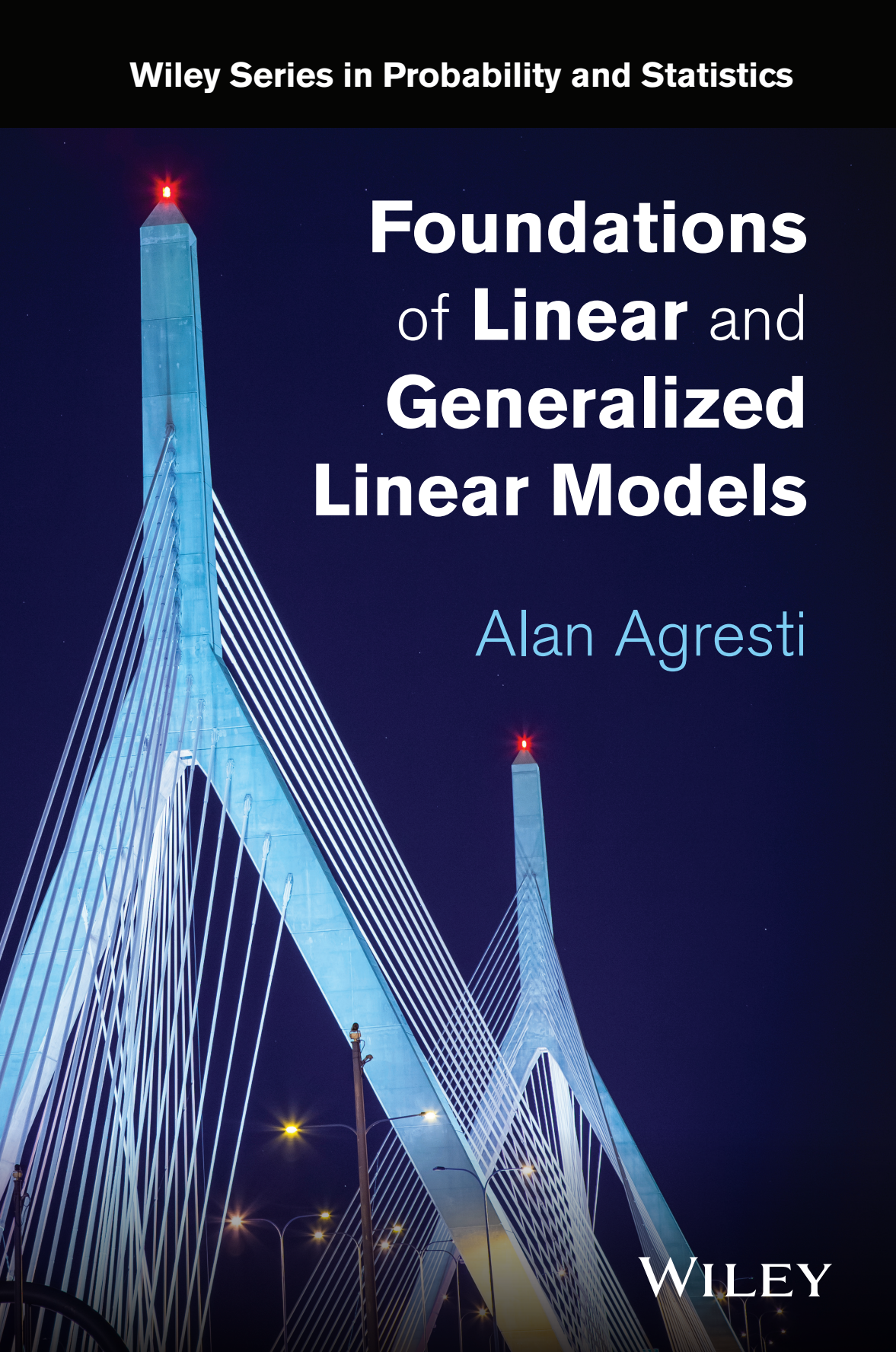


Wiley Series in Probability and Statistics



Foundations
of **Linear** and
Generalized
Linear Models

Alan Agresti

WILEY

Foundations of Linear and Generalized Linear Models

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice,
Geof H. Givens, Harvey Goldstein, Geert Molenberghs, David W. Scott,
Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*

Editors Emeriti: *J. Stuart Hunter, Iain M. Johnstone, Joseph B. Kadane,
Jozef L. Teugels*

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

Foundations of Linear and Generalized Linear Models

ALAN AGRESTI

Distinguished Professor Emeritus
University of Florida
Gainesville, FL

Visiting Professor
Harvard University
Cambridge, MA

WILEY

Copyright © 2015 by John Wiley & Sons, Inc. All rights reserved.

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

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

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

To my statistician friends in Europe

Contents

Preface	xi
1 Introduction to Linear and Generalized Linear Models	1
1.1 Components of a Generalized Linear Model,	2
1.2 Quantitative/Qualitative Explanatory Variables and Interpreting Effects,	6
1.3 Model Matrices and Model Vector Spaces,	10
1.4 Identifiability and Estimability,	13
1.5 Example: Using Software to Fit a GLM,	15
Chapter Notes,	20
Exercises,	21
2 Linear Models: Least Squares Theory	26
2.1 Least Squares Model Fitting,	27
2.2 Projections of Data Onto Model Spaces,	33
2.3 Linear Model Examples: Projections and SS Decompositions,	41
2.4 Summarizing Variability in a Linear Model,	49
2.5 Residuals, Leverage, and Influence,	56
2.6 Example: Summarizing the Fit of a Linear Model,	62
2.7 Optimality of Least Squares and Generalized Least Squares,	67
Chapter Notes,	71
Exercises,	71
3 Normal Linear Models: Statistical Inference	80
3.1 Distribution Theory for Normal Variates,	81
3.2 Significance Tests for Normal Linear Models,	86
3.3 Confidence Intervals and Prediction Intervals for Normal Linear Models,	95

