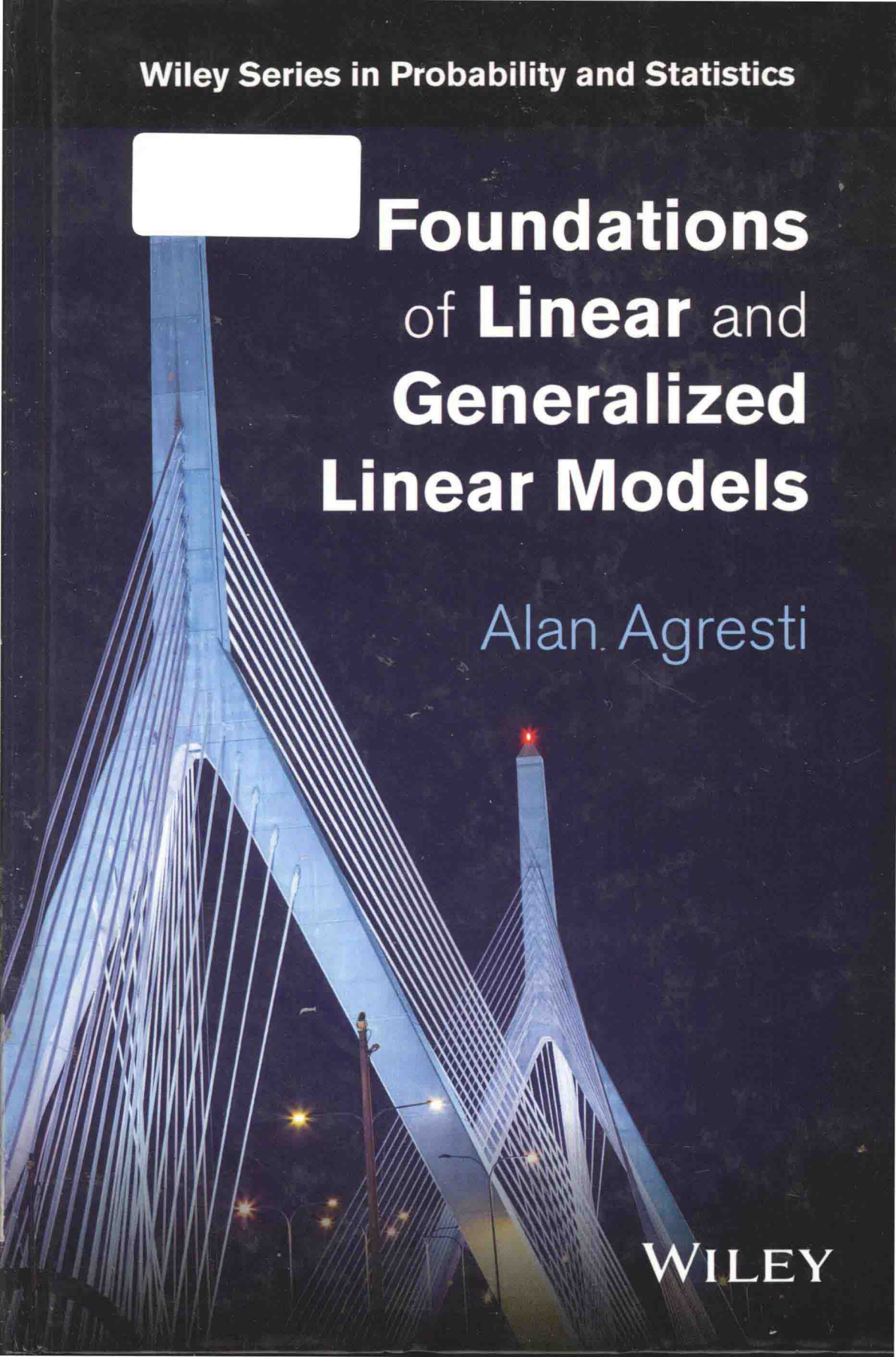
Wiley Series in Probability and Statistics

