+1}$, we also report their standard errors. These can be used to construct prediction intervals of the form BLUP (credibility estimate $\widehat{\mu}_{i,12+1}$) plus and minus multiples of the standard error. We estimate the standard error of prediction, $\widehat{\sigma}_{\widehat{\mu}_{i,12+1}}$, from the data using the common nonparametric estimator:

$$\begin{aligned} \widehat{\sigma}_{\widehat{\mu}_{i,12+1}} &= \left[\widehat{\text{MSE}}(\widehat{\mu}_i) - \widehat{\text{bias}}^2(\widehat{\mu}_i) \right]^{1/2} \\ &= \left[v_{i\bullet}^{-1} \sum_{t=1}^{12} \omega_{it} v_{it} (y_{it} - \widehat{\mu}_{it})^2 - \left(v_{i\bullet}^{-1} \sum_{t=1}^{12} \omega_{it} v_{it} (y_{it} - \widehat{\mu}_{it}) \right)^2 \right]^{1/2}, \end{aligned}$$

where $\widehat{\mu}_{it}$ denotes the credibility estimate obtained from the pursued regression method, $v_{i\bullet}$ is the total number of claims in state i , and ω_{it} is the hard-rejection weight for the observed average loss per claim y_{it} when using the CATL procedure. For nonrobust REML where no data points are truncated we put $\omega_{it} = 1$ as a special case.

Table 9.1. Individual State Predictions for Hachemeister's Bodily Injury Data Based on Various Model-Fitting Procedures (if available, estimated standard errors are provided in parentheses)

Fitting Procedure	Prediction for State				
	1	2	3	4	5
BASE	2436	1650	2073	1507	1759
M-RC ($c = 1.5$)	2437	1650	2073	1507	1759
GM-RC ($k = 1$)	2427	1648	2092	1505	1737
MM-RC	2427	1648	2092	1505	1737
REML	2465 (109)	1625 (122)	2077 (193)	1519 (248)	1695 (77)
CATL	2471 (111)	1545 (74)	2065 (194)	1447 (174)	1691 (57)
REML*	2451 (109)	1661 (123)	2065 (193)	1613 (242)	1706 (78)
CATL*	2450 (113)	1552 (74)	2049 (195)	1477 (172)	1693 (57)

* Based on the revised Hachemeister model (see Section 9.2.2.2).

Several conclusions emerge from Table 9.1. First, we note that the REML and REML* estimates (predictions) are based on Henderson's mixed model equations, and thus they slightly differ from those of Section 9.3. Second, for States 1, 3, and 5, all techniques, standard and robust, yield similar predictions. For the second and fourth states, however, CATL produces slightly lower predictions. For instance, for State 4 it results in 1447 whereas the base model finds 1507 predictions. This can be traced back to the truncation of the suspicious observations #6 and #10 in State 2 and #7 in State 4. Third, CATL also identifies claim #4 in State 5 as an outlier and, as a result, assigns a small discount of -1.47 to each risk. For the revised model (i.e., for REML* and CATL*), prediction patterns are similar.

To illustrate robustness of regression credibility methods that are based on M-RC, MM-RC, and GM-RC estimators for quantifying an individual's risk experience, Pitselis (2008) replaces the last observation of the fifth state, 1,690, by 5,000. We follow the same contamination strategy and summarize our findings in Table 9.2. As one can see, the choice of the model-fitting methodology has a major impact on predictions. Indeed, in the presence of a single outlier, we find that robust procedures provide stability and reasonable adjustment to predictions, whereas standard methods overreact. For example, in the contaminated State 5 the REML and BASE credibility estimates get inflated by the outlying observation and increase from 1,695 to 2542 and from 1759 to 2596, respectively. Further, because outliers usually distort the estimation process of variance components and thus yield too low credibility weights, predictions across all states increase significantly. Note also a dramatic explosion of standard errors (e.g., the standard error of REML in State 5 jumps from 77 to 829), which is due to increased within-risk variability that was caused by the contaminating observation. Furthermore, not all robust credibility predictions react equally to data

Table 9.2. *Individual State Predictions for Contaminated Hachemeister's Data Based on Various Model-Fitting Procedures (if available, estimated standard errors are provided in parentheses)*

Fitting Procedure	Prediction for State				
	1	2	3	4	5
BASE	2501	1826	2181	1994	2596
M-RC ($c = 1.5$)	2755	1979	2396	1841	2121
GM-RC ($k = 1$)	2645	1868	2311	1723	1964
MM-RC	2649	1870	2315	1724	1943
REML	2517 (119)	1852 (150)	2206 (204)	1987 (255)	2542 (829)
CATL	2477 (111)	1550 (74)	2071 (194)	1452 (174)	1689 (60)
REML*	2455 (110)	1949 (166)	2229 (204)	2141 (275)	2629 (818)
CATL*	2459 (112)	1559 (74)	2057 (195)	1484 (172)	1694 (60)

* Based on the revised Hachemeister model (see Section 9.2.2.2).

contamination. The CATL-based predictions change only slightly when compared to the noncontaminated data case, but those of M-RC, GM-RC, and MM-RC shift upward by 10%–20%. For instance, predictions for State 5 change from 1691 to 1689 (for CATL), 1759 to 2121 (for M-RC), 1737 to 1964 (for GM-RC), and 1737 to 1943 (for MM-RC). This variation can be explained by the fact that the latter group of procedures does not provide protection against large claims that influence the between-risk variability. Note also that in all but contaminated State 5 the CATL predictions slightly increase, whereas the corresponding standard errors remain unchanged. In contrast, State 5 prediction is practically unchanged, but the CATL method “penalizes” the state by increasing its standard error (i.e., the standard error has changed from 57 to 60). An increase in standard error implies a credibility reduction for State 5.

As the last point of this discussion, Figure 9.2 illustrates the impact of data contamination on the ratemaking process. The left panel plots claim severity over time for two noncontaminated states, State 1 and State 3, with the last three time periods representing the point and interval predictions. The right panel plots corresponding results for the contaminated State 5. In both cases the REML-based inference leads to elevated point predictions and wider intervals (which are constructed by adding and subtracting one standard error to the point prediction). As one would expect, the REML and CATL predictions are most disparate in State 5.

9.5 Further Reading

The regression-type credibility models of this chapter have been extended and generalized in several directions. To get a broader view of this topic, we encourage the reader to consult other papers and textbooks. For example, to learn how to model

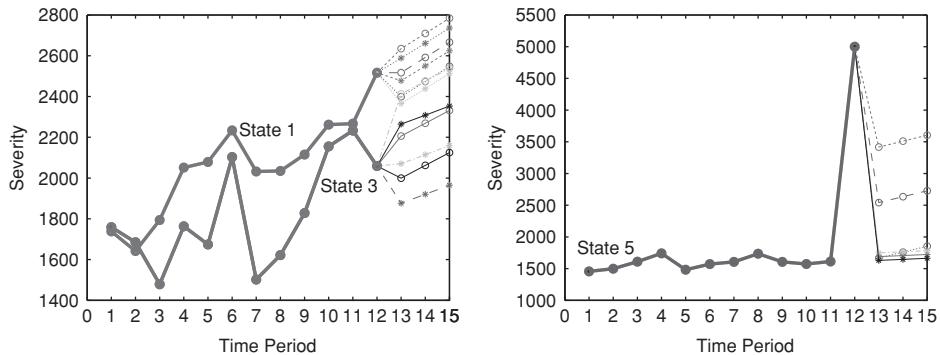


Fig. 9.2. Selected point and interval predictions for contaminated Hachemeister's data. The thin lines connecting “o” denote one-, two-, and three-step predictions using REML. The corresponding CATL predictions are marked by *.

correlated claims data, we recommend reading Frees et al. (1999, 2001) and Frees (2004). Further, to gain a deeper understanding of and appreciation for robust credibility techniques, the reader should review Garrido and Pitselis (2000), Pitselis (2004, 2008, 2012), and Dornheim and Brazauskas (2007, 2011b,a). If the reader is not familiar with the philosophy and methods of robust statistics, then the book by Maronna, Martin, and Yohai (2006) will provide a gentle introduction into the subject. Finally, an introduction to and developments of hierarchical credibility modeling can be found in Sundt (1979, 1980), Norberg (1986), Bühlmann and Jewell (1987), and Belhadj, Goulet, and Ouellet (2009).

9.6 Appendix

Table 9.3. Hachemeister's Bodily Injury Dataset Comprising Average Loss Per Claim, y_{it} , and the Corresponding Number of Claims Per Period, v_{it}

Period	Average Loss Per Claim in State					Number of Claims Per Period in State				
	1	2	3	4	5	1	2	3	4	5
1	1738	1364	1759	1223	1456	7861	1622	1147	407	2902
2	1642	1408	1685	1146	1499	9251	1742	1357	396	3172
3	1794	1597	1479	1010	1609	8706	1523	1329	348	3046
4	2051	1444	1763	1257	1741	8575	1515	1204	341	3068
5	2079	1342	1674	1426	1482	7917	1622	998	315	2693
6	2234	1675	2103	1532	1572	8263	1602	1077	328	2910
7	2032	1470	1502	1953	1606	9456	1964	1277	352	3275
8	2035	1448	1622	1123	1735	8003	1515	1218	331	2697
9	2115	1464	1828	1343	1607	7365	1527	896	287	2663
10	2262	1831	2155	1243	1573	7832	1748	1003	384	3017
11	2267	1612	2233	1762	1613	7849	1654	1108	321	3242
12	2517	1471	2059	1306	1690	9077	1861	1121	342	3425

Source: Hachemeister (1975), figure 3.

References

- Belhadj, H., V. Goulet, and T. Ouellet (2009). On parameter estimation in hierarchical credibility. *ASTIN Bulletin* 39(2), 495–514.
- Bühlmann, H. (1967). Experience rating and credibility. *ASTIN Bulletin* 4, 199–207.
- Bühlmann, H. and A. Gisler (1997). Credibility in the regression case revisited. *ASTIN Bulletin* 27, 83–98.
- Bühlmann, H. and A. Gisler (2005). *A Course in Credibility Theory and Its Applications*. Springer, New York.
- Bühlmann, H. and W. Jewell (1987). Hierarchical credibility revisited. *Bulletin of the Swiss Association of Actuaries* 87, 35–54.
- Bühlmann, H. and E. Straub (1970). Glaubwürdigkeit für Schadensätze. *Mitteilungen der Vereinigung Schweizerischer Versicherungsmathematiker* 70, 111–133.
- Dannenburg, D. R., R. Kaas, and M. J. Goovaerts (1996). *Practical Actuarial Credibility Models*. Institute of Actuarial Science and Economics, University of Amsterdam.
- Dornheim, H. (2009). *Robust-Efficient Fitting of Mixed Linear Models: Theory, Simulations, Actuarial Extensions, and Examples*. Ph.D. thesis, University of Wisconsin-Milwaukee.
- Dornheim, H. and V. Brazauskas (2007). Robust-efficient methods for credibility when claims are approximately gamma-distributed. *North American Actuarial Journal* 11(3), 138–158.
- Dornheim, H. and V. Brazauskas (2011a). Robust-efficient credibility models with heavy-tailed claims: A mixed linear models perspective. *Insurance: Mathematics and Economics* 48(1), 72–84.
- Dornheim, H. and V. Brazauskas (2011b). Robust-efficient fitting of mixed linear models: Methodology and theory. *Journal of Statistical Planning and Inference* 141(4), 1422–1435.
- Dutang, C., V. Goulet, X. Milhaud, and M. Pigeon (2012). Credibility theory features of actuar. <http://cran.r-project.org/web/packages/actuar/index.html>.
- Dutang, C., V. Goulet, and M. Pigeon (2008). Actuar: An R package for actuarial science. *Journal of Statistical Software* 25(7), 1–37.
- Frees, E. W. (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge University Press, Cambridge.
- Frees, E. W., V. R. Young, and Y. Luo (1999). A longitudinal data analysis interpretation of credibility models. *Insurance: Mathematics and Economics* 24, 229–247.
- Frees, E. W., V. R. Young, and Y. Luo (2001). Case studies using panel data models. *North American Actuarial Journal* 5(4), 24–42. Supplemental material is available at: <http://research3.bus.wisc.edu/course/view.php?id=129>.
- Garrido, J. and G. Pitselis (2000). On robust estimation in Bühlmann-Straub's credibility model. *Journal of Statistical Research* 34(2), 113–132.
- Goovaerts, A. S. and W. Hoogstad (1987). *Credibility Theory, Surveys of Actuarial Studies*. National-Nederlanden N.V., Rotterdam.
- Hachemeister, C. A. (1975). Credibility for regression models with applications to trend. In P. M. Kahn (Ed.), *Credibility: Theory and Applications*. Academic Press, New York.
- Kaas, R., D. Dannenburg, and M. Goovaerts (1997). Exact credibility for weighted observations. *ASTIN Bulletin* 27, 287–295.
- Keffer, R. (1929). An experience rating formula. *Transactions of the Actuarial Society of America* 30, 130–139.
- Klugman, S., H. Panjer, and G. Willmot (2012). *Loss Models: From Data to Decisions* (3rd ed.). Wiley, New York.
- Maronna, R. A., D. R. Martin, and V. J. Yohai (2006). *Robust Statistics: Theory and Methods*. Wiley, New York.

- Mowbray, A. H. (1914). How extensive a payroll exposure is necessary to give a dependable pure premium? *Proceedings of the Casualty Actuarial Society I*, 25–30.
- Norberg, R. (1980). Empirical Bayes credibility. *Scandinavian Actuarial Journal 1980*, 177–194.
- Norberg, R. (1986). Hierarchical credibility: Analysis of a random effect linear model with nested classification. *Scandinavian Actuarial Journal 1986*, 204–222.
- Pitselis, G. (2004). A seemingly unrelated regression model in a credibility framework. *Insurance: Mathematics and Economics 34*, 37–54.
- Pitselis, G. (2008). Robust regression credibility: The influence function approach. *Insurance: Mathematics and Economics 42*, 288–300.
- Pitselis, G. (2012). A review on robust estimators applied to regression credibility. *Journal of Computational and Applied Mathematics 239*, 231–249.
- Sundt, B. (1979). A hierarchical regression credibility model. *Scandinavian Actuarial Journal 1979*, 107–114.
- Sundt, B. (1980). A multi-level hierarchical credibility regression model. *Scandinavian Actuarial Journal 1980*, 25–32.
- Whitney, A. W. (1918). The theory of experience rating. *Proceedings of the Casualty Actuarial Society IV*, 275–293.

10

Fat-Tailed Regression Models

Peng Shi

Chapter Preview. In the actuarial context, fat-tailed phenomena are often observed where the probability of extreme events is higher than that implied by the normal distribution. The traditional regression, emphasizing the center of the distribution, might not be appropriate when dealing with data with fat-tailed properties. Overlooking the extreme values in the tail could lead to biased inference for rate-making and valuation. In response, this chapter discusses four fat-tailed regression techniques that fully use the information from the entire distribution: transformation, models based on the exponential family, models based on generalized distributions, and median regression.

10.1 Introduction

Insurance ratemaking is a classic actuarial problem in property-casualty insurance where actuaries determine the rates or premiums for insurance products. The primary goal in the ratemaking process is to precisely predict the expected claims cost which serves as the basis for pure premiums calculation. Regression techniques are useful in this process because future events are usually forecasted from past occurrence based on the statistical relationships between outcomes and explanatory variables. This is particularly true for personal lines of business where insurers usually possess large amount of information on policyholders that could be valuable predictors in the determination of mean cost.

The traditional mean regression, though focusing on the center of the distribution, relies on the normality of the response variable. It is well known that insurance claims data are rarely normal. In actuarial practice, fat-tailed phenomena are often observed where the probability of extreme events is higher than that implied by the normal distribution. Data that sometimes exhibit fat tails are also described as heavy-tailed or long-tailed distributed. This observation motivates the discussion of regression techniques for fat-tailed data in this chapter.

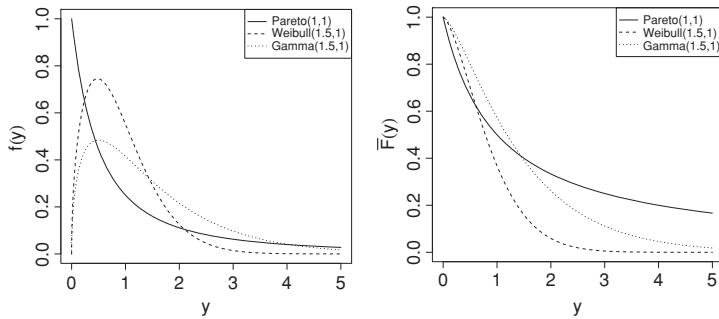


Fig. 10.1. Density and survival function of fat-tailed distributions.

In fact, fat-tailed distributions are not new to the actuarial literature and have been employed for a variety of actuarial applications. For example, in severity modeling, actuaries examine the distribution of individual claims using historically observed data; in loss reserving, valuation actuaries are interested in the distribution of aggregated losses from a single line or multiple lines of business; in risk management, actuaries use risk metrics, such as value-at-risk, of the loss distribution of an insurance portfolio to determine the risk capital required by regulatory bodies. Historically, actuaries have been exposed to fat-tailed data modeling extensively in nonlife insurance where heavy-tailed distributions are often used for the analysis without inclusion of covariates (see Klugman, Panjer, and Willmot 2008). In contrast, this chapter extends such concerns to situations in which covariates are available.

When compared with the normal distribution, tail heaviness of a certain distribution is usually measured by moment-based statistics, such as skewness and kurtosis coefficients. However, moments only provide partial distributional information and are not guaranteed to be finite for many random variables. We focus on another metric of fat tails where the size is measured by distribution functions (see the Appendix for the technical definition). Essentially, we examine how fast the survival function decays as the variable increases. In the following, we define distributions with a lower decay than the normal as fat-tailed distributions. The interpretation is rather intuitive: the fat-tailed random variable has a much higher probability of exceeding the threshold than the normal random variable. For example, one can show that both gamma and Pareto have heavier tails than the normal distribution. To visualize the tail heaviness, we exhibit in Figure 10.1 the density and survival function of three commonly used fat-tailed distributions in actuarial science: gamma, Weibull, and Pareto. The first panel shows that fat tails come along with skewness, and the second panel demonstrates that the survival function approaches zero at different rates.

This figure emphasize the right tail because extreme events, in insurance applications, often represent unusual large claims. However, a similar analysis could be

performed for the left tail of distributions as well. An example could be the downside risk in financial markets.

Note that the normal distribution is not required to be the benchmark when studying tail behaviors. In fact, actuaries might have observed heavy-tailed distributions in insurance risk models (see Panjer and Willmot 1992), where the criterion of subexponential is used to define heavy tail. For regression analysis, we use the normal distribution as the benchmark because it is the key assumption embedded in the traditional mean regression model. Through minimizing a symmetric loss function, the traditional regression focuses on the center of the distribution and downplays extreme values in the tail. Thus a direct application of usual regression to fat-tailed outcomes could lead to biased statistical inference, especially for small sample studies. For example, the hypothesis test of statistical significance in finite sample applications would no longer be valid because of the heavy-tailed distribution. Even in the case of large samples, the tail heaviness could affect the normality of parameter estimates and cause asymptotic efficiency loss.

In this chapter, we introduce four regression techniques for fat-tailed data. We start with transformation techniques. Using transformation, one symmetrizes the distribution of heavy-tailed outcomes so as to apply the usual least squares regression. The second approach is to use generalized linear models (GLMs). Instead of focusing on the link function and the variance function, we view the GLMs as a parametric model where one has the option to select the appropriate distribution for fat-tailed data from the exponential family. Extending the idea of parametric regressions, we further discuss fat-tailed regressions using generalized distributions. In particular, we introduce covariates into a generalized beta distribution that nests many commonly used fat-tailed distribution as special cases. Finally, we present a nonparametric technique – quantile regression. Unlike mean regression, quantile regression looks into the quantiles of the response rather than the mean. We highlight the median regression, a special quantile that is relevant to insurance ratemaking.

10.2 Transformation

When modeling claim size, actuaries often use the natural log of claims rather than the original claims directly in the ordinary regression analysis. This is because insurance claims data often present long tails and the logarithmic transformation helps symmetrize the claims distribution. The approach has its strengths and limitations. It is straightforward to apply, and the regression coefficients can be easily interpreted in terms of proportional change. However, the transformation also means switching from an additive to a multiplicative mean structure, as well as changing the variance structure. These changes might be appropriate for some datasets, but for others, actuaries might only want to change the distribution, not the mean and variance structure.

To formalize the idea of transformation, one transforms the original observations using some suitable nonlinear transformation and then applies the ordinary least squares for the transformed observations. The usual regression relies on restrictive assumptions such as linear mean structure, normality of distributions, and constant error variance. Hence it is not ready to be applied to fat-tailed responses where skewness is typically implied by the correlation of variability with mean. Presumably there exists an appropriate transformation $\psi(y, \theta)$ with parameter θ , in which one expects to have the relation

$$\psi(y, \theta) = E(\psi(y, \theta)|\mathbf{x}) + \varepsilon, \quad (10.1)$$

with a simple structure of mean $E(\psi(y, \theta)|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$ and a normal error term $\varepsilon \sim N(0, \sigma^2)$. Though transformation techniques could be applied to both response and explanatory variables, our discussion refers to functions of dependent variables.

As foreshadowed earlier, the transformation method enjoys the advantage of simple implementation. However, the transformation also changes the mean structure, and thus the interpretation of regression coefficients is not straightforward. In the classical linear regression model, coefficients are interpreted in terms of partial derivatives. In the transformation model (10.1), one could be tempted to write $\partial E(y|\mathbf{x})/\partial x_j = \partial\psi^{-1}(\mathbf{x}'\boldsymbol{\beta}, \theta)/\partial x_j$, such as is the common practice in logarithmic models. However, the difficulty of this practice is obvious because $E\psi(y)$ is not equivalent to $\psi(Ey)$ in general circumstances.

10.2.1 Alternative Transformations

Numerous transformations have been examined in the statistics literature, with an emphasis on those that achieve constancy of error variance. As noted by Bartlett (1947), variance stabilizing transformation often has the effect of improving the closeness of the distribution to normality. In this section, we summarize some transformation methods that have been used to improve normality and stabilize variance in regression analysis:

- Box-Cox: $\psi(y, \theta) = \begin{cases} (y^\theta - 1)/\theta, & \theta > 0 \\ \ln y, & \theta = 0 \end{cases}$ or $\psi(y, \theta) = \begin{cases} \{(y + \theta_2)^{\theta_1} - 1\}/\theta_1, & \theta_1 > 0 \\ \ln(y + \theta_2), & \theta_1 = 0 \end{cases}$
- Signed power: $\psi(y, \theta) = \{\text{sign}(y)|y|^\theta - 1\}/\theta, \theta > 0$
- Modulus: $\psi(y, \theta) = \begin{cases} \text{sign}(y)\{(|y| + 1)^\theta - 1\}/\theta, & \theta \neq 0 \\ \text{sign}(y) \ln(|y| + 1), & \theta = 0 \end{cases}$
- Inverse hyperbolic sine: $\psi(y, \theta) = \sinh^{-1}(\theta y)/\theta, \theta \geq 0$

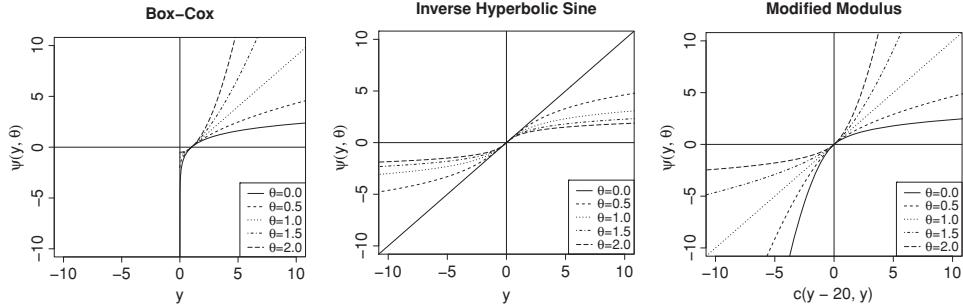


Fig. 10.2. Illustration of transformations. The line type in the legend from top to bottom corresponds to $\theta = 0, 0.5, 1, 1.5, 2$.

- Modified modulus: $\psi(y, \theta) = \begin{cases} \{(y + 1)^\theta - 1\}/\theta, & y \geq 0, \theta \neq 0 \\ \ln(y + 1), & y \geq 0, \theta = 0 \\ -\{(-y + 1)^{2-\theta} - 1\}/(2 - \theta), & y < 0, \theta \neq 2 \\ -\ln(-y + 1), & y < 0, \theta = 2 \end{cases}$

Each transformation family has its own strength and weakness and might be suitable for different types of data. Box-Cox transformation is a popular power family that has been extensively used in practice (Box and Cox 1964). The first form is only valid for positive observations. The second form, handling bounded negative observations through a shift parameter, may sacrifice the asymptotic properties of maximum likelihood method because of the unknown parameter. Many other transformations have been considered to circumvent these difficulties: two examples are the signed power transformation (Bickel and Doksum 1981) and the modulus transformation (John and Draper 1980). Both families are appropriate for either positive or negative outcomes. However, the log-likelihood is not well defined for the signed power transformation when there exist some zeros. In addition, both transformations could lead to bimodal distributions in the case where the support consists of the whole real line, because the Jacobian is not a monotone function when the response changes sign. This issue is addressed by the modified modulus family proposed in Yeo and Johnson (2000), where the modulus transformation is forced to be smooth by imposing different parameters on the positive and negative line. In addition to power functions, variables could be transformed based on other families. The inverse hyperbolic sine due to Burbidge, Magee, and Robb (1988) is an example. This family can be applied to responses assuming both positive and negative values and is shown to be useful in shrinking the extreme observations.

Figure 10.2 illustrates the difference among selected transformations. We observe that the modulus transformation on the positive line could be obtained by shifting the Box-Cox transformation by constant one. Replacing parameter θ with $2 - \theta$ when the

response changes sign, the modified modulus is continuous to the second derivative. Similar to the modulus family, the inverse hyperbolic sine transform suffers from the discontinuity at zero.

10.2.2 Parameter Estimation

For a given transform family, the best transformation could be derived using likelihood-based inference. Considering model (10.1), the log-likelihood function associated with independent original observations (y_1, \dots, y_n) could be shown as

$$l(\theta, \boldsymbol{\beta}, \sigma^2) \propto -\frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (\psi(y_i, \theta) - \mathbf{x}'_i \boldsymbol{\beta}) + \sum_{i=1}^n J(y_i, \theta)$$

with $J(y_i, \theta) = |\psi'(y_i, \theta)|$. First we maximize the log-likelihood function holding θ fixed. Denote $\mathbf{y} = (y_1, \dots, y_n)'$ and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$, one could show

$$\hat{\boldsymbol{\beta}}(\theta) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\psi(\mathbf{y}, \theta) \text{ and}$$

$$\hat{\sigma}^2(\theta) = \psi(\mathbf{y}, \theta)'(I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X})\psi(\mathbf{y}, \theta)/n = S(\mathbf{y}, \theta)/n,$$

where $\hat{\boldsymbol{\beta}}(\theta)$ is the standard least square estimates for dependent variable $\psi(\mathbf{y}, \theta)$ and $S(\mathbf{y}, \theta)$ is the residual sum of squares in the corresponding variance analysis. Then maximizing the concentrated log-likelihood $l(\theta, \hat{\boldsymbol{\beta}}(\theta), \hat{\sigma}^2(\theta))$ provides the maximum likelihood estimate of θ , which, under certain regularity conditions, is strongly consistent and minimizes the Kullback-Leibler information. In addition, the concentrated likelihood will reduce to

$$l(\theta) = -\frac{1}{2} \ln(S(\tilde{\mathbf{y}}, \theta)/n),$$

if one works with rescaled transformation $\tilde{y}_i = y_i/\dot{y}$ with $\dot{y} = \prod_{i=1}^n J(y_i, \theta)$. Here $S(\tilde{\mathbf{y}}, \theta)$ is the residual sum of squares of $\tilde{\mathbf{y}}$, and the maximum likelihood estimate could be obtained by minimizing $S(\tilde{\mathbf{y}}, \theta)$ with respect to θ .

10.3 GLM

The generalized linear model (GLM) is a widely used tool for ratemaking in property-casualty insurance. In the well-known two-part model, actuaries often use Poisson regression to model the frequency of claims and gamma regression to model the size of claims. See Chapter 5 for the theory and technical details of GLMs.

In severity modeling, the fat-tailed claims often come with the dependence of variance on the mean. Thus, actuaries usually go from linear regression to a GLM for a link function for the mean and a variance function. However, it is also useful to think of GLMs as parametric regressions based on the exponential family of distributions.

Table 10.1. Fitting GLM and GAM for Insurance Claims

	Model 1		Model 2		Model 3	
	Estimate	p-value	Estimate	p-value	Estimate	p-value
intercept	4.414	<0.001	3.781	<0.001	5.823	<0.001
year2	-0.023	0.667	-0.022	0.670	-0.019	0.711
year3	0.097	0.066	0.098	0.057	0.102	0.051
year4	0.070	0.184	0.071	0.165	0.075	0.149
year5	0.092	0.082	0.095	0.065	0.100	0.057
ppsm	0.174	<0.001	0.169	<0.001	0.174	<0.001
pci	-0.033	<0.001			-0.167	0.001
s(pci)			2.793	<0.001		
pci ²					0.003	0.010
logLik	-677.354		-672.347		-673.616	
AIC	1370.707		1364.280		1365.232	

Using a GLM, actuaries have the opportunity to select an entire distribution for the claims, not just the mean and variance. For example, one could use a gamma distribution for medium-tail claims and an inverse-Gaussian distribution for long-tail claims. This could be particularly important for prediction purposes.

Another well-known class of statistical models based on the exponential family of distributions is the generalized additive model (GAM). The GAM replaces the linear predictors in GLM with additive predictors. A detailed discussion on GAMs is found in Chapter 15.

Though both are based on the exponential family, the GLM and GAM can be used for different analytic purposes. Using nonparametric and smoothing techniques, a GAM provides good fits to the training data and is more suitable for detecting the relationship between the response and covariates. In contrast, a GLM focuses on estimation and inference with a better interpretability of results. This point is demonstrated through the following data example.

We consider a dataset on insurance claims used in Frees and Wang (2005). It contains automobile bodily injury liability claims from a sample of 29 Massachusetts towns over five years, 1993–1997. The response is the average claims per unit of exposure (the pure premium) adjusted for the effects of inflation. Two explanatory variables are per capita income (pci) and logarithmic population per square mile (ppsm). The original analysis shows that the gamma regression is a reasonable candidate for the insurance claims. Thus we first fit a generalized linear model using the gamma distribution with a logarithmic link function. Time effects are accommodated with binary variables. Estimates are displayed as Model 1 in Table 10.1. Consistently, we observe the significant positive relation with ppsm and negative relation with pci.

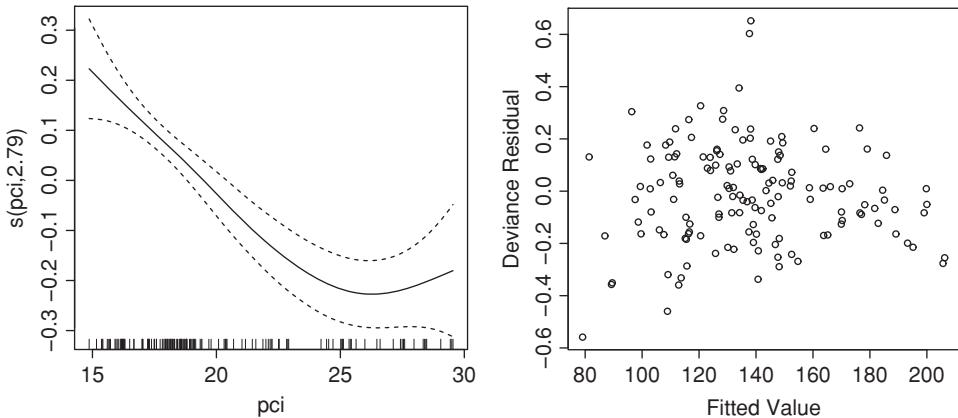


Fig. 10.3. The left panel plots partial residuals from the GAM against covariate. The right panel plots deviance residuals from the GLM against fitted value.

If one wants to explore the effect of pci further, one could fit a generalized additive model with an additive component $s(\text{pci})$, shown as Model 2 in Table 10.1. To facilitate a reasonable comparison, the GAM is also fitted with the gamma distribution and a log link. The small p -value indicates statistical significance, and the degrees of freedom value of 2.793 suggests a quadratic pattern. This relation could also be visualized through an analysis of partial residuals that are the Pearson residuals added to the partial prediction of the smooth term. We show in the left panel of Figure 10.3 the partial residual against pci , where the rug plots at the bottom indicate the value of the covariate. After uncovering the quadratic form of the dependence on pci using a GAM, we can use a GLM to fit and assess the corresponding parametric model. In Model 3, a quadratic term of pci is included in addition to the linear one. As anticipated, the nonlinear effect is well captured in this formulation. The plot of deviance residuals against fitted values is exhibited in the right panel of Figure 10.3, and no particular patterns are found. The goodness-of-fit statistics reported at the bottom of Table 10.1 show that the GAM and the GLM with a quadratic term provide similar fits and both outperform the simple GLM model. An F -test could be used to compare different models, though certain types of approximations are involved. A p -value of 0.09 suggests that the additive structure can be rejected at 5% significance level when comparing Model 2 and Model 3. However, a p -value less than 0.01 implies the superior fit of Model 3 to Model 1.

10.4 Regression with Generalized Distributions

Parametric distributions have been extensively used in modeling heavy-tailed data in actuarial science (Klugman et al. 2008). However, applications of regression models

based on parametric distributions are less common. One can think of, for example, the Pareto distribution. It is probably the most commonly used severity distribution, along with the gamma, by actuaries. The Pareto distribution captures well the situation where the majority of claims are small, but there is the potential for very large claims (heavy tails). It is particularly prevalent in claims modeling for liability insurance and excess of loss reinsurance. However, actuaries often need to split the data to fit separate distributions, thereby allowing for the heterogeneity in the claims of each subpopulation, for instance, male versus female. A more appealing approach is to introduce covariates when fitting a Pareto distribution instead of splitting the data. Relaxing the distributional assumption of an exponential family, this section discusses regression techniques based on more generalized distributions that could accommodate fat-tailed data.

10.4.1 The GB2 Distribution

This section discusses a four-parameter distributional family, the generalized beta of the second kind (*GB2*), and introduces regression techniques for the *GB2* model. Introduced by McDonald (1984), the *GB2* nests more than 20 distributions as limiting cases, including many commonly used skewed distributions in actuarial science.

In general, one could use transformation techniques to create new distributions. A *GB2* distribution can be generated from a beta random variable $B(\phi_1, \phi_2)$ as

$$\ln Y = \mu + \sigma \ln \frac{B(\phi_1, \phi_2)}{1 - B(\phi_1, \phi_2)}.$$

Here $B(\phi_1, \phi_2)/[1 - B(\phi_1, \phi_2)]$ follows the beta prime distribution also known as the beta distribution of the second kind, and Y is known to follow a *GB2* distribution denoted by $GB2(\mu, \sigma, \phi_1, \phi_2)$. This result implies that the *GB2* distribution belongs to the log-location-scale family, a parametric model commonly used for analyzing duration data. Alternatively, using the relation between a beta distribution and an *F*-distribution,

$$\frac{2\phi_2 B(\phi_1, \phi_2)}{2\phi_1(1 - B(\phi_1, \phi_2))} \sim F(2\phi_1, 2\phi_2),$$

a *GB2* distribution can also be constructed from $Y = \exp(\mu)[\phi_1/\phi_2 F(2\phi_1, 2\phi_2)]^\sigma$.

Based on these relations, the density of a *GB2* distribution can be expressed as

$$f(y; \mu, \sigma, \phi_1, \phi_2) = \frac{[\exp(w)]^{\phi_1}}{y|\sigma|B(\phi_1, \phi_2)[1 + \exp(w)]^{\phi_1 + \phi_2}}$$

with $w = (\ln y - \mu)/\sigma$. As we see, a *GB2* distribution has four parameters; in this parametrization μ and σ are location and scale parameters, respectively, and ϕ_1 and ϕ_2 are shape parameters. The *GB2* is flexible in fitting long-tailed data, with the

distribution being left skewed when $\phi_1 > \phi_2$ and right skewed when $\phi_1 < \phi_2$. The k th moment exists $E(y^k) = \exp(k\mu)B(\phi_1 + k\sigma, \phi_2 - k\sigma)/B(\phi_1, \phi_2)$ for $\phi_1 < k\sigma < \phi_2$. Note that there are different ways to parameterize the *GB2* distribution. We choose this particular method to facilitate the regression analysis.

As pointed out already, the *GB2* is a log-location-scale model in the sense that $Z = \ln Y$ has density in the form $f(z) = f_0((z - \mu)/\sigma)/\sigma$. Under this parametrization, we have

$$f_0(z) = \frac{[\exp(z)]^{\phi_1}}{B(\phi_1, \phi_2)[1 + \exp(z)]^{\phi_1 + \phi_2}},$$

The variable Z is known to follow exponential *GB2* distribution denoted by *EGB2*($\mu, \sigma, \phi_1, \phi_2$), and the density f_0 corresponds to *EGB2*(0, 1, ϕ_1, ϕ_2).

Two special cases worth stressing are the Burr XII distribution and the generalized gamma distribution. The former has density

$$f(y; \xi, \tau, \varsigma) = \frac{\delta \xi^\delta \tau y^{\tau-1}}{(\xi + y^\tau)^{\delta+1}},$$

which could be obtained from a *GB2* using parameterizations $\mu = \ln \xi/\tau$, $\sigma = 1/\tau$, $\phi_1 = 1$, $\phi_2 = \delta$. An important special case is the Pareto distribution. For regression purposes, Beirlant et al. (1998) proposed two parameterizations to include covariates, through parameter $\tau = \exp(\mathbf{x}'\boldsymbol{\beta})$ or parameter $\xi = \exp(\tau \mathbf{x}'\boldsymbol{\beta})$. The latter has density

$$f(y; \eta, \omega, \kappa) = \frac{\lambda^\lambda}{\omega y \Gamma(\lambda) \sqrt{\lambda}} \exp\left(\text{sign}(\kappa) u \sqrt{\lambda} - \lambda \exp(\kappa u)\right)$$

with $\lambda = |\kappa|^{-2}$, $u = (\ln y - \eta)/\omega$. It can also be derived from a *GB2* when $\mu = \eta - \omega \sqrt{\lambda} \ln(\phi_1/\phi_2)$, $\sigma = \omega \sqrt{\lambda}$, $\phi_1 = \lambda$, $\phi_2 \rightarrow +\infty$. Special cases include standard gamma, log-normal, and Weibull distributions. The generalized gamma regression has been used in health economics literature to handle skewed medical costs (for example, see Manning, Basu, and Mullahy 2005), where both location parameter $\eta(\mathbf{x})$ and scale parameter $\sigma(\mathbf{x})$ could be allowed to vary with covariates.

10.4.2 Inference for Regression Model

Despite the flexibility of *GB2* in modeling heavy-tailed outcome data, its application in regression analysis is still sparse. McDonald and Butler (1990) first employed the *GB2* regression to investigate the duration of welfare spells. Recently, Sun, Frees, and Rosenberg (2008) applied the *GB2* in the prediction of nursing home utilization with longitudinal data. Frees and Valdez (2008) and Frees, Gao, and Rosenberg (2011) used the *GB2* to capture the long-tailed nature of auto insurance claims in a hierarchical insurance claims model.

A natural way to incorporate explanatory variables is to formulate the location parameter as a linear function of covariates $\mu = \mathbf{x}'\beta$. From the relation $E(Y) = \exp(\mu)B(\phi_1 + \sigma, \phi_2 - \sigma)/B(\phi_1, \phi_2)$, the regression coefficients could be interpreted as a proportional change. In theory, any parameter could be specified as a function of covariates. For example, one could include covariates in the scale parameter σ to allow for heteroskedasticity. Without loss of generality, we focus on this simple formulation. Because the regression model is fully parametric, the likelihood-based approach is the usual choice for estimation and inference. Given $\mathbf{y} = (y_1, \dots, y_n)$, the log-likelihood function is provided by

$$\begin{aligned}\ln L(\Theta|\mathbf{y}) &= \phi_1 \sum_{i=1}^n z_i - \sum_{i=1}^n \ln y_i - n \ln |\sigma| - n \ln B(\phi_1, \phi_2) \\ &\quad - (\phi_1 + \phi_2) \sum_{i=1}^n \ln[1 + \exp(z_i)].\end{aligned}$$

where $\Theta = (\sigma, \phi_1, \phi_2, \boldsymbol{\beta})$ and $z_i = (\ln y_i - \mathbf{x}_i' \boldsymbol{\beta})/\sigma$. The first-order derivatives of the log-likelihood function follow:

$$\begin{aligned}\frac{\partial \ln L(\Theta|\mathbf{y})}{\partial \sigma} &= -\frac{n}{\sigma} - \frac{\phi_1}{\sigma} \sum_{i=1}^n z_i + \frac{\phi_1 + \phi_2}{\sigma} \sum_{i=1}^n \frac{z_i \exp(z_i)}{1 + \exp(z_i)} \\ \frac{\partial \ln L(\Theta|\mathbf{y})}{\partial \phi_1} &= \sum_{i=1}^n z_i + n[\Psi(\phi_1) - \Psi(\phi_1 + \phi_2)] - \sum_{i=1}^n \ln[1 + \exp(z_i)] \\ \frac{\partial \ln L(\Theta|\mathbf{y})}{\partial \phi_2} &= n[\Psi(\phi_2) - \Psi(\phi_1 + \phi_2)] - \sum_{i=1}^n \ln[1 + \exp(z_i)] \\ \frac{\partial \ln L(\Theta|\mathbf{y})}{\partial \beta_j} &= -\frac{\phi_1}{\sigma} \sum_{i=1}^n x_{ij} + \frac{\phi_1 + \phi_2}{\sigma} \sum_{i=1}^n x_{ij} \frac{\exp(z_i)}{1 + \exp(z_i)}, \quad j = 1, \dots, p\end{aligned}$$

Thus, the maximum likelihood estimate $\hat{\Theta}$ of Θ is found by solving

$$\frac{\partial \ln L(\Theta|\mathbf{y})}{\partial \theta_j} \Big|_{\theta_j=\hat{\theta}_j} = 0, \quad j = 1, \dots, p+3$$

subject to constraints $\hat{\theta}_1 > 0$, $\hat{\theta}_2 > 0$, and $\hat{\theta}_3 > 0$. The asymptotic properties of maximum likelihood estimators could be used to construct corresponding t -ratios and confidence intervals for model parameters. Specifically, the asymptotic covariance matrix of $\hat{\Theta}$ could be derived from the Fisher information matrix $I(\Theta)$ and thus could be estimated by $\widehat{Acov}(\hat{\Theta}) = I^{-1}(\hat{\Theta})$. The detailed calculation of the Fisher matrix is provided in Appendix B.

As in the least squares regression, residual analysis could be used for model diagnostics. One way to define residuals is $r_i = (\ln y_i - \mathbf{x}'_i \boldsymbol{\beta})/\sigma$. Straightforward calculations show that r_i are i.i.d $EGB2(0, 1, \phi_1, \phi_2)$. Thus the distributional assumption could be validated by a QQ -plot:

$$\left(F_0^{-1} \left(\frac{i - 0.5}{n} \right), r_{(i)} \right), \quad i = 1, \dots, n$$

where F_0 denotes the distribution function of $EGB2(0, 1, \phi_1, \phi_2)$ and $r_{(i)}$ represents the ordered residuals with $r_{(1)} \leq \dots \leq r_{(n)}$. The distribution F_0 could be calculated based on the relation between the $GB2$ and the beta distribution or the F distribution. In addition, a plot of residuals against fitted values can be used to check the constancy of scale parameter σ .

10.4.3 Medical Care Example

10.4.3.1 Data

As a motivating example, we look into the medical care costs of a probability sample of the U.S. population. Modeling the monetary values of health care utilization is a critical issue in health-related studies, such as cost-effectiveness studies and disease management programs (see Frees et al. 2011). The dataset is from the Medical Expenditure Panel Survey of year 2008. We consider a subsample of non-elderly adults (age between 18 and 64) who have some form of private insurance and no form of public health insurance. In this example, we focus on the medical costs of a particular type of care services: office-based visits. Because of their semi-continuous nature, health care expenditures are often examined within a two-part framework, the frequency and severity. Our interest is the amount; thus all inferences are conditional on positive care consumption. The dataset we analyze includes 7,096 individuals with at least one office-based visit during the year.

The data also contain a set of explanatory variables that could affect health care utilization. These covariates are grouped into three categories: socioeconomic and demographic characteristics, health conditions and insurance, and employment status. The first group includes the age of the respondent, denoted by `age`, as a continuous covariate. The binary variables `female` and `married` indicate the gender and marital status of the respondent, respectively. Ethnicity is controlled by three indicators: `hispanic`, `black`, and `asian`. Residence information is distinguished by whether or not the respondent is from a metropolitan statistical area (`msa`). Other socioeconomic variables include the number of school years (`edu`) and the size of the family (`familysize`). An individual's health status is captured by a binary variable indicating if the person has a functional limitation (`limitation`), the number of chronic

Table 10.2. *Description and Sample Mean of Explanatory Variables*

Covariates	Description	Mean	StdDev
Demographic characteristics			
age	age in years	42.49	12.33
female	=1 if person is female	0.58	
married	=1 if person is married	0.66	
hispanic	=1 if person is Hispanic	0.16	
black	=1 if person is black	0.14	
asian	=1 if person is Asian	0.07	
edu	number of years of education	13.84	2.52
familysize	family size	2.98	1.45
msa	=1 if metropolitan statistical area	0.87	
Health conditions and health insurance			
limitation	=1 if physical limitation	0.18	
chronic	number of chronic conditions	1.60	1.65
excellent	=1 if self-perceived health is excellent	0.28	
verygood	=1 if self-perceived health is very good	0.35	
good	=1 if self-perceived health is good	0.27	
fair	=1 if self-perceived health is fair	0.08	
poor	=1 if self-perceived health is poor	0.02	
hmo	=1 if person is in an HMO	0.45	
Employment status			
employed	=1 if employed	0.85	

conditions (*chronic*), and a self-perceived health status that is indicated by five scales representing excellent, very good, good, fair, and poor.

The description and summary statistics of covariates are exhibited in Table 10.2. Among those who have positive medical care costs associated with office-based visits, more than half are female, 66% are married, and about 87% live in a metropolitan statistical area. The average individual is 42 years old who has a family of three members and receives 14 years of education. Regarding health conditions, 18% are found to have a functional physical limitation, and most individuals perceive themselves in a healthy state. It is not surprising that we observe a high percentage of employment because our sample is limited to individuals with only private insurance, which is often provided through employers as an employment benefit.

Noticeable features of medical care costs are their skewness and long tail, indicating the high likelihood of extremely expensive events. These are demonstrated by the histogram in Figure 10.4. A generalized linear model is a natural choice for the analysis, where the gamma and the inverse-Gaussian have the potential to handle

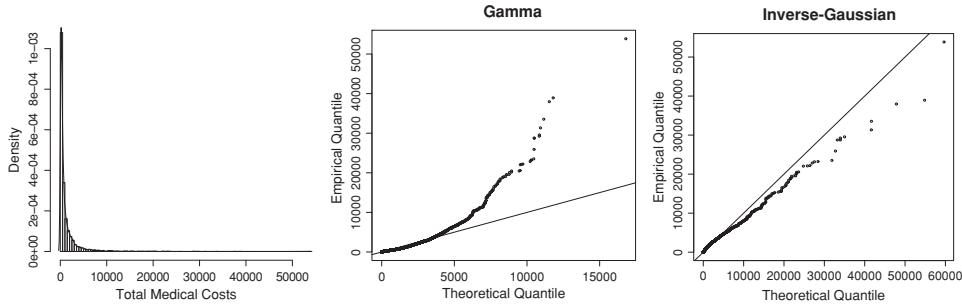


Fig. 10.4. Distribution of medical care costs for office-based visits and corresponding QQ -plots for gamma and inverse-Gaussian distributions.

heavy-tailed data. Figure 10.4 shows the QQ -plot for both distributions. Apparently, neither fits the tail well.

10.4.3.2 Results

Regression techniques based on generalized distributions that provide more flexibility in modeling skewed and fat-tailed data could be useful in such cases. The *GB2* regression model was fit for the medical care costs of office-based visits, and the estimation results are presented in Table 10.3. The location parameter is parameterized as a linear function of covariates. The t -ratios of regression coefficients show that most of the explanatory variables are statistically significant. The effects of some covariates are quite intuitive. For example, an elderly person is expected to have higher medical care costs, and a healthier person, indicated by either health conditions or self-assessed health, is expected to have lower medical costs. For the scale and shape parameters, standard errors are reported to facilitate some hypothesis tests. For instance, if one is interested in the null hypothesis $\psi_1 = 1$ (i.e., whether the *GB2* is significantly better than the Burr XII), a t -statistic, approximately calculated as $(5.883 - 1)/1.874 = 2.61$, could be used to reject the null hypothesis. In general, the estimates of the shape parameters demonstrate the value added by the extra complexity of the *GB2* in fitting heavy-tailed data.

Goodness-of-fit could be assessed by residual analysis. The left panel in Figure 10.5 presents fitted values versus residuals, where no particular pattern is identified. The right panel in Figure 10.5 exhibits the QQ -plot of residuals. The alignment along the 45-degree line suggests that the *GB2* distribution provides a favorable fit, especially for the observations in the right tail. To compare with GLMs, we also report in Table 10.3 the fits from a gamma model and an inverse-Gaussian model. Not surprisingly, we observe comparable estimates of regression coefficients between GLMs and the *GB2*. This is because a logarithmic link function is used for both models. Consistent with

Table 10.3. Comparison between GLM Fits and GB2 Regression

	Gamma		Inverse Gaussian		GB2	
	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
Intercept	5.302	42.182	5.237	23.794	3.794	14.592
age	0.007	4.857	0.008	3.080	0.008	5.176
female	0.415	14.226	0.459	8.567	0.494	15.597
married	0.171	4.779	0.175	2.606	0.177	4.646
hispanic	-0.208	-4.864	-0.216	-2.866	-0.119	-2.605
black	-0.227	-5.271	-0.211	-2.733	-0.264	-5.682
asian	-0.340	-6.041	-0.391	-4.246	-0.297	-4.920
edu	0.063	10.344	0.066	6.337	0.068	10.164
family size	-0.025	-2.204	-0.017	-0.834	-0.064	-5.263
msa	-0.009	-0.201	-0.108	-1.255	0.182	3.864
limitation	0.406	10.299	0.455	5.026	0.364	8.429
chronic	0.104	9.168	0.134	5.415	0.130	10.968
verygood	0.127	3.471	0.078	1.222	0.133	3.358
good	0.221	5.451	0.171	2.348	0.187	4.257
fair	0.411	6.819	0.468	3.631	0.369	5.647
poor	0.892	7.797	0.827	2.451	0.749	5.964
hmo	-0.035	-1.222	-0.013	-0.244	-0.011	-0.354
employed	-0.043	-1.061	-0.042	-0.554	-0.126	-2.848
Scale	Estimate	Std Error	Estimate	Std Error	Estimate	Std Error
	0.706	0.010	0.070	0.001	1.999	0.337
Shape1					5.883	1.874
Shape2					4.763	1.461
Log-likelihood	-56350.99		-56418.18		-55961.60	
AIC	112739.97		112874.35		111965.20	
BIC	112870.45		113004.83		112109.41	

the *QQ*-plots, the goodness-of-fit statistics – *AIC* or *BIC* – support the better fit of the *GB2* model.

10.5 Median Regression

For ratemaking purposes, actuaries are concerned with the correct estimate of the center of the cost distribution. Hence the current practice looks more to the mean regression model, which focuses on the expectation of the response, as does those discussed in previous sections. If one is interested in the center of a distribution, then the sensible question is, Why not use the median instead of the mean to measure the location?

Motivated by this question, this section introduces the median regression that is a distribution-free approach and emphasizes the relation between covariates and

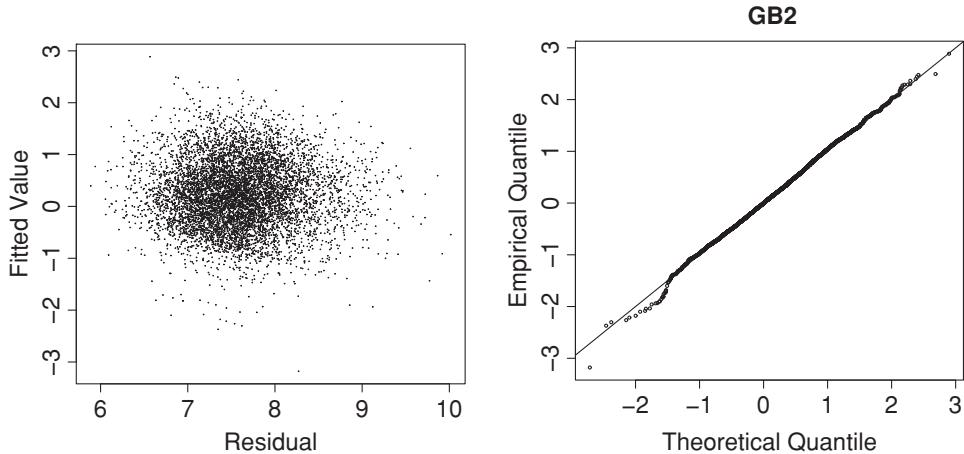


Fig. 10.5. Residual analysis of the *GB2* regression. The left panel depicts residuals versus fitted values, and the right panel is the *QQ*-plot of residuals.

the median of the response variable. The median regression is discussed from the perspectives of both the frequentist and Bayesian approaches. This section concerns the basics of quantile models that are linear in parameters.

10.5.1 Conditional Median

Consider the linear model

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i \quad (10.2)$$

where error ε_i is from an arbitrary distribution F_i . Recall that if one is willing to express the conditional mean of y_i given \mathbf{x}_i as $E[y_i|\mathbf{x}_i] = \mathbf{x}'_i \boldsymbol{\beta}$, then the ordinary least squares calculation suggests that $\boldsymbol{\beta}$ can be estimated by minimizing the sum of squared errors. Similarly, if we specify the conditional median of y_i given \mathbf{x}_i as $Median[y_i|\mathbf{x}_i] = \mathbf{x}'_i \boldsymbol{\beta}$, then parameter $\boldsymbol{\beta}$ can be estimated by solving

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n |y_i - \mathbf{x}'_i \boldsymbol{\beta}| \quad (10.3)$$

This model is known as median regression, where one estimates the conditional median of the response, rather than the conditional mean as in the classic regression model. Unlike least squares regression, the coefficient $\boldsymbol{\beta}$ in the median regression is estimated by minimizing the sum of absolute deviances of observed and fitted values, leading to the LAD (which stands for least absolute deviation) estimator defined in equation (10.3).

The rule of thumb for large samples applies to the regression median $\hat{\beta}$. Specifically, $\hat{\beta}$ is a consistent estimator and follows a normal distribution with mean β and asymptotic variance

$$Avar(\hat{\beta}) = \left(\sum 2f_i(0)\mathbf{x}_i\mathbf{x}'_i \right)^{-1} \left(\sum \mathbf{x}_i\mathbf{x}'_i \right) \left(\sum 2f_i(0)\mathbf{x}_i\mathbf{x}'_i \right)^{-1}.$$

Here f_i is the density function associated with distribution F_i . The approximate normality provides the basis for standard tests of statistical significance. When ε_i are i.i.d. error, one could further show $Avar(\hat{\beta}) = \omega^2(\mathbf{X}'\mathbf{X})^{-1}$, an analogy of the OLS (for ordinary least squares) estimator, where $\omega^2 = 1/(2f(0))^2$ and $\mathbf{X}'\mathbf{X} = \sum \mathbf{x}_i\mathbf{x}'_i$. See Bassett and Koenker (1978) for more details on the asymptotic properties of the LAD estimator.

The median regression is by itself distribution-free, and the LAD estimator does not require the distributional assumption on the error term. Compared with traditional models, median regression enjoys certain advantages in the analysis of fat-tailed data. First, the interpretation of parameters is somewhat simpler. Coefficients of mean regression can be interpreted as partial derivatives. However, as noted in Section 10.2, the situation is complicated in the case of transformation models. In contrast, there is no such interpretation difficulty for any monotone transformation in a quantile regression model. Second, the sensitivity of least square estimator to the outliers makes it a poor estimator for a wide class of non-Gaussian, especially fat-tailed, error models. Median regression is more robust than classical regressions in the sense that the conditional median is not as sensitive to outliers as the conditional mean in estimating the average location of a population.

10.5.2 Bayesian Approach

Bayesian inference has been very useful and attractive in linear models because it provides the entire posterior distribution of the parameter of interest and allows for the incorporation of parameter uncertainty into predictions. This section describes a simple Bayesian approach for median regression employing a likelihood function based on the Laplace distribution.

A random variable Y follows the Laplace distribution (denoted by $Laplace(\mu, \sigma)$) if its density is given by

$$f(y; \mu, \sigma) = \frac{1}{2\sigma} \exp \left\{ -\frac{|y - \mu|}{\sigma} \right\} \quad (10.4)$$

where $\sigma > 0$ is the scale parameter, and $-\infty < \mu < \infty$ is the location parameter. The asymmetric Laplace density is defined on the whole real line, making it a candidate to model both positive and negative responses.

One critical property of random variable Y defined by (10.4) that connects it to median regression is that μ turns out to be the median of Y . Consider the linear model (10.2) again. It is straightforward to show that when the error term follows i.i.d. Laplace distribution with zero location – i.e. $\varepsilon_i \sim \text{Laplace}(0, \sigma)$ – maximizing the associated log-likelihood function is equivalent to the minimization problem in (10.3) and thus yields the regression median $\hat{\beta}$.

In quantile regression, we are interested in the conditional quantile $\text{Median}(y_i | \mathbf{x}_i)$. The above facts motivate the specification of $\text{Laplace}(\mu_i, \sigma)$ as the sampling distribution of y_i with location parameter $\mu_i = \mathbf{x}'_i \beta$, regardless of the original distribution of the data. Given the observations $\mathbf{y} = (y_1, \dots, y_n)$, the posterior distribution of β is given by

$$f(\beta | \mathbf{y}) \propto L(\mathbf{y} | \beta) p(\beta)$$

where $p(\beta)$ denotes the prior distribution of β , and $L(\mathbf{y} | \beta)$ denotes the likelihood function that is based on the Laplace density, which could be expressed as

$$L(\mathbf{y} | \beta) = \frac{1}{(2\sigma)^n} \exp \left\{ -\frac{1}{\sigma} \sum_{i=1}^n |y_i - \mathbf{x}'_i \beta| \right\}.$$

Although a conjugate prior is not available for this quantile regression formulation, the MCMC methods are commonly used to derive approximations for the posterior distributions. In theory, any prior distribution could be used for unknown parameters. In practice, vague priors or improper uninformative priors are often used in the absence of a priori distributional knowledge. Note that Bayesian computation and regression models are discussed in Chapter 13 and 14, and all the theories therein also apply to the quantile regression context, those chapters provide detailed discussion on posterior sampling techniques and convergence analysis.

10.5.3 Quantile Regression

If the median represents a location measure of the distribution of the outcome, how about the 25th percentile, the 75th percentile, or any other quantiles of the response distribution? Extending the idea of regression median, Koenker and Bassett (1978) introduced a new class of statistics termed “regression quantiles.” Unlike the traditional regression that focuses on the center of the distribution $E(y | \mathbf{x})$, the quantile regression looks into the conditional quantiles of the response. Specifically, for a linear quantile function $Q_\tau(y_i | \mathbf{x}_i) = \mathbf{x}'_i \beta(\tau)$ with $0 < \tau < 1$, the regression coefficient $\beta(\tau)$ can be found by solving

$$\hat{\beta}(\tau) = \arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}'_i \beta), \quad (10.5)$$

where $\rho_\tau(u) = u(\tau - I(u \leq 0))$ is known as the check function and $\hat{\beta}(\tau)$ is called the τ th regression quantile. The optimization can be formulated as a nonlinear programming problem and is readily implemented using modern statistical packages.

Quantile regression is a natural extension of conditional mean regression to a basket of conditional quantile functions. Apparently the regression coefficient depends on the fraction τ , reflecting the partial effect of covariates on the τ th conditional quantile of the response. Apparently equation (10.5) reduces to (10.3) when $\tau = 0.5$. Thus the regression median is a special case of regression quantiles. See Koenker (2005) for a book-long review of quantile regression.

The Bayesian quantile regression could be implemented in a similar vein as median regression. Instead of Laplace distribution, one could use a three-parameter asymmetric Laplace density. The relation between quantile regression and the asymmetric Laplace distribution and a detailed discussion on Bayesian modeling are found in Yu and Moyeed (2001).

10.5.4 Medical Care Example: A Revisit

To illustrate quantile regression, let us revisit the example of office-based medical care costs. As shown in Section 10.4, the *GB2* distribution is an appropriate candidate for the heavy-tailed medical cost data. The favorable fit of the log-location-scale model motivates the quantile regression of the accelerated failure time model. Specifically, the conditional median of medical care costs is assumed to take the linear form:

$$\text{Median}(\log(y_i)|\mathbf{x}_i) = \mathbf{x}'_i \boldsymbol{\beta}(\tau).$$

Thus we fit a median regression on the logarithmic of total office-based medical care costs with the same set of covariates used in the *GB2* regression. The choice of the log transformation is dictated primarily by the desire to achieve linearity of parametric specification and its convenience of interpretation. The estimation results of regression coefficients are summarized in Table 10.4.

Next we adopt the Bayesian approach to quantile regression described in Section 10.5.2. To be consistent, the quantile regression is fitted for the logarithmic of medical care costs. Specifically, the sampling distribution is assumed to be Laplace, and vague priors are specified for the regression coefficients; that is, $\boldsymbol{\beta} \sim N(\mathbf{0}, 10^3 \text{ diag}(1, \dots, 1))$. The posterior Bayes estimates and 95% credible intervals are based on 500,000 iterations with first 250,000 discarded as burn-in. For comparison purposes, we also report the estimates of regression coefficients in Table 10.4.

Several observations can be made from these estimates. First, the statistical significance for all parameters from both approaches is consistent with one another. Second, all estimated coefficients from the classical method are within the Bayesian credible

Table 10.4. Estimation Results for Median Regression

	Classical		Bayesian		
	Est.	S.E.	Est.	LB	UB
intercept	3.938	0.159	3.841	3.681	3.999
age	0.010	0.002	0.011	0.008	0.014
female	0.568	0.038	0.572	0.517	0.629
married	0.152	0.048	0.154	0.068	0.218
hispanic	-0.150	0.048	-0.151	-0.227	-0.098
black	-0.307	0.063	-0.304	-0.401	-0.197
asian	-0.358	0.075	-0.394	-0.489	-0.294
edu	0.079	0.007	0.081	0.070	0.091
familysize	-0.056	0.015	-0.052	-0.073	-0.032
msa	0.204	0.057	0.230	0.131	0.314
limitation	0.345	0.056	0.332	0.245	0.421
chronic	0.145	0.015	0.144	0.122	0.168
verygood	0.086	0.046	0.097	0.031	0.164
good	0.150	0.054	0.154	0.075	0.220
fair	0.288	0.080	0.308	0.196	0.425
poor	0.726	0.186	0.760	0.565	0.946
hmo	-0.011	0.038	-0.017	-0.078	0.038
employed	-0.116	0.053	-0.107	-0.188	-0.028

intervals. These results show that the two approaches could be very consistent when no a priori information is incorporated into the Bayesian inference. Another comparison is between the *GB2* regression and the median regression. It is noticeable that the regression coefficients from the two models are of similar size. This relation could be seen from the following observation. In the former, the conditional mean function can be shown to be $E(\log(y_i)|\mathbf{x}_i) = \mathbf{x}'_i \boldsymbol{\beta} + \text{constant}$. In the latter, the conditional median function is $Q_{0.5}(\log(y_i)|\mathbf{x}_i) = \mathbf{x}'_i \boldsymbol{\beta}(0.5)$. In this application, the responses in log scale are more or less symmetric; thus its conditional median and mean are close, and so are the slope parameters in the two regressions.

10.6 Appendix A. Tail Measure

In general, to compare the (right) tails of random variables Y_1 and Y_2 , we look into the following quantity:

$$\lim_{y \rightarrow +\infty} \frac{\Pr(Y_1 > y)}{\Pr(Y_2 > y)} = \lim_{y \rightarrow +\infty} \frac{\bar{F}_{Y_1}(y)}{\bar{F}_{Y_2}(y)} = \lim_{y \rightarrow +\infty} \frac{f_{Y_1}(y)}{f_{Y_2}(y)}. \quad (10.6)$$

In the equation (10.6), $\bar{F}_Y(\cdot)$ and $f_Y(\cdot)$ denote the survival function and density function of random variable Y , respectively. The second equality is attained because

both density and survival functions approach zero as $y \rightarrow \infty$. A limiting value of zero of the ratio in (10.6) indicates that the distribution of Y_2 has a heavier tail than that of Y_1 .

For instance, to compare the right tail of a Weibull to a Pareto distribution, one could easily show that

$$\lim_{y \rightarrow +\infty} \frac{\bar{F}_{\text{Weibull}}(y)}{\bar{F}_{\text{Pareto}}(y)} = \lim_{y \rightarrow +\infty} \frac{\exp(-(y/\lambda)^\tau)}{\theta^\alpha(y+\theta)^{-\alpha}} = \lim_{y \rightarrow +\infty} \exp(-(y/\lambda)^\tau + \alpha \ln(y+\theta)) = 0,$$

suggesting that the Pareto has a heavier tail than the Weibull. As another example, the comparison between a Weibull and a gamma distribution could be performed by examining the ratio of their density functions:

$$\frac{f_{\text{gamma}}(y)}{f_{\text{Weibull}}(y)} = \frac{[\theta^\alpha \Gamma(\alpha)]^{-1} y^{\alpha-1} \exp(-y/\theta)}{\tau \lambda^{-\tau} y^{\tau-1} \exp(-(y/\lambda)^\tau)} \propto \exp((\alpha - \tau) \ln y - y/\theta + (y/\lambda)^\tau).$$

The above ratio shows that the shape of the Weibull distribution changes drastically with the value of τ , with $\tau < 1$ indicating a fatter tail and $\tau > 1$ indicating a thinner tail compared with the gamma distribution. In fact, the Weibull could have a thinner tail even than the normal distribution, since

$$\frac{f_{\text{Weibull}}(y)}{f_{\text{Normal}}(y)} \propto \exp((\tau - 1) \ln y + (y/\lambda)^\tau - (y - \mu)^2/(2\sigma)^2) \rightarrow 0$$

as $y \rightarrow +\infty$ when $\tau > 2$.

10.7 Appendix B. Information Matrix for GB2 Regression

The asymptotic covariance matrix of $\widehat{\Theta}$ defined in Section 10.4.1 could be derived from the Fisher information matrix $\text{Acov}(\widehat{\Theta}) = I^{-1}(\Theta)$, where

$$I(\Theta) = -E \left[\frac{\partial^2 \ln L(\Theta|\mathbf{y})}{\partial \Theta \partial \Theta'} \right].$$

The second-order derivatives of the log-likelihood function are

$$\begin{aligned} \frac{\partial^2 \ln L(\Theta|\mathbf{y})}{\partial \sigma^2} &= \frac{n}{\sigma^2} + \frac{2\phi_1}{\sigma^2} \sum_{i=1}^n z_i - \frac{2(\phi_1 + \phi_2)}{\sigma^2} \sum_{i=1}^n \frac{z_i \exp(z_i)}{1 + \exp(z_i)} \\ &\quad - \frac{\phi_1 + \phi_2}{\sigma^2} \sum_{i=1}^n \frac{z_i^2 \exp(z_i)}{[1 + \exp(z_i)]^2} \end{aligned}$$

$$\frac{\partial^2 \ln L(\Theta|\mathbf{y})}{\partial \phi_1^2} = n[\Psi'(\phi_1) - \Psi'(\phi_1 + \phi_2)]$$

$$\begin{aligned}
\frac{\partial^2 \ln L(\Theta|\mathbf{y})}{\partial \phi_2^2} &= n[\Psi'(\phi_2) - \Psi'(\phi_1 + \phi_2)] \\
\frac{\partial^2 \ln L(\Theta|\mathbf{y})}{\partial \beta_j \beta_k} &= -\frac{\phi_1 + \phi_2}{\sigma^2} \sum_{i=1}^n x_{ij} x_{ik} \frac{\exp(z_i)}{[1 + \exp(z_i)]^2} \\
\frac{\partial^2 \ln L(\Theta|\mathbf{y})}{\partial \sigma \phi_1} &= -\frac{1}{\sigma} \sum_{i=1}^n z_i + \frac{1}{\sigma} \sum_{i=1}^n \frac{z_i \exp(z_i)}{1 + \exp(z_i)} \\
\frac{\partial^2 \ln L(\Theta|\mathbf{y})}{\partial \sigma \phi_2} &= \frac{1}{\sigma} \sum_{i=1}^n \frac{z_i \exp(z_i)}{1 + \exp(z_i)} \\
\frac{\partial^2 \ln L(\Theta|\mathbf{y})}{\partial \sigma \beta_j} &= \frac{\phi_1}{\sigma^2} \sum_{i=1}^n x_{ij} - \frac{\phi_1 + \phi_2}{\sigma^2} \sum_{i=1}^n x_{ij} \frac{\exp(z_i)}{1 + \exp(z_i)} \\
&\quad - \frac{\phi_1 + \phi_2}{\sigma^2} \sum_{i=1}^n x_{ij} \frac{z_i \exp(z_i)}{[1 + \exp(z_i)]^2} \\
\frac{\partial^2 \ln L(\Theta|\mathbf{y})}{\partial \phi_1 \phi_2} &= -n\Psi'(\phi_1 + \phi_2) \\
\frac{\partial^2 \ln L(\Theta|\mathbf{y})}{\partial \phi_1 \beta_j} &= -\frac{1}{\sigma} \sum_{i=1}^n x_{ij} + \frac{1}{\sigma} \sum_{i=1}^n x_{ij} \frac{\exp(z_i)}{1 + \exp(z_i)} \\
\frac{\partial^2 \ln L(\Theta|\mathbf{y})}{\partial \phi_2 \beta_j} &= \frac{1}{\sigma} \sum_{i=1}^n x_{ij} \frac{\exp(z_i)}{1 + \exp(z_i)}
\end{aligned}$$

So we obtain the elements of the Fisher information matrix:

$$\begin{aligned}
I_{1,1}(\Theta) &= -\frac{n}{\sigma^2} - \frac{2n\phi_1}{\sigma^2} [\Psi(\phi_1) - \Psi(\phi_2)] + \frac{2n(\phi_1 + \phi_2)^2}{\phi_1 \sigma^2} [\Psi(\phi_1 + 1) - \Psi(\phi_2)] \\
&\quad + \frac{n(\phi_1 + \phi_2)^2(\phi_1 + \phi_2 - 1)}{\phi_1 \phi_2 \sigma^2} \\
&\quad \times \{[\Psi(\phi_1 + 1) - \Psi(\phi_2 + 1)]^2 + \Psi'(\phi_1 + 1) + \Psi'(\phi_2 + 1)\} \\
I_{2,2}(\Theta) &= n[\Psi'(\phi_1 + \phi_2) - \Psi'(\phi_1)] \\
I_{3,3}(\Theta) &= n[\Psi'(\phi_1 + \phi_2) - \Psi'(\phi_2)] \\
I_{3+j,3+k}(\Theta) &= \frac{(\phi_1 + \phi_2)^2(\phi_1 + \phi_2 - 1)}{\phi_1 \phi_2 \sigma^2} \sum_{i=1}^n x_{ij} x_{ik}, \quad j, k = 1, \dots, p
\end{aligned}$$

$$\begin{aligned}
I_{1,2}(\Theta) &= \frac{n}{\sigma} [\Psi(\phi_1) - \Psi(\phi_2)] - \frac{n(\phi_1 + \phi_2)}{\phi_1 \sigma} [\Psi(\phi_1 + 1) - \Psi(\phi_2)] \\
I_{1,3}(\Theta) &= -\frac{n(\phi_1 + \phi_2)}{\phi_1 \sigma} [\Psi(\phi_1 + 1) - \Psi(\phi_2)] \\
I_{1,3+j}(\Theta) &= -\frac{\phi_1}{\sigma^2} \sum_{i=1}^n x_{ij} + \frac{(\phi_1 + \phi_2)^2}{\phi_1 \sigma^2} \sum_{i=1}^n x_{ij} + \frac{(\phi_1 + \phi_2)^2(\phi_1 + \phi_2 - 1)}{\phi_1 \phi_2 \sigma^2} \\
&\quad \times [\Psi(\phi_1 + 1) - \Psi(\phi_2 + 1)] \sum_{i=1}^n x_{ij}, \quad j = 1, \dots, p \\
I_{2,3}(\Theta) &= n \Psi'(\phi_1 + \phi_2) \\
I_{2,3+j}(\Theta) &= -\frac{\phi_2}{\phi_1 \sigma} \sum_{i=1}^n x_{ij}, \quad j = 1, \dots, p \\
I_{3,3+j}(\Theta) &= -\frac{1}{\sigma} \left(1 + \frac{\phi_2}{\phi_1}\right) \sum_{i=1}^n x_{ij}, \quad j = 1, \dots, p
\end{aligned}$$

References

- Bartlett, M. (1947). The use of transformations. *Biometrics* 3(1), 39–52.
- Bassett, G., Jr. and R. Koenker (1978). Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association* 73(363), 618–622.
- Beirlant, J., Y. Goegebeur, R. Verlaak, and P. Vynckier (1998). Burr regression and portfolio segmentation. *Insurance: Mathematics and Economics* 23(3), 231–250.
- Bickel, P. and K. Doksum (1981). An analysis of transformations revisited. *Journal of the American Statistical Association* 76(374), 296–311.
- Box, G. and D. Cox (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)* 26(2), 211–252.
- Burbidge, J., L. Magee, and A. Robb (1988). Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association* 83(401), 123–127.
- Frees, E., J. Gao, and M. Rosenberg (2011). Predicting the frequency and amount of health care expenditures. *North American Actuarial Journal* 15(3), 377–392.
- Frees, E., P. Shi, and E. Valdez (2009). Actuarial applications of a hierarchical claims model. *ASTIN Bulletin* 39(1), 165–197.
- Frees, E. and E. Valdez (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association* 103(484), 1457–1469.
- Frees, E. and P. Wang (2005). Credibility using copulas. *North American Actuarial Journal* 9(2), 31–48.
- John, J. and N. Draper (1980). An alternative family of transformations. *Applied Statistics* 29(2), 190–197.
- Klugman, S., H. Panjer, and G. Willmot (2008). *Loss Models: From Data to Decisions* (3rd ed.). Wiley.
- Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, Cambridge.

- Koenker, R. and G. Bassett Jr (1978). Regression quantiles. *Econometrica* 46(1), 33–50.
- Manning, W., A. Basu, and J. Mullahy (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics* 24(3), 465–488.
- McDonald, J. (1984). Some generalized functions for the size distribution of income. *Econometrica* 52(3), 647–63.
- McDonald, J. and R. Butler (1990). Regression models for positive random variables. *Journal of Econometrics* 43(1–2), 227–251.
- Panjer, H. and G. Willmot (1992). *Insurance Risk Models*. Society of Acturaries.
- Sun, J., E. Frees, and M. Rosenberg (2008). Heavy-tailed longitudinal data modeling using copulas. *Insurance Mathematics and Economics* 42(2), 817–830.
- Yeo, I. and R. Johnson (2000). A new family of power transformations to improve normality or symmetry. *Biometrika* 87(4), 954–959.
- Yu, K. and R. Moyeed (2001). Bayesian quantile regression. *Statistics & Probability Letters* 54(4), 437–447.

11

Spatial Modeling

Eike Brechmann and Claudia Czado

Chapter Preview. This chapter presents statistical models that can handle spatial dependence among variables. Spatial dependence refers to the phenomenon that variables observed in areas close to each other are often related. Ignoring data heterogeneity due to such spatial dependence patterns may cause overdispersion and erroneous conclusions. In an actuarial context, it is important to take spatial information into account in many cases, such as in the insurance of buildings threatened by natural catastrophes; in health insurance, where diseases affect specific regions; and also in car insurance, as we discuss in an application.

In particular, we describe the most common spatial autoregressive models and show how to build a joint model for claim severity and claim frequency of individual policies based on generalized linear models with underlying spatial dependence. The results show the importance of explicitly considering spatial information in the ratemaking methodology.

11.1 Introduction

It is important to take spatial information related to insurance policies into account when predicting claims and ratemaking. The most prominent example is the modeling of natural catastrophes needed for the insurance of buildings. Another example is health insurance, where spatial information is relevant for an accurate assessment of the underlying risks, because frequencies of some diseases may vary by region. Frequencies in neighbor regions are often expected to be more closely related than those in regions far from each other. This phenomenon is usually referred to as spatial dependence.

Statistically speaking, spatial dependence means that the joint behavior of variables in different regions depends on some kind of underlying distance, the simplest example being the Euclidean distance between two points. If variables in different regions do not behave independently, we speak of a spatial dependence pattern. This pattern

may be the reason for unobserved data heterogeneity, which causes overdispersion. Overdispersion denotes the situation when a higher variance is observed than assumed by the model. In this chapter we demonstrate how such spatial dependence patterns of regions can be accounted for appropriately in an actuarial context. By precisely and flexibly specifying a neighborhood structure of regions, the presented methods go beyond the classical use of territory variables as categorical covariates in regression models, which do not represent explicit territorial characteristics, so that, for example, topographic effects are ignored. Detailed statistical treatments of spatial modeling can be found in the seminal book by Cressie (1993) and also in newer volumes such as Banerjee, Carlin, and Gelfand (2003), Gelfand et al. (2010), and Cressie and Wikle (2011).

To make things more concrete: What exactly do we mean when we talk about *spatial data*? For this, we come back to the earlier mentioned examples. First, when a natural catastrophe such as a flood occurs, it usually does not affect isolated buildings. Therefore, claims due to a natural catastrophe for a portfolio of building insurance contracts are generally not independent. For example, they might depend on the distance of the building to the next river. The statistical challenge here is in predicting flood levels at any location in an area of interest. Then, claim frequency and size can easily be predicted for any building in the portfolio. For obvious reasons, such data are called *point-level* or *point-referenced data*.

Second, the exact location of the occurrence of a hurricane, a wildfire, or an earthquake is usually unknown and can be considered as random. In this case, the set of locations, at which events occur, is random itself. The observed data then simply are equal to 1 for all locations in this set and are called *point-process data*. In addition, the data can be supplemented by covariate information, such as the damage caused by the events (*marked point-process data*).

Third, the frequency of a disease can hardly be assessed at each single location in a region. In such cases, one therefore partitions the area of interest into a certain number of areal locations, such as cells of a grid or geographical regions, which are often directly available because of the type of data. As an example, one may think of skin cancer rates observed in all 50 U.S. states. The quantity of interest is then the sum or average of variables in each areal location. Such data are called *areal* or *lattice data* and are the focus of this chapter. For the sake of simplicity, we refer to areal locations as “regions” for the remainder of this chapter. More information on the other common types of spatial data and statistical models for them can be found in the earlier mentioned references (see, e.g., Cressie 1993, chapter 1, and Banerjee et al. 2003, chapter 1).

A characteristic feature of areal data is its neighborhood structure. This means that for each region we can identify a set of neighbor regions, where the notion of a neighbor can have different meanings. Most often, neighbors are those regions that

share a common border, but the set of neighbors may, for instance, also be defined as the regions within a certain distance of the region of interest.

As an illustrative example, we analyze the full comprehensive car insurance portfolio of a German car insurance company in the year 2000. The dataset includes personal information such as age and gender of the policyholders as well as policy-specific details such as distance driven per year, deductible, and type and age of the car. As in Gschlößl and Czado (2007), we focus on a subset of the data with claims from traffic accidents of policyholders with three models of mid-sized cars. This subset of the data still contains more than 350,000 observations.

According to where policyholders live, they can be attributed to 440 regions in Germany. Clearly, we expect a certain degree of heterogeneity in claim sizes and frequency among the regions. The question that we investigate here is whether there is a spatial pattern in this heterogeneity, which is important to take into account to ensure accurate ratemaking. The left panel of Figure 11.1 shows the average claim size per region. This gives a first indication for spatial dependence: in eastern Germany claim sizes appear to be larger on average than in the middle and in the South-West of Germany. Because some of these effects are possibly due to the population density of the regions, we discuss ways to adjust for this factor.

The average claim size per region is analyzed in a first step in Section 11.4. In Section 11.5 we then describe a more elaborate model for the total loss per region, which combines separate generalized linear models (GLMs) for frequency and severity and accounts for spatial dependence in the data. The exploratory analysis of spatial data is treated in Section 11.2, whereas while appropriate spatial dependence models are discussed in Section 11.3.

11.2 Exploratory Analysis of Spatial Data

In this section we only briefly touch on the exploratory analysis of spatial data. Detailed discussions can be found in any book on spatial statistics (see, in particular, Cressie 1993, and Banerjee et al. 2003).

Let $\mathbf{Y} = (Y_1, \dots, Y_n)'$ be a vector of random variables observed in n different regions. One of the first steps – and at the same time one of the most instructive ones – in such an exploratory analysis is the graphical illustration of the data Y_i , $i = 1, \dots, n$, as in the left panel of Figure 11.1. Here, observations in each regions are color coded so that it is easy to assess whether there is a spatial dependence pattern with similar values in close regions.

Analytically, spatial association can be measured, for example, using Moran's I or Geary's C , which can be regarded as spatial equivalents of common tools in time series analysis (the autocorrelation coefficient and the Durbin-Watson statistic, respectively). To define the two quantities, we need to introduce the concept of the *adjacency matrix*,

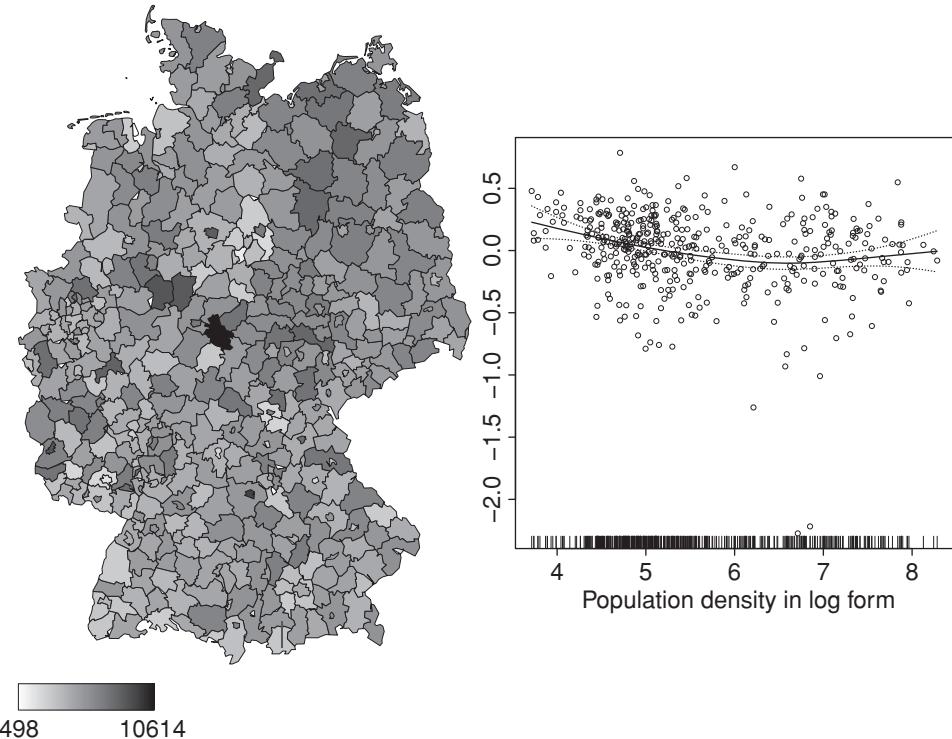


Fig. 11.1. Left panel: Average claim size per region in Germany. The minimum average claim size is 498.17, and the maximum is 10 613.97. Right panel: Fitted third-order cubic smoothing spline of the population density in log form as covariate for average claim sizes also in log form. Dotted lines are pointwise estimated twice-standard-error curves (assuming independent and identically distributed data, and hence ignoring potential spatial dependence).

which encodes the neighborhood structure of \mathbf{Y} . By $N(i)$ we denote the neighbors of variable Y_i , $i = 1, \dots, n$; then the adjacency matrix $W = (w_{ij}) \in \mathbb{R}^{n \times n}$ is defined by

$$w_{ij} = \begin{cases} 1, & j \in N(i) \\ 0, & j \notin N(i) \end{cases}. \quad (11.1)$$

In other words, the entries in row i indicate the neighbors of region i .

Then, Moran's I is given by

$$\rho_I = \frac{n}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i=1}^n (Y_i - \bar{Y})^2},$$

where $\bar{Y} = 1/n \sum_{i=1}^n Y_i$. Under spatial independence, the expectation of Moran's I is $-1/(n-1)$. Positive (negative) values of ρ_I indicate positive (negative) spatial

association. Moran's I is, however, not restricted to $[-1, 1]$, although values usually fall into this range.

Alternatively, Geary's C is available as

$$\rho_C = \frac{n - 1}{2 \sum_{i=1}^n \sum_{j=1}^n w_{ij}} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}(Y_i - Y_j)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \in [0, 2].$$

Contrary to Moran's I , the expectation of Geary's C under spatial independence is 1, and values of ρ_C below (above) 1 indicate positive (negative) spatial association. For both measures, different weight matrices can also be used (e.g., depending on the distance between regions). The adjacency matrix (11.1) is the most typical choice.

The Moran's I and Geary's C calculated for the average claim size data are 0.105 and 0.658, respectively. Estimated standard errors of 0.029 and 0.082, respectively, indicate moderate but significantly positive spatial dependence among regions, as expected from the observed patterns in the left panel of Figure 11.1. For the calculation of ρ_I and ρ_C and the associated standard errors, we used the functions `moran.test` and `geary.test` in the R-package `spdep` by Bivand et al. (2013).

When covariate information is available for the quantity of interest, standard techniques from regression analysis may be used to get some indication of regression effects (see preceding chapters of this volume). However, one has to be careful here, because spatial data are not i.i.d. (except for the boundary case of spatial independence), so that the standard asymptotic theory does not hold and estimated standard errors cannot be trusted.

In our average claim size data, an important covariate to distinguish between rural and urban regions is the population density of a region. To investigate whether there is an effect of population density on the average claim sizes, we fit a generalized additive model (GAM) with a third-order cubic smoothing spline of the population density in log form as a covariate (see Chapter 15). The resulting fit is shown in the right panel of Figure 11.1. The regression curve indicates that there are higher average claim sizes in rural regions, while average claim sizes are smaller in medium-sized to larger cities. The wider pointwise estimated twice-standard-error curves in the right tail of the fit show that the increase observed there may not be significant (assuming independent and identically distributed data). Although a quadratic effect of the population density may be sufficient, a more general model allowing for a cubic effect is considered in the following due to the uncertainty in the estimates of the standard errors.

Finally, note that the spatial dependence in the average claim size data cannot fully be explained through the covariate information: the empirical Moran's I and Geary's C of the residuals of the GAM fit are 0.079 (0.029) and 0.672 (0.080), respectively, where estimated standard errors are given in brackets.

11.3 Spatial Autoregression

The aim of this section is to present appropriate models that can account for spatial dependence among the components of $\mathbf{Y} = (Y_1, \dots, Y_n)'$. In other words, we show how to specify a joint distribution for the random vector \mathbf{Y} .

We discuss the commonly used conditionally and simultaneously autoregressive models. In both models, the joint distribution is a multivariate normal with appropriate covariance structure. The corresponding joint density is denoted by $p(\mathbf{y})$, where $\mathbf{y} = (y_1, \dots, y_n)'$. More details on the two models and on extensions to non-normal spatial data can be found, for example, in Cressie (1993) and Banerjee et al. (2003).

11.3.1 Conditionally Autoregressive Models

A typical feature of spatial data is that the degree of spatial dependence decreases with increasing distance between regions, the simplest case being that the variable Y_i is only influenced by its neighbors $N(i)$. In this case it is natural to work with the so-called full conditional distributions of Y_i , $i = 1, \dots, n$, given all other variables,

$$p(y_i | \mathbf{y}_{-i}) = \frac{p(\mathbf{y})}{p(\mathbf{y}_{-i})},$$

where \mathbf{y}_{-i} is the vector \mathbf{y} with the i th observation removed; that is, $\mathbf{y}_{-i} = (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)'$. We assume that the conditional distribution of Y_i given all other variables depends only on the neighbors $N(i)$ of Y_i ; that is,

$$p(y_i | \mathbf{y}_{-i}) = p(y_i | y_j \in N(i)), \quad i = 1, \dots, n.$$

Especially when the number of regions n is large, this assumption significantly reduces the complexity compared to working with the n -dimensional joint distribution directly. Due to the Markov property known from time series analysis (future values depend only on the present value and not on past values), which is here induced by the neighborhood structure, the random vector \mathbf{Y} is then called a *Markov random field* (see, e.g., Rue and Held 2005, for more details).

The question is, however, whether the collection of conditional distributions $p(y_i | \mathbf{y}_{-i})$, $i = 1, \dots, n$, uniquely identifies the joint distribution $p(\mathbf{y})$ for all \mathbf{y} ; that is, whether the full conditionals are *compatible*. Although the converse is always true, particular conditions have to be met to ensure compatibility. One such case is discussed here.

Definition 11.1 (Conditionally autoregressive model). In the conditionally autoregressive (CAR) model by Besag (1974), the full conditional distribution of $Y_i | \{y_j, j \in N(i)\}$, $i = 1, \dots, n$, is given by a normal distribution with mean

$\mu_i + \sum_{j \in N(i)} c_{ij}(y_j - \mu_j)$ and variance σ_i^2 ,

$$Y_i | \{y_j, j \in N(i)\} \sim \mathcal{N}\left(\mu_i + \sum_{j \in N(i)} c_{ij}(y_j - \mu_j), \sigma_i^2\right),$$

where $\mu_i \in \mathbb{R}$ is the mean parameter, $\sigma_i > 0$ is the conditional standard deviation, and $c_{ij} \in \mathbb{R}$ specifies the degree of dependence between regions i and j .

The CAR model of Definition 11.1 can be thought of as a regression of Y_i on y_j , $j \in N(i)$, observed in the neighbor regions, where the influence of each y_j is directly related to the spatial relationship of regions i and j . For instance, if c_{ij} is constant for all $j \in N(i)$, then the influence of each neighbor is the same. Alternatively, c_{ij} may be a function of the distance between the centers of regions i and j , such that c_{ij} is decreasing with increasing distance, because we usually expect a stronger influence for closer regions. The most common example is a normalized version of the adjacency matrix (11.1), which we discuss later. Before we do so, we return to the question whether the full conditionals are compatible.

Let M_σ be a diagonal matrix with entries $(M_\sigma)_{ii} = \sigma_i^2$, $i = 1, \dots, n$, and $C = (c_{ij}) \in \mathbb{R}^{n \times n}$ be a so-called *proximity* or *neighborhood matrix*, with $c_{ii} = 0$, $c_{ij} = 0$ for all $j \notin N(i)$ and $c_{ij}\sigma_j^2 = c_{ji}\sigma_i^2$ for $i, j = 1, \dots, n$. Then, the Hammersley-Clifford theorem and Brook's lemma (see, e.g., Banerjee et al. 2003) can be used to show that, for the CAR model of Definition 11.1, it holds that Y is jointly normal, in particular

$$\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, (I - C)^{-1} M_\sigma), \quad (11.2)$$

where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)'$ and I is the identity matrix. Here, the condition $c_{ij}\sigma_j^2 = c_{ji}\sigma_i^2$ ensures that $(I - C)^{-1} M_\sigma$ is symmetric. For M_σ a typical choice is $M_\sigma = \sigma^2 M$, where $M = (m_{ij}) \in \mathbb{R}^{n \times n}$ is a diagonal weight matrix; that is, $m_{ij} = 0$ for $i, j = 1, \dots, n$, $i \neq j$. The role of M_σ can be compared to that of the weight matrix in weighted linear regression.