3.4	Example: Normal Linear Model Inference,	99
3.5	Multiple Comparisons: Bonferroni, Tukey, and FDR Methods,	107
	Chapter Notes,	111
	Exercises,	112
4	Generalized Linear Models: Model Fitting and Inference	120
4.1	Exponential Dispersion Family Distributions for a GLM,	120
4.2	Likelihood and Asymptotic Distributions for GLMs,	123
4.3	Likelihood-Ratio/Wald/Score Methods of Inference for GLM Parameters,	128
4.4	Deviance of a GLM, Model Comparison, and Model Checking,	132
4.5	Fitting Generalized Linear Models,	138
4.6	Selecting Explanatory Variables for a GLM,	143
4.7	Example: Building a GLM,	149
	Appendix: GLM Analogs of Orthogonality Results for Linear Models,	156
	Chapter Notes,	158
	Exercises,	159
5	Models for Binary Data	165
5.1	Link Functions for Binary Data,	165
5.2	Logistic Regression: Properties and Interpretations,	168
5.3	Inference About Parameters of Logistic Regression Models,	172
5.4	Logistic Regression Model Fitting,	176
5.5	Deviance and Goodness of Fit for Binary GLMs,	179
5.6	Probit and Complementary Log–Log Models,	183
5.7	Examples: Binary Data Modeling,	186
	Chapter Notes,	193
	Exercises,	194
6	Multinomial Response Models	202
6.1	Nominal Responses: Baseline-Category Logit Models,	203
6.2	Ordinal Responses: Cumulative Logit and Probit Models,	209
6.3	Examples: Nominal and Ordinal Responses,	216
	Chapter Notes,	223
	Exercises,	223
7	Models for Count Data	228
7.1	Poisson GLMs for Counts and Rates,	229
7.2	Poisson/Multinomial Models for Contingency Tables,	235

7.3	Negative Binomial GLMS, 247	
7.4	Models for Zero-Inflated Data, 250	
7.5	Example: Modeling Count Data, 254	
	Chapter Notes, 259	
	Exercises, 260	
8	Quasi-Likelihood Methods	268
8.1	Variance Inflation for Overdispersed Poisson and Binomial GLMs, 269	
8.2	Beta-Binomial Models and Quasi-Likelihood Alternatives, 272	
8.3	Quasi-Likelihood and Model Misspecification, 278	
	Chapter Notes, 282	
	Exercises, 282	
9	Modeling Correlated Responses	286
9.1	Marginal Models and Models with Random Effects, 287	
9.2	Normal Linear Mixed Models, 294	
9.3	Fitting and Prediction for Normal Linear Mixed Models, 302	
9.4	Binomial and Poisson GLMMs, 307	
9.5	GLMM Fitting, Inference, and Prediction, 311	
9.6	Marginal Modeling and Generalized Estimating Equations, 314	
9.7	Example: Modeling Correlated Survey Responses, 319	
	Chapter Notes, 322	
	Exercises, 324	
10	Bayesian Linear and Generalized Linear Modeling	333
10.1	The Bayesian Approach to Statistical Inference, 333	
10.2	Bayesian Linear Models, 340	
10.3	Bayesian Generalized Linear Models, 347	
10.4	Empirical Bayes and Hierarchical Bayes Modeling, 351	
	Chapter Notes, 357	
	Exercises, 359	
11	Extensions of Generalized Linear Models	364
11.1	Robust Regression and Regularization Methods for Fitting Models, 365	
11.2	Modeling With Large p , 375	
11.3	Smoothing, Generalized Additive Models, and Other GLM Extensions, 378	
	Chapter Notes, 386	
	Exercises, 388	

Appendix A	Supplemental Data Analysis Exercises	391
Appendix B	Solution Outlines for Selected Exercises	396
References		410
Author Index		427
Example Index		433
Subject Index		435
Website		

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

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, Ralitzia 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

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

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 observations¹ $\mathbf{y} = (y_1, \dots, y_n)^T$ on that distribution are treated as independent.
- *Linear predictor*: For a *parameter vector* $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ and a $n \times p$ *model matrix* \mathbf{X} that contains values of p explanatory variables for the n observations, the linear predictor is $\mathbf{X}\boldsymbol{\beta}$.
- *Link function*: This is a function g applied to each component of $E(\mathbf{y})$ that relates it to the linear predictor,

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

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, \dots, 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.

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

1.1.2 Linear Predictor of a GLM

For observation i , $i = 1, \dots, n$, let x_{ij} denote the value of explanatory variable x_j , $j = 1, \dots, p$. Let $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$. Usually, we set $x_{i1} = 1$ or let the first variable have index 0 with $x_{i0} = 1$, so it serves as the coefficient of an intercept term in the model. The *linear predictor* of a GLM relates parameters $\{\eta_i\}$ pertaining to $\{E(y_i)\}$ to the explanatory variables x_1, \dots, x_p using a linear combination of them,

$$\eta_i = \sum_{j=1}^p \beta_j x_{ij}, \quad i = 1, \dots, n.$$

The labeling of $\sum_{j=1}^p \beta_j x_{ij}$ as a *linear predictor* reflects that this expression is *linear in the parameters*. The explanatory variables themselves can be nonlinear functions of underlying variables, such as an interaction term (e.g., $x_{i3} = x_{i1}x_{i2}$) or a quadratic term (e.g., $x_{i2} = x_{i1}^2$).

