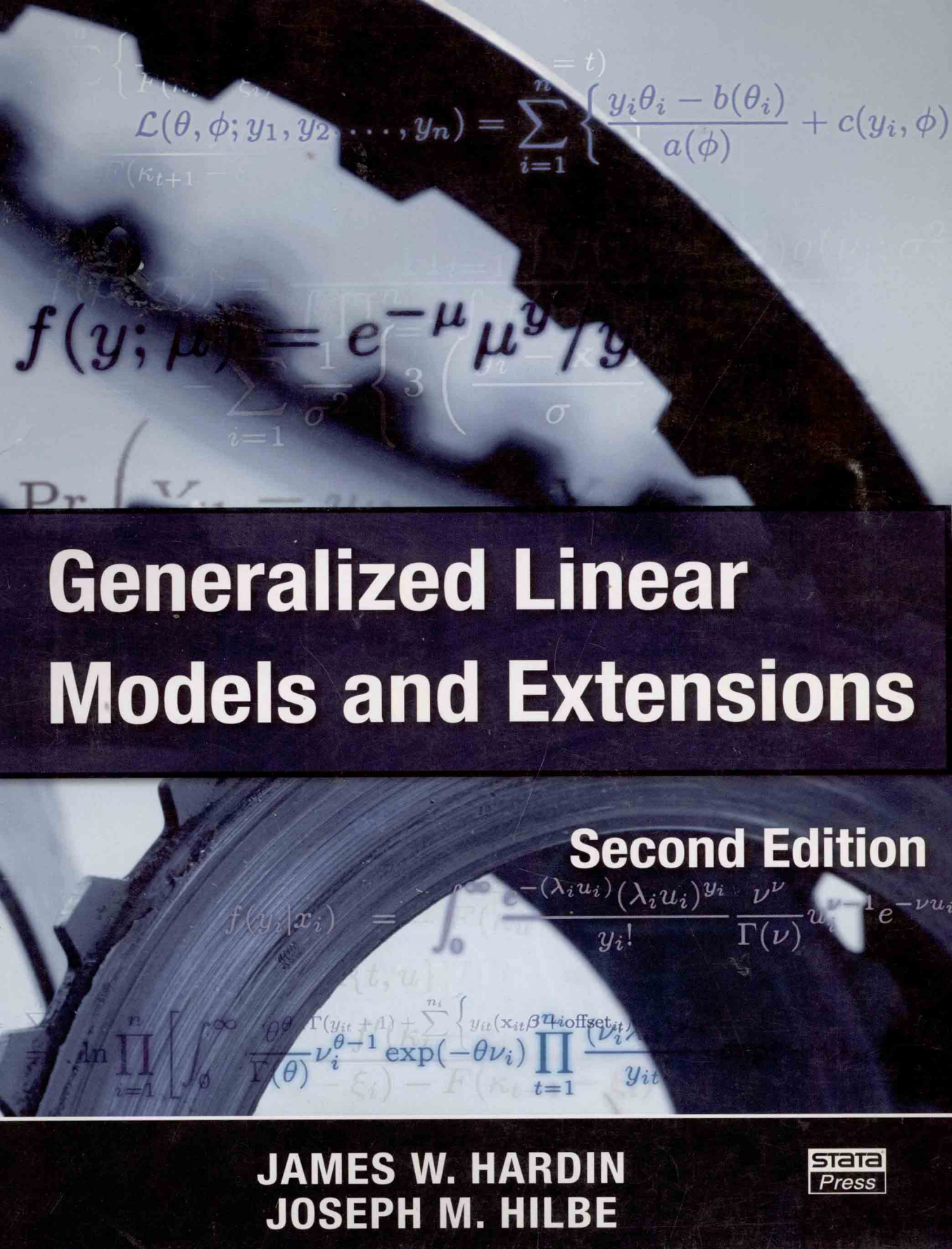




$$\mathcal{L}(\theta, \phi; y_1, y_2, \dots, y_n) = \sum_{i=1}^n \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}$$

$$f(y; \mu) = \frac{e^{-\mu} \mu^y}{y!}$$

$$f(y_i | x_i) = \frac{e^{-(\lambda_i u_i)} (\lambda_i u_i)^{y_i}}{y_i!} \frac{\nu^\nu}{\Gamma(\nu)} u_i^{\nu-1} e^{-\nu u_i}$$