11.3.1.1 Intrinsically Autoregressive Model

The most common CAR specification is the intrinsically autoregressive (IAR) model by Besag, York, and Mollié (1991). For this, we set $m_{ii} = n_i^{-1}$, $i = 1, \dots, n$, where $n_i := |N(i)|$ denotes the number of neighbors of region i , and $C = MW$, where W is the adjacency matrix (11.1) of the regions. The proximity matrix C is thus the *normalized adjacency matrix*. It holds that

$$M_\sigma^{-1}(I - C) = \frac{1}{\sigma^2} M^{-1}(I - MW) = \frac{1}{\sigma^2}(M^{-1} - W).$$

Further, $(M^{-1} - W)$ is singular, since

$$(M^{-1} - W)\mathbf{1} = \mathbf{0},$$

where $\mathbf{0} = (0, \dots, 0)'$ and $\mathbf{1} = (1, \dots, 1)'$. Hence, the joint distribution of \mathbf{Y} is not proper. Nevertheless, the IAR model can still be used for conditionally specified random effects in a hierarchical model specification. The model discussed in Section 11.5 provides an example of such improper prior distributions yielding a proper posterior distribution. In general, the availability of the full conditionals makes CAR models very attractive in the context of Markov chain Monte Carlo (MCMC) methods, as presented in Chapter 13.

One convenient way to remedy this shortcoming of the IAR model is by adding a parameter $\phi \in [0, 1]$ to the specification of the spatial dependence matrix C ; that is, we set $C = \phi M W$ (see, e.g., Sun, Tsutakawa, and Speckman, 1999). If $\phi = 1$, this is the IAR model. If, however, $\phi < 1$, then $(M^{-1} - \phi W)$ is guaranteed to be nonsingular and we obtain a proper distribution. Although the parameter ϕ should not be confused with a correlation parameter, it controls the degree of spatial dependence between the regions. In particular, the case $\phi = 0$ implies independence among the regions.

11.3.2 Simultaneously Autoregressive Models

Using expression (11.2), the CAR model implies $(I - C)\mathbf{Y} \sim \mathcal{N}_n((I - C)\boldsymbol{\mu}, M_\sigma(I - C)')$, and thus the corresponding error formulation holds:

$$\boldsymbol{\varepsilon}_{\text{CAR}} := (I - C)(\mathbf{Y} - \boldsymbol{\mu}) \sim \mathcal{N}_n(\mathbf{0}, M_\sigma(I - C)'). \quad (11.3)$$

This implies that the errors $\boldsymbol{\varepsilon}_{\text{CAR}}$ are not independent. Expression (11.3) therefore motivates the consideration of an alternative model, which can be considered as the spatial equivalent of autoregressive models in time series analysis.

Definition 11.2 (Simultaneously autoregressive model). The simultaneously autoregressive (SAR) model is defined as

$$\boldsymbol{\varepsilon}_{\text{SAR}} := (I - B)(\mathbf{Y} - \boldsymbol{\mu}) \sim \mathcal{N}_n(\mathbf{0}, D_\tau), \quad (11.4)$$

where $\boldsymbol{\tau} = (\tau_1, \dots, \tau_n)'$, D_τ is diagonal with entries $(D_\tau)_{ii} = \tau_i^2$, $i = 1, \dots, n$, and $B = (b_{ij}) \in \mathbb{R}^{n \times n}$ with $b_{ii} = 0$, $i = 1, \dots, n$. Here, in contrast to $\boldsymbol{\varepsilon}_{\text{CAR}}$ in expression (11.3), the components of $\boldsymbol{\varepsilon}_{\text{SAR}}$ are independent. As with the proximity matrix C in the CAR model, we assume $b_{ij} = 0$ if $j \notin N(i)$.

Similar to the CAR model, the SAR model can be interpreted as a regression of Y_i on y_j , $j \in N(i)$, where the impact of the different neighbors is given through the entries of the matrix B . More precisely, we may rewrite model formulation (11.4) as

$$\mathbf{Y} = \boldsymbol{\mu} + B(\mathbf{Y} - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}_{\text{SAR}}.$$

Hence, the expectation of Y_i given its neighbors is

$$E(Y_i | \{y_j, j \in N(i)\}) = \mu_i + \sum_{j \in N(i)} b_{ij}(y_j - \mu_j), \quad i = 1, \dots, n.$$

Another reformulation of (11.4) is

$$\mathbf{Y} = B\mathbf{Y} + (I - B)\boldsymbol{\mu} + \boldsymbol{\varepsilon}_{\text{SAR}},$$

which shows that \mathbf{Y} may be regarded as a spatial weighting of the neighbors and the mean vector. If B is the zero matrix, then no spatial information is used.

Since the role of the matrix B in the SAR model is similar to that of the proximity matrix C in the CAR model, B is also often referred to as a proximity or neighborhood matrix. Although both matrices can be chosen to be the same, other choices may be reasonable when it comes to modeling. This becomes clear when we assume that both B and C are symmetric (which they often are but do not necessarily have to be). Due to the symmetry condition $c_{ij}\sigma_j^2 = c_{ji}\sigma_i^2$ in the CAR model, conditional variances then have to be equal among regions, $\sigma_j^2 = \sigma_i^2$. For the SAR model no such restrictions hold.

Model formulations (11.3) and (11.4) for the CAR and the SAR model, respectively, motivate the notion of autoregressive models, which are common in time series analysis. In particular, although errors in the CAR model formulation (11.3) are not independent, the formulation of the SAR model in (11.4) imitates the usual autoregressive time series modeling by assuming independent errors. It follows immediately that in the SAR model the joint distribution is given by

$$\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, (I - B)^{-1} D_\tau (I - B')^{-1}).$$

This is a proper distribution if $(I - B)$ has full rank and parameters can be conveniently estimated by maximum likelihood.

As noted in Section 11.2, it is often of interest to investigate the influence of covariates on the quantity of interest. A mean specification depending on covariates can easily be introduced in both the CAR and the SAR model by setting

$$\mu_i = \mathbf{x}'_i \boldsymbol{\beta}, \quad i = 1, \dots, n,$$

where $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ik})'$ is a vector of covariates (including an intercept) observed in region i and $\boldsymbol{\beta} = (\beta_0, \dots, \beta_k)'$ is a vector of fixed regression parameters. This reduces the number of parameters in the mean specification from n to $k + 1$. If an intercept model is chosen – that is, $\mu_i = \beta_0$ for all $i = 1, \dots, n$ – we have only one parameter.

Finally a technical remark: Clearly the CAR and the SAR model are closely related (see Table 11.1 for an overview). Most importantly, however, the proximity matrices B and C influence the covariance structure of the models differently. They are equivalent

Table 11.1. Summary of the Main Model Components of the CAR and the SAR Model

Model	$E(Y_i \{y_j, j \in N(i)\})$	$Var(\mathbf{Y})$	$Var(\boldsymbol{\varepsilon}_{\cdot AR})$
CAR	$\mu_i + \sum_{j \in N(i)} c_{ij} (y_j - \mu_j)$	$(I - C)^{-1} M_\sigma$	$M(I - C)'$
SAR	$\mu_i + \sum_{j \in N(i)} b_{ij} (y_j - \mu_j)$	$(I - B)^{-1} D_\tau (I - B')^{-1}$	D_τ

if and only if

$$(I - C)^{-1} M_\sigma = (I - B)^{-1} D_\tau (I - B')^{-1}.$$

Cressie (1993) shows that any SAR model can be represented as a CAR model, whereas the converse is not generally true. Resuming the discussion of the proximity matrices B and C , it is clear that B and C cannot have the same interpretation if the CAR and the SAR model coincide. For further insights and comparisons between SAR and CAR models, we strongly recommend the work by Wall (2004), who points out potential non-intuitive results implied by either models using an illustrative example.

11.4 Average Claim Size Modeling

The previous section discusses two different approaches to specifying spatial autoregressive models. These are now used to analyze the average claim sizes of the German car insurance portfolio presented in the introduction (see the left panel of Figure 11.1). Since both models require normality of the variables of interest, we take average claim sizes in log form, which also ensures that there are no issues with negative values. Hence, we model $Y_i := \log S_i$, $i = 1, \dots, n$, where S_i denotes the average claim size in region i and $n = 440$.

The following specifications of the CAR and the SAR model are chosen in a very similar way:

- (1) *CAR model:* $M_\sigma = \sigma^2 M$, where $M = (m_{ij}) \in \mathbb{R}^{n \times n}$ is diagonal with $m_{ii} = n_i^{-1}$, $i = 1, \dots, n$, and $C = \phi M W$ with $\phi \in [0, 1]$.
- (2) *SAR model:* $D_\tau = \tau^2 D$, where $D = (d_{ij}) \in \mathbb{R}^{n \times n}$ is diagonal with $d_{ii} = n_i^{-1}$, $i = 1, \dots, n$, and $B = \gamma D W$ with $\gamma \in [0, 1]$.

Here, we use the scaled normalized adjacency matrix (see (11.1)) as a proximity matrix in both models. In particular, $\phi = 0$ ($\gamma = 0$) in the CAR (SAR) model corresponds to spatial independence, since then the proximity matrix C (B) is a matrix of zeros.

For the mean, we investigated different specifications: first, an intercept model with $\mu_i = \beta_0$, $i = 1, \dots, n$, and, second, based on the exploratory analysis in the right panel of Figure 11.1, covariate models with intercept β_0 and orthogonal polynomials

Table 11.2. Estimated Parameters and Standard Errors of Linear, CAR, and SAR Models for Average Claim Sizes with Intercept and Covariate Models for the Mean

LM	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$		Log lik.	AIC	BIC
Inter.	8.484 (0.015)					-119.8	243.6	251.8
Lin.	8.484 (0.015)	-1.373 (0.312)				-110.3	226.5	238.8
Quad.	8.484 (0.015)	-1.373 (0.308)	1.011 (0.308)			-104.9	217.8	234.2
Cubic	8.484 (0.015)	-1.373 (0.308)	1.011 (0.308)	0.125 (0.308)		-104.8	219.7	240.1
CAR	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\phi}$	Log lik.	AIC	BIC
Inter.	8.468 (0.021)				0.463 (0.126)	-114.0	234.1	246.4
Lin.	8.484 (0.015)	-1.421 (0.311)			0.001 (0.018)	-110.3	228.5	244.9
Quad.	8.484 (0.015)	-1.415 (0.307)	1.009 (0.307)		0.001 (0.006)	-104.9	219.8	240.3
Cubic	8.484 (0.015)	-1.407 (0.307)	1.010 (0.307)	0.142 (0.308)	0.001 (0.012)	-104.8	221.7	246.2
SAR	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\gamma}$	Log lik.	AIC	BIC
Inter.	8.479 (0.019)				0.220 (0.067)	-114.7	235.4	247.7
Lin.	8.483 (0.018)	-1.394 (0.327)			0.208 (0.068)	-105.8	219.6	236.0
Quad.	8.483 (0.018)	-1.420 (0.323)	0.993 (0.323)		0.196 (0.070)	-101.1	212.3	232.7
Cubic	8.483 (0.018)	-1.411 (0.323)	0.977 (0.324)	0.178 (0.312)	0.197 (0.070)	-101.0	214.0	238.5

Note: The last columns show log likelihoods, AICs, and BICs of the estimated models.

up to order 3 of the population density in log form for each region i . Associated regression parameters are denoted by β_1 , β_2 , and β_3 .

The models are fitted using the function `spautolm` in the R-package `spdep` by Bivand et al. (2013), and estimated parameters and standard errors are reported in Table 11.2, which also shows fits of linear models (LMs) with corresponding mean specifications.

The estimation results underline the differences between both models. Although estimates for the intercept and for the influence of the population density are very similar across both models, there are striking differences in the interpretation of

the degree of spatial dependence. In the SAR model, the inclusion of the spatial information regarding the population density of each region leads to a decrease of up to 11% in the estimated parameter $\hat{\gamma}$. This means that, according to this model, the population density accounts for a certain amount of spatial heterogeneity, but not for all of it. This is in line with the empirical Moran's I and Geary's C of the residuals of the GAM fit in Section 11.2.

The situation is completely different for the CAR model. Here, most of the spatial variability can be accounted for by including the population density as a covariate. In other words, according to the CAR model, the spatial pattern observed in the average claim sizes can be attributed completely to differences in the population density. Note that the parameters ϕ and γ cannot be compared directly because of the different model formulations. For a more detailed discussion on this see Wall (2004).

The extreme change in the parameter estimate $\hat{\phi}$ indicates a potential misspecification of the CAR model. The log likelihoods, AICs, and BICs confirm that the SAR model provides a better fit for the data. The fit is also superior to standard linear regression models with the same covariate specification, which ignore spatial dependence. Figure 11.2 shows fitted average claim sizes per region according to the CAR and SAR models with intercept model and with quadratic covariate model for the mean, since the inclusion of the population density as a covariate significantly improves the model fits. In particular, a quadratic effect of the population density is sufficient, as also indicated in Figure 11.1.

First, looking at the CAR and the SAR model with the pure intercept model (left column of Figure 11.2) shows that the degree of smoothness implied by the SAR model is stronger than for the CAR model. Nevertheless, both models are able to reproduce the main spatial characteristics observed in the data: for instance, there are high average claim sizes in the East and lower ones in the South-West.

When covariate information regarding the population density is included in the model (right column of Figure 11.2), the results of the CAR and the SAR model become closer, although the CAR almost assumes spatial independence, whereas there is still spatial dependence modeled in the SAR model. The main conclusion from this analysis is therefore that the population density plays an important role for explaining average claim sizes. In particular, average claim sizes in mid-sized to larger cities appear to be smaller than those in rural regions of Germany.

Overall, however, the model fits are unsatisfactory. The empirical correlation between observed and fitted values is 26% under spatial independence and can only be increased to 30% by the SAR model with a covariate model for the mean. In the next section, we therefore investigate the benefit of explicitly taking into account spatial information in a more comprehensive modeling approach, including more relevant covariates.

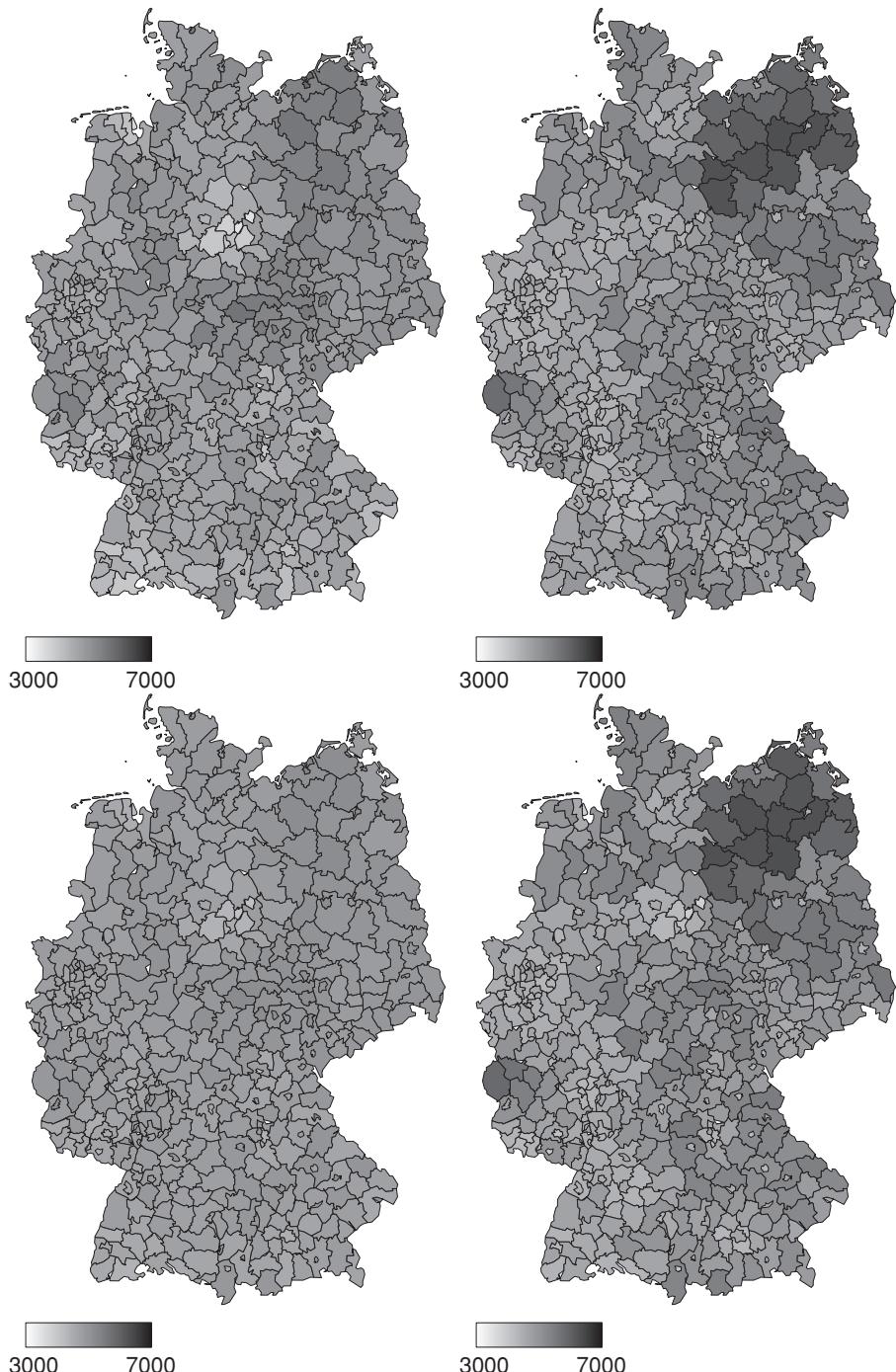


Fig. 11.2. Fitted average claim sizes per region according to CAR (first row) and SAR (second row) models with intercept model (first column) and quadratic covariate model (second column) for the mean. Note that the color coding is different from the left panel of Figure 11.1 in order to emphasize differences more distinctly.

11.5 Hierarchical Model for Total Loss

After the preliminary analysis of the German car insurance data in the previous section, we now study the data in more detail. Car insurance policies are characterized both by their claim frequency, and claim severity. Claim frequencies are the number of occurrences of claims, whereas claim severity reflects the size of each claim. It is common to specify separate distributions for both quantities. In particular, regression models are usually set up to incorporate the influence of significant covariates, such as deductible, type of car, or distance driven per year. Although the model by Gschlößl and Czado (2007) and Gschlößl (2006), which we present here, incorporates those covariates, it also overcomes traditional frequency-severity modeling in two important regards:

- (1) Since Lundberg (1903) introduced the classical compound Poisson model (see Chapter 5), it has been typically assumed that individual claim sizes and numbers of claims are independent, and little attention has been paid to the possibility that this may not be so. In our model, we allow for frequency-severity dependence in the conditional modeling of the claim sizes given the number of claims, and in fact we detect a significant negative relationship between claim frequency and severity.
- (2) Although regression modeling of frequency and severity may remove most of the heterogeneity in the observed claims, it is to be expected that certain spatial dependence patterns cannot be accounted for appropriately, because drivers in certain regions are less risk averse than in others, for instance. We explicitly model such spatial dependence using appropriate proper CAR models in a hierarchical model.

What do we mean by a hierarchical model? This essentially means that we have a general model for the claim sizes and frequencies, with an underlying model specifying the spatial dependence. For parameter estimation, Bayesian methods, as discussed in Chapter 13, are useful for such models.

An alternative hierarchical model for claim frequency and severity with an underlying spatial dependence model was proposed by Dimakos and Frigessi (2002). However, they do not allow for dependence between the frequency and severity components. Moreover, they use an improper spatial autoregressive model.

We now describe and discuss the two separate regression models for the numbers of claims and for the claim sizes. Both models are designed for individual policies, that is, for data specific to individual policyholders and not for aggregated data of groups of policyholders with similar characteristics. Using grouped covariate information, the models are also applicable to such aggregated data.

11.5.1 Model for Claim Frequency

Let N_1, \dots, N_m be the number of claims of m policyholders over one year. These are observed in R different regions (note the change of notation for this different

model-building strategy), where $r(i)$ denotes the region of the i th policyholder. Using this notation, we extend the common Poisson regression model (see Chapter 4) for the number of claims by including spatial random effects. For this, let

$$N_i \sim \text{Poisson}(\mu_i^N), \quad i = 1, \dots, m,$$

where the mean μ_i^N is specified by

$$\mu_i^N = E_i \exp(\mathbf{x}'_i \boldsymbol{\beta} + \gamma_{r(i)}). \quad (11.5)$$

Here, E_i denotes the exposure, $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ik})'$, the covariate vector associated with policyholder i , and $\boldsymbol{\beta} = (\beta_0, \dots, \beta_k)'$, the vector of unknown regression parameters. Further, $\gamma_1, \dots, \gamma_R$ are random effects corresponding to the R regions. They induce, through $\gamma_{r(i)}$, spatial dependence among the variables N_1, \dots, N_m .

In Section 11.3 we discussed appropriate models for spatial autoregression, in particular the CAR and the SAR model. Because of the availability of the full conditionals, we opt here for a CAR model, but with a more sophisticated specification of the proximity matrix C as proposed by Pettitt, Weir, and Hart (2002), which yields a proper joint distribution in contrast to the IAR model discussed in Section 11.3.1.

According to Pettitt et al. (2002), we set in the general CAR model formulation (11.2),

$$c_{jk} = \frac{\psi w_{jk}}{1 + |\psi| n_j}, \quad j, k = 1, \dots, R,$$

if $k \neq j$ and 0 otherwise, where w_{jk} are the elements of the adjacency matrix (as defined in (11.1)) and $\psi \in \mathbb{R}$. As before, we denote by $n_j := |N(j)|$ the number of neighbors of region $j \in \{1, \dots, R\}$.

Further, we define $M_\sigma = \sigma^2 M$ with $m_{jj} = (1 + |\psi| n_j)^{-1}$ for $j = 1, \dots, R$. These choices imply that $c_{jk}\sigma_k^2 = c_{jk}\sigma_j^2$ holds and thus $Q := M^{-1}(I - C)$ is symmetric. Pettitt et al. (2002) also show that Q is positive definite with entries

$$Q_{jk} = \begin{cases} 1 + |\psi| n_j, & j = k \\ -\psi w_{jk}, & j \neq k \end{cases}$$

and therefore for $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_R)'$ it holds that

$$\boldsymbol{\gamma} \sim \mathcal{N}_R(\mathbf{0}, \sigma^2 Q^{-1}), \quad (11.6)$$

which is, in contrast to the IAR specification, a proper joint distribution of $\boldsymbol{\gamma}$, since the inverse of a positive definite matrix is again positive definite. Note that we assume $\boldsymbol{\mu} = \mathbf{0}$ here, because the mean is modeled explicitly through the regression parameters in (11.5).

The parameter ψ allows for a convenient interpretation of spatial dependence. For $\psi = 0$ there is no spatial dependence between a region and its neighbors; in other words, regions are independent. Spatial dependence increases when $|\psi|$ increases.

Additional useful properties of the model can be found in Pettitt et al. (2002). Although not needed here, note that direct maximum likelihood estimation of the parameters is feasible, because the evaluation of the determinant of Q is efficiently tractable.

11.5.2 Models for Claim Severity

For the claim severity, two modeling approaches are possible. One can either set up a model for the individual claim sizes S_{ij} , $j = 1, \dots, N_i$, of each policyholder i , or the average claim size S_i can be modeled, which is defined by

$$S_i := \frac{1}{N_i} \sum_{j=1}^{N_i} S_{ij}, \quad i = 1, \dots, m.$$

For both approaches, we choose appropriate extended Gamma regression models (see Chapters 5 and 6) that include spatial random effects.

11.5.2.1 Individual Claim Size Model

For the individual claim sizes S_{ij} , $j = 1, \dots, N_i$, we assume that they are independently Gamma distributed given the number of claims N_i ,

$$S_{ij}|N_i \sim \text{Gamma}(\mu_i^S, \nu), \quad j = 1, \dots, N_i, \quad i = 1, \dots, m,$$

where the following parametrization of the $\text{Gamma}(\mu, \nu)$ density is used

$$f(s|\mu, \nu) = \frac{\nu}{\mu \Gamma(\nu)} \left(\frac{\nu s}{\mu} \right)^{\nu-1} \exp \left(-\frac{\nu s}{\mu} \right).$$

Then mean and variance of the individual claim size model are given as

$$E(S_{ij}|N_i) = \mu_i^S, \quad \text{and} \quad \text{Var}(S_{ij}|N_i) = \frac{(\mu_i^S)^2}{\nu}.$$

Similar to the model for the number of claims (11.5), a regression on the mean μ_i^S is considered. In particular, we choose a log link and study the following specification

$$\mu_i^S = \exp(\boldsymbol{v}'_i \boldsymbol{\alpha}^S + \zeta_{r(i)}^S), \quad i = 1, \dots, m, \tag{11.7}$$

where $\boldsymbol{v}_i = (v_{i1}, \dots, v_{ip})'$ is a vector of known covariates and $\boldsymbol{\alpha}^S = (\alpha_1^S, \dots, \alpha_p^S)'$ are unknown regression coefficients. The vector of spatial random effects $\boldsymbol{\zeta}^S = (\zeta_1^S, \dots, \zeta_R^S)'$ is modeled by a CAR specification as in (11.6).

Up to this point, we described a hierarchical frequency-severity model with an underlying spatial dependence model. However, the frequency and the severity component are modeled as independent. Dependence between both components can be introduced through the covariate structure. More precisely, because we model the individual claim sizes conditionally on the number of claims N_i , the latter can be used as a covariate. For this, we include the following binary covariates in the model,

$$\delta_{ij} = \begin{cases} 1, & N_i = j \\ 0, & \text{otherwise} \end{cases}, \quad i = 1, \dots, n, \quad j = 2, \dots, N^*, \quad (11.8)$$

where $N^* = \max\{N_1, \dots, N_m\}$ is the maximum number of claims observed. This means that, when the regression parameters associated with δ_{ij} , $j = 2, \dots, N^*$, are significant, then there is dependence between claim frequency and severity. Note that N^* is typically rather small, when individual policies are modeled (in our example $N^* = 4$).

11.5.2.2 Average Claim Size Model

We can model the average claim sizes similarly. Under the assumption that $S_{ij}|N_i$, $i = 1, \dots, m$, are independent and identically distributed, the average claim size S_i is again Gamma distributed, with mean and variance given by

$$E(S_i|N_i) = \mu_i^A = \mu_i^S, \quad \text{and} \quad \text{Var}(S_i|N_i) = \frac{(\mu_i^A)^2}{N_i v},$$

and therefore

$$S_i|N_i \sim \text{Gamma}(\mu_i^A, N_i v), \quad i = 1, \dots, m.$$

As in (11.7) we use a log link for regression on the mean and set

$$\mu_i^A = \exp(\tilde{\boldsymbol{v}}_i' \boldsymbol{\alpha}^A + \zeta_{r(i)}^A), \quad i = 1, \dots, m,$$

where $\tilde{\boldsymbol{v}}_i = (\tilde{v}_{i1}, \dots, \tilde{v}_{ip})'$ is a covariate vector, which is potentially different from \boldsymbol{v} , and $\boldsymbol{\alpha}^A = (\alpha_1^A, \dots, \alpha_p^A)'$ are the associated unknown regression coefficients. As earlier, the number of claims may be used as a covariate here in order to introduce dependence among frequency and severity. For the vector of the random effects $\boldsymbol{\xi}^A = (\zeta_1^A, \dots, \zeta_R^A)'$, a CAR specification as in (11.6) is assumed.

11.5.3 Application

Having described the two-component hierarchical model for the total claim size, we now apply it to the more than 350,000 policies of the German car insurance dataset. For parameter estimation, we follow a fully Bayesian approach. Parameters are estimated using Markov chain Monte Carlo (MCMC) techniques, as presented in Chapter 13.

Table 11.3. *DIC and Scoring Rules for the Total Claim Size Model with and Without Frequency-Severity Dependence and Spatial Effects*

Freq.-sev.	Spatial	DIC	LS	IS _{0.5}	IS _{0.1}	CRPS
yes	yes	269092	-9.5642	-55760	-20412	-2471.9
yes	no	269136	-9.5699	-56125	-20526	-2481.8
no	yes	269122	-9.5669	-55805	-20434	-2474.2
no	no	269175	-9.5734	-56170	-20575	-2484.3

Explicit prior choices for the frequency and the average claim size models can be found in Gschlößl and Czado (2007).

According to a standard Poisson regression analysis ignoring spatial dependence, we include an intercept as well as 17 other variables such as age and gender of the policyholder and deductible as covariates for the frequency model. Most importantly, we also include the population density as a covariate to account for the heterogeneity identified in Section 11.4. Similarly, for the average claim size model, an intercept and a range of 14 covariates are considered after a pre-analysis using a standard Gamma regression model. Furthermore, we include three indicator variables for the number of claims (as defined in (11.8)).

We found a negative relationship between the number of claims and the average claim sizes. This means that the larger the number of claims that are observed the lower the average claim size. A possible interpretation could be that, although more claims occur in cities, they are usually less severe car body incidents in contrast to accidents in more rural areas.

In terms of spatial dependence, we found significantly less accidents in the East and the North than in the South. Conversely, average claim sizes tend to be larger in the East and smaller in the South. Overall, the spatial dependence of the average claim size is mostly weak given the information about the population density. This is in line with the findings in Section 11.4.

Using different model evaluation criteria, the inclusion of an explicit spatial dependence model is in fact shown to substantially improve the model fit. In particular, we consider the deviance information criterion (DIC) by Spiegelhalter et al. (2002) (see Chapter 14), which also takes into account the model complexity, and scoring rules for assessing the quality of probabilistic forecasts (see Gneiting and Raftery 2007): a scoring rule assigns a specific score to a model, which is highest for the true model. Typical choices are the logarithmic score (LS), the interval score at level $\alpha \in (0, 1)$ (IS_α), and the continuous ranked probability score (CRPS). Table 11.3, which is reproduced from Gschlößl (2006), shows the DIC, LS, $IS_{0.5}$, $IS_{0.1}$, and CRPS of our hierarchical model for the total claim size with and without frequency-severity

dependence and spatial effects. From these numbers, the dependence between claim sizes and numbers of claims is also determined to be significant.

The benefit of explicitly modeling spatial dependence is further shown in an analysis of the predictive distribution of the total claim size per region. Ignoring spatial information can lead to an underprediction of total claim sizes of up to 21% in certain regions and, at the same time, to an overprediction of 31% in other regions (see Gschlößl and Czado 2007, section 4.4). Thus, spatial dependence clearly needs to be accounted for in the ratemaking process.

11.6 Discussion and Conclusion

In this chapter we presented models that can account for spatial dependence of areal data, as it is often encountered in an actuarial context. We discussed and applied the model classes of CAR and SAR models to the modeling of average claim sizes in a dataset of German car insurance policies. We also used a specific CAR model as a random effects model to induce spatial dependence into the modeling of claim size and frequency through dependent regression models.

In a regression context, an appropriate covariate with spatial information may not always be identifiable or available (e.g., due to privacy concerns). Then, a CAR or a SAR model can be used as an exploratory tool to detect and analyze potential spatial dependence in the data.

More on spatial count regression models can be found in Gschlößl and Czado (2008) and Czado, Schabenberger, and Erhardt (2013), as well as in the related R-package `spatcounts` by Schabenberger (2009). The modeling of a binary response variable with underlying spatial effects has been studied by Czado and Prokopenko (2008). The class of so-called structured additive regression models (see, e.g., Fahrmeir et al. 2013) provides a comprehensive approach to handling spatial effects in regression problems. Such models, which include GAMs as a special case, can be estimated using Bayesian techniques with the software tool `BayesX` by Belitz et al. (2012).

References

- Banerjee, S., B. P. Carlin, and A. E. Gelfand (2003). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall, London.
- Belitz, C., A. Brezger, T. Kneib, S. Lang, and N. Umlauf (2012). *BayesX – Software for Bayesian inference in structured additive regression models*. Version 2.1, <http://www.stat.uni-muenchen.de/~bayesx>.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B* 36(2), 192–236.
- Besag, J., J. York, and A. Mollié (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 43(1), 1–59.
- Bivand, R. et al. (2013). *spdep: Spatial dependence: weighting schemes, statistics and models*. R package, <http://CRAN.R-project.org/package=spdep>.

- Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York.
- Cressie, N. A. C. and C. Wikle (2011). *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ.
- Czado, C. and S. Prokopenko (2008). Modelling transport mode decisions using hierarchical logistic regression models with spatial and cluster effects. *Statistical Modelling* 8(4), 315–345.
- Czado, C., H. Schabenberger, and V. Erhardt (2013). Non nested model selection for spatial count regression models with application to health insurance. *Statistical Papers*, <http://dx.doi.org/10.1007/s00362-012-0491-9>.
- Dimakos, X. and A. Frigessi (2002). Bayesian premium rating with latent structure. *Scandinavian Actuarial Journal* 3, 162–184.
- Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). *Regression: Models, Methods and Applications*. Springer, Berlin.
- Gelfand, A. E., P. J. Diggle, M. Fuentes, and P. Guttorp (2010). *Handbook of Spatial Statistics*. CRC Press, Boca Raton.
- Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102, 359–378.
- Gschlößl, S. (2006). *Hierarchical Bayesian spatial regression models with applications to non-life insurance*. Ph. D. thesis, Technische Universität München.
- Gschlößl, S. and C. Czado (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal* 107, 202–225.
- Gschlößl, S. and C. Czado (2008). Modelling count data with overdispersion and spatial effects. *Statistical Papers* 49(3), 531–552.
- Lundberg, F. (1903). *Approximerad framställning af sannolikhetsfunktionen. II. återförsäkring af kollektivrisker*. Almqvist & Wiksell's Boktr, Uppsala.
- Pettitt, A. N., I. S. Weir, and A. G. Hart (2002). A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Statistics and Computing* 12, 353–367.
- Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall, London.
- Schabenberger, H. (2009). *spatcounts: Spatial count regression*. R package, <http://CRAN.R-project.org/package=spatcounts>.
- Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B* 64(4), 583–639.
- Sun, D., R. K. Tsutakawa, and P. L. Speckman (1999). Posterior distribution of hierarchical models using CAR(1) distributions. *Biometrika* 86, 341–350.
- Wall, M. M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference* 121, 311–324.

12

Unsupervised Learning

Louise Francis

Chapter Preview. The focus of this chapter is on various methods of unsupervised learning. Unsupervised learning is contrasted with supervised learning, and the role of unsupervised learning in a supervised analysis is also discussed. The concept of dimension reduction is presented first, followed by the common methods of dimension reduction, principal components/factor analysis, and clustering. More recent developments regarding classic techniques such as fuzzy clustering are then introduced. Illustrative examples that use publicly available databases are presented. At the end of the chapter there are exercises that use data supplied with the chapter. Free R code and datasets are available on the book's website.

12.1 Introduction

Even before any of us took a formal course in statistics, we were familiar with *supervised learning*, though it is not referred to as such. For instance, we may read in the newspaper that people who text while driving experience an increase in accidents. When the research about texting and driving was performed, there was a dependent variable (occurrence of accident or near accident) and independent variables or predictors (use of cell phone along with other variables that predict accidents).¹

In finance class, students may learn about the capital asset pricing model (CAPM)

$$R = \alpha + \beta R_M + \varepsilon,$$

where the return on an individual stock R is a constant α plus beta times the return for market R_M plus an error ε . For the CAPM, the dependent variable is the return on individual stocks, and the predictor is the return on the overall market. To fit the model, one needs data (perhaps monthly or weekly) for the return on a an individual stock, and the independent variable is a time series of the return on the market, usually

¹ “Distracted Driving is the New Drunk,” *Philadelphia Inquirer*, Feb 9, 2012.

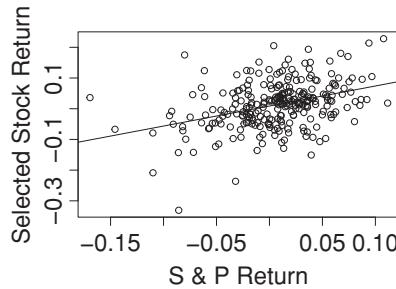


Fig. 12.1. Regression of individual stock return on a market index.

proxied by the return on a widely recognized index composed of many stocks, such as the S&P 500 (see Figure 12.1 where an individual stock's return is regressed on the market return²).

Unsupervised learning, a term coined by artificial intelligence professionals, in contrast, does not involve dependent variables and predictors.

One of the earliest uses of unsupervised learning was in the social and behavioral sciences (for example, see Gorden 1977; Maranell 1974). These early applications used data collected (often from questionnaires) about attitudes, beliefs, and behaviors and constructed sociometric and psychometric *scales* from the data. These scales had no dependent variable per se. A system was used to convert the answers to a set of questions, rankings, and so on, into a single measure believed to quantify a social or psychological concept such as occupational prestige.

Another early application of unsupervised learning in science was in the classification of plants and animals. Data on plant and animal characteristics are collected and used to create a taxonomy. All plants and animals in a particular group, such as species, are expected to be similar to each other and dissimilar to the members of other groups. Many of the techniques discussed in this chapter fall into one of these two types: (1) a scale or quantitative factor based on a weighted sum or average of variables or (2) a classification scheme that organizes the data into (usually) distinct groups based on features they have in common.

In insurance a common example of unsupervised learning is the construction of territories. In certain lines of business such as automobile insurance, it is common to include territorial geographic factors in computing premium rates (Zhang, Wang, and Song 2006). Policyholders are typically included in territories that have some geographic proximity to each other, though increasingly other factors, such as demographic information, are considered in constructing the territories. The key idea is that the policyholders are grouped together based on characteristics they have in common (i.e., similarity on geographic and perhaps other variables). These groupings then can

² Data for Selective Insurance Company and the S&P 500 returns were downloaded from the Yahoo Finance website.

Table 12.1. Dimension Reduction of Claim Database

Occurrence Limit	CSL	Initial Indemnity Reserve	Initial Expense Reserve	Initial Reserve
1,000,000	1,000,000	1,000	1,000	2,000
	500,000	150,000	35,000	185,000
	1,000,000	7,500		7,500
	1,000,000	5,000		5,000
	1,000,000	10,000	10,000	20,000
		17,500	3,500	21,000
	1,000,000	65,000		65,000
	1,000,000	75,000	25,000	100,000
	500,000	5,600		5,600
	1,000,000	15,500		15,500

be used for other purposes such as rating policies or marketing products. Other areas that unsupervised learning has been applied in insurance include the following:

- fraud classification (Brockett et al. 2002; Polon 2008)
- compressing a set of policies to fewer cells for reserving (Freedman and Reynolds 2009)
- pricing or marketing applications, classifying patterns of persistency of life insurance policies (Brzezinski 1981)
- employer rating of actuarial performance (Erin Research 2003)
- predicting mortgage default (Francis and Prevosto 2010)
- variable reduction for predictive modeling (Sanche and Lonergan 2006)
- analyzing the efficiency of financial performance of insurance companies (Mahmoud (2008)
- text mining (Francis and Flynn (2010).

A key concept in unsupervised learning is *dimension reduction*. Table 12.1 displays a snapshot of a typical claims database, in this case closed liability claims posted on the Texas Department of Insurance website. Each column in the exhibit represents a variable or feature. In a typical database used in predictive modeling, there may be hundreds or even thousands of features. Unsupervised learning can be used to condense the features into a much smaller set of *factors* or *components* that capture most of the important information in the many variables. (*Editors' note*: in this context, *factor* refers to an unobserved, or latent, variable. Sometimes, a “factor” refers to a categorical variable as in the statistical program R.)

Each row in the data represents a record, here a liability claim. In a large dataset one may wish to classify similar records together. For instance, an analyst may want to group all records that are “near” each other geographically into territories for further analysis. In classical statistics, variable reduction techniques known as *factor*

analysis or principal components analysis have been used for variable reduction, whereas *cluster analysis* has been used for record reduction.

In sociology and psychology one of the earlier uses of unsupervised learning was to develop *scales*. The scales were intended to measure concepts such as intelligence, ability, and attitude that could not be directly observed or measured, but could be considered related to observed data that could be collected. A number of these scales, such as the *Likert scale*, are based on summing the responses to items on a survey or questionnaire. For instance, the Likert scale typically requires the respondent to select from a range of five ordered options (say from strongly agree to strongly disagree). The answers to the questions can then be added (see Coaley 2010 for an introduction). The items on the questionnaire are often subject to further analysis to validate that they in fact belong in the questionnaire and are related to the concept being measured. One approach is to look at the correlation of the responses on a given question to those of other questions. Items with low correlations may then be rejected. Unidimensional methods such as the Likert scale attempt to measure a single “dimension” or concept.

A similarity between psychometric and sociometric scaling and some of the variable reduction methods in this chapter such as factor analysis and principal components analysis is the use of scores based on summed (possibly weighted) values of the items contributing to the scores. In fact, Spearman’s development of factor analysis was largely due to his work on intelligence scales (p. 150 of Coaley 2010). According to Coaley, Spearman believed that intelligence was attributable to a “general” factor that he labeled “*g*” that enables people to solve problems in a variety of settings. However, others proposed multiple kinds of intelligence. Factor analysis has been used to identify multiple components of a more general concept such as intelligence.

Another similarity is the use of correlation matrices in the construction of the scales. This is pursued more thoroughly in the next section on factor analysis and principal components.

The following sections present the data for this chapter’s illustrated examples and then the major unsupervised learning approaches. The chapter provides an introduction to the major kinds of dimension reduction (in the column dimension/reduction of variables and in the row dimension/reduction of records). More advanced forms of unsupervised learning, such as *k*-nearest neighbor, Kohonen neural networks, association rules, multidimensional scaling, and PRIDIT analysis, are not covered.

12.2 Datasets

Example 12.1 (California Auto Assigned Risk (CAARP) Data). This assigned-risk automobile data were made available to researchers in 2005 for the purpose of studying the effect of constraining the weight of territorial variables on California personal auto premiums. The data contain exposure information (car counts, premium)

and claim and loss information (bodily injury (BI) counts, BI ultimate losses, property damage (PD) claim counts, PD ultimate losses). From these data we have constructed additional variables such as frequency (claim counts/car counts) and severity (ultimate losses/claim counts). The data also contain fields computed by a consultant that capture the effect of proposed regulatory changes (Newpremium1, Newpremium2, etc.). Each record in the data represents a zip code. To create examples that are simple and somewhat exaggerated (compared to the original data), we have also simulated some data roughly based on relationships observed in the original data, but with some parameters exaggerated/ altered to better illustrate our concepts. The code used to create the simulated data is provided on the web site for this book.

Example 12.2 (Texas Closed Claim Data). Download from: <http://www.tdi.texas.gov/reports/report4.html>.

These data are collected annually by the Texas Department of Insurance on closed liability claims that exceed a threshold (i.e., 10,000). The liability claims are from a number of different casualty lines, such as general liability, professional liability, and so on. A key that is downloaded with the data provides information on the variables included in the data. The data include information on the characteristics of the claim, such as report lag, injury type, and cause of loss, as well as data on various financial values such as economic loss, legal expense, and primary insurer's indemnity. Due to the "truncation" models, fit to the data cannot be considered representative of all ground-up losses. However, these data are useful for illustrating and testing various predictive modeling techniques.

Example 12.3 (Simulated Automobile PIP Questionable Claims Data). Francis (2003, 2006) used a Massachusetts Automobile Insurers Bureau research dataset collected in 1993 for research purposes to illustrate data mining techniques, including clustering and their application to modeling questionable claims³ for further action, such as referral to a special investigation unit. We use the description of the data as well as some of the published statistical features to simulate automobile PIP fraud data. These data contain numeric (number of providers, number of treatments, report lag) and categorical variables (injury type, provider type, whether an attorney is involved, whether a police report was filed, whether claimant was treated in the emergency room).

Example 12.4 (Insurance and Economic Indices). These data contain various insurance and economic inflation indices, including workers' compensation severity,

³ Only a small percentage of claims suspected of not being legitimate meet the legal definition of fraud, so such claims are typically not labeled or referred to as fraud.

health insurance, the consumer price index, and various medical-related components of the producer price index. The indices tend to be highly correlated and are useful illustrations of variable reduction methods.

12.3 Factor and Principal Components Analysis

12.3.1 R *princomp* and *factanal* Functions

This section uses the R “*princomp*” and “*factanal*” functions. Venables, Ripley, and Venables (1999) describe the application of *princomp* to principal components analysis, whereas Crawley (2007) describes the application of *factanal* to factor analysis.

Smith (2002) offers an introductory tutorial to principal components analysis (PCA). This tutorial contains a very simple example, including a basic introduction to some needed linear algebra. It then shows an application to computer vision. Kim and Mueller (1978) provide an excellent introduction to factor analysis aimed at social scientists. This introduction covers additional details (such as the topic of factor rotation, which is mentioned briefly in the discussion of factor analysis) that because of the need for brevity, cannot be covered here.

Factor analysis allows us to do variable reduction along the variable dimension. Like the related method, principal components, it can be used to combine many variables into one or perhaps several variables; thus hundreds of variables can be transformed into a handful of predictors. Sometimes the end result is the factors or components themselves, but often they become inputs to a prediction model.

The conceptual model for factor analysis with one factor is⁴

$$x_i = bF + u_i,$$

where x is a variable, b a loading, F a factor, and u a unique component. Because underlying the observed variables is a single unobserved factor, the observed variables are expected to be correlated due to their correlation with the common factor, and the correlation matrix typically plays a role in estimating the factors.

Factor analysis is an unsupervised learning technique that analyzes relationships among a set of variables by reducing the original (manifest) variables to a smaller set of latent (hypothetical) variables. For instance, it is commonly believed in the insurance industry that annual trend costs for many property and casualty lines of business, such as workers' compensation, are higher (and sometimes lower) than overall economic inflation rates observed in the economy, such as payroll inflation and medical cost inflation. This excess cost is often referred to as “social inflation.” Social inflation is more of a concept than an actual empirical item that can be measured. Suppose we

⁴ Notation adopted from p. 67 of Kim and Mueller (1978).

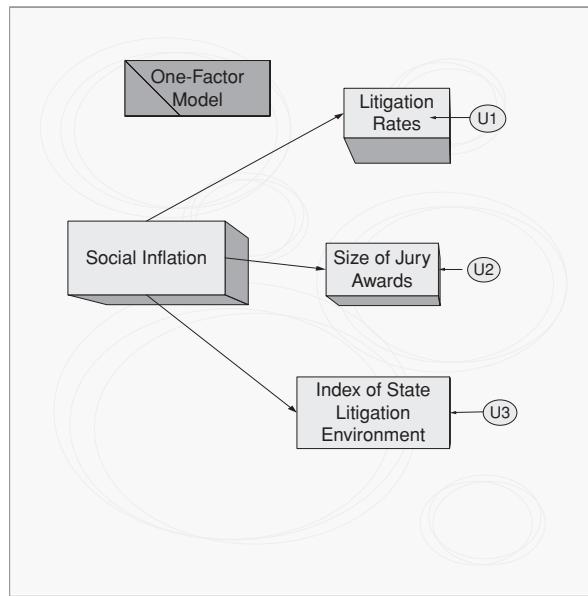


Fig. 12.2. Latent factor (social inflation) and observed variables.

consider “social inflation” to be the unobserved or latent variable and wish to measure it through factors that can be observed (as in Figure 12.2). In this simple illustration we assume three observed variables that can be measured: litigation rates, an index of the legal environment in the state, and the size of jury awards (see Francis 2001, for a fuller development of this particular example).

The factor analysis procedure is applied to data collected on the observed variables to infer values for the unobserved variable.

PCA is a variable reduction technique that does not require the assumption that a latent factor causes the observed results. It simply assumes that two or more variables can be decomposed into components that capture the key features of the data. Figure 12.3 shows an example based on two different medical cost indices. (The two indices, general medical and surgical costs and physician services costs, were downloaded from the Bureau of Labor Statistics website, www.bls.gov, from the producer price section.) The arrow drawn through the scatterplot can be thought of as a “component” that captures the aspects of the indices that are highly correlated with each other. We show later how the two indices can be replaced by their principal component, thus reducing the two variables to one. Figure 12.4 illustrates this in three dimensions. The plot shows three medical indices along with two principal components; it adds an index of pharmaceutical manufacturing costs to indices in Figure 12.3. The three variables can be reduced to the two top components.

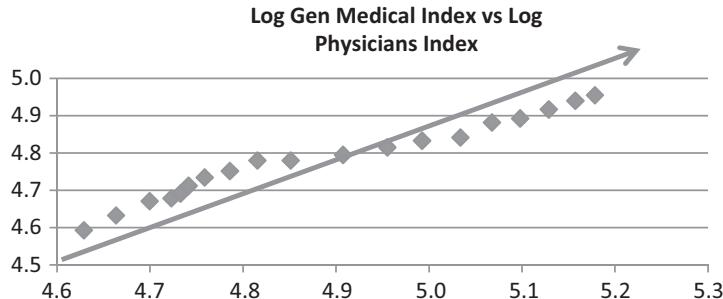


Fig. 12.3. Plot of medical cost indices and a component.

PCA-related methods include singular value decomposition (SVD), another dimensionality reduction method that is frequently included as a multivariate technique in statistical and data mining packages. SVD can be thought of as allowing decomposition of a dataset along the column dimension as with PCA, as well as the row dimension as with cluster analysis. Independent component analysis (ICA) extends PCA by allowing for nonlinear correlations.

Principal component analysis and factor analysis (FA) are related but different methods. PCA decomposes total variance in a set of variables, whereas FA measures shared variance or correlation. In PCA, a principal component is a linear combination of the manifest variables. In FA, a manifest variable is a linear combination of the latent variables. PCA is often viewed as purely exploratory, whereas FA can be both exploratory and confirmatory. In FA, rotations are often needed to help create interpretable factors. In PCA, no interpretations are done on the principal components, so such rotations are not necessary for PCA.

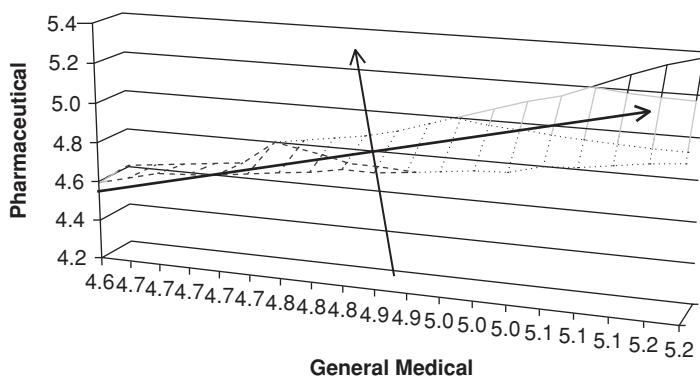


Fig. 12.4. Plot of three medical indices and two principal components.

Table 12.2. Correlation Matrix of Price Indices and WC Severity

	Gen Medical	Health Physicians	Health Pharma	Health Insurance	CPI	WC Compensation	WC Severity
GenMedical	1.000						
Physicians	0.980	1.000					
Pharma	0.988	0.986	1.000				
HealthInsurance	0.994	0.968	0.984	1.000			
CPI	0.990	0.993	0.990	0.985	1.000		
Compensation	0.972	0.988	0.980	0.973	0.993	1.000	
WC Severity	0.952	0.958	0.977	0.962	0.963	0.966	1.000

As shown next, total variance is made up of shared/common variance, unique variance, and error.

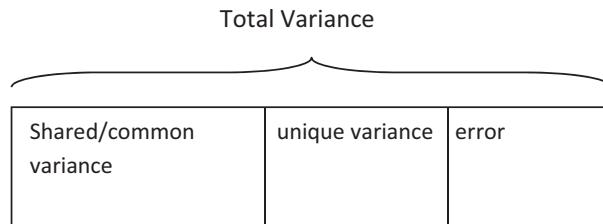


Table 12.2 presents a correlation matrix for several (mostly health-related) price indices and workers' compensation (WC) severity. All variables seem to be relatively highly correlated with each other, though the severity variable, which is more volatile from year to year, tends to show a lower correlation with the other variables.

Let's do a simple example of applying principal components to the two variables. PCA uses either the covariance matrix or correlation matrix. Table 12.3 shows the covariance matrix for the general medical and physicians indices.

The covariance matrix can be decomposed into eigenvalues and eigenvectors:

$$\Sigma = C^T \lambda C, \quad \lambda = \text{eigenvalues}, \quad C = \text{eigenvectors}^5 \quad (12.1)$$

See Smith (2002) for a quick introduction to eigenvectors and eigenvalues.

The first principal component is the column corresponding to the highest eigenvalue. The second principal component corresponds to the lowest eigenvalue. There are as many components as there are variables, but many of the components may add little to modeling the patterns in the data. To fit principal components, we use the `princomp` function on the two medical indices:

⁵ This notation is adopted from Venables et al. (1999, p. 331).

Table 12.3. Covariance Matrix of General Medical and Physicians Indices

	GenMedical	Physicians
GenMedical	0.031	
Physicians	0.018	0.011

```
MedIndices2<-data.frame(Indices$LnGeneralMed, Indices$LnPhysicians)
Simple.Princomp<-princomp(MedIndices2, scores=TRUE)
```

The `princomp` procedure gives us the “loadings” on each of the components. The loadings help us understand the relationship of the original variables to the principal components. Note that both variables are negatively related to the principal component.

```
> Simple.Princomp$loadings
Loadings:
          Comp.1   Comp.2
Indices.LnGeneralMed -0.880  0.475
Indices.LnPhysicians -0.475 -0.880
```

The graph of eigenvalues, produced using `princomp`’s plot function, shows how important each component is. From this graph we can tell that the first component vastly dominates and that component 2 can be discarded without significant loss of the pattern captured by the data. Table 12.4 shows the values of the principal component, produced as a linear sum of the original two variables.

Table 12.4. Original Cost Indices and their Principal Component.
 $\text{Component1} = 0.475 \times \text{Gen Med Index} + 0.88 \times \text{Physicians Index}$

Year	Indices-General Medical	Indices-Physicians	Component 1
1992	4.6	4.6	0.347
1993	4.7	4.6	0.259
1994	4.7	4.7	0.211
1995	4.7	4.7	0.211
1996	4.7	4.7	0.211

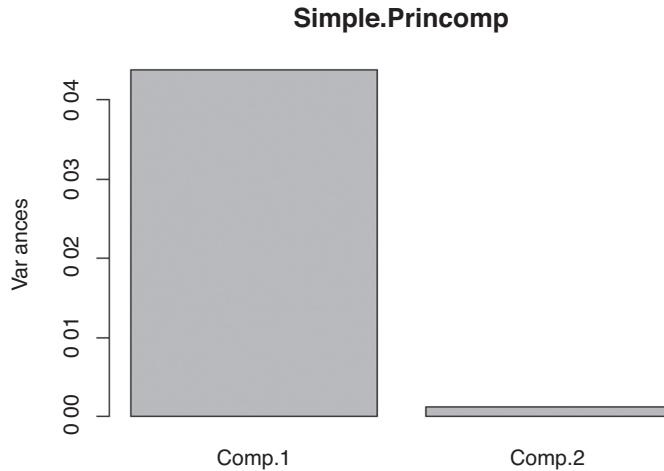


Fig. 12.5. Plot of eigenvalues from principal components procedure.

To give a more complex example, we apply principal components to the seven index variables in Table 12.2. The indices are general medical and surgical costs, physician services, pharmaceutical manufacturing, an index of health insurance costs based on data from www.kkf.org, the CPI, BLS employee compensation costs, and workers' compensation severity computed from data in Bests Aggregates and Averages. From the graph of the eigenvalue (i.e., variance), only the first component can be considered significant. Plots such as Figure 12.6 are typically used to decide how many components to keep. In general, components with low eigenvalues do little to explain the patterns in the data.

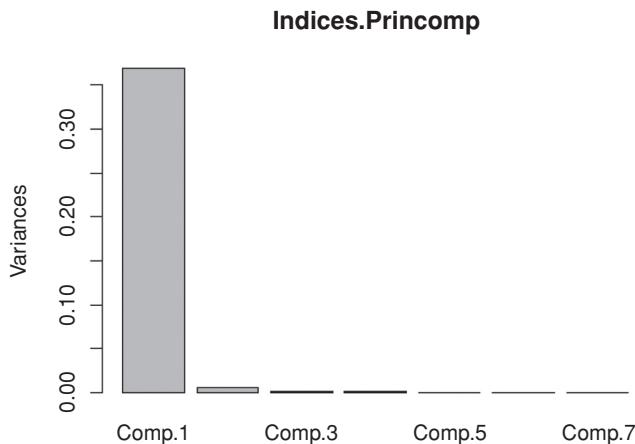


Fig. 12.6. Eigenvalues of components of cost indices.

From the loadings we see that all variables load negatively on the first component, and the workers' compensation severity and the health insurance index have the highest (absolute value) loadings.

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
LnGeneralMed	-0.282	0.276		0.587	0.696		
LnPhysicians	-0.142		-0.656	0.286	-0.222	-0.617	0.185
LnPharma	-0.369		-0.202	-0.746	0.463	-0.165	0.125
LnHealthInsurance	-0.564	0.520	0.435		-0.453	-0.106	
LnCPI	-0.212	0.129	-0.288			0.105	-0.918
LnCompensation	-0.262		-0.477		-0.210	0.750	0.308
LnSeverity	-0.576	-0.789	0.159	0.129			

To illustrate factor analysis we apply factor analysis to the cost indices using the factanal function and ask for two factors.

```
>Indices.Factanal<-factanal(Indices2,factors=2,scores="regression")
```

As mentioned previously, factor loadings are based on correlation between manifest variables and their factors. The loadings are as follows:

	Factor1	Factor2	Factor3
LnGeneralMed	0.982		
LnPhysicians	0.941	0.256	0.102
LnPharma	0.989		
LnHealthInsurance	0.992	-0.103	
LnCPI	0.997		
LnCompensation	0.991		
LnSeverity	0.970		

It appears that the indices load almost equally on the first factor. Also, unlike with principal components analysis, the first factor is positively correlated with the indices. These loadings also assist our interpretation of the factors. For instance, based on the loadings we can think of the first factor as a general inflation factor and the second factor as specific to certain components of medical inflation. Only

physicians and health insurance load on the second factor, and only physicians load on the third. The communality is the proportion of variance explained by the factors. A high communality (> 0.5) indicates that the number of extracted factors adequately explain the variances in the variables. A low communality (< 0.5) indicates that more factors may be needed or that this dataset is not appropriate for factor analysis.

The communalities produced by `factanal` (shown next) indicate that the first factor accounts for more than 95% of the variance.

	Factor1	Factor2	Factor3
SS loadings	6.727	0.090	0.036
Proportion Var	0.961	0.013	0.005
Cumulative Var	0.961	0.974	0.979

The factors derived from factor analysis are not guaranteed to be interpretable. A procedure called factor rotation is used to create new factors and to improve their interpretability. Commonly used methods include orthogonal and oblique methods. Orthogonal rotations are where the axes are kept at a 90-degree angle and include methods referred to as Varimax, Equimax, and Quartimax. Oblique rotations are where X and Y axes are not 90 degrees apart and include the method Promax. Although factor rotation is not covered in this chapter, R offers the `varimax` and `promax` functions for rotating the factor loading outputs of `factanal`.⁶

A factor component score can be assigned to each record. The score is in the form of a standardized value assigned to individual observations. This is useful if follow-up studies are performed. The score may itself become an input to a predictive model.

Weaknesses of factor analysis include the following:

- The following aspects of FA are subjective: rotation method, number of factors, extraction method, and interpretation of factors. As a result, a wide variety of outcomes could result from performing FA.
- It only measures linear relations.
- Outliers may adversely affect the results.

Its strengths include that factor analysis is a well-known and widely used method, especially in the social science fields, and it is Nonparametric.

Factor analysis (and principal components as implemented in the R `princomp` procedure) uses metric/numeric variables. Although categorical variables could be represented by dummy variables, it is good practice to apply FA and PCA to numeric/metric values. Except for the maximum likelihood estimation rotation

⁶ See for instance the R Help material on these functions.

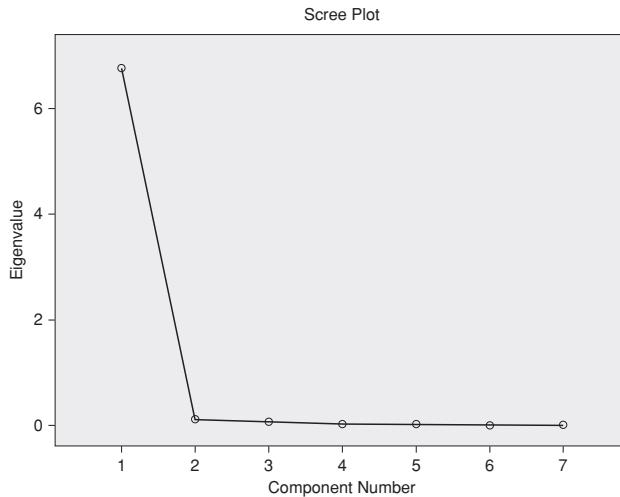


Fig. 12.7. Scree plot of factors from index data.

method, normality is generally not required. However, the results are enhanced if normality is kept.

Although many software packages (e.g., R, SAS, etc.) offer the option of executing PCA within factor analysis functions (e.g., Proc Factor), to avoid confusion it is recommended that PCA-specific functions be used (e.g., Proc Princomp). The two methods are so closely related that some software packages (i.e., SPSS⁷) offer only one of the options. Thus, unless one has a specific purpose, such as identifying latent factors or concepts underlying the variables, it may not matter much which approach is used for variable reduction.

12.3.2 Determining the Number of Factors

A *Scree plot* is used to determine the number of factors to retain. The “elbow” point is where a noticeable drop in eigenvalue occurs (see Figure 12.7). In practice, because multiple elbow points could occur in a scree plot, it is advisable to cross-reference with another criterion to determine which of the points is most appropriate.

Note that different statistical software packages could produce different outcomes even when applying the same statistical technique and input parameters. The differences could be due to various reasons, including implementation and stopping criteria. For instance, the factor analysis results produced by R often differ from those of SAS.

⁷ IBM SPSS Statistics 20 only offers factor analysis.

12.4 Cluster Analysis

12.4.1 R Cluster Library

The “cluster” library from R is used to perform the analyses in this section. Therefore it is necessary to install it from a CRAN mirror site and load the cluster library⁸ at the beginning of your R session. Many of the functions in the library are described in the Kaufman and Rousseeuw (1990) classic book on clustering: *Finding Groups in Data*. Note that at the time that book was written, R was not the predominant software for scholarly statistical analysis, and much of the book’s discussion assumes the user is loading an execute file of code written in Fortran and running it under DOS. Our Appendix contains comprehensive examples of the R code used in this chapter. Also please visit the CRAN website and download the pdf reference file for the cluster library.

12.4.1.1 Books and Papers on Clustering

Because it may be difficult to obtain the *Finding Groups in Data* book on which many of the R functions cited in this chapter are based, other appropriate references are provided in this section. An excellent introductory reference to basic clustering with illustrative clustering applications is *Cluster Analysis* (Aldenderfer and Blashfield, 1984). This book can be read in one or two weekends and, at this time, is available in paperback and e-book form. However, the book predates the R open source movement, so it does not offer any guidance for performing clustering in R. The book *Cluster Analysis* (Everitt et al. 2011) introduces all the basic clustering approaches included in this chapter, along with more advanced topics such as k-nearest neighbors and neural networks. Although it frequently refers to R functions for performing the procedures described in the book, it does not provide the more comprehensive coverage of the R applications contained in the other papers and tutorials cited next. The “Unsupervised Learning” chapter in *An Introduction to Statistical Learning with Applications in R* also provides an introduction to principal components and clustering with illustrated examples and exercises in R.

The original functions in the book *Finding Groups in Data* were written in Fortran. In the late 1990s and early 2000s S - PLUS was a popular vehicle for academic statistical applications. Rousseeuw collaborated with Struyf and Hubert in the paper “Clustering in an Object-Oriented Environment” (Struyf et al. 1997; (<http://www.jstatsoft.org/v01/i04>). It describes seven S - PLUS functions that can be used for clustering and also provides background on the statistics used in the functions. This paper also provides a guide to the graphical outputs found within the R cluster library. S - PLUS is a commercial version of the language S, whereas R is the opens source

⁸ Using the syntax: `load(cluster)`.

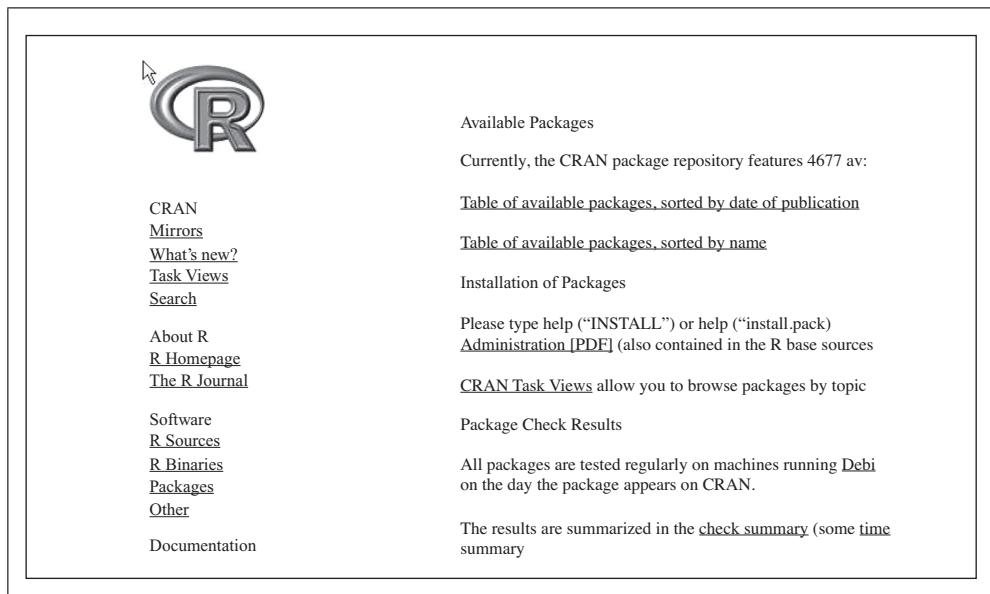
implementation. Most syntax for S-PLUS will work in R. The functions used in this chapter match those described in Struyf et al. (1997) and have been run in R. Additional help with the R cluster functions can be obtained from R's help capability. The user must access the help function from the R command line. For instance to get help with the function `daisy`, type:

```
?daisy
```

Then, documentation of the `daisy` dissimilarity matrix function will appear in an output window.

12.4.1.2 Help for R cluster Library

A pdf file with documentation for all the `cluster` library functions (Maechler 2012) can be obtained from one of the CRAN (Comprehensive R Network) mirror sites, such as <http://lib.stat.cmu.edu/R/CRAN>. When at the site, on the left-hand side click on “packages,” and then select Table of available packages, sorted by name (see the snapshot of the CRAN screen). When the table appears, find the `cluster` function, click on it and then find the `cluster` documentation. Click on it and then download the reference manual. The document provides descriptions for each function in the `cluster` library. Although this file has up-to-date documentation of the functions in the `cluster` library, the Struyf et al. (1997) paper provides a much more comprehensive narrative with worked-out examples.



The screenshot shows the CRAN package repository homepage. On the left, there is a sidebar with links: CRAN Mirrors, What's new?, Task Views, Search, About R, R Homepage, and The R Journal. The main content area is titled "Available Packages" and contains the following text: "Currently, the CRAN package repository features 4677 av:" followed by two links: "Table of available packages, sorted by date of publication" and "Table of available packages, sorted by name". Below this, there is a section titled "Installation of Packages" with the text: "Please type help ("INSTALL") or help ("install.pack") Administration [PDF] (also contained in the R base sources)". It also mentions "CRAN Task Views allow you to browse packages by topic". Further down, there is a section titled "Package Check Results" with the text: "All packages are tested regularly on machines running Deb[...] on the day the package appears on CRAN." and "The results are summarized in the `check summary` (some time summary)".

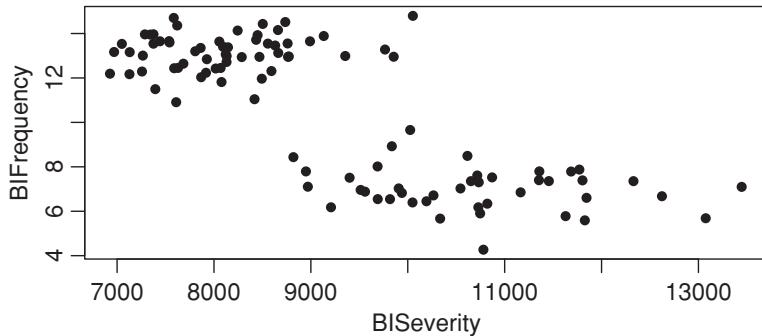


Fig. 12.8. Simple two-cluster example.

12.4.1.3 Tutorials on Clustering

Oksanen (2012) provides a tutorial on hierarchical clustering and fuzzy clustering that includes a library not covered (vegan) in this chapter. See <http://cc.oulu.fi/~jarioksa/opetus/metodi/sessio3.pdf>.

Quick R Tutorial: The Quick-R website provides resources and tutorials for people who are already familiar with statistics to learn how to apply those statistics in R. The Quick-R information on clustering is found at <http://www.statmethods.net/advstats/cluster.html>. Note that by clicking on the link to “*amazing variety*” you can obtain an up-to-date list of all R clustering functions along with the package (noted in the parentheses) that contains the function. Several such as kmeans (similar to pam and clara) are part of basic R and do not require you to load a library, such as cluster. Books such as Crawley (2007) briefly describe some of the built-in R.

12.4.2 Similarity and Dissimilarity

In classical statistics the technique of clustering has been used to partition data. The two main methods of clustering are k -means clustering and hierarchical clustering. Specific clustering approaches are discussed after we introduce the common measures used in the clustering procedures. To motivate an understanding of clustering we show a simple example. Figure 12.8 displays bodily injury frequency versus bodily injury severity for automobile liability claims. These data were simulated, but the parameters of the simulation were based on our analysis of the California automobile data. The data on the graph seem to cluster into two groups, a high frequency/low severity group and a low frequency/high severity group:

- Members of the same group tend to be similar to each other. That is, their values on the analysis variables (here BI frequency and BI severity) tend to be similar to each other.

- Members of different groups tend to be dissimilar. That is, their values on the analysis variables tend to be different.

Thus, to characterize the different groups, measures of similarity and dissimilarity are used. Similarity statistics measure how alike two records are. The Pearson correlation coefficient is a kind of measure of similarity (between two variables, rather than two records). Dissimilarity statistics measure how “far apart” two records are. In general, there is a complementarity between measures of similarity and dissimilarity. The two most common measures of dissimilarity are Euclidean distance and Manhattan distance.⁹

Euclidean Distance: The statistic is based on the record-by-record squared difference between the value of each of the variables for a record and the values for the record it is being compared to. The distance d measured on p numeric variables between two records i and j is

$$d_{ij} = \sqrt{ \sum_{k=1}^p (x_{ik} - x_{jk})^2 }. \quad (12.2)$$

Manhattan Distance: The statistic is based on the record-by-record absolute difference between the values for a record and for the record it is being compared to. The distance d measured on p numeric variables between two records i and j is

$$d_{ij} = \sum_{k=1}^p |x_{ik} - x_{jk}|. \quad (12.3)$$

In R, the `daisy` procedure in the `cluster` library can be used to compute dissimilarity matrices. If no method is selected for computing dissimilarities, the Euclidean method is used as the default.

The R code to compute a dissimilarity matrix for our automobile data (contained in `dataframe ClusterDat1`) is:

```
>distance<-daisy(ClusterDat1)
```

Table 12.5 displays the Euclidian dissimilarity measure for selected records from the simulated auto data. Table 12.1 shows dissimilarity measures for the first two records of our data. The dissimilarity output is typically an $n \times n$ matrix, where n is the number of records in the data.

⁹ See Aldenderfer and Blashfield (1984) for a fuller development of the formulas for dissimilarity.

Table 12.5. View of Dissimilarity Matrix of Auto BI Data

	1	2	3	4	5
1	–	1,400	2,671	4,488	410
2	1,400	–	1,271	3,088	1,810
3	2,671	1,271	–	1,817	3,081
4	4,488	3,088	1,817	–	4,898
5	410	1,810	3,081	4,898	–

For categorical variables a different dissimilarity measure is needed. To motivate the discussion we begin with two binary categorical variables that take on only two values: zero and one. The notation for the discussion is shown next.

		Record 1		
		0	1	Total
Record 2	0	a	b	a+b
	1	c	d	c+d
Total		a+c	b+d	a+b+c+d

in the table, the number of variables that the two claimants are zero on is denoted b , and the number of variables that the two have a 1 on is denoted c . The off-diagonal counts are a and d . The sample matching method computes the proportion of agreements:

Sample matching:

$$d_{ij} = \frac{b + c}{a + b + c + d}.$$

Jaccard's coefficient:

$$d_{ij} = 1 - S_{in}, \quad S_{ij} = \frac{a}{a + b + c}' S = \text{similarity}.$$

Rogers and Tanimoto:

$$d_{ij} = \frac{2(b + c)}{a + d + 2(b + c)}$$

The Rogers and Tanimoto statistic gives more weight to the count of disagreements between records. The Jaccard coefficient does not count the negatives (zero values for binary variables).

Our example of categorical dissimilarity uses the simulated fraud data (see Table 12.6). For this illustration the sample contains only two binary categorical variables:

**Table 12.6. Dissimilarity Matrix for Fraud Data
(two-variable sample of binary variables)**

	1	2	3	4	5
1	0	0	0.5	1	0.5
2	0	0	0.5	1	0.5
3	0.5	0.5	0	0.5	1
4	1	1	0.5	0	0.5
5	0.5	0.5	1	0.5	0

whether the claimant was represented and whether the injury was a soft tissue injury. To obtain this dissimilarity matrix in R, use:

```
>dissim.fraud<-as.matrix( daisy( ClusterDat2,
    metric= "gower" ) )
```

where ClusterDat2 is the simple dataset with two variables from the fraud data.

Since our binary variables are coded as zero for a “no” on a given category and one for a “yes” (i.e., if represented by an attorney, the litigation variable is zero; otherwise it is one), Euclidean distances can be computed. For comparison, we display the dissimilarities computed using Euclidean distances (see Table 12.7). Note that Zhang et al. (2006) point out a number of problems with treating categorical variables as scalar and computing Euclidean or Manhattan distances. For instance the measure is not independent of the scale chosen (i.e., coding no’s as 1 and yes’s as 3, rather than using zeros and ones). For categorical variables with many categories, many dummy binary variables would be required as separate variables for the dissimilarity matrix.

**Table 12.7. Euclidean Dissimilarity Matrix for Fraud Data
(two-variable sample of binary variables)**

	1	2	3	4	5
1	0	0	1	1.414214	1
2	0	0	1	1.414214	1
3	1	1	0	1	1.414214
4	1.414214	1.414214	1	0	1
5	1	1	1.414214	1	0

**Table 12.8. Matrix of Mixed Dissimilarity for Fraud Data
(two-variable categorical and one numeric)**

	1	2	3	4	5
1	0	0.4	0.511	0	0.067
2	0.4	0	0.911	0.4	0.467
3	0.511	0.911	0	0.511	0.444
4	0	0.4	0.511	0	0.067
5	0.067	0.467	0.444	0.067	0

12.4.3 Missing Data

For categorical variables the measures treat a missing value on one or both variables the same as a non-match.

12.4.4 Mixed Data

For mixed categorical and numeric data the gower method is used. The method treats categorical variables in a manner similar to sample matching. For numeric variables, a distance is measure similar to a Manhattan distance.

$$d_{ij} = \frac{|x_{ij} - x_{jk}|}{R_k}, \quad R = \text{range}, \quad S_{ij} = 1 - d_{ij} \quad (12.4)$$

To illustrate this, we use a subset of the simulated fraud data that contains three variables: the two in the previous example plus the number of medical treatments (see Table 12.8). To compute the dissimilarity we use the code:

```
>dissim.fraud3<-as.matrix( daisy( ClusterDat3,
    metric= "gower" ))
```

12.4.5 Standardization

Scale affects the results of clustering. Therefore the analyst often prefers to standardize variables before performing clustering (i.e., for each variable x , subtract its mean and divide by its standard deviation or absolute deviation). The dissimilarity function `daisy` allows for scaling. The `daisy` function uses the absolute deviation rather than the standard deviation when scaling because the absolute deviation is more robust to outliers. To scale the auto BI data, we can standardize the frequency data with the code:

```
>distance2<-as.matrix(daisy(ClusterDat1, stand=TRUE))
```

Table 12.9. Euclidean Dissimilarity Matrix for Scaled BI Variables

	1	2	3	4	5
1	0	2.987	2.141	2.176	2.466
2	2.987	0	2.383	2.393	3.51
3	2.141	2.383	0	0.123	1.511
4	2.176	2.393	0.123	0	1.606
5	2.466	3.51	1.511	1.606	0

12.4.6 k-Means Clustering

k -Means clustering is a common procedure that clusters data by partitioning it into k clusters where k is an input to the cluster procedure. That is, the analysts selects k and then uses the method to segment the data into k groups. Selecting k is somewhat subjective, but is often guided by knowledge of the data and subject. Kaufman and Rousseeuw (1990) discuss a number of methods used to select the optimal number of clusters. One approach they recommend it to examine the silhouette statistics.

An iterative procedure is used to assign each record in the data to one of the k clusters. The iteration begins with the initial centers or *medoids* for k groups. These medoids may be specified by the user, but often they are randomly selected from records in the data. The method uses a dissimilarity measure to assign records to a group and to iterate to a final grouping. For example, in the bodily injury data, suppose we wish to assign each zip code to one of two groups. The following is one way to classify all the records into one of two groups so that the groups will be as homogeneous as possible:

- Randomly select two zip codes from the data.
- For the two records compute the mean for each variable (i.e., BI frequency, BI severity). These then become the initial medoids or means of the two classes.
- For every other record in the data:
 - Compute the distance between that record and the centroids of each of the groups.
 - Assign the record to the group with the smallest distance measure.
- The next or “swap” step searches for two records that are the medoids for two records that provide a better “fit” in the sense that they reduce the average dissimilarity.
- An alternative measure for each group’s medoid is the mean of each variable over all the members of the group.
- Iterate until the centroids do not change significantly or until a maximum number of iterations specified by the user have been completed.

We illustrate using the k -means procedure on the two-variable simulated auto BI data. In the cluster library we use the `pam` (partition against medoids, the more accurate

method) or `clara` (clustering large applications) function. The `clara` function is recommended for large datasets.

```
>BICluster1<-pam(ClusterDat1,2,metric="euclidean")
>BICluster1<-clara(ClusterDat1,2,metric="euclidean")
```

Next we see that the output displays the medoids (or by variable means) for each cluster, thereby identifying a low-frequency/low-severity and high-frequency/high-severity cluster. The low-frequency cluster contains 63% of the records.

```
> BICluster1
Call: clara(x = ClusterDat1, k = 2, metric = "euclidean")
Medoids:
  BIFrequency BISSeverity
[1,]    11.39769   8202.802
[2,]    13.28089  10749.593
Objective function: 577.0351
Clustering vector: int [1:100] 1 1 2 1 1 1 1 1 1 1 1 2 1 2 2 2
1 1 ...
Cluster sizes:      63 37
```

Next we standardize the data used for clustering. The two clusters contain 64% and 37% of the data, about the same as for the unstandardized data. However, the procedure now separates these data into a low-frequency/high-severity and high-frequency low-severity group, similar to the actual data.

```
BICluster2<-clara(ClusterDat1,2,metric="euclidean",stand=TRUE)
Call: clara(x = ClusterDat1, k = 2, metric = "euclidean",
            stand = TRUE)
Medoids:
  BIFrequency BISSeverity
[1,]    13.040838   8043.759
[2,]     7.570882  11019.007
Objective function: 0.625121
Clustering vector: int [1:100] 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2
1 1 ...
Cluster sizes:      64 36
```

It is customary to plot clustered data using principal components as the *X* and *Y* axes of the graph (see Figure 12.9). Principal components are discussed in Section 12.2.

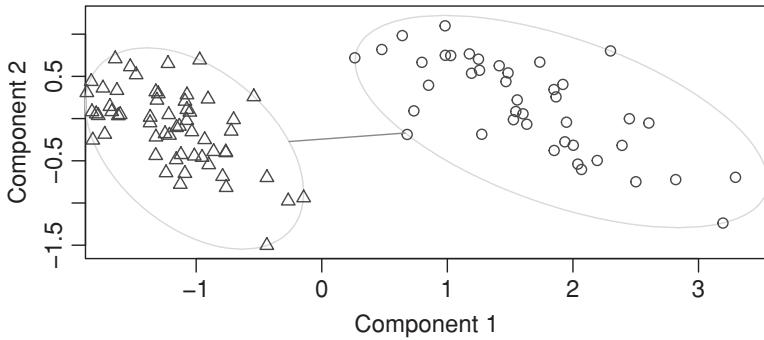


Fig. 12.9. Cluster plot of auto data. These two components explain 100% of the point variability.

In the case of our simple example using simulated data, the two components correspond closely to the two variables that we are clustering on.

```
>plot(BICluster2)
```

The silhouette plot provides information on the goodness of fit of the clusters to the data (see Figure 12.10). The silhouette statistic measures the similarity of members of a cluster relative to other clusters (i.e., the next closest cluster in the case of more than two clusters). The closer the statistic is to 1, the more homogeneous the partitions are. The silhouette plot is a bar plot of the silhouettes of the records for each cluster.

To illustrate clustering on mixed numeric and categorical variables, we apply the `clara` function to the simulated auto questionable claims data. To utilize categorical data we first create a dissimilarity matrix (using the `daisy` function) using the gower

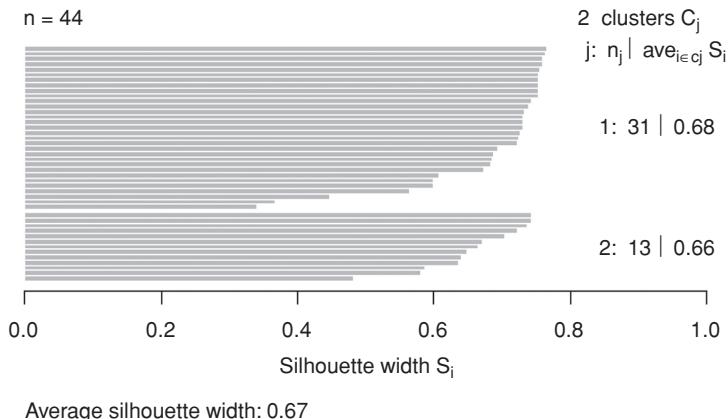


Fig. 12.10. Silhouette plot of simulated auto data.

dissimilarity measure. The matrix then becomes the input to the `pam` function (note the `clara` function does not accept dissimilarity matrices). Figure 12.11 shows the silhouette plot for this data.¹⁰ Because we cannot create a cluster plot for categorical data, we display the “confusion matrix” or the cross-tabulation of actual versus predicted categories.

As can be seen from the cross-tabulation in Table 12.10, the k -means cluster correctly predicts 92% of the smaller (questionable claims) cluster and 96% of the larger (legitimate claims) cluster.

We ran the k -means clustering on the actual auto BI CAARP data and note that the results are much noisier, as can be seen from the cluster plot in Figure 12.12. We clustered on the following variables: BI frequency, PD frequency, log of the BI severity, and log of the PD severity. Because of the skewness of the severity data, we used the log transformation. We ran the procedure with $k = 2, 3$, and 4 . The average silhouette was highest for $k = 2$, suggesting that it is the best partition. The $k = 2$ average silhouette was 0.3 as opposed to 0.23 or less for $k = 3$ and 4 . The results are shown in the box. Note that the PD severity appears to have no meaningful effect on the clustering. The BI and PD frequencies separate into a high-frequency and low-frequency group. The low-frequency group has a higher BI severity (on a log scale).

```
Call: clara(x = AutoBIVars, k = 2, metric = "manhattan",
            stand = TRUE, keep.data = TRUE)
Medoids:
    AutoBI.FREQBI AutoBI.FREQPD AutoBI.LnBiSev AutoBI.LnPdSev
[1,]     11.953501     38.04287     9.080598     7.748904
[2,]      8.347081     25.45903     9.328678     7.763548
Objective function: 3.243358
Clustering vector:  int [1:1459] 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 ...
Cluster sizes:      780 679
Best sample:
```

12.4.7 Fuzzy Clustering

Insurance data tend to be messy. Because of the significant randomness and heavy tailedness of many of the variables, such as claim severity. As a result, the true structure of the data is often obscured by the noise. As we can see from the cluster graph of the auto BI data (Figure 12.15), crisp clusters are often the exception, rather than the rule. Thus it may be better to assign each record a partial membership in

¹⁰ Note that for this plot the sample size was reduced to 100.

Table 12.10. Cross-Tabulation of Actual and Predicted Clusters for Questionable Claims Data

Actual	Predicted		Grand Total
	1	2	
1	659 97%	22 3%	681
2	26 8%	293 92%	319
Grand Total	685	315	1000

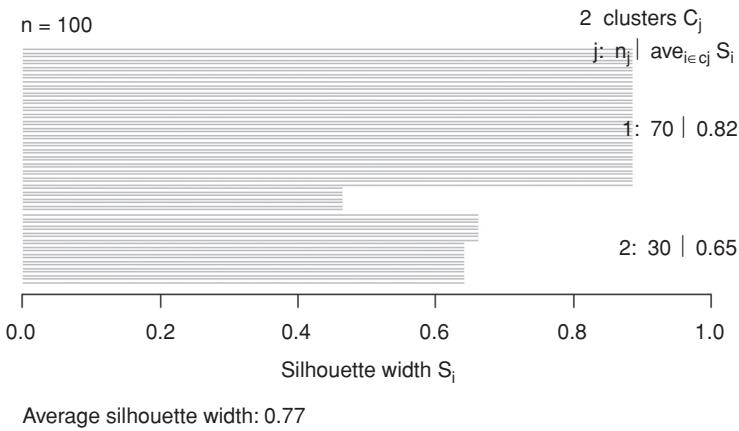


Fig. 12.11. Silhouette plot of questionable claims data.

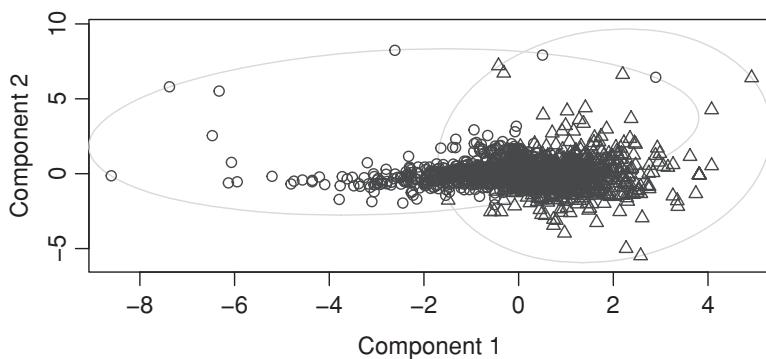


Fig. 12.12. Cluster plot of actual auto data. These two components explain 75.68% of the point variability.

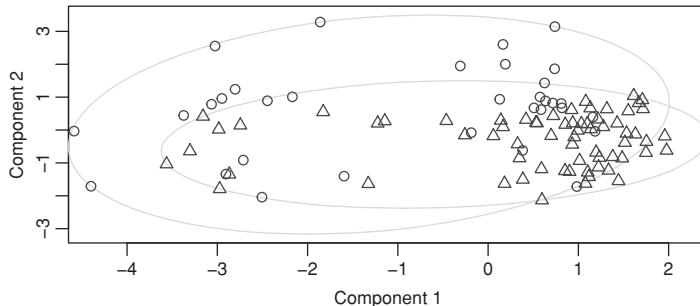


Fig. 12.13. Fuzzy clustering plot of fraud data. These two components explain 40.44% of the point variability.

two or more groups, rather than assigning it to only one group. The procedure that does this is called fuzzy clustering. The R function for fuzzy clustering is `fanny`. Kaufman and Rousseeuw (1990) and Maechler (2012) state that `fanny` minimizes a weighted dissimilarity function where the weights are the fuzzy “membership” function denoting the proportion of its membership in a given cluster. For any record, the membership functions sum to 1. At the limit, the membership functions equal 1 for only one cluster and 0 for all others, yielding the traditional hard clusters:

$$\min \left(\sum_{v=1}^k \frac{\sum_{i=1}^n \sum_{j=1}^n u_{iv}^r u_{jv}^r d(i, j)}{\sum_{j=1}^n u_{jv}^r} \right), \quad (12.5)$$

over $0 < u < 1$ and $\sum u = 1$, $r \geq 1$. Here, n is the number of records, k is the number of clusters, and $d(i, j)$ is the dissimilarity between i and j .

An iterative procedure is used to solve for the clusters and for the membership functions of the clusters. The user must specify the number of clusters or a range of clusters.

To illustrate, we apply `fanny` to our automobile fraud data. This dataset contains a number of binary variables: attorney represented, soft tissue injury, type of medical provider, filing of a police report, use of an emergency room, and whether the claimant had a prior claim. Because `fanny` does not use the gower dissimilarity measure, we treated these variables as numeric. The data also show several numeric variables (number of providers, number of treatments and lag from accident date to report date). All variables were used in the clustering. We show the code next, and Figure 12.13 displays the cluster plot from `fanny`. Table 12.11 shows that `fanny` output assigns a membership value to each record for each cluster.

```
>QuesClaimsDataCluster2<-fanny(ClusterDatFraud, 2, metric="euclidean",
  keep.data=TRUE, memb.exp=1.5)
>plot(QuesClaimsDataCluster2)
```

Table 12.11. First Five Records of Fanny Membership Output

Record	X1	X2	Closest Cluster
1	0.812	0.188	1
2	0.094	0.906	2
3	0.232	0.768	2
4	0.885	0.115	1
5	0.585	0.415	1

Unlike k -means clustering, hierarchical clustering does not create a specific number of clusters. Instead it sequentially partitions the data. At each step additional clusters are created under divisive clustering (see Figure 12.14), or clusters are merged under agglomerative clustering. The results are presented in the form of a tree. Like k -means clustering, hierarchical clustering uses a dissimilarity matrix to split data into groups. Under divisive clustering, the procedure starts with a cluster of size one, containing all the data. It then splits the data into two groups. These two groups are then split into more groups. Ultimately the data can be partitioned into as many clusters as there are records (i.e., each record is its own cluster and no grouping is performed). To perform divisive clustering we use the `diana` function:

```
>BICluster7<-diana(AutoBIVars, metric="manhattan", keep.data=TRUE,
  stand=TRUE)
>plot(BICluster7)
```

Agglomerative clustering is the opposite of divisive clustering. Agglomerative clustering starts with one cluster for each record and then combines records. In

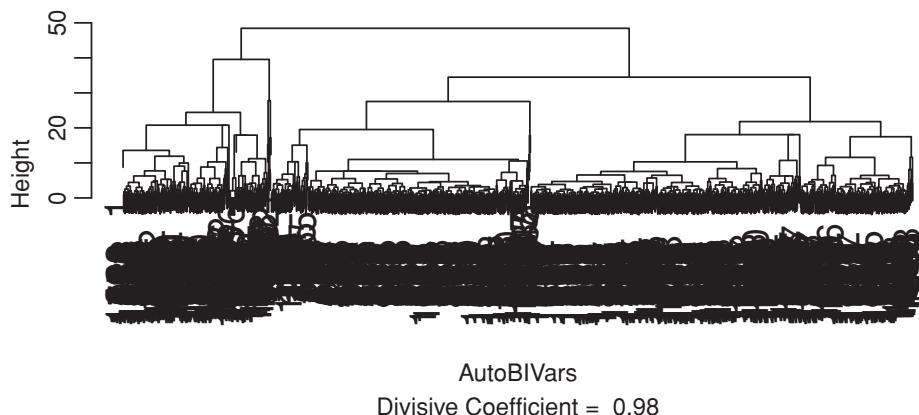


Fig. 12.14. Divisive clustering of auto BI data.

