

Handbook for Applied Modeling: Non-Gaussian and Correlated Data

Designed for the applied practitioner, this book is a compact, entry-level guide to modeling and analyzing non-Gaussian and correlated data. Many practitioners work with data that fail the assumptions of the common linear regression models, necessitating more advanced modeling techniques. This handbook presents clearly explained modeling options for such situations, along with extensive example data analyses. The book explains core models such as logistic regression, count regression, longitudinal regression, survival analysis, and structural equation modeling without relying on mathematical derivations. All data analyses are performed on real and publicly available data sets, which are revisited multiple times to show differing results using various modeling options. Common pitfalls, data issues, and interpretation of model results are also addressed. Programs in both R and SAS are made available for all results presented in the text so that readers can emulate and adapt analyses for their own data analysis needs.

JAMIE D. RIGGS is an adjunct lecturer in the Predictive Analytics program at Northwestern University, Chicago. She specializes in the statistical issues of solar system cratering processes, solar physics, and galactic dynamics, and has collaborated with researchers at the Los Alamos National Laboratory and the Southwest Research Institute. She has held technical and managerial positions at Sun Microsystems, Inc., National Oceanic and Atmospheric Administration, and the Boeing Company, where she applied advanced statistical designs and analyses to manufacturing and business problems. She is the head of the International Astrostatistics Association Solar System and Planetary Sciences Section.

TRENT L. LALONDE is Associate Professor of Applied Statistics at the University of Northern Colorado, and Director of the University's Research Consulting Lab. He has spent a number of years designing and teaching graduate courses covering statistical methods for students in diverse areas such as special education, psychological sciences, and public health. In addition, he has helped direct dissertations in these areas, and has consulted with numerous faculties on publications and funding proposals. He has received awards for both instruction and advising, and has chaired the Applied Public Health Statistics section of the American Public Health Association.

Handbook for Applied Modeling: Non-Gaussian and Correlated Data

Jamie D. Riggs

Northwestern University, Illinois

Trent L. Lalonde

University of Northern Colorado, Colorado



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom
One Liberty Plaza, 20th Floor, New York, NY 10006, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
4843/24, 2nd Floor, Ansari Road, Daryaganj, Delhi – 110002, India
79 Anson Road, #06-04/06, Singapore 079906

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/9781107146990

© Jamie D. Riggs and Trent L. Lalonde 2017

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 2017

Printed in the United States of America by Sheridan Books, Inc.

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

Library of Congress Cataloging-in-Publication Data

Names: Riggs, Jamie. | Lalonde, Trent.

Title: Handbook for applied modeling : non-Gaussian and correlated data /
Jamie Riggs, Northwestern University, Illinois, Trent Lalonde,
University of Northern Colorado.

Description: Cambridge : Cambridge University Press, 2017. |
Includes bibliographical references and index.

Identifiers: LCCN 2017004641 | ISBN 9781107146990 (hardback : alk. paper)

Subjects: LCSH: Mathematical statistics. | Mathematical models. |
Gaussian processes. | Stochastic processes.

Classification: LCC QA276.R5244 2017 | DDC 519.5/3–dc23

LC record available at <https://lcn.loc.gov/2017004641>

ISBN 978-1-107-14699-0 Hardback

ISBN 978-1-316-60105-1 Paperback

Additional resources for this publication at www.cambridge.org/riggslalonde

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.

This book is dedicated to:

Lauren and Jordan. JDR
Amanda, for always listening. TLL

Contents

<i>Preface</i>	xiii
1 The Data Sets	1
1.1 Introduction	1
1.1.1 The School Survey on Crime and Safety	2
1.1.2 The Framingham Heart Study	2
1.1.3 Fire-Climate Interactions in the American West	2
1.1.4 English Wikipedia Clickstream Data	3
1.2 Exploratory Data Analysis	3
1.3 Gauss-Markov Assumptions	4
1.4 Data Summaries and Tables	4
1.5 Graphical Representations	4
1.5.1 Histograms	5
1.5.2 Q-Q Plots	5
1.5.3 Box-Whisker Plots	5
1.5.4 Scatter Plots	6
1.6 Pairwise Correlation	7
1.7 Machine Learning Pattern Recognition	7
1.8 Exploring the Data Sets	8
1.8.1 School Survey on Crime and Safety Data	8
1.8.2 Framingham Heart Study Data	13
1.8.3 Fire-Climate Interactions in the American West Data	17
1.8.4 English Wikipedia Clickstream Data	20
1.9 Summary	23
1.10 Further Reading	24
2 The Model-Building Process	25
2.1 Introduction	25
2.2 The Model-Building Process	26
2.2.1 Exploratory Data Analysis	26
2.2.2 Model Construction	27
2.2.3 Model Fit Diagnostics	28
2.2.4 Model Effects Analysis	28
2.2.5 Model Interpretation and Prediction	29
2.2.6 Effects and Predictive Model Differences	29
2.3 Constant Variance Response Models	30
2.4 Nonconstant Variance Response Models	31