Generalized Linear Models and Extensions

Second Edition

JAMES W. HARDIN
JOSEPH M. HILBE

STATA
Press

Generalized Linear Models and Extensions

Second Edition

James W. Hardin

*Department of Epidemiology and Biostatistics
University of South Carolina*

Joseph M. Hilbe

*Department of Sociology and Statistics
Arizona State University*



A Stata Press Publication
StataCorp LP
College Station, Texas

Stata Press, 4905 Lakeway Drive, College Station, Texas 77845

Copyright © 2001, 2007 by StataCorp LP

All rights reserved. First edition 2001

Second edition 2007

Typeset in L^AT_EX 2_ε

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ISBN-10: 1-59718-014-9

ISBN-13: 978-1-59718-014-6

This book is protected by copyright. All rights are reserved. No part of this book may be reproduced, stored in a retrieval system, or transcribed, in any form or by any means—electronic, mechanical, photocopying, recording, or otherwise—without the prior written permission of StataCorp LP.

Stata is a registered trademark of StataCorp LP. L^AT_EX 2_ε is a trademark of the American Mathematical Society.

For our

wives

Mariaelena Castro-Hardin
Cheryl Hilbe

and

children

Taylor Hardin and Conner Hardin
Michael Hilbe, Mitchell Hilbe,
and Heather Hilbe O'Meara

who were affected by our time away preparing this book (again).

Preface

This second edition of *Generalized Linear Models and Extensions* is written for the active researcher as well as for the theoretical statistician. Our goal has been to clarify the nature and scope of generalized linear models (GLMs) and to demonstrate how all the families, links, and variations of GLMs fit together in an understandable whole.

In a step-by-step manner, we detail the foundations and provide working algorithms that readers can use to construct and better understand models that they wish to develop. In a sense, we offer readers a workbook or handbook of how to deal with data using GLM and GLM extensions.

Many people have contributed to the ideas presented in the new edition of this book. John Nelder has been the foremost influence. Other important and influential people include Peter Bruce, David Collett, David Hosmer, Stanley Lemeshow, James Lindsey, J. Scott Long, Roger Newson, Scott Zeger, Kung-Yee Liang, Raymond J. Carroll, H. Joseph Newton, Henrik Schmiediche, Norman Breslow, Berwin Turlach, Gordon Johnston, Thomas Lumley, Bill Sribney, Vince Wiggins, Mario Cleves, William Greene, and many others. We also thank William Gould, president of StataCorp, for his encouragement in this project. His statistical computing expertise and his contributions to statistical modeling have had a deep impact on this book.

We also thank StataCorp's editorial staff for their equanimity in reading and editing our manuscript, especially Roberto Gutierrez, Patricia Branton, and Lisa Gilmore for their insightful and patient contributions in this area.

Stata Press allowed us to dictate some of the style of this text. In writing this material in other forms for short courses, we have always included equation numbers for all equations rather than only for those equations mentioned in text. Although this is not the standard editorial style for textbooks, we enjoy the benefits of students' being able to more easily (and efficiently) communicate questions and comments on all parts of the material. We hope that readers will find this practice as beneficial as our short-course participants have found it.

James Hardin
Joseph Hilbe

December 2006

Contents

List of Tables	xvii
List of Figures	xix
List of Listings	xxii
Preface	xxiii
1 Introduction	1
1.1 Origins and motivation	1
1.2 Notational conventions	3
1.3 Applied or theoretical?	4
1.4 Road map	4
1.5 Installing the support materials	6
I Foundations of Generalized Linear Models	7
2 GLMs	9
2.1 Components	11
2.2 Assumptions	12
2.3 Exponential family	13
2.4 Example: Using an offset in a GLM	15
2.5 Summary	16
3 GLM estimation algorithms	19
3.1 Newton–Raphson (using the observed Hessian)	25
3.2 Starting values for Newton–Raphson	27
3.3 IRLS (using the expected Hessian)	28
3.4 Starting values for IRLS	31
3.5 Goodness of fit	31
3.6 Estimated variance matrices	32