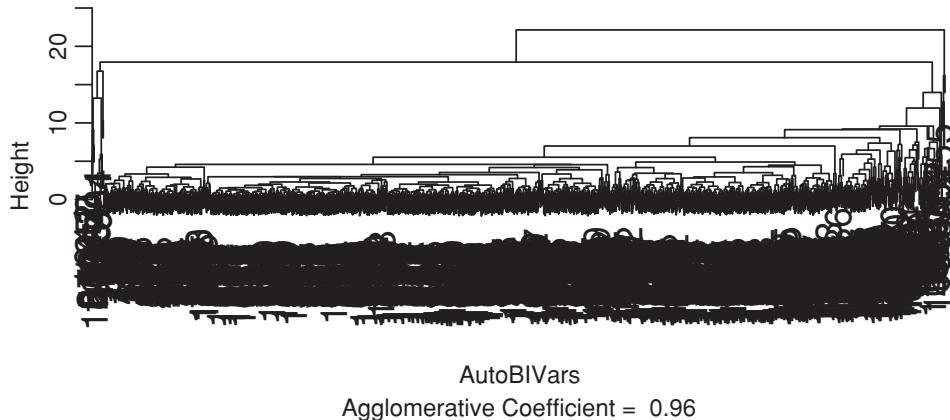


Fig. 12.15. Agglomerative clustering plot.

the first step, the two records with the smallest dissimilarities are combined. In the next step two more records are combined, or a record is joined to the cluster grouping created in step 1. In R the function `agnes` is used to perform agglomerative clustering. The statistic “agglomerative coefficient” is supplied with the `agnes` output, and the “divisive coefficient”¹¹ is supplied with the `Diana` output. According to Kaufman and Rousseeuw (1990), the closer to zero these coefficients are, the less is the natural grouping structure in the data; the closer to one they are, the higher the degree of structure. Higher coefficients are observed for larger datasets, so the coefficients cannot be used to compare data of different sizes.

The graphical output from a hierarchical clustering can be used to assess how many clusters to keep. Note that when your data contain many records, as in this example, the details are blurred at deeper levels of the tree, since ultimately each record is a point on the plot. The length of each branch of the tree (sometimes called an *icycle* graph) is proportional to the distance between the groups. The smaller the length, the less dissimilar the clusters are. Note that under both divisive and agglomerative clustering the first (next to last last) partition has greater depth than subsequent clustering. Under divisive clustering, the second through fourth partitions (i.e., six groups) also seem to be meaningful. Under agglomerative clustering the two-way split of the data from the next to last clustering seems much more significant than further partitions of the data (see Figure 12.15).

12.4.8 Why We Use Clustering

The automobile suspicious/questionable claims data supply one example of a common insurance application of clustering. When an insurance company wishes to develop

¹¹ See pages 210–218 of Kaufman and Rousseeuw (1990) for a more thorough discussion of the construction of the coefficients.

a model of potential fraud and abuse claims, it typically does not have an indicator in its data as to whether a particular claim was considered legitimate or possibly suspicious. That is, there is no dependent variable related to suspicion of a staged accident, exaggeration of damages, or other abuses. Instead, the company can use the data it does have to group the records into those that are similar to each other on a number of claimant, accident, provider, policyholder, geographic, and other variables. After records have been clustered by one of the procedures illustrated in this text, one can then examine the characteristic of the clusters. In the questionable claims example, it turns out that one of the clusters has a much higher proportion of records with attorney representation, soft tissue injury, use of a chiropractor or physical therapist, and history of a prior claim, whereas the other cluster has a much higher percentage of claimants who used the emergency room and filed a police report. As a result, we might label the first cluster as the “suspicious” cluster and the second cluster as “probably legitimate.”

Our examples with the California automobile insurance data also illustrate another way in which unsupervised learning can be used: to develop geographic groupings that can be based on important features of the data other than geographical proximity. Although our simple examples did not show it, we could have used geographic and demographic data in addition to loss data in the construction of geographic groupings. Christopherson and Werland (1996) and Weibel and Walsh (2008) provide more detailed illustrations. Weibel uses clustering along with mixed models in the development of territories. In our illustrations we ignored credibility considerations, and Weibel shows how to take them into account.

12.5 Exercises

Exercise 12.1. Apply both principal components and factor analysis to the simulated auto data in Example 12.3. Keep the first component and the first factor.

Exercise 12.2. Apply factor analysis to the economic and insurance indices data in Example 12.4. How many factors would you keep? How would you use these factors in an insurance pricing analysis?

Exercise 12.3. Using the primary paid loss and total allocated loss adjustment expense variables from the Texas data in Example 12.2, compute dissimilarity matrices using

- (a) the Euclidean measure,
- (b) the Manhattan distance,
- (c) standardized variables with the Manhattan distance.

Exercise 12.4. From the Texas data in Example 12.2, using the Euclidean dissimilarity measure and the Manhattan dissimilarity measure, standardized, choose $k = 2, 3, 4, 5$. Run cluster and silhouette plots. Which k would you select. Why?

Exercise 12.5. Using the injury and cause variables from the Texas data in Example 12.2, compute a dissimilarity matrix and perform k -means clustering using the gower measure. Make sure to convert the injury and cause variables to factors in R, using the `as.factor` function before clustering. Choose $k = 2, 3, 4, 5$. Which k would you choose?

Exercise 12.6. Use the R `mona` function to cluster the categorical variables in the simulated questionable claims data in Example 12.3. Output the clusters to a text file and compare actual to predicted clusters.

Exercise 12.7. Apply hierarchical clustering to the simulated questionable claims data in Example 12.3. Where do you think the best split is?

Exercise 12.8. Apply hierarchical clustering to the CAARP data in Example 12.1 using BI frequency PD frequency, BI severity, and PD severity. Where do you think the best split is?

Exercise 12.9. Use loss variables along with geographic variables (latitude, longitude, and elevation) to cluster the CAARP data in Example 12.1.

Exercise 12.10. Use the scores from Exercise 12.1 to predict cluster membership.

Exercise 12.11. Apply `fanny` to the questionable claims data in Example 12.3. Test 2, 3, and 4 clusters. Output the membership functions and examine them.

References

- Aldenderfer, M. S. and R. K. Blashfield (1984). *Cluster Analysis*. Sage Publications.
- Brockett, P. L., R. A. Derrig, L. L. Golden, A. Levine, and M. Alpert (2002). Fraud classification using principal component analysis of RIDITs. *Journal of Risk and Insurance* 69(3), 341–371.
- Brzezinski, J. R. (1981). Patterns in persistency. *Transactions* 33, 203.
- Christopherson, S. and D. Werland (1996). Using a geographic information system to identify territory boundaries. *Casualty Actuarial Society Forum*.
- Coaley, K. (2010). *An Introduction to Psychological Assessment and Psychometrics*. Sage Publications.
- Crawley, M. (2007). The R book.
- Erin Research (2003). Comparison analysis implications report of employer and member research. *Prepared for the Society of Actuaries*.
- Everitt, B., S. Landan, M. Leese, and D. Stahl (2011). *Cluster Analysis*. Wiley, New York.

- Francis, L. (2001). Neural networks demystified. *Casualty Actuarial Society Forum*, 253–320.
- Francis, L. (2003). Martian chronicles: Is MARS better than neural networks? *Casualty Actuarial Society Forum*, 75–102.
- Francis, L. (2006). Review of PRIDIT. *CAS Ratemaking Seminar*.
- Francis, L. and M. Flynn (2010). Text mining handbook. *Casualty Actuarial Society E-Forum, Spring 2010*, 1.
- Francis, L. and V. R. Prevosto (2010). Data and disaster: The role of data in the financial crisis. *Casualty Actuarial Society E-Forum, Spring 2010*, 62.
- Freedman, A. and C. Reynolds (2009). Cluster modeling: A new technique to improve model efficiency. *CompAct*. The Society of Actuaries.
- Gorden, R. L. (1977). *Unidimensional Scaling of Social Variables: Concepts and Procedures*. Free Press, New York.
- Kaufman, L. and P. J. Rousseeuw (1990). *Finding Groups in Data*. Wiley, New York.
- Kim, J.-O. and C. W. Mueller (1978). *Factor Analysis: Statistical Methods and Practical Issues*, Volume 14. Sage Publications, Thousand Oaks, CA.
- Maechler, M. (2012). Package “cluster” downloaded from CRAN website. www.r-project.org.
- Mahmoud, O. (2008). A multivariate model for predicting the efficiency of financial performance of property liability insurance companies. *CAS Discussion Paper Program*.
- Maranell, G. M. (1974). *Scaling: A Sourcebook for Behavioral Scientists*. Transaction Publishers.
- Oksanen, J. (2012). Cluster analysis: Tutorial with R. <http://www.statmethods.net/index.html>.
- Polon, J. (2008). Dental insurance fraud detection with predictive modeling. Presented at *Society of Actuaries Spring Health Meeting*.
- Sanche, R. and K. Lonergan (2006). Variable reduction for predictive modeling with clustering. *Casualty Actuarial Society Forum*, 89–100.
- Smith, L. (2002). A tutorial on principal components. www.sccg.sk/~haladova/principal_components.pdf.
- Struyf, A., M. Hubert, and P. Rousseeuw (1997). Clustering in an object-oriented environment. *Journal of Statistical Software* 1(4), 1–30.
- Venables, W. N., B. D. Ripley, and W. Venables (1999). *Modern Applied Statistics with S-PLUS*. Springer, New York.
- Weibel, E. J. and J. P. Walsh (2008). Territory analysis with mixed models and clustering. Presented at *Casualty Actuarial Society Spring Meeting*.
- Zhang, P., X. Wang, and P. X.-K. Song (2006). Clustering categorical data based on distance vectors. *Journal of the American Statistical Association* 101(473), 355–367.

Part III

Bayesian and Mixed Modeling

13

Bayesian Computational Methods

Brian Hartman

Chapter Preview. Bayesian methods have grown rapidly in popularity because of their general applicability, structured and direct incorporation of expert opinion, and proper accounting of model and parameter uncertainty. This chapter outlines the basic process and describes the benefits and difficulties inherent in fitting Bayesian models.

13.1 Why Bayesian?

Although the theory underpinning Bayesian methods is about 350 years old (Sir Thomas Bayes' essay was read to the Royal Society after his death; see Bayes and Price 1763), their widespread use was limited by computational power. In models of reasonable size and complexity, large iterated integrals need to be numerically approximated, which can be computationally burdensome. In the early 1990s the development of randomized algorithms to approximate these integrals, including the Gibbs sampler (Gelfand and Smith 1990), and the exponential growth of computing power made Bayesian methods accessible. More recently, with the development and growth of statistical software such as R, WinBUGS, and now PROC MCMC in SAS, Bayesian methods are available and employed by practitioners in many fields (see Albert 2009, for some details and examples). Actuaries have been using Bayesian methods since Whitney (1918) approximated them in a credibility framework. Unfortunately, many have been slow to extend them for use in other natural contexts. Among other benefits, Bayesian methods properly account for model and parameter uncertainty, provide a structure to incorporate expert opinion, and are generally applicable to many methods in this book and throughout the industry.

Actuaries are concerned about extreme events, such as large and rare claims. The predicted size and frequency of those claims are greatly affected by the uncertainty incorporated into the model. Commonly, a model is chosen from a set of candidate models, and parameters for that model are estimated using maximum likelihood. Implicit in this process is the assumption of certainty that both the model chosen and

the estimates of the parameters are correct. Bayesian methods provide an entire distribution for each parameter and probabilities for the various model choices, allowing explicit modeling of both the model and parameter uncertainty. Those uncertainties can have a large impact on the results of the model, especially on the risk measures. For a more thorough discussion of the problem and a few examples of the large effect, see Hartman and Heaton (2011).

Actuaries, like other professionals, use their expert judgment to both assess the validity of model results and adjust those results based on personal knowledge not explicitly incorporated in the model. These adjustments are often performed on an ad hoc basis. Under the Bayesian paradigm, the prior distribution provides a structure to incorporate those expert beliefs efficiently and consistently (see Section 13.5 for further discussion).

Finally, Bayes' framework is readily applicable to and could be beneficial for all the other methods in this book. See Section 13.7 for a list of references describing how to apply the methods in this book in a Bayesian setting.

13.2 Personal Automobile Claims Modeling

The methods presented throughout this chapter are applied to a dataset originally presented in De Jong and Heller (2008). This dataset contains one-year vehicle insurance policies taken out in 2004 or 2005. There are 67,856 total policies in the set, but this chapter focuses on the 4,624 policies (6.8%) that had at least one claim, and therefore a positive total claim amount. The dataset contains many covariates that could easily augment the models in this chapter using the methods in the next chapter. These policies often had exposures smaller than one, so the “annual claim amount” is the claim cost for the policy divided by the exposure. Figure 13.1 presents the kernel density estimates (KDEs) of the annual claim amount both in its original scale and after applying a logarithmic transformation. Notice that the original data are extremely right-skewed. Even after the log transformation some skewness remains. Additionally, there may be more than one mode in the transformed data. Basic summary statistics for both the original and transformed data are presented in Table 13.1.

13.3 Basics of Bayesian Statistics

In classical, or frequentist, statistics the parameters are assumed to be unknown but fixed. Hypothesis tests and confidence intervals are generated by answering the following question, “Assuming that the parameter value is x , what is the probability of obtaining sample statistics as extreme as the current sample?” Essentially, the result is compared to the distribution of all possible samples that could have been taken under certain assumptions.

Table 13.1. Summary Statistics for the Scaled Total Claims Data

Data	Min.	Q_1	Median	Mean	Q_3	Max.	St. Dev.
Original	200	575	1,415	11,080	4,071	5.24 M	119 K
Transformed	5.299	6.354	7.255	7.456	8.312	15.47	1.417

In Bayesian statistics, the parameters are random variables with a distribution. Before any data are observed, the prior distribution of the parameter is denoted $\pi(\theta)$, where θ is the parameter of interest with a range of Θ . After data are observed, the distribution of the parameter is updated by the likelihood of the data, $f(\mathbf{y}|\theta)$. The distribution of the parameter is updated by the data through Bayes' rule:

$$\pi(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{\int_{\Theta} f(\mathbf{y}|\theta)\pi(\theta)d\theta}. \quad (13.1)$$

The updated distribution, $\pi(\theta|\mathbf{y})$, is known as the posterior distribution. Often, especially in high-dimensional cases, the denominator in Bayes' rule has no closed-form expression. It is important to note that, with respect to θ , the denominator is simply a

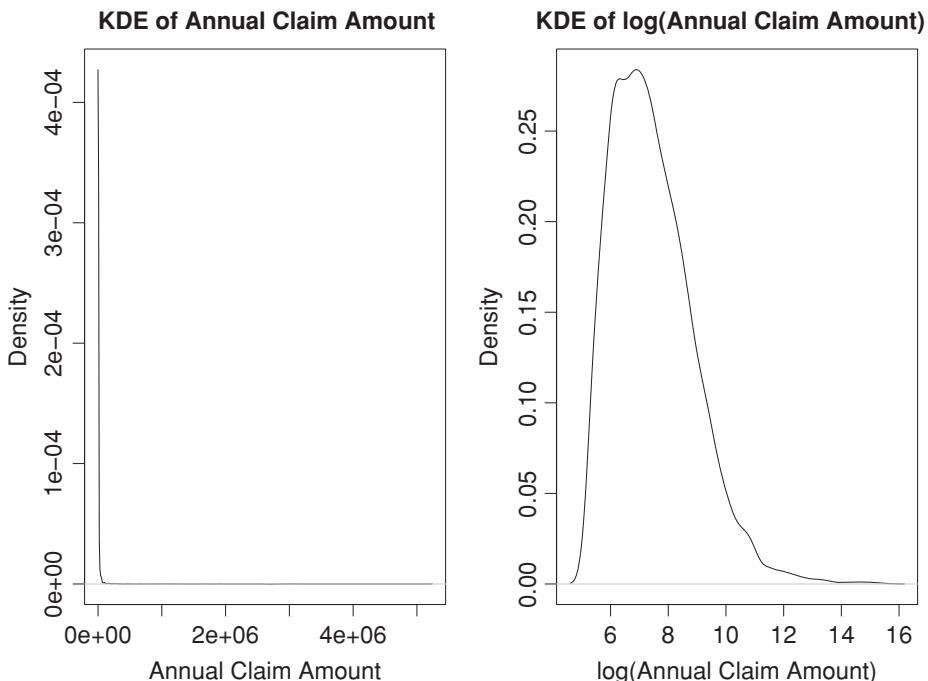


Fig. 13.1. Kernel density estimates of the annual claim amounts.

constant. Therefore,

$$\pi(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{\int_{\Theta} f(\mathbf{y}|\theta)\pi(\theta)d\theta} \propto f(\mathbf{y}|\theta)\pi(\theta). \quad (13.2)$$

All that is missing is a normalization constant (to make the density integrate to one). If $f(\mathbf{y}|\theta)\pi(\theta)$ is proportional to a known distribution, then the normalization will come from that known distribution. A simple example is presented here to illustrate the basic process.

Example 13.1. Assume that claims follow an exponential distribution with mean $1/\lambda$:

$$f(\mathbf{y}|\lambda) = \prod_{i=1}^n \lambda \exp \{-\lambda y_i\} = \lambda^n \exp \left\{ -\lambda \sum_{i=1}^n y_i \right\}. \quad (13.3)$$

The prior distribution of λ is chosen to be exponential as well, with mean $1/\kappa$:

$$\pi(\lambda) = \kappa \exp \{-\kappa \lambda\} \quad (13.4)$$

Using Bayes' rule, the posterior distribution is

$$\pi(\lambda|\mathbf{y}) \propto f(\mathbf{y}|\lambda)\pi(\lambda) = \lambda^n \exp \left\{ -\lambda \sum_{i=1}^n y_i \right\} \times \kappa \exp \{-\kappa \lambda\} \quad (13.5)$$

$$\propto \lambda^n \exp \left\{ -\lambda \left(\sum_{i=1}^n y_i + \kappa \right) \right\} \quad (13.6)$$

which is proportional to the following gamma distribution:

$$\propto \frac{1}{(1/\sum_{i=1}^n y_i + \kappa)^{n+1}} \Gamma(n+1) \lambda^{(n+1)-1} \exp \left\{ \frac{-\lambda}{1/(\sum_{i=1}^n y_i + \kappa)} \right\} \quad (13.7)$$

$$\lambda|\mathbf{y} \sim \Gamma \left(n+1, \frac{1}{\sum_{i=1}^n y_i + \kappa} \right). \quad (13.8)$$

Therefore, only the sum and size of the sample are needed to update the prior distribution to the posterior distribution in this case.

When the prior and posterior distributions are of the same family, they are known as conjugate priors. A list of conjugate priors and their accompanying parameters is available on Wikipedia (2012).

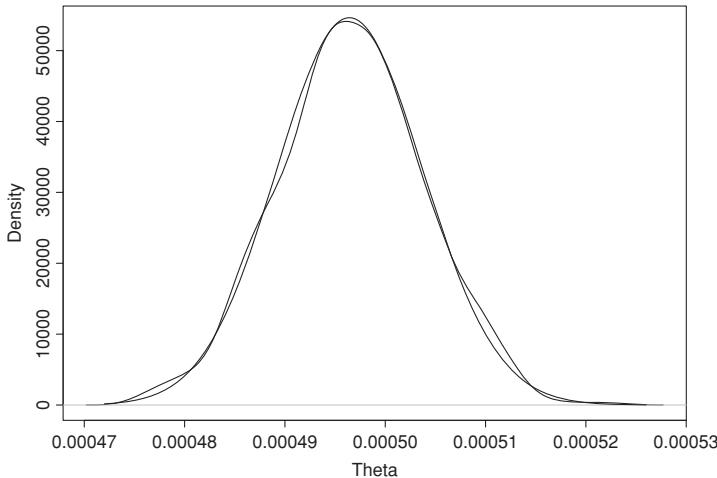


Fig. 13.2. Kernel density estimate of λ .

A simple BUGS script with the following model statement can apply the above conjugate model to the total claims data:

```
model {
    for (i in 1:N) { x[i] ~ dexp(Lambda) }
    Lambda ~ dexp(1)
}
```

Notice that the posterior distribution is not included in the code. BUGS, or its Windows-based version WinBUGS, will take the prior(s) and likelihood and design a sampler to obtain draws from the posterior distribution automatically. Figure 13.2 shows a kernel density estimate of the draws from the posterior distribution generated by the above code and the true posterior distribution calculated above. As expected, the two curves are very similar.

R and WinBUGS code for all the analyses and figures in this chapter is available in the online supplement. The package R2WinBUGS (Sturtz, Ligges, and Gelman 2005) provides the ability to run WinBUGS from R and is used extensively in the code for this chapter. For more information on Bayesian modeling using WinBUGS, see Lunn et al. (2000), Ntzoufras (2011), or Scollnik (2001).

13.4 Computational Methods

Often, the priors are not jointly conjugate. In that case, the posterior distribution of the parameters will not have a closed-form solution and will need to be numerically

estimated. The class of numerical methods that use Markov chains to generate draws from a posterior distribution is collectively known as Markov chain Monte Carlo (MCMC) methods. The two most common methods are the Gibbs sampler (Gelfand and Smith 1990) and the Metropolis-Hastings sampler (Hastings 1970; Metropolis et al. 1953).

13.4.1 Gibbs Sampler

When numerically approximating the normalization constant, often the integral is of high dimension. Gelfand and Smith (1990) showed that, under mild regularity conditions (Besag 1974), the joint estimates of the parameters can be obtained by sequentially estimating each of the parameters, holding all others constant. The distribution of a parameter, holding all other parameters constant, is known as a full conditional distribution. If it is possible to sample from full conditional distribution, the parameters can be drawn directly. Otherwise the parameters can be drawn indirectly using adaptive rejection sampling (Gilks and Wild 1992) or the Metropolis-Hastings sampler described in the next subsection.

Specifically, if the model contains k parameters $(\theta_1, \theta_2, \dots, \theta_k)$, start with arbitrary starting values $(\theta_2^{(0)}, \theta_3^{(0)}, \dots, \theta_k^{(0)})$, and for $t = 1, \dots, T$ perform the following steps:

- Step 1: Draw $\theta_1^{(t)}$ from $\pi(\theta_1 | \theta_2^{(t-1)}, \theta_3^{(t-1)}, \dots, \theta_k^{(t-1)}, \mathbf{y})$
- Step 2: Draw $\theta_2^{(t)}$ from $\pi(\theta_2 | \theta_1^{(t)}, \theta_3^{(t-1)}, \dots, \theta_k^{(t-1)}, \mathbf{y})$
- Step 3: Draw $\theta_3^{(t)}$ from $\pi(\theta_3 | \theta_1^{(t)}, \theta_2^{(t)}, \theta_4^{(t-1)}, \dots, \theta_k^{(t-1)}, \mathbf{y})$
- ⋮
- Step k: Draw $\theta_k^{(t)}$ from $\pi(\theta_k | \theta_1^{(t)}, \theta_2^{(t)}, \dots, \theta_{k-1}^{(t)}, \mathbf{y})$.

The sampler will take some time to move from the starting values to the higher density areas of the posterior distribution and then will take more time to fully explore that distribution. The draws are independent of the starting values when the first n_1 iterations, known as the burn-in period, are discarded. The burn-in period allows the sampler enough time to move from the starting values and start exploring the posterior distribution. The next n_2 draws are used to estimate the posterior distribution and to calculate any quantities of interest. Considerations for deciding the sizes of n_1 and n_2 are discussed later in this section.

13.4.2 Metropolis-Hastings Sampler

Even after holding the other parameters constant, the full conditional distribution may not be of a common form. As such, samples from the distribution cannot be

directly sampled. Luckily, the Gibbs sampler does not require that the full conditional distribution is known, but only that draws from that distribution can be obtained. The Metropolis-Hastings sampler (Hastings 1970; Metropolis et al. 1953) provides those draws.

Assume a current value of the parameter ($\theta_1^{(t-1)}$) and a function (proportional to the full conditional distribution)

$$h(\theta_1) = \pi(\theta_1 | \theta_2, \dots, \theta_k) \pi(y | \theta_1, \dots, \theta_k) \propto \pi(\theta_1 | y, \theta_2, \dots, \theta_k). \quad (13.9)$$

Propose a new value, θ_1^* , drawn from the candidate density $q(\theta_1^* | \theta_1^{(t-1)})$. Next, calculate the acceptance ratio r using

$$r = \frac{h(\theta_1^*) q(\theta_1^{(t-1)} | \theta_1^*)}{h(\theta_1^{(t-1)}) q(\theta_1^* | \theta_1^{(t-1)})}. \quad (13.10)$$

With probability r , accept the proposed value ($\theta_1^{(t)} = \theta_1^*$), and with probability $1 - r$ reject the proposed value and use the current value ($\theta_1^{(t)} = \theta_1^{(t-1)}$). If the proposal distribution is symmetric, then $q(\theta_1^{(t-1)} | \theta_1^*) = q(\theta_1^* | \theta_1^{(t-1)})$, and those terms cancel out, leaving

$$r = \frac{h(\theta_1^*)}{h(\theta_1^{(t-1)})}. \quad (13.11)$$

This is the Metropolis algorithm, of which the Metropolis-Hastings algorithm is an extension. Most commonly, the proposal distribution is a normal distribution centered at the previous value

$$q(\theta^* | \theta^{(t-1)}) = N(\theta^* | \theta^{(t-1)}, \sigma^2). \quad (13.12)$$

In that common case, the only parameter to select is the σ^2 . It may seem desirable to have a value of σ^2 that will make the acceptance probability close to one. In that case, very few draws are wasted, but that is not the most effective selection. Taken to the extreme and trivial case, if $\sigma^2 = 0$ then $\theta^* = \theta^{(t-1)}$ and $h(\theta^*) = h(\theta^{(t-1)})$ so the acceptance probability will always equal 1, but all of the draws will be the same. With small nonzero values of σ^2 , you may have high acceptance probabilities, but the chain will travel slowly around the posterior distribution and the draws will be highly autocorrelated. In contrast, if σ^2 is too large, then the acceptance probabilities will be very low, and the chain will tend to stick (a long series of rejected proposals) and again produce autocorrelated draws. For certain models, the optimal acceptance rates are available (e.g., Gelman, Meng, and Stern 1996), but a general rule of thumb is that the acceptance rate should be somewhere between 20% and 50%. In addition to tracking the acceptance rate, a plot of the value of the parameter against its index in the chain, called a trace plot, can show how well the chain is exploring the posterior

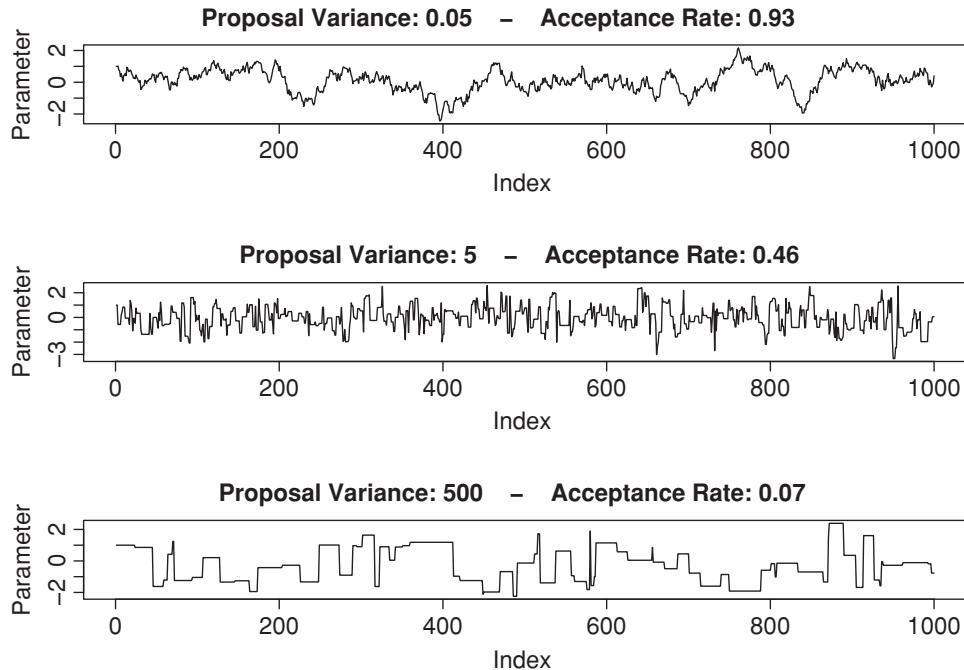


Fig. 13.3. Trace plots of various proposal variances.

distribution. Figure 13.3 shows three trace plots at different values of σ^2 – the first too small, the last too large, and the middle near the theoretical optimum.

In modern applications, the proposal variance is chosen and updated adaptively. For example, monitor the acceptance rate, and if the rate is too low decrease the proposal variance. If the rate is too high, increase the variance. These adaptive proposals can affect the convergence of the sampler. The simplest way to avoid that concern is to adapt the proposal variance only during the burn-in period. WinBUGS does that automatically during the burn-in period for those parameters that require a Metropolis-Hastings sampler.

13.4.3 Convergence Metrics

How many samples should be discarded with the burn-in? Has the chain converged to the posterior distribution? These are difficult questions to which there are no general answers. There are a number of diagnostics that give information about various parts of the chain, but no one metric is perfectly general. The most commonly used convergence metric is the Gelman and Rubin (1992) statistic. To calculate the statistic, run a small number, m , of chains from varying starting points, each for $2N$ iterations. The starting points need to be overdispersed with relation to the posterior distribution,

and the authors give some suggestions on how to define those starting points. The statistic determines whether the variance within the chains for θ (the parameter of interest) is approximately equal to the variance across the chains for the last N iterations. The within-chain variance is

$$W = \frac{1}{m(N-1)} \sum_{j=1}^m \sum_{i=1}^N (\theta_{ij} - \bar{\theta}_{.j})^2, \quad (13.13)$$

where θ_{ij} is the i^{th} draw of θ in chain j and $\bar{\theta}_{.j}$ is the mean of θ from chain j . The between-chain variance is

$$B = \frac{N}{m-1} \sum_{j=1}^m (\bar{\theta}_{.j} - \bar{\theta}_{..})^2, \quad (13.14)$$

where $\bar{\theta}_{..}$ is the overall mean of θ from all the chains. The Gelman-Rubin statistic is

$$\sqrt{\hat{R}} = \sqrt{\left(\frac{N-1}{N} + \frac{m+1}{mN} \frac{B}{W} \right) \frac{df}{df-2}}, \quad (13.15)$$

where df is the degrees of freedom from a t -distribution that approximates the posterior distribution. As N approaches infinity, $\sqrt{\hat{R}}$ must approach 1. This statistic is readily available in R, SAS, and WinBUGS.

Although the Gelman-Rubin approach is valuable, it has been criticized on three counts. First, it is a univariate statistic that therefore must be applied to all the parameters in the model individually and does not incorporate the correlation among the parameters. Second, the approach only focuses on the bias component of the estimates and disregards the variance of the estimates. Finally, the method depends on the overdispersion of the starting values with respect to the true posterior, which if known, there would be no need to estimate the parameters at all!

In summary, various statistics and plots can help determine whether the chain has converged and give suggestions for improvement. Although none of these items can ensure convergence, they can find evidence that the chains have not yet converged. The following statistics may help.

- *Gelman-Rubin statistic:* Even with the shortcomings mentioned earlier, the Gelman-Rubin statistic is easy to calculate with modern statistical packages and provides a simple look into the convergence of the series.
- *Lag-1 Autocorrelation:* High autocorrelation could be a sign of poor mixing. The autocorrelation can also help diagnose a large Gelman-Rubin statistic, either due to poor mixing or possibly multimodality (if the autocorrelation is insignificant).
- *Trace Plots:* For each parameter, overlay the trace plots of all your chains (at least three to five) on the same plot. If the initial values are overdispersed, the plots should start

from different locations. It is hoped that, they will all migrate to the same area and bounce around there.

- *Cross-correlations:* Calculate the correlations between the different parameters. A chain with highly correlated variables may be improved by proposing and sampling the correlated variables together.

Finally, remember to still be an analyst. Using the mean of the posterior distributions for the parameters, compare the density of the fitted model to the original data. Be sure the results make sense. When using a poorly fitting model, the chain may still converge and all the above metrics could give satisfactory results. Obviously, that model should not be used.

13.4.4 Total Claims Data

No longer constrained to use single-parameter models or conjugate priors, the following model can be fit to the total claims data.

$$(\mathbf{y}|\alpha, \beta) \sim \Gamma(\alpha, \beta) \quad (13.16)$$

$$\pi(y_i|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y_i^{\alpha-1} e^{-\beta y_i} \quad (13.17)$$

$$E[Y] = \alpha/\beta \quad (13.18)$$

$$\text{Var}[Y] = \alpha/\beta^2 \quad (13.19)$$

$$\alpha \sim \text{Unif}(0, 1000) \quad (13.20)$$

$$\beta \sim \text{Unif}(0, 1000). \quad (13.21)$$

The choice of prior distributions is explained in the next section. As a pilot run, five simultaneous chains are initialized from various points and updated 20,000 times. The Gelman-Rubin statistic is calculated for both α and β , using only the iterations from one to k . The statistic is then plotted against k (see Figure 13.4) along with an upper limit to determine the length of the burn-in period.

In properly designed MCMC chains, the Gelman-Rubin statistic will converge to 1 as k increases to infinity. Practically speaking, as long as the chain appears to be converging to 1, the chains can be assumed to have reached the posterior distribution when the Gelman-Rubin statistic is less than about 1.1.

In the annual claims data, the statistic drops down close to one around 10,000 iterations. That seems to be a good burn-in length to use. In addition, the trace plots (see Figure 13.5) show that the chains move very quickly (within a few hundred observations) from the initial values to the range of the posterior distribution, so 10,000 iterations should be more than enough.

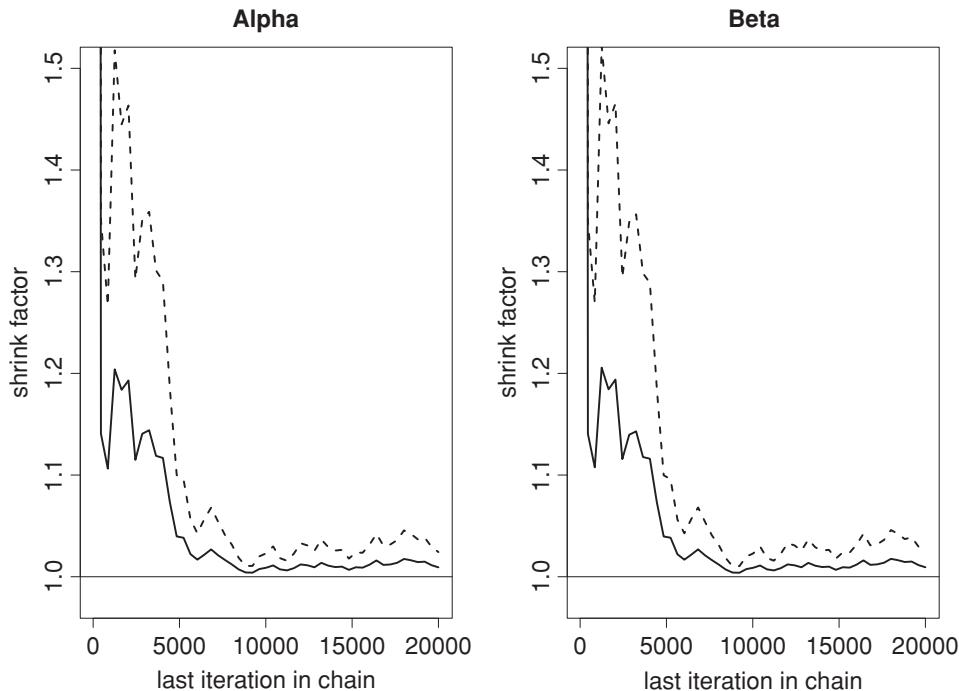


Fig. 13.4. Gelman-Rubin plots for α and β . The solid line is the estimate, and the dashed line is an upper confidence limit. Note: When k is small the statistic is near 15 (far above the limits of this plot). Stretching the limits to display those points would make it hard to distinguish the details of the rest of the plot.

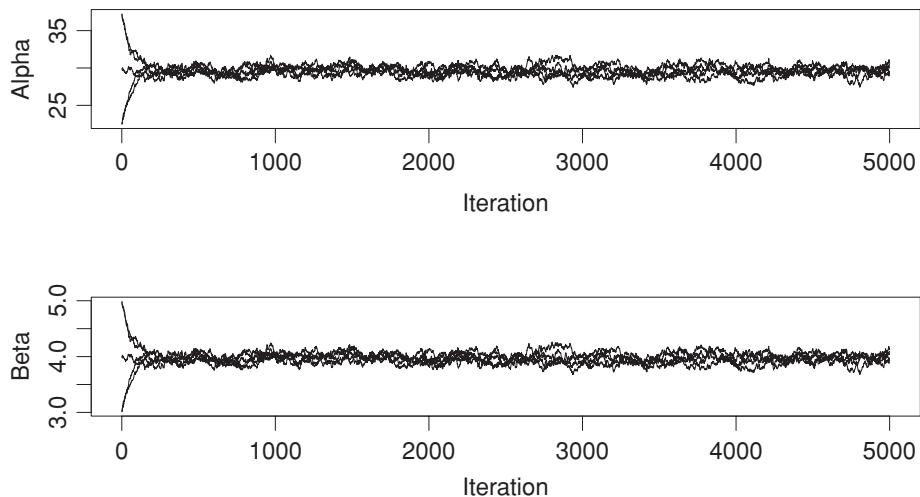


Fig. 13.5. Trace plots for α and β .

Table 13.2. Autocorrelation in the Pilot Run for Various Lags

Lag	α	β
0	1.000	1.000
1	0.9828	0.9828
5	0.9162	0.9163
10	0.8378	0.8377
50	0.4034	0.4033
250	0.0256	0.0257
500	-0.0104	-0.0102

Notice also that both plots for the two parameters are very similar, due to the high cross-correlation between the two parameters. The autocorrelations for various lags in the pilot run are in Table 13.2.

The autocorrelation is very high. The value for lag-250 is acceptable, so the sampler will record only every 250th iteration. This is unusually high. The autocorrelation in most chains is much smaller, but luckily this is a relatively simple model, allowing the computational time to still be palatable even with the excessive thinning. The chosen sampler contains five chains, each with a burn-in of 10,000 iterations and 150,000 samples (recording every 250th). That will provide 3,000 total draws from the posterior distribution. The trace plots (see Figure 13.5) for the pilot run are satisfactory. Summary statistics for the posterior distributions of α and β are available in Table 13.3.

13.5 Prior Distributions

13.5.1 Prior Elicitation

As mentioned in the introduction, one of the biggest advantages of the Bayesian paradigm is the ability to consistently include prior beliefs through the prior distribution. It can be a little difficult to translate an expert's opinion into a distribution function, $\pi(\theta)$. A common approach is to have the expert assign probabilities to various portions of the parameter space. Assume that claim amounts follow a normal distribution. A prior distribution for the mean could be elicited by asking questions of

Table 13.3. Summary Statistics of the Posterior Distributions

Parameter	Q_1	Median	Mean	Q_3	St. Dev.
α	29.30	29.71	29.71	30.12	0.607
β	3.929	3.984	3.984	4.039	0.082

the expert, such as, “What is the chance that the mean claim amount is between \$100 and \$1,000?” The resulting prior distribution can be as specific as desired, based on the number of questions asked.

It is important to note that the range of the posterior distribution is limited by the range of the prior distribution. If there is no probability assigned to a certain region in the prior distribution, that region will also have no probability in the posterior distribution, no matter what the data indicate. Alternatively, a standard distribution with the same range as the parameter of interest could be used instead of a histogram. Unfortunately, it may be difficult to translate the expert opinion into the distribution. It is rather unproductive to ask, “What parameters of a gamma distribution best describe your current feelings about the mean loss?” One option is to ask for summary statistics such as the mean and variance and then translate those answers into parameters of a distribution.

When dealing with human opinion, there is also a natural tendency to overstate confidence in a conclusion. This tendency is known as the overconfidence effect (Pallier et al. 2002). In a well-known example (Adams and Adams 1960), participants were asked to spell difficult words and then state how confident they were in the spelling. When they said they were “100% confident” they misspelled the word 20% of the time. Therefore, when experts are defining prior probabilities, they will tend to put too much probability in the center of the distribution and not enough in the tails. Even after they are warned and educated about the bias, overconfidence still exists (Alpert and Raiffa, 1982). Therefore, to properly include expert opinion in the prior, density may need to be taken from the center of the distribution and moved to the tails.

13.5.2 Noninformative Priors

Strong priors are not always preferred. There may not be any prior information about a particular parameter. Alternatively, with all the inherent concerns about overconfidence, a more objective approach may be preferred even when expert opinion is available. Unfortunately, an objective or noninformative prior can be hard to define. With a finite sample space, a natural choice is to divide the prior density evenly throughout the space. For example, if the parameter is only defined on the integers between 1 and N , then a prior of the form

$$\pi(\theta) = \frac{1}{N} \quad \theta \in \{1, 2, \dots, N\} \quad (13.22)$$

would be noninformative. Also, if the parameter is continuous over a bounded interval, say $[a, b]$, then

$$\pi(\theta) = \frac{1}{b-a} \quad \theta \in [a, b] \quad (13.23)$$

would be noninformative. It is important to note that a uniform prior is not invariant to parameterization. If a uniform prior is desired, be sure to choose the most likely parameterization and set up the uniform prior there.

Beyond the transformation problem, uniform priors on an unbounded interval are improper (i.e., $\int_{\Theta} \pi(\theta) d\theta = \infty$). Under these priors, if the integral of the likelihood function is improper with respect to the parameter θ , then the posterior will be improper. An improper posterior may be hard to detect because the computational methods discussed in this chapter will still return values, but the results can be completely nonsensical. Because of that concern, it is best to use a proper prior. For example, if the parameter of interest is defined over the entire real line, a normal prior with a large variance and a uniform prior over the real line will appear and behave similarly.

A more general formulation of a noninformative prior is the Jeffreys' prior (Jeffreys 1961) and is defined as

$$\pi(\theta) \propto [I(\theta)]^{1/2}, \quad (13.24)$$

where $I(\theta)$ is the Fisher information of the likelihood,

$$I(\theta) = -E \left[\frac{\partial^2}{\partial \theta^2} \log f(y|\theta) \right]. \quad (13.25)$$

The Jeffreys prior is a strong default choice when no prior information is known about the parameters. Additionally, this prior is invariant to transformation. For a more comprehensive review of constructing priors, please Berger (1985, chapter 3).

13.5.3 Prior Sensitivity

It is important to understand the effect of the prior distribution on the posterior distribution of the parameters. This is especially true when the sample size is small, allowing the prior to have a larger impact. When the prior is conjugate, the effect of adjusting the hyperparameters can be seen directly in the posterior distribution. When numerical methods are required to estimate the posterior distribution, it can be computationally difficult to test a wide variety of hyperparameter settings, but doing so is necessary to truly understand the effect of the prior on the overall results.

To exemplify the effect that a prior can have on the posterior distribution, three different models are applied to the annual claims data. In all cases, the prior for β is noninformative ($\beta \sim Unif(0, 1000)$), and the likelihood is

$$\pi(y_i|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y_i^{\alpha-1} e^{-\beta y_i}. \quad (13.26)$$

Table 13.4. Summary Statistics of the Posterior Distributions of α

Prior	Q_1	Median	Mean	Q_3	St. Dev.
1) Non-informative	29.30	29.71	29.71	30.12	0.607
2) Mildly informative	29.27	29.68	29.69	30.08	0.614
3) Strongly informative	30.26	30.58	30.58	30.91	0.477

The three priors for α compared are

- (1) $\alpha \sim \text{Unif}(0, 1000)$ (noninformative)
- (2) $\alpha \sim \text{Exp}(1/32)$ (mildly informative)
- (3) $\alpha \sim \Gamma(1600, 50)$ (strongly informative)

Notice that both the second and third priors have the same mean ($E(\alpha) = 32$), but wildly different variances ($\text{Var}(\alpha) = 1024$ for prior 2 and $\text{Var}(\alpha) = 0.64$ for prior 3).

Plots of the resulting posterior distributions for α are presented in Figure 13.6 and summary statistics are available in Table 13.4. The posterior distributions for

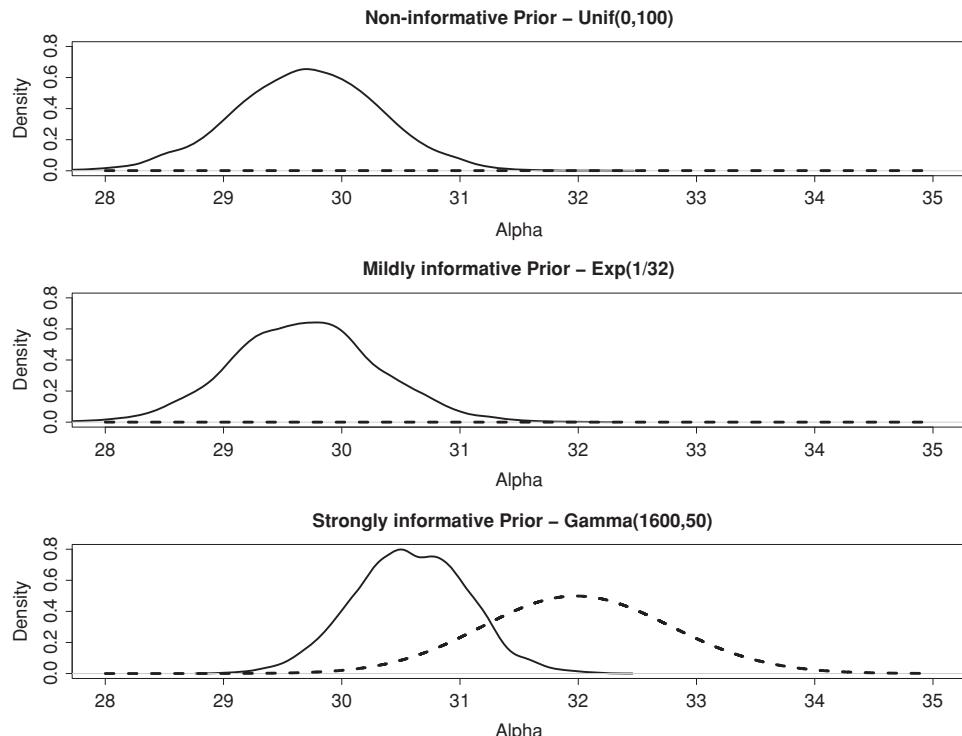


Fig. 13.6. Prior (dashed line) and posterior (solid line) densities for the three different prior specifications.

the mildly informative and non-informative priors are very similar, but the strongly informative prior has a significant effect on the posterior. The value of Q_1 for prior three is greater than the values of Q_3 for priors one and two. If the prior is strong enough, it can even overwhelm large amounts of data (remember $n = 4,624$ in the auto claims data).

13.6 Conclusion

With the rapid increase in computational power, Bayesian methods are far more accessible to actuaries. The methods provide a structure to incorporate prior expert knowledge and properly account for model and parameter uncertainty. The methods are also very natural for actuaries who already use credibility theory and are widely applicable to essentially any model.

This chapter serves as an introduction to the basics of Bayesian statistics, but is limited in scope. There are a number of well-written books devoted entirely to Bayesian statistics that are available for a more thorough introduction (e.g., Carlin and Louis 2009; Congdon 2003; Gelman et al. 2004).

13.7 Further Reading

As mentioned in the introduction, all of the models in this book can be formulated and estimated from a Bayesian perspective. Although there is not sufficient space to elaborate on each model individually, the following list provides several references on how to apply the other models in a Bayesian setting:

- Multiple linear regression – Chapter 14 in this text
- Regression with categorical dependent variables – Chen, Ibrahim, and Yiannoutsos (1999), Genkin, Lewis, and Madigan (2007)
- Regression with count dependent variables – Clyde (2000); El-Sayyad (1973)
- Generalized linear models – Dey, Ghosh, and Mallick (2000); Zeger and Karim (1991)
- Frequency/severity models – Huang, Chin, and Haque (2008); Ma, Kockelman, and Damien (2008); Neil, Fenton, and Tailor (2005)
- Mixed models – Littell (2006); Ronquist and Huelsenbeck (2003)
- Generalized additive models – Hastie and Tibshirani (1990); Wood (2006)
- Fat-tail regression models – Fernández and Steel (1998); Jacquier, Polson, and Rossi (2004)
- Spatial statistics – Eberly et al. (2000); Song et al. (2006)
- Machine learning – Andrieu et al. (2003); Friedman et al. (2000)
- Time series, including Lee-Carter – Czado, Delwarde, and Denuit (2005); Pedroza (2006)
- Longitudinal and panel data models – Cowles, Carlin, and Connell (1996); Daniels and Pourahmadi (2002); Laird and Ware (1982)

- Credibility – Klugman (1992); Makov, Smith, and Liu (1996); Young (1998)
- Survival models – Ibrahim, Chen, and Sinha (2005); Kottas (2006); Mallick, Denison, and Smith (1999)
- Claims triangles – De Alba (2006); Zhang, Dukic, and Guszczca (2012)

Acknowledgments

I would like to thank the three editors and four anonymous referees for their thoughtful and insightful comments. I am also grateful for discussions with and suggestions from Brad Barney, Jenn Hartman, Matt Heaton, and Shujuan Huang. Their work and suggestions greatly improved the final product.

References

- Adams, P. and J. Adams (1960). Confidence in the recognition and reproduction of words difficult to spell. *American Journal of Psychology* 73(4), 544–552.
- Albert, J. (2009). *Bayesian Computation with R*. Springer, New York.
- Alpert, M. and H. Raiffa (1982). A progress report on the training of probability assessors. In A. T. Daniel Kahneman and Paul Slovic (ed.), *Judgment under Uncertainty: Heuristics and Biases*, pp. 294–305. Cambridge University Press, Cambridge.
- Andrieu, C., N. De Freitas, A. Doucet, and M. Jordan (2003). An introduction to MCMC for machine learning. *Machine Learning* 50(1), 5–43.
- Bayes, T. and M. Price (1763). An essay towards solving a problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. *Philosophical Transactions* (1683–1775), 370–418.
- Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, New York.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 192–236.
- Carlin, B. and T. Louis (2009). *Bayesian Methods for Data Analysis*, Volume 78. Chapman & Hall/CRC, Boca Raton, FL.
- Chen, M., J. Ibrahim, and C. Yiannoutsos (1999). Prior elicitation, variable selection and Bayesian computation for logistic regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61(1), 223–242.
- Clyde, M. (2000). Model uncertainty and health effect studies for particulate matter. *Environmetrics* 11(6), 745–763.
- Congdon, P. (2003). *Applied Bayesian Modelling*, Volume 394. Wiley, New York.
- Cowles, M., B. Carlin, and J. Connell (1996). Bayesian Tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness. *Journal of the American Statistical Association*, 86–98.
- Czado, C., A. Delwarde, and M. Denuit (2005). Bayesian Poisson log-bilinear mortality projections. *Insurance: Mathematics and Economics* 36(3), 260–284.
- Daniels, M. and M. Pourahmadi (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika* 89(3), 553–566.
- De Alba, E. (2006). Claims reserving when there are negative values in the runoff triangle: Bayesian analysis using the three-parameter log-normal distribution. *North American Actuarial Journal* 10(3), 45.

- De Jong, P. and G. Heller (2008). *Generalized Linear Models for Insurance Data*. Cambridge University Press, Cambridge.
- Dey, D., S. Ghosh, and B. Mallick (2000). *Generalized Linear Models: A Bayesian Perspective*, Volume 5. CRC, Boca Raton, FL.
- Eberly, L., B. Carlin, et al. (2000). Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Statistics in Medicine* 19(1718), 2279–2294.
- El-Sayyad, G. (1973). Bayesian and classical analysis of Poisson regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 445–451.
- Fernández, C. and M. Steel (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 359–371.
- Friedman, N., M. Linial, I. Nachman, and D. Pe'er (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology* 7(3–4), 601–620.
- Gelfand, A. E. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85(410), 398–409.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL.
- Gelman, A., X. Meng, and H. Stern (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* 6, 733–759.
- Gelman, A. and D. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* 7(4), 457–472.
- Genkin, A., D. Lewis, and D. Madigan (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics* 49(3), 291–304.
- Gilks, W. and P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 41(2), 337–348.
- Hartman, B. M. and M. J. Heaton (2011). Accounting for regime and parameter uncertainty in regime-switching models. *Insurance: Mathematics and Economics* 49(3), 429–437.
- Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Chapman & Hall/CRC, Boca Raton, FL.
- Hastings, W. K. (1970, April). Monte Carlo methods using Markov chains and their applications. *Biometrika* 57(1), 97–109.
- Huang, H., H. Chin, and M. Haque (2008). Severity of driver injury and vehicle damage in traffic crashes at intersections: A Bayesian hierarchical analysis. *Accident Analysis & Prevention* 40(1), 45–54.
- Ibrahim, J., M. Chen, and D. Sinha (2005). *Bayesian Survival Analysis*. Wiley Online Library.
- Jacquier, E., N. Polson, and P. Rossi (2004). Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *Journal of Econometrics* 122(1), 185–212.
- Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press.
- Klugman, S. (1992). *Bayesian Statistics in Actuarial Science: With Emphasis on Credibility*, Volume 15. Springer, New York.
- Kottas, A. (2006). Nonparametric Bayesian survival analysis using mixtures of Weibull distributions. *Journal of Statistical Planning and Inference* 136(3), 578–596.
- Laird, N. and J. Ware (1982). Random-effects models for longitudinal data. *Biometrics*, 963–974.
- Littell, R. (2006). *SAS for Mixed Models*. SAS Publishing, Cary, NC.
- Lunn, D., A. Thomas, N. Best, and D. Spiegelhalter (2000). Winbugs—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* 10(4), 325–337.
- Ma, J., K. Kockelman, and P. Damien (2008). A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis & Prevention* 40(3), 964–975.

- Makov, U., A. Smith, and Y. Liu (1996). Bayesian methods in actuarial science. *The Statistician*, 503–515.
- Mallick, B., D. Denison, and A. Smith (1999). Bayesian survival analysis using a MARS model. *Biometrics* 55(4), 1071–1077.
- Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 21, 1087–1091.
- Neil, M., N. Fenton, and M. Tailor (2005). Using Bayesian networks to model expected and unexpected operational losses. *Risk Analysis* 25(4), 963–972.
- Ntzoufras, I. (2011). *Bayesian Modeling using WinBUGS*, Volume 698. Wiley, New York.
- Pallier, G., R. Wilkinson, V. Danthiir, S. Kleitman, G. Knezevic, L. Stankov, and R. Roberts (2002). The role of individual differences in the accuracy of confidence judgments. *Journal of General Psychology* 129(3), 257–299.
- Pedroza, C. (2006). A Bayesian forecasting model: Predicting US male mortality. *Biostatistics* 7(4), 530–550.
- Ronquist, F. and J. Huelsenbeck (2003). Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12), 1572–1574.
- Scollnik, D. (2001). Actuarial modeling with MCMC and BUGS. *Actuarial Research Clearing House: ARCH* (2), 433.
- Song, J., M. Ghosh, S. Miaou, and B. Mallick (2006). Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis* 97(1), 246–273.
- Sturtz, S., U. Ligges, and A. Gelman (2005). R2winbugs: A package for running WinBUGS from R. *Journal of Statistical Software* 12(3), 1–16.
- Whitney, A. W. (1918). The theory of experience rating. *Proceedings of the Casualty Actuarial Society* 4, 274–292.
- Wikipedia (2012). Conjugate prior – Wikipedia, the free encyclopedia. Accessed May 19, 2012.
- Wood, S. (2006). *Generalized Additive Models: An Introduction with R*, Volume 66. CRC Press, Boca Raton, FL.
- Young, V. (1998). Robust Bayesian credibility using semiparametric models. *ASTIN Bulletin* 28, 187–204.
- Zeger, S. and M. Karim (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, 79–86.
- Zhang, Y., V. Dukic, and J. Guszczza (2012). A Bayesian non-linear model for forecasting insurance loss payments. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

14

Bayesian Regression Models

Luis E. Nieto-Barajas and Enrique de Alba

Chapter Preview. In this chapter we approach many of the topics of the previous chapters, but from a Bayesian viewpoint. Initially we cover the foundations of Bayesian inference. We then describe the Bayesian linear and generalized regression models. We concentrate on the regression models with zero-one and count response and illustrate the models with real datasets. We also cover hierarchical prior specifications in the context of mixed models. We finish with a description of a semi-parametric linear regression model with a nonparametric specification of the error term. We also illustrate its advantage with respect to the fully parametric setting using a real dataset.

14.1 Introduction

The use of Bayesian concepts and techniques in actuarial science dates back to Whitney (1918) who laid the foundations for what is now called empirical Bayes credibility. He noted that the solution of the problem “depends upon the use of inverse probabilities.” This is the term used by T. Bayes in his original paper (e.g., Bellhouse 2004). However, Ove Lundberg was apparently the first one to realize the importance of Bayesian procedures (Lundberg 1940). In addition, Bailey (1950) put forth a clear and strong argument in favor of using Bayesian methods in actuarial science. To date, the Bayesian methodology has been used in various areas within actuarial science; see, for example, Klugman (1992), Makov (2001), Makov, Smith, and Liu (1996), and Scolnik (2001). A brief look at recent issues of the main journals in the field shows that Bayesian applications appear regularly and cover a broad range of actuarial topics; for example, mortality modeling (Cairns et al. 2011); extreme observations (Cabras and Castellanos 2011); mixture models (Bernardi, Maruotti, and Petrella 2012); premium policies (Landriault, Lemieux, and Willmot 2012); and loss reserving (Shi, Basu, and Meyers 2012).

Bayesian methods have several advantages that make them appealing for use in actuarial science. First, they allow the actuary to formally incorporate expert or existing prior information. This prior information can be in the form of global or industry-wide information (experience) or in the form of tables. In this respect it is indeed surprising that Bayesian methods are not used more extensively, because there is a wealth of “objective” prior information available to the actuary. In fact, the “structure distribution” frequently used in credibility was originally formulated in a Bayesian framework (Bühlmann 1967).

The second advantage of Bayesian methods is that the analysis is always done by means of the complete probability distribution for the quantities of interest, either the parameters or the future values of a random variable. Actuarial science is a field where adequate understanding and knowledge of the complete distribution are essential. In addition to expected values we are usually looking at certain characteristics of probability distributions (e.g., ruin probability, extreme values, value-at-risk (VaR), and so on).

From a theoretical point of view, Bayesian methods have an axiomatic foundation and are derived from first principles (Bernardo and Smith 2000). From a practical perspective, Bayesian inference is the process of fitting a probability model to a set of data and summarizing the uncertainty by a probability distribution on the parameters of the model and on unobserved quantities, such as predictions for new observations. A fundamental feature of Bayesian inference is the direct quantification of uncertainty. To carry it out, the actuary must set up a full probability model (a joint probability distribution) for all observable and unobservable quantities in a given problem. This model should be consistent with knowledge about the process being analyzed. Then, Bayesian inference about the parameters in the model or about unobserved data are made in terms of probability statements that are conditional on the observed data (posterior distributions). Hence, these methods provide a full distributional profile for the parameters, or other quantities of interest, so that the features of their distribution are readily apparent; for example, non-normality, skewness, or tail behavior. To obtain these posterior distributions, Bayesian methods combine the prior available information, no matter how limited, with the theoretical models for the variables of interest. Therefore Bayesian models automatically account for all the uncertainty in the parameters.

14.2 The Bayesian Paradigm

Bayesian theory is developed from the axiomatic system of the foundations of decision theory. In some references the dual concepts of probability and utility are formally defined and analyzed. Probabilities are considered “degrees of belief” of the

analyst about the occurrence of a given event, and the criterion of maximizing expected utility is seen to be the only criterion compatible with the axiomatic system. Statistical inference is viewed as a particular decision problem, and whether estimation or prediction, it must follow the laws of probability. As a result, the uncertainty of all unknown quantities is described in terms of probability distributions, which implies that these quantities are treated as random variables. The fact that parameters have a distribution function allows the application of Bayes' theorem to combine information coming from the data with prior information about the parameters. For a comprehensive exposition on the foundations of Bayesian theory, see Bernardo and Smith (2000) and references therein.

The ensuing methodology establishes how to formally combine an initial (prior) degree of belief of a researcher with currently measured, observed data, in such a way that it updates the initial degree of belief. The result is called *posterior belief*. This process is called Bayesian inference because the updating process is carried out through the application of Bayes' theorem. The posterior belief is proportional to the product of the two types of information: the prior information about the parameters in the model and the information provided by the data. This second part is usually thought of as the objective portion of the posterior belief. We explain this process as follows:

Let $\mathbf{y} = (y_1, y_2, \dots, y_n)$ be independent random variables, each of them coming from a probability model with density function $f(y_i|\theta)$, where θ is a parameter vector that characterizes the form of the density. Then $f(\mathbf{y}|\theta) = \prod_{i=1}^n f(y_i|\theta)$ is the joint probability density of \mathbf{y} given θ , which is usually referred to as the likelihood function. Prior available information on the parameter is described through a prior distribution $\pi(\theta)$ that must be specified or modeled by the actuary. Then, from a purely probabilistic point of view, it follows that

$$\pi(\theta | \mathbf{y}) = \frac{f(\mathbf{y} | \theta)\pi(\theta)}{f(\mathbf{y})}$$

where $f(\mathbf{y})$ is the marginal joint density of \mathbf{y} defined as $f(\mathbf{y}) = \int f(\mathbf{y}|\theta) \times \pi(\theta) d\theta$ if θ is continuous, and as $f(\mathbf{y}) = \sum_\theta f(\mathbf{y}|\theta)\pi(\theta)$ if θ is discrete. This is Bayes' theorem, which rules the updating of the information. Considering that $f(\mathbf{y})$ is just a constant for θ , then the updating mechanism can be simply written as $\pi(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)\pi(\theta)$, where \propto indicates proportionality. In other words, the posterior distribution of the parameters, conditional on the observed data, is proportional to the product of the likelihood function and the prior degree of belief. Any inference on the parameters is now carried out using the posterior distribution $\pi(\theta|\mathbf{y})$.

As was mentioned earlier, the only criterion for optimal decision making, consistent with the axiomatic system, is the maximization of the expected utility. Alternatively, this criterion is equivalently replaced by the minimization of a loss function. Therefore,

in the Bayesian framework, parameter estimation is done by minimizing the expected value of a specified loss function $l(\hat{\theta}, \theta)$ with respect to $\hat{\theta}$, where the expected value is taken with respect to the posterior distribution of the parameter θ given the data \mathbf{y} . In particular, a quadratic loss function $l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$ leads to the posterior mean $\hat{\theta} = E(\theta|\mathbf{y})$ as an optimal estimate for the parameter. In contrast, a linear loss function $l(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$ yields the median of the posterior distribution as an optimal estimate $\hat{\theta}$ for θ .

Nuisance parameters are handled in a very straightforward fashion within the Bayesian setting via marginalization. For example, if the parameter has two components, say $\theta = (\phi, \lambda)$ where ϕ is the parameter of interest and λ is the nuisance parameter, inference is done using the marginal posterior distribution $\pi(\phi|\mathbf{y}) = \int \pi(\phi, \lambda|\mathbf{y}) d\lambda$.

When the main purpose of modeling is prediction, then the observed data \mathbf{y} are used to predict future observations y_F by means of the posterior predictive distribution. Assuming continuous random variables to simplify presentation, the predictive distribution is defined as

$$f(y_F|\mathbf{y}) = \int f(y_F|\theta)\pi(\theta|\mathbf{y}) d\theta. \quad (14.1)$$

The parameters in the model have been marginalized (integrated out). Therefore, only information in the observed data is used in prediction. Finally, the optimal point predictor \hat{y}_F , assuming a quadratic loss function, is the mean of the predictive distribution $E(y_F|\mathbf{y})$.

To summarize, the Bayesian inference method can be thought of as comprising the following four principal steps:

1. Specify the prior beliefs in terms of a probability model. This should reflect what is known about the parameter prior to observing the data.
2. Compute the likelihood function in terms of the probability model that gave rise to the data. This contains the observed information about the parameters.
3. Apply Bayes' theorem to derive the posterior density. This posterior belief expresses what we know about the parameters after observing the data together with the prior belief.
4. Derive appropriate inference statements about the parameter from the posterior distribution and about future observations from the posterior predictive distribution.

There is a vast literature on how to specify a prior distribution. One of the most common approaches is to use the family of natural conjugate priors. A prior distribution $\pi(\theta)$ is said to be a natural conjugate for θ if, when combining it with the sample information, $\pi(\theta)$ and the resulting posterior $\pi(\theta|\mathbf{y})$ belong to the same family. These priors can be used to produce vague or diffuse priors, which reflect knowing little or having no prior information about the parameter, or to produce informative priors that

reflect the prior knowledge of the actuary. In either case this is achieved by setting the parameters of the prior to an appropriate value. In particular, vague priors are obtained by letting the prior variance be large, but in fact these priors are just an approximation of what are called noninformative (or objective) priors (e.g., Berger 2006). More details about Bayesian thinking can be found in Chapter 13.

14.3 Generalized Linear Models

14.3.1 Linear Models

The linear regression model is a way of expressing the relationship between a dependent or response variable y and a set of $p - 1$ independent or explanatory variables $\mathbf{x}' = (1, x_1, \dots, x_{p-1})$, via a linear combination with coefficients $\boldsymbol{\beta}' = (\beta_0, \dots, \beta_{p-1})$ of the form $\mathbf{x}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}$. This relationship can be expressed in terms of a linear equation with an additive random error ϵ such that

$$y = \mathbf{x}'\boldsymbol{\beta} + \epsilon, \quad (14.2)$$

where ϵ is interpreted as a measurement error and is assumed to have zero mean and constant precision τ (reciprocal of the variance). If we further assume that the error comes from a normal distribution, then $\epsilon \sim N(0, \tau)$.

The normal assumption in the error term implies that the response variable y , conditional on \mathbf{x} , also follows a normal model and thus can take any possible value on the real line. Generalized linear models extend this assumption to response variables with positive, bounded, or discrete outcomes. For example, if one wants to describe the behavior of the amounts in insurance claims or the number of claims in a certain period of time, the normal assumption would not be adequate in either case because claims cannot be negative and the number of claims is positive and discrete.

The role played by the explanatory variables \mathbf{x} in the linear normal (linear regression) model is to help in understanding the average or mean behavior of the response variable y . This is why the (conditional) expected value $E(y|\mathbf{x})$ is equal to a linear combination of the explanatory variables $\mathbf{x}'\boldsymbol{\beta}$. This property justifies the name *regression to the mean* of the linear regression model (14.2).

The linear regression model is a particular case of the larger class of generalized linear models. We discuss its properties and Bayesian inference in the following sections.

14.3.2 Generalized Linear Models

A Bayesian generalized linear model is a generalized linear model together with a specification of the prior beliefs of the unknown parameters. Generalized linear models

can also be considered regression models to the mean (of y), but in a nonlinear form since the parameter space of $E(y|\mathbf{x})$ is not necessarily the whole real line. Let us start by recalling the form of a generalized linear model. To account for all possible kinds of response variables (positive, bounded, discrete, etc.), the model describes the probabilistic behavior of the responses with a member of the exponential family. Then, for a sample of independent random variables y_1, y_2, \dots, y_n , each of them comes from the model

$$f(y_i \mid \theta_i, \phi_i) = b(y_i, \phi_i) \exp[\phi_i \{y_i \theta_i - a(\theta_i)\}], \quad (14.3)$$

where $a(\cdot)$ and $b(\cdot)$ are two monotonic functions. The parameters θ_i and ϕ_i are known as natural and dispersion parameters, respectively. It is not difficult to show that the mean and variance of y_i can be expressed in terms of derivatives of function $a(\cdot)$ as follows:

$$\mu_i = E(y_i) = a'(\theta_i) \text{ and } \sigma_i^2 = \text{Var}(y_i) = \frac{a''(\theta_i)}{\phi_i}.$$

Here prime and double prime denote the first and second derivative, respectively. Note that generalized linear models have also been described in Chapter 5 with a slightly different parameterization of the exponential family.

Each individual i has its own set of explanatory variables \mathbf{x}_i , $i = 1, \dots, n$. These will be combined in a single value through a linear combination with coefficients $\boldsymbol{\beta}' = (\beta_0, \beta_1, \dots, \beta_{p-1})$, forming what is called the *linear predictor* $\eta_i = \mathbf{x}'_i \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{ip-1}$. We note that linearity means linear in the coefficients because the linear predictor could well be a polynomial of order $p - 1$ of the form $\beta_0 + \beta_1 x_i + \dots + \beta_{p-1} x_i^{p-1}$ with a single explanatory variable for individual i , x_i .

The idea of the generalized linear models is to model the mean of the response variable, $\mu_i = E(y_i)$, in terms of the explanatory variables via the linear predictor η_i and an appropriate transformation $g(\cdot)$; that is, $\eta_i = g(\mu_i)$. The function $g(\cdot)$ is called the *link function* because it links the explanatory variables with the response. At the same time, the link function adjusts the parameter space of μ_i to correspond to the values of the predictor η_i , which is typically the real line. This can be seen as $\mu_i = g^{-1}(\eta_i)$. A particular choice for the link function $g(\cdot)$ is to take $g^{-1}(\cdot) = a'(\cdot)$. In this case the linear predictor η_i becomes equal to the natural parameter θ_i and $g(\cdot)$ is called the canonical link function. Other options for the link function are available, as long as the domain of the function $g(\cdot)$ corresponds to the parameter space of μ_i and the image to the real numbers. Let us consider two examples to illustrate these ideas.

It can be shown that the normal linear regression model is also a generalized linear model. To see this we take $y_i \sim N(\mu_i, \tau_i)$, parameterized in terms of mean μ_i and

precision (reciprocal of the variance) τ_i . The density function is

$$f(y_i | \mu_i, \tau_i) = (2\pi/\tau_i)^{-1/2} \exp \left\{ -\frac{\tau_i}{2}(y_i - \mu_i)^2 \right\}$$

for $y_i \in \mathbb{R}$, $\mu_i \in \mathbb{R}$, and $\tau_i > 0$. Writing the normal density as in (14.3) we get

$$\begin{aligned}\phi_i &= \tau_i, & b(y_i, \phi_i) &= (2\pi/\phi_i)^{-1/2} \exp \left\{ \frac{\phi_i}{2} y_i^2 \right\} \\ \theta_i &= \mu_i, & a(\theta_i) &= \frac{\theta_i^2}{2}\end{aligned}$$

In this case $a'(\theta_i) = \theta_i$ and thus the canonical link is $g(\mu_i) = \mu_i$. Therefore the mean μ_i is modeled directly, with the linear predictor η_i obtaining the linear model $\mu_i = \mathbf{x}'_i \boldsymbol{\beta}$.

A second example, suitable for response variables in the positive real line (as is the case for claim amounts in insurance companies), is to consider a response with gamma distribution. Using a mean parameterization of the gamma – that is, $y_i \sim \text{gamma}(\alpha_i, \alpha_i/\mu_i)$ such that $E(y_i) = \mu_i$ – then the density function is of the form

$$f(y_i | \alpha_i, \mu_i) = \left(\frac{\alpha_i}{\mu_i} \right)^{\alpha_i} \frac{1}{\Gamma(\alpha_i)} y_i^{\alpha_i - 1} e^{-\frac{\alpha_i}{\mu_i} y_i},$$

for $y_i > 0$ and $\alpha_i, \mu_i > 0$. Writing this gamma density as in (14.3) we get

$$\begin{aligned}\phi_i &= \alpha_i, & b(y_i, \phi_i) &= \frac{\phi_i^{\phi_i}}{\Gamma(\phi_i)} y_i^{\phi_i - 1}. \\ \theta_i &= -\frac{1}{\mu_i}, & a(\theta_i) &= \log \left(-\frac{1}{\theta_i} \right).\end{aligned}$$

Computing the derivative of the function $a(\cdot)$ we obtain $a'(\theta_i) = 1/\theta_i$, implying a canonical link $g(\mu_i) = -1/\mu_i$. We note that this link has a problem: its domain is fine because it corresponds to the parameter space of μ_i , but the image of $g(\cdot)$ is the negative numbers and not the real line. An alternative link function that overcomes this flaw is to take $g(\mu_i) = \log(\mu_i)$, where the domain and image are as desired. Another important point about the gamma model, when using the mean parameterization, is that the variance is also a function of the mean (i.e., $\text{Var}(y_i) = \mu_i^2/\alpha_i$). This implies that both the mean and variance will be a function of the explanatory variables, solving in that way the problem of heteroskedasticity that is not accounted for in a normal linear model. More examples for response variables in $\{0, 1\}$ and for count variables are presented in the following sections.

Once we have defined the sampling model for the data, a Bayesian model is completed by assigning prior distributions to the unknown quantities. A typical assumption in the previously defined generalized linear models is to consider a common dispersion parameter $\phi_i = \phi$ for all individuals $i = 1, \dots, n$. In this case, the set of unknown parameters in the model is $(\boldsymbol{\beta}, \phi)$. According to West (1985), conjugate priors for

these parameters are available only in very special cases. In general, posterior distributions are not available in closed forms so the choice of the prior has to do with the simplicity to accommodate prior beliefs. In this context a normal prior for the vector β has been the common choice, together with a gamma prior for the precision ϕ – typically assuming independence a priori among the β_j elements of β and between β and ϕ . The normal and gamma are well-known models that allow the user to specify prior beliefs in terms of mean and precision (reciprocal of variance). That is, we can take a priori $\beta_j \sim N(b_0, t_0)$, where b_0 is the prior mean and t_0 the prior precision for β_j . In the case of little or no information about β_j , we can set $b_0 = 0$ together with t_0 close to zero, say 0.1, 0.01 or 0.001, for $j = 1, \dots, p$. For the sampling precision parameter ϕ , if μ_ϕ and σ_ϕ^2 represent the prior mean and variance, then we can take $\phi \sim \text{gamma}(a_0, a_1)$ with $a_0 = \mu_\phi^2 / \sigma_\phi^2$ and $a_1 = \mu_\phi / \sigma_\phi^2$. Again, in the case of little or no prior information about ϕ , we can set $a_0 = a_1$ equal to a small value, say 0.1, 0.01 or 0.001, in such a way that ϕ has prior mean one and large/small prior variance/precision. Alternatively, more diffuse priors are also considered for the coefficients β_j ; for instance, a student-t prior or even a cauchy prior (Gelman et al. 2008).

Posterior inference of the parameters (β, ϕ) requires combining the information provided by the data, summarized in the likelihood function, and the prior distributions. The likelihood function is constructed by the product of the density (14.3) of the response variables as a function of the explanatory variables and the parameters; that is, $\text{lik}(\beta, \phi) = \prod_{i=1}^n f(y_i | \theta_i, \phi)$. Remember that the explanatory variables enter the model via the natural parameter θ_i , which in the case of using the canonical link $\theta_i = \mathbf{x}'_i \beta$; otherwise θ_i is replaced with an appropriate function of the linear predictor η_i . Finally, the posterior distribution $\pi(\beta, \phi | \text{data})$ is proportional to the product of this likelihood function $\text{lik}(\beta, \phi)$ and the prior distributions $\pi(\beta, \phi)$. Point estimates and credible intervals are obtained as summaries from this posterior distribution. A Bayesian credible interval is also known as a posterior probability interval, which is not to be confused with a frequentist confidence interval. For example, a 95% posterior interval for a parameter is an interval that contains exactly 95% of that parameter's posterior probability. More details on this can be found in Gelman et al. (2008).

Posterior summaries are obtained numerically via a Markov Chain Monte Carlo (MCMC) sampling algorithm or via Estimation Maximization (EM) techniques. The former can be implemented in OpenBugs (<http://www.openbugs.info/>) within R through the library R2OpenBUGS. The latter is implemented in the R command bayesglm from the package arm (data analysis using regression and multilevel/hierarchical models). Both are available in the Comprehensive R Archive Network (CRAN) at <http://www.r-project.org/>. All the examples presented in this chapter were run in OpenBugs within R, and the corresponding code is given in each case. In all cases the Markov chains were run for 20,000 iterations with a burn-in period

of 5,000 and keeping one of every 5th iterations for computing the estimates. See Chapter 13 for the meaning of these numbers.

Another aspect of interest when using generalized regression models is prediction of future outcomes. This inference problem is addressed naturally in the Bayesian approach by computing the predictive distribution for a future observation y_F . If a new individual has explanatory variables \mathbf{x}_F , and assuming the canonical link, then $\theta_F = \mathbf{x}'_F \boldsymbol{\beta}$ and the predictive distribution will be the weighted average of the density $f(y_F | \theta_F, \phi)$ with respect to the posterior distribution $\pi(\boldsymbol{\beta}, \phi | \text{data})$ as in (14.1). Point or interval predictions are produced using summaries from this predictive distribution. This is usually done numerically.

Traditionally, the goodness-of-fit measure used to compare generalized linear models is the deviance (see, Chapter 5). However, in a Bayesian context, model comparison is typically made by using the deviance information criterion (DIC) (Spiegelhalter et al. 2002), which is based on the deviance but includes a penalization for the number of parameters used in the model. Let $D(\theta) = -2 \log f(\mathbf{y}|\theta)$; then $\text{DIC} = 2 E_{\theta|\mathbf{y}}(D) - D(\hat{\theta})$, with $\hat{\theta} = E(\theta|\mathbf{y})$. Smaller values of DIC indicate a better fit.

Example 14.1. The insurance market in Mexico operates in different sectors. Seven of these are accident and sickness (ACC), agriculture and livestock (AGR), automobiles (AUT), major medical expenses (MED), fire (FIR), liability and professional risks (LIA), and health (HEA). It is of interest to the insurance companies to predict claim amounts y_i in terms of the premiums written x_i , both measured in millions of Mexican pesos. The insurance industry regulator in Mexico gathers the information from all different insurance companies every year and makes the information available on its web page <http://www.cnsf.gob.mx/>. The information is available for all 32 Mexican states and in some cases from abroad (in the data 4 sectors showed information from abroad). In total, for the year 2010, we have $i = 1, \dots, n$ with $n = 228$ observations classified by insurance sector. The dataset can be found in the Web Appendix of the book. A dispersion diagram of the 228 observations in logarithmic scale is presented in Figure 14.1. From the graph we can see that all sectors together follow a common pattern and a single line could potentially serve for fitting the data. In fact, the least square estimates are 0.0008 for the intercept and 0.85 for the slope.

Let us assume that the logarithm of the claim amounts $\log(y_i)$ follows a normal distribution with mean μ_i and constant precision τ ; – that is, $\log(y_i) \sim N(\mu_i, \tau)$. We model the mean level μ_i in terms of a linear combination of the premiums written in log scale $\log(x_i)$ and class indicators z_{ji} , for $j = 2, \dots, 7$, where for example z_{2i} takes the value of one if observation i belongs to sector 2 (AGR), and so on, following the order of the sectors in the previous paragraph. These sector indicators will serve to determine possible differences in the intercepts and the slopes

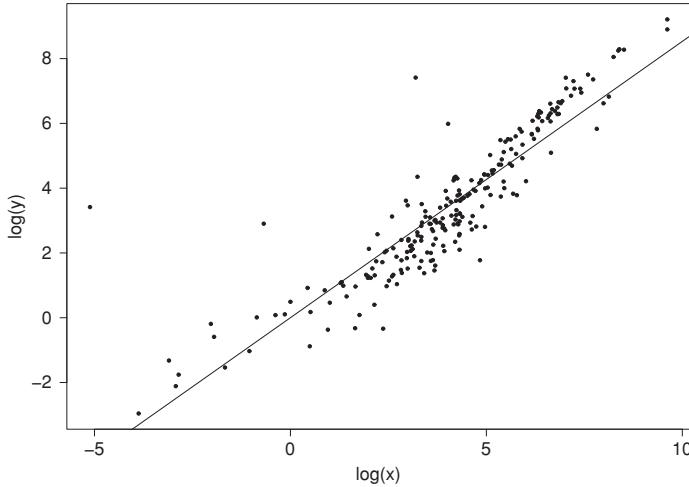


Fig. 14.1. Dispersion diagram of severity amounts y_i versus premium written x_i in logarithmic scale for $i = 1, \dots, 228$ individuals. Straight line corresponds to a least square fit to the data

by including the interactions $\log(x_i) \times z_{ji}$. The mean level is thus modeled as $\mu_i = \alpha_1 + \sum_{j=2}^7 \alpha_j z_{ji} + \beta_1 \log(x_i) + \sum_{j=2}^7 \beta_j \log(x_i) z_{ji}$. Note that to avoid indeterminacy of the model, the indicator for sector one is not present, so sector one has been taken as the baseline. An individual i coming from sector one is identified by assigning zeroes to all sector indicators z_{ji} , $j = 2, \dots, 7$. For the model coefficients we assign vague normal priors centered at zero and with small precision; that is, $\alpha_j \sim N(0, 0.001)$ and $\beta_j \sim N(0, 0.001)$ independently for $j = 1, \dots, 7$. For the common precision of the observations we take $\tau \sim \text{gamma}(0.001, 0.001)$ such that τ has mean one and large variance a priori. The R (BUGS) code of this model is given in Table 14.1.

Posterior estimates of the model coefficients and their credible intervals are presented in the second and third columns in Table 14.2. If the hypothesis of a single regression line for all sectors were true, then coefficients α_j and β_j for $j = 2, \dots, 7$ would all need to be zero. As we can see from the table, except for α_5 , the rest of the coefficients are all different from zero, implying different intercepts $\alpha_1 + \alpha_j$ and different slopes $\beta_1 + \beta_j$ for each sector j . These differences can be better appreciated graphically in Figure 14.2 where each colored line corresponds to a different sector. From the graph it is noticeable that sector ACC, represented by the black line, is the one that deviates the most from the general pattern of Figure 14.1. This large difference is mostly explained by the extreme observation with coordinates $(-5.11, 3.41)$ in logarithmic scale. This observation corresponds to an unfortunate event occurring abroad with a premium of 0.006 millions and a claim amount of 30.45 millions of Mexican pesos. A simple solution to describe the general pattern in the ACC sector would be

Table 14.1. BUGS Code for Model of Example 14.1

```

model{
#Likelihood
for (i in 1:n){
y[i] ~ dnorm(mu[i],tau)
mu[i]<-a[1]+a[2]*z2[i]+a[3]*z3[i]+a[4]*z4[i]+a[5]*z5[i]
+a[6]*z6[i]+a[7]*z7[i]+b[1]*x[i]+b[2]*x[i]*z2[i]
+b[3]*x[i]*z3[i]+b[4]*x[i]*z4[i]+b[5]*x[i]*z5[i]
+b[6]*x[i]*z6[i]+b[7]*x[i]*z7[i]
}
#Priors
for (j in 1:7){
a[j] ~ dnorm(0,0.001)
b[j] ~ dnorm(0,0.001)
}
tau ~ dgamma(0.001,0.001)
}

```

to remove this observation from the analysis. An alternative solution is to consider a model that is more robust to extreme observations such as the semi-parametric regression model with a Polya tree prior for the errors, which is described in Section 14.5. Finally, the posterior mean of the observations precision τ is 1.55 with 95% credible interval (CI) (1.26, 1.86).

Table 14.2. Posterior Estimates and Credible Intervals of Multiple Regression Coefficients in the Insurance Dataset; Parametric and Semi-Parametric Formulations

Coef.	Parametric		Semi-Parametric	
	Mean	95% CI	Mean	95% CI
α_1	1.77	(1.34, 2.20)	0.48	(-0.10, 1.63)
α_2	-1.29	(-2.24, -0.32)	-1.00	(-2.20, -0.23)
α_3	-2.21	(-4.22, -0.15)	-0.45	(-2.01, 0.63)
α_4	-2.92	(-4.01, -1.72)	-1.79	(-2.99, -1.01)
α_5	-0.96	(-2.11, 0.15)	-0.66	(-2.30, 0.54)
α_6	-2.49	(-3.48, -1.54)	-1.47	(-2.46, -0.63)
α_7	-1.57	(-2.11, -1.06)	-0.11	(-1.19, 0.57)
β_1	0.26	(0.14, 0.38)	0.64	(0.41, 0.80)
β_2	0.39	(0.13, 0.64)	0.31	(0.12, 0.57)
β_3	0.76	(0.45, 1.06)	0.32	(0.13, 0.62)
β_4	0.84	(0.61, 1.04)	0.48	(0.27, 0.73)
β_5	0.41	(0.17, 0.67)	0.24	(-0.01, 0.57)
β_6	0.58	(0.35, 0.84)	0.25	(0.01, 0.45)
β_7	0.49	(0.33, 0.65)	0.13	(-0.04, 0.31)

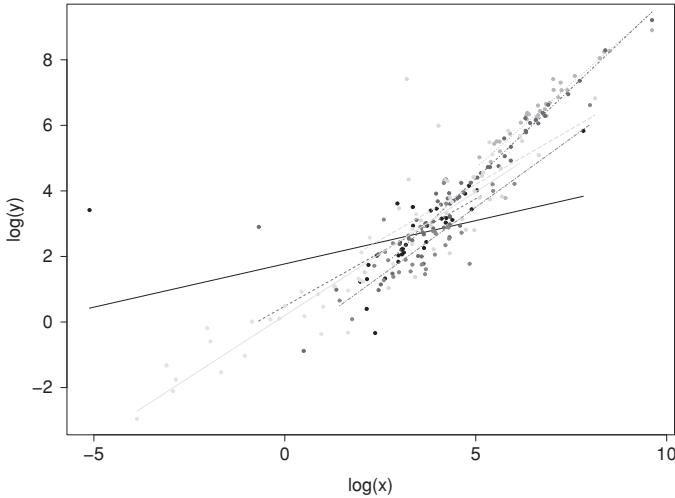


Fig. 14.2. Dispersion diagram of claim amounts y_i versus premiums written x_i in logarithmic scale by class. Straight lines corresponds to model fit by sector. Dispersion diagram of claim amounts y_i versus premiums written x_i in logarithmic scale by class. Straight lines corresponds to model fit by sector. Symbols indicate different sector: empty dots (ACC), triangles (AGR), plus symbols (AUT), crosses (MED), dyamonds (FIR), upside down triangles (LIA) and crossed squares (HEA).

14.3.3 Bayesian Regression with Zero-One Dependent Variables

In actuarial science and risk management, it is of interest to estimate the probability of default. For instance, when assigning a personal credit (loan) to an individual, the financial institution needs to quantify the risk of default, according to the individual's personal characteristics and financial history. This problem can be modeled by assuming a zero-one (Bernoulli) response variable y_i , with probability of success (default) μ_i ; that is, $y_i \sim \text{Bernoulli}(\mu_i)$ with density function given by

$$f(y_i | \mu_i) = \mu^{y_i} (1 - \mu_i)^{1-y_i},$$

for $y_i \in \{0, 1\}$ and $\mu_i \in (0, 1)$. Writing this density as in (14.3), to identify the model as a member of the exponential family, we get

$$\begin{aligned} \phi_i &= 1, & b(y_i, \phi_i) &= 1 \\ \theta_i &= \log \left\{ \frac{\mu_i}{1-\mu_i} \right\}, & a(\theta_i) &= \log(1 + e^{\theta_i}) \end{aligned}$$

The first derivative of function $a(\cdot)$ is $a'(\theta_i) = e^{\theta_i} / (1 + e^{\theta_i})$. Inverting this function to obtain the canonical link we get $g(\mu_i) = \log\{\mu_i/(1 - \mu_i)\}$. See also Chapter 3 for more details on regression models with categorical dependent variables.

A generalized linear model for a Bernoulli response with canonical link is called a *logistic regression model*. Recall that other link functions can be used as long as the domain of function $g(\cdot)$ corresponds to the parameter space and the image to

the real line. Since the parameter space of μ_i is the interval $(0, 1)$, any function that transforms the $(0, 1)$ interval into real numbers is a suitable link function. In our basic probability courses we learned that cumulative distribution functions (c.d.f.) for continuous random variables are functions with real domain and $(0, 1)$ image. Therefore, the inverse of any continuous c.d.f. $F(\cdot)$ can be a link function; that is, $g(\cdot) = F^{-1}(\cdot)$. In particular, if $F = \Phi$, the c.d.f. of a standard normal, then $g(\cdot) = \Phi^{-1}(\cdot)$ produces the *probit regression model*. In fact, the inverse of the canonical link corresponds to the c.d.f. of a logistic random variable; thus the name of logistic regression. Two other common link functions are the log-log link $g(\mu_i) = \log\{-\log(\mu_i)\}$ and complementary log-log link $g(\mu_i) = \log\{-\log(1 - \mu_i)\}$. This latter link corresponds to the inverse of the c.d.f. of a extreme value distribution. Whatever the link function we choose, for a given vector of explanatory variables of individual i , the probability of success is expressed in terms of the explanatory variables as $\mu_i = g^{-1}(\eta_i) = F(\mathbf{x}'_i \boldsymbol{\beta})$.

Sometimes several individuals share the same value of the explanatory variables, or it is also possible that available information is grouped and the covariate information is only available at the group level. This is the case, for example, in insurance groups where it is assumed that all individuals in the same group show similar risk characteristics and the number of claims y_i out of n_i members in group i is reported. In such a case it is of interest to estimate the probability of presenting a claim π_i for group i with characteristics \mathbf{x}_i . These kind of data can also be modeled with a generalized linear model by assuming $y_i \sim \text{binomial}(n_i, \pi_i)$ and $\pi_i = F(\mathbf{x}'_i \boldsymbol{\beta})$ with a specific choice of continuous c.d.f. $F^{-1}(\cdot)$ as the link function.

For both models, Bernoulli and binomial, the precision parameter ϕ_i is equal to one, and for the grouped data, the number of individual in the group n_i is assumed known. This leaves us with one set of unknown parameters $\boldsymbol{\beta}$. For each β_j , $j = 1, \dots, p$, we assign normal and student- t prior distributions as suggested earlier.

Example 14.2. The Mexican Central Bank is responsible for issuing the required number of bills for the effective functioning of the economy. Table 14.3 contains information on the number of bills in circulation and the number of fake bills, both in million pieces, for different denominations ($\$20, \$50, \$100, \200 , and $\$500$ Mexican pesos). This information is available annually from the year 2000 to 2011. Let us disregard temporal dependence and assume that the number of fake bills y_i follows a binomial distribution with parameters n_i , the number of circulating bills, and π_i , the proportion of fake bills with respect to the real bills circulating; that is, $y_i \sim \text{binomial}(n_i, \pi_i)$, for $i = 1, \dots, n$ with $n = 60$ observations. To help understand the information contained in Table 14.3, we present boxplots of the crude proportion of fake bills for every thousand circulating bills; that is, $\hat{\pi} = y_i/n_i \times 1000$. Figure 14.3 shows these fake proportions across the different bill denominations: the $\$20$ bill is

Table 14.3. Number of Bills in Circulation (C) and Number of Fake Bills (F), in Million Pieces, for Different Bill Denominations for Years 2000 to 2011

Year	C20	F20	C50	F50	C100	F100	C200	F200	C500	F500
2000	2182.1	14.8	3141.4	179.7	2779.4	178.5	4163.4	83.9	1100.7	26.6
2001	2092.6	13.1	2900.9	150.5	2795.0	136.8	4745.5	64.4	1335.0	20.9
2002	2182.4	18.1	3026.5	109.7	3155.5	64.2	5192.1	97.3	1802.1	35.7
2003	2449.1	9.4	4245.0	140.9	4455.4	60.1	4870.4	77.6	2352.4	42.9
2004	2545.8	1.5	4031.8	149.2	4951.7	117.8	5087.4	80.5	3028.0	34.0
2005	2707.8	1.0	3420.2	249.3	4411.0	142.9	5422.1	117.8	3522.5	43.6
2006	2877.4	0.7	3615.2	215.1	4625.9	106.5	5935.6	88.8	4190.9	70.9
2007	2959.8	0.6	3847.5	122.5	4768.0	77.1	6358.0	78.6	4889.7	90.5
2008	3360.9	1.0	3892.8	59.4	4830.2	87.6	6850.7	97.7	5682.5	91.7
2009	3578.6	3.2	4129.0	28.3	4872.5	81.0	7314.7	136.3	6934.4	91.2
2010	3707.6	2.7	4197.3	67.9	5210.0	101.2	7505.1	139.7	7799.3	96.4
2011	3858.8	1.3	4375.1	208.9	5416.0	88.7	7528.1	120.1	8907.4	89.7

Source: Banco de México. <http://www.banxico.org.mx/estadisticas/index.html>.

the least falsified bill with low dispersion, and the \$50 bill is the most falsified with a large dispersion across the years. Figure 14.4 presents the fake proportions across the years, showing a decreasing path in time, both in location and dispersion.