2.5	Discrete, Categorical Response Models	32
2.6	Count Response Models	34
2.7	Time-to-Event Response Models	37
2.8	Longitudinal Response Models	39
2.9	Structural Equation Modeling	41
2.10	Effect Size	43
2.11	Model Fit Measures	43
	2.11.1 Measures of Fit	43
	2.11.2 Residual Analyses	45
2.12	Summary	48
2.13	Further Reading	49
3	Constant Variance Response Models	50
3.1	Introduction	50
3.2	School Survey on Crime and Safety	50
3.3	Framingham Heart Study	52
3.4	Fire-Climate Interactions in the American West	53
3.5	English Wikipedia Clickstream Data	55
3.6	Summary	56
3.7	Further Reading	56
4	Nonconstant Variance Response Models	57
4.1	Heterogeneity in Response Variance	57
4.2	Detecting Heteroscedasticity	58
	4.2.1 Descriptive Statistics	58
	4.2.2 Tests for Grouped Data	58
	4.2.3 Tests for Continuous Predictors	59
4.3	Variance-Stabilizing Transformations	59
	4.3.1 Selecting the Transformation	59
	4.3.2 Model Diagnostics	59
4.4	Weighted Least Squares	60
	4.4.1 WLS Estimation	60
	4.4.2 Selecting the Weights	60
4.5	SSOCS Analysis: Annual Suspensions	61
	4.5.1 Exploratory Data Analysis	61
	4.5.2 Normal Linear Model	63
	4.5.3 Outcome Transformations	63
	4.5.4 Weighted Least Squares	65
	4.5.5 Parameter Interpretations	68
	4.5.6 Model Prediction	69
4.6	Fire-Climate Analysis: Decade Averages	70
	4.6.1 Exploratory Data Analysis	70
	4.6.2 Normal Linear Model	71
	4.6.3 Weighted Least Squares	72
	4.6.4 Parameter Interpretations	74
	4.6.5 Model Prediction	74
4.7	Summary	75
4.8	Further Reading	75

Contents

ix

5	Discrete, Categorical Response Models	76
5.1	Categorical Responses	76
5.2	Binary Logistic Regression	76
	5.2.1 Descriptive Statistics for Binary Outcomes	77
	5.2.2 The Logistic Regression Model	78
	5.2.3 Interpreting Model Coefficients	78
	5.2.4 Model Fit	79
5.3	Nominal Multinomial Models	81
5.4	Ordinal Multinomial Models	82
	5.4.1 Cumulative Logit Model	83
	5.4.2 Adjacent Categories Model	83
	5.4.3 Continuation Ratio Model	84
5.5	FHS Analysis: Probability of Hypertension	85
	5.5.1 Exploratory Data Analyses	85
	5.5.2 Logistic Regression Model	86
	5.5.3 Logistic Regression Model Fit	87
	5.5.4 Model Parameter Interpretations	89
	5.5.5 Model Prediction	90
5.6	SSOCS Analysis: Probability of Bullying	93
	5.6.1 Exploratory Data Analysis	93
	5.6.2 Ordinal Multinomial Model	94
	5.6.3 Ordinal Multinomial Model Fit	96
	5.6.4 Model Parameters Interpretations	97
	5.6.5 Model Prediction	99
5.7	Clickstream Analysis: Probability of Redlink	101
	5.7.1 Exploratory Data Analysis	102
	5.7.2 Logistic Regression Model	102
	5.7.3 Logistic Regression Model Fit	103
	5.7.4 Model Parameter Interpretations	104
	5.7.5 Model Prediction	105
5.8	Summary	106
5.9	Further Reading	107
6	Count Response Models	108
6.1	Introduction	108
6.2	Modeling Count Data	109
	6.2.1 Poisson Models	109
	6.2.2 Overdispersion	110
	6.2.3 Coefficient Interpretations	111
	6.2.4 Negative Binomial Models	113
	6.2.5 Zero-Inflated Models	114
	6.2.6 Zero-Deflated Models	114
	6.2.7 Hurdle Models	115
6.3	Fire-Climate Analysis: Decade Counts	115
	6.3.1 Exploratory Data Analysis	115
	6.3.2 Poisson Model	116
	6.3.3 Negative Binomial Models	118
	6.3.4 Zero-Inflated NB Models	119