3.6.1	Hessian	34
3.6.2	Outer product of the gradient	35
3.6.3	Sandwich	35
3.6.4	Modified sandwich	36
3.6.5	Unbiased sandwich	37
3.6.6	Modified unbiased sandwich	38
3.6.7	Weighted sandwich: Newey–West	38
3.6.8	Jackknife	40
3.6.8.1	Usual jackknife	40
3.6.8.2	One-step jackknife	41
3.6.8.3	Weighted jackknife	41
3.6.8.4	Variable jackknife	41
3.6.9	Bootstrap	42
3.6.9.1	Usual bootstrap	42
3.6.9.2	Grouped bootstrap	43
3.7	Estimation algorithms	43
3.8	Summary	44
4	Analysis of fit	47
4.1	Deviance	48
4.2	Diagnostics	49
4.2.1	Cook’s distance	49
4.2.2	Overdispersion	49
4.3	Assessing the link function	50
4.4	Checks for systematic departure from the model	51
4.5	Residual analysis	52
4.5.1	Response residuals	53
4.5.2	Working residuals	53
4.5.3	Pearson residuals	54
4.5.4	Partial residuals	54
4.5.5	Anscombe residuals	54

4.5.6	Deviance residuals	55
4.5.7	Adjusted deviance residuals	55
4.5.8	Likelihood residuals	55
4.5.9	Score residuals	55
4.6	Model statistics	55
4.6.1	Criterion measures	56
4.6.1.1	AIC	56
4.6.1.2	BIC	57
4.6.2	The interpretation of R^2 in linear regression	58
4.6.2.1	Percent variance explained	58
4.6.2.2	The ratio of variances	58
4.6.2.3	A transformation of the likelihood ratio	59
4.6.2.4	A transformation of the F test	59
4.6.2.5	Squared correlation	59
4.6.3	Generalizations of linear regression R^2 interpretations	59
4.6.3.1	Efron's pseudo- R^2	60
4.6.3.2	McFadden's likelihood-ratio index	60
4.6.3.3	Ben-Akiva and Lerman adjusted likelihood-ratio index	60
4.6.3.4	McKelvey and Zavoina ratio of variances	61
4.6.3.5	Transformation of likelihood ratio	61
4.6.3.6	Cragg and Uhler normed measure	61
4.6.4	More R^2 measures	62
4.6.4.1	The count R^2	62
4.6.4.2	The adjusted count R^2	62
4.6.4.3	Veall and Zimmermann R^2	62
4.6.4.4	Cameron-Windmeijer R^2	62

II Continuous-Response Models **65**

5 The Gaussian family **67**

5.1	Derivation of the GLM Gaussian family	68
-----	---	----

5.2	Derivation in terms of the mean	68
5.3	IRLS GLM algorithm (nonbinomial)	70
5.4	Maximum likelihood estimation	73
5.5	GLM log-normal models	74
5.6	Expected versus observed information matrix	75
5.7	Other Gaussian links	77
5.8	Example: Relation to OLS	77
5.9	Example: Beta-carotene	79
6	The gamma family	89
6.1	Derivation of the gamma model	90
6.2	Example: Reciprocal link	92
6.3	Maximum likelihood estimation	95
6.4	Log-gamma models	96
6.5	Identity-gamma models	100
6.6	Using the gamma model for survival analysis	101
7	The inverse Gaussian family	105
7.1	Derivation of the inverse Gaussian model	105
7.2	The inverse Gaussian algorithm	107
7.3	Maximum likelihood algorithm	107
7.4	Example: The canonical inverse Gaussian	108
7.5	Noncanonical links	109
8	The power family and link	113
8.1	Power links	113
8.2	Example: Power link	114
8.3	The power family	115
III	Binomial Response Models	117
9	The binomial–logit family	119
9.1	Derivation of the binomial model	120
9.2	Derivation of the Bernoulli model	123