To identify the bill denomination we construct auxiliary dummy variables, say x_{ji} , $j = 1, \dots, 5$ for each of the five bill denominations \$20, \$50, \$100, \$200, and \$500, respectively, such that x_{ji} takes the value of one if observation i corresponds to

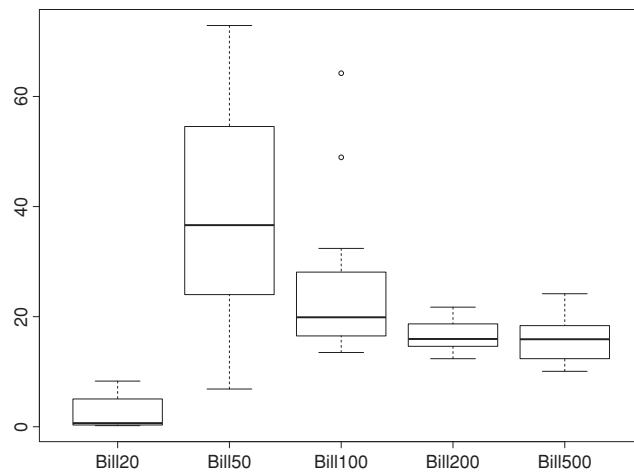


Fig. 14.3. Boxplots of crude fake proportions multiplied by 1,000 for the different bill denominations in Table 14.3.

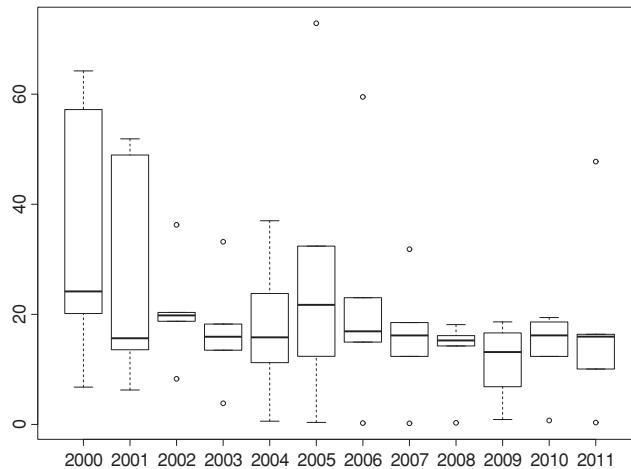


Fig. 14.4. Boxplots of crude fake proportions multiplied by 1,000 for the different years in Table 14.3.

the bill denomination j and zero otherwise. We then define a first model (Model A) with a linear predictor of the form $\eta_i = \alpha + \sum_{j=1}^5 \beta_j x_{ji}$. By considering a different parameter β_j for each denomination plus an intercept, Model A is overparametrized and the β_j 's will not be estimable. To identify the parameters we include the constraint $\sum_j \beta_j = 0$. We then compare the logistic and standard normal links. These two links imply $\pi_i = e^{\eta_i} / (1 + e^{\eta_i})$ and $\pi_i = \Phi(\eta_i)$, respectively. For the prior distributions on α and the β_j 's, we consider two alternatives $N(0, 0.001)$ and $St(0, 0.001, 3)$. Note that the third parameter in the student-t distribution corresponds to the degrees of freedom, which have to be greater than 2. We chose 3 to avoid numerical problems. This model is translated into R (BUGS) code as shown in Table 14.4. Note that the last two rows of the code describe the inclusion of the constraint on the parameters.

To compare the different models, we compute the deviance information criterion (DIC). This value is easily obtained from the library R2OpenBUGS. Table 14.5 shows the DIC values for the competing models. As can be seen from the table, the logit link is preferred to the probit link for this particular dataset, regardless of the prior distribution used. When comparing the two priors, the fit is practically the same, with a negligible advantage for the normal model. For producing inference of the model parameters, we select the logit-normal model since it achieved the smallest DIC value.

We define the rate of fake bills for every thousand circulating bills as $p_j = e^{\alpha+\beta_j} / (1 + e^{\alpha+\beta_j}) \times 1,000$, for each of the five bill denominations $j = 1, \dots, 5$. Table 14.6 contains posterior point estimates (posterior means) as well as 95% credible intervals. Interpreting the estimated rates, the \$50 pesos bill has the largest fake rate, with 37.5 fake bills for every 1,000 circulating bills. In contrast, the \$20 pesos

Table 14.4. BUGS Code for Model A of Example 14.1

```

model{
#Likelihood
for (i in 1:n){
y[i] ~ dbin(pi[i],e[i])
logit(pi[i])<-a+b[1]*x20[i]+b[2]*x50[i]+b[3]*x100[i]+b[4]*x200[i]
+b[5]*x500[i]
#probit(pi[i])<-a+b[1]*x20[i]+b[2]*x50[i]+b[3]*x100[i]
+b[4]*x200[i]+b[5]*x500[i]
}
#Priors
a ~ dnorm(0,0.001)
for (j in 1:5){
b[j] ~ dnorm(0,0.001)
#b[j] ~ dt(0,0.001,3)}
#Estimable parameters
a.adj<-a+mean(b[])
for (j in 1:5){b.adj[j]<-b[j]-mean(b[])}

```

bill shows the smallest rate, with almost 2 fake bills for every 1,000 circulating. According to Figure 14.3, the distribution of crude proportions for the \$200 and \$500 denominations seems to overlap. However, inference derived from the binomial-logistic-normal model shows that these rates have 95% credible intervals that do not intersect (13.23, 15.27) for the \$500 bill and (15.73, 17.63) for the \$20 bill. In fact, the posterior probability of $p_4 > p_5$ is 0.9995.

As suggested by Gelman et al. (2004), an important role of Bayesian inference is to assess model fitting. Apart from the numerical measures, such as the DIC, it is possible to obtain posterior predictive draws from the fitted model and compare them to the actual observed data. This is known as posterior predictive checking. We sampled

Table 14.5. Deviance Information Criterion (DIC) for Four Binomial Generalized Models Fitted to the Mexican Central Bank Data

Prior	Model A		Model B	
	Logit	Probit	Logit	Probit
Normal	1295.87	1297.45	901.98	919.38
Student-t	1295.98	1297.49	902.02	919.44

Table 14.6. *Posterior Inference for the Rate of Fake Bills for Every Thousand Circulating Bills, p_i , $i = 1, \dots, 5$, Using the Mexican Central Bank Data, Posterior Mean and 95% Credible Intervals*

\$	Rate	Model A			Model B		
		Mean	2.5%	97.5%	Mean	2.5%	97.5%
20	p_1	1.95	1.53	2.45	1.95	1.51	2.44
50	p_2	37.50	35.75	39.28	37.10	35.33	38.90
100	p_3	23.77	22.49	25.10	23.87	22.56	25.22
200	p_4	16.66	15.73	17.63	16.65	15.72	17.63
500	p_5	14.24	13.23	15.27	15.48	14.37	16.63

y_i^F , $i = 1, \dots, 60$ observations from the posterior predictive distribution of Model A to mimic the observed data. We computed posterior predictive proportions for one thousand circulating $\hat{\pi}_i^F = y_i^F/n_i \times 1,000$ and compared them with the observed ratios for the five bill denominations. The first boxplot in each panel of Figure 14.5 corresponds to the observed proportions, and the following five boxplots correspond to five replicates of posterior predictive proportions from Model A. As can be seen, these predictive proportions are highly concentrated and do not capture the variability shown by the data.

To overcome these deficiencies of Model A, we propose a second model (Model B) that accounts for the differences in the number of fake bills across time. As we did for the bill denominations, to identify the years we define auxiliary variables t_{ki} , $k = 1, \dots, 12$ for years from 2000 to 2011, respectively, where t_{ki} takes the value of one if observation i occurred in year k and zero otherwise. Therefore, the linear predictor for Model B has the form $\eta_i = \alpha + \sum_{j=1}^5 \beta_j x_{ji} + \sum_{k=1}^{12} \gamma_k t_{ki}$, together with the estimability constrain $\sum_{k=1}^{12} \gamma_k = 0$. As for Model A, we consider independent normal and student-t prior distributions for α , β_j 's, and γ_k 's and also compare between the logit and probit links. DIC values are reported in Table 14.5. Again, the normal-logit model achieves the best fit. Additionally, we can compare the fitting with respect to Models A and B. The DIC values for Model B are a lot smaller than those obtained with Model A, so it is expected to produce better inferences.

In the last three columns of Table 14.6 we present posterior estimates of the rates of fake bills for every thousand circulating bills p_j , for $j = 1, \dots, 5$. Point and interval estimates are consistent with those obtained with Model A, except for p_5 , the proportion of fake \$500 pesos bills. Model B estimates a slightly larger proportion than Model A. In fact, when comparing the proportions of \$200 and \$500 bill denominations, the probability of $p_4 > p_5$ is 0.9376, not as large as that of Model A. Finally, we produce posterior predictive checks for Model B and computed $\hat{\pi}_i^F$ and compare them with the observed proportions. The last five boxplots in each panel of

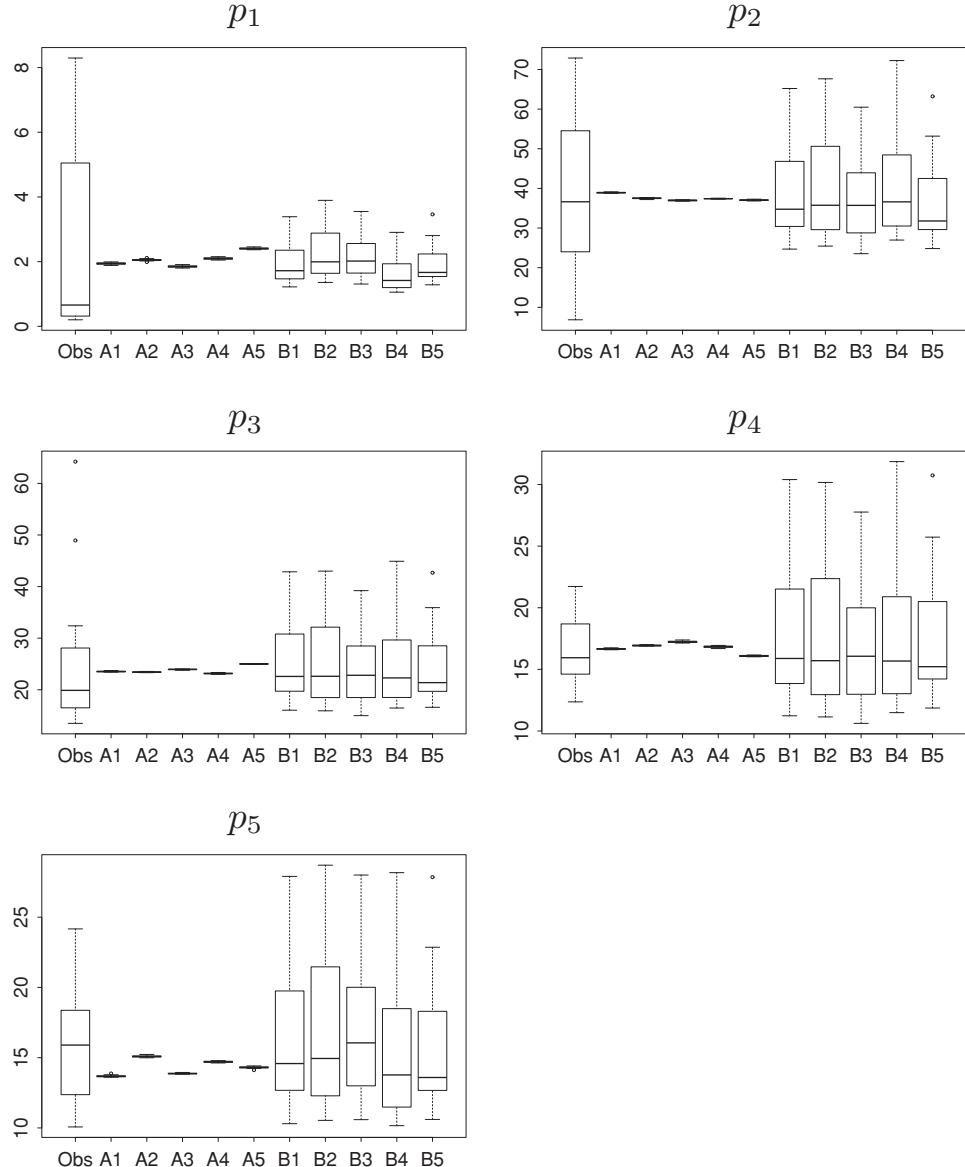


Fig. 14.5. Boxplots of posterior predictive checks for the proportion of fake bills for every one thousand circulating bills, p_i , $i = 1, \dots, 5$. Reported are observed rates (Obs), five replicates from Model A (A_1, \dots, A_5), and five replicates from Model B (B_1, \dots, B_5).

Figure 14.5 correspond to five replicates from Model B. As is clear from the figure, Model B better captures the variability of the data and the inferences obtained from it are more reliable. Although Model B has improved on Model A, the model is a little deficient in capturing the variability in the \$20 pesos bills p_1 .

14.3.4 Bayesian Regression with Count Dependent Variables

Another common problem in actuarial science is the study of counting or count data; for example, the number of claims that an insured individual can file during a calendar year. The natural assumption for a counting response variable y_i is a Poisson model; that is, $y_i \sim \text{Poisson}(\mu_i)$. This model has density function given by

$$f(y_i | \mu_i) = e^{-\mu_i} \frac{\mu_i^{y_i}}{y_i!},$$

for $y_i = 0, 1, \dots$ and $\mu_i > 0$. Identifying this density as (14.3) we obtain

$$\begin{aligned}\phi_i &= 1, & b(y_i, \phi_i) &= \frac{1}{y_i!}. \\ \theta_i &= \log(\mu_i), & a(\theta_i) &= e^{\theta_i}.\end{aligned}$$

Thus the canonical link obtained as the inverse of the derivative of function $a(\cdot)$ is $g(\mu_i) = \log(\mu_i)$. Therefore, the mean of the response variable is modeled as $\mu_i = e^{\mathbf{x}_i^\beta}$. This model is also known as *Poisson regression model* (see Chapter 4 for regression models with count dependent variables).

Sometimes, instead of modeling the mean μ_i of a Poisson response variable, it is of interest to model the rate of occurrence of events λ_i relative to a known number exposed or at risk e_i , such that $\mu_i = e_i \lambda_i$. In this case the rate is modeled through the explanatory variables as $\lambda_i = e^{\mathbf{x}_i^\beta}$. For instance, in mortality studies, the maternity mortality ratio is defined as the rate of maternity deaths for every 100,000 births. In such studies the response variable y_i is the number of maternity deaths, e_i is the number of births (in 100,000s), and thus λ_i becomes the maternity mortality ratio.

As mentioned before, in regression models with count dependent variables, the Poisson model is the common assumption; however, this model assumes that the mean and variance of the responses are the same; that is, $E(y_i) = \text{Var}(y_i) = \mu_i$. In practice, this assumption is not always satisfied by the data, due to an effect of overdispersion ($\text{Var}(y_i) > E(y_i)$). To account for overdispersion in a dataset, a different model for the responses has to be used. The negative binomial is the typical alternative for modeling counting data in the presence of overdispersion (i.e., $y_i \sim \text{NB}(r_i, \pi_i)$). To give the parameters of the negative binomial the same interpretation as in the Poisson model, the integer parameter r_i has to coincide with the number of exposed e_i and the probability of success π_i with $1/(1 + \lambda_i)$. This implies that $E(y_i) = e_i \lambda_i$ and $\text{Var}(y_i) = e_i \lambda_i (1 + \lambda_i)$. The quantity $1 + \lambda_i = \text{Var}(y_i)/E(y_i)$ is a measure of the amount of overdispersion present in the data. Finally, the rate λ_i is modeled in terms of the explanatory variables \mathbf{x}_i as in the Poisson model.

In the Bayesian literature (e.g., Bernardo and Smith 2000), it is well known that a negative binomial distribution is a particular case of a Poisson-gamma distribution. The latter gets its name because it can be obtained as a mixture of a Poisson distribution

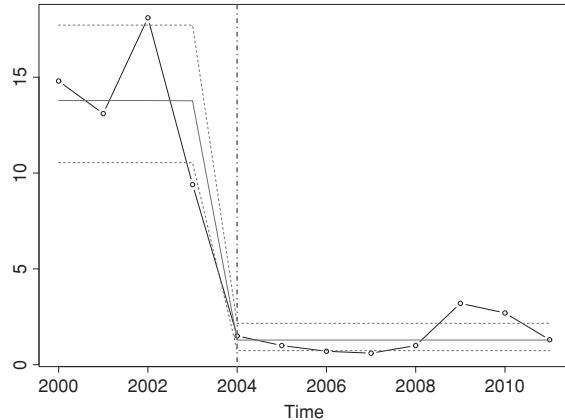


Fig. 14.6. Number of fake \$20 peso bills from 2000 to 2011. Empty dots (linked with solid black lines) represent observed values (F20). Red straight lines correspond to rate estimates (solid line) and 95% credible intervals (dotted lines). Vertical grey dashed-dotted line corresponds to year change estimate.

with respect to a gamma distribution. For the particular parameterization of our negative binomial model $y_i \sim NB(e_i, 1/(1 + \lambda_i))$, we can obtain the same model by considering a conditional Poisson distribution $y_i|t_i \sim \text{Poisson}(t_i\lambda_i)$ and marginal distribution $t_i \sim \text{gamma}(e_i, 1)$. Writing the negative binomial in this form allows us to consider the overdispersion case within the non-overdispersed Poisson setting by taking $t_i = e_i$ fixed if no overdispersion is present and $t_i \sim \text{gamma}(e_i, 1)$ random in the case of overdispersion. This construction of the negative binomial represents a hierarchical model that is explained in detail in Section 14.4.

Here, as in the previous model with Bernoulli or binomial response variables, the only set of unknown parameters is β , so normal and student-t distributions are used to represent prior knowledge.

Example 14.3. Consider the bills dataset presented in Table 14.3. The number of fake \$20 pesos bills is reported in variable F20. These numbers are shown as empty dots in Figure 14.6 and linked with a solid black line. As can be seen, there is a drastic drop of level in the number of fake bills around year 2003. This change is explained by the fact that in the early months of 2003 the Bank of Mexico released a new \$20 pesos bill made of polymer, which is more difficult to counterfeit than the previous bill made out of regular money paper. To model these data we propose a generalized linear Poisson regression model that accounts for a change in the level. Specifically, we assume that the number of fake \$20 pesos bills y_i follows a Poisson distribution with rate or intensity μ_i ; that is, $y_i \sim \text{Poisson}(\mu_i)$, with $\log(\mu_i) = \beta_1 + \beta_2 I(t_i \geq \alpha)$, for observations $i = 1, \dots, 12$. Note that t_i corresponds to the year reported in Table 14.3. We consider $N(0, 0.001)$ independent priors for β_j , $j = 1, 2$ and a uniform discrete

Table 14.7. BUGS Code for Model of Example 14.3

```

model{
#Likelihood
for (i in 1:n){
y[i] ~ dpois(mu[i])
log(mu[i])<-b[1]+b[2]*step(t[i]-a)}
#Priors
a<-c+1999
c ~ dcat(p[])
for (j in 1:12){p[j]<-1/12}
for (j in 1:2){b[j] ~ dnorm(0,0.001)}
}

```

prior for α on the set $\{2000, 2001, \dots, 2011\}$. The corresponding R (Bugs) code is presented in Table 14.7.

The red lines, solid and dotted in Figure 14.6, correspond to the rate μ_i point estimates and 95% credible intervals, respectively, for all years. Years 2000 to 2003 (inclusive) have a common rate of 13.71 million fake pieces per year with a 95% credible interval of (10.17, 17.52), whereas from 2004 to 2011 the rate drops to 1.28 million pieces with a 95% credible interval of (0.69, 2.12). The posterior distribution for the change year, α , concentrates its probability in only two values: 2003 with a probability of 0.0006, and 2004 with a probability of 0.9994. The estimated change year is denoted with a light grey (dotted-dashed) vertical line in Figure 14.6. Although the new \$20 pesos bills were introduced at the beginning of 2003, the impact on the counterfeit rate was reflected from 2004 onward. The number of fake bills in 2003 was more similar to the previous years than the number in the following years. This was captured by the model.

14.4 Mixed and Hierarchical Models

14.4.1 Mixed Models

In the previous section, general regression models for different forms of the response variable were introduced. Those models are also known as *fixed effects models* and assume that the observed individuals (y_i, \mathbf{x}_i) , $i = 1, \dots, n$ are independent. In some applications, response variables are observed over time (longitudinal models) or in space (spatial models) or are clustered in groups (repeated measurements). All these cases assume certain kind of dependence among observations. *Mixed effects models*, or simply *mixed models*, account for dependence among observations by introducing random (unobserved) effects in the model. The general specification of a mixed model

assumes two sets of explanatory variables \mathbf{x}_i and \mathbf{z}_i such that the former is associated to fixed coefficients $\boldsymbol{\beta}$ and the latter to random coefficients $\boldsymbol{\alpha}_i$. Thus, in the context of a generalized linear model, the mixed model has a linear predictor of the form $\eta_i = \mathbf{x}'_i \boldsymbol{\beta} + \boldsymbol{\alpha}'_i \mathbf{z}_i$. This is again linked to the response variable through an appropriate link function such that $g(\mu_i) = \eta_i$.

To better understand how dependence is introduced in a mixed model, let us consider a nested structure for the observations, say y_{ij} , where $j = 1, \dots, n_i$ and $i = 1, \dots, n$. For example, individual i could file a total of n_i number of claims during a year, with j denoting the specific claim. In this case, a mixed model for the claim amounts y_{ij} , in a generalized linear setting, would have linear predictor $\eta_{ij} = \mathbf{x}'_{ij} \boldsymbol{\beta} + \alpha_i$. Note that the parameter $\boldsymbol{\beta}$ is the fixed effect component common to all individuals, whereas α_i is a random effect common to all claims j made by the same individual i – and thus introducing a dependence in those claims made by the same individual.

Specifications for the random effects α_i 's may vary according to the application. They could simply be $\alpha_i \stackrel{iid}{\sim} N(0, \tau)$, which is the typical specification in repeated measurements, clustered observations, and longitudinal models. Alternative specifications include spatial effects $(\alpha_1, \dots, \alpha_n) \sim \mathcal{CAR}(\rho, \tau)$, where \mathcal{CAR} stands for a conditionally autoregressive model with association parameter ρ and precision τ . This model is a multivariate normal whose precision matrix is based on the spatial neighborhood structure. We refer the reader to Chapter 11 or alternatively to Banerjee, Carlin, and Gelfand (2004) for details. These specifications also include temporal effects $\alpha_i = \gamma \alpha_{i-1} + \nu_i$, with $\nu_i \stackrel{iid}{\sim} N(0, \tau)$, following a dynamic equation to account for dependence over time. We refer the reader to West and Harrison (1997) for details. In the following subsection we describe an alternative specification for the random effects that is based on the idea of exchangeability. See Chapter 8 for more details on mixed models.

14.4.2 Hierarchical Models

According to Gelman et al. (2004) hierarchical models are the most powerful tool for data analysis. Hierarchical specifications are usually helpful in specifying a joint prior distribution for a set of parameters. However, they are also useful for specifying the distribution of random effects in a mixed model.

To describe the construction of a hierarchical model, let us consider a simple scenario with response variables y_i , for $i = 1, \dots, n$, where the distribution of each y_i depends on a parameter θ ; that is, $f(y_i | \theta)$. This scenario assumes that there is a unique parameter θ common to all individuals. So inference on θ will be based on a prior distribution $\pi(\theta)$ and all observations y_i 's (as in a traditional Bayesian analysis

with i.i.d. observations). A completely different specification of the problem would be to assume that the distribution of each individual i has its own parameter θ_i ; that is, $f(y_i|\theta_i)$. In this case, if we further take independent priors $\pi(\theta_i)$ for all $i = 1, \dots, n$, inference on θ_i will only depend on its prior and the single observation y_i , like having n separate analyses. Hierarchical models present a compromise between these two extreme scenarios by allowing (i) heterogeneity in the parameters by keeping a different θ_i for each y_i and (ii) the pooling of strength across different observations to increase precision in the estimation of the θ_i 's.

We achieve (i) and (ii) by considering an exchangeable prior distribution for the vector $\boldsymbol{\theta}' = (\theta_1, \dots, \theta_n)$. Exchangeability can be interpreted as a symmetric condition in the prior such that each θ_i has the same marginal distribution and the dependence among any pair (θ_i, θ_j) is the same. We achieve this symmetry in the prior with a two-level hierarchical representation of the form:

$$\begin{aligned}\theta_i | \psi &\stackrel{\text{iid}}{\sim} \pi(\theta_i | \psi), \quad i = 1, \dots, n \\ \psi &\sim \pi(\psi).\end{aligned}$$

The parameter ψ is called the hyper-parameter and plays the role of an anchor of the θ_i 's. Conditional on ψ , the θ_i 's are independent, and when ψ is marginalized the θ_i 's become dependent (i.e., $\pi(\theta_1, \dots, \theta_n) = \int \prod_{i=1}^n \pi(\theta_i | \psi) \pi(\psi) d\psi$). The hierarchical model is completed by specifying the distribution of the data, which in general would be $y_i \sim f(y_i | \theta_i)$ independently for $i = 1, \dots, n$. This specification is analogous to the so-called structure distribution that is frequently used in actuarial science, specifically in credibility theory.

Hierarchical models are particularly useful for meta-analysis, where information coming from different studies y_i is linked via a hierarchical prior distribution on the different parameters $(\theta_1, \dots, \theta_n)$. Global or population inference from all studies is usually summarized in terms of the hyper-parameter ψ .

Example 14.4. Regarding the bills dataset of Table 14.3, consider now that the individuals i are the different bill denominations for $i = 1, \dots, n$ with $n = 5$. For each bill denomination i we have $n_i = 12$ observations $j = 1, \dots, n_i$ corresponding to the 12 years. For each observed number of fake bills y_{ij} we assume a Poisson model of the form $y_{ij} \sim \text{Poisson}(\mu_i)$ with $\log(\mu_i) = \beta_i$. Here we have two options: take independent priors for each β_i , say $\beta_i \sim N(0, 0.001)$ for $i = 1, \dots, 5$, or take an exchangeable prior for the vector $(\beta_1, \dots, \beta_5)$ with hierarchical representation given by $\beta_j | \beta_0, \tau \sim N(\beta_0, \tau)$ with $\beta_0 \sim N(0, 0.001)$ and $\tau \sim \text{gamma}(10, 1)$. Here the crucial parameter is τ . Since τ is a precision parameter for the β_i 's, a small value would imply a large uncertainty, allowing a broad combination of information across different individuals i , whereas a large value reduces the uncertainty around β_0 and constrains

Table 14.8. BUGS Code for Model of Example 14.4

```

model{
  #Likelihood
  for (i in 1:5){
    for (j in 1:n){y[i,j] ~ dpois(mu[i])}
    log(mu[i])<-b[i]
  }
  #Priors
  for (i in 1:5){
    b[i] ~ dnorm(0,0.001)
    #b[i] ~ dnorm(b0,tau)
  }
  #b0 ~ dnorm(0,0.001)
  #tau ~ dgamma(10,1)
}

```

the sharing of information across different i 's. The prior we took for τ , $\text{gamma}(10, 1)$ is a slightly informative prior that allows a moderate sharing of information. The R (BUGS) code of this model is presented in Table 14.8.

Posterior estimates for the two prior choices, independent and hierarchical, are reported in Table 14.9 and presented in Figure 14.7. Numerical values are very similar when using both priors. If we concentrate in the point estimates of the μ_i 's, we can see that for those denominations with the smallest rates (F20 and F500) their point estimates increase when using a hierarchical prior with respect to those with independent priors. In contrast, for those denominations with the largest rates (F50 and F100) their point estimates decrease. These effects are the result of sharing information between models; the estimates tend to compromise among all pieces

Table 14.9. Posterior Estimates of Fake Rates for Different Bill Denominations Under Distinct Scenarios

Coef.	Variable	Independent		Hierarchical	
		Mean	95% CI	Mean	95% CI
μ_1	F20	5.62	(4.39, 7.06)	5.74	(4.47, 7.20)
μ_2	F50	140.16	(133.65, 147.10)	140.15	(133.70, 147.10)
μ_3	F100	103.57	(97.79, 109.40)	103.53	(97.85, 109.30)
μ_4	F200	98.58	(93.03, 104.05)	98.54	(92.96, 104.20)
μ_5	F500	61.17	(56.83, 65.68)	61.22	(56.81, 65.73)
μ	F	81.76	(79.54, 84.12)	58.05	(32.04, 96.06)

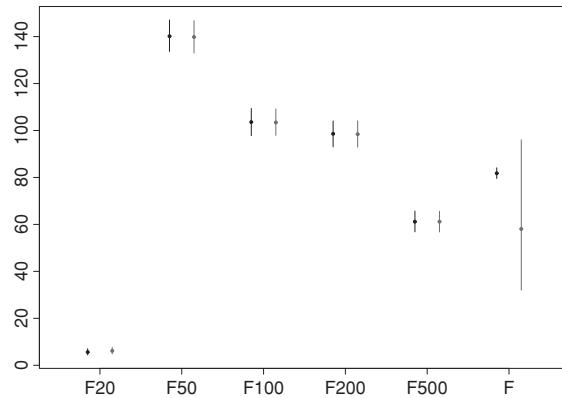


Fig. 14.7. Posterior estimates of fake rates for different bill denominations. Vertical lines correspond to 95% credible intervals and big dots to posterior means. Black (left) lines are obtained with independent priors and red (right) lines with the hierarchical prior.

of information, but at the same time respect the differences. An advantage of using a hierarchical prior is that the mean parameter of the coefficients, β_0 , concentrates the population information coming from all parameters β_i . The posterior estimate of $\mu = e^{\beta_0}$, the population counterfeit rate, is reported in the last row of Table 14.9. This estimate is compared with that obtained from considering that all observations come from the same model; that is, $y_{ij} \sim \text{Poisson}(\mu)$ with $\log(\mu) = \beta$ and prior $\beta \sim N(0, 0.001)$. The estimate of μ from this latter model is also included in the last row in Table 14.9 under the column of Independent (prior). These two estimates show great differences. The model that assumes that all observations come from the same model with a single rate μ produces an interval estimate that is very narrow showing an enormous precision, whereas the interval estimate obtained with the hierarchical model acknowledges the uncertainty coming from the different denomination rates μ_i producing an overall counterfeit rate estimate for μ that is more realistic with a lot less precision. This effect can better be appreciated in the last pair of intervals in Figure 14.7.

14.5 Nonparametric Regression

14.5.1 Bayesian Nonparametric Ideas

The concepts of parametric and nonparametric statistics refer to assumptions that are placed on the distribution of the available observations. One might assume that a particular dataset was generated from a normal distribution with unknown mean and precision. This is a parametric assumption since the $N(\mu, \tau)$ defines a parametric family. In general, a parametric assumption would mean that the dataset is assumed to

be generated from a member of a parametric family. Once a parametric assumption has been placed on a dataset, the objective is to estimate the unknown quantities, which are typically a finite number of parameters that define the model. A nonparametric assumption would imply that the dataset is not generated from a member of a particular parametric family, but from an unknown density (distribution) $f(F)$. Since the whole f is unknown, we can say that the number of parameters to estimate is infinity: all $f(y)$'s values at any specific point y .

The way the Bayesian paradigm treats the unknown quantities is to determine a prior distribution and, via Bayes' theorem, update the prior knowledge with the information given by the data. In a nonparametric assumption the unknown quantities are the whole f (or F), so one is required to place a prior distribution on f . The way we achieve this is by using stochastic processes whose paths are density (or distribution) functions. This leads to the concept of nonparametric priors or random probability measures, because once a stochastic process has been chosen for f , any probability calculated from it, say $P(Y \in B) = \int_B f(y)dy$, is a random variable. According to Ferguson (1973) a nonparametric prior must have large support in the sense that any fixed probability distribution can be arbitrarily approximated by a realization from the prior.

14.5.2 Polya Tree Prior

One of the simplest and most powerful nonparametric priors is the Polya tree (Lavine 1992). To start building the picture, let us consider a probability histogram. This is an estimate of a density function where the sampling space is partitioned into intervals and a bar is placed on top of each interval whose area represents the probability of lying in that particular interval. Imagine that the area assigned to each interval is a random variable such that the sum of all areas (random variables) is constrained to be one (almost surely); then this would be a realization of a (finite) Polya tree. This behavior is illustrated in Figure 14.8. The formal definition of a Polya tree is given in Appendix 14.6.

14.5.3 Semi-Parametric Linear Regression

Let us recall the linear regression model of Section 14.3, where the conditional distribution of the response variable y_i given a set of explanatory variables \mathbf{x}_i is given by $y_i | \mathbf{x}_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \tau)$. If we consider the linear equation with an additive error $\epsilon_i \sim N(0, \tau)$, the same model is reexpressed as $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$. This linear parametric model can be converted into a semi-parametric model by relaxing the distribution f_ϵ of the errors ϵ_i to be nonparametric, for instance, a Polya tree. The normal model can be our prior mean, and we can control how uncertain we are about the normal

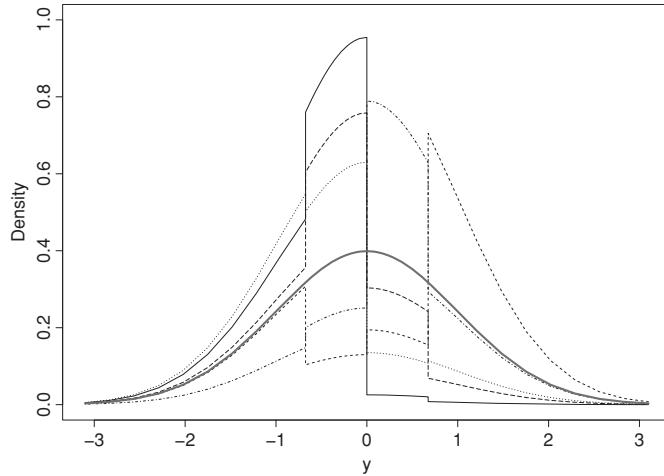


Fig. 14.8. Five realizations (black thin lines) of a finite \mathcal{PT} with $M = 2$ levels and centered on a $N(0, 1)$ density (red thick line), with $a = 1$.

distribution of the errors by controlling the precision parameter a . For $a \rightarrow \infty$ we go back to the parametric normal regression model as a limiting case.

The semi-parametric regression model is then defined as

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \mid f_\epsilon \sim f_\epsilon, \quad f_\epsilon \sim \mathcal{PT}(\Pi, \mathcal{A}), \quad (14.4)$$

with $f_0 = N(0, \tau)$, $\alpha_{mj} = am^2$, and $a > 0$ (see Appendix 14.6 for details). A further adjustment needs to be made. Since $E(f_\epsilon) = f_0$ this implies that $\epsilon_i \sim N(0, \tau)$ marginally (on average). That is, $E(\epsilon_i) = 0$ not always, but only on average. We can force the Polya tree to be centered at zero always (with probability 1) by fixing the first partition of the tree to be $B_{11} = (-\infty, 0]$ and $B_{12} = (0, \infty)$, and taking $\theta_{11} = \theta_{12} = 1/2$ with probability 1. This implies that the median of the random density f_ϵ of each ϵ_i is zero. This is verified by noting that the median of ϵ_i is zero iff $P(\epsilon_i \leq 0) = P(\epsilon_i \in B_{11}) = \theta_{11} = 1/2$. Therefore, the semi-parametric regression model (14.4) is not a mean regression model but a *median regression model* since $\mathbf{x}'_i \boldsymbol{\beta}$ becomes the median of the response variable y_i .

Model (14.4) has two unknown quantities: $\boldsymbol{\beta}$ and f_ϵ . Our prior knowledge on f_ϵ has been placed through the Polya tree, so we also require a prior distribution for $\boldsymbol{\beta}$. The common assumption is to take $\beta_j \sim N(b_0, t_0)$, independently for $j = 1, \dots, p - 1$, as in most (generalized) regression models. The likelihood function for this semi-parametric model is a function of the prior parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ as follows:

$$\text{lik}(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{x}) = \prod_{i=1}^n f(y_i \mid \mathbf{x}_i) = \prod_{i=1}^n f_\epsilon(y_i - \mathbf{x}'_i \boldsymbol{\beta}) = \prod_{i=1}^n \prod_m \theta_{m, j_{\epsilon_i}},$$

with $\epsilon_i = y_i - \mathbf{x}'_i \boldsymbol{\beta}$. Posterior distributions for $(\boldsymbol{\beta}, f_\epsilon)$ will be characterized conditionally. $f_\epsilon | \boldsymbol{\beta}, \mathbf{y}$ is another Polya tree with the distribution of the parameters θ updated with ϵ_i 's as observations. $\boldsymbol{\beta} | f_\epsilon, \mathbf{y}$ is just proportional to the product of the likelihood and the prior, and Metropolis-Hastings steps will be required for sampling from it. More details on this semi-parametric model can be found in Walker and Mallick (1999). Fortunately, posterior inference with this semi-parametric model is implemented in the function `PTlm` from the R library `DPPackage`.

Example 14.5. Consider the insurance dataset described in Example 14.1. This dataset consisted of claim amounts y_i , premiums written x_i , and class indicators z_{ji} , for $j = 1, \dots, 7$ sectors and $i = 1, \dots, 228$ observations. In Example 14.1 a linear model in the logarithmic scale was suggested to describe severity amounts in terms of the premiums written by sector. Let us consider the same linear predictor, but instead of assuming normality of the errors (responses) we will consider nonparametric errors with a Polya tree prior. The new model becomes $y_i = \alpha_1 + \sum_{j=2}^7 \alpha_j z_{ji} + \beta_1 \log(x_i) + \sum_{j=2}^7 \beta_j \log(x_j) * z_{ji} + \epsilon_i$, with $\epsilon_i | f_\epsilon \sim f_\epsilon$ and $f_\epsilon \sim \mathcal{PT}(\Pi, \mathcal{A})$.

The Polya tree is centered on $f_0 = N(0, \tau)$, as in the parametric model, with $\alpha_{mj} = am^2$. We took $a = 1$ and assigned a prior to the error precision $\tau \sim \text{gamma}(0.01, 0.01)$. It is worth mentioning that this latter prior induces a different specification of the Polya tree partitions $\Pi = \{B_{mj}\}$ for every value of τ , producing a mixing over the partitions and thus implying smoothed paths of the tree. We specify a finite tree with a number of partition levels $M = 6$. For the model coefficients we use the same priors as in Example 14.1, that is, $\alpha_j \sim N(0, 0.001)$ and $\beta_j \sim N(0, 0.001)$, for $j = 1, \dots, 7$. The specifications for implementing this model in R with the use of the `DPPackage` library are presented in Table 14.10.

Remember that the semi-parametric regression model just defined is a median regression model with an enhanced flexibility in the specification of the errors. Posterior estimates of model coefficients are included in the last two columns of Table 14.2. The point estimates of all coefficients are numerically different from those obtained with the parametric model, but only few are statistically different. Estimates of β_1 and β_7 present intervals that do not intercept between the parametric and the nonparametric scenarios, thus implying a difference in the slope relationships between $\log(y_i)$ and $\log(x_i)$ for insurance sectors `ACC` and `HEA` comparing parametric and nonparametric fit. Figure 14.9 compares the parametric and the nonparametric fittings for the seven sectors in different panels. The major differences among the two fittings are in the first two sectors `ACC` and `AGR`. In both cases there is an extreme observation that pulls the parametric fit, whereas the nonparametric model is less sensitive to extreme observations, producing a more realistic fit consistent with nonextreme observations.

Table 14.10. R Code for Model of Example 14.5

```

# Initial state
state<-NULL
# MCMC parameters
nburn<-500; nsave<-5000; nskip<-2; ndisplay<-500
mcmc<-list(nburn=nburn,nsave=nsave,nskip=nskip,ndisplay=ndisplay)
# Prior information
prior<-list(alpha=1,beta0=rep(0,14),Sbeta0=diag(1000,14),
             tau1=0.01,tau2=0.01,M=6)
# Fit the model
fit<-PTlm(formula=log(y) ~ z2+z3+z4+z5+z6+z7+log(x)
           +z2*log(x)+z3*log(x)+z4*log(x)
           +z5*log(x)+z6*log(x)+z7*log(x),prior=prior,mcmc=mcmc,state=state,
           status=TRUE)
# Summary with HPD and Credibility intervals
summary(fit)

```

14.6 Appendix. Formal Definition of a Polya Tree

To formally define a Polya tree prior let $\Pi = \{B_{mj}\}$ be a set of binary nested partitions of \mathbb{R} such that at level $m = 1, 2, \dots$ we have a partition of \mathbb{R} with 2^m elements, and the index j , $j = 1, \dots, 2^m$, identifies the element of the partition at level m . For example, at level one ($m = 1$), we have a partition of 2^1 elements B_{11} and B_{12} .

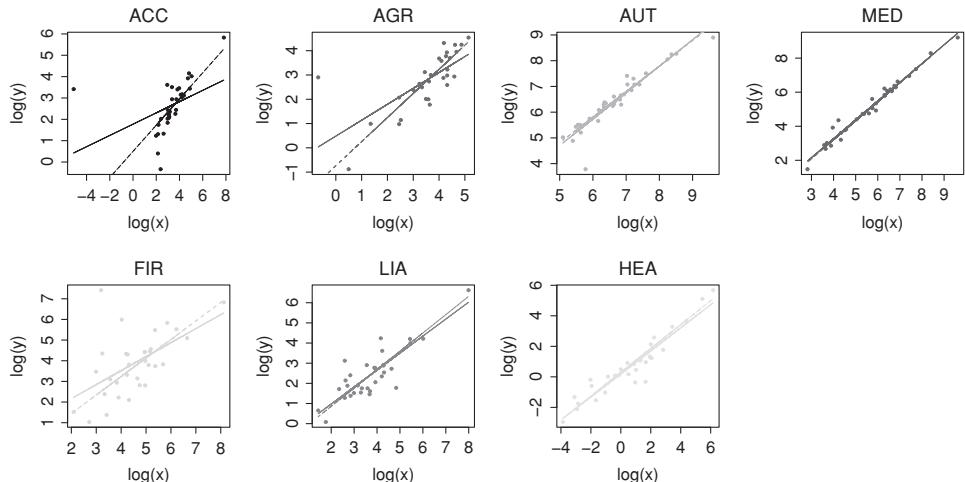


Fig. 14.9. Dispersion diagrams of severity amounts y_i versus premiums written x_i by insurance class $j = 1, \dots, 7$. Solid line (parametric fitting) and dotted line (nonparametric fitting).

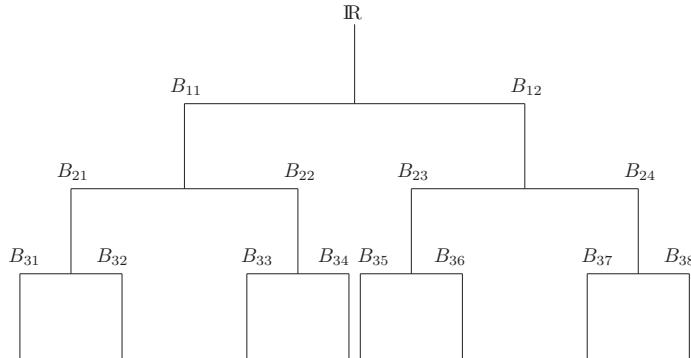


Fig. 14.10. Diagram of nested partitions of \mathbb{R} for three levels.

At level two ($m = 2$) we have a partition of $2^2 = 4$ elements B_{21}, B_{22}, B_{23} , and B_{24} such that (B_{21}, B_{22}) are a partition of B_{11} and (B_{23}, B_{24}) are a partition of B_{12} . In general, at level m , B_{mj} is partitioned into $(B_{m+1,2j-1}, B_{m+1,2j})$ at level $m + 1$ with $B_{m+1,2j-1} \cap B_{m+1,2j} = \emptyset$. Figure 14.10 presents a diagram of these nested partitions for levels $m = 1, 2, 3$.

Let $\boldsymbol{\theta} = \{\theta_{mj}\}$ be a set of parameters such that each θ_{mj} is associated with the set B_{mj} . The parameter θ_{mj} determines the conditional probability of a random variable Y being in the set B_{mj} given that it belongs to the father, $B_{m,(j+1)/2}$ if j is odd, or $B_{m,j/2}$ if j is even. For example, $\theta_{21} = P(Y \in B_{21}|Y \in B_{11})$. Since the two subsets of a father set form a partition of the set, the conditional probabilities must sum to one. In the example, $\theta_{21} + \theta_{22} = 1$, where $\theta_{22} = P(Y \in B_{22}|Y \in B_{11})$. In general $\theta_{m,2j} = 1 - \theta_{m,2j-1}$ for $j = 1, \dots, 2^{m-1}$. Therefore, for the sets at level m , the probability of Y belonging to the set B_{mj} is just the product of all conditional probabilities θ_{mj} , one for each level, to where the set B_{mj} belongs. In notation,

$$P(Y \in B_{mj}) = \prod_{k=1}^m \theta_{m-k+1,r(m-k+1)},$$

where $r(k-1) = \lceil(r(k)/2)\rceil$ is a recursive decreasing formula whose initial value is $r(m) = j$ and that locates the set B_{mj} with its ancestors upward in the tree. $\lceil \cdot \rceil$ denotes the ceiling function. For example, $P(Y \in B_{21}) = \theta_{21}\theta_{11}$. If we continue the partitions down to infinity, we can define the density $f(y|\boldsymbol{\theta})$ for every $y \in \mathbb{R}$ in terms of the parameters $\boldsymbol{\theta}$.

The Polya tree is then defined as the prior distribution for the density $f(y|\boldsymbol{\theta})$. Since $\boldsymbol{\theta}$ is an infinite set, then $f(y|\boldsymbol{\theta})$ is nonparametric (or infinitely parametric). Because θ_{mj} are (conditional) probabilities, they must be in the interval $(0, 1)$, so a natural prior is a beta distribution. Therefore $\theta_{mj} \sim \text{Be}(\alpha_{m,j}, \alpha_{m,j+1})$. If we denote by $\mathcal{A} = \{\alpha_{mj}\}$

the set of all α parameters, then we can denote by $\mathcal{PT}(\Pi, \mathcal{A})$ a Polya tree prior for the density $f(y)$ or for the probability measure $P(\cdot)$.

The Polya tree prior is defined in terms of the partitions Π and non-negative parameters \mathcal{A} . These two sets must reflect our prior knowledge about the unknown density $f(\cdot)$. If we know that the true $f(\cdot)$ should be around a $f_0(\cdot)$ density (e.g., a $N(0, 1)$ density), we can make the prior satisfy $E(f) = f_0$ in the following way (e.g., Hanson and Johnson 2002). Take the partition elements B_{mj} to correspond to the dyadic quantiles of f_0 , i.e.,

$$B_{mj} = \left(F_0^{-1} \left(\frac{j-1}{2^m} \right), F_0^{-1} \left(\frac{j}{2^m} \right) \right], \quad (14.5)$$

for $j = 1, \dots, 2^m$, with $F_0^{-1}(0) = -\infty$ and $F_0^{-1}(1) = \infty$, and F_0 the distribution function corresponding to density f_0 ; then take $\alpha_{mj} = a m^2$ (constant within each level m) such that $\theta_{m,2j-1} \sim Be(am^2, am^2)$ independently for $j = 1, \dots, 2^{m-1}$. This particular choice of the α_{mj} parameters defines an almost surely continuous prior (Ferguson 1974). The parameter a plays the role of a precision parameter: larger values of a make the prior concentrate closer to the mean f_0 , whereas smaller values make the prior “more nonparametric” since the prior will place larger variance around the mean f_0 .

To better understand the Polya tree, Figure 14.8 presents five realizations of a finite Polya tree prior with a total of $M = 2$ levels, producing four elements partitioning the real line. These subsets were defined using the quartiles of an $N(0, 1)$ density as in (14.5). Since we stop partitioning at a finite level M , the density of the points inside the sets B_{Mj} needs to be spread, either uniformly (forming a histogram) or according to f_0 , as in Figure 14.8 with $f_0 = N(0, 1)$. This is achieved by defining $f(y | \theta) = 2^M f_0(y) \prod_{m=1}^M \theta_{m,j_y}$, where the pair (m, j_y) identifies the set B at level m that contains the point y . Each realization of a (finite) Polya tree corresponds to a “histogram” that results from a random perturbation of the centering density f_0 in the sets at level M , B_{Mj} .

In addition to its intuitive definition, a Polya tree has the advantage that its posterior representation is conjugate, following another Polya tree with updated parameters \mathcal{A} . For a sample y_1, \dots, y_n of size n such that $y_i | f \sim f$ and $f \sim \mathcal{PT}(\Pi, \mathcal{A})$, then $f | \mathbf{y} \sim \mathcal{PT}(\Pi, \mathcal{A}^*)$ with $\alpha_{mj}^* = \alpha_{mj} + n_{mj}$ where $n_{mj} = \sum_{i=1}^n I(y_i \in B_{mj})$ is the number of observations y_i 's that fall in the set B_{mj} .

Acknowledgments

The authors are grateful to the comments from the editors and two anonymous referees. The authors also acknowledge support from “Asociacion Mexicana de Cultura, A.C.”

References

- Bailey, A. (1950). Credibility procedures: Laplace's generalization of Bayes' rule and the combination of collateral knowledge with observed data. *Proceedings of the Casualty Actuarial Society* 37, 7–23.
- Banerjee, S., B. Carlin, and A. Gelfand (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall, Boca Raton.
- Bellhouse, D. (2004). The Reverend Thomas Bayes, frs: A biography to celebrate the tercentenary of his birth. *Statistical Science* 19(1), 3–43.
- Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis* 1, 385–402.
- Bernardi, M., A. Maruotti, and L. Petrella (2012). Skew mixture models for loss distributions: A Bayesian approach. *Insurance: Mathematics and Economics* 51(3), 617–623.
- Bernardo, J. and A. Smith (2000). *Bayesian Theory*. Wiley, New York.
- Bühlmann, H. (1967). Experience rating and probability. *ASTIN Bulletin* 4, 199–207.
- Cabras, S. and M. E. Castellanos (2011). A Bayesian approach for estimating extreme quantiles under a semiparametric mixture model. *ASTIN Bulletin* 41(1), 87–106.
- Cairns, A. J., D. D. K. Blake, G. D. Coughlan, and M. Khalaf-Allah (2011). Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin* 41(1), 29–59.
- Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1, 209–230.
- Ferguson, T. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* 2, 615–629.
- Gelman, A., J. Carlin, H. Stern, and D. Rubin (2004). *Bayesian Data Analysis*. Chapman and Hall, Boca Raton.
- Gelman, A., A. Jakulin, M. Grazia-Pittau, and Y.-S. Su (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics* 2, 1360–1383.
- Hanson, T. and W. Johnson (2002). Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association* 97, 1020–1033.
- Klugman, S. (1992). *Bayesian Statistics in Actuarial Science: With Emphasis on Credibility*. Huebner International Series on Risk, Insurance, and Economic Security. Kluwer Academic Publishers, Boston.
- Landriault, D., C. Lemieux, and G. Willmot (2012). An adaptive premium policy with a Bayesian motivation in the classical risk model. *Insurance: Mathematics and Economics* 51(2), 370–378.
- Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *Annals of Statistics* 20, 1222–1235.
- Lundberg, O. (1940). *On Random Processes and Their Application to Sickness and Accident Statistics*. Almqvist and Wiksell, Uppsala.
- Makov, U. E. (2001). Principal applications of Bayesian methods in actuarial science: A perspective. *North American Actuarial Journal* 5(4), 53–73.
- Makov, U. E., A. F. M. Smith, and Y.-H. Liu (1996). Bayesian methods in actuarial science. *Journal of the Royal Statistical Society. Series D (The Statistician)* 45(4), 503–515.
- Scollnik, D. (2001). Actuarial modeling with MCMC and BUGS. *North American Actuarial Journal* 5(4), 96–125.
- Shi, P., S. Basu, and G. Meyers (2012). A Bayesian log-normal model for multivariate loss reserving. *North American Actuarial Journal* 16(1), 29–51.
- Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* 64, 583–639.

- Walker, S. and B. Mallick (1999). Semiparametric accelerated life time model.
Biometrics 55, 477–483.
- West, M. (1985). Generalized linear models: Scale parameters, outlier accommodation and prior distributions (with discussion). In J. M. Bernardo, M. H. DeGroot, and A. Smith (Eds.), *Bayesian Statistics 2*, pp. 531–558, Wiley, New York.
- West, M. and J. Harrison (1997). *Bayesian Forecasting and Dynamic Models*. Springer, New York.
- Whitney, A. W. (1918). The theory of experience rating. *Proceedings of the Casualty Actuarial Society* 4, 274–292.

15

Generalized Additive Models and Nonparametric Regression

Patrick L. Brockett, Shuo-Li Chuang, and Utai Pitaktong

Chapter Preview. Generalized additive models (GAMs) provide a further generalization of both linear regression and generalized linear models (GLM) by allowing the relationship between the response variable y and the individual predictor variables x_j to be an additive but not necessarily a monomial function of the predictor variables x_j . Also, as with the GLM, a nonlinear link function can connect the additive concatenation of the nonlinear functions of the predictors to the mean of the response variable, giving flexibility in distribution form, as discussed in Chapter 5. The key factors in creating the GAM are the determination and construction of the functions of the predictor variables (called smoothers). Different methods of fit and functional forms for the smoothers are discussed. The GAM can be considered as more data driven (to determine the smoothers) than model driven (the additive monomial functional form assumption in linear regression and GLM).

15.1 Motivation for Generalized Additive Models and Nonparametric Regression

Often for many statistical models there are two useful pieces of information that we would like to learn about the relationship between a response variable y and a set of possible available predictor variables x_1, x_2, \dots, x_k for y : (1) the statistical strength or explanatory power of the predictors for influencing the response y (i.e., predictor variable worth) and (2) a formulation that gives us the ability to predict the value of the response variable y_{new} that would arise under a given set of new observed predictor variables $x_{1,new}, x_{2,new}, \dots, x_{k,new}$ (the prediction problem).

Linear regression models as presented in Chapter 2 were the first to address these problems and did so under the assumption that the mean response y relates to the predictors x_j 's via an additive linear functional of the x 's (i.e., the value of y is an additive summation of the parametric monomials $\beta_j x_j$'s plus some random error).

Thus, the i^{th} observation $(y_i, x_{i1}, x_{i2}, \dots, x_{ik})$ is assumed to follow the parametric relationship

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i,$$

where the error terms ε_i are independent and identically distributed (i.i.d.) with $E(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$. Accordingly, the mean value of y satisfies $\mu_y = E[y|x_1, \dots, x_k] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$. Estimation of the β parameters is performed by least squares. For inferential purposes, the ε_i variates are usually assumed to be i.i.d. normal random variables, so y also becomes normal (conditional on the observed values of the x_i 's). Normality is not needed for prediction, but distributions are needed for inference.

Linear regression is widely used because it provides a simple, easy linear way to explain and interpret the relationship, and because the contribution of the predictor variable is crystallized into one number, namely the predictor variable's β coefficient. Under the usual normality assumption, the sampling distribution of estimates of the β coefficient can be determined, which makes inference concerning the significance of the predictor variable easy to assess. Moreover, the relative predictive strength among the k predictors can be examined and compared by simply looking at the standardized magnitudes of their coefficients β_j . Using such an equation for prediction is also simple. If we are given a new data point with predictor variables $\{x_{1,new}, x_{2,new}, \dots, x_{k,new}\}$, the prediction of y_{new} is simply the expected value or mean of y from the above equation, namely $\hat{y}_{new} = \mu_y = \beta_0 + \beta_1 x_{1,new} + \beta_2 x_{2,new} + \cdots + \beta_k x_{k,new}$.

That simple linear model can be relaxed in several ways. First one can preserve the additive linear parametric monomial structure for the contribution of the x_i 's, but instead of predicting the mean response μ_y of the dependent variable y , one predicts a function of the mean of y in a linear and additive fashion (i.e., $g(E[y]) = g(\mu_y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$). Chapter 5 on generalized linear models (GLMs) addressed those problems wherein the response variable y , does not necessarily have normal distribution and the mean of y is not a simple linear function of the x variables, but rather $g(\mu_y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ or, stated differently, $\mu_y = g^{-1}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$. The function g in this relationship is called the *link function*. Consequently, in the GLM model, the mean or expected value of y is related to the x 's only as a function $g^{-1}(\eta)$ where $\eta(x) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$ (here $\eta(x)$ is called the *systematic component*). It follows that the impact of x exhibits itself through the linear concatenation η , and inference can be pursued under the statistical assumption that y belongs to an exponential family of distributions. The process is still parametric, but is more flexible

than simple linear regression. Instead of least squares estimation, maximum likelihood estimation is used once the exponential family has been specified.

A second way to relax the assumptions in the simple linear model is to relax the parametric structure on the right hand side of the regression equation (i.e., replace the systematic component $\eta(x)$ by a more general function of the x_j 's). This leads to *nonparametric regression*. Nonparametric regression poses the model $y = F(x_1, \dots, x_k) + \varepsilon$ where the problem is to find a multivariate function F that “appropriately fits” the collection of given observations on y and x_1, \dots, x_k , neither overfitting nor oversmoothing the relationship. In its generality, this model can be very difficult to address for more than a single predictor variable, and it requires much larger data samples as the number of predictor variables increases (the “curse of dimensionality”). Hence, more structure is usually assumed. A compromise between the fully general nonparametric regression model $y = F(x_1, \dots, x_k) + \varepsilon$ and the linear parametric model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$ is to fit a model that retains the additive structure but does not impose the strict monomial linear structure $\beta_j x_j$ term accounting for the contribution of the j^{th} variable. The impact of x_j on y is now a nonparametric function of x_j . Replacing $\beta_j x_j$ by a more general function $f_j(x_j)$, which is then summed over the predictor variables, provides a more structured form of nonparametric regression that is referred to as the *additive model* (discussed subsequently)

$$y = \alpha + \sum_{j=1}^k f_j(x_j) + \varepsilon, \quad (15.1)$$

where the error ε is independent of the x_j 's, $E(\varepsilon) = 0$, $\text{Var}(\varepsilon) = \sigma^2$, and the functions f_j are constructed to appropriately fit the relationship. For example, the impact of driver age x on accident claim size y in automobile insurance is not linear, but rather is a more U-shaped function $f(x)$, with higher claims at very young and at very old ages, and smaller claims at intermediate ages. Additive models seek to determine the appropriate functions f_j from the data.¹ To avoid identifiability problems in (15.1) associated with constant terms in the f_j confounding the estimation of α , we center the f_j and assume $E[f_j(X_j)] = 0$ so $E[y] = \alpha$.

The final generalization we pursue is to introduce a link function into (15.1), and we then obtain what is known as the *generalized additive model* (GAM). The GAM has the same relationship to the additive model as the generalized linear model (GLM) has to the simple linear model. As does the GLM, it allows additional flexibility in the

¹ These structure additive models can be extended to additively incorporate additional functions $f_{ij}(x_i, x_j)$ into the regression equation (15.1) if there is significant joint influence of (x_i, x_j) on y . We do not pursue this extension here.

choice of y (such as a binary variable) via the link function as described in Chapter 5, as well as flexibility to have highly nonlinear relationships between the predictor variables and the dependent variable as in additive models.

A brief synopsis of this chapter is as follows. In the next section we introduce the additive model as a structured form of nonparametric regression and discuss methods for determining the functions f_j . This is done first for a single x and y ($k = 1$ in (15.1), or general univariate nonparametric regression); subsequently we discuss the more general situation of $k > 1$ in (15.1). The functions in (15.1) are called *smoothers*, and several different possible types are discussed.

Next we discuss algorithms for fitting the model and the trade-off between fit and smoothness. After the discussion of the additive model, we introduce a link function analogous to the GLM link function so as to obtain the GAM formulation. Fit, smoothness, and inference are discussed. We illustrate the techniques along the way with an example involving the prediction of coronary heart disease, an important consideration in life insurance.

15.2 Additive Models for Nonparametric Regression

The *additive nonparametric regression model* (or simply additive model) is defined to be

$$y = \alpha + \sum_{j=1}^k f_j(x_j) + \varepsilon, \quad (15.2)$$

where ε is random error and the functions are not determined ahead of time, but rather are flexible and are sought to be defined to fit the relationship. For simplicity, usually each smoother f_j is selected as a univariate function.² The individual functions f_j in the additive model can be thought of as analogous to the coefficients in simple regression, and one encounters all the interpretation pitfalls found in linear regression; sometimes the interpretation difficulties are more severe (Hastie and Tibshirani 1999). The smoothers are concatenated in an additive manner in describing the overall relationship of y to the x 's in the additive model. However, actually writing down a closed-form mathematical formula for the relationship between x and y is infeasible because it may be defined by segment, so graphical representations are often used for displaying the estimated f_j 's.

The individual functions f_j in (15.2) are the building blocks that attempt to smoothly capture (possibly nonlinear) trends and patterns between the response

² Nonparametric regression (curve fitting of the response surface) can require substantial data and is hard to interpret and to present graphically in higher dimensions. The additive model is a compromise between relaxing the linearity assumption in multivariate regression and unrestricted curve fitting in general high-dimensional nonparametric regression.

variable y and the predictor variables x_j . If the f_j 's are globally polynomials, then (15.2) is polynomial regression, and the problem becomes a parametric optimization in the polynomial coefficients. In general, however, the structural relationship between y and $f_j(x_j)$ may change as the values of x_j change over its range (i.e., the same polynomial or Box-Cox transformation may not fit the relationship over the entire range of x values). Therefore, in the spline regression approach to the additive model the estimate of f_j is defined in a piecewise fashion in local neighborhoods of the x_j values. Since the relationship between y and $f_j(x_j)$ is assumed to be smooth and continuous, the functions f_j are called *smoothers*.

Three main decisions involving smoothers are the choice of the type or class of smoothers, the choice of the size of the local neighborhoods within which the relationship is fit, and the selection of how smooth the globally piecewise patched together function will be. Smoothers basically differ on the way they summarize the trend or average the response y within the area of fit. The size of the neighborhood regions used for fitting involves a fundamental trade-off between the bias and variance. Large neighborhoods tend to mitigate the variance (be smoother), whereas small regions keep the bias low (more closely reproduce the observed data).

15.2.1 Elementary Smoothers: Univariate Nonparametric Regression Case

The additive model does not start with a specific global parametric form for variables over the entire estimation interval (as in linear regression), but instead it utilizes some smooth functions to form the estimation. We observe the data and let the data show what smooth functions we should use, since no parametric forms will be postulated a priori. For exposition purposes we discuss the process of finding the appropriate function f in the univariate case first (i.e., one y and one x with $y = \alpha + f(x) + \varepsilon$ where ε is the error term) and then discuss multivariate extensions.

A scatterplot is a simple but powerful diagrammatic tool used to demonstrate the relationship between two variables. After observing the data, however, we may want to develop a line or curve, f , through the scatterplot that parsimoniously illustrates the pattern between the variables represented in the scatterplot. Intuitively, scatterplot smoothers (such as the running average and the running line smoother, to be discussed subsequently) are candidates for numerically obtaining an estimate of this f .

The simplest smoother when x is a categorical variable is to take an average of y for each distinct value of x . If we apply this type of smoother to a predictor x , the smoother would be the average value of y for each category of x , and the plot of x versus y would look piecewise constant (like a bar chart). A bin smoother extends this idea to noncategorical predictors by partitioning the values of x into a finite number of bins or neighborhoods, and then assigning the average value of y to each bin. The resulting function again looks like piecewise constant, and the function f so

developed does not provide a very smooth estimator because there will be jumps at each cut point (bin boundary).

Running average smoothers (also called moving mean smoothers) attempt to create a smoother estimation by fading out the abrupt transition at the cut points between bins by utilizing overlapping regions in the estimation. Suppose all the values of x are sorted and there are no replications in the data. One method for determining a region for estimating y at the predictor point x_0 is to define a neighborhood that covers the target data point x_0 together with a chosen number k of points to the left of x_0 and k points to the right of x_0 . The value of this smoothing function is the mean of y over the neighborhood region that centers at x_0 . The choice of the number of points to the right and left of x_0 controls the responsiveness of the running average smoother to the data. If $k = 0$ points are to the right and left, you have a very rough estimator since in this case you are only considering the single-point neighborhood (y_0, x_0) . You then obtain a “connect-the-dots” estimator, which perfectly reproduces the observed data, but is also perfectly useless since it is reproducing the noise as well as the relationship in the data that you wish to uncover. If you take $k = n/2$ points and include all the data in the estimate (making the neighborhood consist of the entire range of x values), then you have smoothed the response so much that a single mean is produced that may not reflect any actual local variation present in the response of y to x . A compromise is to define the neighborhoods by taking a proportion h (between zero and one) of the n data points for each neighborhood. The number h is called the *span* (it is also called the bandwidth), and it controls the smoothness. The neighborhood is defined by including $k = \frac{[hn]-1}{2}$ data points to the left and right of each predictor point. A larger h will introduce more data for each neighborhood, and therefore the resulting estimation will be smoother.

There are two problems with the running average smoother: first, at the endpoints of the interval, the neighborhoods are not symmetrical so there is estimation bias at the end points, and second, in fact, this smoother is often not very smooth (see Figure 15.1). Thus, we are led to another neighborhood based smoother around x_0 called the *running line smoother*. The running line is a variation of the previous approach, but instead of averaging the response variable y over the local interval as in the running average, it fits a regression line over the neighborhood region centered at x_0 and then uses this regression line to predict a smoother value for that y observation corresponding to x_0 .

One of the reasons for the lack of smoothness in the running average smoother is the “all or nothing” contribution of the various observations (x_i, y_i) , with each y_i corresponding to a point in the neighborhood of x_0 contributing equally to the mean estimate of y at x_0 . If y_i is from a point in the neighborhood (regardless of if its x_i is very close to x_0 or is at the end of the neighborhood far from x_0), it contributes to the average, whereas it contributes nothing if it is outside the neighborhood. To further smooth running average predictions, weights can be applied to the adjacent

data points. The highest weight is given to x_0 itself, and, because we think points that are close together are similar, the weights are then defined in a decreasing and smooth fashion as they move farther away from x_0 toward the boundaries or breakpoints of the neighborhood that define the local bin. This approach is along the line of a popular smoother called *locally weighted running-line smoother*, LOcal regrESSion, or LOESS, implemented by Cleveland (1979).

A weight function is used in a Loess smoother to address the issue that, if we calculate locally, those data that are distant from the designed target point x_0 at which we want to estimate the response y should make only a relatively small contribution to the estimate (i.e., correct the all-or-nothing problem referred to earlier for running average smoothers). The weight function in a Loess is designed to weight each datum within the neighborhood according to the distance between the datum and the designated point – with higher weights for the points near the designated point and lower weights for more distant points. Compared to scatterplot smoothers, a Loess can provide a more general and smoother pattern. However, a Loess could estimate an incorrect pattern due to the influential outliers. Accordingly, Cleveland (1981) proposes a more sophisticated weight function to reduce the impacts from the outliers, which he calls the Lowess smoother. A Lowess is a robust version of a Loess, and it can be iteratively calculated in a short time, which is one reason that it is popular. The Lowess iteratively reweights the Loess estimate to improve performance. Lowess is now available for use in statistical packages (e.g., SAS and mgcv available in R have these smoothers as options when performing additive modeling).

Example 15.1 (Predicting Coronary Heart Disease:). Examples of Running Average, Running Line, Loess, and Lowess Estimators Coronary heart disease and cancer combined constitute more than half of all deaths among people over the age of 40 (Johnson et al. 2005, p. 109), and hence life insurers and annuity providers are clearly interested in noninvasive methods of detecting coronary heart disease in applicants for insurance. Accordingly, for a numerical illustration of the techniques described earlier we use a heart disease dataset from the University of California at Irvine Machine Learning Repository. The data are available at <http://archive.ics.uci.edu/ml/datasets/Heart+Disease>. The variables we use are shown in Table 15.1. In this example we examine the different types of smoothers using the continuous variable predictors, with diagnosis of heart disease as the response variable (the incorporation of the categorical predictors being considered later).

Figure 15.1 shows the running average and running line smoother while Figure 15.2 shows the Loess and Lowess smoother for the continuous predictor variables. The span h is set as 0.35. The purpose of this set of graphs is to help us observe the relationship between the response variable and an explanatory variable and to compare the different types of smoothers.

Table 15.1. Definitions for variables used in Heart Disease Prediction (from University of California at Irvine Machine Learning Repository)

Variable Name	Definition
Sex	0: Female 1: Male
Cp	Chest pain type 1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic
Fbs	Fasting blood sugar 0: > 120 mg/dl 1: ≤ 120 mg/dl
Res	Resting electrocardiographic results 0: normal 1: having ST-T wave abnormality 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
Num	Diagnosis of heart disease 0: $< 50\%$ diameter narrowing 1: $> 50\%$ diameter narrowing
Age	Age in years
Threstbps	Resting blood pressure (in mm/Hg on admission to the hospital)
Chol	Serum cholesterol in mg/dl
Thalach	Maximum heart rate achieved

Computations were done using computer package R, which is available for download at <http://cran.r-project.org/bin/windows/base/>. The package `mgcv` is designed for applying GAMs. As can be seen, the running average does not smooth the data sufficiently, and the running line technique is preferred.

15.2.2 Kernel Smoothers: Univariate Case

The kernel smoother explicitly defines a set of local weights by describing the shape of the weight function via a density function³ with a scale parameter that adjusts the size and the form of the weights near a target predictor point x_0 . The kernel is a continuous, bounded, and symmetric real function that integrates to one. Kernel smoothers usually define weights in a smoothly decreasing manner as one moves away from the target point x_0 whose value y is to be estimated. The Gaussian density

³ The weight function $W(z) = (1 - |z|^3)^3$ for $z < 1$ (zero otherwise) is used in the Loess smoother for weighting the impact of the points x_i a distance $|z|$ from the target point x_0 , and this is not a probability density.

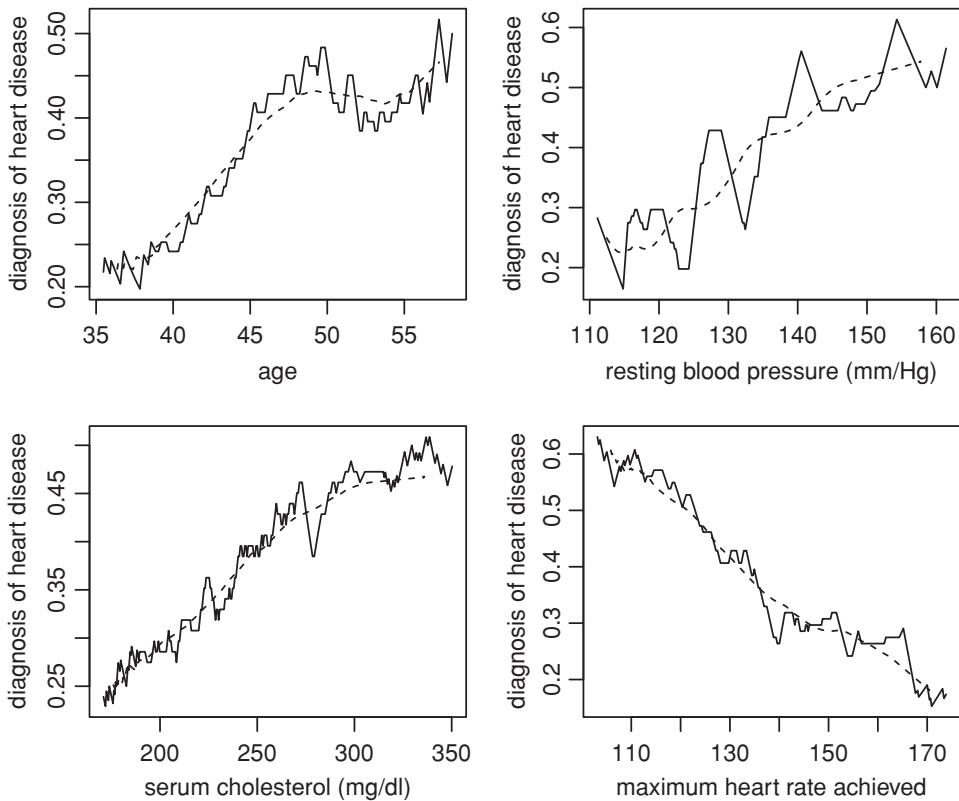


Fig. 15.1. Running average and running line smoothers for the continuous predictor variables. The solid line is the running average, and the dashed line is the running line smoother.

is one candidate, and the resulting estimator is usually called the Gaussian kernel smoother. Running average smoothers can also be considered as kernel smoothers that use a “box” kernel. Most often the kernel is chosen to have bounded support.⁴

Kernel smoothers use a weight function producing the value of the estimator at point x_0 as a weighted sum of the responses. The weight function has its peak at the point of estimation x_0 and gradually decreases as it moves away to both sides. The weights are usually standardized so that they sum to unity. A simple weight function for weighting the contribution of the point x_i to the estimation of y at the point x_0 is, $S(x_i, x_0, \lambda) = \frac{1}{\lambda} K\left(\frac{x_0 - x_i}{\lambda}\right)$, where $K(\cdot)$ is supported on $[-1, 1]$ and peaks at zero.⁵

⁴ The use of kernels with unbounded support can produce some global bias difficulty in the estimators, an issue that can be restricted to boundary points only if one uses the bounded support kernels (cf., Eubank 1999, exercises, p. 217). Thus, bounded supported kernels are generally preferred.

⁵ Note that any kernel with finite support can be scaled to have support on $[-1, 1]$.

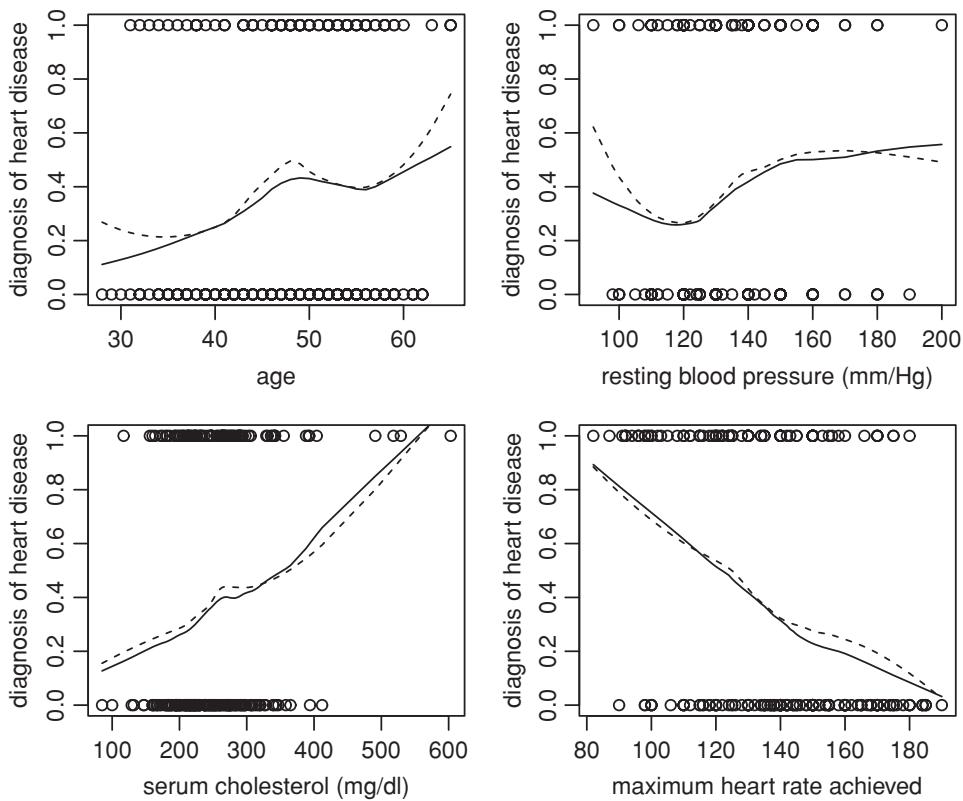


Fig. 15.2. Loess and Lowess smoothers for the continuous predictor variables. The solid line is the Lowess, and the dashed line is the Loess smoother.

To ensure that the predicted response value of y corresponding to the predictor value x_0 is most heavily influenced by the actual value y_0 , the kernel estimators are typically computed by

$$f(x_0) = \frac{1}{n\lambda} \sum_{i=1}^n K\left(\frac{x_0 - x_i}{\lambda}\right) y_i. \quad (15.3)$$

The parameter λ is called the estimator bandwidth, and it determines the amount of smoothing. It is allowed to be any non-negative number. If $K(u)$ has support on $[-1, 1]$, then $K(u/\lambda)$ has support on $[-\lambda, \lambda]$. This means the bandwidth determines how far away observations are allowed to be from x_0 and still contribute to the estimate $f(x_0)$ of y at x_0 . In this case the parameter λ is the scaling parameter of this density function. The bandwidth also controls the pointiness of the weight function, affecting the level of dependency of the estimator on the observations near the target x_0 . A small value of λ will cause the estimate to depend heavily on the data near x_0 , and hence will result in a rougher or more irregular nonsmooth estimate as one moves

Table 15.2. Common Second-Order Kernels for Use in
Additive Model Estimation (from Eubank 1999)

Kernel	K	V	M_2
Rectangle	$K(u) = \frac{1}{2}I_{[-1,1]}(u)$	$\frac{1}{2}$	$\frac{1}{3}$
Quadratic ^a	$K(u) = \frac{3}{4}(1-u^2)I_{[-1,1]}(u)$	$\frac{3}{5}$	$\frac{1}{5}$
Biweight	$K(u) = \frac{15}{16}(1-u^2)^2I_{[-1,1]}(u)$	$\frac{5}{7}$	$\frac{1}{7}$

^a The quadratic kernel $K(u) = \frac{3}{4}(1-u^2)I_{[-1,1]}(u)$ is also called the Epanechnikov kernel (Epanechnikov 1969). It minimizes asymptotic mean squared error of the estimate.

from point to point along the x axis. In contrast, a larger λ will allow the estimate to be averaged over a larger region and provide a smoother estimate. The choice of λ is a decision to be made ahead of time, often by trial and error, using several values of λ and choosing the estimator whose smoothness and performance are considered to adequately describe the relationship.

The kernel K is usually selected to have the following properties:

$$\int_{-1}^1 K(u)du = 1, \quad (15.4)$$

$$\int_{-1}^1 uK(u)du = 0. \quad (15.5)$$

Condition (15.4) is to have weights sum to one, whereas (15.5) is a symmetry condition that addresses estimator bias. For the purpose of inference, K is sometimes selected to have two additional properties:

$$M_2 = \int_{-1}^1 u^2 K(u) du \neq 0, \quad (15.6)$$

$$V = \int_{-1}^1 K(u)^2 du < \infty. \quad (15.7)$$

Conditions (15.6) and (15.7) are added to (15.4) and (15.5) to improve the asymptotic behavior of the estimate \hat{f} of f as $n \rightarrow \infty$ and $\lambda \rightarrow 0$, with (15.6) used to control the asymptotic bias and (15.7) used to make sure the estimator has finite asymptotic variance (see Eubank 1999). A function K with all these properties combined is called a *second order kernel*.

Examples of second-order kernels are shown in Table 15.2.

15.2.3 Regression Smoothers: Univariate Case

The use of smoothers as we discussed with a locally weighted running line or local linear regression extends to polynomial regression. The poor global performance of the fit using just a single polynomial over the entire interval of possible x values is overcome by fitting the polynomial functions separately over local regions (neighborhoods), and then obtaining a global fit by forcing the piecewise polynomials to join together smoothly at the breakpoints of the local regions. Constraints on the polynomial functions' parameters can be introduced such that the curves coincide smoothly at the neighborhood boundaries. This is called a *regression spline*. The regions of fit in a regression spline are separated by breakpoints, which are called *knots*. The resultant piecewise polynomial components are forced to join smoothly at these knots by adding constraints that the first and second derivatives of the neighboring polynomial functions match at the adjacent knots. The most popular regression spline involves piecewise cubic polynomials constrained to be continuous and have continuous first and second derivative at the specified knots.⁶

15.2.4 Cubic Smoothing Splines: Univariate Case

A general measure of fit of an arbitrary smoother f to the data is given by

$$\sum_{i=1}^n \{y_i - f(x_i)\}^2 + \lambda \int_a^b \{f''(t)\}^2 dt, \quad (15.8)$$

where λ is a fixed constant and $''$ represents twice differentiation. The first term measures the closeness of fit of the predicted value $f(x_i)$ to the observed value y_i , whereas the second term penalizes the fit for overall nonsmoothness. Of course one could always fit a curve that exactly went through the data points (y_i, x_i) , but such a curve would be very irregular and rough due to random variations in the data, and it would capture the idiosyncrasies of the particular sample as opposed to the sought-after more general relationship between the response variable and the predictor variable, which is assumed to be smooth and continuous. The relative trade-off between fit and smoothness is governed by the parameter λ . It can be shown that the general optimization problem of minimizing (15.8) for a fixed λ has an explicit unique minimizer that is a natural cubic spline (cf., Hastie and Tibshirani 1999, p. 27; Reinsch 1967, Wasserman 2006, Theorem 5.73; Wood 2006, p. 148).

⁶ A useful variation is natural cubic splines. These are cubic splines with the additional constraint that the function must be linear after the knots at the endpoints to cope with the high variance near the boundaries. Because of data limitations at the endpoints, kernel smoothers suffer from bias at the estimation interval endpoints and thus require more consideration of how to handle the estimation at the endpoints. The forced linearity at the endpoints addresses this concern.

While knowing that the minimizer of (15.8) is a natural cubic spline is useful, it does not make it clear how to actually explicitly obtain the function f . We need a computationally effective manner of doing this, and one such way is based on Theorem 5.74 of Wasserman (2006).

Consider the function space consisting of all cubic splines with knots at specified points $\{t_1, \dots, t_q\}$ where $a \leq x_1 \leq t_1 < t_2 < \dots < t_q \leq x_n \leq b$. Wasserman (2006; Theorem 5.74) proves that a basis set for this function space is the set of functions $b_0(x) = 1$, $b_1(x) = x$, $b_2(x) = x^2$, $b_3(x) = x^3$, and $b_j(x) = (x - t_{j-3})^3$ for $j = 4, 6, \dots, q + 3$. Since the functions $\{b_0, b_1, \dots, b_{q+3}\}$ form a basis, any member of this space can be written as a linear combination of these basis elements.⁷ In particular the cubic spline f that minimizes (15.8) is a linear combination of the above basis elements, i.e.,

$$f(x) = \sum_{i=0}^{q+3} \beta_i b_i(x). \quad (15.9)$$

For a fixed smoothness parameter λ , the determination of the cubic spline minimizing (15.8) involves the computation of the β parameters in (15.9). Within this basis-oriented approach, the fact that $f(\cdot)$ is linear in the basis element coefficients β parameters allows one to formulate the optimization in matrix notation and exploit linear algebraic methodology to optimize, as shown in Wood (2006) and Wasserman (2006). The formulation is as follows.

Since $f(\cdot)$ is linear in the β parameters, the penalty term can be written as a quadratic form in β as

$$\int_a^b \{f''(t)\}^2 dt = \beta' A \beta,$$

where A is a matrix of known coefficients. To see this, write $\mathbf{Q}'(x)$ as the (row) matrix of second derivatives of the basis functions $\{b_j''(x) : j = 0, \dots, q + 3\}$ so $f''(x) = \mathbf{Q}'(x)\beta$ and the penalty term is $\beta' A \beta$ where

$$A = \int \mathbf{Q}'(x) \mathbf{Q}(x) dx, \quad (15.10)$$

(i.e., $A_{i,j}$ is the number $\int b_i''(x) b_j''(x) dx$, which is computable since b_j'' is at most linear). Note also that the first two rows and columns of A are $\mathbf{0}$ because these involve second derivatives of basis elements b_0 and b_1 , which are at most linear. A large λ then puts most weight on these two b_j 's.

⁷ One can construct other bases for the space of cubic splines, and these could be used in (15.9) instead. Wood (2006, chapter 4) discusses several bases and notes the differences that may arise in computation when using different bases. Since our goal is to provide an intuitive theoretical framework, we do not delineate these other bases here.

The first term in (15.8) – the squared error of the fit at the knot points – can also be written in matrix form.⁸ Therefore the penalized regression spline fitting of (15.8) is to minimize β over

$$\|\mathbf{y} - \mathbf{M}\beta\|^2 + \lambda\beta' \mathbf{A}\beta,$$

where \mathbf{M} is the matrix with (i, j) entry $m_{i,j} = b_i(x_j)$ (cf., Hastie and Tibshirani 1999; Wood 2006). Given λ , the formal expression of the penalized least squares estimate of β is

$$\hat{\beta} = (\mathbf{M}'\mathbf{M} + \lambda\mathbf{A})^{-1}\mathbf{M}'\mathbf{y}. \quad (15.11)$$

Since this is a regression, the corresponding “hat matrix” derived from normal equations⁹ can be written (Wood 2006, p. 129) as

$$\mathbf{S}_\lambda = \mathbf{M}(\mathbf{M}'\mathbf{M} + \lambda\mathbf{A})^{-1}\mathbf{M}', \quad (15.12)$$

and

$$\hat{\mathbf{y}} = \mathbf{M}\hat{\beta} = \mathbf{S}_\lambda\mathbf{y}.$$

For actual computation, using orthogonal matrix formulations offers greater stability in computation than the intuitive matrices \mathbf{M} and \mathbf{A} (Wood 2006, p. 129). Computer programs such as those written in R (e.g., the program mgcv discussed in Wood 2006, chapter 5) exploit these representations for computation.

In many applications the knot locations $\{t_1, \dots, t_q\}$ are not prespecified, but rather are taken to be the data points x_1, \dots, x_n themselves (plus the endpoints a and b). This smoothing spline has a knot for each data point (and endpoints). Although in doing this the number of internal knots is the same as the number of observations, the smoothing penalty term in (15.8) will impose a penalty for some spline coefficients to enforce smoothness. A high enough value of λ in the objective function (15.8) results in a much smoother function than n degree of freedom might suggest. This occurs since some coefficients will be shrunk in magnitude because of λ , so that the effective degree of freedom in the fit is much less. If we let $\lambda \rightarrow \infty$, then the fit produces linear regression with constant slope.¹⁰ If we take $\lambda \rightarrow$ zero, the change of slope is less penalized so that smoothness hardly matters, and the fit produces more irregular curves, approaching an interpolation fit.

⁸ Writing the least squares (first) component of (15.8) in matrix form is similar to ordinary least squares, but where the observed independent variables are the basis elements evaluated at the knots.

⁹ The matrix that transforms the observed vector of \mathbf{y} 's into the predicted values of \mathbf{y} 's denoted by $\hat{\mathbf{y}}$ is called the *hat matrix* (since it transforms the \mathbf{y} 's to the $\hat{\mathbf{y}}$) in ordinary least squares regression. Since $\hat{\mathbf{y}} = \mathbf{M}\hat{\beta} = \mathbf{S}_\lambda\mathbf{y}$, the matrix \mathbf{S}_λ in (15.12) is also called the hat matrix in the additive model context. In ordinary regression, the hat matrix is $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ where \mathbf{X} is the design matrix.

¹⁰ The first two rows and columns of \mathbf{A} are $\mathbf{0}$, so a larger value of λ puts much more weight on the first two basis elements, which are the linear part of the linear basis set.

15.2.5 Choosing the Smoothing Parameter λ : Univariate Case

Although the cubic smoothing spline allows great flexibility in fitting data, selecting an appropriate value of the smoothing parameter λ for a given set of data generates a new problem. One can adopt a trial-and-error process, fitting the model for various values of λ and selecting the smoothing parameter λ that gives what appears to be a parsimonious trade-off between smoothness and fidelity to the data (and a graphical relationship between x and y that is consistent with domain area knowledge of the amount of smoothness that one would expect should be present in the relationship). There are, however, more formal methods for determining λ .

One method for selecting λ suggested in Hastie and Tibshirani (1999) is to first decide how many degrees of freedom (effective dimensionality) one is willing to use for creating a model that fits the data. By analogy with ordinary linear models where one can write $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ where \mathbf{H} is the so-called hat matrix,¹¹ and where the degrees of freedom in the ordinary linear model equal the trace $Tr(\mathbf{H})$ of \mathbf{H} , in this more general setting one defines the *effective degrees of freedom* as the trace of the hat matrix S_λ in (15.12). Although not exactly identical to the usual definition of degrees of freedom, the effective degrees of freedom defined this way can certainly be thought of as a measure of dimensionality of the model.¹² After choosing the effective degrees of freedom to be used in the fit, one then estimates λ so that $Tr(S_\lambda)$ is approximately equal to the selected desired degrees of freedom.

Other formal methods for obtaining λ generally try to minimize the mean square error of the estimate using a form of cross-validation. We discuss this next.

The smoothing parameter λ controls the smoothness of the estimate \hat{f} of f . We do not want λ to be too large since that leads to oversmoothing and a possibly bad fit that misses turns and trends in the true relationship. Nor do we want it to be too small, resulting in undersmoothing and too rough a curve that will capture noise as well as the postulated relationship. We need λ to be appropriately chosen.

Similarly to how we found the estimates for parameters, we ideally want to choose a λ that minimizes a mean squared error (MSE) loss function obtained when this λ is used to construct an estimate $\hat{f}(x_i)$ to predict $f(x_i)$:

$$MSE_\lambda = \frac{1}{n} \sum_{i=1}^n (\hat{f}(x_i) - f(x_i))^2.$$

¹¹ The hat matrix \mathbf{H} in the linear model corresponds to the hat matrix S_λ in (15.12) without the penalty part $\lambda\mathcal{A}$.

¹² Hastie and Tibshirani (1999) define two other estimates for the effective degrees of freedom based on analogy with the linear model, namely, $n - Tr(2S_\lambda - S_\lambda S'_\lambda)$ and $Tr(S_\lambda S'_\lambda)$. For linear models and for smoothers with S_λ symmetric, these matrices are the same, and this includes linear and polynomial regression smoothers and regression splines. They are also the same for running line smoothers (cf., Hastie and Tibshirani 1999, p. 54).

The problem here is that f is unknown, so we cannot use this directly. Instead of obtaining the estimation from MSE, we consider minimizing the predictive squared error (PSE):

$$PSE_\lambda = \frac{1}{n} \sum_{i=1}^n (\hat{f}(x_i) - y_i)^2.$$

Of course this could be minimized by taking $\hat{f}(x_i) = y_i$, which has no smoothness ($\lambda = 0$ in (15.8)) and would not fit the data (or match f) well out of sample. We are interested in more global predictive estimation at points that are not used in forming the estimation.

One way to construct such a predictive estimator is to not use the observed datum point (x_i, y_i) in our estimator of the predicted value $\hat{f}(x_i)$ of f at x_i and instead estimate $\hat{f}(x_i)$ using the other $(n - 1)$ data points excluding x_i . In this way we get a predictive estimate $\hat{f}_{-i}(x_i)$ of f at each of the n data points (a jackknifed fit), which can then be validated by comparison against the actual points (x_i, y_i) . This is called the leaving-one-out method and is similar to having a testing and training set: we fit the data using the training set and test the fit using the testing sample. Here the testing set consists of a single datum (x_i, y_i) with the datum points being taken sequentially. Thus, instead of minimizing MSE_λ , we minimize the ordinary cross-validation score obtained by sequentially leaving one observation out and fitting the model to the other points:

$$CV_\lambda = \frac{1}{n} \sum_{i=1}^n (\hat{f}_{-i}(x_i) - y_i)^2.$$

Since we assumed that $y_i = f(x_i) + \varepsilon_i$ with $E(\varepsilon_i) = 0$ and $E(\varepsilon_i^2) = \sigma^2$, we can replace $f(x_i)$ by $y_i - \varepsilon_i$. Hence CV_λ can be rewritten as

$$\begin{aligned} CV_\lambda &= \frac{1}{n} \sum_{i=1}^n (\hat{f}_{-i}(x_i) - f(x_i) - \varepsilon_i)^2 \\ &= \frac{1}{n} \sum_{i=1}^n (\hat{f}_{-i}(x_i) - f(x_i))^2 - 2(\hat{f}_{-i}(x_i) - f(x_i))\varepsilon_i + \varepsilon_i^2. \end{aligned}$$

The expected value of CV_λ is

$$E(CV_\lambda) = E\left(\frac{1}{n} \sum_{i=1}^n [\hat{f}_{-i}(x_i) - f(x_i)]^2\right) + \sigma^2.$$

Intuitively, because for a large sample n the influence of any single point is not large, we should have $\hat{f}(x_i) \approx \hat{f}_{-i}(x_i)$. Accordingly, it should be the case that

$E(CV_\lambda) \approx E(MSE_\lambda) + \sigma^2$. Since σ^2 does not depend on λ , the λ that minimizes $E(CV_\lambda)$ provides a reasonable estimator of the λ that minimizes $E(MSE_\lambda)$. Similarly, if we find a λ to minimize CV_λ this λ will be a good estimator of the λ that minimizes MSE_λ .

