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

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://lccn.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

This book is dedicated to:

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

Contents

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,

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.