9.3	The binomial regression algorithm	124
9.4	Example: Logistic regression	126
9.4.1	Model producing logistic coefficients: The heart data	127
9.4.2	Model producing logistic odds ratios	128
9.5	GOF statistics	129
9.6	Interpretation of parameter estimates	132
10	The general binomial family	141
10.1	Noncanonical binomial models	141
10.2	Noncanonical binomial links (binary form)	142
10.3	The probit model	143
10.4	The clog-log and log-log models	148
10.5	Other links	155
10.6	Interpretation of coefficients	156
10.6.1	Identity link	156
10.6.2	Logit link	156
10.6.3	Log link	157
10.6.4	Log complement link	158
10.6.5	Summary	159
10.7	Generalized binomial regression	159
11	The problem of overdispersion	165
11.1	Overdispersion	165
11.2	Scaling of standard errors	170
11.3	Williams' procedure	175
11.4	Robust standard errors	178
IV	Count Response Models	181
12	The Poisson family	183
12.1	Count response regression models	183
12.2	Derivation of the Poisson algorithm	184
12.3	Poisson regression: Examples	189

12.4	Example: Testing overdispersion in the Poisson model	192
12.5	Using the Poisson model for survival analysis	194
12.6	Using offsets to compare models	195
12.7	Interpretation of coefficients	197
13	The negative binomial family	199
13.1	Constant overdispersion	201
13.2	Variable overdispersion	203
13.2.1	Derivation in terms of a Poisson–gamma mixture	203
13.2.2	Derivation in terms of the negative binomial probability function	206
13.2.3	The canonical link negative binomial parameterization	207
13.3	The log-negative binomial parameterization	209
13.4	Negative binomial examples	211
13.5	The geometric family	215
13.6	Interpretation of coefficients	218
14	Other count data models	221
14.1	Count response regression models	221
14.2	Zero-truncated models	224
14.3	Zero-inflated models	227
14.4	Hurdle models	232
14.5	Heterogeneous negative binomial models	235
14.6	Generalized Poisson regression models	239
14.7	Censored count response models	241
V	Multinomial Response Models	249
15	The ordered-response family	251
15.1	Ordered outcomes for general link	252
15.2	Ordered outcomes for specific links	254
15.2.1	Ordered logit	254
15.2.2	Ordered probit	255
15.2.3	Ordered clog-log	255

15.2.4	Ordered log-log	256
15.2.5	Ordered cauchit	256
15.3	Generalized ordered outcome models	257
15.4	Example: Synthetic data	258
15.5	Example: Automobile data	263
15.6	Partial proportional-odds models	269
15.7	Continuation ratio models	273
16	Unordered-response family	279
16.1	The multinomial logit model	280
16.1.1	Example: Relation to logistic regression	280
16.1.2	Example: Relation to conditional logistic regression	281
16.1.3	Example: Extensions with conditional logistic regression	283
16.1.4	The independence of irrelevant alternatives	284
16.1.5	Example: Assessing the IIA	285
16.1.6	Interpreting coefficients	287
16.1.7	Example: Medical admissions—introduction	287
16.1.8	Example: Medical admissions—summary	289
16.2	The multinomial probit model	295
16.2.1	Example: A comparison of the models	297
16.2.2	Example: Comparing probit and multinomial probit	299
16.2.3	Example: Concluding remarks	302
VI	Extensions to the GLM	305
17	Extending the likelihood	307
17.1	The quasilikelihood	307
17.2	Example: Wedderburn's leaf blotch data	308
17.3	Generalized additive models	316
18	Clustered data	319
18.1	Generalization from individual to clustered data	319
18.2	Pooled estimators	320