In matrix form, we express the linear predictor as

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

where $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^T$, $\boldsymbol{\beta}$ is the $p \times 1$ column vector of model parameters, and \mathbf{X} is the $n \times p$ matrix of explanatory variable values $\{x_{ij}\}$. The matrix \mathbf{X} is called the *model matrix*. In experimental studies, it is also often called the *design matrix*. It has n rows, one for each observation, and p columns, one for each parameter in $\boldsymbol{\beta}$. In practice, usually $p \leq n$, the goal of *model parsimony* being to summarize the data using a considerably smaller number of parameters.

GLMs treat y_i as random and \mathbf{x}_i as fixed. Because of this, the linear predictor is sometimes called the *systematic component*. In practice \mathbf{x}_i is itself often random, such as in sample surveys and other observational studies. In this book, we condition on its observed values in conducting statistical inference about effects of the explanatory variables.

1.1.3 Link Function of a GLM

The third component of a GLM, the *link function*, connects the random component with the linear predictor. Let $\mu_i = E(y_i)$, $i = 1, \dots, n$. The GLM links η_i to μ_i by $\eta_i = g(\mu_i)$, where the link function $g(\cdot)$ is a monotonic, differentiable function. Thus, g links μ_i to explanatory variables through the formula:

$$g(\mu_i) = \sum_{j=1}^p \beta_j x_{ij}, \quad i = 1, \dots, n. \quad (1.1)$$

In the exponential family representation of a distribution, a certain parameter serves as its *natural parameter*. This parameter is the mean for a normal distribution, the log of the odds for a binomial distribution, and the log of the mean for a Poisson distribution. The link function g that transforms μ_i to the natural parameter is called the *canonical link*. This link function, which equates the natural parameter with the

linear predictor, generates the most commonly used GLMs. Certain simplifications result when the GLM uses the canonical link function. For example, the model has a concave log-likelihood function and simple sufficient statistics and likelihood equations.

1.1.4 A GLM with Identity Link Function is a “Linear Model”

The link function $g(\mu_i) = \mu_i$ is called the *identity link function*. It has $\eta_i = \mu_i$. A GLM that uses the identity link function is called a *linear model*. It equates the linear predictor to the mean itself. This GLM has

$$\mu_i = \sum_{j=1}^p \beta_j x_{ij}, \quad i = 1, \dots, n.$$

The standard version of this, which we refer to as the *ordinary linear model*, assumes that the observations have constant variance, called *homoscedasticity*. An alternative way to express the ordinary linear model is

$$y_i = \sum_{j=1}^p \beta_j x_{ij} + \epsilon_i,$$

where the “error term” ϵ_i has $E(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma^2$, $i = 1, \dots, n$. This is natural for the identity link and normal responses but not for most GLMs.

In summary, ordinary linear models equate the linear predictor directly to the mean of a response variable y and assume constant variance for that response. The normal linear model also assumes normality. By contrast, a GLM is an extension that equates the linear predictor to a link-function-transformed mean of y , and assumes a distribution for y that need not be normal but is in the exponential family. We next illustrate the three components of a GLM by introducing three of the most important GLMs.

1.1.5 GLMs for Normal, Binomial, and Poisson Responses

The class of GLMs includes models for continuous response variables. Most important are ordinary normal linear models. Such models assume a normal distribution for the random component, $y_i \sim N(\mu_i, \sigma^2)$ for $i = 1, \dots, n$. The natural parameter for a normal distribution is the mean. So, the canonical link function for a normal GLM is the identity link, and the GLM is then merely a linear model. In particular, standard regression and analysis of variance (ANOVA) models are GLMs assuming a normal random component and using the identity link function. Chapter 3 develops statistical inference for such normal linear models. Chapter 2 presents model fitting for linear models and shows this does not require the normality assumption.

Many response variables are binary. We represent the “success” and “failure” outcomes, such as “favor” and “oppose” responses to a survey question about legalizing

same-sex marriage, by 1 and 0. A *Bernoulli trial* for observation i has probabilities $P(y_i = 1) = \pi_i$ and $P(y_i = 0) = 1 - \pi_i$, for which $\mu_i = \pi_i$. This is the special case of the binomial distribution with the number of trials $n_i = 1$. The natural parameter for the binomial distribution is $\log[\mu_i/(1 - \mu_i)]$. This is the log odds of response outcome 1, the so-called *logit* of μ_i . The logit is the canonical link function for binary random components. GLMs using the logit link have the form:

$$\log\left(\frac{\mu_i}{1 - \mu_i}\right) = \sum_{j=1}^p \beta_j x_{ij}, \quad i = 1, \dots, n.$$

They are called *logistic regression models*, or sometimes simply *logit models*. Chapter 5 presents such models. Chapter 6 introduces generalized logit models for multinomial random components, for handling categorical response variables that have more than two outcome categories.

Some response variables have counts as their possible outcomes. In a criminal justice study, for instance, each observation might be the number of times a person has been arrested. Counts also occur as entries in contingency tables. The simplest probability distribution for count data is the Poisson. It has natural parameter $\log \mu_i$, so the canonical link function is the log link, $\eta_i = \log \mu_i$. The model using this link function is

$$\log \mu_i = \sum_{j=1}^p \beta_j x_{ij}, \quad i = 1, \dots, n.$$

Presented in Chapter 7, it is called a *Poisson loglinear model*. We will see there that a more flexible model for count data assumes a negative binomial distribution for y_i .