This approach is called *ordinary cross-validation*.¹³ Utilizing the leaving-one-out formulation, cross-validation gives us the ability to predict response data at points not included in the equation-forming estimation. It also provides an estimator λ^* that can be used in creating our final cubic spline estimate of f . This straightforward implementation of the leaving-one-out method, however, is computationally inefficient for larger values of n because it requires fitting the smoothing estimator n times in the leaving-one-out to form the calculation.

Fortunately, it can be shown (after some algebra, cf. Wood 2006, §3.6.1, or Liu 2008, p. 12, for details) that for the linear representation $\hat{f} = S_\lambda y$ using the hat matrix S_λ

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n \frac{(\hat{f}(x_i) - y_i)^2}{(1 - S_{ii}(\lambda))^2} \quad (15.13)$$

where $\hat{f}(x_i) = \hat{f}_\lambda(x_i)$ is the estimate obtained from fitting all the data, and $S_{ii}(\lambda)$ is the diagonal element of the hat matrix S_λ . In practice, the individual terms $1 - S_{ii}(\lambda)$ are replaced by the mean $Tr[I - S_\lambda]/n$, providing what is known as the *generalized cross-validation measure* (GCV) (Craven and Wahba 1979; Golub, Heath, and Wahba 1979):

$$GCV(\lambda) = \frac{\frac{1}{n} \sum_{i=1}^n \{\hat{f}(x_i) - y_i\}^2}{\left[\frac{1}{n} Tr[I - S_\lambda]\right]^2}. \quad (15.14)$$

GCV is computationally more tractable than ordinary CV, and it also has the advantage of invariance (Wahba 1990). The estimator of λ that minimizes the GCV is denoted by λ_{GCV} , and in terms of the estimator's loss performance, the choice of λ_{GCV} is asymptotically the best possible value of the smoothing parameter (Eubank 1999), p. 240; Li 1986). Computer packages allow the computation of λ_{GCV} .

15.2.6 Additive Models: Multivariate Case and the Backfitting Algorithm

The previous discussion considered the situation of fitting the additive model in the case of a single predictor variable x and a response variable y (essentially univariate nonparametric regression). The additive model in (15.2), however, has k possible predictors x_1, \dots, x_k for the response variable y . Here we address the steps necessary to incorporate k predictor variables (the backfitting algorithm).

¹³ Stone (1977) shows that cross-validation is asymptotically equivalent to the Akaike information criterion (AIC).

To motivate the backfitting algorithm, consider first the familiar linear model:

$$E(y|\mathbf{X} = \mathbf{x}) = \alpha + \sum_{j=1}^k \beta_j x_j.$$

The expected value conditional on $X_j = x_j$ is

$$\begin{aligned} E(y|X_j = x_j) &= E(E(y|X_1, \dots, X_k)|X_j = x_j) \\ &= E\left(\alpha + \sum_{m=1}^k \beta_m X_m | X_j = x_j\right) \\ &= \beta_j x_j + E\left(\alpha + \sum_{m \neq j} \beta_m X_m | X_j = x_j\right). \end{aligned}$$

Therefore, we find

$$\begin{aligned} \beta_j x_j &= E(y|X_j = x_j) - E\left(\alpha + \sum_{m \neq j} \beta_m X_m | X_j = x_j\right) \\ &= E\left(\left\{y - \alpha - \sum_{m \neq j} \beta_m X_m\right\} | X_j = x_j\right). \end{aligned}$$

The term inside the expectation is the residual error in predicting y using all the information about y that is contained in all the other predictor variables x_m except for x_j . Therefore the idea is this. First run the partial regression predicting y using all the other predictors except x_j , and then calculate the residual of this regression – call it e . The resulting equation is $E(e) = \beta_j x_j$ so this suggests running a simple regression with x_j and the partial residual e to obtain an estimate of β_j . The problem, however, is that initially we do not know any of the other β_m 's to get e . Therefore, instead we take an initial guess of the β_m 's, $m \neq j$, and use these to get an estimate β_j . We do this process for each of the β_m 's in turn, which at the end provides an updated estimate of each of the β_m 's. The same process can then be repeated again, but now using these updated estimates of the β_m coefficients to create a second round of partial residual regressions, which are in turn regressed with the omitted variable to get another updated estimate of the β_m 's; this process is continued until convergence takes place (i.e., the β_m 's do not change from iteration to iteration). In this way we end up having estimated all the β_m 's in the original linear regression.

An example involving only two explanatory variables should clarify this process. Suppose that we have a model

$$y_i = \alpha + x_{i1}\beta_1 + x_{i2}\beta_2 + e_i.$$

We center the variables, so we construct our regression as

$$y_i = \bar{y} + \beta_1^{(k)}(x_{i1} - \bar{x}_1) + \beta_2^{(k)}(x_{i2} - \bar{x}_2).$$

We observe that $\hat{\alpha} = \bar{y}$. Let $y_i^* = y_i - \bar{y}$, $x_{i1}^* = x_{i1} - \bar{x}_1$, and $x_{i2}^* = x_{i2} - \bar{x}_{12}$. Compute the partial residuals for x_1 and x_2 at iteration k as

$$\begin{aligned} e_{i[1]}^{(k)} &= y_i^* - \beta_2^{(k-1)}x_{i2}^* = e_i + \beta_1^{(k-1)}x_{i1}^*, \\ e_{i[2]}^{(k)} &= y_i^* - \beta_1^{(k-1)}x_{i1}^* = e_i + \beta_2^{(k-1)}x_{i2}^*. \end{aligned}$$

We run a regression involving $e_{i[1]}^{(k)}$ and x_{i1}^* to obtain $\beta_1^{(k)}$. Similarly, we can obtain $\beta_2^{(k)}$ with a second partial residual regression. Thus, we first start with an initial estimate $\beta_1^{(0)}$ and $\beta_2^{(0)}$ (e.g., we could just assume them to be zero), and then we run the procedure to get updated estimates $\beta_1^{(1)}$ and $\beta_2^{(1)}$. The same process can be done again, this time starting from estimates $\beta_1^{(1)}$ and $\beta_2^{(1)}$. This iteration continues until the $\beta_1^{(k)}$ and $\beta_2^{(k)}$ estimates become stable (do not change from iteration to iteration).

To apply this same intuitive process to the fitting of the additive model, we observe that the important assumption that was used in this process was the additivity, not the fact that the predictor function involving x_j was the monomial $\beta_j x_j$. We can use functions $f_j(x_j)$ in place of the terms $\beta_j x_j$ and can perform a similar iterative fit. We show this now.

Suppose that we have n observations and a simple additive model (15.2). Each of the individual functions f_j in (15.2) can be initially estimated using a smoother (kernel, cubic spline, etc.) as described earlier. Suppose $S_j(\cdot)$ is the smoother chosen for the j^{th} predictor variable in (15.2). It will provide a function f_j for the general model (15.2). Since the smoothers as discussed previously have a constant term in their formulation and so does the model (15.2), to have identifiability (unique estimates), we center variables by adding the constraint $E[\sum_{j=1}^k \hat{f}_j(x_j)] = 0$ or equivalently $E(y) = \hat{\alpha}$ when we perform our estimations. We improve the resultant fit when the k smoothers are put together, by iteratively considering the functions to be fit simultaneously in a manner similar to what we did with linear regression.

Assuming that the additive model (15.2) is correct and $y = \alpha + \sum_{j=1}^k f_j(x_j) + e$, it follows that for any j ,

$$f_j(x_j) = E[y - \alpha - \sum_{m \neq j} f_m(x_m) | x_j]. \quad (15.15)$$

The term inside the expectation in (15.15) is the residual left in y (holding x_j fixed) after accounting for the other predictors $f_m(x_m)$, $m \neq j$. Thus, as before, the intuition is that if for each $m \neq j$ we can find good estimates $\hat{f}_m \approx f_m$, then we can use these to get a good estimate \hat{f}_j of f_j .

This suggests the following iterative algorithm for fitting the model (15.2) similar to that described previously for linear regression. First, make initial estimates of the f_m , $m = 1, \dots, k$, $m \neq j$ (say, linear regression estimates of x_m on y , or even set them equal to zero) and set the initial $\hat{\alpha} = \bar{y}$. Use these to estimate \hat{f}_j . This new estimate of \hat{f}_j can then be used in a similar manner to (15.15) to get new estimates of the other f_m . In each iterative step, we smooth out the response variable y after removing the effect from other predictors:

$$\hat{f}_j = S_j \left[y - \hat{\alpha} - \sum_{l \neq j} \hat{f}_l(x_l) | x_j \right].$$

We also center each smoother \hat{f}_j by removing the average

$$\hat{f}_j \leftarrow \hat{f}_j - n^{-1} \sum_{i=1}^n \hat{f}_j(x_{ij}),$$

and add the average to $\hat{\alpha}$. The algorithm iterates until all \hat{f}_j 's can no longer be improved; that is, the algorithm continues until no change in the \hat{f}_j 's occur.¹⁴ Formally, the steps are as follows.

- (1) Initialize the estimates: $\hat{\alpha}^{(0)} = \bar{y}$, $\hat{f}_j = \hat{f}_j^{(0)}$, $j = 1, \dots, k$. The $\hat{f}_j^{(0)}$ initial estimates could be linear estimates or some estimate derived from another process (domain knowledge, for example). The better the initial estimate, the faster the convergence.
- (2) Cycle over $j = 1, \dots, k$: At step p , use the smoothers S_j to update

$$\hat{f}_j^{*(p)} = S_j(y - \hat{\alpha}^{(p-1)} - \sum_{m \neq j} \hat{f}_m^{(p-1)} | x_j)$$

Centering:

$$\begin{aligned} \hat{f}_j^{(p)} &= \hat{f}_j^{*(p)} - \frac{1}{n} \sum_{i=1}^n \hat{f}_j^{*(p)}(x_{ij}) \\ \hat{\alpha}^{(p)} &= \hat{\alpha}^{(p-1)} + \frac{1}{n} \sum_{i=1}^n \hat{f}_j^{*(p)}(x_{ij}) \end{aligned}$$

- (3) Repeat step 2 until convergence: $\|\hat{f}_j^{(p)} - \hat{f}_j^{(p-1)}\| < \delta$ for a specified accuracy level δ .

This process is called the *backfitting algorithm*, and by performing the backfitting algorithm we have accounted for the influence of the other predictors \hat{f}_m on \hat{f}_j ,

¹⁴ Schimek (2000) gives a discussion of convergence and methods to speed convergence.

yielding a better global estimate of \mathbf{y} . The backfitting algorithm converges in many practical situations (although not always in general). When smoothers are linear operators (i.e., the function for estimating $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_n)'$ can be written as $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$, where \mathbf{S} does not depend on \mathbf{y}), Buja, Hastie, and Tibshirani (1989) show the backfitting algorithm is the Gauss-Seidel method for solving the additive model's normal equations. This proves convergence for a wide class of smoothers that includes the cubic spline and kernel smoothers discussed previously. What is so appealing about this algorithm is that it simultaneously estimates all the smooth terms using algorithms that individually estimate each smoother term and maintain the additive property of the model.

15.2.7 Penalized Least-Square Framework for the Multivariate Additive Model

A multidimensional additive model is defined by (15.2) where $k > 1$. To fit this model, we extend the one-dimensional penalized least-square framework of (15.8) to incorporate each of the k functions (cf., Hastie and Tibshirani 1999, p. 110). Now all of the functions f_m work together at the point x_i to produce the response y_i . Thus we now seek to find functions f_m to minimize

$$\sum_{i=1}^n \left\{ y_i - \sum_{m=1}^k f_m(x_{im}) \right\}^2 + \sum_{m=1}^k \lambda_m \int \left\{ f_m''(t) \right\}^2 dt, \quad (15.16)$$

where the f_m 's are twice continuous differentiable. Each λ_m is used for penalizing each smooth function f_m . Again, the solution to (15.16) occurs when the smoother f_m for the predictor x_m is a cubic spline (Schimek 2000, p. 283). Equation (15.16) can be rewritten as

$$\left(\mathbf{y} - \sum_{m=1}^k f_m \right)' \left(\mathbf{y} - \sum_{m=1}^k f_m \right) + \sum_{m=1}^k \lambda_m f_m' \mathbf{A}_m f_m, \quad (15.17)$$

where the \mathbf{A}_j 's are penalty matrices of each predictor, as previously detailed in the one-dimensional setting. Differentiating (15.17) with respect f_j to produces $-2(\mathbf{y} - \sum_m f_m) + 2\lambda_j \mathbf{A} f_j = 0$, so that (cf., Hastie and Tibshirani 1999, p. 111),

$$\hat{f}_j = (I + \lambda_j \mathbf{A}_j)^{-1} \left(\mathbf{y} - \sum_{m \neq j} \hat{f}_m \right), \quad j = 1, \dots, k. \quad (15.18)$$

Utilizing the individual smoothers (hat matrices), $\mathbf{S}_j = (\mathbf{I} + \lambda_j \mathbf{A}_j)^{-1}$, $j = 1, \dots, k$, and rearranging terms, we may rewrite the entire set of equations (15.16) in a giant $(nk \times nk)$ matrix form as

$$\begin{bmatrix} \mathbf{I} & \mathbf{S}_1 & \mathbf{S}_1 & \cdots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \mathbf{S}_2 & \cdots & \mathbf{S}_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_k & \mathbf{S}_k & \mathbf{S}_k & \cdots & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_k \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1 \mathbf{y} \\ \mathbf{S}_2 \mathbf{y} \\ \vdots \\ \mathbf{S}_k \mathbf{y} \end{bmatrix}. \quad (15.19)$$

In short form, we let $\mathbf{f} = (f'_1, \dots, f'_k)'$ be the nk dimensional vector obtained by concatenating the n dimensional functions f_1, \dots, f_k , and we may write the normal equation of the penalized least-square additive model in matrix form as

$$\mathbf{P}\mathbf{f} = \mathbf{Q}\mathbf{y}. \quad (15.20)$$

Solving (15.19) or (15.20) yields the desired smoothers for (15.2). According to Hastie and Tibshirani (1999, p. 111), the second summation (the penalty term in (15.17)) can be interpreted as a down-weighting of each component of \mathbf{f}_j . The down-weighting is determined by the corresponding eigenvalue of that component and λ_j . In practice, the `mgcv` program written by Simon Wood and available in the computer package R will supply the resulting estimated functions f_m .

15.3 The Generalized Additive Model

We have shown that an additive model is a nonparametric extension from linear regression to nonparametric regression obtained by generalizing each parametric monomial predictor in the linear regression to a general function while preserving the additive property of the contribution of the individual variables. Thus, the mean of the response variable is modeled as an additive sum of the individual predictor functions. This generalization can be further extended to cope with differences in the structure of the response variable y by adopting a generalized linear model (GLM) approach for the response variable, but in an additive model setting for the predictor variables. GLM assumes linear effects on the predictors, but generalizes the distribution of the response variable, as well as the link between the predictors and this distribution. The generalized additive model (GAM) retains all the nice features of the GLM and extends the additive sum of the predictors to be the additive sum of more general than linear predictor functions. Thus, the GAM can be written as

$$g(\mathbb{E}[y]) = \alpha + \sum_{j=1}^k f_j(x_j), \quad (15.21)$$

where $g(\cdot)$ is a nonlinear differentiable and strictly monotonic link function. Since the link function $g(\cdot)$ is invertible, the equation (15.21) can also be rewritten as

$$E[y] = g^{-1} \left[\alpha + \sum_{j=1}^k f_j(x_j) \right]. \quad (15.22)$$

Equation (15.22) looks very similar to (5.6) in the GLM framework, except the systematic function inside the brackets in GLM is a linear model and in GAM it is an additive model. The introduction of the additive nonparametric regression inside the brackets in (15.22) complicates the estimation, but utilizing an estimation approach similar to that used in GLM (Fisher scoring) and combining it with the estimation method previously described for the additive model (backfitting), we are able to obtain estimates in the GAM. The method is called local scoring and can be performed using `mgcv` in R. Details and proofs can be found in Wood (2006) or Wasserman (2006).

15.3.1 Viewing GAM as a GLM and Penalized Likelihood Fit

Using the approach of Section 15.2.4 for cubic smoothing spline estimators, there is a way to view a GAM as a type of GLM with a sum of the smooth functions of covariates when using cubic splines for our smooth estimators (cf., Wood 2006, p. 163). The model structure is

$$g(\mu_i) = \alpha + f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_k(x_{ki}), \quad (15.23)$$

where $\mu_i \equiv E(y_i)$ and the distribution of y_i is some exponential family distribution as described in Chapter 5. The link function g is a monotonic and twice differentiable and known. The f_j 's are smooth functions of the predictor variables x_j 's. For linear smoothers, as noted in Section 15.2.4, we can specify a basis set for the estimation of each of the functions f_j ; hence, $g(\mu_i)$ in (15.23) can be expressed as a linear combination of the concatenation of these basis functions. Let b_{ji} be a set of basis functions for each function f_j . As in (15.9) the smoother function f_j can be represented in this larger basis set as

$$f_j(x_j) = \sum_{i=0}^{q_j} \beta_{ji} b_{ji}(x_j),$$

where x_j are the predictors and the β_{ji} are coefficients of the smoother f_j that are to be estimated. Once the basis is selected (and λ_j is determined), as in Section 15.2.4, for each j we have a matrix M_j so that $\hat{f}_j = M_j \hat{\beta}_j$. For identifiability we center the f_j by making the sum (or mean) of the elements of f_j be zero (and we incorporate the constant term into our estimate of α in the general formula). Thus we can create

a large model matrix \mathbf{X} and a set of parameters $\boldsymbol{\beta}$ where $\mathbf{X} = [\mathbf{1}, \mathbf{M}_1, \mathbf{M}_2, \dots, \mathbf{M}_k]$, $\mathbf{1}$ is a vector of 1's, and $\boldsymbol{\beta} = (\alpha, \boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_k)'$. We can now write the model as

$$g(\boldsymbol{\mu}_i) = \mathbf{X}\boldsymbol{\beta},$$

and we recognize this is a GLM.

As in Chapter 5, once the link function has been determined and an exponential family distribution model specified, the log-likelihood function $l(\boldsymbol{\beta})$ can be written down accordingly as in Chapter 5. Denoting the smoothing parameter of each smoother function f_j as λ_j , we define a penalized log-likelihood $l_p(\boldsymbol{\beta})$ for a model as

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2} \sum_j \lambda_j \boldsymbol{\beta}' \mathbf{A}_j \boldsymbol{\beta}, \quad (15.24)$$

where \mathbf{A}_j is the penalty matrix for predictor j discussed in Section 15.2.4. The penalty is needed so that the fit is smooth (otherwise the fit could be a very rough interpolative fit of a saturated model that maximizes the unpenalized likelihood). The vector of λ_j 's controls the trade-off between the roughness of smoother functions and the goodness of fit of the model. The optimal $\hat{\boldsymbol{\beta}}$ vector is solved for by maximizing $l_p(\boldsymbol{\beta})$ given a set of λ_j 's. In practice, all this is done in the computer program in the packages that fit GAMs.

Example 15.2 (GAM Fit to Coronary Heart Disease with Logit Link Function). Returning to the example of coronary heart disease prediction (Example 15.1), we can apply the GAM methodology to determine the influence of the continuous predictor variables on the likelihood of an individual having coronary heart disease. In this analysis we are interested in estimating the probability of an individual having heart disease, rather than the zero-one response variable gathered in the data. Therefore, as in Chapters 3 and 5, we introduce a link function (the logit link function g in a GAM), which results in an ability to estimate this probability using the logistic distribution. Refer to the discussion in Chapters 3 and 5 for details of this logit link function.

Figure 15.3 displays the results of fitting a cubic spline function to the data and plots the results against the log odd-ratio that would result from using a logit link function g in a GAM. The response variable of interest in this case is the probability of an individual having coronary heart disease given the set of predictor variables. This probability is best displayed as a log odds-ratio, which is the left-hand side of a GAM with the function g as the logit function $g(y) = \ln\left(\frac{y}{1-y}\right)$. As can be seen, the cubic spline fits exhibit the nonlinear nature of the fit to the log odds, and hence the influence of each predictor variable on log odds is not a simple function.

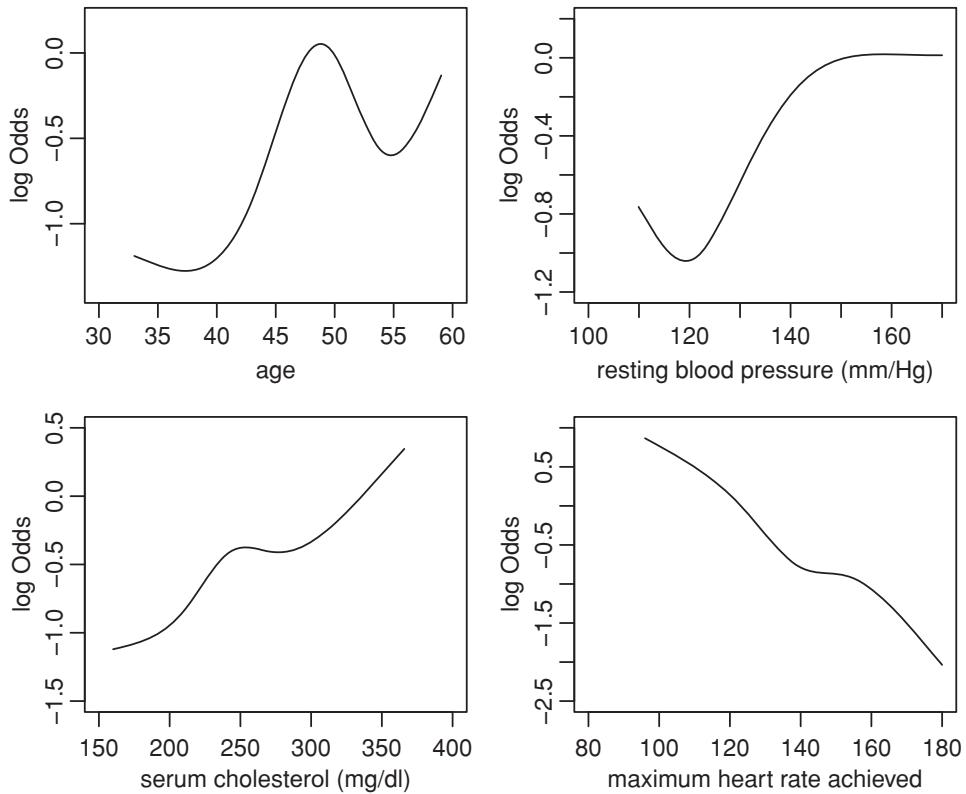


Fig. 15.3. Fit of the continuous predictors variables to the log odds-ratio of coronary heart disease using a GAM with the logit link function.

15.3.2 Choosing between a GAM and a GLM

A GAM provides a more general and flexible framework for model fitting than does a GLM, but it is more complicated and not as intuitively interpretable. If there are similar fit characteristics, a GLM, in most circumstances, is the preferred model form because of its simplicity and because it is easier to interpret its implications. If, however, the fit of the GAM is sufficiently better in improving the predictability of the response variable based on the nonlinear characteristics of the predictor variables, then the GAM is preferred.

Several hypothesis tests have been developed to determine whether the smoothing functions in a GAM can significantly improve the predictive performance. The null hypothesis is that only a parametric model form like the GLM is needed. The alternative is that predictor variable smoothing can be applied to significantly improve the model fit.

Recall that for smoothing splines the penalty for lack of smoothness is λA , where A is the penalty matrix discussed in (15.10) whose first two rows and columns are

zero corresponding to the constant term and linear term in the basis representation. A larger value of λ means a smoother function should be applied, and if λ is infinity, then all the weight must go to the first two basis elements, implying the model is a linear model and the fit is a GLM. Therefore, in testing between a GAM and a GLM the hypothesis to be tested can be phrased as

$$\begin{aligned} H_0 &: \lambda = \infty, \\ H_a &: \lambda < \infty. \end{aligned}$$

Wang (2011) presents a statistical test of this hypothesis using a ratio test statistic based on using the smoothing parameter λ_{GCV} that minimizes the GCV measure previously discussed in 15.2.5:

$$t_{GCV} = \frac{GCV(\lambda_{GCV})}{GCV(\lambda = \infty)}.$$

H_0 is rejected if t_{GCV} is too small. A Monte Carlo method is used to compute an estimate of the p -value.¹⁵ There are other likelihood-based tests but they require distributional assumptions.

15.3.3 GAM Inference: Effective Degrees of Freedom for the Predictor Variables

To fit a parsimonious GAM we would like to determine which predictor variables add significantly to the prediction and how much dimensionality (degrees of freedom) they add to the model. To this end we wish to determine the effective degrees of freedom of each predictor in a GAM. We now show how to do this.

In Section 15.2.5 we provided the one-dimensional GAM analogy to the degrees of freedom in a linear model (called the effective degrees of freedom) as $Tr(S_\lambda)$ where $S_\lambda = M(M'M + \lambda A)^{-1}M'$ is the so-called hat matrix, $\hat{f} = M\hat{\beta} = S_\lambda y$, and $\hat{\beta} = (M'M + \lambda A)^{-1}M'y$ (this trace formulation coincides with the usual degrees of freedom when we are in the linear model). Now, in the multidimensional GAM setting, let $A = \sum \lambda_j A_j$ and define $P = (M'M + A)^{-1}M'$, so $\hat{\beta} = Py$ and $\hat{f} = MPy$. Let P_i^0 denote the matrix that has same i^{th} row as P , but with all the other rows equal to zero, so $P = \sum P_i^0$ and $P_i^0 y = \beta_i$. Accordingly, the total effective degrees of freedom can be decomposed as $Tr(S_\lambda) = Tr(MP) = \sum Tr(MP_i^0)$. Summing the $Tr(MP_i^0)$ values over the basis elements that constitute the basis representation for the individual predictor variable x_j (the basis discussed in Section 15.2.4 in the univariate case and formulated in a very large vector framework in Section 15.3.1) yields the effective degrees of freedom of x_j . Intuitively, in the usual linear model each β_i would add a degree of freedom, but because of the additional smoothing enforced by the

¹⁵ See Wang (2011) for details of the simulation.

smoothness penalty λ_j in the hat matrix, the trace can shrink and the effective degrees of freedom can be less. This was the shrinkage of effective degrees of freedom alluded to at the end of Section 15.2.4.

15.3.4 GAM Inference: Choosing between GAM Formulations

Sometimes we would like to test between two nested GAMs to see if adding another predictor variable adds significantly to the overall fit. In GLM, the log-likelihood ratio test is the test statistic that plays the central role. The deviance discussed in Section 5.7.2 in GLM evaluation for a fitted model represented by $\hat{\eta}$ is defined by

$$D(y; \hat{\eta}) = 2(l(\eta_{max}; y) - l(\hat{\eta}; y)),$$

where η_{max} is the parameter value that maximizes the log-likelihood $l(\eta; y)$ of the saturated model, or the model with one parameter per data point. This saturated model has the highest possible value $l(\eta_{max}; y)$ that the likelihood could possibly have given the data, because with one parameter per data point, it is possible to reproduce the data. Deviance can be interpreted as being similar to the residual sum of squares in ordinary linear modeling and is used for evaluating goodness of fit and for comparing models.

The asymptotic distribution theory for GLM has been well studied. For η_1 and η_2 defining two linear models, with η_1 nested within η_2 , the ordinary likelihood ratio test, computed as twice log-likelihood ratio statistic, $2 \ln \left[\frac{L(\eta_2; y)}{L(\eta_1; y)} \right] = 2[l(\eta_2; y) - l(\eta_1; y)]$, asymptotically has a chi-squared distribution with the degrees of freedom equal to the difference in degrees of freedom of the models. Given the definition of deviance, this can be expressed as $D(\eta_1; \eta_2) = D(y; \eta_1) - D(y; \eta_2)$. If η_1 is correct, under certain regularity conditions, $D(\eta_1; \eta_2)$ has an asymptotic χ^2 distribution with the degrees of freedom equal to the difference in the dimensions of these two models:

$$D(\eta_1; \eta_2) \sim \chi^2_{df_2 - df_1}.$$

For a GAM, it still makes sense to consider deviance as the means to measure the difference when comparing the two models, but now we need to define what we mean by a saturated model, and we need to use the penalized log-likelihood function l_p instead of the log-likelihood function l . However, the deviance as defined earlier with the penalized log-likelihood is not generally provable to be asymptotically χ^2 -distributed, and its distribution theory is not yet well developed. In practice, however, there is some empirical justification (Hastie and Tibshirani 1999) for informally using the deviance test by referencing the χ^2 distribution with degrees of freedom equal to the difference of the two models' effective degrees of freedom.

Table 15.3. *GAM Model 1 with Four Continuous Predictor Variables, $g = \text{Logit}$ Link Function, Predictor Variables Smoothed Individually*

	Estimate	Std. Error	z Value	p -Value
Intercept	−0.5975 edf	0.1421 Chi sq.	−4.206 p -value	0
age	1	0.001	0.97285	
threstbps	3.396	4.827	0.34039	
chol	1	7.671	0.00561	
thalach	1	20.568	0	
AIC	312.18			

An intuitive justification for using deviance in model comparisons involving GAM fitting is also well described in Wood (2006). When we construct test statistics for a GAM that are similar to those we have done in a GLM, we need to use penalized likelihood, and this involves the smoothing parameter λ . Although λ is unknown and numerical approximation is needed, we can perform the test and treat λ as a known parameter. Wood (2006) argues that the ratio comparison results of two smoothing splines in the GAM formulation can be approximated by comparing two GAMs without considering the penalty parameters, and he argues (and shows empirically by examples) that the comparison results are very similar. The test statistic that is pertinent when we ignore the smoothing parameter λ is that corresponding to a GLM, and we know in this case that the log-likelihood ratio can be compared to a χ^2 distribution with degrees of freedom equal to the difference of the two models' degrees of freedom. Since this is true for the very good approximating models without λ , the distribution of the ratio when using the penalty parameters should at least be approximately correct. Therefore, if testing the nested hypothesis, the approximate p -value is given by the p -value for the model without the penalty. Since the purpose of a GAM is to find a model with the best predicting power, we may also calculate the Akaike information criterion (AIC) from each model and compare them.

Example 15.3 (Fitting a GAM to Coronary Heart Disease Data). Returning again to the coronary heart disease data, we fit various GAM models involving the different configurations of predictor variables. The results are presented in Tables 15.3–15.5. The computer implementations provide the AIC values of the various GAM model fits, as well as the effective degrees of freedom and the approximate p -values corresponding to the approximating χ^2 distribution for determining whether the addition of a predictor variable adds significantly to the fit. The actual computations are beyond the scope of this chapter, but the interested reader is referred to Wood (2006) and Clark (2013), which provide detailed instructions for computing and interpreting the

**Table 15.4. GAM Model 2 with Two Significant Continuous Predictor Variables,
 $g = \text{Logit Link Function}$, Two Predictor Variables Smoothed Together**

	Estimate	Std. Error	z Value	p-Value
Intercept	-0.5903	0.1409	-4.189	0
	edf	Chi.sq	<i>p</i> -value	
chol, thalach	2	33.59	0	
AIC	310.18			

results of GAM analysis in R. We used the R program `mgcv` designed for applying GAMs that is available for download at <http://cran.r-project.org/bin/windows/base/>.

The first model (GAM 1) uses the four continuous predictor variables discussed earlier, and the logit link function (so the log odds-ratio of the likelihood of exhibiting coronary heart disease is the dependent variable in the GAM model). The results obtained using the R program package `mgcv` are presented in Table 15.3. As can be seen, the variables `age` and `threstbps` are not significant.

Since the two variables `age` and `threstbps` were not significant, in the interests of parsimony, another model was created with just the two significant continuous predictor variables included. In Table 15.4 the two variables were smoothed. As can be seen, by eliminating the nonsignificant variables we obtain a better model (lower AIC) than in Table 15.3.

Since we also have other categorical variables available in the dataset, we can include them in the model. The categorical variables do not need smoothing, so they are just included in the GAM formulation. Table 15.5 show the GAM formulation that has the two significant continuous predictor variables `chol` and `thalach` each smoothed individually, as well as the categorical predictor variables.

Thus, the final GAM formulation involving the two significant continuous predictor variables as well as the additional categorical variables produces the best fit (lowest AIC).

15.4 Conclusion

Generalized additive models (GAMs) combine the flexibility of modeling the relationship between the response variable and the predictor variables exhibited in GLM with the flexibility of allowing nonlinear relationships between individual predictors and the response variable exhibited in nonparametric regression models. The additive model, a compromise between the full flexibility of multidimensional nonparametric regression and fully parametric linear models, provides the basis for the GAM

**Table 15.5. GAM Model 3 with Two Significant Continuous Predictor Variables
Smoothed Individually and Categorical Predictor Variables Included,
Link Function $g = \text{Logit Link}$**

	Estimate	Std. Error	<i>z</i> Value	<i>p</i> -Value
Intercept	−4.9986	0.7391	−6.763	0
Sex	1.0624	0.407	2.61	0.00904
Cp	1.1221	0.1969	5.699	0
Fbs	1.5154	0.6398	2.369	0.01785
Res	−0.2715	0.3649	−0.744	0.45679
	edf	Chi.sq	<i>p</i> -value	
chol	1	4.23	0.0397	
thalach	1	6.802	0.0091	
AIC	257.69			

that provides enough structure to allow inference as well as improved model fit for prediction. We have shown the intuition behind the models in this chapter, but the computation of the GAM is substantial and requires care. Fortunately, such computational considerations have been incorporated into statistical packages such as SAS and R, and the computing power now available has made such computations and model fits available to researchers without trouble. Books such as Clark (2013) and Wood (2006) can provide instruction for using the freeware in R and can provide access to improved modeling of complex relationships between response variables and predictors in actuarial science, insurance, and other domains.

References

- Buja, A., T. Hastie, and R. Tibshirani (1989). Linear smoothers and additive models (with discussion). *Annals of Statistics* 17(2), 453–510.
- Clark, M. (2013). *Generalized Additive Models: Getting Started with Additive Models in R*. Center for Social Research, University of Notre Dame, Available at <http://www3.nd.edu/~mclark19/learn/GAMS.pdf>.
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 74(368), 829–836.
- Cleveland, W. S. (1981). Lowess: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician* 35(1), 54.
- Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* 31, 377–403.
- Epanechnikov, V. A. (1969). Nonparametric estimation of a multidimensional probability density. *Theory of Probability and its Application* 14(1), 156–161.
- Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing* (2nd ed.). Marcel Dekker, New York.

- Golub, G. H., M. Heath, and G. Wahba (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21(2), 215–223.
- Hastie, T. J. and R. J. Tibshirani (1999). *Generalized Additive Models*. Chapman & Hall, Boca Raton, FL.
- Johnson, M. L., V. L. Bengtson, P. G. Coleman, and T. B. Kirkwood (2005). *The Cambridge Handbook of Age and Ageing*. Cambridge University Press, Cambridge.
- Li, K.-C. (1986). Asymptotic optimality of CL and generalized cross-validation in ridge regression with application to spline smoothing. *Annals of Statistics* 14(3), 1101–1112.
- Liu, H. (2008). *Generalized Additive Model*. Department of Mathematics and Statistics, University of Minnesota Duluth, Duluth, MN 55812.
- Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische Mathematik* 10(3), 177–183.
- Schimek, M. G. (2000). *Smoothing and Regression: Approaches, Computation, and Application*. John Wiley & Sons, New York.
- Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika* 64(1), 29–35.
- Wahba, G. (1990). *Spline Models for Observational Data*. Number 59. Siam.
- Wang, Y. (2011). *Smoothing Splines: Methods and Applications*. Taylor & Francis Group, New York.
- Wasserman, L. (2006). *All of Nonparametric Statistics*, Volume 4. Springer, New York.
- Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall, Boca Raton, FL.

16

Nonlinear Mixed Models

Katrien Antonio and Yanwei Zhang

Preview of the Chapter. We start with a discussion of model families for multilevel data outside the Gaussian framework. We continue with generalized linear mixed models (GLMMs), which enable generalized linear modeling with multilevel data. The chapter includes highlights of estimation techniques for GLMMs in the frequentist as well as Bayesian context. We continue with a discussion of nonlinear mixed models (NLMMs). The chapter concludes with an extensive case study using a selection of R packages for GLMMs.

16.1 Introduction

Chapter 8 (Section 3) motivates predictive modeling in actuarial science (and in many other statistical disciplines) when data structures go beyond the cross-sectional design. Mixed (or multilevel) models are statistical models suitable for the analysis of data structured in nested (i.e., *hierarchical*) or non-nested (i.e., cross-classified, *next to* each other, instead of hierachically nested) clusters or levels. Whereas the focus in Chapter 8 is on *linear* mixed models, we now extend the idea of mixed modeling to outcomes with a distribution from the exponential family (as in Chapter 5 on generalized linear models (GLMs)) and to mixed models that generalize the concept of linear predictors. The first extension leads to the family of generalized linear mixed models (GLMMs), and the latter creates nonlinear mixed models (NLMMs). The use of mixed models for predictive modeling with multilevel data is motivated extensively in Chapter 8. These motivations also apply here. We focus in this chapter on the formulation and interpretation of mixed models for non-normal outcomes and nonlinear modeling problems. Estimation, inference, and prediction with a range of numerical techniques are discussed. Readers who are not interested in the technicalities of estimation techniques can skip Sections 16.3.3.1 to 16.3.3.3.

16.2 Model Families for Multilevel Non-Gaussian Data

Section 2 of Chapter 8 explains the connection between the marginal and hierarchical interpretation of a linear mixed model (LMM). This feature is a consequence of the nice properties of the multivariate normal distribution, but it no longer exists when outcomes are of the non-Gaussian type. Thus, with outcomes of non-Gaussian type we explicitly distinguish so-called *marginal* (cfr. *infra*) versus *random effects* models for clustered (or multilevel) non-normal data. This implies that the fixed effects β have different interpretations in both approaches. Estimates obtained with one of those both model families may differ substantively. Molenberghs and Verbeke (2005) distinguish three model families for handling non-Gaussian clustered data: *marginal*, *conditional*, and *subject-specific* models. Generalized estimating equations (GEEs) are a well-known computational tool for *marginal models*. With GEEs the marginal mean $\mu = E[y] = g^{-1}(X\beta)$ should be correctly specified, in combination with a working assumption about the dependence structure. $g(\cdot)$ is the link function introduced in Chapter 5. Even though this working assumption may be wrong, the GEE estimator of β has nice properties (consistency, asymptotic normality with mean β , and covariance matrix as in Liang and Zeger (1986)). Applications of GEEs in actuarial predictive modeling are found in Purcaru, Guillén, and Denuit (2004) and Denuit et al. (2007). With this approach, there is interest only in the effect of certain covariates on the marginal response; no cluster-specific inference or prediction is possible. The class of *conditional models* is a second group of models where y is modeled conditional on (a subset of) the other outcomes. We do not discuss these models here. Our focus – from Section 16.3 on – is on *subject or cluster-specific models*, more specifically on generalized and nonlinear mixed models (GLMMs and NLMMs) incorporating random, subject, or cluster-specific effects.

16.3 Generalized Linear Mixed Models

16.3.1 Generalized Linear Models

Generalized linear models (GLMs) have numerous applications in actuarial science, ranging from ratemaking over loss reserving to mortality modeling (see Haberman and Renshaw 1996, for an overview). Chapter 5 explains in detail the use of GLMs with cross-sectional data. A GLM is a regression model specified within the distributional framework of the exponential family. A member of this family has a density of the form

$$f_Y(y) = \exp\left(\frac{y\theta - \psi(\theta)}{\phi} + c(y, \phi)\right). \quad (16.1)$$

$\psi(\cdot)$ and $c(\cdot)$ are known functions; θ is the natural and ϕ the scale parameter. Using vector notation the following relations hold:

$$\boldsymbol{\mu} = E[\mathbf{y}] = \psi'(\boldsymbol{\theta}) \quad \text{and} \quad \text{Var}[\mathbf{y}] = \phi\psi''(\boldsymbol{\theta}) = \phi V(\boldsymbol{\mu}), \quad (16.2)$$

where derivatives are with respect to $\boldsymbol{\theta}$ and $V(\cdot)$ is the so-called variance function. The latter function captures the relationship between the mean and variance of \mathbf{y} . GLMs provide a way around transforming data by specifying a linear predictor for a transformation of the mean regression parameters

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \quad (16.3)$$

with $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ the vector of regression parameters and \mathbf{X} ($m \times p$) the design matrix. g is the link function and $\boldsymbol{\eta}$ the so-called linear predictor. Estimates for $\boldsymbol{\beta}$ follow by solving the maximum likelihood equations with an iterative numerical technique (such as Newton-Raphson). Likelihood ratio and Wald tests are available for inference purposes. If the scale parameter ϕ is unknown, we estimate it by maximum likelihood or by dividing the deviance or Pearson's chi-square statistic by its degrees of freedom.

16.3.2 Extending GLMs with Random Effects

GLMMs extend GLMs by adding random effects $\mathbf{Z}\mathbf{u}$ to the linear predictor $\mathbf{X}\boldsymbol{\beta}$. Motivations for this extension are similar to those in Section 8.1: the random effects enable cluster-specific prediction, allow for heterogeneity between clusters, and structure correlation within clusters. Conditional on a q -dimensional vector \mathbf{u}_i of random effects for cluster i , GLMM assumptions for the j th response on cluster or subject i , y_{ij} , are

$$\begin{aligned} y_{ij} | \mathbf{u}_i &\sim f_{Y_{ij}|\mathbf{u}_i}(y_{ij} | \mathbf{u}_i) \\ f_{Y_{ij}|\mathbf{u}_i}(y_{ij} | \mathbf{u}_i) &= \exp\left(\frac{y_{ij}\theta_{ij} - \psi(\theta_{ij})}{\phi} - c(y_{ij}, \phi)\right) \\ \mathbf{u}_i &\sim f_U(\mathbf{u}_i), \end{aligned} \quad (16.4)$$

with \mathbf{u}_i independent among clusters i . The following conditional relations hold:

$$\mu_{ij} = E[y_{ij} | \mathbf{u}_i] = \psi'(\theta_{ij}) \quad \text{and} \quad \text{Var}[y_{ij} | \mathbf{u}_i] = \phi\psi''(\theta_{ij}) = \phi V(\mu_{ij}). \quad (16.5)$$

A transformation of the mean μ_{ij} is linear in both the fixed ($\boldsymbol{\beta}$) and random effects (\mathbf{u}_i) parameter vectors

$$g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i, \quad (16.6)$$

with \mathbf{u}_i the vector of random effects for cluster i , and \mathbf{x}_{ij} and \mathbf{z}_{ij} the p and q dimensional vectors of known covariates corresponding with the fixed and random

effects, respectively. A distributional assumption for the random effects vector \mathbf{u}_i , say $f_U(\mathbf{u}_i)$, completes the specification of a GLMM. Most applications use normally distributed random effects, but other distributional assumptions for the random effects are possible.

The model assumptions in (16.4), (16.5), and (16.6) imply the following specifications for marginal mean and variance:

$$\begin{aligned} E[y_{ij}] &= E[E[y_{ij}|\mathbf{u}_i]] = E[g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i)] \\ \text{Var}(y_{ij}) &= \text{Var}(E[y_{ij}|\mathbf{u}_i]) + E[\text{Var}(y_{ij}|\mathbf{u}_i)] \\ &= \text{Var}(\mu_{ij}) + E[\phi V(\mu_{ij})] \\ &= \text{Var}(g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i)) + E[\phi V(g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i))]. \end{aligned} \quad (16.7)$$

In general, simplification of these expressions is not possible. The GLMM regression parameters $\boldsymbol{\beta}$ do not have a marginal interpretation; they express the effect of a set of covariates on the response, conditional on the random effects \mathbf{u}_i . Indeed, $E[y_{ij}] = E[E[y_{ij}|\mathbf{u}_i]] = E[g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i)] \neq g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta})$. Example 16.1 shows explicit calculation of the marginal mean, variance, and covariance within a Poisson GLMM.

Example 16.1 (A Poisson GLMM). Conditional on a random intercept $u_i \sim N(0, \sigma^2)$, y_{ij} is Poisson distributed with $\mu_{ij} = E[y_{ij}|u_i] = \exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + u_i)$. Thus, the link function g is the logarithm. The corresponding likelihood is

$$L(\boldsymbol{\beta}, \sigma | \mathbf{y}) = \prod_{i=1}^m \int_{-\infty}^{+\infty} \left(\prod_{j=1}^{n_i} \frac{\mu_{ij} e^{-\mu_{ij}}}{y_{ij}!} \right) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}u_i^2} du_i. \quad (16.8)$$

Straightforward calculations using mean and variance of a lognormal distribution show

$$\begin{aligned} E(y_{ij}) &= E(E(y_{ij}|u_i)) = E(\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + u_i)) \\ &= \exp(\mathbf{x}'_{ij}\boldsymbol{\beta}) \exp(\sigma^2/2) \end{aligned} \quad (16.9)$$

and

$$\begin{aligned} \text{Var}(y_{ij}) &= \text{Var}(E(y_{ij}|u_i)) + E(\text{Var}(y_{ij}|u_i)) \\ &= E(y_{ij})(\exp(\mathbf{x}'_{ij}\boldsymbol{\beta})[\exp(3\sigma^2/2) - \exp(\sigma^2/2)] + 1), \end{aligned} \quad (16.10)$$

and

$$\begin{aligned} \text{Cov}(y_{ij}, y_{ik}) &= \text{Cov}(E(y_{ij}|u_i), E(y_{ik}|u_i)) + E(\text{Cov}(y_{ij}, y_{ik}|u_i)) \quad (j \neq k) \\ &= \exp(\mathbf{x}'_{ij}\boldsymbol{\beta}) \exp(\mathbf{x}'_{ik}\boldsymbol{\beta}) (\exp(2\sigma^2) - \exp(\sigma^2)). \end{aligned} \quad (16.11)$$

The expression in round parentheses in (16.10) is always greater than 1. Thus, although $y_{ij}|u_i$ follows a regular Poisson distribution, the marginal distribution of y_{ij} is over-dispersed. According to (16.11), due to the random intercept, observations on the same subject are no longer independent, as is desirable for clustered data. Actuarial literature on ratemaking (see, e.g., Denuit et al. 2007, and Antonio and Valdez 2012) often uses a slightly modified version of the normality assumption, namely $u_i \sim N(-\frac{\sigma^2}{2}, \sigma^2)$. This leads to

$$\begin{aligned} E[y_{ij}] &= E[E[y_{ij}|u_i]] = \exp(\mathbf{x}'_i \boldsymbol{\beta} - \frac{\sigma^2}{2} + \frac{\sigma^2}{2}) \\ &= \exp(\mathbf{x}'_i \boldsymbol{\beta}), \\ E[y_{ij}|u_i] &= \exp(\mathbf{x}'_i \boldsymbol{\beta} + u_i). \end{aligned} \tag{16.12}$$

In actuarial parlance, the so-called a priori premium ($E[y_{ij}]$), specified as $\exp(\mathbf{x}'_i \boldsymbol{\beta})$, uses only a priori measurable risk factors (such as gender, age, and car capacity). It is the marginal mean of y_{ij} and is therefore correct on average. The a posteriori correction factor, $\exp(u_i)$, adjusts the a priori tariff based on the observed claim history of the insured. We estimate this factor by predicting u_i .

Example 16.2 (An Illustration of Shrinking). We consider a claim frequency model using the auto claim data from Yip and Yau (2005), where we specify a log-linear Poisson model with `Jobclass` as a random effect. In particular, we are interested in how the estimate for each job class level differs between the mixed model and the GLM where `Jobclass` enters as a factor fixed effect. This is the difference between a partial pooling approach (with mixed models) and the “no pooling” approach (with cluster-specific intercepts); see our discussion in Chapter 8. Figure 16.1 shows such a comparison on the estimation of job class levels. The horizontal dotted line corresponds to the estimated intercept from the mixed model and represents the average effect for all job categories because all the random effects have zero means. That is, it is roughly the estimate when all job categories are pooled together. In contrast, the estimates from the generalized linear model (the points) can be viewed as the individual estimate for each job class level, ignoring the other levels – indeed, fitting a GLM with only the job class as a predictor is equivalent to fitting eight separate GLMs on each subset of data with a unique job class because of the orthogonal design matrix corresponding to the job class. We see that the mixed model (the triangle) shrinks the separate estimates from the GLM toward the pooled group-level estimate across all the job classes. The shrinkage is most significant for Lawyer, Professional, and Student. Therefore, the generalized linear mixed model captures the core insight of the credibility models, where the estimates from the mixed models can be viewed as the weighted average between the pooled group-level estimate and the separate

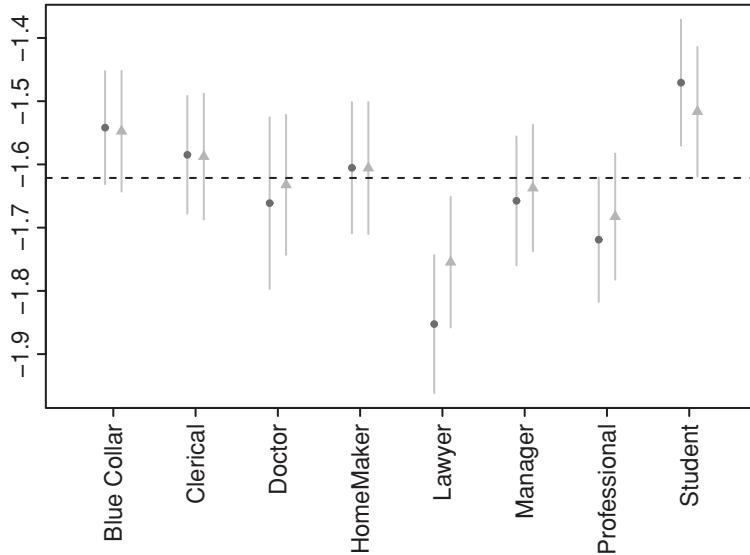


Fig. 16.1. The job class estimates from the generalized linear model (●) and the Poisson mixed models (△) in the auto insurance frequency model. The horizontal line is the average estimate for all job classes, and the vertical lines show the uncertainty intervals based on \pm one standard error.

individual estimates. As a result, the mixed model produces less extreme estimates while still accounting for the heterogeneity across the various levels.

16.3.3 Estimation

Using the model specifications in (16.4) it is straightforward to write down the likelihood of the corresponding GLMM:

$$L(\boldsymbol{\beta}, \mathbf{D} | \mathbf{y}) = \int f_{\mathbf{Y}|\mathbf{u}}(\mathbf{y}|\mathbf{u}) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u}, \quad (16.13)$$

where the integral goes over the random effects vector \mathbf{u} (with covariance matrix \mathbf{D}). The presence of the integral in (16.13) hinders maximum likelihood estimation and prohibits explicit expressions for estimators and predictors, such as those derived for LMMs. Only so-called *conjugate* distributional specifications lead to a closed-form solution in (16.13), a normal distribution for the response, combined with normally distributed random effects (as with LMMs), being one example. More general model assumptions require approximate techniques to estimate $\boldsymbol{\beta}$, \mathbf{D} and predict the random effect for cluster i , \mathbf{u}_i . As in Molenberghs and Verbeke (2005) we distinguish three approaches to tackle this problem: approximating the integrand, approximating the data, and approximating the integral (through numerical integration). Having

Pinheiro and Bates (2000), McCulloch and Searle (2001), (chapters 8 and 10) and Tuerlinckx et al. (2006) as main references, we discuss next some highlights of these methods. This discussion will help readers understand the differences among different R packages available for data analysis with GLMMs (as demonstrated in Section 16.6). Section 16.3.4 presents pros and cons of the techniques mentioned in 16.3.3.1, 16.3.3.2, and 16.3.3.3, as well as references to other techniques (not discussed here). We postpone a discussion of a Bayesian approach to Section 16.5.

16.3.3.1 Approximating the Likelihood: the Laplace Method

The Laplace method (see Tierny and Kadane 1986) approximates integrals of the form

$$\int e^{h(\mathbf{u})} d\mathbf{u}. \quad (16.14)$$

for some function h of a q -dimensional vector \mathbf{u} . The method relies on a second-order Taylor expansion of $h(\mathbf{u})$ around its maximum $\hat{\mathbf{u}}$:

$$h(\mathbf{u}) \approx h(\hat{\mathbf{u}}) + \frac{1}{2}(\mathbf{u} - \hat{\mathbf{u}})'h''(\hat{\mathbf{u}})(\mathbf{u} - \hat{\mathbf{u}}), \quad (16.15)$$

with

$$\frac{\partial h(\mathbf{u})}{\partial \mathbf{u}}|_{\mathbf{u}=\hat{\mathbf{u}}} = \mathbf{0}, \quad (16.16)$$

and $h''(\hat{\mathbf{u}}) = \left. \frac{\partial^2 h(\mathbf{u})}{\partial \mathbf{u} \partial \mathbf{u}'} \right|_{\mathbf{u}=\hat{\mathbf{u}}}$ the matrix with second-order derivatives of h , evaluated at $\hat{\mathbf{u}}$.

We replace $h(\mathbf{u})$ with the approximation from (16.15):

$$\int e^{h(\mathbf{u})} d\mathbf{u} \approx \int e^{h(\hat{\mathbf{u}}) + \frac{1}{2}(\mathbf{u} - \hat{\mathbf{u}})'h''(\hat{\mathbf{u}})(\mathbf{u} - \hat{\mathbf{u}})} d\mathbf{u}. \quad (16.17)$$

Approximating the density of \mathbf{u} with a multivariate Gaussian distribution $\mathcal{N}(\hat{\mathbf{u}}, (-h''(\hat{\mathbf{u}}))^{-1})$ leads to

$$\int e^{h(\mathbf{u})} d\mathbf{u} \approx (2\pi)^{q/2} |-h''(\hat{\mathbf{u}})|^{-1/2} e^{h(\hat{\mathbf{u}})}. \quad (16.18)$$

This technique is readily available to approximate the likelihood in a GLMM (see Breslow and Clayton 1993, and McCulloch and Searle 2001, among other references):

$$\begin{aligned} \ell &= \log \int f_{Y|U}(\mathbf{y}|\mathbf{u}) f_U(\mathbf{u}) d\mathbf{u} \\ &= \log \int e^{\log f_{Y|U}(\mathbf{y}|\mathbf{u}) + \log f_U(\mathbf{u})} d\mathbf{u} \\ &= \log \int e^{h(\mathbf{u})} d\mathbf{u}, \end{aligned} \quad (16.19)$$

with $h(\mathbf{u}) := \log f_{Y|U}(y|\mathbf{u}) + \log f_U(\mathbf{u}) = \log f_{Y|U}(y|\mathbf{u}) - \frac{1}{2}\mathbf{u}'\mathbf{D}^{-1}\mathbf{u} - \frac{q}{2}\log 2\pi - \frac{1}{2}\log |\mathbf{D}|$. Expression (16.16) should be solved numerically and requires

$$\begin{aligned} \frac{\partial h(\mathbf{u})}{\partial \mathbf{u}} &= \frac{\partial \log f_{Y|U}(y|\mathbf{u})}{\partial \mathbf{u}} - \mathbf{D}^{-1}\mathbf{u} = \mathbf{0} \\ &\Updownarrow \\ \frac{1}{\phi} \mathbf{Z}' \mathbf{W} \Delta (\mathbf{y} - \boldsymbol{\mu}) - \mathbf{D}^{-1}\mathbf{u} &= \mathbf{0}, \end{aligned} \quad (16.20)$$

where \mathbf{W} and Δ are diagonal matrices with elements $[V(\mu_i)(g'(\mu_i))^2]^{-1}$ and $g'(\mu_i)$, respectively.¹ Hereby $g(\mu_i)$ and $V(\mu_i)$ are the mean and variance of y_i , conditional on \mathbf{u}_i , as introduced in (16.2).

The matrix of second-order derivatives is (see (16.18))

$$\frac{\partial^2 h(\mathbf{u})}{\partial \mathbf{u} \partial \mathbf{u}'} = -\frac{1}{\phi} \mathbf{Z}' \mathbf{W} \Delta \frac{\partial \boldsymbol{\mu}}{\partial \mathbf{u}'} + \frac{1}{\phi} \mathbf{Z}' \frac{\partial \mathbf{W} \Delta}{\partial \mathbf{u}'} (\mathbf{y} - \boldsymbol{\mu}) - \mathbf{D}^{-1}. \quad (16.21)$$

The random vector corresponding with the second term in this expression has expectation zero, with respect to $f_{Y|U}(y|\mathbf{u})$, and will be ignored. Therefore,

$$\begin{aligned} -\frac{\partial^2 h(\mathbf{u})}{\partial \mathbf{u} \partial \mathbf{u}'} &= \frac{1}{\phi} \mathbf{Z}' \mathbf{W} \Delta \Delta^{-1} \mathbf{Z} + \mathbf{D}^{-1} \\ &= \left(\frac{1}{\phi} \mathbf{Z}' \mathbf{W} \mathbf{Z} \mathbf{D} + \mathbf{I} \right) \mathbf{D}^{-1}. \end{aligned} \quad (16.22)$$

Using this expression, an approximation to the log-likelihood in (16.19) follows:

$$\begin{aligned} \ell &\approx \log f_{Y|U}(y|\hat{\mathbf{u}}) - \frac{1}{2}\hat{\mathbf{u}}'\mathbf{D}^{-1}\hat{\mathbf{u}} - \frac{q}{2}\log 2\pi - \frac{1}{2}\log |\mathbf{D}| \\ &\quad + \frac{q}{2}\log 2\pi - \frac{1}{2}\log |(\mathbf{Z}' \mathbf{W} \mathbf{Z} \mathbf{D}/\phi + \mathbf{I}) \mathbf{D}^{-1}| \\ &= \log f_{Y|U}(y|\hat{\mathbf{u}}) - \frac{1}{2}\hat{\mathbf{u}}'\mathbf{D}^{-1}\hat{\mathbf{u}} + \frac{1}{2}\log |\mathbf{Z}' \mathbf{W} \mathbf{Z} \mathbf{D}/\phi + \mathbf{I}|. \end{aligned} \quad (16.23)$$

This expression should be maximized with respect to β . Assuming \mathbf{W} is not changing a lot as a function of β , the last term can be ignored and

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{\phi} \mathbf{X}' \mathbf{W} \Delta (\mathbf{y} - \boldsymbol{\mu}). \quad (16.24)$$

¹ Derivations are similar to those in Chapter 5 on GLMs, and basically go as follows:

$$\begin{aligned} \frac{\partial \log f_{Y|U}(y|\mathbf{u})}{\partial \mathbf{u}} &= \frac{1}{\phi} \sum_i \left(y_i \frac{\partial \theta_i}{\partial \mathbf{u}} - \frac{\partial \psi(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mathbf{u}} \right) \\ &= \frac{1}{\phi} \sum_i (y_i - \mu_i) \frac{1}{V(\mu_i)} \frac{1}{g'(\mu_i)} \mathbf{z}'_i. \end{aligned}$$

Therefore, the following set of equations has to be solved simultaneously with respect to β and \mathbf{u} (using a numerical optimization method):

$$\begin{aligned} \frac{1}{\phi} \mathbf{X}' \mathbf{W} \Delta(\mathbf{y} - \boldsymbol{\mu}) &= \mathbf{0} \\ \frac{1}{\phi} \mathbf{Z}' \mathbf{W} \Delta(\mathbf{y} - \boldsymbol{\mu}) &= \mathbf{D}^{-1} \mathbf{u}. \end{aligned} \quad (16.25)$$

This set of equations also arises by jointly maximizing (with respect to β and \mathbf{u})

$$\log f_{Y|U}(\mathbf{y}|\mathbf{u}) - \frac{1}{2} \mathbf{u}' \mathbf{D}^{-1} \mathbf{u}, \quad (16.26)$$

which is a quasi-likelihood term, $f_{Y|U}(\mathbf{y}|\mathbf{u})$, augmented with a penalty term, $\mathbf{u}' \mathbf{D} \mathbf{u}$. Hence, the name penalized quasi-likelihood (PQL) for (16.26). Breslow and Clayton (1993) present a Fisher scoring algorithm and its connection with Henderson's mixed model equations for simultaneous solution of the set of equations in (16.25). This approach is discussed in the next section.

16.3.3.2 Approximating the Data: Pseudo-Likelihood (PL)

Wolfinger and O'Connell (1993) develop pseudo-likelihood (PL) (or restricted pseudo-likelihood, REPL) in the context of GLMMs. This approach generalizes the idea of a *working variate*, introduced for maximum likelihood estimators with GLMs, to the case of GLMMs (see Breslow and Clayton 1993, and McCulloch and Searle 2001). In the context of GLMs Nelder and Wedderburn (1972) define a working variate t_i as follows:

$$\begin{aligned} t_i &= g(\mu_i) + g'(\mu_i)(y_i - \mu_i) \\ &= \mathbf{x}_i' \beta + g'(\mu_i)(y_i - \mu_i). \end{aligned} \quad (16.27)$$

Estimates of β follow from iteratively fitting a weighted linear regression of t on \mathbf{X} , until the estimates converge. In a GLMM we generalize the notion of a working variate t_i as follows:

$$t_i = \mathbf{x}_i' \beta + \mathbf{z}_i' \mathbf{u} + g'(\mu_i)(y_i - \mu_i). \quad (16.28)$$

This is a first-order Taylor expansion of $g(y_i)$ around the conditional mean μ_i . In matrix notation the vector of working variates, \mathbf{t} , becomes

$$\mathbf{t} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \Delta(\mathbf{y} - \boldsymbol{\mu}), \quad (16.29)$$

with Δ a diagonal matrix with entries $g'(\mu_i)$. Calculating the variance of \mathbf{t} is complicated because of the dependence of Δ on $\boldsymbol{\mu}$ (and therefore on the random vector \mathbf{u}).

A simplification is possible by replacing $\boldsymbol{\mu}$ with $\hat{\boldsymbol{\mu}}$ in the variance matrix (see Wolfinger and O'Connell 1993). Consequently,

$$\begin{aligned}\text{Var}(\mathbf{t}) &= \mathbf{ZDZ}' + \Delta_{\hat{\boldsymbol{\mu}}} \text{Var}(\mathbf{Y} - \boldsymbol{\mu})_{\hat{\boldsymbol{\mu}}} \Delta_{\hat{\boldsymbol{\mu}}} \\ &:= \mathbf{ZDZ}' + \Sigma_{\hat{\boldsymbol{\mu}}}.\end{aligned}\quad (16.30)$$

The working variate \mathbf{t} approximately follows a linear mixed model (as in Chapter 8), with design matrices \mathbf{X} (fixed effects), \mathbf{Z} (random effects), \mathbf{D} the covariance matrix of the random effects, and Σ the covariance matrix of the error terms. In this LMM it is straightforward to estimate $\boldsymbol{\beta}$, \mathbf{u} , and the unknown variance components. Therefore, the pseudo-likelihood algorithm goes as follows. Starting from initial estimates of $\boldsymbol{\beta}$, \mathbf{u} , and the variance components, the working variates in (16.29) are evaluated. Consequently, using LMM methodology, updated estimates follow from (16.29) and (16.30). These steps are repeated until there is convergence of the estimates.

16.3.3.3 Approximating the Integral: Numerical Integration Techniques

Approximating the integral in (16.13) with a so-called (adaptive) quadrature rule for numerical integration is based on Liu and Pierce (1994). For ease of explanation we consider next the case of a one-dimensional integral. The case with multidimensional integrals is documented in Tuerlinckx et al. (2006).

Nonadaptive Gauss–Hermite quadrature. *Non-adaptive* Gauss–Hermite quadrature approximates an integral of the form

$$\int_{-\infty}^{+\infty} h(z) \exp(-z^2) dz, \quad (16.31)$$

with a weighted sum, namely

$$\int_{-\infty}^{+\infty} h(z) \exp(-z^2) dz \approx \sum_{l=1}^Q w_l h(z_l). \quad (16.32)$$

Q is the order of the approximation, the z_l are the zeros of the Q th order Hermite polynomial, and the w_l are corresponding weights. The nodes (or quadrature points) z_l and the weights w_l are tabulated in Abramowitz and Stegun (1972, p. 924). The quadrature points used in (16.32) do not depend on h . Therefore, it is possible that only a very few nodes lie in the region where most of the mass of h is, which would lead to poor approximations.

Adaptive Gauss–Hermite quadrature. With an *adaptive* Gauss–Hermite quadrature rule, the nodes are rescaled and shifted such that the integrand is sampled in a

suitable range. Assume $h(z)\phi(z; 0, 1)$ is unimodal, and consider the numerical integration of $\int_{-\infty}^{+\infty} h(z)\phi(z; 0, 1)dz$. Let $\hat{\mu}$ and $\hat{\nu}$ be

$$\hat{\mu} = \text{mode}[h(z)\phi(z; 0, 1)] \quad \text{and} \quad \hat{\nu}^2 = \left[-\frac{\partial^2}{\partial z^2} \ln(h(z)\phi(z; 0, 1)) \Big|_{z=\hat{\mu}} \right]^{-1} \quad (16.33)$$

Acting as if $h(z)\phi(z; 0, 1)$ were a Gaussian density, $\hat{\mu}$ and $\hat{\nu}$ would be the mean and variance of this density. The quadrature points in the adaptive procedure, z_l^* , are centered at $\hat{\mu}$ with spread determined by $\hat{\nu}$, namely

$$z_l^* = \hat{\mu} + \sqrt{2}\hat{\nu}z_l \quad (16.34)$$

with ($l = 1, \dots, Q$). Now rewrite $\int_{-\infty}^{+\infty} h(z)\phi(z; 0, 1)dz$ as

$$\int_{-\infty}^{+\infty} \frac{h(z)\phi(z; 0, 1)}{\phi(z; \mu, \nu)} \phi(z; \mu, \nu) dz, \quad (16.35)$$

where $\phi(z; \mu, \nu)$ is the Gaussian density function with mean μ and variance ν^2 . Using simple manipulations it is easy to see that for a suitably regular function v ,

$$\begin{aligned} \int_{-\infty}^{+\infty} v(z)\phi(z; \mu, \nu) dz &= \int_{-\infty}^{+\infty} v(z)(2\pi\nu^2)^{-1/2} \exp\left(-\frac{1}{2}\left(\frac{z-\mu}{\nu}\right)^2\right) dz \\ &= \int_{-\infty}^{+\infty} \frac{v(\mu + \sqrt{2}\nu z)}{\sqrt{\pi}} \exp(-z^2) dz \\ &\approx \sum_{l=1}^Q \frac{v(\mu + \sqrt{2}\nu z_l)}{\sqrt{\pi}} w_l. \end{aligned} \quad (16.36)$$

Using $\frac{h(z)\phi(z; 0, 1)}{\phi(z; \mu, \nu)}$ instead of $v(z)$ and replacing μ and ν with their estimates from (16.33), the following quadrature formula results:

$$\begin{aligned} \int_{-\infty}^{+\infty} h(z)\phi(z; 0, 1) dz &\approx \sqrt{2}\hat{\nu} \sum_{l=1}^Q w_l \exp(z_l^2) \phi(z_l^*; 0, 1) h(z_l^*) \\ &= \sum_{l=1}^Q w_l^* h(z_l^*), \end{aligned} \quad (16.37)$$

with adaptive weights $w_l^* := \sqrt{2}\hat{\nu}w_l \exp(z_l^2) \phi(z_l^*; 0, 1)$. Expression (16.37) is an *adaptive Gauss-Hermite quadrature formula*.

Link with Laplace approximation. We illustrate the connection between the Laplace approximation (from Section 16.3.3.1) and the adaptive Gauss–Hermite quadrature with a single node. Indeed, when $Q = 1$ (i.e., the case of a single node), $z_1 = 0$ (from the Hermite polynomial) and $w_1 = 1$. The corresponding adaptive

node and weight are $z_1^* = \hat{\mu}$ and $w_1^* = \sqrt{2}\hat{v}\phi(z_1^*; 0, 1)$, respectively. The adaptive GH quadrature formula then becomes

$$\int h(z)\phi(z; 0, 1)dz \approx \sqrt{2}\hat{v} \exp(\log(\phi(z_1^*; 0, 1)h(z_1^*))) \\ \times (2\pi)^{1/2} \underbrace{\left| -\frac{\partial^2}{\partial z^2} \log(h(z)\phi(z; 0, 1)) \right|_{z=\hat{\mu}}^{-1/2}}_{\hat{v}} \exp\{\log(\phi(\hat{\mu}; 0, 1)h(\hat{\mu}))\}, \quad (16.38)$$

where $\hat{\mu} = z_1^*$ maximizes $h(z)\phi(z; 0, 1)$. This corresponds with the Laplace formula from (16.18).

Adaptive Gauss–Hermite quadrature for GLMMs. We describe the case of a GLMM with a single, normally distributed random effect $u_i \sim N(0, \sigma^2)$ for each cluster i . The use of an adaptive Gauss–Hermite quadrature with GLMMs starts from determining the posterior mode of u_i . Since this posterior distribution depends on unknown fixed effects and variance parameters, we replace the unknown β , ϕ , and σ with their current estimates: $\hat{\beta}^{(c)}$, $\hat{\phi}^{(c)}$, and $\hat{\sigma}^{(c)}$. Using these current estimates \hat{u}_i maximizes

$$f(\mathbf{y}_i|u_i)f(u_i|\hat{\sigma}^{(c)}), \quad (16.39)$$

which is proportional to the posterior density of u_i , given \mathbf{y}_i :

$$f(u_i|\mathbf{y}_i) = \frac{f(\mathbf{y}_i|u_i)f(u_i|\hat{\sigma}^{(c)})}{\int f(\mathbf{y}_i|u_i)f(u_i|\hat{\sigma}^{(c)})du_i} \\ \propto f(\mathbf{y}_i|u_i)f(u_i|\hat{\sigma}^{(c)}). \quad (16.40)$$

Therefore \hat{u}_i is the posterior mode of u_i . We also determine (numerically) \hat{v}_i^2 as

$$\hat{v}_i^2 = \left[-\frac{\partial^2}{\partial u_i^2} \ln(f(\mathbf{y}_i|u_i)f(u_i|\hat{\sigma}^{(c)})) \Big|_{u_i=\hat{u}_i} \right]^{-1}. \quad (16.41)$$

Using an adaptive Gauss–Hermite quadrature rule we approximate the likelihood contribution of cluster i as follows (with $\delta_i := \sigma^{-1}u_i \sim N(0, 1)$):

$$\int f_{Y|U}(\mathbf{y}_i|u_i)f_U(u_i)du_i = \int f_{Y|U}(\mathbf{y}_i|\delta_i)\phi(\delta_i|0, 1)d\delta_i \\ = \int \left(\prod_{j=1}^{n_i} f_{Y|U}(y_{ij}|\delta_i) \right) \phi(\delta_i|0, 1)d\delta_i \\ \approx \sum_{l=1}^Q w_l^* \left(\prod_{j=1}^{n_i} f_{Y|U}(y_{ij}|z_l^*) \right), \quad (16.42)$$

with adaptive weights $w_l^* = \sqrt{2}\hat{v}_l w_l \exp(z_l^2)\phi(z_l^*; 0, 1)$ and $z_l^* = \hat{\delta}_l + \sqrt{2}\hat{v}_l z_l$. In this expression the linear predictor corresponding with $f_{Y|U}(y_{ij}|\delta_i)$ and $f_{Y|U}(y_{ij}|z_l^*)$, respectively, is $\mathbf{x}'_{ij}\boldsymbol{\beta} + \sigma\delta_i$ and $\mathbf{x}'_{ij}\boldsymbol{\beta} + \sigma z_l^*$. Multiplying (16.42) over all clusters i leads to the total likelihood. Maximizing the latter over the fixed effects regression parameters, the dispersion parameter, and the variance components leads to updated parameter estimates $\hat{\boldsymbol{\beta}}^{(c+1)}$, $\hat{\phi}^{(c+1)}$, and $\hat{\sigma}^{(c+1)}$. We predict the cluster-specific random effects with the posterior modes from (16.39).

16.3.4 Pros and Cons of Various Estimation Methods for GLMMs

Laplace and PQL methods for estimation within GLMMs rely on quite a few approximations. Breslow and Lin (1995) and Lin and Breslow (1996) investigate settings in which PQL (which results in the iterative approach from Section 16.3.3.2) performs poorly, and they discuss the limits of this approach. Based on this, McCulloch and Searle (2001, p. 283) conclude, “*We thus cannot recommend the use of simple PQL methods in practice.*” The Gauss–Hermite quadrature approach is more accurate than PQL but is limited to GLMMs with a small number of nested random effects. It is not possible to handle a large number of random effects, crossed random effects, or high levels of nesting with this approach. Moreover, a Gauss–Hermite quadrature is explicitly designed for normally distributed random effects, although other quadrature formulas exist (not discussed here). Kim et al. (2013) provide a recent comparison of the estimation methods in the context of logistic regression.

The (Monte Carlo) EM algorithm and simulated maximum likelihood or Monte Carlo integration (see McCulloch and Searle 2001, chapter 10, or Tuerlinckx et al. 2006) are alternative methods for estimation with GLMMs. We discuss a Bayesian implementation of (G)LMMs in Section 16.5. This is a way to circumvent the estimation problems discussed earlier.

16.3.5 Statistical Inference with GLMMs

The general ideas on statistical inference with LMMs carry over to GLMMs, where fitting is based on maximum likelihood principles. Wald, score, and likelihood ratio tests (LRT) are available for testing fixed effects parameters, as well as variance components. However, closed-form expressions – for example, for the covariance matrix of $\hat{\boldsymbol{\beta}}$ – are no longer available. Numerical evaluation of the inverse Fisher information matrix is required for precision estimates. When using the PL method as described in Section 16.3.3.2, the original likelihood expression should be used in an LRT, and not the likelihood of the LMM that is specified for the pseudo-data. As with LMMs, testing the necessity of a random effect is problematic, because the

corresponding null hypothesis constrains the variance of the random effect to the boundary of its parameter space. With respect to inference with (G)LMMs a Bayesian analysis has some additional features (see Section 16.5 for discussion).

16.4 Nonlinear Mixed Models

LMMs and GLMMs model the mean (in LMMs) or a transformation of the (conditional) mean (in GLMMs) as *linear* in the fixed effects parameters β and the random effects u . Nonlinear mixed models (NLMM) extend the concept of linear predictors. In an NLMM the conditional distribution of y_{ij} (being the j th response on cluster i), given u_i , belongs to the exponential family with mean structure

$$E[y_{ij}|u_i] = h(x_{ij}, \beta, z_{ij}, u_i), \quad (16.43)$$

where $h(\cdot)$ is an arbitrary function of covariates, parameters, and random effects. A distributional assumption for the random effects completes the model assumptions; typically $u_i \sim N(\mathbf{0}, D)$. GLMMs are therefore a subclass of the general class of NLMMs. (Adaptive) Gauss–Hermite quadrature is available for ML estimation within NLMMs. A fully Bayesian analysis is an alternative approach.

16.5 Bayesian Approach to (L,GL,NL)MMs

The presence of random effects is an essential feature in the hierarchical model formulation of a mixed model. A link with Bayesian statistics is then straightforward, because the random effects have explicit distributional assumptions. In addition to the distribution of the random effects u and the distributional framework for the response y , a Bayesian analysis requires prior distributions for β (ϕ in GLMMs) and D . Inference is based on simulated samples from the posterior distribution of the parameters, which is (with m clusters)

$$\begin{aligned} & f(\beta, D, \phi, u_1, \dots, u_m | y_1, \dots, y_m) \\ & \propto \prod_{i=1}^m f_i(y_i | \beta, \phi, u_1, \dots, u_m) \cdot \prod_{i=1}^m f(u_i | D) \cdot f(D) \cdot f(\beta) \cdot f(\phi). \end{aligned} \quad (16.44)$$

We refer to Chapters 13 and 14 on Bayesian concepts and regression models for an overview of useful concepts and simulation methods. For GLMMs in particular Zhao et al. (2006) and the references herein are a good starting point for Bayesian (L,G,NL)MMs.

Bayesian multilevel models have some very nice features. As discussed in Chapter 8 on LMMs (see Section 8.2), precision estimates based on MLE require variance components estimates to be plugged in and are therefore not able to account for all sources of randomness. A fully Bayesian approach, with a prior specified for

each parameter (vector), solves this issue and provides a way to circumvent otherwise intractable calculations. The likelihood approximations discussed in Section 16.3.3 are replaced in a Bayesian analysis with general MCMC methodology for sampling from posterior distributions. This allows specification of more complex hierarchical model structures, such as the spatial structures in Chapter 11 or the three-level count data models in Antonio, Frees, and Valdez (2010). Moreover, the Bayesian methodology is not limited to Gaussian random effects. For predictive modeling in actuarial science, Bayesian statistics are particularly useful for simulation from the posterior (predictive) distribution of quantities of interest, such as a policy's random effect or the number of claims in a future time period.

16.6 Example: Poisson Regression for Workers' Compensation Insurance Frequencies

We analyze the data from Example 8.3 (see Chapter 8) on claim counts reported by 133 occupation classes with respect to workers' compensation insurance policies. Each occupation class is followed over seven years. The response variable of interest is Count_{ij} , the number of claims registered per occupation class i during year j . To enable out-of-sample predictions, we split the data in a training (without Count_{i7}) versus validation set (the Count_{i7} observations). We remove observations with zero payroll. Models are estimated on the training set, and centering of covariate `Year` is applied. Since the data are claim counts, we investigate the use of Poisson regression models. Throughout our analysis we include $\log(\text{Payroll}_{ij})$ as an offset in the regression models, because the number of accidents should be interpreted relative to the size of the risk class.

From the discussion in Section 16.3.3 we are aware of (at least) three ways to tackle the problem of likelihood optimization with GLMMs. Correspondingly, multiple R packages are available for calibrating GLMMs to data. We illustrate hereafter the packages `lme4`, `glmmML`, and the function `glmmPQL` from library `MASS` for likelihood-based estimation of a Poisson model with random effects. The illustration ends with a demonstration of a Bayesian analysis of this Poisson regression model.

Complete pooling. Similar to our approach in Chapter 8, we start with a *complete pooling model*, ignoring the clustering of data in occupation classes. This is a simple Poisson regression model, with an overall intercept β_0 and an overall slope β_1 for the effect of `Year`:

$$\begin{aligned} \text{Count}_{ij} &\sim \text{POI}(\text{Payroll}_{ij} \cdot \lambda_{ij}) \\ \lambda_{ij} &= \exp(\beta_0 + \beta_1 \cdot \text{Year}_{ij}). \end{aligned} \tag{16.45}$$

A Poisson regression is an example of a generalized linear model (see Chapter 5) for which the `glm` function in R is available.

```
> fitglm.CP <- glm(count~yearcentr, offset=log(payroll), family=poisson,
                     data=wcfit)
> summary(fitglm.CP)

Call:
glm(formula = count ~ yearcentr, family = poisson, data = wcfit,
     offset = log(payroll))

Deviance Residuals:
    Min      1Q  Median      3Q      Max 
-26.8194 -1.0449   0.2456   2.3197  18.1740 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -3.702172   0.008648 -428.105 <2e-16 ***
yearcentr   -0.010155   0.005098   -1.992   0.0464 *  
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1    1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 12274  on 766  degrees of freedom
Residual deviance: 12270  on 765  degrees of freedom
AIC: 14904

Number of Fisher Scoring iterations: 5
```

The estimate $\hat{\beta}_0$ for the intercept is -3.702 (with s.e. 0.00865) and $\hat{\beta}_1 = -0.0102$ (with s.e. 0.00510).

No pooling. We continue the analysis with a fixed effects Poisson model that specifies an occupation class specific intercept, say $\beta_{0,i}$, for each of the 113 occupation classes in the dataset, as well as a global fixed Year effect. The intercepts $\beta_{0,i}$ are unknown, but fixed. We fit the model in R with the `glm` function, identifying the occupation class as a factor variable:

$$\text{Count}_{ij} \sim \text{POI}(\text{Payroll}_{ij} \cdot \lambda_{ij})$$

$$\lambda_{ij} = \exp(\beta_{0,i} + \beta_1 \cdot \text{Year}_{ij}). \quad (16.46)$$

```

> fitglm.NP <- glm(count~0+yearcentr+factor(riskclass),
  offset=log(payroll),
  family=poisson(), data=wcFit)
> summary(fitglm.NP)
Call:
glm(formula = count ~ 0 + yearcentr + factor(riskclass),
  family = poisson(),
  data = wcFit, offset = log(payroll))

Deviance Residuals:
    Min      1Q   Median      3Q      Max
-4.2403 -0.8507 -0.1629  0.7186  7.1909

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
yearcentr      9.918e-03  5.157e-03   1.923 0.054448 .
factor(riskclass)1 -2.578e+00  2.425e-01 -10.630 < 2e-16 ***
factor(riskclass)2 -3.655e+00  4.082e-01  -8.952 < 2e-16 ***
factor(riskclass)3 -3.683e+00  1.374e-01 -26.810 < 2e-16 ***
factor(riskclass)4 -1.309e+01  2.103e+03  -0.006 0.995035
factor(riskclass)5 -2.737e+00  9.325e-02 -29.347 < 2e-16 ***
...
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1    1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 980297.4  on 767  degrees of freedom
Residual deviance: 1192.9  on 636  degrees of freedom
AIC: 4084.4

Number of Fisher Scoring iterations: 14

```

Comparing the deviance of the *complete* and the *no pooling* model results in a drop-in-deviance of $12,270 - 1,192 = 11,078$, with a difference in degrees of freedom of 129. In R the anova function is available for this comparison. With a *p*-value of $< 2.2 \cdot 10^{-16}$ the ‘no pooling’ model outperforms the ‘complete pooling’ model.

```
> anova(fitglm.CP,fitglm.NP,test="Chisq")
Analysis of Deviance Table

Model 1: count ~ yearcentr
Model 2: count ~ 0 + yearcentr + factor(riskclass)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       765    12270.4
2       636    1192.9 129     11078 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

GLMMs: Random intercepts, Laplace approximation with `lmer`. We investigate a Poisson regression model with random occupation class specific intercepts as a meaningful alternative to the no pooling model. The model formulation is

$$\begin{aligned} \text{Count}_{ij}|u_{i,0} &\sim \text{POI}(\text{Payroll}_{ij} \cdot (\lambda_{ij}|u_{i,0})) \\ \lambda_{ij}|u_{i,0} &= \exp(\beta_0 + u_{i,0} + \beta_1 \cdot \text{Year}_{ij}) \\ u_{i,0} &\sim N(0, \sigma_u^2). \end{aligned} \quad (16.47)$$

We first fit this random intercepts model with the `lmer` (or: `glmer`) function from the R library `lme4`. By default `lmer` uses Laplace approximation (see Section 16.3.3.1) to approximate the Poisson likelihood. Adaptive Gauss–Hermite quadrature is also available within the `lme4` package (see *infra*).²

```
> hml1 <- glmer(count~(1|riskclass)+yearcentr+offset(log(payroll)) ,
+                  family=poisson(link="log") , data=wcFit)
> print(hml1)
Generalized linear mixed model fit by the Laplace approximation
Formula: count ~ (1 | riskclass) + yearcentr + offset(log(payroll))
Data: wcFit
AIC  BIC logLik deviance
1771 1785 -882.6      1765
Random effects:
 Groups      Name        Variance Std.Dev.
 riskclass (Intercept) 0.80475  0.89708
Number of obs: 767, groups: riskclass, 130
```

² The `glmmML` package is another R package for Laplace approximation and Gauss–Hermite quadrature for Binomial and Poisson random effects models; see the book’s website for sample code.

```

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.562125   0.083489 -42.67   <2e-16 ***
yearcentr    0.009730   0.005156    1.89    0.0592 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Correlation of Fixed Effects:
        (Intr)
yearcentr -0.010

```

From the R output just given, we conclude that $\hat{\beta}_0 = -3.5621$ (with s.e. 0.0835), $\hat{\beta}_1 = 0.00973$ (with s.e. 0.00516), and $\hat{\sigma}_u^2 = 0.805$ is the estimate of the variance of random intercepts. Extracting the estimates of fixed and random effects, as well as prediction intervals for the random intercepts, goes as follows.

```

> ## get fixed effects
> fixef(hlm1)
(Intercept)      yearcentr
-3.562124594  0.009730364

> ## get random intercepts
> int <- ranef(hlm1)$riskclass

> ## get prediction intervals for r.e.'s
> str(rr1 <- ranef(hlm1, condVar = TRUE))
> # s.e. for 'riskclass' r.e.
> my.se.risk = sqrt(as.numeric(attributes(rr1$riskclass)$postVar))
> # get prediction intervals for random intercepts (per riskclass)
> lower.risk <- rr1$riskclass[[1]]-1.96*my.se.risk
> upper.risk <- rr1$riskclass[[1]]+1.96*my.se.risk
> int.risk <- cbind((lower.risk),(rr1$riskclass[[1]]),(upper.risk))
> colnames(int.risk) <- c("Lower","Estimate R.E.","Upper")
# you can use these to create error bar plots
> int.risk[1:5,]
      Lower Estimate R.E.      Upper
[1,]  0.4407844  0.9146787935 1.3885732
[2,] -0.8002651 -0.0767152172 0.6468347
[3,] -0.3834920 -0.1177250934 0.1480418

```

```
[4,] -1.7583808 -0.0009170246 1.7565468
[5,]  0.6340478  0.8166379517 0.9992281
>
> ## variance components
> VarCorr(hlm1)
$riskclass
  (Intercept)
(Intercept) 0.8047458
attr(,"stddev")
(Intercept)
  0.8970762
attr(,"correlation")
  (Intercept)
(Intercept)      1

attr(,"sc")
[1] NA
```

We are now ready to compare model (16.46) (no pooling) to (16.47) (random intercepts). The left panel in Figure 16.2 shows the intercepts and corresponding error bars from the no pooling model, together with one standard error, against the total size of the occupation class (i.e., $\sum_j \text{Payroll}_{ij}$). Figure 16.2 (right) shows the point predictions of the random intercepts. To create this plot we refit the random effects model and

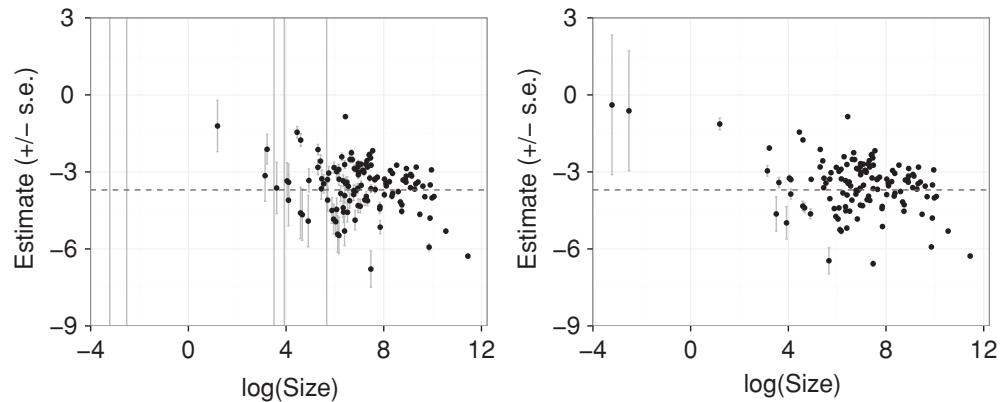


Fig. 16.2. Point estimates for occupation class specific intercepts, plus/minus one standard error. Results are from no pooling approach (left) and Poisson random intercepts model (right). The dashed line is $y = -3.702$ (i.e., the overall intercept from the complete pooling model).

do not include an intercept.³ The light grey dashed line in the Figure is $y = -3.702$, the overall intercept from the complete pooling model.

As discussed in the introduction to Chapter 8 the no pooling model results in unreasonable estimates for certain occupation classes. The output printed next compares the size, random intercept estimate, fixed intercept estimate, and corresponding standard errors for a selection of occupation classes.

```
## occupation class 122 (our numbering)
# random intercept model
num    size      lower   estimate     upper   logsize
122 33.32 -6.228015 -4.635567 -3.04312  3.506158
# no pooling model
num    size      lower   estimate     upper   logsize
122 33.32 -854.0756 -17.96726 818.1411  3.506158
# data for this class (i.e. zero claims on 33.32 payroll total)
riskclass year count      payroll
       122     1     0      3.28
       122     2     0      5.69
       122     3     0      4.51
       122     4     0      4.80
       122     5     0      9.07
       122     6     0      5.97
## occupation class (our numbering)
# random intercepts model
num    size      lower   estimate     upper   logsize
61 23.16 -3.840979 -2.955994 -2.071008  3.142427
# no pooling model
num    size      lower   estimate     upper   logsize
61 23.16 -4.144029 -3.144028 -2.144028  3.142427
# data for this class (i.e. 1 claim on 23.26 payroll total)
riskclass year count payroll
       61     1     0      3.12
       61     2     0      3.68
       61     3     0      3.76
       61     4     0      3.83
       61     5     1      4.99
       61     6     0      3.78
## occupation class (our numbering)
# random intercepts model
num    size      lower      estimate      upper      logsize
52     0.08    -3.5900335 -0.6187492    2.3525350  -2.525729
```

³ As explained in Chapter 8 on LMMs, `lme4` evaluates the variance of the distribution of $\mathbf{u}|\mathbf{y}$, conditional on the maximum likelihood estimates for unknown parameters.

```
# no pooling model
num  size      lower       estimate        upper      logsize
52    0.08 -2117.1442456 -13.7817186     2089.5808085   -2.525729
# data for this class (i.e. 0 claims on 0.08 payroll total)
riskclass year count payroll
52          4     0     0.08
```

GLMMs: Random intercepts, adaptive Gauss–Hermite quadrature with lmer.

Adaptive Gauss–Hermite quadrature (see Section 16.3.3.3) is available within the `lme4` package. We estimate the random intercepts model from (16.47) again for this technique and opt for 15 quadrature points (see `nAGQ=15`).

```
> hlm2 <- lmer(count ~ (1|riskclass)+yearcentr+offset(log(payroll)),
+                  family=poisson(), data=wcfit, nAGQ=15)
> print(hlm2)
Generalized linear mixed model fit by the adaptive Gaussian
  Hermite approximation
Formula: count ~ (1 | riskclass) + yearcentr + offset(log(payroll))
Data: wcfit
AIC  BIC logLik deviance
1771 1785 -882.4     1765
Random effects:
Groups      Name        Variance Std.Dev.
riskclass (Intercept) 0.80733  0.89851
Number of obs: 767, groups: riskclass, 130

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.561974  0.083614 -42.60  <2e-16 ***
yearcentr    0.009731  0.005156    1.89   0.0591 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Correlation of Fixed Effects:
      (Intr)
yearcentr -0.010
```

The parameter estimates and standard errors obtained with GH quadrature are very close to those obtained with Laplace approximation (see the results of `hlm1`). Changing the number of quadrature points has a very minor impact on the results.

GLMMs: Random intercepts, approximating the data with glmmmpQL. To illustrate the approach from Section 16.3.3.2 (i.e., repetitive fits of a linear mixed model to pseudo-data), the function `glmmPQL` from library (MASS) is available in R. For the Poisson random intercepts model from (16.47), convergence is reached in 10 iterations. Parameter estimates and corresponding standard errors are printed next. They are different from but close to the results obtained with `lme4`. Based on the discussion in Section 16.3.4, however, we prefer the use of adaptive Gauss–Hermite quadrature whenever possible for the model under consideration.