6.4	SSOCS Analysis: Annual Suspensions	123
6.4.1	Hurdle Negative Binomial Model	123
6.4.2	Model Fit	124
6.4.3	Model Interpretations	124
6.5	Clickstream Analysis: Site Pairings	126
6.5.1	Exploratory Data Analysis	126
6.5.2	Left-truncated Count Model	126
6.5.3	Count Model Fit	128
6.5.4	Coefficient Interpretations	129
6.6	Summary	130
6.7	Further Reading	131
7	Time-to-Event Response Models	132
7.1	Time-to-Event Data	132
7.2	Time-to-Event Models	133
7.3	FHS Analysis: Time to Hypertension	135
7.3.1	Life Tables	135
7.3.2	Kaplan-Meier Method	138
7.3.3	Cox Proportional Hazards Models	140
7.3.4	Time-Dependent Cox Models	145
7.4	Summary	150
7.5	Further Reading	150
8	Longitudinal Response Models	152
8.1	Longitudinal Data	152
8.2	Autocorrelation in Longitudinal Data	153
8.2.1	Descriptive Analysis	153
8.2.2	Scatter plots	153
8.2.3	Autocorrelation Plots	154
8.2.4	Variograms	155
8.2.5	Modeling Longitudinal Data	156
8.3	Marginal Models	156
8.3.1	Generalized Estimating Equations	157
8.3.2	Working Correlation Structure	157
8.3.3	Marginal Model Fit	159
8.4	Conditional Models	160
8.4.1	Random-Intercept Models	160
8.4.2	Random-Slopes Models	161
8.4.3	Conditional Model Fit	162
8.5	FHS Analysis: Probability of Hypertension	163
8.5.1	Exploratory Data Analysis	163
8.5.2	Marginal Longitudinal Model	166
8.5.3	Examining the Autocorrelation	166
8.5.4	Marginal Longitudinal Model Fit	168
8.5.5	Model Parameter Interpretations	168
8.5.6	Model Prediction	170
8.6	Fire-Climate Analysis: Decade Counts	172
8.6.1	Exploratory Data Analysis	172

Contents

xi

8.6.2	Autocorrelation in Decade Counts	175
8.6.3	Conditional Models for Decade Counts	175
8.6.4	Conditional Longitudinal Model Fit	176
8.6.5	Model Parameter Interpretations	178
8.6.6	Model Prediction	179
8.7	Summary	181
8.8	Further Reading	181
9	Structural Equation Modeling	183
9.1	Introduction	183
9.1.1	SEM Variable Categories	184
9.1.2	Model Types	185
9.1.3	SEM Paths	185
9.1.4	Confirmatory Factor Analysis	187
9.1.5	Evaluating Model Fit	188
9.2	FHS Analysis: Latent Stress	189
9.3	SSOCS Analysis: School Climate and Academic Success	194
9.4	Summary	201
9.5	Further Reading	201
10	Matching Data to Models	202
10.1	The Decision Process of Modeling	202
10.2	Results of Model Application	207
10.2.1	School Survey on Crime and Safety	207
10.2.2	Framingham Heart Study	208
10.2.3	Fire-Climate Interactions in the American West	208
10.2.4	English Wikipedia Clickstream	209
10.3	Perspectives on Modeling	209
	<i>Bibliography</i>	211
	<i>Index</i>	213

Preface

Modern society is data driven. When you buy – or even shop for – a shirt on the Internet, the next time you enter the web, you’ll be inundated with advertisements for more shirts, all the outcome of data collection, analysis, and targeted marketing. Global networks have been designed specifically to deliver stock market and commodities market data for near real-time trading. Public services depend heavily on censuses for allocation of government funding and assistance programs to the populations that need them. These same censuses determine the districts needed for so-called enfranchisement, at least in the United States. Travel, particularly international, is regulated based on personal information collected by government agencies. Large chain retailers collect cash-out data to stock according to collective shopping habits. Educators undertake quantitative assessments of new instructional methods to determine best practice. Health policy administrators analyze data to allocate resources according to the timing and volume of patient needs. These applications are just a hint of the universal use of data in both public and private spheres.