Table 1.1 lists some GLMs presented in Chapters 2–7. Chapter 4 presents basic results for GLMs, such as likelihood equations, ways of finding the ML estimates, and large-sample distributions for the ML estimators.

1.1.6 Advantages of GLMs versus Transforming the Data

A traditional way to model data, introduced long before GLMs, transforms y so that it has approximately a normal conditional distribution with constant variance. Then, the least squares fitting method and subsequent inference for ordinary normal linear

Table 1.1 Important Generalized Linear Models for Statistical Analysis

Random Component	Link Function	Model	Chapters
Normal	Identity	Regression	2 and 3
		Analysis of variance	2 and 3
Exponential family	Any	Generalized linear model	4
Binomial	Logit	Logistic regression	5
Multinomial	Generalized logits	Multinomial response	6
Poisson	Log	Loglinear	7

Chapter 4 presents an overview of GLMs, and the other chapters present special cases.

models presented in the next two chapters are applicable on the transformed scale. For example, with count data that have a Poisson distribution, the distribution is skewed to the right with variance equal to the mean, but \sqrt{y} has a more nearly normal distribution with variance approximately equal to 1/4. For most data, however, it is challenging to find a transformation that provides both approximate normality and constant variance. The best transformation to achieve normality typically differs from the best transformation to achieve constant variance.

With GLMs, by contrast, the choice of link function is separate from the choice of random component. If a link function is useful in the sense that a linear model with the explanatory variables is plausible for that link, it is not necessary that it also stabilizes variance or produces normality. This is because the fitting process maximizes the likelihood for the choice of probability distribution for y , and that choice is not restricted to normality.

Let g denote a function, such as the log function, that is a link function in the GLM approach or a transformation function in the transformed-data approach. An advantage of the GLM formulation is that the model parameters describe $g[E(y_i)]$, rather than $E[g(y_i)]$ as in the transformed-data approach. With the GLM approach, those parameters also describe effects of explanatory variables on $E(y_i)$, after applying the inverse function for g . Such effects are usually more relevant than effects of explanatory variables on $E[g(y_i)]$. For example, with g as the log function, a GLM with $\log[E(y_i)] = \beta_0 + \beta_1 x_{i1}$ translates to an exponential model for the mean, $E(y_i) = \exp(\beta_0 + \beta_1 x_{i1})$, but the transformed-data model² $E[\log(y_i)] = \beta_0 + \beta_1 x_{i1}$ does not translate to exact information about $E(y_i)$ or the effect of x_{i1} on $E(y_i)$. Also, the preferred transform is often not defined on the boundary of the sample space, such as the log transform with a count or a proportion of zero.

GLMs provide a unified theory of modeling that encompasses the most important models for continuous and discrete response variables. Models studied in this text are GLMs with normal, binomial, or Poisson random component, or with extended versions of these distributions such as the multinomial and negative binomial, or multivariate extensions of GLMs. The ML parameter estimates are computed with an algorithm that iteratively uses a weighted version of least squares. The same algorithm applies to the entire exponential family of response distributions, for any choice of link function.

1.2 QUANTITATIVE/QUALITATIVE EXPLANATORY VARIABLES AND INTERPRETING EFFECTS

So far we have learned that a GLM consists of a random component that identifies the response variable and its distribution, a linear predictor that specifies the explanatory variables, and a link function that connects them. We now take a closer look at the form of the linear predictor.

²We are not stating that a model for log-transformed data is never relevant; modeling the mean on the original scale may be misleading when the response distribution is very highly skewed and has many outliers.

1.2.1 Quantitative and Qualitative Variables in Linear Predictors

Explanatory variables in a GLM can be

- quantitative, such as in simple linear regression models.
- qualitative factors, such as in analysis of variance (ANOVA) models.
- mixed, such as an interaction term that is the product of a quantitative explanatory variable and a qualitative factor.

For example, suppose observation i measures an individual’s annual income y_i , number of years of job experience x_{i1} , and gender x_{i2} (1 = female, 0 = male). The linear model with linear predictor

$$\mu_i = \beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + \beta_3x_{i1}x_{i2}$$

has quantitative x_{i1} , qualitative x_{i2} , and mixed $x_{i3} = x_{i1}x_{i2}$ for an interaction term. As Figure 1.1 illustrates, this model corresponds to straight lines $\mu_i = \beta_0 + \beta_1x_{i1}$ for males and $\mu_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_{i1}$ for females. With an interaction term relating two variables, the effect of one variable changes according to the level of the other. For example, with this model, the effect of job experience on mean annual income has slope β_1 for males and $\beta_1 + \beta_3$ for females. The special case, $\beta_3 = 0$, of a lack of interaction corresponds to parallel lines relating mean income to job experience for females and males. The further special case also having $\beta_2 = 0$ corresponds to identical lines for females and males. When we use the model to compare mean incomes for females and males while accounting for the number of years of job experience as a covariate, it is called an *analysis of covariance* model.

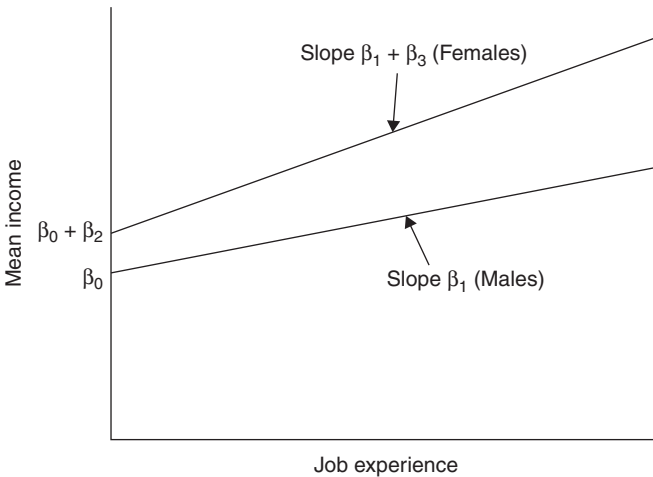


Figure 1.1 Portrayal of linear predictor with quantitative and qualitative explanatory variables.