```
> library(MASS)
> PQL1 <- glmmPQL(count ~ yearcentr + offset(log(payroll)),
+                     random = ~ 1 | riskclass, family = poisson, data = wcFit)
iteration 1
iteration 2
iteration 3
iteration 4
iteration 5
iteration 6
iteration 7
iteration 8
iteration 9
iteration 10
> summary(PQL1)
Linear mixed-effects model fit by maximum likelihood
Data: wcFit
AIC BIC logLik
NA NA     NA

Random effects:
Formula: ~1 | riskclass
            (Intercept) Residual
StdDev:    0.9290198 1.50974

Variance function:
Structure: fixed weights
Formula: ~invwt

Fixed effects: count ~ yearcentr + offset(log(payroll))
               Value Std.Error DF   t-value p-value
(Intercept) -3.489754 0.08911097 636 -39.16189 0.0000
yearcentr    0.009496 0.00656277 636    1.44688 0.1484
```

```

Correlation:
  (Intr)
yearcentr -0.012

Standardized Within-Group Residuals:
    Min         Q1        Med         Q3        Max
-2.5749914 -0.5294022 -0.1518360  0.4497736 12.6121268

Number of Observations: 767
Number of Groups: 130

```

GLMMs: Random intercepts, a Bayesian approach. Finally, we present a Bayesian analysis of the random intercepts model in (16.47). We analyze this example using WinBUGS and its interface with R, namely the function `bugs` from library `BRugs`. On the book's website we also demonstrate the use of R package `glmmBUGS`. In WinBUGS the random intercepts model in (16.47) is coded as follows.

```

model;

for(i in 1:895)
mu[i] <- payroll[i]*exp(beta0*yearcentred[i]+b[riskclass[i]])
count[i] ~ dpois(mu[i])

#specify distribution for fixed effects
beta0 ~ dnorm(0,0.0001)
beta1 ~ dnorm(0,0.0001)
#specify distribution for random effects
for(i in 1:133)
b[i] ~ dnorm(beta0,taub)

taub ~ dgamma(0.01,0.01)
sigma2b <- 1/taub

```

where we use normal $N(0, 10^{-4})$ priors for β_0 and β_1 , and a $\Gamma(0.01, 0.01)$ prior for $(\sigma_u^2)^{-1}$. The posterior densities of these parameters are illustrated in Figure 16.3.

With respect to predictive modeling, a Bayesian approach is most useful, because it provides the full predictive distribution of variables of interest (here: Count_{i7}). We illustrate this in Figure 16.4 for a selection of risk classes. Histograms are based on

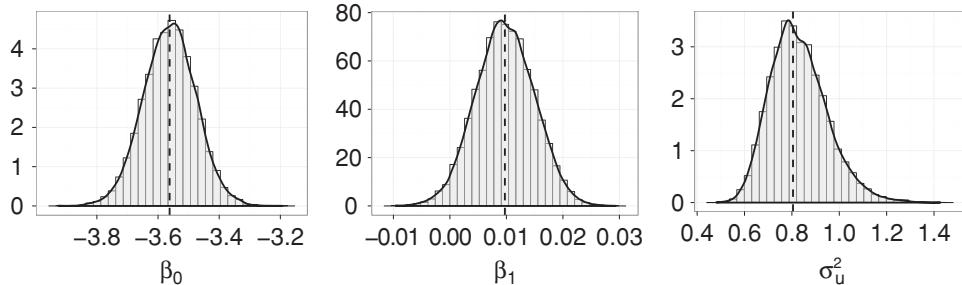


Fig. 16.3. Posterior simulations for parameters used in (16.47) (from left to right: β_0 , β_1 , and σ_u^2), workers' compensation insurance (frequencies). Results are based on two chains, 50,000 simulations each, a thinning factor of 5, and burn-in of 2,000 simulations.

50,000 simulations from the relevant predictive distribution (using model (16.47)). For each risk class the observed number of claims is indicated, as well as the point prediction obtained with a frequentist approach, using Laplace approximation from (g) `lmer`.

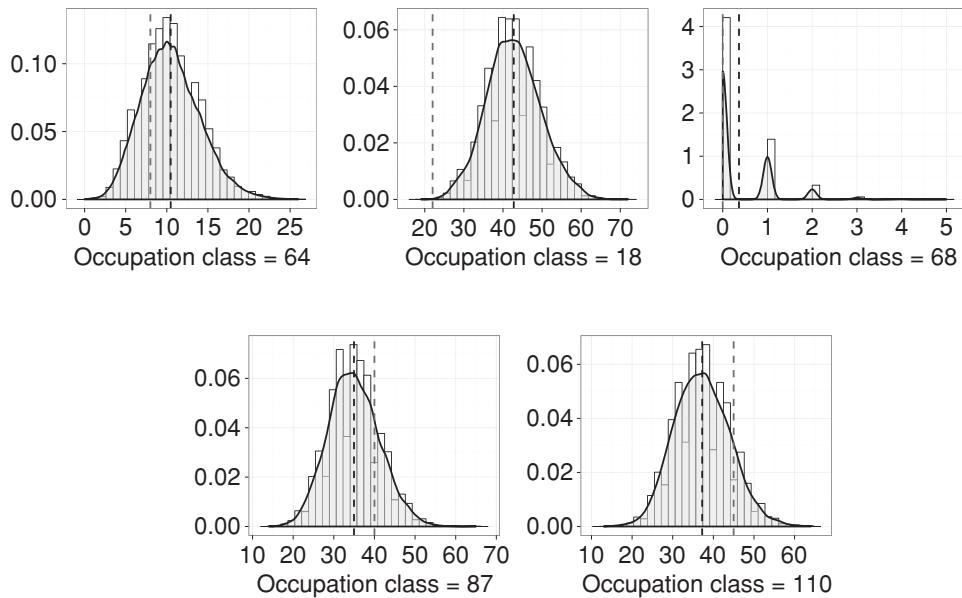


Fig. 16.4. Posterior predictive simulations for the number of claims in year 7 for a selection of risk classes. Simulations are based on Bayesian analysis of (16.47), using two chains, 50,000 simulations each, a thinning factor of 5, and burn-in of 2,000 simulations. The dashed light grey line is the point prediction as obtained with (g) `lmer`; the dashed light grey line is the observed number of claims.

References

- Abramowitz, M. and I. Stegun (1972). *Handbook of Mathematical Functions: With Formulas, Graphs and Mathematical Tables*. Dover, New York.
- Antonio, K., E. Frees, and E. Valdez (2010). A multilevel analysis of intercompany claim counts. *ASTIN Bulletin*: 40(1), 151–177.
- Antonio, K. and E. Valdez (2012). Statistical aspects of *a priori* and *a posteriori* risk classification in insurance. *Advances in Statistical Analysis* 96(2), 187–224.
- Breslow, N. and D. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 88(421), 9–25.
- Breslow, N. and X. Lin (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika* 82, 81–91.
- Denuit, M., X. Maréchal, S. Pitrebois, and J.-F. Walhin (2007). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Scales*. Wiley, New York.
- Haberman, S. and A. Renshaw (1996). Generalized linear models and actuarial science. *The Statistician* 45(4), 407–436.
- Kim, Y., Y.-K. Choi, and S. Emery (2013). Logistic regression with multiple random effects: A simulation study of estimation methods and statistical packages. *The American Statistician* 67(3), 171–182.
- Liang, K. and S. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73(1), 13–22.
- Lin, X. and N. Breslow (1996). Analysis of correlated binomial data in logistic-normal models. *Journal of Statistical Computation and Simulation* 55, 133–146.
- Liu, Q. and D. Pierce (1994). A note on Gauss-Hermite quadrature. *Biometrika* 81(3), 624–629.
- McCulloch, C. and S. Searle (2001). *Generalized, Linear and Mixed Models*. Wiley Series in Probability and Statistics, Wiley, New York.
- Molenberghs, G. and G. Verbeke (2005). *Models for Discrete Longitudinal Data*. Springer Series in Statistics, Springer, New York.
- Nelder, J. and R. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* 135, 370–384.
- Pinheiro, J. and D. Bates (2000). *Mixed Effects Models in S and S-Plus*. Springer, New York.
- Purcaru, O., M. Guillén, and M. Denuit (2004). Linear credibility models based on time series for claim counts. *Belgian Actuarial Bulletin* 4(1), 62–74.
- Tierny, L. and J. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* 81, 82–86.
- Tuerlinckx, F., F. Rijmen, G. Verbeke, and P. D. Boeck (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology* 59, 225–255.
- Wolfinger, R. and M. O'Connell (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* 48, 233–243.
- Yip, K. C. and K. K. Yau (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics* 36(2), 153–163.
- Zhao, Y., J. Staudenmayer, B. Coull, and M. Wand (2006). General design Bayesian generalized linear mixed models. *Statistical Science* 21, 35–51.

Part IV

Longitudinal Modeling

17

Time Series Analysis

Piet de Jong

Chapter Preview. This chapter deals with the analysis of measurements over time, called time series analysis. Examples of time series include inflation and unemployment indices, stock prices, currency cross rates, monthly sales, the quarterly number of claims made to an insurance company, outstanding liabilities of a company over time, internet traffic, temperature and rainfall, and the number of mortgage defaults. Time series analysis aims to explain and model the relationship between values of the time series at different points of time. Models include ARIMA, structural, and stochastic volatility models and their extensions. The first two classes of models explain the level and expected future level of a time series. The last class seeks to model the change over time in variability or volatility of a time series. Time series analysis is critical to prediction and forecasting. This chapter explains and summarizes modern time series modeling as used in insurance, actuarial studies, and related areas such as finance. Modeling is illustrated with examples, analyzed with the R statistical package.

17.1 Exploring Time Series Data

17.1.1 Time Series Data

A time series is a sequence of measurements y_1, y_2, \dots, y_n made at consecutive, usually regular, points in time. Four time series are plotted in Figure 17.1 and explained in detail later. Each time series is “continuous,” meaning each y_t can attain any value in some interval of the line. This chapter deals with continuous time series and ignores time series where, for example, y_t is non-numeric or a count variable¹.

Time series analysis seeks to measure, explain, model, and exploit the relationships between y_t and y_{t-s} for $s = \pm 1, \pm 2 \dots$; that is, the values of the time series at different

¹ The distinction between counts and continuous variables disappears if the counts are large. For example, the GNP of a country is a “count” of, say, dollars. However, for practical purposes GNP is regarded as a continuous variable.

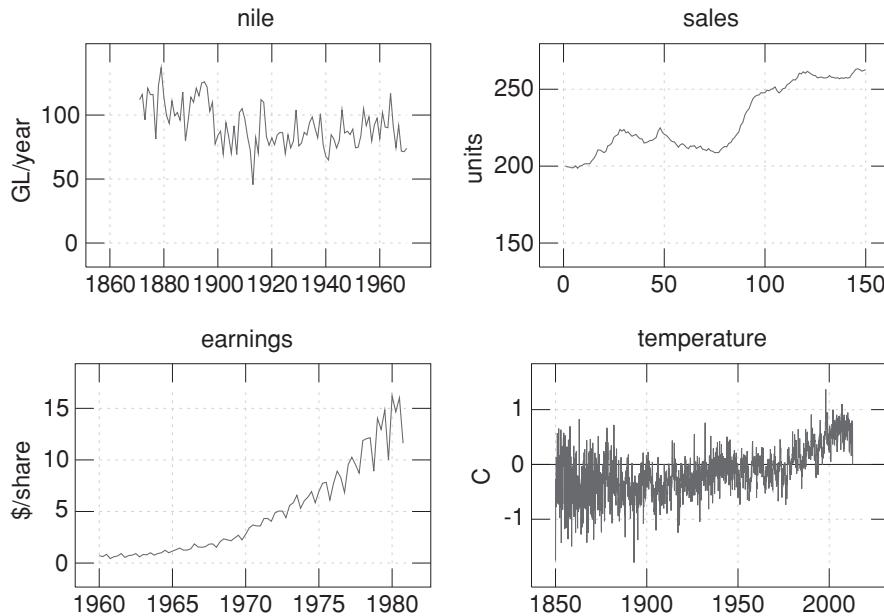


Fig. 17.1. Four time series described at the end of Section 17.1.1 and used to illustrate time series analysis in this chapter.

points of time. These relationships are of interest for one or more of the following reasons:

Forecasting. Forecasting aims to predict future outcomes or distributions. It requires knowledge and understanding of the relationship between what is happening or has happened and what will happen in the future. Forecasting requires a special case of predictive modeling as generally discussed in this book: the events to be predicted are in the future, and data used for prediction are the similar outcomes in the past or present. This is well illustrated with the temperature series in Figure 17.1 that measures global average temperature. Of interest is the likely progression of future temperature based on historical temperature. Forecasts often provide a basis for action. For example, the forecast that insurance claims will rise next quarter might necessitate an increase in insurance reserves.

Understanding. We may simply want to know the “why’s” and “wherefore’s” in the movements of a time series.

Description. Can the information in a time series be compressed into a few meaningful “summary statistics”. Any summary must hold across time relationships.

Control. Controlling future outcomes requires knowledge of the relationship between variables happening now and the outcomes of these same or other variables in the future.

Signal Extraction. Many time series are “noisy” indicators of underlying “signals.” For example, the temperature time series has movements that appear to mask an underlying trend. Signals often have some kind of continuity in time, and estimating the relationship between measurements at different points of time makes for a better recovery of a signal.

Inference. Inferring the effect of one variable on another requires partialing out the effects of other, possibly across time, effects. For example, to detect “outliers” in the Nile river flow time series, one has to understand the dynamics of Nile river flow.

Time series analysis with R. The analyses described in this chapter require detailed calculations facilitated by a computer package such as R (R Development Core Team 2010). The R package provides a large number of time series datasets, especially if the `tseries` package is loaded. Therefore, ensure the commands are accessible using the `library(tseries)`. Other R packages for time series analysis include `forecast`, `ltsa`, `FitAR`, and `FitARMA`. Many time series are available in the R datasets package. Issue the command `data()` to list the available time series, including the first three plotted in Figure 17.1 and listed next.

- `nile` – Annual flow of the river Nile 1871–1970
- `sales` – Sales data from Box–Jenkins
- `earnings` – Quarterly earning of Johnson & Johnson 1960–1980
- `temperature` – Monthly global temperature for the years 1850–2013

The final series is read into R as follows:

```
> f=read.table(url,fill=TRUE)
> ny=nrow(f)/2;x=c(t(f[2*1:ny-1,2:13]))
> temperature=ts(x,start=1850,frequency=12,end=c(2013,8))
```

where `url` is the http address in Jones (2010). The series is continually updated and extended, requiring the updating of `c(2013, 8)`.

17.1.2 Plotting and Charting Time Series

A popular and recommended initial way to study a time series is to look at its graph. This is informative, but is notoriously unreliable. Example graphs are displayed in Figure 17.1. The Nile time series displays a relatively stable mean and variation often called “stationary.” Stationarity is formally defined in the next section. The sales time series is much smoother, but appears more prone to wandering, suggesting possible “nonstationarity” in the mean. The earning time series has a strong growth pattern, seasonal variation, and increasing variation. The temperature time series is subject to much erratic fluctuation, with a decrease in variance or volatility over time and an increase in the mean or level. Each of these series is analysed in more detail later.

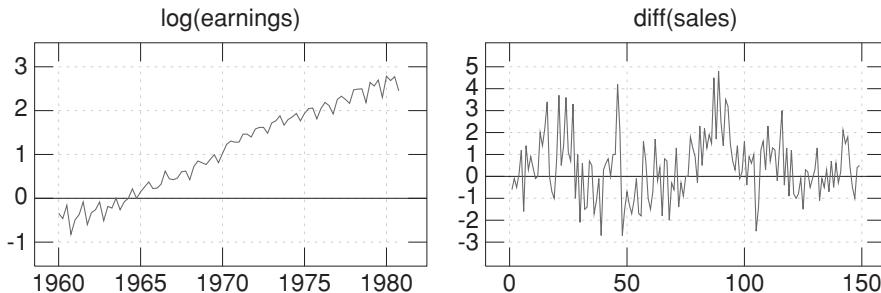


Fig. 17.2. Log of earnings and first difference of sales.

The detailed visual inspection of financial time series has led to the development of so-called charting techniques popular in stock market circles. Here, features of the graph are examined, and certain standard shapes are identified such as “resistance levels,” “cups,” and so on. On the basis of such standard shapes the likely future progression of the time series is forecast. However, blind experiments involving randomly generated series suggest that charting techniques are worthless.

17.1.3 Frequency, Scale, and Differencing

Prior to detailed analysis of a time series, decisions need be made regarding frequency, scale, and differencing. Frequency refers to how often measurements are made: hourly, daily, monthly, and so on. It has an impact on the relationships between the measurements. In many situations the frequency is given. In other situations it is convenient to choose an annual series to avoid, for example, the complicating factor of seasonal fluctuations. However, other things equal, it is better to measure a time series more frequently.

Scale decisions refer to whether the analysis proceeds in terms of the actual measurements or some transformation thereof. The linear transformations of adding or multiplying all observations by a constant do not generally affect the analyses discussed in this chapter. A common nonlinear transformation is natural logarithms.² Logarithms are appropriate if the variation in the time series is proportional to the level. Consider, for example, the earnings series in the bottom left panel of Figure 17.1. Here variability appears to increase with the mean. The logarithm of earnings is displayed in the left panel of Figure 17.2, and the log transform has “stabilized” the variance.

Differencing concerns whether the original series y_t or its (first) differences $y_t - y_{t-1}$, or higher order differences (differences of differences), or seasonal differences such as $y_t - y_{t-4}$, should be analyzed. For example, with economic time series it is often useful to compare each quarter with the same quarter in the previous year. The first difference of the sales data is plotted in the right panel of Figure 17.2; its relatively constant mean and variance suggest the difference in sales is reasonably “stationary.” Changing scale, differencing, and other transformations are often called filtering.

² All logarithms discussed and used in this chapter are natural logarithms.

The first difference of the natural logarithm of a time series has a special significance:

$$\ln(y_t) - \ln(y_{t-1}) = \ln\left(\frac{y_t}{y_{t-1}}\right) = \ln\left(1 + \frac{y_t - y_{t-1}}{y_{t-1}}\right) \approx \frac{y_t - y_{t-1}}{y_{t-1}}, \quad (17.1)$$

where the approximation is accurate if the difference $y_t - y_{t-1}$ is small compared to y_{t-1} . Hence analyzing differences in logs is tantamount to analyzing approximate percentage changes.

17.2 Modeling Foundations

17.2.1 Stationarity and the Autocorrelation Function

Many time series have no trend in either in the mean or variance and are often termed “stationary.” The only time series in Figure 17.1 that appears stationary is the Nile river flow in the top left panel. The three others have means that appear to increase with time and, in the case of the earnings series, increasing variance. Nonstationary time series can often be made stationary with an appropriate transformation.

Formally, a time series is stationary if its mean, variance, and covariances are finite and independent of time. The observed or measured time series $y_t, t = 1, 2, \dots, n$ is here regarded as the realization of a sequence of random variables. The means and covariances here are unconditional – that is, without conditioning on any previous or other values. This definition is often called “covariance,” “second order,” “weak,” or “wide sense” stationarity. “Strict,” “strong,” or “distributional” stationarity is where, for every s , the (unconditional) probability distribution of $y_t, y_{t-1}, \dots, y_{t-s}$ is the same for every t .

Given a (covariance) stationary time series y_t , the mean and covariances are denoted

$$\mu \equiv E(y_t), \quad \rho_s \equiv E\{(y_t - \mu)(y_{t-s} - \mu)\}, \quad s = 0, \pm 1, \pm 2, \dots$$

Note that $\rho_s = \rho_{-s}$ and $\rho_0 \geq 0$ is the constant variance of the time series. The sequence ρ_s as a function of s is called the auto or lag covariance function. Lag covariances ρ_s satisfy other properties such as $|\rho_s| < \rho_0$. For convenience it is often assumed that y_t denotes the value of the time series after subtracting its mean, implying that the mean of the adjusted series is $\mu = 0$ and $\rho_s = E(y_t y_{t-s})$.

The covariance ρ_s divided by the product of the standard deviations of y_t and y_{t-s} yields a correlation, called the lag correlation. By stationarity, the two standard deviations both equal $\sqrt{\rho_0}$, and hence the lag correlation lag s is ρ_s/ρ_0 . Lag correlation as a function of the s is called the auto or lag correlation function (ACF). The adjective “auto” emphasizes that it is the correlation of y_t with itself at another point in time.

Autocovariance ρ_s is estimated as

$$\hat{\rho}_s \equiv \frac{1}{n} \sum_{t=s+1}^n (y_t - \bar{y})(y_{t-s} - \bar{y}), \quad s = 0, 1, \dots, n-1. \quad (17.2)$$

Here \bar{y} is the sample mean of the time series. The divisor in $\hat{\rho}_s$ is n , even though the sum contains only $n - s \leq n$ terms.

The estimated ACF at lag s is $r_s \equiv \hat{\rho}_s / \hat{\rho}_0$. Computing r_s is equivalent to estimating the autocovariance function of the time series $(y_t - \bar{y})/\sqrt{\hat{\rho}_0}$, called the standardized time series. The standardized time series has sample mean 0 and variance 1.

Computing the ACF with R. Use `acf(y)` where y is the time series. Optional parameters include the maximum lag and whether the auto covariance or autocorrelation function is computed. Estimated ACFs corresponding to the time series in Figures 17.1 and 17.2 are displayed in Figure 17.3.

Interpreting the ACF requires care because ρ_s is the unconditional covariance between y_t and y_{t-s} . Other variables, including the intermediate variables $y_{t-1}, \dots, y_{t-s+1}$, are not held “constant.” A large value for ρ_s may be a consequence of relationships of y_t and y_{t-s} to these other variables, as is further discussed later.

The ACFs in Figure 17.3 are those for each series in Figure 17.1 as well as their first differences $y_t - y_{t-1}$. Also plotted are the horizontal $\pm 2/\sqrt{n}$ confidence bounds discussed later. The nile series displays mild autocorrelation, whereas the sales series has more persistent autocorrelation even in the first differences. The logarithm of earnings indicates strong autocorrelation and periodicity in the autocorrelation of the first differences, with the ACF peaking at lags 4, 8, … corresponding to an annual cycle. The temperature series also has strong autocorrelation, which largely disappears once differences are taken.

17.2.2 Noise and Random Walks

A time series y_t is noise (or white noise) if it is stationary and $\rho_s = 0$ for $s \neq 0$. In other words there is no covariance or correlation between values of the time series at different points in time. A noise time series is often denoted $\varepsilon_t \sim (0, \sigma^2)$, indicating noise with mean 0 and variance σ^2 .

To judge if a time series is noise, the estimated ACF is examined, and a judgment is made whether deviations of the ACF from 0 at $s \neq 0$ are reasonably explained in terms of sampling fluctuation. This judgment is made using the sampling distribution of the estimated ACF. With noise, the sampling distribution of ACF for $s \neq 0$ is, for large n , approximately normal with mean 0 and variance $1/n$. Hence for noise, the estimated ACF at each s is expected to be within about $\pm 2/\sqrt{n}$ of zero for 95% of samples. Values of the ACF outside these bounds suggest the time series is not noise. The 95% confidence limits are plotted in each of the panels in Figure 17.3. Note that all the autocorrelations for small lags corresponding to each of the series are “significant.” However, the significance of many of the autocorrelations disappears once differences are taken.

A variant of noise is the random walk, which is also called integrated noise. Time series y_t is a random walk if the first difference series $y_t - y_{t-1}$ is noise or $y_t = y_{t-1} + \varepsilon_t$. A random walk is nonstationary because, using repeated substitution, $y_t = y_0 + \varepsilon_1 + \dots + \varepsilon_t$. Hence the variance of y_t is $\text{var}(y_0) + t\sigma^2$, diverges as it becomes large, provided $\sigma^2 \neq 0$. A random

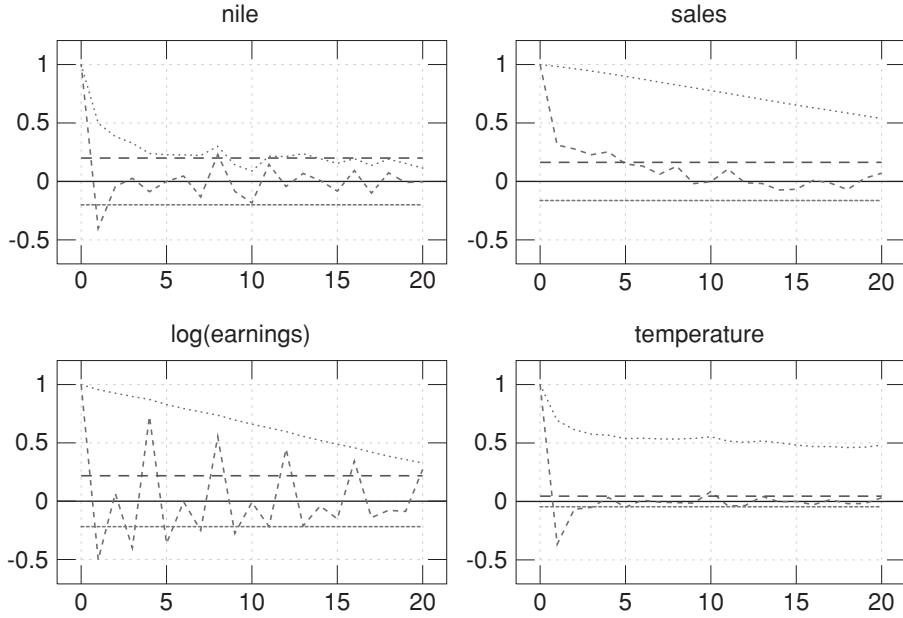


Fig. 17.3. Estimated ACFs of the series in Figure 17.1 as well as their first differences. For earnings, the estimates are for the log of the series. Lines in each panel approaching zero more rapidly correspond to the first differences. Horizontal lines in each graph are the $\pm 2/\sqrt{n}$ approximate 95% limits for testing zero autocorrelation. In R use `plot(acf(y))` or `plot(acf(diff(y)))` where y is the series.

walk with drift μ is where $y_t = y_{t-1} + \mu + \varepsilon_t$. To test for a random walk, take the first differences $y_t - y_{t-1}$ and see if the resulting time series is noise. From Figure 17.3 the nile series does not appear to be a random walk because the ACF of the first differences at lag 1 lies outside the 95% limits. Similarly, none of the other three series appear to be random walks.

Overdifferencing. If y_t is noise ε_t then the difference $y_t - y_{t-1} = \varepsilon_t - \varepsilon_{t-1}$ has an ACF that is zero everywhere, except at lag $s = 1$ where it is $-1/2$. Thus differencing can induce autocorrelation where previously there was none, a situation often called “overdifferencing.” Thus negative autocorrelation at lag 1 in the first difference of a series may be a consequence of differencing, rather than of the model. For example, consider the ACFs of the first differences in Figure 17.3. The nile, (log) earnings, and temperature series have negative autocorrelation in the first differences that is partially explainable by differencing.

Geometric random walk (GRW). Suppose the percentage changes in the right hand side of (17.1) have mean μ and variance σ^2 . Taking the exponential of (17.1) and rearranging yields $y_t = y_{t-1}e^{\mu+\varepsilon_t}$ with $\varepsilon_t \sim (0, \sigma^2)$. In this equation exact equality has replaced approximate equality, appropriate if μ and σ are small, which is in turn enforced if the frequency of observation is large. Thus the log of a GRW is a random walk with drift μ .

17.3 Autoregressive, Moving Average (ARMA) Models

17.3.1 ARMA Models

Practically all stationary time series can be thought of as being generated by an ARMA model³:

$$y_t = a_1 y_{t-1} + \cdots + a_p y_{t-p} + \varepsilon_t + b_1 \varepsilon_{t-1} + \cdots + b_q \varepsilon_{t-q}, \quad (17.3)$$

where $\varepsilon_t \sim (0, \sigma^2)$ is unobserved noise and $a_p \neq 0, b_q \neq 0$. Equation (17.3) is assumed to hold for all t and defines the ARMA(p, q). Here p is the number of autoregressive (AR) components or lags in y_t , and q is the order of the moving average (MA) component: lags in ε_t . The ARMA($p, 0$) and ARMA($0, q$) are denoted and called the AR(p) and MA(q), respectively.

Mean or intercept effects. In (17.3) it is typically assumed that $E(y_t) \equiv \mu = 0$. If $\mu \neq 0$ then each of the y_t in (17.3) is replaced by $y_t - \mu$. Rearranging the resulting equation shows that mean effects are incorporated by including the intercept $(1 - a_1 - \cdots - a_p)\mu$ in the right-hand side of (17.3). Since stationary time series analysis aims to model autocorrelation, mean effects are usually not of interest and are dealt with by subtracting out the sample mean of the series before doing any analysis and assuming $\mu = 0$.

AR(1). Here $y_t = a y_{t-1} + \varepsilon_t$. Thus y_t is a fraction of the previous value plus the noise. Since the equation holds for all t , y_{t-1} can be substituted out of the right-hand side of the equation to yield $y_t = a^2 y_{t-2} + \varepsilon_t + a \varepsilon_{t-1}$. In turn, successively substituting out y_{t-2}, y_{t-3}, \dots yields

$$y_t = \varepsilon_t + a \varepsilon_{t-1} + a^2 \varepsilon_{t-2} + \cdots, \quad (17.4)$$

which is an MA(∞). The moving average (17.4) only makes sense if $|a| < 1$; otherwise the coefficients in (17.4) blow up. The condition $|a| < 1$ ensures the variance of y_t is finite and is called the *stationarity* condition for an AR(1). Stationarity conditions are further discussed later. Note the ACF of an AR(1) is never zero because y_t and y_{t-s} have, according to (17.4) and for all s , common noise terms and hence are correlated.

MA(1). Here $y_t = \varepsilon_t + b \varepsilon_{t-1}$ is again assumed to hold for all t , and hence $\varepsilon_{t-1} = y_{t-1} - b y_{t-2}$ with similar expressions for $\varepsilon_{t-2}, \varepsilon_{t-3}$, and so on. Successively substituting these expressions for past errors leads to

$$y_t = \varepsilon_t + b y_{t-1} + b^2 y_{t-2} + \cdots,$$

which is an AR(∞). This representation only makes sense if $|b| < 1$, which is called the *invertibility* condition for an MA(1). Again ε_t represents the “new” part of the series at time t . Note an MA(1) is always stationary, but need not be invertible. The ACF of an MA(1) is zero for lags $s > 1$ because for such lags y_t and y_{t-s} are based on distinct noise terms.

MA(q). Here $y_t = \varepsilon_t + \cdots + b_q \varepsilon_{t-q}$ has a lag covariance function ρ_s that is zero for $s > q$ because y_t and y_{t-s} for $s > q$ involve different noise terms that are uncorrelated. Further if $s < q$, then $\rho_s = \sigma^2 \sum_j b_j b_{j-s}$. This shows that every MA(q) is stationary provided $q < \infty$. An AR(p) is equivalent to an MA(∞), and hence the ACF does not cut out. More generally the ACF of ARMA(p, q) does not cut out if $p > 0$.

³ Models other than ARMA models can also produce stationary time series.

ARMA(1,1). This model combines an AR(1) and MA(1):

$$y_t = ay_{t-1} + \varepsilon_t + b\varepsilon_{t-1}, \quad (17.5)$$

assumed to hold for all t . If $v_t = \varepsilon_t + b\varepsilon_{t-1}$, then as with the AR(1)

$$y_t = v_t + av_{t-1} + a^2v_{t-2} + \cdots = \varepsilon_t + (a+b)\varepsilon_{t-1} + (a^2+ab)\varepsilon_{t-2} + \cdots.$$

Thus as with the AR(1), if $|a| < 1$ the model can be sensibly written as an MA(∞) and $|a| < 1$ is again called the “stationarity” condition. Alternatively, using a similar argument, expression (17.5) can be written as an AR(∞):

$$y_t = (a+b)y_{t-1} - b(a+b)y_{t-2} + b^2(a+b)y_{t-3} + \cdots + \varepsilon_t.$$

This representation makes sense if $|b| < 1$, which is called the invertibility condition for an ARMA(1,1).

Any ARMA(p, q) can be *back substituted* to arrive at an AR(∞) or MA(∞). Back substitution becomes increasingly tedious as p and q increase. Fortunately there is a streamlined notation and treatment using lag operators, discussed in Section 17.3.3.

17.3.2 Fitting ARMA Models

Given p and q and an observed time series y_t , the coefficients of an ARMA(p, q) are estimated by maximizing the likelihood, assuming the ε_t are normally distributed. This is closely related to least squares: minimizing with respect to the AR and MA parameters the sum of squares $\sum_t e_t^2$ where the e_t are the computed residuals calculated recursively as⁴

$$e_t = y_t - a_1 y_{t-1} - \cdots - a_p y_{t-p} - b_1 \varepsilon_{t-1} - \cdots - b_q \varepsilon_{t-q}. \quad (17.6)$$

Numerical procedures are used to find the minimum of $\sum_t e_t^2$, which is the main significant component in the normal-based likelihood calculation. The recursion (17.6) is “seeded” with assumed initial values y_0, \dots, y_{1-p} and $\varepsilon_0, \dots, \varepsilon_{1-q}$. For long time series the assumed seed values do not have a practical impact on the estimates. For shorter time series the seeds may be material, requiring a more careful setting of the seeds, a technique sometimes called *back casting*.

Fitting ARMA models with R. Back casting and the numerical search for the optimum values of the AR and MA coefficients are implemented in R with the command `arima(y, c(p, d, q))`. For example, `arima(nile, c(1, 0, 1))` fits an ARMA(1,1) model to the nile data mentioned earlier. In the `arima(y, c(p, d, q))` command `y` specifies the time series and `p` and `q` specify p and q , respectively. The `d` value is the degree of differencing to be performed on the series prior to fitting. This is further discussed later.

⁴ Note the right-hand side equals ε_t if the AR and MA coefficients are the true coefficients. The notation e_t is used here to indicate that the AR and MA coefficients are estimates.

```

Coefficients:
      ar1     ma1  intercept
      0.8611 -0.5177   920.5567
  s.e.  0.1067   0.1908   46.6736

sigma^2 estimated as 19892:  log likelihood = -637.04

```

Fig. 17.4. Output from R command `arima(nile,c(1,0,1))`.

Figure 17.4 displays R output when an ARIMA(1,0,1) is fit to the nile time series. With R, the estimate of the mean is called the intercept, and hence the fitted model is

$$(y_t - 920.56) = 0.86(y_{t-1} - 920.56) + \varepsilon_t - 0.52\varepsilon_{t-1}, \quad \varepsilon_t \sim (0, 141^2). \quad (17.7)$$

The standard errors on the AR and MA coefficients suggest significance: both estimates are more than 2 standard deviations away from 0. These standard errors are derived on the basis of large sample arguments linking standard errors to the shape of the log-likelihood near the maximum. The estimate $\hat{\sigma} = 141$ compares to the standard deviation of the series of about 168. Hence the ARMA(1,1) model explains about $1 - (141/168)^2 = 0.30$ or 30% of the variability in the series.

17.3.3 Lag Operators

The ARMA(p, q) models illustrate that each model can be rewritten into equivalent forms. Moving from one model to another is achieved by back substitution. With higher order ARMA(p, q), back substitution is awkward and tedious. However, matters are simplified and streamlined with the use of the *lag operator* notation:

$$a(L)y_t = b(L)\varepsilon_t, \quad (17.8)$$

$$a(L) \equiv 1 - a_1L - \cdots - a_pL^p, \quad b(L) \equiv 1 + b_1L + \cdots + b_qL^q.$$

Equation (17.8) is shorthand for (17.3), with all y_t terms taken to the left-hand side and L , the lag operator $L^k y_t = y_{t-k}$, serving to lag the time index on the y_t or ε_t terms. The ARMA(1,1) has $a(L) = 1 - aL$ and $b(L) = 1 + bL$, whereas the AR(1) has $a(L) = 1 - aL$ and $b(L) = 1$, and for the MA(1) $a(L) = 1$ and $b(L) = 1 + bL$.

In terms of the lag operator notation, the AR and MA representations of an ARMA(p, q) are, symbolically,

$$y_t = \frac{b(L)}{a(L)} \varepsilon_t, \quad \frac{a(L)}{b(L)} y_t = \varepsilon_t, \quad (17.9)$$

respectively. As suggested by the AR(1), MA(1), and ARMA(1,1), the representations in (17.9) are valid only for certain values of the AR or MA coefficients. These conditions are explored in the next section.

17.3.4 Stationarity and Invertibility of an ARMA Model

Given the model (17.8), suppose $a(L)$ considered as a polynomial in L is factored:

$$a(L) = (1 - z_1 L) \cdots (1 - z_p L), \quad (1 - z_i L)^{-1} = 1 + z_i L + (z_i L)^2 + \cdots,$$

where the $1/z_i$ are the roots (real or complex) of $a(L)$. Both of these equations follow from algebra. The expansion on the right is cogent if $|z_i| < 1$ or, in terms of the root, $|1/z_i| > 1$. Further, $1/a(L)$ makes sense as an expansion in powers of L if all the factors can be expanded; that is, if all the roots of $a(L)$ lie outside the unit circle. This guarantees that the ARMA can be written as a (convergent) MA. Since MAs with convergent coefficients are stationary, all roots of $a(L)$ outside the unit circle are called the stationarity condition for an ARMA(p, q).

The same argument applied to $b(L)$ yields the invertibility condition for an ARMA(p, q): an ARMA(p, q) can be written as an AR(∞) if all the roots of $b(L)$ lie outside the unit circle. Invertibility has no bearing on the issue of stationarity.

The random walk is an ARMA(1,0) model with $a(L) = 1 - L$ that has a single root 1, on the unit circle. Hence a random walk is not stationary. The model $y_t = \varepsilon_t + \varepsilon_{t-1}$ is an ARMA(0,1) with $b(L) = 1 + L$. The model is stationary but not invertible because the root of $1 + L$ is -1 . Hence the latter model cannot be written as a pure AR. Nevertheless it is obviously stationary. To see the issue, note $\varepsilon_t = y_t - \varepsilon_{t-1}$. Lagging once and substituting into the equation for y_t yields $y_t = \varepsilon_t + y_{t-1} - \varepsilon_{t-2}$. Repeating this argument by substituting out ε_{t-2} and so on shows

$$y_t = \varepsilon_t + y_{t-1} - y_{t-2} + y_{t-3} - \cdots,$$

where the coefficients attaching to y values in the remote past are ± 1 . Obviously this is not a satisfactory representation: the model is noninvertible. Practically speaking, a noninvertible model implies that ε_t cannot be written as the difference between y_t and some linear combination of previous y_t values.

17.3.5 Autocorrelation Properties of Stationary ARMA Models

Given a stationary time series, what is an appropriate ARMA(p, q) model? This question is usually answered with the aid of the ACF and the partial autocorrelation function (PACF) explained later.

For an MA(q), the ACF at lags $s = 0, 1, \dots, q$ are nonzero because at such lags, y_t and y_{t-s} contain common noise terms and hence are correlated. For $s > q$ the ACF is zero because y_t and y_{t-s} contain no common noise terms. Hence, as displayed in Table 17.1, the ACF of an MA(q) is zero for lags $s > q$; that is, it cuts out. Since a stationary AR(p) is equivalent to an MA(∞), the ACF of an AR(p) does not cut out: it is said to “die out” – that is, it approaches zero, but does not eventually become and stay equal to zero. The same holds for a stationary ARMA(p, q) because this model is equivalent to an MA(∞). These ACF properties are displayed in the second column of Table 17.1.

The third column in Table 17.1 deals with the partial autocorrelation function (PACF). The partial autocorrelation between y_t and y_{t-s} , denoted π_s , is the correlation between these

Table 17.1. Auto and Partial Autocorrelations

Model	ACF	PACF
MA(q)	cuts out: zero for $s > q$	dies out
AR(p)	dies out	cuts out: zero for $s > p$
ARMA(p, q)	dies out	dies out

two variables after removing any effects of the intermediate variables $y_{t-1}, \dots, y_{t-s+1}$. If $s = 1$ there are no intermediate variables, and hence $\pi_1 = \rho_1/\rho_0$. At lag $s = 2$ the partial autocorrelation removes the effect of the single intermediate variable y_{t-1} . If the time series is an AR(1) $y_t = ay_{t-1} + \varepsilon_t$, then removing the effect of y_{t-1} from y_t leaves $y_t - ay_{t-1} = \varepsilon_t$, and the correlation between the latter and $y_{t-2} - cy_{t-1}$ is 0 because ε_t is noise, unrelated to the past. Hence the PACF for an AR(1) is zero at $s = 2$. Similarly at lag $s > 2$ an AR(1) has PACF equal to zero because removing the intermediate variables is equivalent to removing just y_{t-1} , which leaves just ε_t , as $s = 2$. Hence the PACF for an AR(1) cuts out: all values are zero for $s = 2, 3, \dots$.

The PACF for an AR(p) cuts out for $s > p$. This holds because removing from y_t the effects of the intermediate variables $y_{t-1}, \dots, y_{t-s+1}$ for $s > p$ leaves ε_t , which is uncorrelated to any previous values y_{t-1}, y_{t-2}, \dots . Since invertible MA(q) and ARMA(p, q) time series are equivalent to AR(∞), it follows that the PACF for models with MA components never cuts out (i.e., dies out). These results are displayed in the third column of Table 17.1.

The results in Table 17.1 are used to identify appropriate values for p and q given an ACF and PACF function. Thus cutting-out behavior in either the ACF or PACF indicates a pure AR or MA model, with p and q determined by the cutting-out point. If neither cuts out, an ARMA(p, q) is indicated with neither p nor q equal to zero. In all cases p and q are assumed small if not zero.

In practice, estimates of the ACF and PACF are available. Estimated PACFs of the series in Figure 17.1 are displayed in Figure 17.5. The PACF lag s is estimated, in effect, by correlating the residuals from the regressions of y_t and y_{t-s} on the intermediate values $y_{t-1}, \dots, y_{t-s+1}$. These regressions and subsequent correlations are recursively computed in R with the `pacf(y)` command.

For example, with the nile time series, the PACF for the undifferenced series appears to cut out quite rapidly, whereas that of the differenced series has a negative early spike. Similar to the ACF, the approximate 95% confidence limits for testing whether the PACF is zero are $\pm 2/\sqrt{n}$; these horizontal lines are indicated in each of the panels of Figure 17.5. The PACFs for sales and log earnings cut out very early on, whereas for temperature there is no early cutting out. The PACFs of the differences of log earnings and temperature have negative spikes, suggesting possible overdifferencing.

The $\pm 2/\sqrt{n}$ limits on the ACF and PACF are applicable for large n and assume that all population correlations for lags $s > 0$ are zero: that is the series is noise. If the time series is not noise – then the confidence interval is adjusted to $\pm \sqrt{c/n}$ where, for the case of an AR(p),

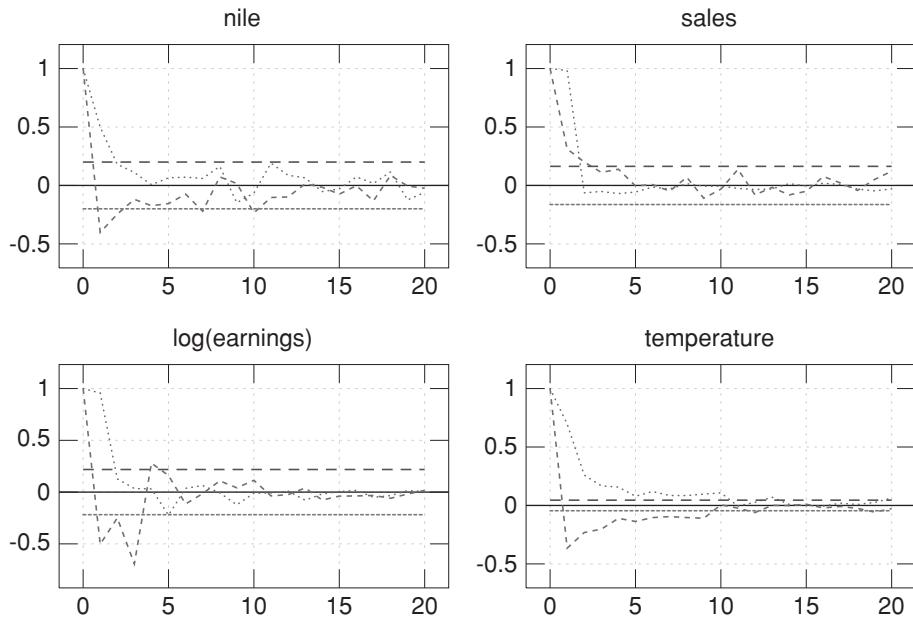


Fig. 17.5. Estimated PACFs of the four series in Figure 17.1 for the original series and first differences. For each series the lines approaching zero more rapidly correspond to the first differences. Horizontal lines in each graph are the $\pm 2/\sqrt{n}$ approximate 95% limits for testing zero correlation. In R use `plot(pacf(y))` or `plot(pacf(diff(y)))` where y is the series.

in the case of the ACF, to $\pm 2\sqrt{c/n}$ where

$$c = 1 + \sum_{s=1}^{k-1} r_s^2. \quad (17.10)$$

where r_s is the sample autocorrelation lag s . An analogous adjustment pertains to constructing a confidence interval for the PACF, with the estimated autocorrelations r_s in (17.10) replaced by the estimated partial autocorrelations $\hat{\pi}_s$.

Testing the ACF or PACF at each lag s , in isolation of what happens at other lags, suffers from *multiple test* bias. That is, there are bound to be rejections if a large number of tests are performed. For the case of the ACF a simultaneous test of the hypotheses $\rho_1 = \dots = \rho_k = 0$ is achieved with

$$n(n+2) \sum_{s=1}^k \frac{r_s^2}{n-s}, \quad (17.11)$$

called the Ljung–Box test statistic. If (17.11) is large, compared to the chi-square with k degrees of freedom, then the hypothesis that there is zero autocorrelation at lags $s = 1, \dots, k$ is rejected. Use `Box.test(y, k, "Ljung-Box")` in R to compute (17.11) for time series y .

17.3.6 ARIMA Modeling

ARMA models aim to model stationary time series. ARIMA models (note the “I”) aim to model nonstationary time series. Nonstationarity is dealt with by assuming that first or higher order differencing will remove any nonstationarity (i.e., yield a stationary time series). A time series whose d th difference follows a stationary ARMA(p, q) is called an ARIMA(p, d, q) time series; here d is the lowest order of differencing that makes the series stationary. The “I” in ARIMA refers to “integrated” because any model in terms of differences can be written as a sum or “integral” of stationary terms.

In terms of the lag operator notation, the ARIMA(p, d, q) is

$$a(L)(1 - L)^d(y_t - \mu) = b(L)\varepsilon_t, \quad \varepsilon_t \sim (0, \sigma^2),$$

where $a(L)(1 - L)^d$ has d roots on the unit circle and all remaining roots – that is, the roots of $a(L)$ – outside the unit circle.

ARIMA models are also called Box–Jenkins models. With these models, nonstationarity, if present, is assumed to be induced by roots on the unit circle, called unit roots. Non-unit roots inside the unit circle are always excluded on practical grounds. The argument is that roots inside the unit circle induce explosive behavior and few time series display such behavior.

The appropriate degree of differencing d is judged from the plot of the series and the behavior of the estimated ACF and PACF. Evidence of unit roots is provided when both the ACF and PACF converge very slowly to zero. This suggests $d \geq 1$. If the first difference has an ACF and PACF that also converge very slowly, then perhaps $d \geq 2$, and so on. Practically speaking d is rarely greater than 2. Once the appropriate degree of differencing is identified, the stationary model for the differences is estimated as in the previous section.

The Box–Jenkins methodology proceeds according to the following steps:

- (1) **Check for and remove nonstationarity.** Methods of detecting nonstationarity include plotting the data against time and examining the estimated ACF and PACF. Loosely speaking, nonstationarity can also refer to outliers and so on. Nonstationarity in the mean is generally removed by differencing. Nonstationarity in variance may be removed by taking, for example, logarithms. Other ways of removing trends include regression against time or the use of dummy variable regression to take out outliers.
- (2) **Identify the AR and MA orders p and q .** Typically there are two or three orders at the most. Overfitting, and eliminating nonsignificant AR or MA coefficients, is also used to determine the appropriate p and q .
- (3) **Estimate the AR and MA coefficients.** This step uses the techniques discussed in Section 17.3.2.
- (4) **Check the fitted model.** Diagnostic checking focuses on the fitted residual $y_t - \hat{y}_t$. With a proper model – that is, a model that correctly captures across time dependence – the residuals behave similarly to noise. The ACF and PACF of the residuals are thus examined directly or with (17.11) to confirm noise. Also the sizes of the standard errors in relation to the estimates are examined to see if any estimated coefficients can be set to zero.

Table 17.2. ARIMA Fits for Sales Data

	ARIMA(1, 1, 0)	ARIMA(1, 1, 1)	ARIMA(2, 1, 0)	ARIMA(2, 1, 1)				
coef	0.31	–	0.84	–0.61	0.25	0.20	0.77	0.04
se	0.08	–	0.08	0.12	0.08	0.08	0.20	0.12
$\hat{\sigma}, \ell$	1.37	–258	1.32	–253	1.34	–255	1.32	–253

To illustrate, consider the sales time series plotted in the top right panel of Figure 17.1 and in the right panel of Figure 17.2. The ACF and PACF of the series and its first differences are plotted in Figure 17.3, suggesting the series is nonstationary, but the first difference may be regarded as stationary. Results from fitting ARIMA($p, 1, q$) models are displayed in Table 17.2. For each fit the AR and MA coefficients are presented (in order), with standard errors reported below each coefficient estimate. The final row reports the estimated σ and the log-likelihood value for each fitted model.

The highest log-likelihood is achieved with the (1,1,1) and (2,1,1) models. With the latter, the second AR coefficient is not significant, suggesting that the more parsimonious ARIMA(1,1,1) is adequate. Here overfitting is used to show that a higher order model does not provide statistically significant additional explanation.

The value of the statistic (17.11) computed from the ACF of the residuals of the ARIMA(1,1,1) model with $k = 10$ is 8.20. Using a chi-squared distribution with $k - 2$ parameters (2 degrees of freedom are subtracted because two ARMA parameters are estimated) the value 8.20 indicates no statistical significant autocorrelation in the residuals. Hence the ARIMA(1,1,1) model appears adequate in explaining the across time dependence.

The estimated mean of the first differences $y_t - y_{t-1}$ is 0.04. The preferred model is thus

$$(y_t - y_{t-1} - 0.04) = 0.84(y_{t-1} - y_{t-2} - 0.04) + \varepsilon_t - 0.61\varepsilon_{t-1}.$$

This equation is rearranged to yield an ARMA(2,1) equation appropriate for forecasting:

$$y_t = 0.006 + 1.84y_{t-1} - 0.84y_{t-2} + \varepsilon_t - 0.61\varepsilon_{t-1}, \quad \varepsilon \sim (0, 1.32^2).$$

This model is nonstationary: $1 - 1.84L + 0.84L^2 = (1 - L)(1 - 0.84L)$ has a unit root.

17.3.7 Further ARMA Modeling Issues and Properties

Model multiplicity and parsimony. Every ARIMA model can be rewritten in an infinite number of ways. This is well illustrated with an AR(1), $y_t = ay_t + \varepsilon_t$. A single back substitution yields $y_t = a^2y_{t-2} + \varepsilon_t + a\varepsilon_{t-1}$. This is an ARMA(2,1) with constraints on the coefficients. A further substitution produces an ARMA(3,2) model, and so on. Similarly an MA(1) can be written as $y_t = by_{t-1} + \varepsilon_t - b^2\varepsilon_{t-2}$, an ARMA(1,2). Thus simple low-order models can be rewritten as more complicated higher order models with more, albeit constrained, parameters. This leads to the principle that models should be kept simple – parsimonious in the number of parameters.

Common factors. Model multiplicity also arises in the context of common factors. Suppose $a(L)y_t = b(L)\varepsilon_t$. Multiplying both sides by, say, $1 + cL$ leads to a different, though equally valid representation. Obviously the resulting model is not parsimonious. In the model $(1 + cL)a(L)y_t = (1 + cL)b(L)\varepsilon_t$, the term $1 + cL$ is a common factor that must be canceled out.

One step ahead prediction. Given the ARMA(p, q) model (17.3), the one step ahead forecast of y_{n+1} given y_1, \dots, y_n is

$$\hat{y}_{n+1} = a_1 y_n + \dots + a_p y_{n+1-p} + b_1 \varepsilon_n + \dots + b_q \varepsilon_{n+1-q}. \quad (17.12)$$

This differs from (17.3) with $t = n$ only in that ε_{n+1} is missing from the right-hand side. This omission is justified given that ε_{n+1} is the innovation or new element entering at $t = n + 1$ and nothing is known about it other than it has mean zero and variance σ^2 . The values $\varepsilon_n, \dots, \varepsilon_{n+1-q}$ in the right-hand side are estimated with $y_t - \hat{y}_t$ where \hat{y}_t is the one step ahead predictor of y_t given its past. In this context, the variance of $y_{n+1} - \hat{y}_{n+1}$ is called the mean square error of one step ahead prediction. Once y_{n+1} is available, the next ε_{n+1} is calculated. Thus one step ahead prediction is recursive: as each new observation becomes available, the error of prediction is calculated, which in turn is used in the one step prediction of subsequent observations.

MMSE prediction. One step ahead prediction discussed earlier is an example of minimum mean square error (MMSE) prediction, in that the predictor minimizes the average squared error: $E\{(y_t - \hat{y}_t)^2\}$. If the noise terms ε_t are normal, then \hat{y}_t and the associated mean square error are the conditional mean and variance y_t given the past data. MMSE prediction is, however, subject to criticism. First the MSE criterion penalizes over- and underprediction equally. Further, squaring the error implies that the predictor is focused on avoiding large errors. Despite these criticisms MMSE prediction is a favorite because it is relatively simple.

More than one-step-ahead prediction. The prediction of y_{n+m} given the observations up to time $t = n$ is also computed recursively. Here $m > 1$ is called the “lead time” of the prediction and n is fixed. Consider (17.3) with $t = n + m$. In the right of this equation any y_t not observed is replaced by its prediction given y_1, \dots, y_n ; any “future” ε_t – that is, for $t = n + 1, n + 2, \dots$ – is replaced by zero and the “present” and “past” errors $\varepsilon_t \equiv y_t - \hat{y}_t$ for $t = n, n - 1, \dots$: the one step ahead prediction errors. The mean square error of these m step ahead predictions is found by substituting out successively $y_{n+m-1}, \dots, y_{n+1}$, yielding y_{n+m} as a linear combination of $\varepsilon_{n+m}, \dots, \varepsilon_{n+1}$ plus a remainder term in $\varepsilon_n, \varepsilon_{n-1}, \dots$ and y_n, y_{n-1}, \dots . The future error terms are unknown, whereas the remainder term is known. The variance of the prediction is thus σ^2 times the sum of squares of the coefficients associated with the future noise terms. As with one step ahead prediction, uncertainty associated with the AR and MA parameters is ignored.

Non-normal noise. ARIMA modeling assumes the approximate normality and independence of the noise terms ε_t . A histogram plot of the computed noise terms $\hat{\varepsilon}_t$ as in (17.6), computed with the estimated ARMA coefficients, will display departures from normality, if they are present. Many financial time series are “leptokurtic” in that, despite symmetry, the tails of the

distribution are fat compared to the normal. Leptokurtic computed residuals e_t suggest that an ARIMA model does not adequately capture the large shocks and that stochastic volatility modeling, for example, is needed, as discussed in Section 17.4.2.

17.3.8 Seasonal ARIMA Modelling

The variation in many time series is partly explained by seasonal effects. Seasonal factors typically show up in the time series plots of the series and in the ACF. For example, a monthly time series with a yearly cycle shows up as strong correlations at lag 12, 24, and so on. The PACF may be similarly affected.

Seasonal versions of the ARMA model are of the form

$$a(L)\alpha(L^s)y_t = b(L)\beta(L^s)\varepsilon_t,$$

where $\alpha(L^s)$ and $\beta(L^s)$ are low-order polynomials in the L^s , with s corresponding to the length of the season. If necessary, y_t is differenced to stationarity before any ARMA fit. The attraction of models such as these is that there are relatively few parameters explaining both short term dependence via $a(L)$ and $b(L)$ and those extending from one season to the next, via $\alpha(L^s)$ and $\beta(L^s)$.

For example a monthly series may be modeled as

$$(1 - aL)(1 - \alpha L^{12})(1 - L^{12})y_t = (1 - \beta L^{12})\varepsilon_t.$$

This is an ARIMA $\{(1, 0, 0), (1, 1, 1)_{12}\}$ model with the first triplet specifying the usual (p, d, q) and the second specifying the orders of the seasonal polynomials and the degree of seasonal differencing. This model specifies seasonal differencing $1 - L^{12}$ and has three parameters: a , α , and β .

Seasonal modeling typically requires more observations than an ordinary ARIMA fit because accurate estimation of the seasonal effect requires the series to extend across a substantial number of seasons.

17.4 Additional Time Series Models

17.4.1 Structural Models

Structural models focus on modeling nonstationarity in the mean. This contrasts with ARIMA modeling where nonstationarity is regarded as a nuisance and is dealt with by appropriate differencing.

A basic structural model is

$$y_t = \mu_t + \varepsilon_t, \quad \mu_{t+1} = \mu_t + \delta_t + a\eta_t, \quad \delta_{t+1} = \delta_t + b\nu_t. \quad (17.13)$$

Only y_t is observed and ε_t , η_t , and ν_t are mutually uncorrelated noise or disturbance terms with common variance σ^2 . Further μ_t is the “level” of the series, whereas δ_t is the “slope,” both varying with t . If (17.13) applies, then $(1 - L)^2y_t$ is a linear combination of the noise terms ε_t , η_t , and ν_t and hence is stationary. The time series thus has two unit roots.

Variances:

level	slope	seas	epsilon
1.033e-03	3.471e-05	2.194e-03	1.404e-03

Fig. 17.6. Output from R command `StructTS(log(earnings))`.

The parameters a and b indicate the importance of level and slope changes or “shocks” η_t and ν_t , as compared to the measurement shocks ε_t . Components μ_t and δ_t are removed from the model, and others, such as seasonal components, are added where appropriate. For example, to remove the slope component δ_t , set $\delta_0 = b = 0$.

To add a seasonal component to, for example, a quarterly series, component γ_t is added in the right-hand side equation for y_t in (17.13) where

$$\gamma_{t+1} = -\gamma_t - \gamma_{t-1} - \gamma_t + c\xi_t, \quad \xi_t \sim (0, \sigma^2). \quad (17.14)$$

This states that the sum of four consecutive seasonal components γ_t adds to $c\xi_t$ where ξ_t is approximately zero, independent of the other disturbance terms. If $c = 0$ the components add exactly to zero, corresponding to seasonal components modeled as in dummy variable regression. Initial γ_0 , γ_{-1} , and γ_{-2} are estimated, because they determine subsequent seasonal effects.

Structural models frame time series in terms of concrete and interpretable constructs: level μ_t , slope or trend δ_t , seasonal γ_t , and the associated parameters a , b , and c . This interpretability contrasts with ARMA models where the interpretation attached to AR or MA coefficients is more obscure. With structural models the emphasis is as much on estimating components such as μ_t and δ_t as on estimating a and b . For example, with the temperature time series a pertinent question is, What is the latest trend δ_t ?

Fitting structural models with R. The command `StructTS(y)` fits a structural model to time series y including, where appropriate, a seasonal component. To eliminate components, use keywords such as "level", which eliminates slope and seasonal components, or "trend", which eliminates just the seasonal.

Figure 17.6 displays the output from the R when fitting the structural model to the log earnings time series displayed in Figure 17.2. The output is in terms of variances of ε_t , $a\eta_t$, and $b\nu_t$. Thus,

$$\hat{\sigma} = \sqrt{0.001404} = 0.037, \quad \hat{a} = \sqrt{\frac{1.033}{1.404}} = 0.86, \quad \hat{b} = \sqrt{\frac{3.471}{140.4}} = 0.157.$$

The fit includes a seasonal component as in (17.14). Thus $\hat{c} = \sqrt{2.194/1.404} = 1.250$.

Straight line trend. If in (17.13), $a = b = 0$, then $y_t = \mu_0 + t\delta_0 + \varepsilon_t$, a straight line observed with error. Fitting with these constraints is tantamount to regressing y_t on t , and hence (17.13) generalizes the straight line trend model by allowing both the intercept and slope to vary over time. If additionally $\delta_0 = 0$, then y_t is noise with constant mean μ_0 .

Random walk with fixed drift. If $a \neq 0$ and $b = 0$ then $\delta_t = \delta_0$ is constant and y_t is a random walk with fixed drift δ_0 . Zero drift results if $\delta = 0$.

Starting conditions. The straight line and random walk with drift special cases indicate the relevance of μ_0 and δ_0 . These “starting conditions” are not necessarily nuisance parameters, but are often practically important and warrant estimation.

Testing restrictions. The significance of a component in (17.13) is judged by computing the drop in the log-likelihood when it is eliminated. For example, suppose `r=StructTS(log(earnings))` is the R command. Then `r$loglik` displays the value of the log-likelihood of the fitted model – in this case 289.559. With `r=StructTS(log(earnings), "trend")` the seasonal terms γ_t are eliminated; in this case `r$loglik` is 258.446, implying a drop of $289.559 - 258.446 = 31.113$. The significance of the drop is evaluated with the likelihood ratio test:

$$2 \times \ln \left(\frac{\text{Unrestricted Likelihood}}{\text{Restricted Likelihood}} \right) = 2 \times (289.559 - 258.446) = 62.226.$$

This test statistic is compared to the chi-squared distributions with degrees of freedom equal to the number of parameters set to zero: in this case $c = \gamma_0 = \gamma_{-1} = \gamma_{-2} = 0$, a total of four parameters. Clearly the drop is very significant, and there appears to be a seasonal effect.

Stationary components. ARMA components can be included in the model to model stationary dependence. For example ε_t in (17.13) may be replaced by an AR(1) if there is short-term transitory dependence in y_t on top of the dependence induced by the nonstationary components.

Kalman filter. This is an algorithm for the recursive computing of one step ahead predictions for models of the form (17.13) and its generalizations, often called state-space models. The predicted value given the past, \hat{y}_t , and the realized value y_t are combined with ancillary information accumulated with the filter to derive the next one-step-ahead prediction \hat{y}_{t+1} . The weights in the weighted average are automatically adjusted as more observations come to hand. The algorithm at the same time computes variance of the prediction error, as well as estimates of the “state” (in this case μ_t and δ_t), and their prediction variances and covariances. Calculations assume values of a , b , c , and σ^2 .

Maximum likelihood estimation of structural models. The parameters of (17.13) and its seasonal extension (17.14) are estimated iteratively. For example, with (17.13), given a guess of the unknown parameters a , b , and σ^2 , the one step ahead MMSE predictor \hat{y}_t is calculated as well as the MSE: $\sum_{t=1}^n (y_t - \hat{y}_t)^2 / n$. Maximizing the likelihood is equivalent to minimizing the MSE plus second-order terms. Hence maximum likelihood estimation proceeds by repeatedly guessing the unknown parameters, evaluating the MSE and the second-order adjustment, and using a numerical search to find the maximizing parameters. The search is facilitated by using numerical optimization.

Smoothing filter. This recursive algorithm computes estimates of μ_t , δ_t , and the noise terms, for $t = 1, \dots, n$ based on all available data y_1, \dots, y_n , not just the past. For example, given a time series y_1, \dots, y_n the slope δ_t near the middle of the series may be required as well as its

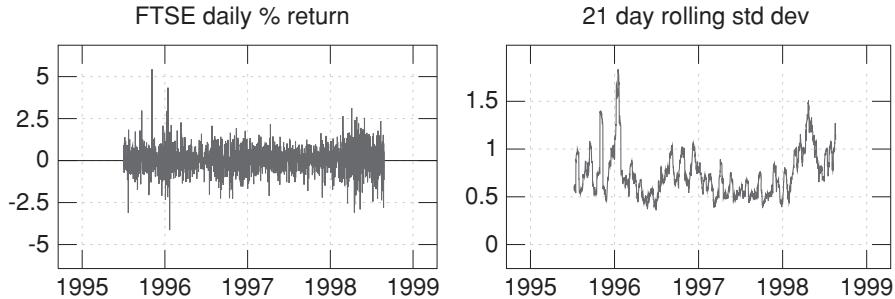


Fig. 17.7. Daily return FTSE index and standard deviation.

error variance. These and similar calculations are obtained with a single pass of the smoothing filter through y_n, \dots, y_1 .

Fitting the basic structural model to the temperature data yields $\hat{b} = 0$. Hence the slope $\delta_t = \delta_0 \equiv \delta$ is constant and

$$y_t = \mu_t + \varepsilon_t, \quad \mu_{t+1} = \mu_t + \delta + 0.387\eta_t.$$

Hence δ , the initial value, is critical because it indicates the constant slope of the process. The initial value is estimated as $\hat{\delta} = 0.0007$, corresponding to an increase of $100 \times 12 \times 0.0007 = 0.84$ C per century.

17.4.2 Stochastic Volatility Models

Stochastic volatility models seek to explain nonconstant variance or volatility in a time series. Volatility is critical in risk calculations where higher volatility means higher risk. Varying volatility occurs in stock price data in which periods of relative stability are followed by high volatility. A good example is provided in Figure 17.7, which displays, in the left panel, daily returns computed from the daily closing prices of the FTSE index, 1991–1998, available with the `FTSE=EuStockMarkets[, "FTSE"]` command in R. Returns are computed as the differences in the logarithm. The estimated mean and standard deviation (average volatility) are 0.0432% and 0.796%, respectively.

The right-hand panel in Figure 17.7 displays the 21-day rolling average of the standard deviation in daily returns. The rolling standard deviation displays strong autocorrelation over and above that due to averaging. For example, the sample ACF (not shown here) of the rolling standard deviations at lags larger than 10, the half-width of the moving average, is highly significant. Hence volatility appears to be nonconstant and highly autocorrelated.

A stochastic volatility model is the GARCH(1,1) model stated in terms of noise $\varepsilon_t \sim (0, 1)$ and time-varying volatility σ_t :

$$y_t = \mu + \sigma_t \varepsilon_t, \quad \sigma_{t+1}^2 = a_0 + a_1 \sigma_t^2 + b_1(y_t - \mu)^2. \quad (17.15)$$

```

garch(x = 100 * diff(log(FTSE)), order = c(1, 1))