The ubiquity of data-driven decisions means that our personal and collective lives are affected daily by how data are analyzed and interpreted. When data are interpreted accurately, we expect fair treatment. When data are improperly collected, analyzed, or interpreted, not only is our quality of life diminished, but the faulty information can debilitate or even kill. Clearly, then, we want data analysts who, conscious of the consequences of poor or incorrect analyses, have the knowledge to extract information from data – properly and with a healthy awareness of any uncertainties that should qualify interpretation.

To support this kind of mastery, we have written this handbook to overcome two common limitations in tutorial resources for practicing data analysts.

- **We make a broad selection of the most useful basic models, from a range of disciplines and domains.** Applied disciplines that use statistical analysis sometimes rely on a restricted set of tools particular to the discipline. Although this practice has advantages at the entry level, it can encourage overreliance on familiar methods to the exclusion of viable, even superior, alternatives. This danger is compounded if discipline-specific software entrenches an unchanging set of models. Our approach is to look at a variety of data that is typical of modern applications and to present the models most likely to extract meaningful information. Our goal is not to present all possible useful models, but to build your facility with a range of core methods so that you are equipped to tackle new data with new or adapted models.

- **We deal with data as it comes, which is often non-Gaussian and often correlated.** Common practice, especially with large data sets, has been to assume that the data are close enough to Gaussian and uncorrelated even when these assumptions can be shown to be untrue. Misapplied analyses then produce tangles of misinformation. Our approach is to guide you to and through the statistical methods that best match the characteristics of the data under consideration, in particular methods suited to the prevalent non-Gaussian forms of observational data. Our goal is for you to become confident in building models for your real-world purposes.

This handbook is for data analysts with a grounding in basic statistics, biostatistics, econometrics, business statistics, social science statistics, or predictive analytics who want to develop their modeling skills beyond the commonly used, idealized setting of independent Gaussian analyses. We assume you are practiced in the use of descriptive statistics, analysis of variance, and regression.

All data analyses are performed on **real and publicly available data sets**, which are revisited multiple times to show differing results using various modeling options. You will see concrete examples of common pitfalls, issues that arise from messy data, and interpretation of model results. To encourage your hands-on engagement and so you can replicate any of the analyses, **code for all analyses is provided as both R and SAS commands**, available online at www.cambridge.org/riggslalonde.

The modeling methods are presented from a data analyst's perspective. We use basic mathematics to summarize model structure, basic model diagnostics, model effects interpretation, and predictive ability; however, our emphasis is always on the application of methods, rather than study of the methods themselves. We demonstrate how the methods are (or are not) appropriate, and how weak or strong is a model's performance with a given data set. This includes effects interpretation, predictive strength, and model aptness for model-to-model comparisons. While the book is well suited as a text for graduate-level methods courses, we present models through "standalone" discussions, so that you can use any single chapter as a self-contained resource for the models covered there.

Chapter 1 is the gateway to the rest of the book. In it the four driving data sets are described and explored, including all relevant variables used for analyses in later chapters. Chapter 2 gives a review of all the model types used in the book. Then, after a review of ordinary least squares estimation models in Chapter 3, we progress in the heart of the book through remedial methods under such violations of least squares assumptions as heteroscedasticity, serial correlation, and endogenous variables as found in panel data types, in addition to models for nonnormal responses, autocorrelated responses, time-to-event responses, and ending with structural equation modeling. The final chapter gives a point-by-point system for matching data to models.

Acknowledgements

We wish to thank Dr. Joseph M. Hilbe for sharing his vast knowledge and insightful wisdom. His encouragement in this endeavor was invaluable. Dr. Annyce Stone listened, commented,

Preface

xv

and encouraged us. Her top-down view gave further incentive to our work. Andrea Sorrells applied her artistic and critical eye to our cover art and images. We cannot thank her enough. Lucy Edwards of Cambridge University Press, New York, was patient, encouraging, cajoling, and, in general, a wonderful editor. She made this book a reality. Diana Gillooly believed in our data-centric approach to convey complex information.