A quantitative explanatory variable x is represented by a single βx term in the linear predictor and a single column in the model matrix \mathbf{X} . A qualitative explanatory variable having c categories can be represented by $c - 1$ indicator variables and terms in the linear predictor and $c - 1$ columns in the model matrix \mathbf{X} . The R software uses as default the “first-category-baseline” parameterization, which constructs indicators for categories 2, \dots , c . Their parameter coefficients provide contrasts with category 1. For example, suppose racial–ethnic status is an explanatory variable with $c = 3$ categories, (black, Hispanic, white). A model relating mean income to racial–ethnic status could use

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

with $x_{i1} = 1$ for Hispanics and 0 otherwise, $x_{i2} = 1$ for whites and 0 otherwise, and $x_{i1} = x_{i2} = 0$ for blacks. Then β_1 is the difference between the mean income for Hispanics and the mean income for blacks, β_2 is the difference between the mean income for whites and the mean income for blacks, and $\beta_1 - \beta_2$ is the difference between the mean income for Hispanics and the mean income for whites. Some other software, such as SAS, uses an alternative “last-category-baseline” default parameterization, which constructs indicators for categories 1, \dots , $c - 1$. Its parameters then provide contrasts with category c . All such possible choices are equivalent, in terms of having the same model fit.

Shorthand notation can represent terms (variables and their coefficients) in symbols used for linear predictors. A quantitative effect βx is denoted by X , and a qualitative effect is denoted by a letter near the beginning of the alphabet, such as A or B . An interaction is represented³ by a product of such terms, such as $A.B$ or $A.X$. The period represents forming component-wise product vectors of constituent columns from the model matrix. The crossing operator $A*B$ denotes $A + B + A.B$. *Nesting* of categories of B within categories of A (e.g., factor A is states, and factor B is counties within those states) is represented by $A/B = A + A.B$, or sometimes by $A + B(A)$. An intercept term is represented by 1, but this is usually assumed to be in the model unless specified otherwise. Table 1.2 illustrates some simple types of linear predictors and lists the names of normal linear models that equate the mean of the response distribution to that linear predictor.

Table 1.2 Types of Linear Predictors for Normal Linear Models

Linear Predictor	Name of Model
$X_1 + X_2 + X_3 + \dots$	Multiple regression
A	One-way ANOVA
$A + B$	Two-way ANOVA, no interaction
$A + B + A.B$	Two-way ANOVA, interaction
$A + X$ or $A + X + A.X$	Analysis of covariance

³In R, a colon is used, such as $A:B$.

1.2.2 Interval, Nominal, and Ordinal Variables

Quantitative variables are said to be measured on an *interval scale*, because numerical intervals separate levels on the scale. They are sometimes called *interval variables*. A qualitative variable, as represented in a model by a set of indicator variables, has categories that are treated as unordered. Such a categorical variable is called a *nominal variable*.

By contrast, a categorical variable whose categories have a natural ordering is referred to as *ordinal*. For example, attained education might be measured with the categories (<high school, high school graduate, college graduate, postgraduate degree). Ordinal explanatory variables can be treated as qualitative by ignoring the ordering and using a set of indicator variables. Alternatively, they can be treated as quantitative by assigning monotone scores to the categories and using a single βx term in the linear predictor. This is often done when we expect $E(y)$ to progressively increase, or progressively decrease, as we move in order across those ordered categories.

1.2.3 Interpreting Effects in Linear Models

How do we interpret the β coefficients in the linear predictors of GLMs? Suppose the response variable is a college student's math achievement test score y_i , and we fit the linear model having x_{i1} = the student's number of years of math education as an explanatory variable, $\mu_i = \beta_0 + \beta_1 x_{i1}$. Since β_1 is the slope of a straight line, we might say, "If the model holds, a one-year increase in math education corresponds to a change of β_1 in the expected math achievement test score." However, this may suggest the inappropriate causal conclusion that if a student attains another year of math education, her or his math achievement test score is expected to change by β_1 . To validly make such a conclusion, we would need to conduct an experiment that adds a year of math education for each student and then observes the results. Otherwise, a higher mean test score at a higher math education level (if $\beta_1 > 0$) could at least partly reflect the correlation of several other variables with both test score and math education level, such as parents' attained educational levels, the student's IQ, GPA, number of years of science courses, etc. Here is a more appropriate interpretation: If the model holds, when we compare the subpopulation of students having a certain number of years of math education with the subpopulation having one fewer year of math education, the difference in the means of their math achievement test scores is β_1 .