Foundations of Linear and Generalized Linear Models

Alan Agresti



WILEY

Foundations of Linear and Generalized Linear Models

ALAN AGRESTI

Distinguished Professor Emeritus
University of Florida
Gainesville, FL

Visiting Professor
Harvard University
Cambridge, MA

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data

Agresti, Alan, author.
Foundations of linear and generalized linear models / Alan Agresti.
pages cm. – (Wiley series in probability and statistics)
Includes bibliographical references and index.
ISBN 978-1-118-73003-4 (hardback)
1. Mathematical analysis–Foundations. 2. Linear models (Statistics) I. Title.
QA299.8.A37 2015
003′.74–dc23

2014036543

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Foundations of Linear and Generalized Linear Models

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

To my statistician friends in Europe

Preface

PURPOSE OF THIS BOOK

Why yet another book on linear models? Over the years, a multitude of books have already been written about this well-traveled topic, many of which provide more comprehensive presentations of linear modeling than this one attempts. *My book is intended to present an overview of the key ideas and foundational results of linear and generalized linear models.* I believe this overview approach will be useful for students who lack the time in their program for a more detailed study of the topic. This situation is increasingly common in Statistics and Biostatistics departments. As courses are added on recent influential developments (such as "big data," statistical learning, Monte Carlo methods, and application areas such as genetics and finance), programs struggle to keep room in their curriculum for courses that have traditionally been at the core of the field. Many departments no longer devote an entire year or more to courses about linear modeling.

Books such as those by Dobson and Barnett (2008), Fox (2008), and Madsen and Thyregod (2011) present fine overviews of both linear and generalized linear models. By contrast, my book has more emphasis on the theoretical foundations—showing how linear model fitting projects the data onto a model vector subspace and how orthogonal decompositions of the data yield information about effects, deriving likelihood equations and likelihood-based inference, and providing extensive references for historical developments and new methodology. In doing so, my book has less emphasis than some other books on practical issues of data analysis, such as model selection and checking. However, each chapter contains at least one section that applies the models presented in that chapter to a dataset, using R software. The book is not intended to be a primer on R software or on the myriad details relevant to statistical practice, however, so these examples are relatively simple ones that merely convey the basic concepts and spirit of model building.

The presentation of linear models for continuous responses in Chapters 1–3 has a geometrical rather than an algebraic emphasis. More comprehensive books on linear models that use a geometrical approach are the ones by Christensen (2011) and by

Seber and Lee (2003). The presentation of generalized linear models in Chapters 4–9 includes several sections that focus on discrete data. Some of this significantly abbreviates material from my book, *Categorical Data Analysis* (3rd ed., John Wiley & Sons , 2013). Broader overviews of generalized linear modeling include the classic book by McCullagh and Nelder (1989) and the more recent book by Aitkin et al. (2009). An excellent book on statistical modeling in an even more general sense is by Davison (2003).

USE AS A TEXTBOOK

This book can serve as a textbook for a one-semester or two-quarter course on linear and generalized linear models. It is intended for graduate students in the first or second year of Statistics and Biostatistics programs. It also can serve programs with a heavy focus on statistical modeling, such as econometrics and operations research. The book also should be useful to students in the social, biological, and environmental sciences who choose Statistics as their minor area of concentration.

As a prerequisite, the reader should be familiar with basic theory of statistics, such as presented by Casella and Berger (2001). Although not mandatory, it will be helpful if readers have at least some background in applied statistical modeling, including linear regression and ANOVA. I also assume some linear algebra background. In this book, I recall and briefly review fundamental statistical theory and matrix algebra results where they are used. This contrasts with the approach in many books on linear models of having several chapters on matrix algebra and distribution theory before presenting the main results on linear models. Readers wanting to improve their knowledge of matrix algebra can find on the Web (e.g., with a Google search of "review of matrix algebra") overviews that provide more than enough background for reading this book. Also helpful as background for Chapters 1–3 on linear models are online lectures, such as the MIT linear algebra lectures by G. Strang at `http://ocw.mit.edu/courses/mathematics` on topics such as vector spaces, column space and null space, independence and a basis, inverses, orthogonality, projections and least squares, eigenvalues and eigenvectors, and symmetric and idempotent matrices. By not including separate chapters on matrix algebra and distribution theory, I hope instructors will be able to cover most of the book in a single semester or in a pair of quarters.