Coefficient(s):
      a0        a1        b1
 0.008723  0.045322  0.941862

```

Fig. 17.8. Output from R with indicated garch command.

Here μ is the mean of the series. The equation for σ_{t+1}^2 is analogous to an ARMA(1,1) with $(y_t - \mu)^2$ serving as the “innovation” or noise term. The rationale is that $(y_t - \mu)^2$ is the latest available estimate of the variance at time t , and it is combined with the “permanent” component σ_t^2 . Constraints are required on a_0 , a_1 , and b_1 to ensure that σ_t stays positive and finite. If $a_1 = b_1 = 0$, then y_t is noise with mean μ .

Fitting GARCH models with R. The command `garch(y, c(1, 1))` fits model (17.15) to y . The `garch` command requires `library(tseries)`.

The quantity $a_0/(1 - a_1 - b_1)$ is the long-run average σ_t^2 because it is the constant solution to the right-hand-side equation (17.15), if $(y_t - \mu)^2$ is replaced by its expectation.

Figure 17.8 displays output from fitting a GARCH model to the daily percentage returns of the FTSE index plotted in Figure 17.7. In this fit, μ is set to the sample mean return: $\hat{a}_1 = 0.045$ indicates moderate autocorrelation in the standard deviation, whereas $\hat{b}_1 = 0.942$ indicates high reaction to the latest volatility $(y_t - \mu)^2$. The long-run average standard deviation is $\sqrt{0.009/(1 - 0.045 - 0.942)} = 0.825$, marginally above the empirical standard deviation of 0.796. The standard errors of the coefficients are 0.003, 0.007, and 0.010, respectively, for a_0 , a_1 , and b_1 . These are available using the `vcov(x)` where x contains the results of the `garch` command. Hence all three coefficients are significant.

GARCH stands for generalized autoregressive conditional heteroskedasticity. GARCH(1,1) indicates one lag in both the σ_t^2 and $(y_t - \mu)^2$ terms. The GARCH(p, q) model extends this to p and q lags. The ARCH(q) model contains only lags in $(y_t - \mu_t)^2$ and hence is GARCH(0, q).

Estimating GARCH models is similar to the estimation of ARMA models. For any set of a and b parameters, as well as starting conditions, y_0 and σ_0 , the GARCH equation for σ_t^2 is used to recursively calculate $(y_t - \hat{y}_t)/\sigma_t$. These standardized errors, as well as related information, are combined to evaluate the normal based likelihood. The likelihood is maximized with respect to the unknown parameters using a numerical search.

Forecasting future volatility with a GARCH model requires of iterating the σ_t^2 equation (17.15) forward from the latest available time point $t = n$. In this iteration, future $(y_t - \mu_t)^2$ are replaced by their latest conditional expectation.

17.4.3 Vector Time Series and the Lee–Carter Model

The developments of the previous section all suppose that y_t is a scalar time series – that is, one value is measured at each point of time. In many situations, however, many time

series are measured simultaneously. In this case y_t is a vector of measurements with different components corresponding to, for example, different stock prices, a variety of economic indicators, or mortality at different ages. The concept of serial correlation is now extended to correlation between different components of y_t at different points of time.

The Lee-Carter (LC) model is a special, idiosyncratic multiple time series model much used in the analysis of mortality. The time series measurements are mortality deaths at different ages $i = 0, 1, \dots, p$, say, where p is the maximum age and it is assumed that

$$y_t = a + b\tau_t + \varepsilon_t.$$

Here a and b are vectors with as many components as there are ages, whereas τ_t is an unobserved scalar time series modeling the trend in overall mortality, and ε_t is a vector of mutually uncorrelated noise terms. The vector of mortalities y_t is usually taken as the logarithm of the death rate; that is, the number of deaths divided by exposure. The aim of the analysis is to estimate a and b as well as make inferences about the overall trend τ_t and, if required, forecast τ_t to obtain an understanding of likely future death rates.

17.5 Further Reading

Box, Jenkins, and Reinsel (1994) is the classic text in ARIMA modeling, describing in detail the many steps and techniques of ARIMA or Box–Jenkins modeling. Harvey (1989) gives a detailed treatment of structural time series modeling. Chatfield (1989) and Kendall and Ord (1990) are classic texts describing stationary time series analysis. Shumway and Stoffer (2011) describe time series analysis using R.

References

- Box, G. E. P., G. M. Jenkins, and G. C. Reinsel (1994). *Time Series Analysis: Forecasting and Control* (3rd ed.). Prentice-Hall, Englewood Cliffs, NJ.
- Chatfield, C. (1989). *The Analysis of Time Series: An Introduction* (2nd ed.). Chapman and Hall, London.
- Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.
- Jones, P. (2010, April). Climate Research Unit at the University of East Anglia.
[http://www.cru.uea.ac.uk/cru/data/temperature/
CRUTEM3-g1.dat](http://www.cru.uea.ac.uk/cru/data/temperature/CRUTEM3-g1.dat).
- Kendall, M. G. and J. K. Ord (1990). *Time Series* (3rd ed.). E. Arnold, Seven Oaks, UK.
- R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Shumway, R. H. and D. S. Stoffer (2011). *Time Series Analysis and Its Applications: With R Examples*. Springer, New York.

18

Claims Triangles/Loss Reserves

Greg Taylor

Chapter Preview. This chapter considers the application of predictive models to insurance claims triangles and the associated prediction problem of loss reserving (Section 18.1). This is approached initially by reference to the chain ladder, a widely used heuristic reserving algorithm. Rigorous predictive models, in the form of Generalized linear models (GLMs) that reproduce this algorithm, are then explored (Section 18.2). The chain ladder has a restricted model structure and a number of embellishments are considered (Section 18.3). These include the incorporation in the model of accident year effects through the use of exposure data (e.g., accident year claim counts) (Section 18.3.2), and also the incorporation of claim closure data (Section 18.3.3). A subsequent section considers models that incorporate claim closure data on an operational time basis (Section 18.3.4). In each of these cases emphasis is placed on the ease of inclusion of these model features in the GLM structure. All models in Sections 18.1 to 18.3 relate to conventional claims triangles. These datasets are aggregate, as opposed to unit record claim datasets that record detail of individual claims. The chapter closes with a brief introduction to individual claim models (18.4). On occasion these models use survival analysis as well as GLMs.

18.1 Introduction to Loss Reserving

18.1.1 Meaning and Accounting Significance of Loss Reserving

Typically, property & casualty (**P&C**) insurance will indemnify the insured against events that occur within a defined period. From an accounting viewpoint, a liability accrues to the insurer at the date of occurrence of an indemnified event. From that date the insurer's accounts are therefore required to recognize the quantum of this liability to the extent that it has not been settled.

However, there will be an inevitable delay between the occurrence of the event and notification of an associated claim to the insurer, and a further delay between notification and settlement of the claim. These delays can be considerable. For example, in

casualty classes of business, such as auto bodily injury liability, a typical period from claim occurrence to settlement might be 3 or 4 years, with some claims remaining unsettled for 10 to 20 years.

During the period between occurrence and notification, even the existence of the claim will be unknown to the insurer. Between notification and settlement its existence will be known, but its ultimate quantum will be unknown because it may depend on the outcome of negotiation between insurer and insured or on an adversarial legal process. The insurer is obliged to make accounting provision for the claim despite these shortfalls in knowledge.

The balance sheet item recognizing the insurer's liability is called the *loss reserve*, and the determination of this reserve is *loss reserving*. The reserve will certainly include the liability for the cost of the claims themselves and depending on the accounting regime, it may also include allowance for other associated costs such as loss adjustment expenses.

For long-tail lines of insurance, the loss reserve may amount to several times the annual premium income, and therefore it may amount to a large multiple of typical annual profit. As a result, a modest proportionate error in the loss reserve can impinge heavily on annual profit, and so accuracy of the reserve is a matter of considerable importance.

The estimation of loss reserves in practice relies on a wide variety of models and methodologies. It is not feasible to review them all here; for that review, refer to Taylor (2000) and Wüthrich and Merz (2008). Of all models in use, the so-called *chain ladder* is the most commonly encountered. For this reason this chapter begins with this model (Section 18.2), and then proceeds to examine variations of it as well as some other models.

18.1.2 Data and Notation

A typical P&C claim is generated by the occurrence of some event. The generating event may be an accident (e.g., automobile bodily injury insurance), or it may be something else (e.g., professional liability insurance, where the date of occurrence is the date on which some lapse of professional conduct occurred). Nonetheless, the date of this event is referred to generically as the *accident date*.

For the purpose of data analysis it is common to group claims according to the calendar periods in which their accident dates fall. These periods may be years, quarters, months, or any other convenient choice and are referred to as *accident periods*. For convenience, the present chapter assumes that they are *accident years*.

Some lapse of time usually occurs between the generating event and the subsequent payment of the claim. Time measured from the accident year is referred to as *development time*, and years measured from the accident year are referred to as *development*

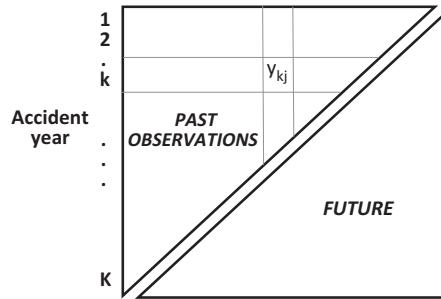


Fig. 18.1. Illustration of a claims triangle.

years. The accident year itself is labeled development year 1, and subsequent years development years 2, 3, and so on.

Consider the situation at the end of calendar year K . Accident year K will have just been completed (i.e., development year 1 will have been completed). For accident year $K - 1$, development years 1 and 2 will have been completed. In general, for accident year k ($\leq K$), development years $1, 2, \dots, K - k + 1$ will have been completed.

Observations on accident years $1, 2, \dots, K$ thus form a triangular structure, called a *claims triangle*, as illustrated by the upper part of Figure 18.1. Accident years are represented by rows and labeled $k = 1, 2, \dots, K$; development years are represented by columns and labeled $j = 1, 2, \dots, K$; and y_{kj} denotes the observation on claims experience in development year j of accident year k . Diagonals in this triangle, represented by $k + j = \text{const}$, denote contemporaneous years of calendar time and are referred to subsequently as *calendar years*.

The precise nature of the observations y_{kj} are left unspecified for the moment. They may comprise counts of claim notifications or closures, amounts of paid losses or case estimates, or some alternative. Specific examples are given later in this chapter.

Let the claims triangle in Figure 18.1 be denoted $\mathcal{D}_K = \{y_{kj} : 1 \leq k \leq K, 1 \leq j \leq K - k + 1\}$. Future observations on the same accident years, up to and including development year, K , form the lower triangle in Figure 18.1, denoted $\mathcal{D}_K^c = \{y_{kj} : 2 \leq k \leq K, K - k + 1 < j \leq K\}$. Also let

$$\mathcal{D}_K^+ = \mathcal{D}_K \cup \mathcal{D}_K^c.$$

This chapter is concerned with the prediction of \mathcal{D}_K^c on the basis of observed \mathcal{D}_K . Row sums in \mathcal{D}_K^c are of particular interest. These are the quantities

$$R_k = \sum_{j=K-k+2}^K y_{kj}, \quad k = 2, 3, \dots, K. \quad (18.1)$$

Define the *cumulative row sums*:

$$x_{kj} = \sum_{i=1}^j y_{ki}. \quad (18.2)$$

Also define, for $k = 2, 3, \dots, K$,

$$R_k = \sum_{j=K-k+2}^K y_{kj} = x_{kK} - x_{k,K-k+1}, \quad (18.3)$$

$$R = \sum_{k=2}^K R_k. \quad (18.4)$$

Note that R is the sum of the (future) observations in \mathcal{D}_K^c . In the case that the y_{kj} denote paid losses, R is referred to as the total amount of *outstanding losses*. Likewise, R_k denotes the amount of outstanding losses in respect of accident year k . The objective stated earlier is to forecast the R_k and R .

In regarding R as the total amount of outstanding losses, it is assumed that no claims activity occurs beyond development year K . If this is not so, estimated outstanding losses will require increase accordingly. Such an increase will not, however, be obtainable from the available data, which do not involve development years beyond K , and the details of this issue are beyond the scope of the present chapter. In general terms, one would need to rely on the following:

- benchmark data from some external source,
- case estimates of liabilities beyond development year K , or
- some form of extrapolation (e.g., an assumption that, for each k , the y_{kj} , $j = J, J + 1, \dots$ form a geometric sequence).

18.1.3 Modeling and Forecasting

A loss reserve requires a forecast of future claim payments. The forecast of future claims experience, including the forecast of outstanding losses, is approached by means of modeling past experience. As noted in Section 18.1.1, a loss reserve may include components other than future paid losses, but estimation of these additional components is outside the scope of this chapter.

The required model of \mathcal{D}_K^+ is of the general form

$$y_{kj} \sim Dist(h_{kj}(b)), \quad (18.5)$$

which indicates that y_{kj} is a random variable with some distribution $Dist$ defined by a vector of parameters $h_{kj}(b)$ – h_{kj} being some function specific to cell (k, j) and b being a vector of parameters relating to the entire array \mathcal{D}_K^+ rather than to cell (k, j) .

The fact that y_{kj} is a random variable in the model definition means that the model is *stochastic*. Notwithstanding that this is essential if the stochastic properties of model forecasts are to be understood, many conventional actuarial loss reserving models are not of this type (see, e.g., Taylor 2000). One such model is encountered in Section 18.2.1.

18.2 The Chain Ladder

18.2.1 Heuristic Model

The chain ladder has been the mainstay of the loss reserving toolkit for many years. It was not formulated as a stochastic model until 1975. Before then it was an algorithm rather than a model (see the following box). Taylor (1986) traces its provenance as far back as Harnek (1966), but it is likely to have been in use earlier.

Chain Ladder Algorithm

Define the following **age-to-age factors**:

$$\hat{f}_j = \frac{\sum_{k=1}^{K-j} x_{k,j+1}}{\sum_{k=1}^{K-j} x_{kj}} j = 1, 2, \dots, K-1. \quad (18.6)$$

Then define the forecasts of $y_{kj} \in \mathcal{D}_K^c$:

$$\hat{y}_{kj} = x_{k,K-k+1} \hat{f}_{K-k+1} \hat{f}_{K-k+2} \cdots \hat{f}_{j-2} (\hat{f}_{j-1} - 1). \quad (18.7)$$

These then yield the following forecasts of quantities (18.2)–(18.4):

$$\hat{x}_{kj} = x_{k,K-k+1} \hat{f}_{K-k+1} \hat{f}_{K-k+2} \cdots \hat{f}_{j-1} \quad (18.8)$$

$$\hat{R}_k = \sum_{j=K-k+2}^K \hat{y}_{kj} = \hat{x}_{kK} - x_{k,K-k+1} \quad (18.9)$$

$$\hat{R} = \sum_{k=2}^K \hat{R}_k. \quad (18.10)$$

The procedure described by (18.6)–(18.10) is called the *chain ladder algorithm*.

In practice actuaries may apply judgmental adjustments to this algorithm. For example,

- If the \hat{f}_j do not form a smooth sequence over j , they may hand smooth the sequence.
- If a particular \hat{f}_j appears unduly affected by a single observation $x_{k,j+1}$ (for that same j), they may adjust the \hat{f}_j .

Section 18.3 comments further on such practices.

18.2.2 Formal Stochastic Models

The chain ladder algorithm was shown to be derivable from a formal stochastic model by Hachemeister and Stanard (1975). A slightly generalized form of their model, introduced by England and Verrall (2002) and subsequently labeled the *ODP cross-classified model* (Taylor 2011), is described by the following properties.

ODP Cross-Classified Model

- (ODPCC1) The random variables $y_{kj} \in D$ are stochastically independent.
- (ODPCC2) For each, $k, j = 1, 2, \dots, K$
 - (a) $y_{kj} \sim ODP(\alpha_k \gamma_j, \varphi)$ for some parameters $\alpha_k, \gamma_j, \varphi > 0$;
 - (b) $\sum_{j=1}^K \gamma_j = 1$.

Here $y \sim ODP(\mu, \varphi)$ means that y is subject to an *overdispersed Poisson distribution* (ODP) with mean μ and dispersion parameter φ (i.e., $y/\varphi \sim Poisson(\mu/\varphi)$). This yields $E[y] = \mu$ and $\text{Var}[y] = \varphi\mu$, where, in practice, often $\varphi > 1$ (overdispersion). The model of Hachemeister and Stanard (1975) assumes $\varphi = 1$ (i.e., y_{kj} Poisson distributed).

England and Verrall (2002) show that maximum likelihood estimation (MLE) of the ODP cross-classified model parameters, followed by forecast of \mathcal{D}_K^+ by means of these estimates, produces the same result as the chain ladder algorithm.

ODP Mack Model

- (ODPM1) Accident years are stochastically independent (i.e., $y_{k_1 j_1}, y_{k_2 j_2}$ are stochastically independent if $k_1 \neq k_2$).
- (ODPM2) For each, $k = 1, 2, \dots, K$ the x_{kj} (j varying) form a Markov chain.
- (ODPM3) For each $k = 1, 2, \dots, K$ and for each $j = 1, 2, \dots, K - 1$,
 $y_{k,j+1}|x_{kj} \sim ODP(x_{kj} f_j, \varphi_j)$ for some parameters $f_j, \varphi_j > 0$.

An alternative model, the *ODP Mack model*, was introduced by Taylor (2011). It consists essentially of the earlier distribution-free *Mack model* (Mack 1993) with an ODP distribution added for each observation. This model is very different from the ODP cross-classified model; the differences concerned with stochastic independence are particularly striking. The Markov chain in (ODPM2) ensures dependency between observations within the same row, whereas all observations are independent in the cross-classified model. Nonetheless, once again, MLE of the model parameters, followed by forecast of \mathcal{D}_K^c by means of these estimates, produces the same result as the chain ladder algorithm.

Table 18.1. *Chain Ladder Models as GLMs*

GLM Characteristic	ODP Cross-Classified Model	ODP Mack Model
Response variate	y_{kj}	$y_{k,j+1} x_{kj}$
Covariates (categorical)	Accident years k , development years j	Accident years k , development years j
Link function	Log	Log
Error structure	ODP	ODP
Weights	None	Column dependent permitted
Offset	None	$\ln x_{kj}$
Linear response	$\ln \alpha_k + \ln \gamma_j$	$\ln f_j$

18.2.3 GLM Formulation

It is worthy of note that each of these two formulations of chain ladder models – the ODP cross-classified model and the ODP Mack model – is capable of expression as a GLM (see Chapter 5), as shown in Table 18.1. Note that these formulations make use of the two advantages of the GLM over ordinary linear regression:

1. nonlinearity of the relation between response and covariates, and
2. nonnormality of error structure.

18.3 Models of Aggregate Claims Triangles

18.3.1 More Flexible Predictive Model Structures

Note that the ODP cross-classified model contains $2K$ parameters, of which one will be aliased, leaving $2K - 1$ independent parameters. The ODP Mack model specifies only $K - 1$ parameters, though Verrall (2000) points out that the observations $x_{k,K-k+1}$, $k = 1, 2, \dots, K$ function as K additional parameters.

The literature contains numerous comments that chain ladder models of this sort are over-parameterized. Might it be possible to represent the categorical variates γ_j of (ODPCC2) by a continuous function of j and reduce the number of parameters? And to do the same with the α_k ?

The GLM formulation of the chain ladder models in Table 18.1 prompts a number of other questions. For example, should a trend be modeled over calendar years? Should all accident years be modeled as subject to the same schedule of γ_j , or can there be differences?

In short, is there a need to adapt the very rigid structures of the chain ladder models to more flexible forms? The chain ladder algorithm, consisting of row sums and column ratios, precludes extensions of this sort. However, many of them are easily feasible within a GLM structure. The following subsections provide some examples.

18.3.1.1 Development Year Profiles

Consider the earlier suggestion that the categorical variates γ_j be represented by a continuous functional form. In this case the linear predictor set out in Table 18.1 is replaced by

$$\eta(k, j) = \alpha_k + f(j), \quad (18.11)$$

where $f(\cdot)$ is a function yet to be defined. If it is to form part of a linear predictor, it must be linear in any unknown parameters.

The form (18.11) is a particular case of a *generalized additive model* (Hastie and Tibshirani 1990), for which the linear predictor in general takes the form

$$\eta(k, j) = \sum_i \beta_i f_i(k, j),$$

with the $f_i(\cdot)$ predefined parametric functional forms and the β_i regression coefficients requiring estimation.

Generalized additive models were introduced into the loss reserving literature by Verrall (1996). The idea of loss reserving by means of regression with parametric functions as independent variables can also be found in earlier papers (e.g., De Jong and Zehnwirth 1983; Frees and Wang 1990; Reid 1978). Some of these approaches are discussed by Wüthrich and Merz (2008).

It is sometimes appropriate in practice to suppose that the profile of the γ_j over j follows a gamma-like shape:

$$\gamma_j = A j^b \exp(-cj), \quad (18.12)$$

where A, b, c are parameters to be estimated. In this case (18.11) becomes

$$\eta(k, j) = \alpha_k + a + b \ln j + c(-j), \quad (18.13)$$

with $a = \ln A$.

The linear predictor $\eta(k, j)$ is indeed linear in all unknown parameters, and the number of those parameters has been reduced from the $2K - 1$ of the categorical model to $K + 2$ (again one parameter will be aliased).

It is of particular interest that the adoption of a smooth functional form $f(j)$ in (18.11) ensures both a smooth development year profile and a smooth sequence of estimated age-to-age factors. It would replace the hand smoothing of the sequence of \hat{f}_j mentioned at the end of Section 18.2.1. However, the parameters defining $f(j)$ would be estimated objectively, which may be preferable to ad hoc smoothing.

18.3.1.2 Accident Year Profiles

The same sort of reasoning may be applied to the model's accident year structure. If there were no indication of dependence of the linear predictor on accident year, then

the categorical parameters α_k could simply be dropped from (18.13) and the number of parameters would be reduced to three.

Alternatively, if the linear response appeared to exhibit a linear trend in k (exponential dependence of observations on k), then the linear predictor would become

$$\eta(k, j) = mk + a + b \ln j + c(-j), \quad (18.14)$$

with m a new unknown parameter. The number of model parameters is now four.

18.3.1.3 Calendar Year Profiles

Inflationary trends operate over calendar years. That is, in the presence of an inflation factor of e^φ between calendar years $k + j$ and $k + j + 1$ but no other accident year trend, the expectation of each observation from calendar year $k + j + 1$ will be a multiple e^φ of that of the same development year in calendar year $k + j$.

A constant rate of inflation would be added to model (18.13), and the accident year trend omitted, by means of the following linear predictor:

$$\eta(k, j) = a + b \ln j + c(-j) + (k + j)\varphi. \quad (18.15)$$

Note that (18.15) is, as required, linear in the unknown parameters a , b , c , and φ .

If the y_{kj} represent claim amounts, then they may have been corrected for inflation before modeling in an attempt to eliminate the calendar year trend from the model. This correction is typically carried out using some standard form of monetary inflation; for example, adjusting workers' compensation claim payments by reference to wage inflation or adjusting automobile property damage claim amounts by reference to a mixture of price inflation (vehicle parts) and wage inflation (repair labor). On such a basis the y_{kj} would be adjusted to common dollar values, such as those relating to the end of calendar year K , the date of evaluation of the loss reserve.

Claim amounts may sometimes escalate at a rate different from the seemingly "natural" inflation index. The difference between the observed rates of inflation and those associated with the inflation index is called *superimposed inflation* (SI). In such cases a calendar year trend is required in the model of claim amounts and is used to estimate SI.

SI is often found to progress irregularly over time, and so similarly for the total inflation of claim amounts. In this case a constant rate φ will not exist as in (18.15), and it will be necessary for (18.15) to include a more general function of $(k + j)$. See, for example, Section 18.4.2.

18.3.2 Use of Exposure Data

For some lines of insurance it is possible to identify a time series $\{e_k\}$ with respect to accident year such that the claims experience of accident year k can be expected to be

proportional to e_k . For example, the counts of auto bodily injury claim notifications of accident year k might be proportional to the average number of vehicles insured over that year. The counts of workers' compensation claim notifications of accident year k might be proportional to the total payroll of the workforce over that year.

In such cases the quantity e_k is referred to as the *exposure* associated with accident year k . Since claims experience is postulated to be proportional to it, exposure replaces α_k in (18.11) thus:

$$\eta(k, j) = \ln e_k + f(j). \quad (18.16)$$

This substitution may appear trivial. However, it should be noted that e_k in (18.16) is a known quantity, whereas α_k in (18.11) is an unknown quantity requiring estimation. In other words, $\ln e_k$ forms an *offset* within the GLM, whereas α_k is a parameter.

Recall that the model under discussion uses a log link, in which case $\ln E[y_{kj}] = \eta(k, j)$ and so, by (18.16),

$$\ln \{E[y_{kj}]\}/e_k = f(j). \quad (18.17)$$

The quantity within the braces represents *claim experience per unit exposure*. If, for example, y_{kj} denotes the amount of paid losses in the (k, j) cell, then the quantity within braces represents *payments per unit exposure*. If $e_k = \hat{N}_k$ (i.e., the estimated number of claims incurred in accident year k), then the quantity within braces represents *payments per claim incurred*, as illustrated in section 4.2 of Taylor (2000).

By linearity of the expectation operator, (18.17) yields

$$\ln \{E[x_{kk}]\}/e_k = \sum_{j=1}^K f(j), \quad (18.18)$$

and if y_{kj} denotes the claim notification count in the (k, j) cell, then the quantity within braces represents *ultimate claim frequency* for accident year k (e.g., claim frequency per vehicle insured in the auto bodily injury case mentioned at the start of this subsection).

The right side of (18.18) is independent of k and so represents the case of claim frequency that is constant over accident years. Although the exposure term was introduced into (18.16) as a proxy for the accident year effect α_k , it will be an inadequate proxy if claim frequency exhibits a trend over accident years. In this case it will be necessary to reintroduce α_k to capture that trend, so that (18.18) becomes

$$\ln \{E[x_{kk}]\}/e_k = \alpha_k + \sum_{j=1}^K f(j), \quad (18.19)$$

or, even more generally,

$$\eta(k, j) = \ln e_k + \alpha_k + f(j), \quad (18.20)$$

which is a generalization of (18.16). Here the total accident year effect is $\ln e_k + \alpha_k$ consisting of a known effect, $\ln e_k$ and an unknown effect α_k requiring estimation.

Just as for the development year effect, it may be possible to replace the categorical covariate α_k by a parametric form $g(k)$. The simplest such form would be $g(k) = \alpha k$ for unknown parameter α , in which case (18.19) yields

$$E[x_{kK}] / e_k = (\exp \alpha)^k \times \exp \left[\sum_{j=1}^K f(j) \right], \quad (18.21)$$

and claim experience per unit exposure exhibits an exponential trend over accident year.

18.3.3 Use of Claim Closure Data

The present subsection is specific to the case in which y_{kj} denotes the amount of paid losses in the (k, j) cell. Suppose that the number of claim closures in the cell is available and is denoted c_{kj} . For certain lines of business, it may be reasonable to assume that y_{kj} will be related to c_{kj} ; specifically, y_{kj} may be modeled as

$$\ln \{E[y_{kj}] / c_{kj}\} = \alpha_k + f(j), \quad (18.22)$$

or, in a form parallel to (18.20),

$$\ln E[y_{kj}] = \ln c_{kj} + \alpha_k + f(j), \quad (18.23)$$

where the quantity in the braces in (18.22) is the expected amount of *paid losses per claim closure* (PPCC) (see section 4.3 of Taylor 2000) and the term $\ln c_{kj}$ in (18.22) is now an offset. PPCCs are sometimes referred to as *payments per claim finalized* (PPCF).

18.3.4 Operational Time

Define

$$d_{kj} = \sum_{i=1}^j c_{ki}, \quad (18.24)$$

which is the cumulative number of claims closed up to the end of development year j of accident year k . Note that $d_{k\infty} = N_k$, the number of claims incurred in accident year k , because all of these claims must ultimately close.

Table 18.2. Illustrative Example of Operational Time and Paid Losses per Claim Closure

Development Year	Accident Year k		Accident Year $k + 1$	
	Operational Time at End of Development Year	Expected Payment per Claim Closure	Operational Time at End of Development Year	Expected Payment per Claim Closure
1	10%	\$ 5,000	20%	\$ 7,500
2	20%	10,000	50%	30,000
3	50%	30,000	80%	50,000
4	80%	50,000	100%	100,000
5	100%	100,000		

Now define *operational time* (OT) associated with accident year k and corresponding to elapsed time j (integral) from the beginning of that accident year as

$$t_k(j) = d_{kj}/N_k, \quad (18.25)$$

which is just the proportion of incurred claims closed up to time j .

Although (18.25) defines OT only for integral j , the definition continues to be sensible for non-integral j if the definition of d_{kj} is extended to mean the cumulative number of claims closed up to elapsed time $j \in [0, \infty]$. Then $t_k(j)$ is also defined for $j \in [0, \infty]$.

The relevance of OT can be illustrated by means of a simple hypothetical example. Suppose that, for accident years k , expected PPCCs within each development year are as given in Table 18.2.

Now consider accident year $k + 1$. Its essential feature is the acceleration of OT, whereby claims that had previously been settled in the first two development years (20% of claims incurred) are now settled in the first year. The expected PPCC in development year 1 become $[10\% \times \$5,000 + 10\% \times \$10,000]/20\% = \$7,500$. All subsequent claim payments are advanced by a year.

If one were to consider expected PPCC as a function of development year, then a radical change in claims experience is observed between accident years k and $k + 1$. In fact, however, there has been no change in expected PPCC, but only a change in the timing of settlements. This becomes evident when expected PPCC is considered as a function of OT.

It follows that one may choose to model PPCC in terms of OT. If, for example, y_{kj} denotes PPCC, then one might assume that

$$y_{kj} \sim \text{Gamma}(\mu_{kj}, \nu_{kj}), \quad (18.26)$$

$$\lambda(k, j) = \ln \mu_{kj} = f(t_k(j)), \quad (18.27)$$

where μ_{kj} , v_{kj} are the mean and coefficient of variation, respectively, of y_{kj} and f is some function requiring specification. By (18.26), the model has a log link, and as previously, $\lambda(k, j)$ denotes the linear predictor.

Accident year and calendar year effects may be included as necessary, just as in Section 18.3.1.

18.3.5 Interactions

From Chapter 5, the linear response of a GLM takes the vector form

$$\eta = X\beta, \quad (18.28)$$

or, in component form,

$$\eta_i = \sum_m x_{im} \beta_m, \quad (18.29)$$

with x_{im} the value of the m -th explanatory variable for the i -th observation.

If one now wishes to add a further effect to the linear response, involving say the m_1 -th and m_2 -th explanatory variables, and this effect cannot be expressed as a linear combination of functions of these variables, then the linear response takes the form

$$\eta_i = \sum_m x_{im} \beta_m + f(x_{im_1}, x_{im_2}), \quad (18.30)$$

for some nonlinear function f .

In this case, the nonlinear effect $f(x_{im_1}, x_{im_2})$ is called an *interaction* between the m_1 -th and m_2 -th explanatory variables. Examples were given in Chapter 2. The definition of an interaction can be extended to involve more than two explanatory variables.

Consider the model of PPCC for accident year k in Table 18.2. The linear response takes the form

$$\eta(k, j) = (\ln 5000) I(0 \leq t_k(j) < 0.1) + (\ln 10000) I(0.1 \leq t_k(j) < 0.2) + \dots, \quad (18.31)$$

where $I(\cdot)$ is an indicator function defined as follows:

$$\begin{aligned} I(c) &= 1 \text{ if condition } c \text{ is satisfied,} \\ &= 0 \text{ otherwise.} \end{aligned} \quad (18.32)$$

Now suppose that this model holds for accident years up to and including 2005. A legislative change was introduced at the start of 2006 that, with one exception, halved the expected PPCCs. The exception affected the final 20% of closures. These are, by and large, the most serious claims, and the legislation did not change the conditions governing their entitlements. The results are shown in Table 18.3.

Table 18.3. *Illustrative Example of Operational Time and Paid Losses per Claim Closure*

Operational Time	Expected Payment per Claim Closure	
	Accident Years 2005 and Earlier	Accident Years 2006 and Earlier
%	\$	\$
0–10	5,000	2,500
10–20	10,000	5,000
20–50	30,000	15,000
50–80	50,000	25,000
80–100	100,000	100,000

The linear response (18.31) now becomes

$$\begin{aligned}\eta(k, j) = & [(\ln 5000) I(0 \leq t_k(j) < 0.1) + (\ln 10000) I(0.1 \leq t_k(j) < 0.2) + \dots] \\ & + (\ln 1/2) I(k \geq 2006) I(0 \leq t_k(j) < 0.8).\end{aligned}\quad (18.33)$$

The term $(\ln 1/2) I(k \geq 2006)$ results in the halving of all PPCCs for accident years 2006 and later, and the additional factor $I(0 \leq t_k(j) < 0.8)$ ensures that this does not apply to OTs above 80%. The model covariates here are shown in Table 18.4. Note that the term added to the linear predictor in (18.32) is a non-linear function of this set of covariates. It is therefore of the same nature as the final member of (18.30) and constitutes an interaction.

18.3.6 Forecasts

The purpose of the modeling just described is, as discussed in Section 18.1, ultimately the forecast of \mathcal{D}_K^c on the basis of the model of \mathcal{D}_K . The general form of a GLM of

Table 18.4. *Set of Covariates in Model (18.33)*

$I(0 \leq t_k(j) < 0.1)$
$I(0.1 \leq t_k(j) < 0.2)$
$I(0.2 \leq t_k(j) < 0.5)$
$I(0.5 \leq t_k(j) < 0.8)$
$I(0.8 \leq t_k(j) < 1)$
$I(t_k(j) \geq 2006)$

\mathcal{D}_K is given in Chapter 5 as follows:

$$\mathbf{y} \sim EDF, \text{ with } E[\mathbf{y}] = g^{-1}(\mathbf{X}\boldsymbol{\beta}), \quad (18.34)$$

where \mathbf{y} denotes the set of $y_{kj} \in \mathcal{D}_K$ arranged as a vector, $\mathbf{y} \sim EDF$ indicates that it is sampled from a member of the exponential dispersion family, $\boldsymbol{\beta}$ denotes the vector of model parameters, \mathbf{X} a design matrix, and g a link function that operates component-wise on its vector argument.

Let the set of $y_{kj} \in \mathcal{D}_K^c$ be arranged in a vector, \mathbf{y}^c , which may be represented in a form parallel to (18.34):

$$\mathbf{y}_{model}^c \sim EDF, \text{ with } E[\mathbf{y}_{model}^c] = g^{-1}(\mathbf{X}^c\boldsymbol{\beta}), \quad (18.35)$$

where \mathbf{X}^c is the matrix of covariate values that apply to the future. The notation \mathbf{y}_{model}^c is used to recognize that the right side of (18.35) is an idealized representation of reality and that observation may not (almost certainly will not) conform with it precisely. In other words, \mathbf{y}^c is distinguished from \mathbf{y}_{model}^c .

A forecast \mathbf{y}^{c*} of \mathbf{y}^c is given by an estimate of $E[\mathbf{y}_{model}^c]$ from the model, which is

$$\mathbf{y}^{c*} = g^{-1}(\mathbf{X}^c\hat{\boldsymbol{\beta}}) \quad (18.36)$$

by (18.35), with $\hat{\boldsymbol{\beta}}$ an estimate of $\boldsymbol{\beta}$.

The matrices \mathbf{X} and \mathbf{X}^* may be illustrated by the following example. One may write $\mathbf{y} = [y_{11}, y_{12}, \dots, y_{1K}, y_{21}, y_{22}, \dots, y_{2,K-1}, \dots, y_{K1}]'$ with the prime denoting transposition. Now consider, for example, the ODP cross-classified model described in Table 18.1, in which case, $g(\cdot) = \ln(\cdot)$, $\boldsymbol{\beta} = [\ln \alpha_1, \ln \alpha_2, \dots, \ln \alpha_K, \ln \gamma_1, \ln \gamma_2, \dots, \ln \gamma_K]'$, and

$$\mathbf{X} = \left[\begin{array}{cccccc|cccccc} 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & & & & & & & & \\ 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 1 & \cdots & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & & & & & & & & \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 \\ & & & & \vdots & & & & \vdots \\ 0 & 0 & \cdots & 1 & 1 & 0 & \cdots & 0 & 0 \end{array} \right] \quad \begin{matrix} \} & K \text{ rows} \\ \} & K-1 \text{ rows} \\ \} & 1 \text{ row} \end{matrix} \quad (18.37)$$

Let the set of $y_{kj} \in \mathcal{D}_K^c$ be arranged in a vector $\mathbf{y}^c = [y_{2K}, y_{3,K-1}, y_{3K}, \dots, y_{K2}, y_{K3}, \dots, y_{KK}]'$. Then

$$\mathbf{X}^c = \left[\begin{array}{ccccccccc|c} 0 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 \\ & & & & \vdots & & & & & \\ 0 & 0 & 0 & \cdots & 1 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 & 1 & \cdots & 0 & 0 \\ & & & & \vdots & & & & & \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 & 0 & \cdots & 0 & 1 \end{array} \right] \quad \begin{array}{l} \} \quad 1 \text{ row} \\ \} \quad 2 \text{ rows} \\ \vdots \\ \} \quad K-1 \text{ rows} \end{array} \quad (18.38)$$

18.3.7 Forecast Error

The forecast (18.36) depends on $\hat{\boldsymbol{\beta}}$, which is a function of the set of observations \mathcal{D}_K . Thus \mathbf{y}^{c*} is a random variable. It generates a *forecast error* or *prediction error* relative to the model (18.35) of

$$\boldsymbol{\epsilon}^{c*} = \mathbf{y}^{c*} - \mathbf{y}^c = g^{-1}(\mathbf{X}^c \hat{\boldsymbol{\beta}}) - [g^{-1}(\mathbf{X}^c \boldsymbol{\beta}) + \boldsymbol{\epsilon}^c] + [\mathbf{y}_{model}^c - \mathbf{y}^c]. \quad (18.39)$$

This may be rearranged as

$$\boldsymbol{\epsilon}^{c*} = \underbrace{[g^{-1}(\mathbf{X}^c \hat{\boldsymbol{\beta}}) - g^{-1}(\mathbf{X}^c \boldsymbol{\beta})]}_{\text{parameter error}} - \underbrace{\boldsymbol{\epsilon}^c}_{\text{process error}} + \underbrace{[\mathbf{y}_{model}^c - \mathbf{y}^c]}_{\text{model error}}, \quad (18.40)$$

thereby decomposing the forecast error into three components.

It has already been noted that the model (18.34) and its forecast form (18.35) are idealizations and unlikely to capture reality with full precision. The component of forecast error so arising is the *model error*, also sometimes called the *model specification error* or just the *specification error*.

Even if the model were to correspond fully to the real world, its calibration is affected by sampling error; $\hat{\boldsymbol{\beta}}$ is an estimate of $\boldsymbol{\beta}$ and differs from it by sampling error. This generates the *parameter error* component of forecast error.

Even if the model were correctly specified and the sample size were increased without limit so as to eliminate parameter error, the model (18.34) (or (18.35)) is inherently stochastic and the forecast \mathbf{y}^{c*} is merely an estimate of the mean of \mathbf{y}^c .

The actual observation y^c will also contain the random error (i.e., noise) ϵ^c , referred to as *process error*.

Parameter error and process error are independent because the former depends on only past observations and the latter on only future observations, and all observations are independently modeled with a GLM. It will be convenient, and perhaps not greatly inaccurate, to assume that both parameter error and process error are independent of model error.

If all three components of forecast error are stochastically independent, then (18.40) yields

$$\text{Var}[\epsilon^{c*}] = \text{Var}[\epsilon_{\text{model}}^{c*}] + \text{Var}[\epsilon_{\text{parameter}}^{c*}] + \text{Var}[\epsilon_{\text{process}}^{c*}], \quad (18.41)$$

where the three arguments on the right side are the model, parameter, and process errors from (18.40).

The quantity $\text{Var}[\epsilon^{c*}]$ is referred to as the *mean squared error of prediction* and is often written as $\text{MSEP}[\epsilon^{c*}]$. It is sometimes referred to as the forecast (or prediction) error, and its components referred to as model error, parameter error, and process error, respectively. Such usage leads to confusion with the earlier definition of these terms, and so the present chapter retains those earlier definitions, referring to the components of (18.41) as *model MSEP*, *parameter MSEP*, and *process MSEP*, respectively.

18.3.8 Estimation of Forecast Error

As illustrated in Sections 18.3.1 to 18.3.5, the link function of model (18.34) will usually be nonlinear, and hence $\hat{\beta}$ will be a nonlinear function of β . The forecast y^{c*} given by (18.36) then becomes a complicated function of the observations, and the estimation of MSEP is consequently difficult.

It is true that the Mack chain ladder model is analytically tractable to the extent that estimates of its model MSEP and parameter MSEP are obtainable analytically (Mack 1993). In almost all other cases, however, analytical estimates will not be available, and resort to bootstrap estimation will be necessary. The bootstrap has the advantage that it estimates the entire distribution of the loss reserve estimate, rather than just its second moment.

Suppose that the bootstrap involves N replications, with $\hat{\beta}_{(n)}$ denoting the estimate $\hat{\beta}$ for the n -th replication. Further, let $\epsilon_{(n)}^{c*}$ denote a drawing from the distribution of y^c as assumed in the GLM representation of y^c .

If a parametric bootstrap is in use, then $\epsilon_{(n)}^{c*}$ will be simply a drawing from the assumed distribution parameterized by $\hat{\beta}_{(n)}$, with the components of $\epsilon_{(n)}^{c*}$ drawn independently for each $n = 1, \dots, N$. If a nonparametric bootstrap is in use, then $\epsilon_{(n)}^{c*}$ will

be obtained by drawing an i.i.d. sample of standardized residuals from the vector of residuals associated with the estimate $\hat{\beta}$.

According to (18.36), the n -th replication of the forecast of \mathbf{y}^c is

$$\mathbf{y}_{(n)}^{c*} = g^{-1} \left(\mathbf{X}^c \hat{\beta}_{(n)} \right), \quad (18.42)$$

and the quantity

$$\tilde{\mathbf{y}}_{(n)}^{c*} = \mathbf{y}_{(n)}^{c*} + \boldsymbol{\epsilon}_{(n)}^{c*} = g^{-1} \left(\mathbf{X}^c \hat{\beta}_{(n)} \right) + \boldsymbol{\epsilon}_{(n)}^{c*} \quad (18.43)$$

is a simulated value of the future observation \mathbf{y}^c including process error, based on the n -th bootstrap replication. The set $\{\tilde{\mathbf{y}}_{(n)}^{c*}, n = 1, 2, \dots, N\}$ provides an empirical distribution of the forecast \mathbf{y}^{c*} from which all stochastic properties of the forecast may be estimated.

Consider, in particular, the variance of this empirical distribution:

$$\text{Var} [\tilde{\mathbf{y}}_{(n)}^{c*}] = \text{Var} \left[g^{-1} \left(\mathbf{X}^c \hat{\beta}_{(n)} \right) + \boldsymbol{\epsilon}_{(n)}^{c*} \right] = \text{Var} \left[g^{-1} \left(\mathbf{X}^c \hat{\beta}_{(n)} \right) \right] + \text{Var} [\boldsymbol{\epsilon}_{(n)}^{c*}], \quad (18.44)$$

where the last equality is obtained by recognizing that the two members on the right represent contributions from past data and future noise, respectively, and, as already mentioned, they are stochastically independent.

Comparison of (18.44) with (18.40) indicates that the two members on the right of the former equation are estimates of parameter MSEP and process MSEP, respectively. The bootstrap has thus provided estimates of two of the three components of total forecast MSEP. A numerical example can be found in chapter 11 of Taylor (2000).

The remaining component, model MSEP, is not obtainable from the bootstrap because it does not contemplate model designs other than the single one according to which the forecasts have been made (see (18.34)). The estimation of model error remains problematic, with little coverage in the existing literature, despite the fact that in many cases its magnitude will be highly material, possibly even exceeding the total of parameter and process MSEP. One substantial, albeit subjective, attempt to address the issue is found in O'Dowd, Smith, and Hardy (2005).

18.3.9 Models with Multiple Submodels

Consider the PPCC model introduced in Section 18.3.3. The models discussed there and in Section 18.3.4 are adequate for the forecast of future PPCCs. However, loss reserving requires the forecast of future claim payments, which take the form

$$y_{kj} = c_{kj} p_{kj}, \quad (18.45)$$

where c_{kj} denotes the number of claim closures (as previously) and p_{kj} denotes the observed average PPCC. The corresponding forecast is then

$$\hat{y}_{kj} = \hat{c}_{kj} \hat{p}_{kj}, \quad (18.46)$$

where \hat{c}_{kj} , \hat{p}_{kj} denote forecasts of the relevant quantities.

It is evident that the PPCC model consists in fact of two submodels, one of the numbers of closures and one of the PPCCs themselves. The prediction error of \hat{y}_{kj} will compound the prediction errors of \hat{c}_{kj} and \hat{p}_{kj} . Assessment of the introduction of submodels must therefore take account of the additional prediction error introduced by them.

It is sometimes argued that the PPCC model, with its two submodels, contains two sources of uncertainty and is therefore unreliable. However, this issue should be approached quantitatively by quantifying the uncertainty in each submodel and thus the total uncertainty of the forecasts of the combined model. It can happen that the dispersion of PPCCs is less than that of payments in a simpler model by a considerable margin and that the combination of submodel prediction errors arising from (18.46) is less than the prediction error of the simpler model based just on claim payments.

The PPCC model is not the only one consisting of multiple submodels. (Taylor (2000, section 4.4) also gives the example of the *projected case estimates model*.

18.4 Models of Individual Claims

18.4.1 Motivation

All modeling considered until here in this chapter has assumed that data are available in the form of the claims triangle illustrated by Figure 18.1. However, such data are highly summarized. The claims data held by an insurance company actually exist in unit record form (i.e., as one or more records specific to each claim lodged with the company). A typical format has the following elements:

- A *claims header file*. This consists of one record per claim, containing static information about the claim, such as the following:
 - Claimant information:
 - policy number;
 - name;
 - date of birth;
 - preinjury earnings (for lines of business providing income replacement).
 - Claim information:
 - date of accident;
 - date of notification;
 - peril generating the claim (e.g., for automobile insurance, collision, damage, theft);
 - type of injury (in lines of business providing bodily injury coverage).

- A *transaction file*. This consists of one record per claim transaction, containing information (often financial information) about that transaction, such as the following:
 - policy number;
 - date of transaction;
 - type of transaction, such as a claim payment or a change of claim status (e.g., from open to closed);
 - amount of claim payment (if there is one);
 - type of payment (e.g., income replacement, medical expense).

For modeling purposes, the claims header file, and transaction file are typically be merged into a single *claim data file*, containing one record per claim. That one record may contain details of many transactions.

The claims triangle of Figure 18.1 dates from an age in which modern data handling facilities were not available and there was a need to summarize data into a convenient form for analysis. To the extent that one is concerned with the analysis of claims experience, this motivation of claim triangles no longer exists, and the more natural dataset for loss reserving may often be the unit record claim data file just described (though note the comment on IBNR claims in Section 18.4.4).

The use of individual claim information may, of course, involve the manipulation of large datasets. For example, a portfolio with a 10-year history may contain hundreds of thousands of claims. These would provide hundreds of thousands of observations for analysis, with each observation accompanied by a possibly large number of explanatory variables. The claims triangle would summarize this data into just 55 observations, each with just two explanatory variables: accident year and development year. However, this compression of data is liable to eliminate much useful predictive information.

18.4.2 Predictive Models

With modern computing facilities there is no particular difficulty in performing analysis of the unit record claim file. The natural medium for this is regression and, in particular, for the same reasons as in Section 18.2.3, GLM regression. The formulation is just as in (18.34), i.e.,

$$\mathbf{y} = g^{-1}(\mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\epsilon},$$

where \mathbf{y} is now the vector of observations on individual claims.

This regression may take many forms. For example, y_i might denote the delay from accident to notification date, in which case the regression might be used to predict, for each past accident year, the number of future claims to be notified, or in other words, the number of incurred but not reported (IBNR) claims.

Alternatively, the dataset represented by y might comprise only *closed claims*, with y_i denoting the *total cost* of the i -th claim. In this case the regression could be used to forecast the ultimate sizes at closure of currently open claims. This example is evidently relevant to loss reserving, and some details of it are pursued next.

Example 18.1 (Australian Automobile Bodily Injury Scheme). Consider the claims triangles in Table 18.5 and Table 18.6. These are annual summaries of data that are collected quarterly that summarize selected data from the automobile bodily injury scheme of one Australian state. The scheme is tort-based but subject to some controls:

- For accident dates from October 2000 the amounts that could be awarded as plaintiff legal costs were subject to prescribed ranges that were related to the amounts of general damages awarded (e.g., no plaintiff costs could be awarded if the amount of general damages was less than \$25,000).
- For accident dates from December 2002 all injuries were awarded a point score under scheme guidelines, and this score then determined the amount of general damages to be awarded, diminishing judicial discretion.

The scheme began in October 1994, and so the tables cover almost its entire history to date. The injury associated with each claim is assigned one of six severities ranging from 1 (lowest) to 5 (catastrophic injury) and 6 (fatality).

Table 18.5 displays the average claim size of legally represented claims closed in each cell with injury severity 1 (the lowest severity). All claim sizes are adjusted to common dollar values according to the wage inflation index of the state concerned. Average claim sizes beyond development year 11 are not shown because they become erratic due to sampling error.

Although Table 18.5 presents average claim sizes in the conventional triangular form, this is for convenience of presentation only. Modeling is carried out in terms of individual claims. There are more than 125,000 claims.

Table 18.6 displays the OTs of the same subset of claims attained at the ends of the development years represented by the columns are not shown beyond development year 6 because they approach 100% and so do not inform the present discussion.

An earlier version of the same dataset appeared in Taylor and McGuire (2004), together with a detailed discussion of its modeling, and a further discussion was given by McGuire (2007).

The tables display three distinct and obvious characteristics that require incorporation in any model:

- (1) Claim size increases steadily with increasing OT.
- (2) There is a discontinuity in experience between accident years 2002 and 2003. Legislation that took effect in December 2002 made fundamental changes to the conditions

Table 18.5. Numerical Example: Average Sizes of Closed Claims (Inflation Adjusted)

Table 18.6. Numerical Example: Operational Times

Accident Year	Development Year					
	1 %	2 %	3 %	4 %	5 %	6 %
1995	7	32	59	80	91	96
1996	6	36	66	84	92	95
1997	5	42	71	86	92	96
1998	6	41	68	82	90	95
1999	9	43	68	84	92	96
2000	6	35	67	84	92	96
2001	4	36	64	84	92	96
2002	5	31	59	80	92	96
2003	1	16	45	76	90	95
2004	1	18	49	76	90	95
2005	2	20	56	81	91	95
2006	1	21	56	79	90	95
2007	2	22	54	77	90	
2008	2	22	57	81		
2009	2	23	61			
2010	1	26				
2011	2					

of the scheme, that eliminated many smaller claims. This elimination shifted the development patterns of both OTs and average claim size.

- (3) The changes discussed in (2) result in average claim sizes that increase over time. These sizes were already adjusted for inflation, so a question arises as to whether the increase amounts to SI. This is discussed later, but a period of very rapid increase is particularly noticeable over the shaded area of Table 18.5.

The model is a single GLM of average claim size with log link and comprises sub-models of the six different injury severities. The model also includes the following effects:

- OT (calculated separately for each severity);
- whether the claim was legally represented;
- SI;
- the 2002 legislation;
- some other lesser legislative change in 2000.

Noteworthy details of the model are as follows.

Operational time effect. As noted in Taylor and McGuire (2004), the logarithm of average claim size can be modeled as a linear function of OT over a substantial part

of its range (in fact, $0.2 \leq OT \leq 0.8$). That is, (18.34) for an individual observation reads

$$y_i = \exp\left(\dots \beta_{m_i}^{(OT)} X_{im_i}^{(OT)} \dots\right) + \epsilon_i, \quad (18.47)$$

with

$$X_{im_i}^{(OT)} = \max(0.2, \min(0.8, t_{im_i})), \quad (18.48)$$

where m_i is the severity of the i -th claim, t_{im_i} is the (severity) m_i OT of that claim at closure calculated on the basis of severity m_i claims only (i.e., only severity m_i claims included in the numerator and denominator of (18.25)), $\beta_m^{(OT)}$ is the rate of increase per unit of OT in log average claim size of severity m , and the dots indicate that there are many other terms in the regression.

Superimposed inflation. It is a feature of this scheme that SI is uniform over neither time nor injury severity. Claims experience exhibits periods of SI quiescence, punctuated by short and possibly sharp bursts of SI. These dynamics reflect the shifting ascendancy of claimants and their legal representatives, on the one hand, and the scheme's cost containment measures, on the other.

When SI does occur, its impact in terms of proportional increase tends to be greatest for smaller claims. Large claims, involving catastrophic injury, are subject to reasonably clear precedent and show relatively little increase in real terms over time. Smaller claims, in contrast, have headroom within which the boundaries of damages claimed can be tested.

Analysis of the claim sizes underlying Table 18.5 indicates SI of severity 1 claims over the two-year period 2006Q2 (the second quarter of 2006) to 2008Q1. The rate of SI decreased linearly with increasing OT, attaining a zero value at OT=100%. Thus (18.34) for an individual observation reads

$$y_i = \exp\left(\dots \beta_1^{(SI)} X_{i1}^{(SI)} \dots\right) + \epsilon_i, \quad (18.49)$$

with

$$X_{i1}^{(SI)} = (1 - t_{i1}) \max(q', \min(q_i, q'')), \quad (18.50)$$

and where q_i denotes the calendar quarter of closure of the i -th claim, q' , q'' denote 2006Q2 and 2008Q1, respectively, and $\beta_1^{(SI)}$ is the rate of SI for severity 1 claims at OT zero. Note that the contribution to the linear predictor appearing in (18.49) is constant for $q_i < q'$ and, $q_i > q''$, indicating no SI over these time intervals.

Claims of higher injury severities were differently affected by SI. Generally, the effects were less (in proportional terms), and severities 4 to 6 were not affected at all.

The 2002 legislation. Two major effects of the legislation are apparent. First, Table 18.6 shows a marked reduction in the rate of claim closure under the new

legislation, followed by a partial recovery. Second, Table 18.5 shows a marked reduction in the sizes of closed claims, followed by an increase.

It is not immediately obvious whether the changes in claim sizes are genuine or induced by the changed rates of closure. However, one can, for example, calculate a cumulative average size of claims closed in the first two development years of accident year 2002 as $1\% \times \$18,013 + (16\% - 1\%) \times \$34,404 = \$31,925$. This relates to OT 31%.

The corresponding calculation for the first three development years of accident year 2003 gives a cumulative average claim size of \$21,618 at OT 45%. This confirms a genuine reduction in claim sizes, because the data indicate that average claim size increases monotonically with increasing OT.

The reductions in both closure rates and claim sizes appear to narrow as one goes to later development years, suggesting that larger claims have been less affected by the legislative change.

Such calculations are tedious, and the legislative effect on claim sizes is much better quantified by the inclusion of an accident year effect in the GLM. The inclusion of an accident year effect does indeed confirm a confinement of the effect to smaller claims, which was indeed the intention of the legislation.

Further investigation of the data indicates several simultaneous effects of the legislation on claims experience:

- Some claims were eliminated from experience entirely. This was not as obvious as might be expected because claim frequency was declining anyway, both before and after 2003.
- Naturally, the eliminated claims were typically those that would have been small in the absence of the legislation. This elimination of claims occurred only for injury severity 1.
- These eliminations meant that the relation between OT and average claim size differed between pre-2003 and post-2002 accident years. Claims that would have been closed at OT t in the older accident years were closed at OT $h(t)$ in the later ones, where $h(\cdot)$ is a contraction distortion in the sense that $h(0) = 0$, $h(1) = 1$, $h(t) \leq t$ for $0 < t < 1$.
- The degree of claims elimination varied as the effect of the legislation matured. As a result the distortion function varied over the accident periods 2003 and later, before eventually stabilizing.
- There were genuine reductions in claim size superimposed on the above changes. These were observed in only the lower injury severities (but not only for severity 1). For the severities affected, the proportionate changes varied by severity and diminished with increasing OT, with zero reduction above some OT threshold. Thus (18.34) for an individual observation reads

$$y_i = \exp \left(\cdots \beta_{m_i}^{(Lh)} X_{im_i}^{(Lh)} \cdots \right) + \epsilon_i, \quad (18.51)$$

with

$$X_{im_i}^{(Lh)} = \max [0, h_{k_i, m_i}(T_{m_i}) - h_{k_i, m_i}(t_{im_i})], \quad (18.52)$$

and where k_i denotes the accident quarter to which the i -th claim belongs, h_{k_i, m_i} the distortion function for severity m_i in accident quarter k_i (= identity function for $m_i \neq 1$), and T_{m_i} the OT threshold for severity m_i .

Discussion. The model just described contains 73 parameters. It is unlikely that the information contained in the 17×17 triangles of which Table 18.5 and Table 18.6 are excerpts (153 observations per triangle) would be sufficient for the reliable estimation of all parameters. Possibly the information would be insufficient even for the formulation of some of the effects described.

This situation would be considerably eased by the use of quarterly data, which are available because insurers lodge data quarterly. This would convert the 17×17 triangles to 68×68 triangles (2,346 observations per triangle). Even in this case, however, fine gradations of OT would be lacking. As can be seen from Table 18.6, OT increases by 20–30% per annum in the early development years, which implies increases of about 5–7% per quarter. The modeling described earlier was based on 2% divisions of OT.

In short, the use of individual claim data (>125,000 observations per “triangle”) creates a richness of information that facilitates both the detection of data features and their subsequent modeling.

18.4.3 Stochastic Case Estimation

Stochastic case estimation (or *stochastic case estimate*, as the context requires) (SCE) consists of the formulation of the estimated ultimate claim cost of individual claims on the basis of a predictive model applied to information specific to those claims. In other words, it is a mapping, $S : I_i \rightarrow c_i$, where I_i denotes the total information in respect of the i -th claim and c_i the estimated ultimate claim cost of that claim.

The predictive model is stochastic. By definition, the model underlying SCE is an individual claims model. In fact, there is nothing in the definition that is any more specific than the individual claims models discussed in Section 18.4. However, the name SCE is usually reserved for models that employ detailed data specific to the individual claim.

The SCEs are intended to function as case estimates (i.e., as substitutes for “manual” or “physical” case estimates, estimates of ultimate cost made by claims experts after consideration of all the facts and circumstances of the claims).

The more data specific to a claim used as input to the SCE (subject to significance and the generally sound construction of the predictive model), the more accurate those

estimates can be made and the greater will be the potential dispersion of the SCEs, taken over the whole body of claims. An SCE that predicts the same claim size for every claim is of little value.

Examples of these SCE models are found in Brookes and Prevett (2004), Taylor and Campbell (2002), and Taylor, McGuire, and Sullivan (2006). The first two of these relate to Australian workers' compensation insurance and the last to U.S. medical malpractice.

Taylor et al. (2006) classify explanatory variables into a number of categories:

- *Static variables*: constant over the life of a claim (e.g., date of birth of claimant);
- *Dynamic variables*: liable to change over the life of a claim, which consist of two subcategories:
 1. *Time variables*: directly related to just the passage of time, whose future values are therefore fully predictable (e.g., development period);
 2. *Unpredictable variables*: future changes not predictable with certainty (e.g., development period of claim closure).

The last category, unpredictable variables, requires special treatment. If they are included in a model, then any forecast of the model will require forecasts of these explanatory variables. The advantages and disadvantages of the inclusion of unpredictable variables were discussed in relation to the timing of claim closures in Section 18.3.9.

Example 18.2 (Australian Workers' Compensation Insurance). Workers' compensation claims are characterized by periods of *incapacity*, during which the insurance provides income maintenance, and spells of *activity*, when income maintenance is not required.

In Taylor and Campbell (2002), for example, income maintenance is modeled on the basis of a stochastic process of claim status (incapacitated or active). A spell of incapacity is modeled as a survival process in which the hazard is recovery to an active status. Likewise, a spell of activity is modeled as a survival process in which the hazard is relapse to incapacity.

The hazard rates of the survival processes depend on many explanatory variables specific to the claim. Brookes and Prevett (2004) categorize these as follows:

- Claimant characteristics: age, gender, occupation, marital and dependant status, wage rate, etc.
- Employer characteristics: industry, wages, location, etc.
- Claim status: claim is open/closed/reopened/disputed, work status, etc.
- Claim characteristics: injury nature, location, etc.
- Claim history: payments and rates of payment, time lost, etc.

The hazard rates of Taylor and Campbell (2002) are modeled by means of *survival analysis* (Kalbfleisch and Prentice 1980), specifically *proportional hazards* survival models in which the probability of survival of the i -th claim to time $t + dt$, conditional on survival to t , is, $1 - \nu_i(t) dt$ where the *hazard rate* $\nu_i(t)$ takes the form

$$\nu_i(t) = \nu_0(t) \exp(x_i^T \boldsymbol{\beta}). \quad (18.53)$$

Here $\nu_0(t)$ is independent of i and is referred to as the *baseline hazard function*, $\boldsymbol{\beta}$ is a vector of parameters also independent of i , x_i is a vector of explanatory variables with values specific to the i -th claim, and the upper T denotes transposition. Fitting of the model requires estimates of $\nu_0(t)$ and $\boldsymbol{\beta}$.

As an example of (18.53), survival might represent continuation of incapacity (survival of the incapacitated status), in which case the hazard is recovery and the hazard rate $\nu_i(t) dt$ is the i -th claim's probability of recovery to active status over an infinitesimal period after a duration t of continuous incapacity.

The baseline $\nu_0(t)$ may take a specified parametric form or may be left unspecified and its form estimated from the data. In the latter case the model becomes semi-parametric and is referred to as a Cox regression model (Cox 1972).

Workers' compensation provides benefits other than income replacement, such as indemnity for the cost of medical treatment. There may be dependency between payments of different types. For example, continued payment of income replacement usually requires periodic proof of continuing incapacity, so that higher medical costs are more likely to be associated with periods of incapacity than with periods of activity. The SCE model of Taylor and Campbell (2002) includes details of all such payment types and the dependencies between them.

Example 18.3 (U.S. Medical Malpractice Insurance). U.S. medical malpractice is a form of tort-based liability insurance. Completion of a claim therefore results from either a court verdict or a negotiated settlement between the parties. Although a claim certainly does not commonly involve just a single payment, it does tend to be subject to a high concentration of the claim cost in a small number of payments in the vicinity of the date of closure.

It is therefore a reasonable approximation to model individual claims as involving a single payment of the total claim cost on or about the closure date. The survival issues associated with periodic payments do not arise, and claim size may be modeled as the response variate of a GLM.

As in Section 18.4.2, the GLM takes the form (18.34). However, as indicated by Section 18.4.1, the explanatory variables present in design matrix X are relatively large in number and include detailed information on the individual claims. The case considered by Taylor et al. (2006) was the special one in which case estimates were included as an unpredictable dynamic variable.

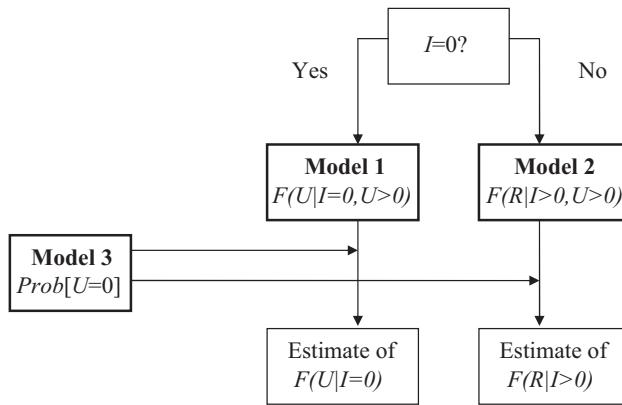


Fig. 18.2. Structure of case estimation forecast model.

The explanatory variables included in the model were as follows:

- Static variables:
 - Specialty of practice;
 - Geographic region of practice;
 - Policy limit of coverage;
 - Whether coverage of practitioner is combined with coverage of hospital attended by practitioner;
 - Whether coverage includes a corporate layer covering vicarious liability;
 - Quarter of occurrence;
 - Delay from occurrence to notification.
- Dynamic variables:
 - Time variables:
 - Development quarter;
 - Unpredictable:
 - Year of claim closure;
 - Operational time at closure;
 - Case estimate (development quarter by development quarter).

The model contained a number of features of note. First, its main thrust was to model ultimate claim cost as a multiple of the case estimate held at the valuation date. However, this required recognition of the special cases of either nil case estimate or nil ultimate claim cost. Figure 18.2, which reproduces Figure 4.1 from Taylor et al. (2006), sets out the structure of the model, involving three submodels. Here the symbols I and U denote the case estimate and the actual amount, respectively, of a claim's ultimate incurred cost, $R = U/I$, and F denotes some distribution (different for Models 1 and 2).

The dataset includes a history of case estimates for each of its claims, so that each closed claim generates multiple observations of the development of the case estimate from one epoch to the ultimate claim cost. These, however, are unlikely to be stochastically independent, as assumed by a GLM, and so the dataset modeled includes just a single observation from each claim, randomly selected from those available.

18.4.4 IBNR Claims

A loss reserve is intended to cover the unpaid liability in respect of all claims that occurred on or before the valuation date, whether or not the insurer has been notified of them. These claims are referred to as *incurred claims* at the valuation date. Those of the incurred claims *not* notified to the insurer by the valuation date are referred to as *incurred but not reported*, commonly abbreviated as *IBNR*.

Many loss reserving models provide estimates of liability including IBNR claims. All of the models discussed in this chapter other than SCE are of this type. It is in the nature of the SCE systems described in Section 18.4.3, however, that they generate individual estimates of ultimate liability only in respect of those claims known to the insurer at the valuation date. The IBNR liability requires separate estimation.

One may begin with the simple observation that

$$\text{IBNR liability} = \text{number of IBNR claims} \times \text{average size of IBNR claims}$$

Decompose this by accident year and write it symbolically as

$$i_k = n_k a_k. \quad (18.54)$$

The ultimate number of claims m_k notified in respect of each accident year, and hence the number of IBNR claims, is often estimated by means of the chain ladder (Section 18.2) with y_{kj} denoting the number of notifications in cell (k, j) or some variant of it. If m_k, n_k are estimated by \hat{m}_k, \hat{n}_k respectively, then $\hat{m}_k = x_{k, K-k+1} + \hat{n}_k$.

Various approaches to the estimation of the average claim size in (18.55) are possible and a couple are discussed next.

18.4.4.1 Estimation by Reference to Aggregate Models

One simple possibility consists of the application of an aggregate model to form an estimate \hat{u}_k of the ultimate claims cost u_k in respect of each accident year k . This might, for example, be one of the models discussed in Sections 18.2 and 18.3. Then, a_k is estimated by

$$\hat{a}_k = \frac{\hat{u}_k}{\hat{m}_k}. \quad (18.55)$$

The estimate of u_k can now be discarded and only \hat{n}_k, \hat{a}_k retained for insertion in (18.55). This approach has the advantage of simplicity, but may involve an implicit assumption that the average size of IBNR claims is the same as for claims already notified. This may or may not be true.

18.4.4.2 Estimation by Reference to Individual Models

The possibly unwarranted assumption just referred to may be eliminated by a more detailed approach, but at the cost of an increase in complexity.

Let the number of IBNR claims in respect of accident year k be estimated by \hat{n}_k just as earlier, but now assume that its components $\hat{y}_{kj}, j = K - k + 2, \dots, J$ are available. For each $(k, j) \in \mathcal{D}_K^c$, draw a random sample of N_{kj} claims from past accident years that were notified in development year j .

The number of past accident years would need to be sufficiently large to yield a sample of at least N_{kj} . Subject to this, more recent accident years would usually be preferred over earlier ones as being more relevant to forecasts.

Let $s_{kji}, i = 1, 2, \dots, N_{kj}$ denote the ultimate claim size of the i -th of these N_{kj} claims, estimated by \hat{s}_{kji} , comprising all claim payments up to the valuation date and the SCE of any remaining liability. Some of these claims may have been closed so that their sizes may be treated as known with certainty ($\hat{s}_{kji} = s_{kji}$), provided that the likelihood of their reopening is small.

The average size of claims of accident year k notified in development year j may then be estimated by

$$\hat{a}_{kj} = \frac{\sum_{i=1}^{N_{kj}} \hat{s}_{kji}}{N_{kj}}, \quad (18.56)$$

and the quantity i_k in (18.55) is estimated by

$$\hat{i}_k = \sum_{j=K-k+2}^J \hat{y}_{kj} \hat{a}_{kj}. \quad (18.57)$$

References

- Brookes, R. and M. Prevett (2004). Statistical case estimation modelling – an overview of the NSW WorkCover model. *Institute of Actuaries of Australia Xth Accident compensation Seminar*. <http://www.actuaries.asn.au/Library/Accident\%20Compensation\%20Seminar\%20SCE\%20Paper\%20RGB\%2022Nov04\%20final.pdf>.
- Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, B* 34, 187–220.
- De Jong, P. and B. Zehnwirth (1983). Claims reserving state space models and the Kalman filter. *Journal of the Institute of Actuaries* 110, 157–181.

- England, P. D. and R. J. Verrall (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal* 8, 443–518.
- Frees, E. and P. Wang (1990). A stochastic method for claims reserving in general insurance. *Journal of the Institute of Actuaries* 117, 677–731.
- Hachemeister, C. A. and J. N. Stanard (1975). IBNR claims count estimation with static lag functions. Spring Meeting of the Casualty Actuarial Society.
- Harnek, R. F. (1966). Formula loss reserves. *Insurance Accounting and Statistical Association Proceedings*.
- Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Chapman and Hall, London.
- Kalbfleisch, J. D. and R. L. Prentice (1980). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York.
- Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin* 23, 213–225.
- McGuire, G. (2007). Individual claim modelling of CTP data. *Institute of Actuaries of Australia XIth Accident Compensation Seminar*, Melbourne, Australia. http://www.actuaries.asn.au/Library/6.a\ACS07_paper\McGuire\Individual\%20claim\%20modellingof\%20CTP\%20data.pdf.
- O'Dowd, C., A. Smith, and P. Hardy (2005). A framework for estimating uncertainty in insurance claims cost. *Proceedings of the 15th General Insurance Seminar*. Institute of Actuaries of Australia. <http://www.actuaries.asn.au/Library/gipaper\odowd-smith-hardy0510.pdf>.
- Reid, D. H. (1978). Claims reserves in general insurance. *Journal of the Institute of Actuaries* 105, 211–296.
- Taylor, G. C. (1986). *Claim Reserving in Non-Life Insurance*. North-Holland, Amsterdam.
- Taylor, G. C. (2000). *Loss Reserving – An Actuarial Perspective*. Kluwer Academic Publishers, Boston.
- Taylor, G. C. (2011). Maximum likelihood and estimation efficiency of the chain ladder. *ASTIN Bulletin* 41(1), 131–155.
- Taylor, G. and M. Campbell (2002). Statistical case estimation. *Research Paper No 104 of the Centre for Actuarial Studies, University of Melbourne*. <http://www.economics.unimelb.edu.au/ACT/html/no104.pdf>.
- Taylor, G. and G. McGuire (2004). Loss reserving with GLMs: a case study. *Casualty Actuarial Society 2004 Discussion Paper Program*, 327–392.
- Taylor, G., G. McGuire, and J. Sullivan (2006). Individual claim loss reserving conditioned by case estimates. *Research Paper Commissioned by the Institute of Actuaries*. <http://www.actuaries.org.uk/research-and-resources/documents/individual-claim-loss-reserving-conditionedcase-estimates>.
- Verrall, R. J. (1996). Claims reserving and generalised additive models. *Insurance: Mathematics and Economics* 19(1), 31–43.
- Verrall, R. J. (2000). An investigation into stochastic claims reserving models and the chain-ladder technique. *Insurance: Mathematics and Economics* 26, 91–99.
- Wüthrich, M. V. and M. Merz (2008). *Stochastic Claims Reserving Methods in Insurance*. John Wiley & Sons Ltd, Chichester, UK.

19

Survival Models

Jim Robinson

Chapter Preview. Survival modeling focuses on the estimation of failure time distributions from observed data. Failure time random variables are defined on the non-negative real numbers and might represent time to death, time to policy termination, or hospital length of stay. There are two defining aspects to survival modeling. First, it is not unusual to encounter distributions incorporating both parametric and nonparametric components, as is seen with proportional hazard models. Second, the estimation techniques accommodate incomplete data (i.e., data that are only observed for a portion of the time exposed as a result of censoring or truncation). In this chapter, we apply R's survival modeling objects and methods to complete and incomplete data to estimate the distributional characteristics of the underlying failure time process. We explore parametric, nonparametric, and semi-parametric models; isolate the impact of fixed and time-varying covariates; and analyze model residuals.

19.1 Survival Distribution Notation

Frees (2010) provides an excellent summary of survival model basics. This chapter adopts the same notation.

Let y denote the failure time random variable defined on the non-negative real numbers. The distribution of y can be specified by any of the following functions:

- $f(t)$ = the density of y
- $F(t) = \Pr(y \leq t)$, the cumulative distribution of y
- $S(t) = \Pr(y > t) = 1 - F(t)$, the survival function
- $h(t) = f(t)/S(t)$, the hazard function
- $H(t) = -\ln(S(t))$, the cumulative hazard function

So, the distribution of y can be defined in terms of the density function, the survival function, or the hazard function. These functions may be continuous, discrete, or a mixture. They may also be parametric, nonparametric, or a mixture.

19.2 Survival Data Censoring and Truncation

The observed values of y are typically incomplete in some respect. If the final value of y is not observed by the end of the observation period, the observation is right-censored. If the subject is observed only if the value of y exceeds some value, then it is left-truncated. If the observations are independent, then each makes the following contribution to the likelihood function:

- If the subject is left-truncated at C_L and observed to fail at y , then use $f(y)/S(C_L)$ or $h(y)S(y)/S(C_L)$.
- If the subject is left-truncated at C_L and right-censored at C_U , then use $S(C_U)/S(C_L)$.
- If there is no left-truncation, then set C_L to zero.

These three cases can be combined if we define $\delta = I(y \geq C_U)$, an indicator that the observation is right-censored, and $t = \min(y, C_U)$, the ending duration of the observation's exposure. Then, the contribution of the observation, (δ, t) , to the likelihood function is $h(t)^{1-\delta} S(t)/S(C_L)$. The hazard rate is included if failure is observed. The ratio of survival functions is included for all observed subjects. Note that this common term is an exponential function of the accumulated hazard from entry to termination of the subject's exposure. This assumes that C_L and C_U are either nonrandom values (which may differ for each subject) or are random values with distributions independent of y (i.e., non-informative).

In the next section, we introduce an example of failure time data exhibiting both non-informative right-censoring and left-truncation.

19.3 National Nursing Home Survey

In this chapter, we demonstrate survival modeling techniques using data from the National Nursing Home Survey (NNHS). The random variable of interest is the length of stay in a nursing home, from admission to discharge. The 1999 NNHS is a stratified random sample of U.S. nursing homes and, within that, a random sample of current residents and residents discharged in the year prior to the survey date. The current and discharged resident samples can be combined to serve as a right-censored and left-truncated sample of nursing home stays. The surveys also include resident information such as gender, age, marital status, type of care required, and payer, as well as facility characteristics such as type of care provided, urban vs. rural location, and ownership type, which can serve as candidate explanatory factors in models for discharge rates. In addition, the discharge survey includes the discharge status of the resident (i.e., deceased, discharged to community, discharged to another nursing home).¹

¹ Two aspects of the 1999 NNHS have been simplified to provide tractable examples of survival modeling. First, the stratified sampling structure of the NNHS implies that each resident has a different probability of being sampled. The survey provides weights inversely proportional to the sampling probability for the resident. To eliminate the need to work with weighted observations, we extracted a new sample of records from the NNHS current and

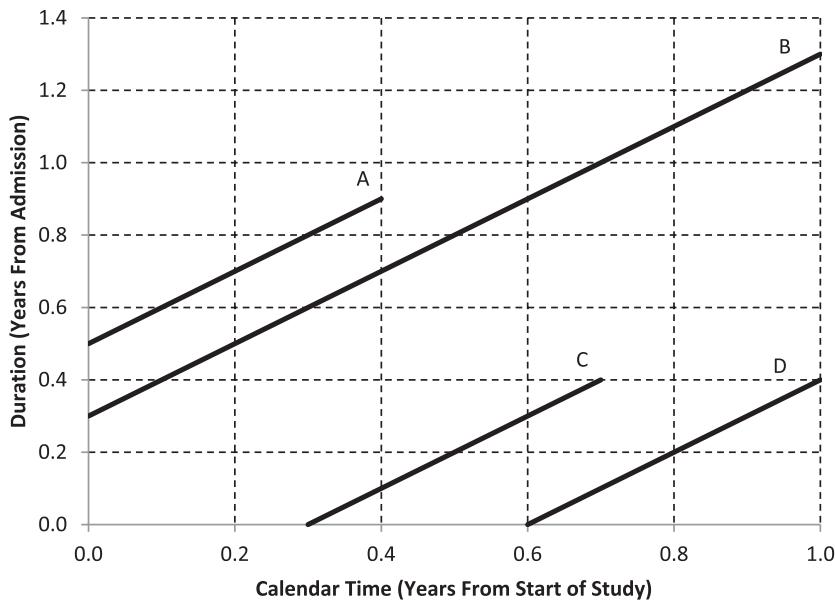


Fig. 19.1. Population diagram.

All current resident observations are right-censored at the nursing home duration recorded at the end of the one-year survey period. Current residents who are more than one year since admission are also left-truncated at the censoring duration minus 365 days. Current residents who are within one year of admission are not left-truncated. Discharged residents are not right-censored. They are, however, left-truncated if the simulated entry duration is greater than zero. The population diagram in Figure 19.1 summarizes the survey observation period and four possible censoring/truncation situations.

- At the start of the study year, Resident A is 0.5 years from admission and remains in the nursing home for an additional 0.4 years. This is an example of a left-truncated uncensored observation. The contribution to the likelihood function should be $h(0.9)S(0.9)/S(0.5)$.
- Resident B enters the study at duration 0.3 years and remains in the nursing home throughout the study year. This is an example of a left-truncated right-censored observation with a likelihood contribution equal to $S(1.3)/S(0.3)$.

discharged resident files in which the probability of selecting a record is proportional to the survey weight. The resulting resampled records can be considered equally weighted in that each record represents the same number of nursing home residents (or discharges) in the general population. Second, the discharge records in the NNHS do not indicate the date of discharge in the year leading up to the survey. The duration since admission is provided (in days), but we do not know when each discharged resident entered the one-year study period. All we know is that the duration at which each resident's exposure started was within 365 days prior to the recorded duration at discharge. To create a more conventional left-truncation example for this chapter, we randomly generated a starting duration for each discharge record from a uniform distribution on the interval from the greater of zero and the discharge duration minus 365 days to the discharge duration. Building estimation techniques appropriate for the original NNHS data (i.e., for weighted samples with interval-censored entry durations) from basic principles, would unnecessarily complicate the presentation of survival modeling concepts.

- Resident C is admitted to the nursing home 0.3 years after the start of the study and is discharged 0.4 years later. This is a complete (uncensored without truncation) observation with a likelihood contribution of $h(0.4)S(0.4)$.
- Resident D is admitted 0.6 years into the study and remains through the remainder of the study period. This is a nontruncated right-censored observation with a likelihood contribution of $S(0.4)$.

Throughout the remainder of this chapter, we use R to fit and assess a variety of survival models to this nursing home length of stay data. The resampled NNHS data were assembled using SAS and imported into an R data object named nnhs using the following code.

```
> library(foreign)
> nnhs<-read.xport("transport")
```

Next, we loaded R's OISurv object library, set selected NNHS variables as categorical using R's factor method, and created filtering variables for later use.

```
> library(OISurv)
> attach(nnhs)

> nnhs$CERT<-factor(CERT)
> nnhs$FACID<-factor(FACID)
> nnhs$RESID<-factor(RESID)
> nnhs$LIVLOC<-factor(LIVLOC)
> nnhs$LIVWITH<-factor(LIVWITH)
> nnhs$LOC<-factor(LOC)
> nnhs$MARRIED<-factor(MARRIED)
> nnhs$MSA<-factor(MSA)
> nnhs$OWNER<-factor(OWNER)
> nnhs$PAYOR<-factor(PAYOR)
> nnhs$SEX<-factor(SEX)

> smpl<- (runif(length(nnhs$ENTRY))<0.01)
> nnhs$SMPL<-smpl
> nnhs$ADM<- (ENTRY==0)
> nnhs$T18<- (PAYOR==4)

> detach(nnhs)
> attach(nnhs)
```

SMPL is a vector of randomly generated Boolean values used to select a 1% sample of records to demonstrate modeling results for a small sample of observations.

ADM is another vector of indicators for observations that are observed from admission (i.e., records that are not left-truncated).

T18 is a vector of indicators for observations where the subject was a Medicare resident before discharge or on the survey date.

R's `Surv` method creates a survival data object formatted for use by other survival modeling methods. The code below creates the survival data object `nnhs_surv`, using variables `ENTRY`, `DAYS`, and `DISCHG`.

```
> nnhs_surv<-Surv(ENTRY,DAYS,(DISCHG!=0))
```

`ENTRY` is the duration at which exposure starts. `DAYS` is the duration at which exposure ends, either as an observed discharge or by censoring on the current resident survey date. `DISCHG` is nonzero if the resident is discharged; otherwise it is zero if the resident is still in the facility on the survey date (censored). The survival data object displays each observation in the format `(start,end]` for a discharge or `(start,end+]` for a censored subject. The code below displays the first 100 observations in this format.

```
> nnhs_surv[1:100,]
 [1] ( 2,16 ] ( 2,16 ] ( 2,16 ] ( 2,16 ] ( 2,16 ] ( 2,16 ]
 [7] ( 2,16 ] ( 2,16 ] ( 2,16 ] ( 0, 3 ] ( 0, 3 ] ( 0, 3 ]
[13] ( 0, 3 ] ( 0, 3 ] ( 0, 3 ] ( 0, 3 ] ( 0,14 ] ( 0,14 ]
[19] ( 0,14 ] ( 0,14 ] ( 0,14 ] ( 0,14 ] ( 0,14 ] ( 0,14 ]
[25] ( 0,14 ] ( 0,14 ] ( 0, 5 ] ( 0, 5 ] ( 0, 5 ] ( 0, 5 ]
[31] ( 0, 5 ] ( 0, 5 ] ( 0, 5 ] (17,32 ] (17,32 ] (17,32 ]
[37] (17,32 ] (17,32 ] (17,32 ] ( 0,13 ] ( 0,13 ] ( 0,13 ]
...
...
```

We see that only 2 of the first 100 observations are right-censored. Applying the `summary` method, we obtain some quick sample statistics.

```
> summary(nnhs_surv)
      start          stop         status
Min.   : 0.0   Min.   : 1.0   Min.   :0.0000
1st Qu.: 0.0   1st Qu.: 20.0  1st Qu.:0.0000
Median : 0.0   Median : 102.0 Median :1.0000
Mean   : 373.4 Mean   : 519.6 Mean   :0.6053
3rd Qu.: 304.0 3rd Qu.: 607.0 3rd Qu.:1.0000
Max.   :11695.0 Max.   :12060.0 Max.   :1.0000
```

We see that roughly 61% of the observations are discharges and 39% are censored. We also note that more than half of the subjects are observed from admission.

19.4 Nonparametric Estimation of the Survival Function

It can be useful to calibrate likelihood methods with nonparametric methods that do not rely on a parametric form of the distribution. The product-limit estimator due to Kaplan and Meier (1958) is a well-known estimator of the distribution in the presence of censoring.

To introduce this estimator, we consider the case of right-censored data. Let $t_1 < \dots < t_c$ be distinct time points at which an event of interest occurs and let d_j be the number of events at time point t_j . Further, the corresponding "risk set" is the number of observations that are active at an instant just prior to t_j . Using notation, the risk set is $R_j = \sum_{i=1}^n I(y_i \geq t_j)$.

With this notation, the product-limit estimator of the survival function is

$$\hat{S}(t) = \begin{cases} 1 & t < t_1 \\ \prod_{t_j \leq t} \left(1 - \frac{d_j}{R_j}\right) & t \geq t_1 \end{cases}$$

Greenwood (1926) derived the formula for the variance of the product-limit estimator:

$$\widehat{\text{Var}}(\hat{S}(t)) = (\hat{S}(t))^2 \sum_{t_j \leq t} \frac{d_j}{R_j(R_j - d_j)}.$$

R's `survfit` method takes a survival data object and creates a new object containing the Kaplan-Meier estimate of the survival function along with confidence intervals. The code below demonstrates this process using NNHS subjects observed from admission. Note that the `Surv` method is applied without the `ENTRY` field when the observations are not left-truncated. The `~1` tells the method to fit a single combined survival function. The subsequent `print` method provides summary statistics for the fitted survival curve.

```
> nnhs_KM<-survfit(Surv(DAYS, (DISCHG!=0)) ~ 1, subset=ADM, data=nnhs)
> print(nnhs_KM,rmean="common")
Call: survfit(formula = Surv(DAYS, (DISCHG != 0)) ~ 1,
  data = nnhs, subset = ADM)

      records      n.max      n.start      events      *rmean
    9016.00    9016.00    9016.00    6474.00    103.95

  *se(rmean)      median      0.95LCL      0.95UCL
    1.56        30.00       29.00       32.00
  * restricted mean with upper limit = 365
```

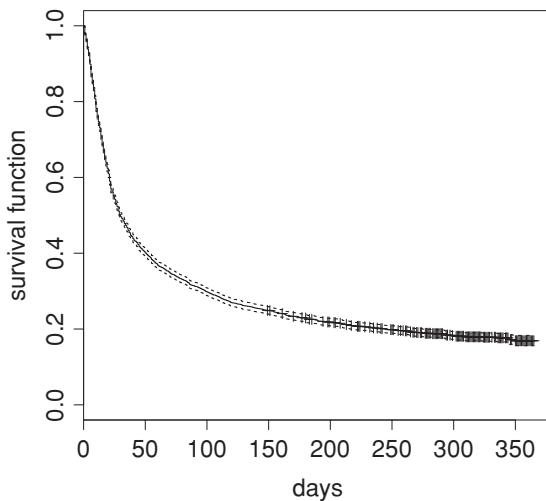


Fig. 19.2. Kaplan-Meier estimate – admissions only.

The results include the restricted mean of 104 days (i.e., the area under the fitted survival curve from admission through 365 days). The median length of stay is 30 days, significantly less than the mean and consistent with a skewed distribution.

The `plot` method is used to graph the fitted survival function, including the 95% confidence interval at each duration.

```
> plot(nnhs_KM, main="Kaplan-Meier estimate - Admissions only",
       xlab="days", ylab="survival function")
```

The resulting graph in Figure 19.2 indicates that about 17% of admissions remain in the nursing home for at least one year. We next expand the sample to include left-truncated residents.

```
> nnhs_KM<-survfit(Surv(ENTRY,DAYS, (DISCHG!=0)) ~ 1, data=nnhs)
> print(nnhs_KM,rmean="common")
Call: survfit(formula = Surv(ENTRY, DAYS, (DISCHG != 0)) ~ 1,
  data = nnhs)

  records      n.max      n.start      events      *rmean
 15127.00    9016.00      0.00    9157.00    229.09

*se(rmean)      median      0.95LCL      0.95UCL
      5.59      30.00      29.00      32.00
* restricted mean with upper limit = 12060
```

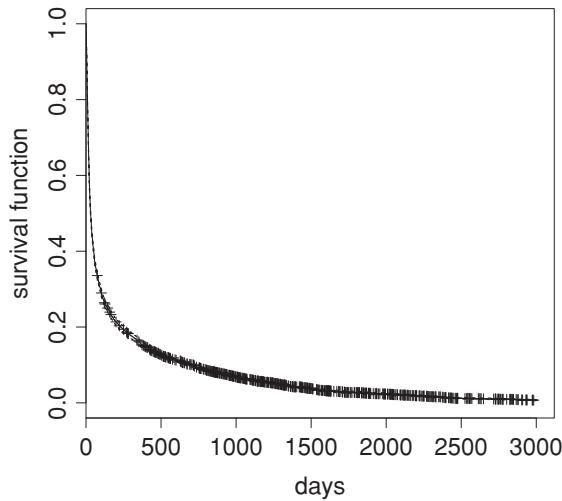


Fig. 19.3. Kaplan-Meier estimate – full data.

The restricted mean is now 229 days, representing the area under the survival curve from admission to 12,060 days (the largest observed duration). The median remains at 30 days, because much of the additional left-truncated information relates to resident survival beyond the first year. Figure 19.3 displays the fitted survival curve through 3,000 days, along with a 95% confidence interval.

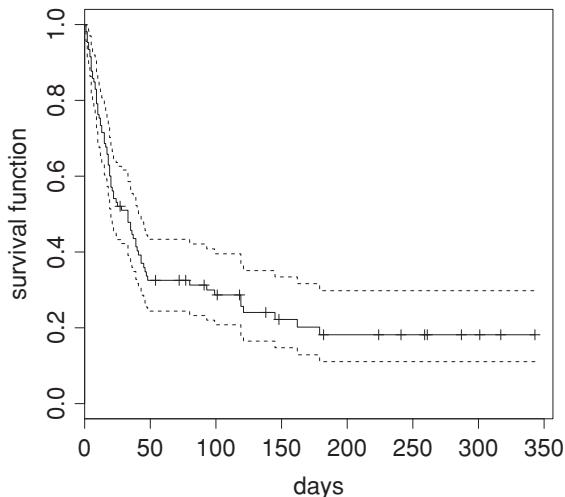


Fig. 19.4. Kaplan-Meier estimate – admission sample.

```
> plot(nnhs_KM, main="Kaplan-Meier estimate - Full data",
      xmax=3000,
      xlab="days", ylab="survival function")
```

We next fit the survival curve to a 1% sample of those observed from admission.

```
> nnhs_KM<-survfit(Surv(DAYS, (DISCHG!=0)) ~ 1, subset=SMPL&ADM,
  data=nnhs,
  conf.type="log", conf.int=0.95)
> print(nnhs_KM,rmean="common")
Call: survfit(formula = Surv(DAYS, (DISCHG != 0)) ~ 1, data = nnhs,
subset = SMPL & ADM, conf.type = "log", conf.int = 0.95)

  records      n.max      n.start      events      *rmean
    106.0      106.0      106.0       77.0      94.2

  *se(rmean)      median      0.95LCL      0.95UCL
    13.5        33.0        20.0        43.0
  * restricted mean with upper limit =  343

> plot(nnhs_KM, main="Kaplan-Meier estimate - Admission Sample",
      xlab="days", xmax=400, ylab="survival function", conf.int=TRUE)
```

In Figure 19.4 we see that, although the confidence interval has significantly increased, the estimated survival curve and restricted mean are quite similar to their counterparts from the complete sample.

19.4.1 Kaplan-Meier, Nelson-Aalen, and Handling Ties

The Kaplan-Meier method (`type='kaplan-meier'`) is used by default to construct an estimate of the survival curve. The resulting discrete survival function has point masses at the observed event durations (discharge dates), where the probability of an event given survival to that duration is estimated as the number of observed events at the duration divided by the number of subjects exposed or 1 “at risk” just prior to the event duration. Two alternate types of estimation are also available for the `survfit` method. The first (`type='fleming-harrington'`) uses the Nelson-Aalen (see

Aalen 1978) estimate of the cumulative hazard function to obtain an estimate of the survival function. The estimated cumulative hazard starts at zero and is incremented at each observed event duration by the number of events divided by the number at risk. The second alternative (`type='fh2'`) handles ties, in essence, by assuming that multiple events at the same duration occur in some arbitrary order. So, if 3 of 10 nursing home residents are discharged at the same duration, rather than increment the cumulative hazard by $3/10$, we increment by $1/10 + 1/9 + 1/8$ at that duration. Note that when this ordering of tied results is applied to the Kaplan-Meier estimate of the survival function it has no effect. In that case, we multiply the survival function by $7/10$ (allowing for ties) or by $9/10 \times 8/9 \times 7/8 = 7/10$ (arbitrarily ordering the tied discharges). The following code and Figure 19.5 plot the estimated cumulative hazard function for all three options.

```
> nnhs_KM <- survfit(Surv(DAYS, (DISCHG!=0)) ~ 1, subset=(SMPL&ADM),
  data=nnhs, type="kaplan-meier")
> nnhs_NA <- survfit(Surv(DAYS, (DISCHG!=0)) ~ 1, subset=(SMPL&ADM),
  data=nnhs, type="fleming-harrington")
> nnhs_FH2<- survfit(Surv(DAYS, (DISCHG!=0)) ~ 1, subset=(SMPL&ADM),
  data=nnhs, type="fh2")
> plot(nnhs_KM, conf.int=TRUE, col=2, mark.time=FALSE, xmax=365,
  fun="cumhaz", main="Product-Limit Variations - Admission
  Sample",
  xlab="Days", ylab="Cumulative Hazard")
> lines(nnhs_NA, mark.time=FALSE, lty=2, xmax=365, fun="cumhaz")
> lines(nnhs_FH2, mark.time=FALSE, lty=3, xmax=365, fun="cumhaz")
> legend("bottomright",c("Kaplan-Meier","Nelson-Aalen",
  "Fleming-Harrington"),lty=1:3, col=c(2,1,1))
```

Even with the small sample, we can see in Figure 19.5 that the three options yield similar estimates. Note that the plot also displays the 95% confidence interval for the Kaplan-Meier estimate based on Greenwood's estimate of the survival variance, taking into account the log transformation needed to obtain the cumulative hazard function. This confidence interval is a pointwise estimate specific to each duration along the curve. We cannot say that the corridor formed by the interval will capture the entire hazard curve 95% of the time. To obtain such a simultaneous confidence band, we employ the `confBands` method. In the following example, we plot both the pointwise and simultaneous 95% confidence intervals for the survival function.

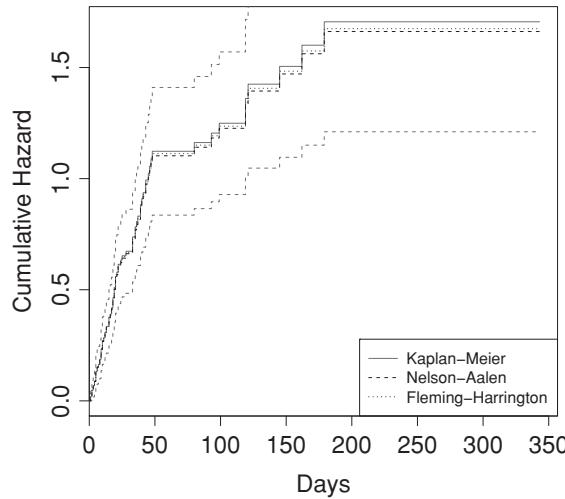


Fig. 19.5. Product-limit variations – admission sample.

```
> nnhs_KM<-survfit(Surv(DAYS, (DISCHG!=0)) ~ 1, subset=SMPL&ADM,
+ conf.type="log-log", conf.int=0.95)
> nnhs_KM_cb<-confBands(Surv(DAYS, (DISCHG!=0)) [SMPL&ADM] ,
+ confType = "log-log", confLevel = 0.95, type = "ep")
> plot(nnhs_KM, main="Kaplan-Meier with 95% CI - Admission Sample",
+ xlab="Days", ylab="Survival Function", conf.int=TRUE,
+ mark.time=FALSE)
> lines(nnhs_KM_cb, lty=3)
> legend("topright", c("KM Estimate","Pointwise Conf. Interval",
+ "Simultaneous Conf. Interval"), lty=c(1,2,3))
```

As expected, Figure 19.6 shows that the simultaneous confidence interval is wider than the pointwise counterpart. Also note that we have used the `log-log` transformation option to construct the confidence intervals, which assumes that the log of minus the log of the survival function estimate is approximately normal. Note that the `confBands` method requires the `OIsurv` package and does not work with left-truncated data.

We next compare the nonparametric estimates of survival for subsets of nursing home residents (e.g., by gender, type of care, and payer). The `survfit` method allows us to stratify the analysis by specifying one or more categorical variables after the survival object. The first example constructs the Kaplan-Meier estimates of the survival curves for male versus females residents based on those admitted during the year prior to the survey.

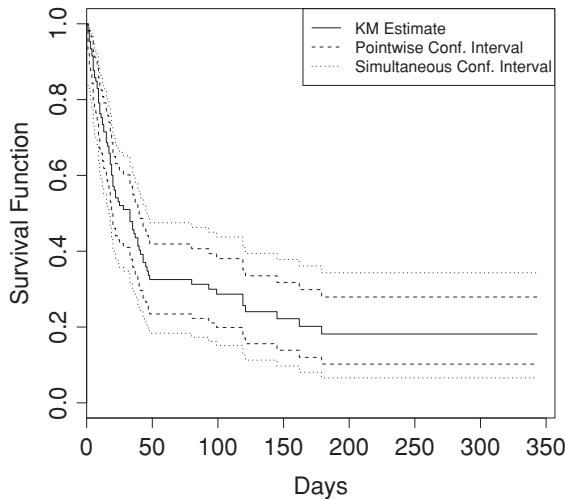


Fig. 19.6. Kaplan-Meier with 95% confidence interval – admission sample.

```
> nnhs_KM<-survfit(Surv(DAYS, (DISCHG!=0)) ~ SEX, subset=ADM,
  data=nnhs)
Call: survfit(formula = Surv(DAYS, (DISCHG != 0)) ~ SEX,
  data = nnhs, subset = ADM)

  records n.max n.start events *rmean *se(rmean) median
SEX=1      3426    3426     3426    2631     87.4       2.30     24
SEX=2      5590    5590     5590    3843    114.3       2.07     36

  0.95LCL 0.95UCL
SEX=1        23        26
SEX=2        33        38
* restricted mean with upper limit = 365

> plot(nnhs_KM, main="Male vs. Female KM Estimates -
  All Admissions",
  xlab="Days", ylab="Survival Function", lty=c(1,2),
  mark.time=FALSE)
> legend("topright",c("Male","Female"),lty=1:3)
```

The print method provides sex-distinct estimates of the median and restricted mean time to discharge (using a common upper limit of 365 days for both males and females). The plot method graphs the Kaplan-Meier survival curves. From

Figure 19.7 it is evident that females are more likely than males to remain in the nursing home for at least one year.

The following example shows survival estimates for combinations of care level (skilled and intermediate care) and payment source (Medicare and Medicaid), again for new admissions during the year prior to the survey.

```
> nnhs_KM<-survfit(Surv(DAYS, (DISCHG!=0)) ~ PAYOR+LOC,
  subset=(ADM&((PAYOR==4) | (PAYOR==5)) &(LOC!=3) &(LOC!=4)) ,
  data=nnhs)
> print(nnhs_KM,rmean="common")
Call: survfit(formula = Surv(DAYS, (DISCHG != 0)) ~ PAYOR + LOC,
  data = nnhs, subset = (ADM & ((PAYOR == 4) | (PAYOR == 5)))
  & (LOC != 3) & (LOC != 4)))

      records n.max n.start events *rmean *se(rmean)
PAYOR=4, LOC=1     3565   3565     3565    3111    42.6    1.32
PAYOR=4, LOC=2      366    366      366     251    115.3    7.88
PAYOR=5, LOC=1     1255   1255     1255     731    159.0    4.48
PAYOR=5, LOC=2      992    992      992     372    233.4    5.22

      median 0.95LCL 0.95UCL
PAYOR=4, LOC=1      19      19      20
PAYOR=4, LOC=2      41      33      60
PAYOR=5, LOC=1     110      97     121
PAYOR=5, LOC=2     345     298      NA
  * restricted mean with upper limit = 364

> sum(nnhs_KM$time*(nnhs_KM$n.event+nnhs_KM$n.censor))
[1] 407843
> sum(nnhs_KM$n.event)
[1] 4465
> plot(nnhs_KM, main="KM Estimates by Payor and Level of Care
  All Admissions", xlab="Days", ylab="Survival Function",
  lty=1:4, mark.time=FALSE)
> legend("topright",c("Medicare-Skilled","Medicare-Interm",
  "Medicaid-Skilled","Medicaid-Interm"),lty=1:4)
```

We see in Figure 19.8 that first-year discharge rates vary significantly. Medicare residents have shorter nursing home stays than Medicaid residents, and within each of these payment sources, skilled-care residents have shorter stays than intermediate-care residents.

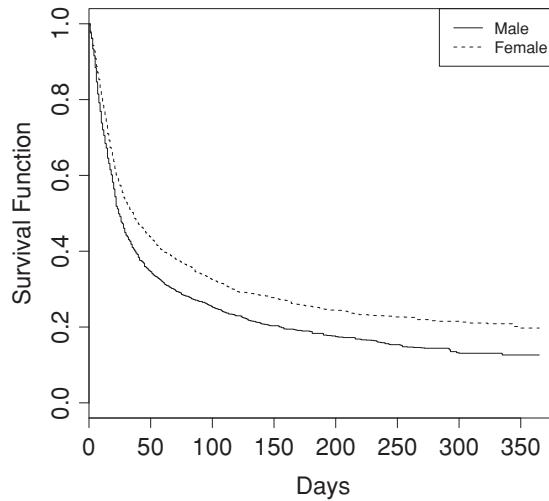


Fig. 19.7. Male versus female KM estimates – all admissions.

Plotting the log-log Kaplan-Meier survival function (or the log of the cumulative hazard function) versus the log of time can be used to assess two assumptions. First, if discharges follow an exponential distribution, the curve should be linear with unit slope. Second, if discharges for different resident populations satisfy the proportional hazards assumption (to be discussed later in this chapter), then the curves will be parallel. The next example focuses on new admissions of Medicare residents.

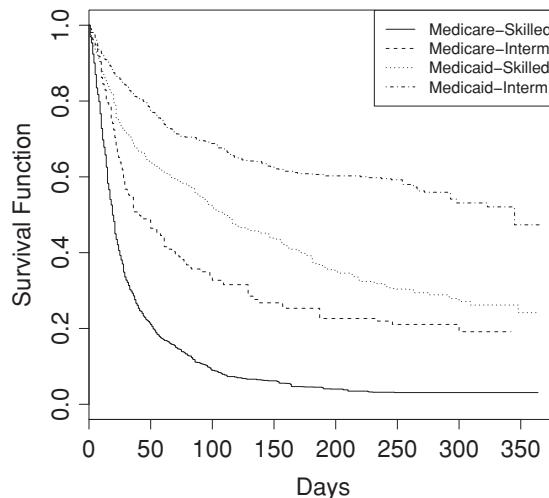


Fig. 19.8. KM Estimates by payer and level of care – all admissions.

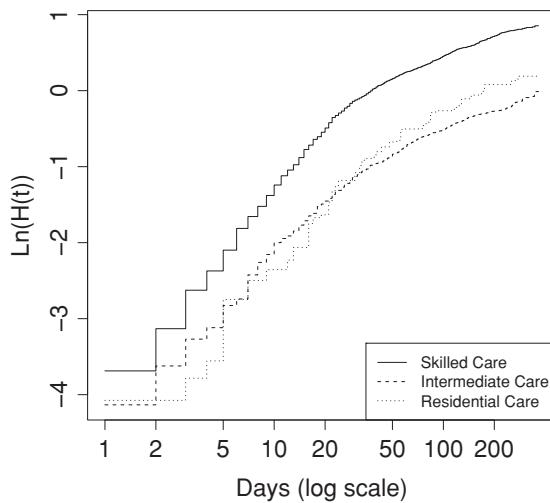


Fig. 19.9. KM log-cumulative hazard by level of care – admissions.

```
> nnhs_KM<-survfit(Surv(DAYS, (DISCHG!=0)) ~ LOC, subset=ADM&(LOC!=4),
  data=nnhs)
> plot(nnhs_KM, main="KM Log-Cumulative Hazard by Level of Care
  All Admissions", xlab="Days (log scale)", ylab="Ln(H(t))",
  lty=c(1,2,3), fun="cloglog", mark.time=FALSE)
> legend("bottomright",c("Skilled Care","Intermediate Care",
  "Residential Care"),lty=1:3)
```

The results in Figure 19.9 indicate that neither the exponential assumption (i.e., constant hazard rate) nor the proportional hazards assumption appears to be appropriate for this population.

19.4.2 Nonparametric Tests of the Equivalence of Survival Functions

To assess the difference in stratified survival curve estimates, two common tests are available using the `survdiff` method. With option `rho=0` the log-rank test (i.e., the Mantel-Haenzel test) is applied. With option `rho=1`, the Wilcoxon test is applied. Both tests measure the overall distance between the stratified survival curve estimates. The log-rank test weights the differences at each duration along the curve equally. The Wilcoxon test places greater weight on the differences at early durations. The following example compares male and female discharge survival for a 1% sample of new admissions.

```
> survdiff(Surv(DAYS, (DISCHG!=0)) ~ SEX, subset=ADM&SMPL,
  data=nnhs, rho=0)
Call:
survdiff(formula = Surv(DAYS, (DISCHG != 0)) ~ SEX,
  data = nnhs, subset = ADM & SMPL, rho = 0)

      N Observed Expected (O-E)^2/E (O-E)^2/V
SEX=1 44       34       31     0.281     0.48
SEX=2 62       43       46     0.190     0.48

Chisq= 0.5 on 1 degrees of freedom, p= 0.488
```

The log-rank p -value of 49% indicates that we cannot reject the null hypothesis that males and females have the same discharge rates. By rerunning the example with $\rho=1$, the reader can confirm that p -value of the Wilcoxon test is 55%.

With these tests, we can explore subpopulations of residents based on, say, combinations of gender, age, level of care, payment source, and ADLs impaired (i.e., the number of activities of daily living for which assistance is required). The party library offers an automated tool to perform this search: the conditional inference tree, using the `ctree` method. This method does not work with left-truncated data. The following example demonstrates the results for new admissions in the year prior to the nursing home survey. The `plot` output in Figure 19.10 provides a convenient visual display of the results.

```
> library(party)
> nnhs_KM_ctree <- ctree(Surv(DAYS, (DISCHG!=0)) ~ SEX+LOC+PAYOR
+AGE+ADL, subset=ADM&(LOC!=4)&((PAYOR==4) | (PAYOR==5)),
  data=nnhs)
> #Define a larger graphics page
> windows(width=25,height=10,pointsize=8)
> plot(nnhs_KM_ctree, main="Conditional Inference Tree
- All Admissions")
```

The method looks for the most significant split first, in this case based on Medicare versus Medicaid payment. Each population is then split again. The process continues until no significant split remains. This results in a partitioning of residents showing the estimated survival curve and the number of residents in each grouping. Note that the p -values are specific to each split decision (node) and that the probability of a Type 1 (or Type 2) error accumulates throughout the process.

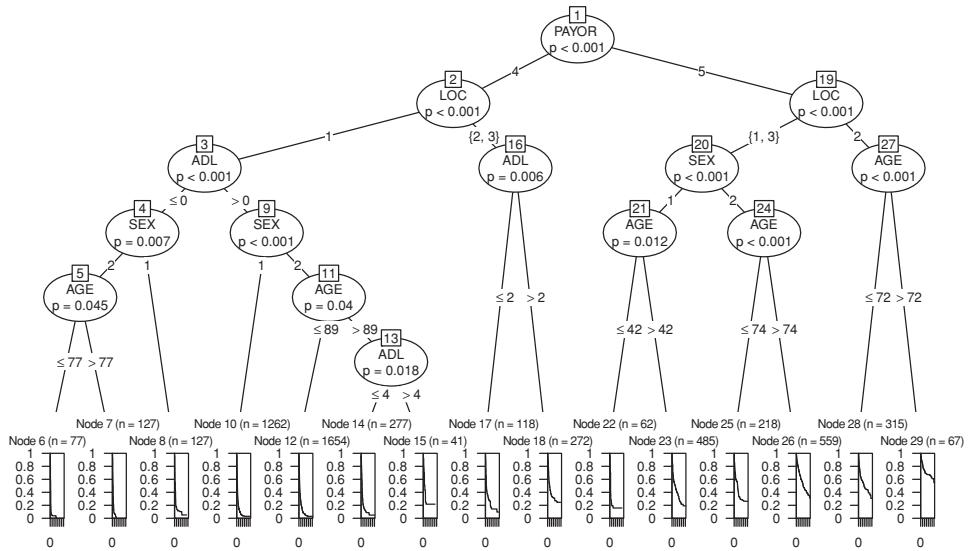


Fig. 19.10. Conditional inference tree – all admissions.

19.5 Proportional Hazards Model

In the proportional hazards model due to Cox (1972), one assumes that the hazard function can be written as the product of some baseline hazard and a function of a linear combination of explanatory variables. To illustrate, we use $h_i(t) = h_0(t) \exp(x_i'\beta)$, where $h_0(t)$ is the baseline hazard.

For right-censored data, the resulting likelihood function is

$$\begin{aligned} L(\beta, h_0) &= \prod_{i=1}^n f(y_i)^{1-\delta_i} S(y_i)^{\delta_i} = \prod_{i=1}^n h(y_i)^{1-\delta_i} \exp(-H(y_i)) \\ &= \prod_{i=1}^n (h_0(y_i) \exp(x_i'\beta))^{1-\delta_i} \exp(-H_0(y_i) \exp(x_i'\beta)). \end{aligned}$$

Left-truncation simply requires division by the survival function at the truncation point for each individual. This leads to a difference in the baseline cumulative hazard in the last term. Conventional maximum likelihood estimators are not available because the baseline hazard function is not parametrically defined. Cox (1972), Breslow (1974), Efron (1977), and Kalbfleisch and Prentice (2002) offer a variety of estimation techniques designed to maximize this likelihood. The techniques vary in how discrete baseline probabilities are handled, especially with regard to ties in the observed event times. In all cases, estimations of the baseline hazard function and the regression coefficients are performed separately. Inspection of the likelihood as a function of the baseline hazard function (fixing the regression coefficients) suggests minimizing the function at all durations except those at which failure events are

observed. The terms of the log-likelihood can be arranged to isolate each discrete jump in the baseline cumulative hazard function. The components can be differentiated and set to zero to find optimal choices for each discrete baseline jump (as functions of the regression coefficients). These optimal jump values can then be substituted into the likelihood to obtain a partial likelihood function that depends only on the regression coefficients. Breslow's technique leads to the following partial likelihood function:

$$L_P(\beta) = \prod_{i=1}^n \left(\frac{\exp(x_i' \beta)}{\sum_{j \in R(y_i)} \exp(x_j' \beta)} \right)^{1-\delta_i}$$

where $R(y)$ is the risk set at duration y (i.e., those individuals exposed at duration y).

With regression coefficients that maximize the partial likelihood, a variety of techniques have been suggested to estimate the baseline hazard function. Many of these baseline estimators can be considered variations of the product-limit or Nelson-Aalen estimators in which each exposed individual is assigned a weight proportional to his or her hazard multiplier. Again, the variations reflect different approaches to handling ties and discrete baseline jumps.

The `coxph` method allows the user to specify one of three methods to estimate the regression coefficients; Breslow (`method='breslow'`), Efron (`method = 'efron'`), and an exact method assuming a discrete logistic form for the hazard function (`method='exact'`). The default is the `Efron` method and is used for all of the examples in this chapter.

The following example fits a proportional hazard model to all the Medicare residents against gender, level of care, and ADL impairment count.

```
> nnhs_PH <- coxph(Surv(ENTRY, DAYS, (DISCHG!=0)) ~ SEX + LOC + ADL,
  subset=T18, data=nnhs, weights=, x=TRUE, method="efron")
> summary(nnhs_PH)

Call:
coxph(formula = Surv(ENTRY, DAYS, (DISCHG != 0)) ~ SEX + LOC +
ADL, data = nnhs, subset = T18, method = "efron", x = TRUE)

n= 4938, number of events= 4088

      coef  exp(coef)  se(coef)      z Pr(>|z|)
SEX2 -0.17458    0.83981   0.03239 -5.389 7.07e-08 ***
LOC2 -0.71314    0.49010   0.05647 -12.628 < 2e-16 ***
LOC3 -0.86208    0.42228   0.18500 -4.660 3.16e-06 ***
LOC4  0.16646    1.18112   0.21470   0.775   0.438
ADL  -0.05809    0.94356   0.01139 -5.099 3.42e-07 ***

---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

	exp(coef)	exp(-coef)	lower .95	upper .95
SEX2	0.8398	1.1907	0.7882	0.8949
LOC2	0.4901	2.0404	0.4387	0.5475
LOC3	0.4223	2.3681	0.2939	0.6068
LOC4	1.1811	0.8467	0.7754	1.7991
ADL	0.9436	1.0598	0.9227	0.9649

Concordance= 0.57 (se = 0.005)
 Rsquare= 0.053 (max possible= 1)
 Likelihood ratio test= 270.7 on 5 df, p=0
 Wald test = 236.3 on 5 df, p=0
 Score (logrank) test = 243.6 on 5 df, p=0

The summary method displays the regression coefficient estimates, standard errors, and *p*-values. The model indicates that female discharge rates are 84% of male rates; that intermediate-care and residential-care discharge rates are 49% and 42% of skilled-care rates, respectively; and that discharge rates decrease roughly 6% for each ADL impaired. This assumes that the proportional hazard and other model assumptions are correct. All three of the tests for model significance reject the null hypothesis that all the regression coefficients are zero.

To test the proportional hazards assumption, we can use the cox.zph method to assess the constancy of the coefficients over time. The method employs rescaled Schoenfeld residuals to generate and plot local regression coefficient estimates over time. If the proportional hazards assumption is correct, plots of residuals over time should be horizontal. Correlation coefficients between the residuals and observation time are used to compute test statistics for each coefficient. The following example shows results for the proportional hazards model fit in the previous example.

```
> nnhs_PH_zph <- cox.zph(nnhs_PH, transform="km")
> nnhs_PH_zph
      rho    chisq      p
SEX2   0.0221  2.0167 0.155580
LOC2   0.0198  1.5878 0.207643
LOC3   0.0282  3.2852 0.069907
LOC4   0.0045  0.0832 0.772952
ADL    0.0519 11.5805 0.000666
GLOBAL     NA 18.4179 0.002466
```

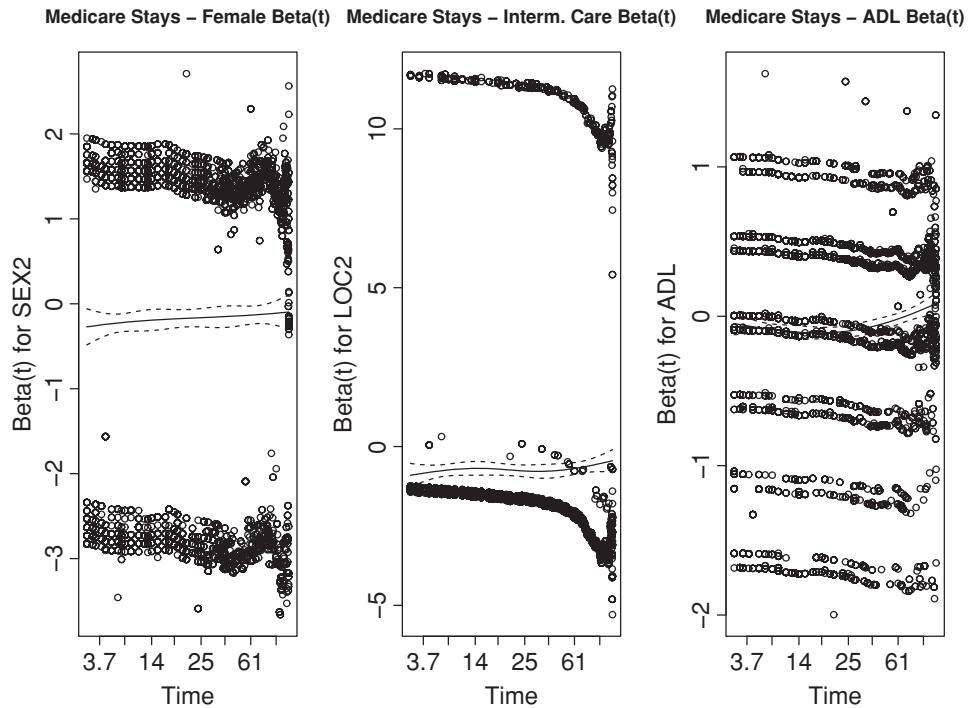


Fig. 19.11. Medicare stays – Cox ZPH plots.

We see that the p -value for the null hypothesis that the female gender effect is uncorrelated with time is 16%, supporting the proportional hazards assumption. In contrast, there is a significant correlation of the ADL effect with time because the p -value is less than 0.1%. The plots in Figure 19.11 show the residual patterns over time for the female gender, intermediate-care, and ADL effects.

```
> par(mfrow=c(1,3), cex=.7)
> plot(nnhs_PH_zph, var="SEX2", main="Medicare Stays
- Female Beta(t)")
> plot(nnhs_PH_zph, var="LOC2", main="Medicare Stays
- Interm. Care Beta(t)")
> plot(nnhs_PH_zph, var="ADL", main="Medicare Stays
- ADL Beta(t)")
> par(mfrow=c(1,1), cex=1)
```

Each plotted point can be considered the contribution of an observed failure (discharge) to the estimated regression coefficient. The fitted moving average and confidence interval provide a visual indication of the varying effect of each factor over time. We see that ADL impairment at early durations is associated with reduced discharge rates, whereas the reverse is true at later durations. This may be due to the increasing role of mortality as the reason for discharge at later nursing home stay durations.

19.5.1 Estimation of the Baseline Survival and Hazard Functions

Once the regression coefficients of the proportional hazards model are estimated, we may wish to estimate the survival curve (or cumulative hazard function) associated with a specified set of factors. When applied to a `coxph` object (resulting from applying the `coxph` method to a survival data object), the `survfit` method, by default, uses one of three techniques to estimate the fitted survival curve depending on the tie-handling method used in the estimation of the regression coefficients.