Now suppose the model adds x_{i2} = age of student and x_{i3} = mother's number of years of math education,

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

Since $\beta_1 = \partial\mu_i/\partial x_{i1}$, we might say, "The difference between the mean math achievement test score of a subpopulation of students having a certain number of years of math education and a subpopulation having one fewer year of math education equals β_1 , when we keep constant the student's age and the mother's math education." Controlling variables is possible in designed experiments. But it is unnatural and

possibly inconsistent with the data for many observational studies to envision increasing one explanatory variable while keeping all the others fixed. For example, x_1 and x_2 are likely to be positively correlated, so increases in x_1 naturally tend to occur with increases in x_2 . In some datasets, one might not even observe a 1-unit range in an explanatory variable when the other explanatory variables are all held constant. A better interpretation is this: “The difference between the mean math achievement test score of a subpopulation of students having a certain number of years of math education and a subpopulation having one fewer year equals β_1 , when both subpopulations have the same value for $\beta_2 x_{i2} + \beta_3 x_{i3}$.” More concisely we might say, “The effect of the number of years of math education on the mean math achievement test score equals β_1 , *adjusting*⁴ for student’s age and mother’s math education.” When the model also has a qualitative factor, such as $x_{i4} = \text{gender}$ (1 = female, 0 = male), then β_4 is the difference between the mean math achievement test scores for female and male students, adjusting for the other explanatory variables in the model. Analogous interpretations apply to GLMs for a link-transformed mean.

The effect β_1 in the equation with a sole explanatory variable is usually not the same as β_1 in the equation with multiple explanatory variables, because of factors such as confounding. The effect of x_1 on $E(y)$ will usually differ if we ignore other variables than if we adjust for them, especially in observational studies containing “lurking variables” that are associated both with y and with x_1 . To highlight such a distinction, it is sometimes helpful to use different notation⁵ for the model with multiple explanatory variables, such as

$$\mu_i = \beta_0 + \beta_{y1 \cdot 23} x_{i1} + \beta_{y2 \cdot 13} x_{i2} + \beta_{y3 \cdot 12} x_{i3},$$

where $\beta_{yj \cdot k\ell}$ denotes the effect of x_j on y after adjusting for x_k and x_ℓ .

Some other caveats: In practice, such interpretations use an *estimated* linear predictor, so we replace “mean” by “estimated mean.” Depending on the units of measurement, an effect may be more relevant when expressed with changes other than one unit. When an explanatory variable also occurs in an interaction, then its effect should be summarized separately at different levels of the interacting variable. Finally, for GLMs with nonidentity link function, interpretation is more difficult because β_j refers to the effect on $g(\mu_i)$ rather than μ_i . In later chapters we will present interpretations for various link functions.

1.3 MODEL MATRICES AND MODEL VECTOR SPACES

For the data vector \mathbf{y} with $\boldsymbol{\mu} = E(\mathbf{y})$, consider the GLM $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ with link function g and transformed mean values $\boldsymbol{\eta} = g(\boldsymbol{\mu})$. For this GLM, \mathbf{y} , $\boldsymbol{\mu}$, and $\boldsymbol{\eta}$ are points in n -dimensional Euclidean space, denoted by \mathbb{R}^n .

⁴For linear models, Section 2.5.6 gives a technical definition of *adjusting*, based on removing effects of x_2 and x_3 by regressing both y and x_1 on them.

⁵Yule (1907) introduced such notation in a landmark article on regression modeling.

1.3.1 Model Matrices Induce Model Vector Spaces

Geometrically, model matrices of GLMs naturally induce *vector spaces* that determine the possible $\boldsymbol{\mu}$ for a model. Recall that a vector space S is such that if \mathbf{u} and \mathbf{v} are elements in S , then so are $\mathbf{u} + \mathbf{v}$ and $c\mathbf{u}$ for any constant c .

For a particular $n \times p$ model matrix \mathbf{X} , the values of $\mathbf{X}\boldsymbol{\beta}$ for all possible vectors $\boldsymbol{\beta}$ of model parameters generate a vector space that is a linear subspace of \mathbb{R}^n . For all possible $\boldsymbol{\beta}$, $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ traces out the vector space spanned by the columns of \mathbf{X} , that is, the set of all possible linear combinations of the columns of \mathbf{X} . This is the *column space* of \mathbf{X} , which we denote by $C(\mathbf{X})$,

$$C(\mathbf{X}) = \{\boldsymbol{\eta} : \text{there is a } \boldsymbol{\beta} \text{ such that } \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}\}.$$

In the context of GLMs, we refer to the vector space $C(\mathbf{X})$ as the *model space*. The $\boldsymbol{\eta}$, and hence the $\boldsymbol{\mu}$, that are possible for a particular GLM are determined by the columns of \mathbf{X} .

Two models with model matrices \mathbf{X}_a and \mathbf{X}_b are equivalent if $C(\mathbf{X}_a) = C(\mathbf{X}_b)$. The matrices \mathbf{X}_a and \mathbf{X}_b could be different because of a change of units of an explanatory variable (e.g., pounds to kilograms), or a change in the way of specifying indicator variables for a qualitative predictor. On the other hand, if the model with model matrix \mathbf{X}_a is a special case of the model with model matrix \mathbf{X}_b , for example, with \mathbf{X}_a obtained by deleting one or more of the columns of \mathbf{X}_b , then the model space $C(\mathbf{X}_a)$ is a vector subspace of the model space $C(\mathbf{X}_b)$.

1.3.2 Dimension of Model Space Equals Rank of Model Matrix

Recall that the *rank* of a matrix \mathbf{X} is the number of vectors in a *basis* for $C(\mathbf{X})$, which is a set of linearly independent vectors whose linear combinations generate $C(\mathbf{X})$. Equivalently, the rank is the number of linearly independent columns (or rows) of \mathbf{X} . The *dimension* of the model space $C(\mathbf{X})$ of $\boldsymbol{\eta}$ values, denoted by $\dim[C(\mathbf{X})]$, is defined to be the rank of \mathbf{X} . In all but the final chapter of this book, we assume $p \leq n$, so the model space has dimension no greater than p . We say that \mathbf{X} has *full rank* when $\text{rank}(\mathbf{X}) = p$.