Each chapter contains exercises for students to practice and extend the theory and methods and also to help assimilate the material by analyzing data. Complete data files for the text examples and exercises are available at the text website, `http://www.stat.ufl.edu/~aa/glm/data/`. Appendix A contains supplementary data analysis exercises that are not tied to any particular chapter. Appendix B contains solution outlines and hints for some of the exercises.

I emphasize that this book is not intended to be a complete overview of linear and generalized linear modeling. Some important classes of models are beyond its scope; examples are transition (e.g., Markov) models and survival (time-to-event) models. I intend merely for the book to be an overview of the *foundations* of this subject—that is, core material that should be part of the background of any statistical scientist. I

invite readers to use it as a stepping stone to reading more specialized books that focus on recent advances and extensions of the models presented here.

ACKNOWLEDGMENTS

This book evolved from a one-semester course that I was invited to develop and teach as a visiting professor for the Statistics Department at Harvard University in the fall terms of 2011–2014. That course covers most of the material in Chapters 1–9. My grateful thanks to Xiao-Li Meng (then chair of the department) for inviting me to teach this course, and likewise thanks to Dave Harrington for extending this invitation through 2014. (The book's front cover, showing the Zakim bridge in Boston, reflects the Boston-area origins of this book.) Special thanks to Dave Hoaglin, who besides being a noted statistician and highly published book author, has wonderful editing skills. Dave gave me detailed and helpful comments and suggestions for my working versions of all the chapters, both for the statistical issues and the expository presentation. He also found many errors that otherwise would have found their way into print!

Thanks also to David Hitchcock, who kindly read the entire manuscript and made numerous helpful suggestions, as did Maria Kateri and Thomas Kneib for a few chapters. Hani Doss kindly shared his fine course notes on linear models (Doss 2010) when I was organizing my own thoughts about how to present the foundations of linear models in only two chapters. Thanks to Regina Dittrich for checking the R code and pointing out errors. I owe thanks also to several friends and colleagues who provided comments or datasets or other help, including Pat Altham, Alessandra Brazzale, Jane Brockmann, Phil Brown, Brian Caffo, Leena Choi, Guido Consonni, Brent Coull, Anthony Davison, Kimberly Dibble, Anna Gottard, Ralitza Gueorguieva, Alessandra Guglielmi, Jarrod Hadfield, Rebecca Hale, Don Hedeker, Georg Heinze, Jon Hennessy, Harry Khamis, Eunhee Kim, Joseph Lang, Ramon Littell, I-Ming Liu, Brian Marx, Clint Moore, Bhramar Mukherjee, Dan Nettleton, Keramat Nourijelyani, Donald Pierce, Penelope Pooler, Euijung Ryu, Michael Schemper, Cristiano Varin, Larry Winner, and Lo-Hua Yuan. James Booth, Gianfranco Lovison, and Brett Presnell have generously shared materials over the years dealing with generalized linear models. Alex Blocker, Jon Bischof, Jon Hennessy, and Guillaume Basse were outstanding and very helpful teaching assistants for my Harvard Statistics 244 course, and Jon Hennessy contributed solutions to many exercises from which I extracted material at the end of this book. Thanks to students in that course for their comments about the manuscript. Finally, thanks to my wife Jacki Levine for encouraging me to spend the terms visiting Harvard and for support of all kinds, including helpful advice in the early planning stages of this book.