If the categorical factors are set to their default values and the continuous covariates are set to zero, we obtain the baseline survival or hazard functions. If no factors or covariates are specified, then the `survfit` method will use the average value of each factor or covariate. This default curve should be used with caution, because the factor combinations assumed may not be obtainable for any individual (e.g., 30% male and 70% female). The user can specify one or more combinations of factors/covariates using the `newdata` option. The following example shows the fitted male Medicare cumulative hazard functions for five combinations of level of care (skilled and intermediate) and ADL impairment (none, one, and five impairments).

```
> nnhs_PH_xsurv <- survfit(nnhs_PH,
  newdata=list(SEX=factor(c(1,1,1,1,1),levels=1:2),
  LOC=factor(c(1,1,1,2,2),levels=1:4),ADL=c(0,1,5,1,5)))
> plot(nnhs_PH_xsurv, fun="cumhaz", xmax=2000, conf.int=FALSE,
  lty=c(1:5), mark.time=FALSE, ylim=c(0,5),
  main="Medicare Stays Male Cumulative Hazard", xlab="Days",
  ylab="Cumulative Hazard")
> legend("bottomright",c("Baseline (Skilled ADL=0)",
  "Skilled ADL=1","Skilled ADL=5","Intermediate ADL=1",
  "Intermediate ADL=5"), lty=c(1:5), cex=.9)
```

We see in Figure 19.12 that skilled-care residents have a significantly greater discharge rate than intermediate-care residents and that, on average, higher ADL

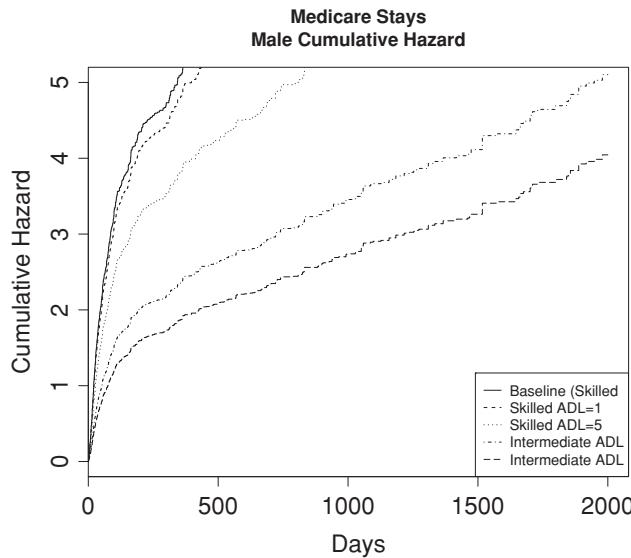


Fig. 19.12. Medicare stays – male cumulative hazard by level of care.

impairment is associated with lower discharge rates. Recall, however, that the early ADL effect was found to reverse at later durations when we allowed the regression coefficients to vary over time.

19.5.2 Residuals

A variety of residuals are available to assess the proportional hazards model assumptions. The Cox-Snell residual for an observation is the estimated difference in the cumulative hazard function from entry to exit. If the model is correctly specified, these residuals will behave like a censored/truncated sample from a unit exponential survival process. So, a plot of the Nelson-Aalen estimator of the cumulative hazard function for the Cox-Snell residuals versus the residuals should be a straight line with unit slope.

The `residuals` method applied to a `coxph` object allows the user to select several residual types. Although the Cox-Snell residuals are not immediately available, a closely related residual, the martingale residual, is an option. The martingale residual is defined as the difference between an event indicator and the Cox-Snell residual for each observation. If the model is correctly specified, the martingale residuals can be considered the difference over time of the actual number of failures minus the expected number of failures. As with conventional regression residuals, plots of the martingale residuals versus a covariate can be used to determine the appropriateness of the functional form used in the hazard multiplier.

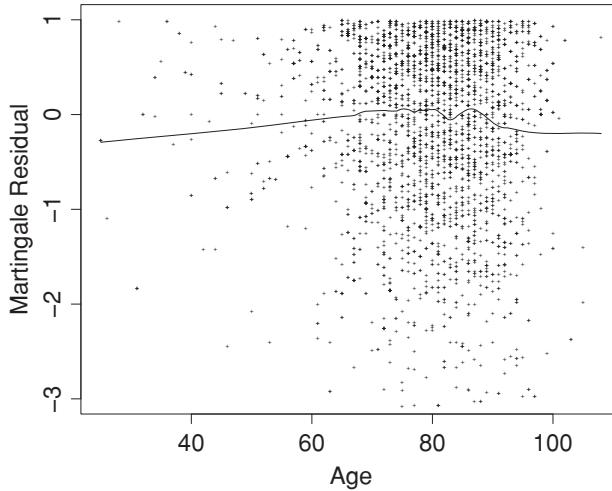


Fig. 19.13. Medicare stays – Cox regression: martingale residuals by age.

The following example plots martingale residuals for Medicare data versus age, a variable not included in the model.

```
> nnhs_PH <- coxph(Surv(ENTRY, DAYS, (DISCHG!=0)) ~ SEX + LOC
+ ADL, subset=T18, data=nnhs, method="efron")
> nnhs_PH_resids <- residuals(nnhs_PH, weighted="FALSE",
  type="martingale")
> scatter.smooth(AGE[T18], nnhs_PH_resids, ylim=c(-3,1),
  pch="+", evaluation=300, family="gaussian", span=.25, cex=.4,
  main="Medicare Stays - Cox Regression Martingale Residuals
  by Age",
  xlab="Age", ylab="Martingale Residual")
```

The plot in Figure 19.13 indicates that actual discharges at very young and very old ages tend to be somewhat less than expected by the fitted model. The data are sparse at these extremes, however. So, adding age to the model appears to be unnecessary.

While the martingale residuals have mean zero if the model is correctly specified, the distribution about zero is highly skewed, making outlier detection difficult. A standardized version of the martingale residual, the deviance residual, attempts to normalize the distribution. Specifically, $D = \text{sign}(r)\sqrt{-2(r + (1 - \delta)\ln(1 - \delta - r))}$, where r is the martingale residual for the observation and δ is an indicator that the observation is right-censored. The next example shows the deviance residuals plotted against age.

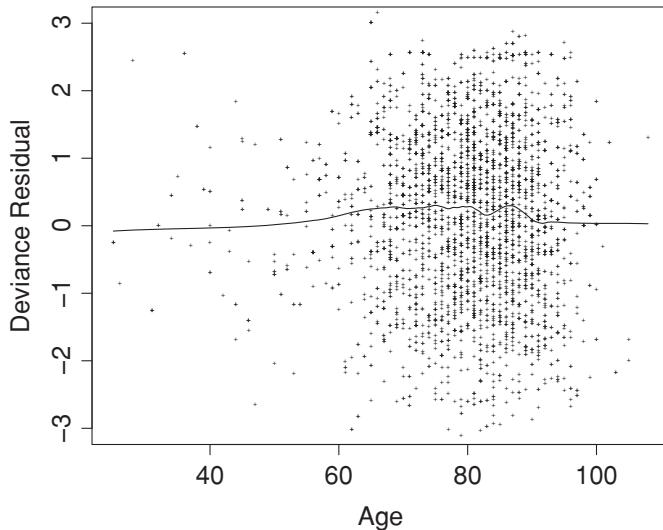


Fig. 19.14. Medicare stays – Cox regression: deviance residuals by age.

```

> nnhs_PH <- coxph(Surv(ENTRY, DAYS, (DISCHG!=0)) ~ SEX + LOC
+ ADL, subset=T18, data=nnhs, weights=, method="efron")
> nnhs_PH_dev <- residuals(nnhs_PH, weighted="FALSE",
+ type="deviance")
> scatter.smooth(AGE[T18],nnhs_PH_dev, ylim=c(-3,3), pch="+",
+ evaluation=300, family="gaussian", span=.25,cex=.4,
+ main="Medicare Stays - Cox Regression Deviance Residuals
+ by Age",
+ xlab="Age", ylab="Deviance Residual")

```

In Figure 19.14 we can see more clearly with deviance residuals than with martingale residuals which observations are outliers in both directions.

Several other residuals are available from the `residuals` method, including scores residuals (`score`, derived from the observations' contribution to the score vector), rescaled score residuals (`dfbeta`, which estimate the impact of removing an observation on each regression coefficient), rescaled `dfbeta` residuals (`dfbetas`), Schoenfeld residuals (`schoenfeld`, a part of the score residual based on the actual less expected values of covariates at observed failure times), and rescaled Schoenfeld residuals (`scaledsch`, which are used in `coxph.zph` to generate local estimates of the regression coefficients).

19.5.3 Testing the Estimated Regression Coefficients

The `summary` method applied to a `coxph` object generates the usual array of test statistics for individual coefficients and for the entire model. The following example shows results for the proportional hazards model fit to Medicare data.

```
> nnhs_PH <- coxph(Surv(ENTRY, DAYS, (DISCHG!=0)) ~ SEX + LOC
+ ADL, subset=T18, data=nnhs, weights=, method="efron")
> summary(nnhs_PH)

Call:
coxph(formula = Surv(ENTRY, DAYS, (DISCHG != 0)) ~ SEX + LOC +
ADL, data = nnhs, subset = T18, method = "efron")

n= 4938, number of events= 4088

      coef exp(coef) se(coef)      z Pr(>|z|)    
SEX2 -0.17458   0.83981  0.03239 -5.389 7.07e-08 ***
LOC2 -0.71314   0.49010  0.05647 -12.628 < 2e-16 ***
LOC3 -0.86208   0.42228  0.18500 -4.660 3.16e-06 ***
LOC4  0.16646   1.18112  0.21470  0.775  0.438    
ADL  -0.05809   0.94356  0.01139 -5.099 3.42e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1    1

      exp(coef) exp(-coef) lower .95 upper .95    
SEX2     0.8398     1.1907    0.7882    0.8949    
LOC2     0.4901     2.0404    0.4387    0.5475    
LOC3     0.4223     2.3681    0.2939    0.6068    
LOC4     1.1811     0.8467    0.7754    1.7991    
ADL      0.9436     1.0598    0.9227    0.9649    

Concordance= 0.57  (se = 0.005 )
Rsquare= 0.053  (max possible= 1 )
Likelihood ratio test= 270.7 on 5 df,  p=0
Wald test           = 236.3 on 5 df,  p=0
Score (logrank) test = 243.6 on 5 df,  p=0
```

If additional testing of combinations of coefficients is necessary, the `coxph` object contains the needed coefficient estimates and related variance-covariance estimates. For example, to test a single coefficient (say, the effect of ADL impairments), the following code can be applied.

```

> beta=nnhs_PH$coeff
> va=nnhs_PH$var
>
> #Test ADL beta
> C <- matrix(c(0, 0, 0, 0, 1), nrow=1, byrow=TRUE)
> d <- rep(0, 1)
> t1 <- C %*% beta - d
> t2 <- C %*% va %*% t(C)
> XW2 <- c(t(t1) %*% solve(t2) %*% t1)
> pchisq(XW2, 1, lower.tail=FALSE)
[1] 3.418314e-07

```

We see that this results in the same p -value for ADLs as generated by the `summary` method. To test the significance of all of the level of care coefficients, we apply the following code.

```

> C <- matrix(c(0, 1, 0, 0, 0,
+               0, 0, 1, 0, 0,
+               0, 0, 0, 1, 0), nrow=3, byrow=TRUE)
> d <- rep(0, 3)
> t1 <- C %*% beta - d
> t2 <- C %*% va %*% t(C)
> XW2 <- c(t(t1) %*% solve(t2) %*% t1)
> pchisq(XW2, 3, lower.tail=FALSE)
[1] 2.072037e-38

```

So, we easily reject the null hypothesis that all of the level of care coefficients are zero.

19.5.4 Time-Varying Covariates

The `coxph` method allows for time-varying covariates. This is accomplished by using the time transforming function, `tt()`, as part of the model design. For example, we might include an interaction term for ADL count and time to test the earlier finding that the ADL effect changes with time from admission.

```

> nnhs_PH2 <- coxph(Surv(ENTRY, DAYS, (DISCHG!=0)) ~ SEX + LOC
+ ADL + tt(ADL), subset=T18, data=nnhs, x=TRUE,
method="efron", tt = function(x, t, ...) (t*x))

```

```

Call:
coxph(formula = Surv(ENTRY, DAYS, (DISCHG != 0)) ~ SEX + LOC +
    ADL + tt(ADL), data = nnhs, subset = T18, method = "efron",
    x = TRUE, tt = function(x, t, ...) (t * x))

n= 4938, number of events= 4088

            coef  exp(coef)   se(coef)      z Pr(>|z|)
SEX2     -1.770e-01 8.378e-01 3.240e-02 -5.462 4.70e-08 ***
LOC2     -7.110e-01 4.911e-01 5.650e-02 -12.585 < 2e-16 ***
LOC3     -8.157e-01 4.423e-01 1.850e-01 -4.410 1.03e-05 ***
LOC4      1.533e-01 1.166e+00 2.148e-01  0.714 0.475490
ADL      -7.182e-02 9.307e-01 1.196e-02 -6.007 1.89e-09 ***
tt(ADL)   1.955e-04 1.000e+00 5.677e-05  3.444 0.000573 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

            exp(coef)  exp(-coef) lower .95 upper .95
SEX2       0.8378     1.1936   0.7863   0.8927
LOC2       0.4911     2.0361   0.4397   0.5487
LOC3       0.4423     2.2608   0.3078   0.6356
LOC4       1.1657     0.8579   0.7651   1.7760
ADL        0.9307     1.0745   0.9091   0.9528
tt(ADL)    1.0002     0.9998   1.0001   1.0003

Concordance= 0.57  (se = 0.037 )
Rsquare= 0.056  (max possible= 1 )
Likelihood ratio test= 285.1 on 6 df,  p=0
Wald test          = 248.5 on 6 df,  p=0
Score (logrank) test = 256.2 on 6 df,  p=0

```

We see that the coefficient of the ADL-time interaction term is statistically significant. It is also evident that the ADL effect reverses sign after 367 days.

19.6 Parametric Survival Modeling

So far, we have explored nonparametric and semi-parametric approaches to modeling survival data. We now consider parametric options, which allow conventional maximum likelihood estimation to be applied to the censored/truncated likelihood functions discussed at the start of the chapter. The `survreg` method allows the users to select from several parametric distributions, including the exponential, the Weibull, logistic, log-logistic, and the log-normal.

Unfortunately, the `survreg` method does not allow for left-truncated data. So, in the example that follows, we limit the data to Medicare admissions in the year prior to the nursing home survey. To assess the adequacy of each of the candidate parametric distributions, we first compute and plot the Kaplan-Meier estimate of the cumulative hazard function. The code that follows then fits each of the parametric distributions and plots the cumulative hazard function on the same graph.

```
> plot(survfit(Surv(DAYS, (DISCHG!=0)) ~ 1, data=nnhs,
+ subset=(T18&ADM)), conf.int=FALSE, lty=1, mark.time=FALSE,
+ fun="cumhaz", main="Medicare Admissions - Parametric
+ Regression Fits", xlab="Days", ylab="Cumulative Hazard")
> nnhs_par_exp <- survreg(Surv(DAYS, (DISCHG!=0)) ~ 1, data=nnhs,
+ subset=(T18&ADM), dist="exponential")
> lines(predict(nnhs_par_exp, type="quantile",
+ p=seq(.001,.999,by=.001)) [1,], -log(seq(.999,.001,by=-.001)),
+ lty=2)
> nnhs_par_wei <- survreg(Surv(DAYS, (DISCHG!=0)) ~ 1, data=nnhs,
+ subset=(T18&ADM), dist="weibull")
> lines(predict(nnhs_par_wei, type="quantile",
+ p=seq(.001,.999,by=.001)) [1,], -log(seq(.999,.001,by=-.001)),
+ lty=3)
> nnhs_par_logistic <- survreg(Surv(DAYS, (DISCHG!=0)) ~ 1,
+ data=nnhs,
+ subset=(T18&ADM), dist="logistic")
> lines(predict(nnhs_par_logistic, type="quantile",
+ p=seq(.001,.999,by=.001)) [1,], -log(seq(.999,.001,by=-.001)),
+ lty=4)
> nnhs_par_loglog <- survreg(Surv(DAYS, (DISCHG!=0)) ~ 1,
+ data=nnhs,
+ subset=(T18&ADM), dist="loglogistic")
> lines(predict(nnhs_par_loglog, type="quantile",
+ p=seq(.001,.999,by=.001)) [1,], -log(seq(.999,.001,by=-.001)),
+ lty=5)
> nnhs_par_logn <- survreg(Surv(DAYS, (DISCHG!=0)) ~ 1, data=nnhs,
+ subset=(T18&ADM), dist="lognormal")
> lines(predict(nnhs_par_logn, type="quantile",
+ p=seq(.001,.999,by=.001)) [1,], -log(seq(.999,.001,by=-.001)),
+ lty=6)
> legend("bottomright", c("Kaplan-Meier", "Exponential", "Weibull",
+ "Logistic", "Log-Logistic", "Log-Normal"), lty=1:6)
```

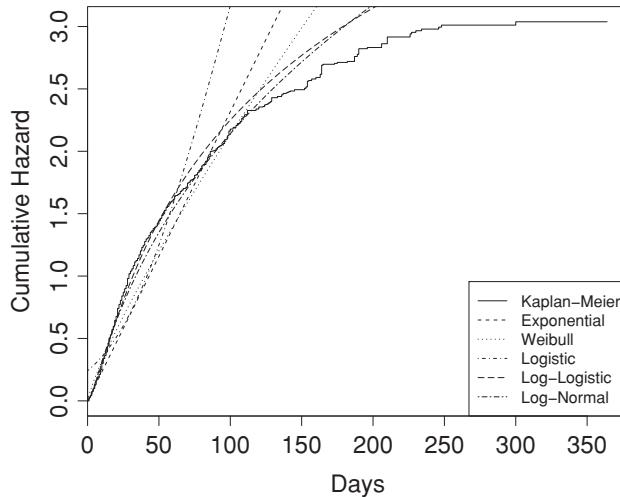


Fig. 19.15. Medicare admissions – parametric regression fits.

In Figure 19.15 all of the parametric fits fail to match the concavity of the Kaplan-Meier hazard function. The log-logistic distribution does fit reasonably well over the first 100 days and shows more concavity at later durations. So, we adopt this form for the next example.

Prior proportional hazards modeling clearly indicates that gender, level of care, and ADL impairment all have an impact of nursing home discharge rates. To incorporate these effects into the log-logistic model, we employ a transformation of the time axis. Specifically, we define $S(t|z) = S_0(te^{-\beta'z})$. If $\beta'z$ is large and positive, the survival function is increased and the hazard rates are deflated. This is known as an accelerated failure-time model.

The following example uses the `survreg` method to fit a log-logistic accelerated failure-time model to new Medicare admission data with effects for gender, level of care, and ADL impairment.

```
> nnhs_par_loglog2 <- survreg(Surv(DAYS, (DISCHG!=0)) ~ SEX +
  LOC + ADL, data=nnhs, subset=(T18&ADM), dist="loglogistic")
> summary(nnhs_par_loglog2)

Call:
survreg(formula = Surv(DAYS, (DISCHG != 0)) ~ SEX + LOC + ADL,
  data = nnhs, subset = (T18 & ADM), dist = "loglogistic")
              Value Std. Error      z      p
(Intercept) 2.6244      0.0530 49.514 0.00e+00
```

SEX2	0.2098	0.0402	5.213	1.86e-07
LOC2	0.8532	0.0734	11.617	3.36e-31
LOC3	0.9713	0.2584	3.759	1.70e-04
LOC4	0.2426	0.4031	0.602	5.47e-01
ADL	0.0781	0.0136	5.752	8.84e-09
Log (scale)	-0.3570	0.0143	-24.899	7.60e-137

Scale= 0.7

Log logistic distribution

Loglik(model)= -15546 Loglik(intercept only)= -15652.9
 Chisq= 213.71 on 5 degrees of freedom, p= 0
 Number of Newton-Raphson Iterations: 3
 n= 3967

We see that the signs of the regression coefficients are positive rather than negative (as they were in the proportional hazards model), because they act to increase survival and decrease discharge rates. To compare the resulting fitted model to the comparable proportional hazards model, we apply the following code. This plots the fitted cumulative hazard functions for females using skilled care and having three ADL impairments.

```
> xvals=data.frame(SEX=2,LOC=1,ADL=3);
  xvals$SEX=as.factor(xvals$SEX);
  xvals$LOC=as.factor(xvals$LOC)
> plot(survfit(coxph(Surv(DAYS, (DISCHG!=0)) ~ SEX+LOC+ADL,
  data=nnhs, subset=(T18&ADM), weights=), newdata=xvals),
  lty=1, fun="cumhaz", conf.int=FALSE, mark.time=FALSE,
  main="Medicare Admissions - Log-Logistic Regression
  Female Skilled Care ADL=3", xlab="Days",
  ylab="Cumulative Hazard")
> lines(predict(nnhs_par_loglog2,newdata=xvals,type="quantile",
  p=seq(.001,.999,by=.001)), -log(seq(.999,.001,by=-.001)),
  lty=2)
> legend("bottomright",c("Cox Model","Log-Logistic Model"),
  lty=1:2)
```

In Figure 19.16 we see that the resulting cumulative hazard functions are close for the first 100 days. After 100 days, the log-logistic model generates consistently lower survival rates.

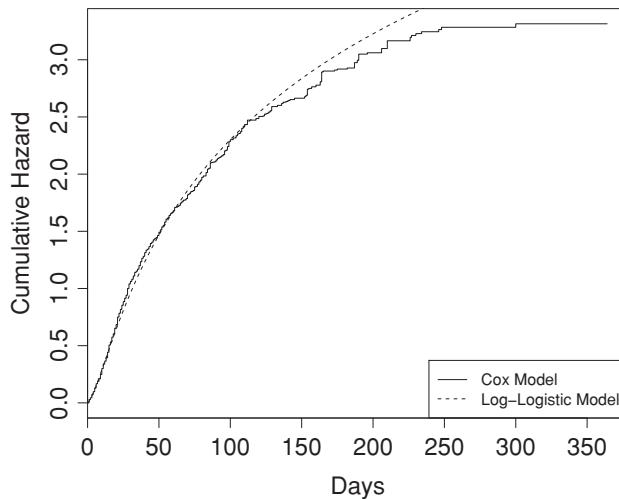


Fig. 19.16. Medicare admissions – log-logistic regression: female skilled care ADL = 3.

19.7 Further Reading

In this chapter, data from the National Nursing Home Survey were used to demonstrate the application of various survival models using tools available in R. Readers seeking additional discussion and examples of survival modeling methods and techniques are referred to Moeschberger and Klein (2005) and Kalbfleisch and Prentice (2002). Additional documentation of R survival modeling tools can be found at cran.r-project.org/web/views/Survival.html.

19.8 Exercises

Exercise 19.1. Use R to explore the contents of the nnhs data object. The Appendix provides a brief description of the NNHS variables.

Exercise 19.2. Explore the structure of nnhs_KM. Use R to compute the exposure in resident-days for this sample using the time, n.event, and n.censor components of the nnhs_KM object.

Exercise 19.3. Compare the first-year survival functions of new admissions by level of care for Skilled Care (LOC=1), Intermediate Care (LOC=2), and Residential Care (LOC=3) residents.

Exercise 19.4. Compare the first-year survival functions of new admissions by payment source for Personal (PAYOR=1), Insurance (PAYOR=2), Medicare (PAYOR=4), and Medicaid (PAYOR=5) residents.

Exercise 19.5. Stratified baseline model: If it is clear that a single baseline hazard function cannot accommodate all of the data, the `strata` option can be used in the `coxph` method to allow for distinct baseline hazard functions for each strata within the data. The following exercise shows results when a proportional hazards model is fit to Medicare and Medicaid resident data, allowing for separate baseline hazard functions for each payment source. The regression coefficient estimates related to gender, level of care, and ADL are pooled across both strata.

Use the following code to fit a stratified proportional hazards model to Medicare and Medicaid resident data, display the fitted regression coefficients, plot the average Medicare and Medicaid baseline cumulative hazard functions, and test the constancy of the regression coefficients over time.

```
> nnhs_PH <- coxph(Surv(ENTRY, DAYS, (DISCHG!=0)) ~ SEX + LOC
+ ADL +strata(T18), subset=(T18 | (PAYOR==5)), data=nnhs,
x=TRUE, method="efron")
> summary(nnhs_PH)
> nnhs_PH_base <- survfit(nnhs_PH)
> plot(nnhs_PH_base, fun="cloglog", xmax=5000, lty=c(1,2),
mark.time=FALSE, conf.int=FALSE,
main="Medicare and Medicaid Stays Stratified Base Hazards",
xlab="Days (log scale)", ylab="log(H(t))"
> legend("bottomright",c("Medicaid","Medicare"),lty=c(1,2))
> nnhs_PH_zph<-cox.zph(nnhs_PH, transform="km")
> nnhs_PH_zph
> par(mfrow=c(1,3), cex=.7)
> plot(nnhs_PH_zph, var="SEX2", main="Medicare & Medicaid Stays
- Female Beta(t)")
> plot(nnhs_PH_zph, var="LOC2", main="Medicare & Medicaid Stays
- Interm. Care Beta(t)")
> plot(nnhs_PH_zph, var="ADL", main="Medicare & Medicaid Stays
- ADL Beta(t)")
> par(mfrow=c(1,1), cex=1)
```

Exercise 19.6. Repeat the log-log survival plots for all Medicare residents (i.e., not limited to new admissions) over the first 3,000 days in the nursing home.

Exercise 19.7. Confirm that the log-rank and Wilcoxon tests for the difference in discharge survival for a 1% sample of skilled-care and intermediate-care admissions both have p -values less than 0.5%.

Exercise 19.8. Use the following code to generate a boxplot of the Martingale residuals versus ADL impairment count.

```
> boxplot(nnhs_PH_resids ~ as.factor(ADL[T18]), ylim=c(-3,1),
  xlab="ADL", main="Medicare Stays - Cox Regression
  Martingale Residual Boxplot by ADL")
```

Exercise 19.9. Use the following code to generate the rescaled Schoenfeld residuals. Also, verify that the correlation coefficients of the residuals with time are the same as those generated by the `cox.zph` method in testing the constancy of the regression coefficients.

```
> nnhs_PH <- coxph(Surv(ENTRY, DAYS, (DISCHG!=0)) ~ SEX + LOC +
  ADL, subset=T18,
  data=nnhs, weights=, method="efron")
> nnhs_PH_schoen2 <- residuals(nnhs_PH, weighted="FALSE",
  type="scaledsch")
> cor.test(sort(DAYS[T18&(DISCHG!=0)]), nnhs_PH_schoen2[,1])
> cor.test(sort(DAYS[T18&(DISCHG!=0)]), nnhs_PH_schoen2[,2])
> cor.test(sort(DAYS[T18&(DISCHG!=0)]), nnhs_PH_schoen2[,3])
> cor.test(sort(DAYS[T18&(DISCHG!=0)]), nnhs_PH_schoen2[,4])
> cor.test(sort(DAYS[T18&(DISCHG!=0)]), nnhs_PH_schoen2[,5])
> nnhs_PH_zph <- cox.zph(nnhs_PH, transform="identity")
> nnhs_PH_zph
```

19.9 Appendix. National Nursing Home Survey Data

The following fields were abstracted from the current resident and discharged resident survey files for the 1999 National Nursing Home Survey for use in the examples and exercises of this chapter. A complete listing of the data and additional documentation for the NNHS is available on the book's website.

Variable	Description
RESID	Resident identification (ID)
SEX	Sex
AGE	Age at admission
MARRIED	Marital status
LIVLOC	Living location prior to admission
LIVWITH	Living status prior to admission
DAYS	Length of stay
LOC	Level of care in nursing home
PAYOR	Payor (recent)
MSA	Rural versus urban (Metropolitan Statistical Area)
OWNER	Nursing home ownership
BEDS	Nursing home size (bed count)
CERT	Nursing home certification
DISCHG	Reason for discharge
FACID	Facility Identification (ID)
ENTRY	Days since admission at start of observation period
ADL	ADL score
SMPL	1% sample Indicator
ADM	Admitted during observation period
T18	Payer is Medicare

References

- Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics* 6(4), 701–726.
- Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 89–99.
- Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 187–220.
- Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American statistical Association* 72(359), 557–565.
- Frees, E. W. (2010). *Regression Modeling with Actuarial and Financial Applications*. Cambridge University Press, Cambridge.
- Greenwood, M. (1926). The errors of sampling of the survivorship tables. In *Reports on Public Health and Statistical Subjects*, Volume 33. Her Majesty's Stationery Office, London.
- Kalbfleisch, J. D. and R. L. Prentice (2002). *The statistical analysis of failure time data*. John Wiley & Sons, New York.
- Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53(282), 457–481.
- Moeschberger, M. L. and J. P. Klein (2005). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, New York.

20

Transition Modeling

Bruce Jones and Weijia Wu

Chapter Preview. This chapter provides an introduction to transition modeling. Consider a situation where an individual or entity is, at any time, in one of several states and may from time to time move from one state to another. The state may, for example, indicate the health status of an individual, the status of an individual under the terms of an insurance policy, or even the “state” of the economy. The changes of state are called *transitions*. There is often uncertainty associated with how much time will be spent in each state and which state will be entered on each transition. This uncertainty can be modeled using a multistate stochastic model. Such a model may be described in terms of the rates of transition from one state to another. Transition modeling involves the estimation of these rates from data.

Actuaries often work with contracts involving several states and financial implications associated with presence in a state or transition between states. A life insurance policy is a simple example. A multistate stochastic model provides a valuable tool to help the actuary analyze the cash flow structure of a given contract. Transition modeling is essential to the creation of this tool.

This chapter is intended for practitioners, actuaries, or analysts who are faced with a multistate setup and need to estimate the rates of transition from available data. The assumed knowledge – only basic probability and statistics as well as life contingencies – is minimal. After a discussion of some relevant actuarial applications in Section 20.1, this chapter gives a brief introduction to multistate stochastic models in Section 20.2. We follow the notation, terminology, and approach of (Dickson et al. 2009, chapter 8). Although our introduction is intended to stand on its own, the reader is encouraged to consult Dickson et al. (2009) for a more detailed presentation. After we introduce multistate models, we discuss how to estimate the rates of transition under different assumptions about the model and the data in Sections 20.3 and 20.4.



Fig. 20.1. Alive-dead model.

20.1 Multistate Models and Their Actuarial Applications

Many actuarial calculations involve a multistate setup. Although they use multistate models, the multistate aspect of some approaches is implicit.

Consider, for example, the survival model used for basic life insurance and life annuity calculations. As shown in Figure 20.1, this model can be viewed as a multistate model or, more specifically, as a two-state model. Accordingly, the individual is, at any time, in one of two states: alive (0) or dead (1). Notice that we adopt the convention of Dickson et al. (2009), where states are numbered using consecutive integers, with the initial state given the number 0. In this model, state 1 is an absorbing state. That is, an individual who enters state 1 can never leave. Therefore, an individual may make no more than one transition – the transition from state 0 to state 1. The rate of transition is, of course, the mortality rate.

A second example is the two-life model used for joint life insurance and joint life annuity policies. In this example, the states indicate the survival status of the two lives who are ages x and y at the time origin. We therefore do not say that an individual is in a given state at a given time. Instead, we recognize that our multistate model describes a *stochastic process* – a collection of random variables representing the state at each point in time – and we say that the process is in a given state at a given time. In the two-life model, the process is, at any time, in one of four states. Using the shorthand (x) and (y) to mean “a life aged x ” and “a life aged y ,” the four states are defined in Figure 20.2.

State 3 is, of course, an absorbing state. Notice that the figure allows for a direct transition from state 0 to state 3, implying the simultaneous death of the two lives. This is not normally allowed in models for joint lifetimes. However, by including this possibility, we allow for a “common shock,” which is an event that causes the two

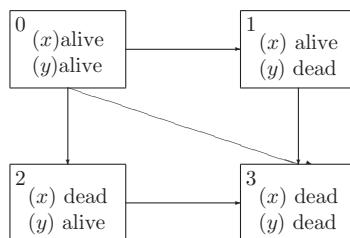


Fig. 20.2. Two-life model.

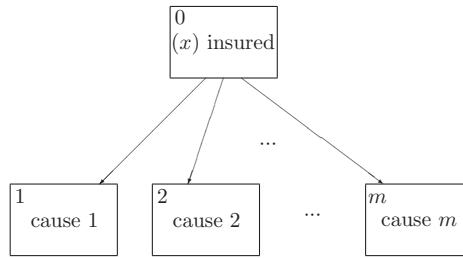


Fig. 20.3. Multiple decrement model.

lives to die at the same time. This is one way of introducing positive dependence of the two lifetimes. Another way is to assume that the rates of transition from state 2 to state 3 are greater than the rates of transition from state 0 to state 1 and that the rates of transition from state 1 to state 3 are greater than the rates of transition from state 0 to state 2. That is, mortality rates are higher after one's spouse dies than before.

A third example of a multistate model is the multiple decrement model shown in Figure 20.3. The process begins in state 0 and may move to one of m absorbing states. Such a model can be used, for example, to represent mortality by cause of death or the mode of termination of an insurance policy.

The three-state illness-death model shown in Figure 20.4 is another example of a multistate model that is useful in actuarial practice. Methods for multistate models are particularly helpful in this situation, because individuals may return from the sick state to the healthy state. Therefore, probabilities associated with the process being in a given state at a given future time cannot be calculated directly. However, the methods presented in Section 20.2.1 allow us to determine these probabilities.

Finally, the claims payment process can be modeled using a multistate model as shown in Figure 20.5. In practice, we must allow for partial payment of the claim and the possibility that payment is required by a legal decision after it is denied. This more complicated situation may still be represented as a multistate model.

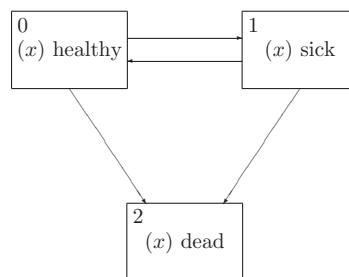


Fig. 20.4. Three-state illness-death model.

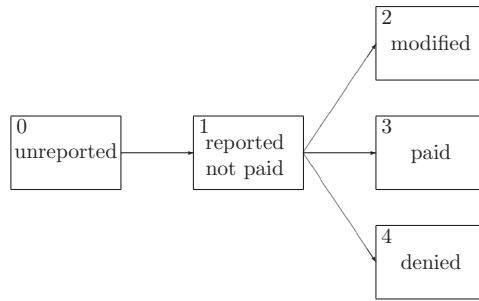


Fig. 20.5. Claims payment process model.

The application of multistate models to actuarial problems was introduced by Hoem (1969). Significant contributions to the analysis of present values in a multistate model framework were made by Waters (1984), Ramlau-Hansen (1988), and Norberg (1995). Haberman and Pitacco (1999) present a substantial review of the literature in their book *Actuarial Models for Disability Insurance*. In addition, a number of interesting actuarial applications have been studied, including critical illness insurance, which is discussed in a pair of papers by Macdonald et al. (2005a) and Macdonald et al. (2005b). Relevant statistical methods are reviewed by Macdonald (1996a,b).

20.2 Describing a Multistate Model

We use the notation, terminology, and assumptions in Dickson et al. (2009) and refer the reader to this textbook for additional details.

Let $\{Y(t), t \geq 0\}$ be a multistate stochastic process. It is simply a collection of random variables, $Y(t)$, each representing the state of the process at a given time t . We assume that the number of states, s , is finite, and the states are numbered $0, 1, \dots, s - 1$. For shorthand, we simply use Y to refer to the multistate process – the entire collection of random variables.

For Y to be useful to actuaries, we need a model that allows us to calculate probabilities associated with the process being in a particular state at a given future time. If we can do this, then we can generally perform the actuarial calculations required by our application.

It is often reasonable to assume that, at any time, the probabilities associated with the future of the process depend only on the current state. This assumption is referred to as the Markov property, and multistate processes that have this property are known as Markov processes or continuous-time Markov chains. We assume throughout this chapter that our multistate process is a Markov process. When it is not appropriate to assume that the Markov property holds, we can often construct an appropriate Markov

process using the “splitting of states” technique described by Haberman and Pitacco (1999).

For $i, j = 0, 1, \dots, s - 1$ and $z, t \geq 0$, we define the transition probability functions

$${}_t p_z^{ij} = \Pr\{Y(z + t) = j | Y(z) = i\}. \quad (20.1)$$

Notice that this is an extension of the life contingencies notation ${}_t p_z$, which represents the probability that an individual will survive to age $z + t$ given that the individual has survived to age z . In this context, it is common to view age z as the time origin. In our multistate modeling context, time 0 is the time origin, and z is just a number of years since time 0. This is important to recognize when our multistate model describes something different from the progression of an individual through a number of states. For example, consider the two-life model shown in Figure 20.2. At time 0, we have a life aged x and a life aged y . Suppose a joint life insurance policy is issued at this time, and we wish to value the policy 10 years later, when one life is aged $x + 10$ and the other is aged $y + 10$. Our calculations then involve probabilities denoted by ${}_t p_{10}^{00}$.

It is also useful to define the probabilities,

$${}_t p_z^{\bar{i}\bar{i}} = \Pr\{Y(z + w) = i \text{ for all } w \in [0, t] | Y(z) = i\}.$$

In words, ${}_t p_z^{\bar{i}\bar{i}}$ simply represents the probability that a process in state i at time z will remain in state i until at least time $z + t$. The probability ${}_t p_z^{\bar{i}\bar{i}}$ differs from (and is no larger than) the probability ${}_t p_z^{ii}$, because the latter includes the probability that a process in state i at time z will leave state i before time $z + t$ and return to state i by time $z + t$ (for multistate processes in which this can occur). For certain multistate processes, the probabilities ${}_t p_z^{\bar{i}\bar{i}}$ will help us in developing expressions for the ${}_t p_z^{ij}$ probabilities.

The transition intensity functions or, simply, transition intensities are also very useful. For $i, j = 0, 1, \dots, s - 1$ with $i \neq j$, and $z > 0$, define

$$\mu_z^{ij} = \lim_{h \rightarrow 0^+} \frac{{}_h p_z^{ij}}{h},$$

where the right-hand limit is used because h represents a time period and therefore must be positive. The transition intensities represent the instantaneous rates of transition from one state to another at a given time. They are analogous to the force of mortality and are sometimes referred to as the forces of transition. It is convenient to characterize a multistate model in terms of the transition intensities due to their interpretability. We often have a fairly clear idea of how these functions should behave, whereas the behavior of the transition probability functions may be less obvious.

The probabilities, ${}_t p_z^{\bar{i}i}$ can be expressed in terms of the transition intensities. We have

$${}_t p_z^{\bar{i}i} = \exp \left\{ - \int_0^t \sum_{j=0, j \neq i}^{s-1} \mu_{z+w}^{ij} dw \right\}. \quad (20.2)$$

Equation (20.2) is analogous to the well-known relation from life contingencies:

$${}_t p_z = \exp \left\{ - \int_0^t \mu_{z+w} dw \right\}.$$

This makes sense because ${}_t p_z^{\bar{i}i}$ is the probability of remaining in state i (rather than remaining alive), and $\sum_{j=0, j \neq i}^{s-1} \mu_{z+w}^{ij}$ is the total force of transition out of state i (rather than the force of mortality).

20.2.1 Calculating Transition Probabilities from Transition Intensities

In this section, we assume that we know the transition intensities, and we wish to determine the transition probability functions in order to perform an actuarial calculation. For processes in which repeat visits to states are not possible, it is relatively straightforward to determine expressions for the transition probability functions. We consider this case first, and then we present a convenient numerical approach that can be used when repeat visits are possible.

Consider, for example, the two-life model in Figure 20.2. Clearly,

$${}_t p_z^{00} = {}_t p_z^{\bar{0}\bar{0}},$$

$${}_t p_z^{11} = {}_t p_z^{\bar{1}\bar{1}},$$

$${}_t p_z^{22} = {}_t p_z^{\bar{2}\bar{2}},$$

and each of these functions can be determined using (20.2). To obtain an expression for ${}_t p_z^{01}$, we integrate over the possible times of transition from state 0 to state 1. We have

$${}_t p_z^{01} = \int_0^t {}_w p_z^{\bar{0}\bar{0}} \mu_{z+w}^{01} {}_{t-w} p_{z+w}^{\bar{1}\bar{1}} dw.$$

The idea here is that the process remains in state 0 from time z until time $z + w$, then makes a transition to state 1 during the (infinitesimally small) time period from $z + w$ to $z + w + dw$, and then remains in state 1 until time $z + t$. The integral then

sums over all of the small time intervals of length dw during which the transition may occur. Similarly,

$${}_t p_z^{02} = \int_0^t {}_w p_z^{\overline{00}} \mu_{z+w}^{02} {}_{t-w} p_z^{\overline{22}} dw.$$

Finally,

$${}_t p_z^{03} = 1 - {}_t p_z^{00} - {}_t p_z^{01} - {}_t p_z^{02},$$

$${}_t p_z^{13} = 1 - {}_t p_z^{11},$$

and

$${}_t p_z^{23} = 1 - {}_t p_z^{22}.$$

Note that it will often be necessary to evaluate the above integrals numerically. Dickson et al. (2009) suggest using repeated Simpson's rule for this.

An alternative approach to determining the transition probability functions from the transition intensities involves solving the Kolmogorov forward differential equations. This approach can be used even when repeat visits to states are possible. For $i, j = 0, 1, \dots, s-1$ and $z, t \geq 0$, the Kolmogorov forward differential equations are given by

$$\frac{d}{dt} {}_t p_z^{ij} = \sum_{k=0, k \neq j}^{s-1} ({}_t p_z^{ik} \mu_{z+t}^{kj} - {}_t p_z^{ij} \mu_{z+t}^{jk}). \quad (20.3)$$

These equations can be solved numerically using an approach called Euler's method. We begin by noting that, if (20.3) holds, then for small h

$$\frac{{}_t+h p_z^{ij} - {}_t p_z^{ij}}{h} \approx \sum_{k=0, k \neq j}^{s-1} ({}_t p_z^{ik} \mu_{z+t}^{kj} - {}_t p_z^{ij} \mu_{z+t}^{jk}),$$

and therefore,

$${}_t+h p_z^{ij} \approx {}_t p_z^{ij} + h \sum_{k=0, k \neq j}^{s-1} ({}_t p_z^{ik} \mu_{z+t}^{kj} - {}_t p_z^{ij} \mu_{z+t}^{jk}). \quad (20.4)$$

We can then approximate values of ${}_t p_z^{ij}$ by starting with

$${}_0 p_z^{ii} = 1 \text{ for } i = 0, 1, \dots, s-1,$$

and

$${}_0 p_z^{ij} = 0 \text{ for } i \neq j,$$

choosing a small value of h (e.g., $h = 1/12$), and using (20.4) to calculate ${}_h p_z^{ij}$ values, ${}_{2h} p_z^{ij}$ values, and so on. See Dickson et al. (2009) for examples of these calculations as well as a derivation of the Kolmogorov forward differential equations.

Since we can determine the probabilities we need from the transition intensities, it is sufficient to describe a multistate model by specifying the transition intensities. We next turn to the main focus of this chapter: estimating the transition intensities from data.

20.3 Estimating the Transition Intensity Functions

20.3.1 Our Objective

In this section, our goal is to estimate the transition intensities for a multistate model using data about the multistate process of interest. We recognize that our model must reflect the impact of various factors that may affect the transition intensities, and we need to be able to quantify our uncertainty about the resulting intensity estimates.

This uncertainty arises due to the possibility that any mathematical structure imposed on the transition intensities is incorrect and also due to incompleteness and variability associated with the data used. When we use parametric models for the transition intensities, a structure is imposed. Through appropriate methods of model selection and validation, we hope that this structure is reasonable. Variability associated with the data can be quantified by estimating the variances of the parameter estimators.

20.3.2 Data

For the purpose of estimating transition intensities, the ideal data would consist of a large number of complete realizations or outcomes of the process Y . That is, for each outcome of the process, we would know the value of $Y(t)$ from $t = 0$ until the process enters an absorbing state or until the end of the time horizon of interest. It is rarely the case that we have ideal data. In this chapter, we consider two possibilities in terms of the available data. In the first case, the process is observed continuously over some interval of time for each realization. In the second case, the process is observed only at distinct time points for each realization, as is the case with panel data. The latter case is discussed in Section 20.4.

Suppose that we are interested in estimating the transition intensity from state i to state j ; that is, μ_z^{ij} for a specific i and j . Now assume that our data consist of n outcomes of the process Y observed continuously over the time interval $[0, \tau]$. We denote these outcomes by y_1, y_2, \dots, y_n , where for outcome ℓ , the collection of observed states is $\{y_\ell(t), 0 \leq t \leq \tau\}$. Also, suppose that there are m distinct times at

which an $i \rightarrow j$ transition is observed. Let these times be denoted t_1, t_2, \dots, t_m , with $t_1 < t_2 < \dots < t_m$. Let d_w be the number of $i \rightarrow j$ transitions that occur at time t_w . That is, d_w is the number of outcomes for which an $i \rightarrow j$ transition occurs at time t_w . Note that, for our continuous-time process, the probability that two or more outcomes have transitions at the same time is assumed to be 0. However, in practice we are limited by the precision with which time is measured and recorded, and we may have more than one transition with a given recorded time. Finally, let R_w represent the number of outcomes for which the process is in state i just before time t_w . Notice that we have omitted i and j from the notation used here. This is for simplicity, but we must keep in mind that this notation corresponds to a particular $i \rightarrow j$ transition.

20.3.3 Nonparametric Models

A nonparametric model does not impose a mathematical structure on the transition intensities. Such a model is useful in exploring a data set graphically and in understanding the behavior of the transition intensities. While it is difficult to obtain a nonparametric model for the transition intensities themselves, the well-known Nelson-Aalen (NA) estimator introduced in Chapter 19 provides nonparametric models for their integrals. Specifically, we can estimate the cumulative intensity function

$$M^{ij}(t) = \int_0^t \mu_z^{ij} dx.$$

Omitting the ij superscripts, the NA estimator of $M(t)$ is given by

$$\widehat{M}(t) = \sum_{w:t_w \leq t} \frac{d_w}{R_w}. \quad (20.5)$$

The NA estimator is consistent ($\widehat{M}(t)$ approaches $M(t)$ as n gets large), and although it is biased downward, it is asymptotically unbiased and asymptotically normally distributed. The reason it is biased downward is that there is a nonzero probability that, during a given time interval, the process will not be in state i for any of the n realizations. In this case, there can be no $i \rightarrow j$ transitions during this time interval, even though the transition intensity may be quite high. A large sample variance estimator is given by

$$\widehat{\text{Var}}[\widehat{M}(t)] = \sum_{w:t_w \leq t} \frac{d_w}{R_w^2}. \quad (20.6)$$

The NA estimator is associated with survival analysis, which is discussed in Chapter 19. Standard textbooks on the subject include Kalbfleisch and Prentice (2002) and Lawless (2003), and a thorough treatment of the methods applied to multistate models is provided by Andersen (1993). The NA estimator was proposed by Nelson (1969)

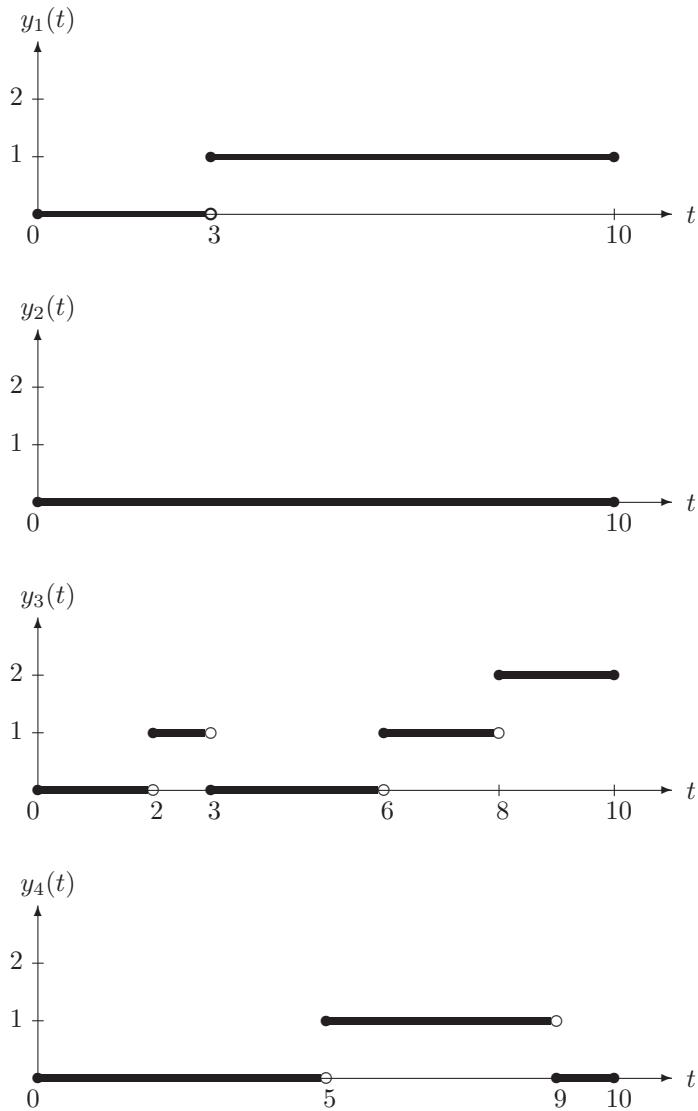


Fig. 20.6. Example outcomes of the three-state illness-death process.

as a graphical method, and its properties were investigated by Aalen (1978). The NA estimator is also discussed by Klugman et al. (2012).

To illustrate the method, we introduce a simple example.

Example 20.1. Consider the illness-death process shown in Figure 20.4 and imagine that we observe the four outcomes given in Figure 20.6. Suppose we are interested in the transition intensity from state 0 to state 1.

We summarize the information necessary to calculate the NA estimates in the following table. In this example we have just four distinct times at which $0 \rightarrow 1$ transitions occur.

w	t_w	R_w	d_w
1	2	4	1
2	3	3	1
3	5	3	1
4	6	2	1

And according to (20.5), the NA estimates are

$$\widehat{M}(t) = \begin{cases} 0, & t < 2 \\ \frac{1}{4}, & 2 \leq t < 3 \\ \frac{1}{4} + \frac{1}{3} = \frac{7}{12}, & 3 \leq t < 5 \\ \frac{7}{12} + \frac{1}{3} = \frac{11}{12}, & 5 \leq t < 6 \\ \frac{11}{12} + \frac{1}{2} = \frac{17}{12}, & 6 \leq t < 10 \end{cases} .$$

The computations are easily performed using functions in the `survival` package in R. The following commands from an R session produce the plot in Figure 20.7. Notice that the survival data object contains six observations, even though we have observed just four outcomes of the process. This is because the four outcomes include a total of six visits to state 0. Our survival data object requires an observation for each interval of time spent in state 0 in order to correctly calculate the R_w and d_w values for each transition time t_w . The `survfit` function determines the estimates.

```
> library(survival) # load survival package
Loading required package: splines
> # Create the survival data object.
> start <- c(0,0,0,3,0,9) # interval start times
> end <- c(3,10,2,6,5,10) # interval end times
> event <- c(1,0,1,1,1,0) # indicator of 0->1 transitions
> data.surv <- Surv(start,end,event)
> data.surv
[1] (0, 3]  (0,10+] (0, 2]  (3, 6]  (0, 5]  (9,10+]
> # Calculate and plot NA estimates.
> fit.NA <- survfit(data.surv~ 1,type="f1")
> plot(fit.NA, fun="cumhaz",xlab="t",ylab="M(t)",
+ main="NA Estimates",conf.int=FALSE)
```

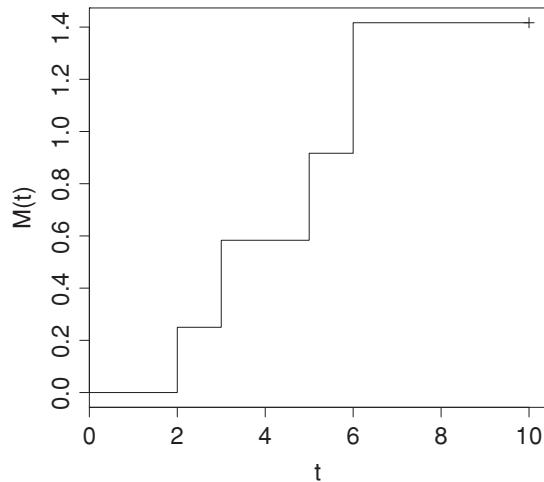


Fig. 20.7. Nelson-Aalen estimates of the cumulative transition intensity from state 0 to state 1 in the illness-death model example.

Example 20.2. In this example, we consider real data from a pension plan. Plan members are observed over two consecutive calendar years. For each individual who was a member of the plan at any time during a given calendar year, we observe the date of birth, the date of hire, the date of termination (if terminated), the sex of the individual, and the annual salary during the year. The dataset contains a total of 725 observations, and 418 unique members were observed – many were observed in both years.

Suppose we are interested in estimating the intensity of termination for reasons other than death. We can gain an understanding of the behavior of this function by plotting the NA estimates of the cumulative intensity of termination (as a function of years of service) separately for males and females. These estimates are shown in Figure 20.8 along with approximate 95% confidence intervals. The confidence intervals are based on the variance estimates obtained from (20.6) along with a normal distribution assumption.

The graph suggests that males have a significantly lower cumulative intensity of termination than females. We can justify this conclusion by noting that we are unable to hypothesize a single cumulative intensity function for both genders that lies mostly between the confidence limits for males and females. Also, the concave behavior of the estimated cumulative intensity of termination suggests that the intensity of termination is a decreasing function (of years of service).

Note that, in using the NA estimator in this example, we must assume that termination for reasons other than death occurs independently of other types of termination (e.g., death). This assumption is required more generally – competing transitions must occur independently of one another.

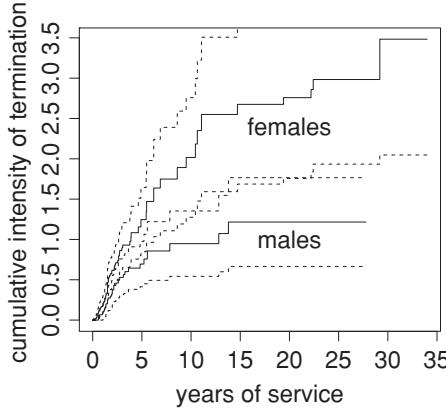


Fig. 20.8. Nelson-Aalen estimates of the cumulative intensity of termination for females and males along with approximate 95% confidence intervals.

20.3.4 Parametric Models

Unlike a nonparametric model, a parametric model imposes a mathematical structure and involves a fixed (usually small) number of parameters that must be estimated. Once we have an understanding of the behavior of the transition intensities, we can choose a suitable parametric model for each. The parameters can be estimated by the method of maximum likelihood.

To use maximum likelihood estimation, we need to first write the likelihood function, a function of the parameters to be estimated. The likelihood function is proportional to the probability or probability density associated with observing the outcomes we actually observed. Assuming that the outcomes are independent, the likelihood function is then the product of the likelihood contributions associated with the observed outcomes. The likelihood contribution of the ℓ th outcome is the conditional probability density of the observed outcome given the initial time and state. The likelihood function is then

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{\ell=1}^n L_\ell(\boldsymbol{\theta}) \\ &= \prod_{\ell=1}^n \Pr\{Y(t) = y_\ell(t), 0 < t \leq \tau | Y(0) = y_\ell(0), \boldsymbol{\theta}\}, \end{aligned} \quad (20.7)$$

where $\boldsymbol{\theta}$ is a vector of parameters to be estimated. Note that the probability in expression (20.7) is actually a probability density; the probability of a specific outcome of the process is 0. The calculation of the required probability densities is illustrated in the next example.

Standard asymptotic likelihood theory can be used to make inferences about $\boldsymbol{\theta}$ and functions of $\boldsymbol{\theta}$, provided that our sample of outcomes is sufficiently large. An overview

of statistical methods for multistate models, including likelihood construction, as well as an extensive list of references is given by Andersen and Keiding (2002).

Example 20.3. We now consider the likelihood construction for the data in Example 20.1, shown in Figure 20.6. For the first outcome, the process is in state 0 from time 0 until time 3, when a transition to state 1 occurs. The process then remains in state 1 until time 10. The likelihood contribution is then the conditional probability density of the process remaining in state 0 until time 3, moving to state 1 at time 3, and then staying in state 1 until time 10, given that the process is in state 0 at time 0. According to our multistate model notation, this density is

$$L_1(\boldsymbol{\theta}) = {}_3p_0^{\overline{00}} \mu_3^{01} {}_7p_3^{\overline{11}}.$$

Using similar arguments, the other likelihood contributions are

$$L_2(\boldsymbol{\theta}) = {}_{10}p_0^{\overline{00}},$$

$$L_3(\boldsymbol{\theta}) = {}_2p_0^{\overline{00}} \mu_2^{01} {}_1p_2^{\overline{11}} \mu_3^{10} {}_3p_3^{\overline{00}} \mu_6^{01} {}_2p_6^{\overline{11}} \mu_8^{12},$$

and

$$L_4(\boldsymbol{\theta}) = {}_5p_0^{\overline{00}} \mu_5^{01} {}_4p_5^{\overline{11}} \mu_9^{10} {}_1p_9^{\overline{00}}.$$

The vector $\boldsymbol{\theta}$ depends on the parametric models we have chosen for our transition intensities. In the simplest case, each transition intensity is assumed to be constant. That is, $\mu_z^{ij} = \mu^{ij}$ and we have the parameter vector $\boldsymbol{\theta} = (\mu^{01}, \mu^{02}, \mu^{10}, \mu^{12})$. Using equation 20.2, the likelihood function is, after simplification,

$$L(\boldsymbol{\theta}) = \exp\{-24(\mu^{01} + \mu^{02}) - 14(\mu^{10} + \mu^{12})\} (\mu^{01})^4 (\mu^{10})^2 \mu^{12}. \quad (20.8)$$

The maximum likelihood estimates are obtained in the usual way by taking the log of the likelihood function, differentiating the log-likelihood with respect to each parameter, and setting these derivatives to 0. We have

$$\begin{aligned}\hat{\mu}^{01} &= \frac{4}{24} = \frac{1}{6}, \\ \hat{\mu}^{10} &= \frac{2}{14} = \frac{1}{7}, \\ \hat{\mu}^{12} &= \frac{1}{14}.\end{aligned}$$

The MLE of μ^{02} cannot be found using this approach, because the likelihood is a strictly decreasing function of μ^{02} . The likelihood is then maximized when this parameter equals 0, the left endpoint of the range of possible values. Thus,

$$\hat{\mu}^{02} = 0.$$

Also, reasoning intuitively, we observe the process in state 0, but no transitions from state 0 to state 2 are observed. So a transition intensity estimate of 0 makes sense. It is interesting to note that, when the transition intensities are constant as in this example,

$$\hat{\mu}^{ij} = \frac{\text{the total number of } i \rightarrow j \text{ transitions}}{\text{the total time spent in state } i},$$

where both totals are taken over all observed outcomes.

20.3.5 Models Involving Explanatory Variables

Often there are several explanatory variables or covariates that affect the behavior of the transition intensities. The methods of survival analysis provide us with approaches for modeling the impact of these covariates on our transition intensities. The most popular of these methods involves the proportional hazards model. In our context, this model assumes that the transition intensities from state i to state j for different values of the covariates are proportional. That is,

$$\mu_{t|\mathbf{x}(t)}^{ij} = \mu_{t,0}^{ij} \exp\{\boldsymbol{\beta}' \mathbf{x}(t)\}, \quad (20.9)$$

where $\mathbf{x}(t)$ is a q -dimensional vector of covariates whose components may change as t changes, $\boldsymbol{\beta}$ is a q -dimensional vector of parameters to be estimated, and $\mu_{t,0}^{ij}$ is called the baseline transition intensity. The exponential function guarantees that the baseline intensity is multiplied by a positive number. This model is suitable only if the transition intensities are indeed proportional for different values of the covariate vector. Since $\mathbf{x}(t)$ depends on t , we are allowing the covariates to be time-varying. Time-varying covariates are quite common in practice, and the next example shows how they can be handled in estimating the model parameters.

One of the appealing features of the proportional hazards model is that the parameter vector $\boldsymbol{\beta}$ can be estimated without estimating the baseline intensity. This is described in Chapter 19. This aspect of the model is convenient when one wants only to investigate the impact of the covariates. The proportional hazards model can be fit using the `coxph` function in the `survival` package in R, and standard likelihood methods can be used to make inferences about the parameter vector.

Example 20.4. Consider the pension plan data in Example 20.2. We wish to model the intensity of termination as a function of time since hire. Once again, a total of 418 plan members are observed over two consecutive calendar years. An observation is available for each member in each calendar year, resulting in 725 observations, because many members were observed in both years. The following is taken from an R session in which we investigate the impact of several covariates on the intensity of termination as a function of time since hire. In this R session, `start.time` is a vector (of length 725) giving the time since hire at which we begin to observe each member

during the calendar year. The vector `end.time` gives the time since hire at which we stop observing each member during the calendar year, and the vector `end.event` contains a 1 for each member who terminated at the time given in `end.time` and a 0 for each member who did not. The three vectors `start.time`, `end.time`, and `end.event` are used to create a survival data object corresponding to the 725 observations.

```
> library(survival)
Loading required package: splines
> term.surv <- Surv(start.time,end.time,end.event)
```

The values of the covariates we consider are contained in the vectors `hire.age` (the age at hire), `sex` (F or M), `salary` (the annual salary during the calendar year), and `year` (0 for the first calendar year and 1 for the second calendar year). We use R to estimate β in the proportional hazards model in (20.9), with these covariates making up the components of the vector $\mathbf{x}(t)$ for each observation.

Note that the annual salary of the member is a time-varying covariate. However, it is assumed to be constant within each calendar year. The data are structured in such a way that there is an observation for each member in each of the two calendar years. So, for the interval of time represented by each observation, the salary is constant and can therefore be treated as a fixed covariate, as can the other covariates.

Estimation of the components of β is performed by the `coxph` function, and its output is summarized by the `summary` function. Only part of the information produced by the `summary` function is displayed next. In specifying the model, we take the logarithm of the salaries because the salary distribution is highly skewed and we obtain poor results if the salary values themselves are used.

```
> term.coxpath1 <- coxph(term.surv~hire.age+sex+log(salary)+year)
> summary(term.coxpath1)

Call:
coxph(formula = term.surv ~ hire.age + sex + log(salary) + year)

n= 725, number of events= 96

            coef      exp(coef)    se(coef)      z     Pr(>|z| )
hire.age   -0.003913   0.996095   0.011859   -0.330   0.74145
sexM       -0.120873   0.886146   0.265267   -0.456   0.64863
log(salary) -0.737475   0.478320   0.231664   -3.183   0.00146**
year        0.108676   1.114801   0.208062    0.522   0.60144
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1
```

The results show that 96 of the 725 observation intervals ended with a termination. In the table provided by the summary function, we see the estimates of the β components corresponding to the four covariates in the column headed `coef` (for coefficient), as well as the exponentiated values of these estimates and the standard errors associated with the estimates. The final two columns provide the information necessary to perform individual tests of the hypotheses that the β components are 0, implying that the covariate is not statistically significant in modeling the intensity of termination. The column headed `z` gives the values of the coefficients divided by their standard errors. Under the hypothesis that a given coefficient is 0, the corresponding value of this test statistic is distributed (asymptotically) standard normal. Under this assumption, the final column of the table then gives the *p*-values of the hypotheses.

The information in the above table suggests that only `log(salary)` has a significant effect on the intensity of termination. This seems to contradict our earlier observation that this intensity function is different for males and females. Further analysis reveals a relationship between sex and salary: the male salaries are generally higher, and it appears that sex is not an important variable when the effect of salary is reflected.

When we fit a proportional hazards model with just the covariate `log(salary)`, we obtain the following.

```
> term.coxpath2 <- coxph(term.surv ~ log(salary))
> summary(term.coxpath2)
Call:
coxph(formula = term.surv ~ log(salary))

n= 725, number of events= 96

            coef exp(coef)  se(coef)      z Pr(>|z|)
log(salary) -0.8146    0.4428   0.1815 -4.489 7.16e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1
```

We can perform a likelihood ratio test of the hypothesis that the coefficients corresponding to age at hire, sex, and year are all 0. The maximum log-likelihood obtained with all four variables in the model is -381.9689 (obtained from `term.coxph`), and the maximum log-likelihood obtained with just `log(salary)` in the model is -382.2580 (obtained from `term.coxpath2`). The likelihood ratio statistic is then $2(382.2580 - 381.9689) = 0.5781$. When comparing this with the quantiles of a chi-squared distribution with 3 degrees of freedom, we conclude that there is no evidence against the hypothesis (*p*-value = 0.9014).

When including covariates in a model, the model may be improved by using some function of the covariate value, rather than the value itself. Since `log(salary)` was

chosen rather arbitrarily to address the fact that the salary distribution is skewed, we explore various other functions of salary. We find that the log-likelihood is larger (equal to -378.8770) when we use an indicator that the salary is greater than 36,200, rather than $\log(\text{salary})$. We obtain the following in this case

```
> high.sal.ind <- as.numeric(salary>36200)
> term.coxpath3 <- coxph(term.surv~high.sal.ind)
> summary(term.coxpath3)

Call:
coxph(formula = term.surv ~ high.sal.ind)

n= 725, number of events= 96

            coef  exp(coef)   se(coef)      z Pr(>|z|)    
high.sal.ind -1.1347     0.3215    0.2125 -5.341 9.25e-08 *** 
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1    1
```

The intensity of termination is then

$$\mu_{t|\text{salary}} = \mu_{t,0} \exp\{\beta I(\text{salary} > 36,200)\}.$$

And we find that $\hat{\beta} = -1.1347$ so that $\exp\{\hat{\beta}\} = 0.3215$. This indicates that plan members in the high-salary group have a termination rate that is about 32% of those in the low-salary group. The standard error estimate of our estimator of β is 0.2125.

It is important to remember that the appropriateness of this analysis depends on the reasonableness of the proportional hazards assumption. Chapter 19 discusses how to check this assumption.

Although it is convenient to be able to investigate the impact of covariates without worrying about the baseline transition intensity, we require fully specified transition intensities if we wish to perform calculations. Nonparametric models can be used for the baseline intensities. However, parametric models may be preferred because of their simplicity. A fully parametric proportional hazards model may be suitable. In this case, we simply need to specify the baseline intensities of transition up to a small number of parameters to be estimated.

The well-known accelerated failure-time model can also be used in the context of estimating a transition intensity. It is a fully parametric model that reflects the impact of covariates. Although it is normally presented as an expression for the survival function,

we can write the following nonstandard expression for the transition intensity:

$$\mu_{t|\mathbf{x}(t)}^{ij} = f_0\left(\frac{\log t - u(\mathbf{x}(t))}{\sigma}\right) \Big/ \left[\sigma t S_0\left(\frac{\log t - u(\mathbf{x}(t))}{\sigma}\right)\right], \quad (20.10)$$

where $S_0(\cdot)$ is a fully specified survival function of a random variable that can take any real value, $f_0(\cdot)$ is the corresponding probability density function, $u(\mathbf{x}(t)) = \beta_0 + \boldsymbol{\beta}'\mathbf{x}(t)$, and $\sigma > 0$ is a parameter to be estimated, as are β_0 and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)'$. This specification of the model allows for time-dependent covariates, which are not generally used with the accelerated failure-time model. In fact, the accelerated failure-time interpretation is lost when time-dependent covariates are present. Cox and Oakes (1984) elaborate on this.

Common choices for $S_0(\cdot)$ include the survival functions of the extreme value, logistic, and normal distributions. These lead to Weibull, loglogistic, and lognormal transition intensities, respectively.

Example 20.5. We fit a number of fully parametric regression models to the pension plan termination data. We show results for two of them in this example.

The first is a proportional hazards model with a parametric form specified for the baseline intensity of termination. Since we expect the intensity of termination to decrease with increasing service, we consider an exponentially decreasing baseline intensity. We refer to this as a Gompertz intensity, even though it is normally increasing. The single covariate in the model is $I(\text{salary} > 36,200)$, which is 1 when the salary is greater than 36,200 and 0 otherwise. Thus we have the model

$$\mu_{t|\text{salary}} = \exp\{\beta_0 + \beta_1 I(\text{salary} > 36,200)\}c^t,$$

where t is the time since hire and β_0 , β_1 , and c are parameters to be estimated, with $-\infty < \beta_0, \beta_1 < \infty$, and $0 < c < 1$, respectively. Constructing the likelihood, and maximizing with respect to the three parameters, gives the estimates $\hat{\beta}_0 = -0.4997$, $\hat{\beta}_1 = -1.0531$, and $\hat{c} = 0.8425$. The resulting cumulative intensity of termination is shown in Figure 20.9, which indicates that the fit is rather poor.

The second parametric model we discuss is a lognormal accelerated failure-time model with the same covariate. For this model,

$$\mu_{t|\text{salary}} = \frac{\phi\left(\frac{\log t - \beta_0 - \beta_1 I(\text{salary} > 36,200)}{\sigma}\right)}{\sigma t \left[1 - \Phi\left(\frac{\log t - \beta_0 - \beta_1 I(\text{salary} > 36,200)}{\sigma}\right)\right]},$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal pdf and cdf, respectively, and β_0 , β_1 , and σ are parameters, with $-\infty < \beta_0, \beta_1 < \infty$, and $\sigma > 0$. The maximum likelihood estimates of these parameters are $\hat{\beta}_0 = 0.4669$, $\hat{\beta}_1 = 1.2997$, and $\hat{\sigma} = 1.1788$. The resulting cumulative intensity of termination is shown in Figure 20.9 and appears to fit somewhat better than that of the proportional hazards model with Gompertz

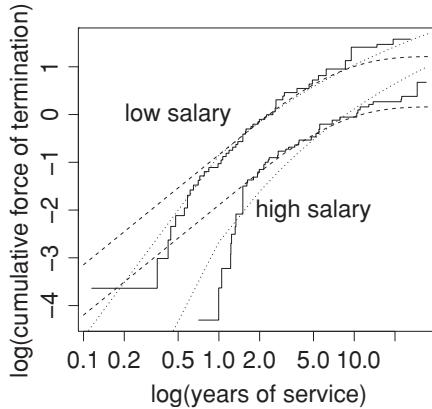


Fig. 20.9. Fitted cumulative forces of termination for low-salary and high-salary employees based on a proportional hazards model with Gompertz baseline force (dashed) and a lognormal accelerated failure-time model (dotted).

baseline intensity for the early months of service, but the lognormal model may not be satisfactory.

Further analysis is required to determine a suitable parametric model for the force of termination for this pension plan.

A variety of graphical methods and formal statistical tests can be used in selecting an appropriate parametric model and checking its fit. These are discussed by Lawless (2003) and Klugman et al. (2012).

20.4 Estimating the Transition Intensities with Outcomes Observed at Distinct Time Points

The estimation methods discussed in the previous section assume that the multistate process is observed continuously over some interval of time. In practice, it may not be possible to observe continuously. Instead, the state of the process may be observed only at specific points in time. This is the case with *panel data*, which are discussed in Chapter 7. We can still estimate the parameters of our multistate model using maximum likelihood estimation, but there is greater uncertainty associated with our estimates. The likelihood contribution of a given outcome is the probability that the process will be in the observed state at the second and later observation times given its state at the first observation time. This is illustrated in the following example.

Example 20.6. Consider the data in Example 20.1 (see Figure 20.6). Now suppose that each outcome of the process is observed only at times 0, 6, and 10. The data are shown in Table 20.1.

Table 20.1. *Example Outcomes of
Three-State Illness Death Process When the
State is Observed at Times 0, 6, and 10*

Outcome (ℓ)	$y_\ell(0)$	$y_\ell(6)$	$y_\ell(10)$
1	0	1	1
2	0	0	0
3	0	1	2
4	0	1	0

The likelihood contribution corresponding to the first outcome is the conditional probability that the process is in state 1 at time 6 and at time 10, given that it is in state 0 at time 0. That is,

$$L_1(\boldsymbol{\theta}) = \Pr\{Y(6) = 1, Y(10) = 1 | Y(0) = 0\}.$$

We can write this conditional probability as

$$\begin{aligned} & \Pr\{Y(6) = 1, Y(10) = 1 | Y(0) = 0\} \\ &= \Pr\{Y(6) = 1 | Y(0) = 0\} \Pr\{Y(10) = 1 | Y(0) = 0, Y(6) = 1\} \\ &= \Pr\{Y(6) = 1 | Y(0) = 0\} \Pr\{Y(10) = 1 | Y(6) = 1\}, \end{aligned}$$

where the last step follows from the Markov property. Therefore,

$$L_1(\boldsymbol{\theta}) = {}_6 p_0^{01} {}_4 p_6^{11}.$$

Similarly,

$$L_2(\boldsymbol{\theta}) = {}_6 p_0^{00} {}_4 p_6^{00},$$

$$L_3(\boldsymbol{\theta}) = {}_6 p_0^{01} {}_4 p_6^{12},$$

and

$$L_4(\boldsymbol{\theta}) = {}_6 p_0^{01} {}_4 p_6^{10}.$$

Suppose, again, that the transition intensities are constant such that $\boldsymbol{\theta} = (\mu^{01}, \mu^{02}, \mu^{10}, \mu^{12})$. Then, given estimates of these intensities, the likelihood contributions can be determined. The likelihood function $L(\boldsymbol{\theta}) = \prod_{\ell=1}^4 L_\ell(\boldsymbol{\theta})$ can be maximized using numerical methods, resulting in the estimates

$$\hat{\mu}^{01} = 0.2425,$$

$$\hat{\mu}^{02} = 0,$$

$$\hat{\mu}^{10} = 0.1554,$$

Table 20.2. Maximum Likelihood Estimates of Transition Intensities in the Illness-Death Example Assuming Constant Intensities and Several Assumptions about the Observation Times

Transition	Set of Observation Times			
	{0, 6, 10}	{0, 2, 4, 6, 8, 10}	{0, 1, 2, ..., 10}	[0, 10]
0 → 1	0.2425	0.2599	0.1995	0.1667
0 → 2	0	0	0	0
1 → 0	0.1554	0.2442	0.1781	0.1429
1 → 2	0.0555	0.0664	0.0689	0.0714

and

$$\hat{\mu}^{12} = 0.0555.$$

These estimates are quite different from those obtained when we observe each outcome continuously from time 0 to time 10. Table 20.2 shows the MLEs obtained with several sets of observation times. It helps us better understand the impact of the information loss associated with distinct observation times.

The variability associated with the maximum likelihood estimates can be assessed using standard asymptotic likelihood theory. If one maximizes the log-likelihood by using the `optim` function in R with the optional argument `hessian = TRUE`, then the output provides the observed information matrix evaluated at the maximum likelihood estimates. This can be used to obtain variance and covariance estimates for the maximum likelihood estimators. A good discussion of asymptotic likelihood theory is provided by Lawless (2003). Asymptotic methods are not suitable for our simple example involving just four outcomes of the multistate process.

References

- Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics* 6(4), 701–726.
- Andersen, P. K. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.
- Andersen, P. K. and N. Keiding (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research* 11(2), 91–115.
- Cox, D. D. R. and D. Oakes (1984). *Analysis of Survival Data*, Volume 21. CRC Press, Boca Raton, FL.
- Dickson, D. C., M. R. Hardy, and H. R. Waters (2009). *Actuarial Mathematics for Life Contingent Risks*. Cambridge University Press, Cambridge.
- Haberman, S. and E. Pitacco (1999). *Actuarial Models for Disability Insurance*. CRC Press, Boca Raton, FL.

- Hoem, J. M. (1969). Markov chain models in life insurance. *Blätter der DGVFM* 9(2), 91–107.
- Kalbfleisch, J. D. and R. L. Prentice (2002). *The Statistical Analysis of Failure Time Data* (2nd ed.). John Wiley & Sons, New York.
- Klugman, S. A., H. H. Panjer, and G. E. Willmot (2012). *Loss Models: From Data to Decisions* (4th ed.), Volume 715. John Wiley & Sons, New York.
- Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data* (2nd ed.). John Wiley & Sons, New York.
- Macdonald, A. (1996a). An actuarial survey of statistical models for decrement and transition data-i: Multiple state, Poisson and binomial models. *British Actuarial Journal* 2(1), 129–155.
- Macdonald, A. (1996b). An actuarial survey of statistical models for decrement and transition data: II: Competing risks, non-parametric and regression models. *British Actuarial Journal*, 429–448.
- Macdonald, A. S., H. R. Waters, and C. T. Wekwete (2005a). A model for coronary heart disease and stroke with applications to critical illness insurance underwriting I: The model. *North American Actuarial Journal* 9(1), 13–40.
- Macdonald, A. S., H. R. Waters, and C. T. Wekwete (2005b). A model for coronary heart disease and stroke with applications to critical illness insurance underwriting II: Applications. *North American Actuarial Journal* 9(1), 41–56.
- Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology* 1(1).
- Norberg, R. (1995). Differential equations for moments of present values in life insurance. *Insurance: Mathematics and Economics* 17(2), 171–180.
- Ramlau-Hansen, H. (1988). The emergence of profit in life insurance. *Insurance: Mathematics and Economics* 7(4), 225–236.
- Waters, H. R. (1984). An approach to the study of multiple state models. *Journal of the Institute of Actuaries* 111(2), 363–374.

Index

adjacency matrix, 263, 266, 274
asymptotic covariance matrix, 246, 256

backward elimination, 76
baseline choice, *see* reference category
baseline hazard, 476, 497
Bayes' theorem, 336
Bayesian approach for mixed models, 411
Bayesian credible intervals, 255
best linear unbiased predictor (BLUP), 197, 222
bootstrap, 465

case control, 162
check function, 254
claim frequency, 273
claim severity, 275
 average claim size, 269–271, 276
 individual claim size, 275–276
claims payment process, 517
classification table, 74
coefficient of determination, *see* R^2
collinearity, 18, 27, 30, 32, 51
 VIF, 30, 32
conditional mean, 251
conditional quantile, 253
conditionally autoregressive model, *see* spatial
 autoregression
confidence intervals, 316
contagion, true and apparent, 97
covariance matrix, 192
 autoregressive, 193
 diagonal, 193
 nondiagonal, 193
 Toeplitz, 193
 unstructured, 193
covariates, 529
 time-varying, 529
credibility, 173, 217, 334
 Bühlmann model, 207, 220
 Bühlmann-Straub model, 208, 221
 Hachemeister model, 221
 multidimensional, 189

robust, 229
shrinkage estimator, 198

data
 clustered, 180, 183
 count, 87
 cross-classified, 190
 cross-sectional, 183
 grouped, 146
 hierarchical, 183, 188
 individual, 146
 longitudinal, 167, 219
 missing, 180
 multilevel, 180, 183
 nested, *see also* hierarchical
 non-Gaussian, 399
 non-nested, *see* cross-classified
 panel, 167, 183, 534
 spatial, 261
 survival, 482
 time series, 427
 transitions, 515, 522
 two-part, 141
 unbalanced, 171, 180
data example
 Annual Flow of the River Nile, 429
 Australian Automobile Bodily Injury Scheme, 469
 Australian Workers' Compensation Insurance, 475
 Automobile PIP Questionable Claims Data, 284
 California Auto Assigned Risk (CAARP) Data, 284
 Credit Insurance Data, 190, 212
 Cross Selling, 122
 German Car Insurance Data, 262, 269, 276
 Group Term Life, 175
 Hachemeister Data, 187, 206, 218, 224
 Heart Disease Data, 373, 390, 394
 Losses and Expenses, 142
 Massachusetts Automobile Claims, 152, 242
 Medical Expenditure Panel Survey (MEPS), 15, 247, 254
 Mexican Central Bank, 346, 353, 356

- data example (*cont.*)
 Mexico Insurance Market, 342, 361
 Monthly Global Temperature, 429
 Motor Insurance, 69, 74, 80, 83
 National Nursing Home Survey (NNHS), 482, 486,
 513
 pension plan terminations, 526, 529, 533
 Personal Automobile Claims, 316
 Quarterly Earnings of Johnson & Johnson, 429
 Sales Data from Box-Jenkins, 429
 Singapore Automobile Claims, 91
 Texas Closed Claim Data, 284
 U.S. Medical Malpractice Insurance, 475, 476
 U.S. Workers' Compensation Insurance, 187, 190,
 201, 412
 dependent lifetimes
 common shock, 516
 dimension reduction, 282
 distribution tail
 criteria of subexponential, 238
 left tail, 238
 right tail, 237
 tail heaviness, 237
 distributions
F –, 244
 asymmetric Laplace, 254
 Bernoulli, 66
 beta prime, *see* generalized beta of the second kind
 (GB2)
 binomial, 88, 112, 133
 bivariate normal, 170
 Burr XII, 245
 exponential, 507
 exponential GB2, 245
 exponential family, 110, 146, 161, 238, 241
 extreme value, 533
 fat-tailed distributions, 237
 gamma, 100, 112, 134, 146, 237, 242, 245
 generalized beta, *see* generalized beta of the second
 kind (GB2)
 generalized beta of the second kind (GB2), 244
 generalized gamma, 245
 geometric, 94
 inverse-Gaussian, 100, 112, 242
 Laplace, 252
 location-scale, 227
 log location-scale family, 244
 log-normal, 100, 245
 logistic, 507, 533
 negative binomial, 93
 negative binomial *p* (NB*p*), 93
 normal, 21, 112, 237, 533
 over-dispersed Poisson (ODP), 454
 Pareto, 161, 237, 244, 245
 Poisson, 87, 111
 Poisson-inverse Gaussian, 94
 Poisson-lognormal, 95
 Tweedie, 128, 150
 Weibull, 228, 237, 245, 507, 533
 zero-inflated, 102
 zero-inflated Poisson (ZIP), 102
- estimation
 corrected adaptively truncated likelihood (CATL),
 228
 generalized least squares (GLS), 196
 Laplace method, 404
 least absolute deviations (LAD), 251
 least squares, 21, 241
 maximum likelihood, 7, 70, 72, 90, 93, 95, 110, 118,
 135, 149, 177, 241, 527, 528, 534
 minimum bias technique, 6, 87
 penalized least-square, 387
 penalized quasi-likelihood (PQL), 406
 pseudo-likelihood (PL), 406
 restricted maximum likelihood (REML), 199, 219,
 228
 restricted pseudo-likelihood (REPL), 406
 robust, 228
 exchangeability, 355, 356
 expected utility, 336
 exposure, 102, 144, 187, 221
 false negative, 75
 false positive, 75
 Fisher information matrix, 246, 256
 full conditional distribution, 265
 Gelman-Rubin statistic, 322
 generalized estimating equation (GEE), 399
 generalized linear mixed model (GLMM), 400
 generalized linear model (GLM), 7, 60, 107, 141, 238,
 241, 338, 369, 399, 455, 462
 gamma regression, 117, 126, 150, 241, 249,
 275–276
 linear model, 338
 link function, 68, 90, 109, 142, 369, 399
 logistic regression, 67, 149, 345
 mixed, 179
 offset, 124, 142
 Poisson regression, 89, 150, 241, 274, 352
 probit regression, 78
 systematic component, 369
 Tweedie, 128, 150
 goodness-of-fit statistic, 75, 243
 generalized cross-validation (GCV), 383
 adjusted R^2 , 32, 33, 42
 mean squared error, 381
 predictive squared error, 382
 pseudo- R squared, 76
 R^2 , 25, 30, 33, 52
- hypothesis testing, 316
- influential point
 Cook's distance, 40, 41, 42
 leverage, 34, 36
 outlier, 34, 36, 37, 38, 40
- information statistic
 Akaike information criterion, *AIC*, 32, 33, 42, 52,
 98, 132, 250, 383
 Akaike information criterion, corrected, *AICC*, 132
 Bayesian information criterion, *BIC*, 98, 132, 250

- deviance information criterion, *DIC*, 277, 342, 348, 393
 Kullback-Leibler, 241
 Schwarz Bayesian criterion
 see Akaike information criterion, *AIC*, 132
- Kolmogorov forward differential equations, 521
 numerical solution using Euler's method, 521
- learning
 supervised, 280
 unsupervised, 281
- life annuity and insurance, 516
- log-odds, 68
- logistic regression, *see* generalized linear model (GLM)
- logit, 68, 390
- loss reserving, 3, 399, 450
 chain ladder algorithm, 453
 generalized linear model (GLM), 455
 models of individual claims, 467
 over-dispersed Poisson (ODP) cross-classified model, 454, 463
 over-dispersed Poisson (ODP) Mack model, 454
- MSE*, *see* residual variance
- Markov chain Monte Carlo, 320
 Gibbs sampler, 320
 Metropolis-Hastings sampler, 320
- Markov chain Monte Carlo methods, 267, 276
- Markov process, 518
- Markov random field, 265
- median regression, 360
- medoid, 301
- meta-analysis, 356
- model
 GB2 regression, 244, 245, 249
 accelerated failure times, 532, 533
 Bayesian approach for median regression, 252
 Bayesian quantile regression, 254
 Burr XII regression, 249
 cluster analysis, 283
 complete pooling, 183, 201
 conditional logit, 81, 84
 cross-sectional, 182
 cumulative logit, 79
 factor analysis, 283
 frequency-severity, 148
 generalized additive model (GAM), 60, 242, 369, 388, 456
 generalized logit, 81
 hierarchical, 185, 273, 355
 inverse-Gaussian regression, 249
 linear fixed effects, 173, 183
 linear mixed (LMM), 182, 183, 184, 192, 220
 linear probability, 67
 linear random effects, 173, 183
 MCMC, *see* Markov chain Monte Carlo
 mean regression, 236
 median regression, 238, 250
 mixed, 354
 multilevel, 184, 185
 multinomial logistic regression, 83
 multinomial logit, 81, 180
 no pooling, 183, 202
 nominal categorical dependent variable, 81
 non-nested, 185
 nonlinear growth curve, 191
 nonparametric, 523
 parametric, 527, 533
 penalized splines (P-splines), 214
 principal components, 283
 proportional hazards, 529, 533
 quantile models, 251
 quantile regression, 238, 253
 random intercepts, 173, 203
 random intercepts and slopes, 205
 semi-parametric regression, 214
 Tobit, 149
 two-part, 241
 variance components, 186
 varying intercepts, 184
 varying slopes and intercepts, 184
- multistate model, 516
 absorbing state, 516
 illness-death, 517, 524, 528, 534
 multiple decrement, 517
 multiple correlation coefficient, 25
- natural parameter, 341
- Newton optimization, 72
- nonlinear mixed model (NLMM), 411
- nonparametric regression, 358, 370
 backfitting algorithm, 383
 bandwidth, 372
 cubic smoothing splines, 378
 curse of dimensionality, 369
 kernel smoother, 374
 knots, 378
 locally weighted running-line smoother (LOESS), 373
 multivariate additive models, 383
 smoother, 370, 371
- nuisance parameters, 337
- odds, 68
- odds-ratio, 69, 163
- offset, 146, 458
- out-of-sample comparisons, 158
- out-of-sample statistic
 PRESS, 42, 43
 SSPE, 42, 43
- forecast error, 464
- mean squared error of prediction (MSEP), 465
- model MSEP, 465, 466
- parameter MSEP, 465, 466
- process MSEP, 465, 466
- over-dispersion, 102, 261, 352, 454
- parameter
 location, 244, 245, 246, 253
 scale, 146, 244, 245
 shape, 244

- plot
 - LOESS* curve, 38
 - ggplot2* package, 182
 - boxplot, 19
 - gains chart, 58
 - Gelman-Rubin, 324
 - icycle, 308
 - multiple time series, 218
 - quantile-quantile (*QQ*), 39, 247, 249
 - receiver operating characteristic (ROC) curve, 77
 - residual, 36
 - rug, 243
 - scatterplot, 16
 - scatterplot matrix, 49
 - scree, 293
 - time series, 427, 429
 - trace, 324
 - trellis, 175, 187
- Poisson regression, *see* generalized linear model (GLM)
- Polya tree, 344
- posterior distribution, 253, 317
- posterior predictive, 337, 349, 350
 - checks, 349, 350
- prior
 - conjugate, 253, 337, 340
 - degrees of belief, 335
 - diffuse, 337, 341, 346, 348, 350
 - hierarchical, 334, 355, 356
 - improper uninformative, 253
 - informative, 335, 357
 - nonparametric, 344, 359
 - Polya tree, 359, 362
 - vague, 253, 338, 341, 343, 346, 348, 350, 353
- prior distribution, 253, 317
 - conjugate, 318
 - elicitation, 326
 - Jeffreys', 328
 - non-informative, 327
- probit regression model, *see* generalized linear model (GLM)
- proper distribution, 267, 274
- proportional hazards
 - assumption, 494, 497
 - baseline survival function, 501
 - model, 497
 - model residuals, 502
 - time-varying covariates, 506
- proportional hazards model, 476
- pure premium, 127, 155
- random effect
 - spatial, 274
- random effects
 - binary outcomes, 177
 - crossed random effects, 196
 - generalized linear mixed model
 - see* generalized linear model (GLM) 180
 - intercepts, 173
 - multiple random effects per level, 195
 - nested random effects, 196
 - single random effect per level, 195
- ratemaking, 3, 128, 133, 140, 147, 167, 222, 250, 399
 - credibility, 5, 182, 187
- receiver operating characteristic (ROC), 77
- reference category, 17, 66
- regression quantile, 254
- relativity, 90, 142
- residual, 16, 22, 28, 36, 37, 53, 218, 228, 247
 - MSE*, 26
 - Cox-Snell, 502
 - deviance, 503
 - deviance residual, 243
 - martingale, 502
 - outlier, 219, 227, 231
 - partial residual, 243
 - Pearson residual, 243
 - residual standard error, 32
 - residual sum of squares, 241
 - residual variance, 29, 32, 33
 - standard error, 23, 26
 - standardized residual, 36
- risk factors, 65
- risk-ratio, 69
- sampling, 140
 - frequency-severity, 141
 - over-, 162
 - two-part, 148
- sensitivity, 75
- shrinkage estimator, 198
- simultaneously autoregressive model, *see* spatial autoregression
- spatial autoregression
 - conditionally autoregressive model, 265–267, 274
 - intrinsically autoregressive model, 266–267
 - proximity matrix, 266, 268, 274
 - simultaneously autoregressive model, 267–268
- spatial statistic
 - Geary's *C*, 264
 - Moran's *I*, 263
- specificity, 75
- standard error of regression estimator, 27, 32
- stochastic process, 516
- survival data
 - left-truncated, 482
 - right-censored, 482
- survival function
 - Kaplan-Meier estimator, 486
 - Nelson-Aalen estimator, 490, 523
 - nonparametric estimator, 486
 - product-limit estimator, 486
- survival model
 - cumulative hazard function, 481
 - hazard function, 481
- test
 - F*-test, 243
 - Hosmer-Lemeshow, 77
 - likelihood ratio test (LRT), 76, 95, 200, 411, 531, 536
 - Mantel-Haenzel, 495

- score, 95
- Wald, 95, 200
- Wilcoxon, 495
- threshold, 74
- time series inference
 - autocorrelation, 431, 437
 - autocovariance, 431
 - back casting, 435
 - Kalman filter, 445
 - prediction, 428, 442
- time series model
 - generalized autoregressive conditional heteroskedasticity (GARCH), 447
 - autoregressive of order 1 (AR(1)), 434
 - autoregressive, integrated, moving average (ARIMA), 440
 - autoregressive, moving average (ARMA), 434
 - autoregressive, moving average of order (p, q) , ARMA(p, q), 435
 - Box-Jenkins, 440
 - geometric random walk, 433
 - Lee-Carter, 447
 - moving average model of order q (MA(q)), 434
 - random walk, 432
 - seasonal ARIMA, 443
 - state space, 445
 - stochastic volatility, 446
 - structural, 443
 - white noise, 432
- time series model properties
 - invertibility, 437
 - stationarity, 429, 431, 434, 437
- transformation, 238
 - Box-Cox, 239, 371
 - inverse hyperbolic sine, 239
 - log-log, 491
 - logarithmic, 238
 - modified modulus, 240
 - modulus, 239
 - nonlinear, 239
 - signed power, 239
- time series differencing, 430, 433, 440
- time series lag operator, 435, 436, 440
- variance stabilizing, 239
- transition intensity function, 519, 520, 522
 - cumulative intensity function, 523
- transition probability function, 178, 519, 520
- two-part distribution, 148
- underwriting, 3
- value-at-risk, 237
- variables
 - binary, 17, 66, 67
 - categorical, 17, 65, 66, 281
 - components, 282
 - dummy, 17
 - dynamic, 475
 - factor, 282
 - indicator, 17
 - Likert scale, 283
 - ordinal categorical, 79
 - scale, 281, 283
 - static, 475
 - unpredictable, 475