When \mathbf{X} has less than full rank, the columns of \mathbf{X} are linearly dependent, with any one column being a linear combination of the other columns. That is, there exist linear combinations of the columns that yield the $\mathbf{0}$ vector. There are then nonzero $p \times 1$ vectors $\boldsymbol{\zeta}$ such that $\mathbf{X}\boldsymbol{\zeta} = \mathbf{0}$. Such vectors make up the *null space* of the model matrix,

$$N(\mathbf{X}) = \{\boldsymbol{\zeta} : \mathbf{X}\boldsymbol{\zeta} = \mathbf{0}\}.$$

When \mathbf{X} has full rank, then $\dim[N(\mathbf{X})] = 0$. Then, no nonzero combinations of the columns of \mathbf{X} yield $\mathbf{0}$, and $N(\mathbf{X})$ consists solely of the $p \times 1$ zero vector, $\mathbf{0} = (0, 0, \dots, 0)^T$. Generally,

$$\dim[C(\mathbf{X})] + \dim[N(\mathbf{X})] = p.$$

When X has less than full rank, we will see that the model parameters β are not well defined. Then there is said to be *aliasing* of the parameters. In one way this can happen, called *extrinsic aliasing*, an anomaly of the data causes the linear dependence, such as when the values for one predictor are a linear combination of values for the other predictors (i.e., perfect *collinearity*). Another way, called *intrinsic aliasing*, arises when the linear predictor contains inherent redundancies, such as when (in addition to the usual intercept term) we use an indicator variable for each category of a qualitative predictor. The following example illustrates.

1.3.3 Example: The One-Way Layout

Many research studies have the central goal of comparing response distributions for different groups, such as comparing life-length distributions of lung cancer patients under two treatments, comparing mean crop yields for three fertilizers, or comparing mean incomes on the first job for graduating students with various majors. For c groups of independent observations, let y_{ij} denote response observation j in group i , for $i = 1, \dots, c$ and $j = 1, \dots, n_i$. This data structure is called the *one-way layout*.

We regard the groups as c categories of a qualitative factor. For $\mu_{ij} = E(y_{ij})$, the GLM has linear predictor,

$$g(\mu_{ij}) = \beta_0 + \beta_i.$$

Let μ_i denote the common value of $\{\mu_{ij}, j = 1, \dots, n_i\}$, for $i = 1, \dots, c$. For the identity link function and an assumption of normality for the random component, this model is the basis of the *one-way ANOVA* significance test of $H_0: \mu_1 = \dots = \mu_c$, which we develop in Section 3.2. This hypothesis corresponds to the special case of the model in which $\beta_1 = \dots = \beta_c$.

Let $\mathbf{y} = (y_{11}, \dots, y_{1n_1}, \dots, y_{c1}, \dots, y_{cn_c})^T$ and $\beta = (\beta_0, \beta_1, \dots, \beta_c)^T$. Let $\mathbf{1}_{n_i}$ denote the $n_i \times 1$ column vector consisting of n_i entries of 1, and likewise for $\mathbf{0}_{n_i}$. For the one-way layout, the model matrix X for the linear predictor $X\beta$ in the GLM expression $g(\mu) = X\beta$ that represents $g(\mu_{ij}) = \beta_0 + \beta_i$ is

$$X = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_c} & \mathbf{0}_{n_c} & \mathbf{0}_{n_c} & \cdots & \mathbf{1}_{n_c} \end{pmatrix}.$$

This matrix has dimension $n \times p$ with $n = n_1 + \dots + n_c$ and $p = c + 1$.

Equivalently, this parameterization corresponds to indexing the observations as y_h for $h = 1, \dots, n$, defining indicator variables $x_{hi} = 1$ when observation h is in group i and $x_{hi} = 0$ otherwise, for $i = 1, \dots, c$, and expressing the linear predictor for the link function g applied to $E(y_h) = \mu_h$ as

$$g(\mu_h) = \beta_0 + \beta_1 x_{h1} + \cdots + \beta_c x_{hc}.$$

In either case, the indicator variables whose coefficients are $\{\beta_1, \dots, \beta_c\}$ add up to the vector $\mathbf{1}_n$. That vector, which is the first column of X , has coefficient that is

the intercept term β_0 . The columns of X are linearly dependent, because columns 2 through $c + 1$ add up to column 1. Here β_0 is intrinsically aliased with $\sum_{i=1}^c \beta_i$. The parameter β_0 is *marginal* to $\{\beta_1, \dots, \beta_c\}$, in the sense that the column space for the coefficient of β_0 in the model lies wholly in the column space for the vector coefficients of $\{\beta_1, \dots, \beta_c\}$. So, β_0 is redundant in any explanation of the structure of the linear predictor.

Because of the linear dependence of the columns of X , this matrix does not have full rank. But we can achieve full rank merely by dropping one column of X , because we need only $c - 1$ indicators to represent a c -category explanatory variable. This model with one less parameter has the same column space for the reduced model matrix.

1.4 IDENTIFIABILITY AND ESTIMABILITY

In the one-way layout example, let d denote any constant. Suppose we transform the parameters β to a new set,

$$\beta^* = (\beta_0^*, \beta_1^*, \dots, \beta_c^*)^T = (\beta_0 + d, \beta_1 - d, \dots, \beta_c - d)^T.$$

The linear predictor with this new set of parameters is

$$g(\mu_{ij}) = \beta_0^* + \beta_i^* = (\beta_0 + d) + (\beta_i - d) = \beta_0 + \beta_i.$$

That is, the linear predictor $X\beta$ for $g(\mu)$ is exactly the same, for any value of d . So, for the model as specified with $c + 1$ parameters, the parameter values are not unique.