ALAN AGRESTI

Brookline, Massachusetts, and Gainesville, Florida
June 2014

Contents

Website

Data sets for the book are at `www.stat.ufl.edu/~aa/glm/data`

CHAPTER 1

Introduction to Linear and Generalized Linear Models

This is a book about *linear models* and *generalized linear models*. As the names suggest, the linear model is a special case of the generalized linear model. In this first chapter, we define generalized linear models, and in doing so we also introduce the linear model.

Chapters 2 and 3 focus on the linear model. Chapter 2 introduces the *least squares* method for fitting the model, and Chapter 3 presents statistical inference under the assumption of a *normal* distribution for the response variable. Chapter 4 presents analogous model-fitting and inferential results for the generalized linear model. This generalization enables us to model non-normal responses, such as categorical data and count data.

The remainder of the book presents the most important generalized linear models. Chapter 5 focuses on models that assume a *binomial* distribution for the response variable. These apply to binary data, such as "success" and "failure" for possible outcomes in a medical trial or "favor" and "oppose" for possible responses in a sample survey. Chapter 6 extends the models to multicategory responses, assuming a *multinomial* distribution. Chapter 7 introduces models that assume a *Poisson* or *negative binomial* distribution for the response variable. These apply to count data, such as observations in a health survey on the number of respondent visits in the past year to a doctor. Chapter 8 presents ways of weakening distributional assumptions in generalized linear models, introducing *quasi-likelihood* methods that merely focus on the mean and variance of the response distribution. Chapters 1–8 assume *independent* observations. Chapter 9 generalizes the models further to permit *correlated* observations, such as in handling *multivariate* responses. Chapters 1–9 use the traditional *frequentist* approach to statistical inference, assuming probability distributions for the response variables but treating model parameters as fixed, unknown values. Chapter 10 presents the *Bayesian* approach for linear models and generalized linear models, which treats the model parameters as random variables having their

Foundations of Linear and Generalized Linear Models, First Edition. Alan Agresti.

own distributions. The final chapter introduces extensions of the models that handle more complex situations, such as *high-dimensional* settings in which models have enormous numbers of parameters.

1.1 COMPONENTS OF A GENERALIZED LINEAR MODEL

The ordinary linear regression model uses linearity to describe the relationship between the mean of the response variable and a set of explanatory variables, with inference assuming that the response distribution is normal. *Generalized linear models* (GLMs) extend standard linear regression models to encompass non-normal response distributions and possibly nonlinear functions of the mean. They have three components.

- *Random component*: This specifies the response variable y and its probability distribution. The observations[1] $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ on that distribution are treated as independent.
- *Linear predictor*: For a *parameter vector* $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$ and a $n \times p$ *model matrix* $\boldsymbol{X}$ that contains values of p explanatory variables for the n observations, the linear predictor is $\boldsymbol{X\beta}$.
- *Link function*: This is a function g applied to each component of $E(\boldsymbol{y})$ that relates it to the linear predictor,

$$g[E(\boldsymbol{y})] = \boldsymbol{X\beta}.$$

Next we present more detail about each component of a GLM.

1.1.1 Random Component of a GLM

The *random component* of a GLM consists of a response variable y with independent observations $(y_1, \ldots, y_n)$ having probability density or mass function for a distribution in the *exponential family*. In Chapter 4 we review this family of distributions, which has several appealing properties. For example, $\sum_i y_i$ is a sufficient statistic for its parameter, and regularity conditions (such as differentiation passing under an integral sign) are satisfied for derivations of properties such as optimal large-sample performance of maximum likelihood (ML) estimators.

By restricting GLMs to exponential family distributions, we obtain general expressions for the model likelihood equations, the asymptotic distributions of estimators for model parameters, and an algorithm for fitting the models. For now, it suffices to say that the distributions most commonly used in Statistics, such as the normal, binomial, and Poisson, are exponential family distributions.

[1]The superscript T on a vector or matrix denotes the transpose; for example, here $\boldsymbol{y}$ is a column vector. Our notation makes no distinction between random variables and their observed values; this is generally clear from the context.