18.3	Fixed effects	321
18.3.1	Unconditional fixed-effects estimators	322
18.3.2	Conditional fixed-effects estimators	323
18.4	Random effects	325
18.4.1	Maximum likelihood estimation	325
18.4.2	Gibbs sampling	329
18.5	GEEs	330
18.6	Other models	333
VII	Stata Software	337
19	Programs for Stata	339
19.1	The glm command	340
19.1.1	Syntax	340
19.1.2	Description	341
19.1.3	Options	341
19.2	The predict command after glm	345
19.2.1	Syntax	345
19.2.2	Options	345
19.3	User-written programs	347
19.3.1	Global macros available for user-written programs	347
19.3.2	User-written variance functions	348
19.3.3	User-written programs for link functions	350
19.3.4	User-written programs for Newey–West weights	352
19.4	Remarks	353
19.4.1	Equivalent commands	353
19.4.2	Special comments on family(Gaussian) models	353
19.4.3	Special comments on family(binomial) models	353
19.4.4	Special comments on family(nbinomial) models	354
19.4.5	Special comment on family(gamma) link(log) models	354
A	Tables	355

References	369
Author index	379
Subject index	383

Tables

2.1	Predicted values for various choices of variance function	13
9.1	Binomial regression models	119
9.2	Common binomial link functions	120
9.3	Variables for heart data	126
10.1	Common binomial noncanonical link functions	141
10.2	Noncanonical binomial link functions ($\eta = \mathbf{x}\boldsymbol{\beta} + \text{offset}$)	143
10.3	1964 microplot data of carrot fly damage	160
14.1	Other count data models	222
14.2	Variance functions V for count data models; ϕ and α above are constants	223
14.3	Poisson and negative binomial panel data models	223
19.1	Resulting standard errors	343
19.2	Statistics for predict	347
19.3	Equivalent Stata commands	353
A.1	Variance functions	355
A.2	Link and inverse link functions ($\eta = \mathbf{X}\boldsymbol{\beta} + \text{offset}$)	356
A.3	First derivatives of link functions ($\eta = \mathbf{X}\boldsymbol{\beta} + \text{offset}$)	357
A.4	First derivatives of inverse link functions ($\eta = \mathbf{X}\boldsymbol{\beta} + \text{offset}$)	358
A.5	Second derivatives of link functions where $\eta = \mathbf{X}\boldsymbol{\beta} + \text{offset}$ and $\Delta = \partial\eta/\partial\mu$	359
A.6	Second derivatives of inverse link functions where $\eta = \mathbf{X}\boldsymbol{\beta} + \text{offset}$ and $\nabla = \partial\mu/\partial\eta$	360
A.7	Log likelihoods	361

A.8	Weight functions (kernels) for weighted sandwich variance estimates . . .	362
A.9	Pearson residuals	363
A.10	Anscombe residuals	364
A.11	Squared deviance residuals and deviance adjustment factors $\rho_3(\theta)$. . .	365
A.12	Cameron–Windmeijer Kullback–Leibler divergence	366
A.13	Cameron–Windmeijer R^2	367
A.14	Interpretation of power links	368

Figures

5.1	Pearson residuals obtained from linear model	80
5.2	Normal scores versus sorted Pearson residuals obtained from linear model	81
5.3	Pearson residuals versus kilocalories; Pearson residuals obtained from linear model	82
5.4	Pearson residuals obtained from log-normal model (two outliers removed)	84
5.5	Pearson residuals versus fitted values from log-normal model (two outliers removed)	84
5.6	Pearson residuals from lognormal model (log-transformed outcome, two outliers removed, and zero outcome removed)	86
5.7	Pearson residuals versus fitted values from lognormal model (log-transformed outcome, two outliers removed, and zero outcome removed)	86
5.8	Normal scores versus sorted Pearson residuals obtained from log-normal model (log-transformed outcome, two outliers removed, and zero outcome removed)	87
5.9	Pearson residuals versus kilocalories; Pearson residuals obtained from lognormal model (log-transformed outcome, two outliers removed, and zero outcome removed)	88
6.1	Anscombe residuals versus log (variance)	94
9.1	Sample proportions of girls reaching menarche for each age category .	136
9.2	Predicted probabilities of girls reaching menarche for each age category	137
9.3	Predicted probabilities and sample proportions of girls reaching menarche for each age category	138
10.1	Probit and logit functions	144
10.2	Predicted probabilities for probit and logit link function in proportional binary models. The observed (sample) proportions are included as well.	148