1.4.1 Identifiability of GLM Model Parameters

For this model, because the value for β is not unique, we cannot estimate β uniquely even if we have an infinite amount of data. Whether we assume normality or some other distribution for y , the likelihood equations have infinitely many solutions. When the model matrix is not of full rank, β is not *identifiable*.

Definition. For a GLM with linear predictor $X\beta$, the parameter vector β is *identifiable* if whenever $\beta^* \neq \beta$, then $X\beta^* \neq X\beta$.

Equivalently, β is identifiable if $X\beta^* = X\beta$ implies that $\beta^* = \beta$, so this definition tells us that if we know $g(\mu) = X\beta$ (and hence if we know μ satisfying the model), then we can also determine β .

For the parameterization just given for the one-way layout, β is not identifiable, because $\beta = (\beta_0, \beta_1, \dots, \beta_c)^T$ and $\beta^* = (\beta_0 + d, \beta_1 - d, \dots, \beta_c - d)^T$ do not have different linear predictor values. In such cases, we can obtain identifiability and eliminate the intrinsic aliasing among the parameters by redefining the linear predictor with fewer parameters. Then, different β values have different linear predictor values $X\beta$, and estimation of β is possible.

For the one-way layout, we can either drop a parameter or add a linear constraint. That is, in $g(\mu_{ij}) = \beta_0 + \beta_i$, we might set $\beta_1 = 0$ or $\beta_c = 0$ or $\sum_i \beta_i = 0$ or $\sum_i n_i \beta_i = 0$. With the first-category-baseline constraint $\beta_1 = 0$, we express the model as $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$ with

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \mathbf{1}_{n_3} & \mathbf{0}_{n_3} & \mathbf{1}_{n_3} & \cdots & \mathbf{0}_{n_3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_c} & \mathbf{0}_{n_c} & \mathbf{0}_{n_c} & \cdots & \mathbf{1}_{n_c} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_c \end{pmatrix}.$$

When used with the identity link function, this expression states that $\mu_1 = \beta_0$ (from the first n_1 rows of \mathbf{X}), and for $i > 1$, $\mu_i = \beta_0 + \beta_i$ (from the n_i rows of \mathbf{X} in set i). Thus, the model parameters then represent $\beta_0 = \mu_1$ and $\{\beta_i = \mu_i - \mu_1\}$. Under the last-category-baseline constraint $\beta_c = 0$, the parameters are $\beta_0 = \mu_c$ and $\{\beta_i = \mu_i - \mu_c\}$. Under the constraint $\sum_i n_i \beta_i = 0$, the parameters are $\beta_0 = \bar{\mu}$ and $\{\beta_i = \mu_i - \bar{\mu}\}$, where $\bar{\mu} = (\sum_i n_i \mu_i)/n$.

A slightly more general definition of identifiability refers instead to linear combinations $\boldsymbol{\ell}^T \boldsymbol{\beta}$ of parameters. It states that $\boldsymbol{\ell}^T \boldsymbol{\beta}$ is identifiable if whenever $\boldsymbol{\ell}^T \boldsymbol{\beta}^* \neq \boldsymbol{\ell}^T \boldsymbol{\beta}$, then $\mathbf{X}\boldsymbol{\beta}^* \neq \mathbf{X}\boldsymbol{\beta}$. This definition permits a subset of the terms in $\boldsymbol{\beta}$ to be identifiable, rather than treating the entire $\boldsymbol{\beta}$ as identifiable or nonidentifiable. For example, suppose we extend the model for the one-way layout to include a quantitative explanatory variable taking value x_{ij} for observation j in group i , yielding the analysis of covariance model

$$g(\mu_{ij}) = \beta_0 + \beta_i + \gamma x_{ij}.$$

Then, without a constraint on $\{\beta_i\}$ or β_0 , according to this definition $\{\beta_i\}$ and β_0 are not identifiable, but γ is identifiable. Here, taking $\boldsymbol{\ell}^T \boldsymbol{\beta} = \gamma$, different values of $\boldsymbol{\ell}^T \boldsymbol{\beta}$ yield different values of $\mathbf{X}\boldsymbol{\beta}$.

1.4.2 Estimability in Linear Models

In a non-full-rank model specification, some quantities are unaffected by the parameter nonidentifiability and can be estimated. In a linear model, the adjective *estimable* refers to certain quantities that can be estimated in an unbiased manner.

Definition. In a linear model $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, the quantity $\boldsymbol{\ell}^T \boldsymbol{\beta}$ is *estimable* if there exist coefficients \mathbf{a} such that $E(\mathbf{a}^T \mathbf{y}) = \boldsymbol{\ell}^T \boldsymbol{\beta}$.

That is, some linear combination of the observations estimates $\boldsymbol{\ell}^T \boldsymbol{\beta}$ unbiasedly.

We show now that if $\boldsymbol{\ell}^T \boldsymbol{\beta}$ can be expressed as a linear combination of means, it is estimable. Recall that \mathbf{x}_i denotes row i of the model matrix \mathbf{X} , corresponding to observation y_i , for which $E(y_i) = \mathbf{x}_i \boldsymbol{\beta}$. Letting $\boldsymbol{\ell}^T = \mathbf{x}_i$ and taking \mathbf{a} to be identically 0 except for a 1 in position i , we have $E(\mathbf{a}^T \mathbf{y}) = E(y_i) = \mathbf{x}_i \boldsymbol{\beta} = \boldsymbol{\ell}^T \boldsymbol{\beta}$ for all $\boldsymbol{\beta}$. So $E(y_i) =$