

Vector Generalized Linear and Additive Models

With an Implementation in R

This book presents a statistical framework that expands generalized linear models (GLMs) for regression modelling. The framework shared in this book allows analyses based on many semi-traditional applied statistics models to be performed as a coherent whole. This is possible through the approximately half-a-dozen major classes of statistical models included in the book and the software infrastructure component, which makes the models easily operable.

The book's methodology and accompanying software (the extensive VGAM R package) are directed at these limitations, and this is the first time the methodology and software are covered comprehensively in one volume. Since their advent in 1972, GLMs have unified important distributions under a single umbrella with enormous implications. The demands of practical data analysis, however, require a flexibility that GLMs do not have. Data-driven GLMs, in the form of generalized additive models (GAMs), are also largely confined to the exponential family. This book treats distributions and classical models as generalized regression models, and the result is a much broader application base for GLMs and GAMs.

The book may be used in senior undergraduate and first-year postgraduate courses on GLMs and regression modeling, including categorical data analysis. It may also serve as a reference on vector generalized linear models and as a methodology resource for VGAM users. The methodological contribution of this book stands alone and does not require use of the VGAM package. In the second part of the book, the R package VGAM makes applications of the methodology immediate. R code is integrated in the text, and datasets are used throughout. Potential applications include ecology, finance, biostatistics, and social sciences.

Statistics

ISBN 978-1-4939-2817-0



9 781493 928170



Part I General Theory

1	Introduction	3
1.1	Introduction	3
1.1.1	Outline of This Book and a Quick Start	4
1.2	Six Illustrative Models	4
1.2.1	The Linear Model	5
1.2.2	Poisson and Negative Binomial Regression	6
1.2.3	Bivariate Odds Ratio Model	8
1.2.4	Proportional Odds and Multinomial Logit Models	11
1.3	General Framework	13
1.3.1	Vector Generalized Linear Models	13
1.3.2	Vector Generalized Additive Models	14
1.3.3	RR-VGLMs	15
1.3.4	QRR-VGLMs and Constrained Ordination	16
1.3.5	RR-VGAMs and Constrained Ordination	17
1.3.6	RCIMs	17
1.4	An Overview of VGAM	17
1.5	Some Background Topics	18
1.5.1	The Penalty Function Approach	19
1.5.2	Snippets of the S Language	20
1.5.3	More on the Number of Parameters and Terms	29
1.6	Summary	29
	Exercises	30
2	LMs, GLMs and GAMs	33
2.1	Introduction	33
2.2	LMs	33
2.2.1	The Hat Matrix	36
2.2.2	WLS and GLS	38
2.2.3	Fitting LMs in R	39
2.3	GLM Basics	39
2.3.1	Inference	42
2.3.2	GLM Residuals and Diagnostics	43

2.3.3	Estimation of ϕ	44
2.3.4	Fitting GLMs in R	45
2.3.5	Quasi-Likelihood Models	45
2.3.6	Binary Responses	46
2.4	Univariate Smoothing Methods	49
2.4.1	The Classical Smoothing Problem	49
2.4.2	Polynomial Regression	51
2.4.3	Regression Splines	52
2.4.4	Smoothing Splines	59
2.4.5	P-Splines	64
2.4.6	Local Regression	66
2.4.7	Some General Theory	74
2.5	Generalized Additive Models	81
2.5.1	Why Additive Models?	82
2.5.2	Binomial ‘example’	83
	Exercises	85
3	VGLMs	91
3.1	Introduction	91
3.2	Iteratively Reweighted Least Squares	92
3.2.1	Computation †	94
3.2.2	Advantages and Disadvantages of IRLS	96
3.3	Constraints on the Component Functions	96
3.3.1	Fitting Constrained Models in VGAM	99
3.4	The <code>xij</code> Argument	103
3.4.1	The Central Formula	105
3.4.2	Using the <code>xij</code> Argument in VGAM	106
3.4.3	More Complicated Examples	107
3.4.4	Smoothing	110
3.4.5	Last Word	114
3.5	Other Topics	115
3.5.1	Multiple Responses	115
3.5.2	Deviance	116
3.5.3	Convergence	117
3.5.4	Half-Stepping	118
3.6	Inference	120
3.6.1	Regularity Conditions	120
3.6.2	Parameter Link Functions	120
3.6.3	Hypothesis Testing	120
3.6.4	Residual Degrees of Freedom	120
3.7	Residuals and Diagnostics	121
3.7.1	Working Residuals	121
3.7.2	Pearson Residuals	122
3.7.3	Response Residuals	122
3.7.4	Deviance Residuals	123
3.7.5	Hat Matrix	123
	Exercises	125

4	VGAMs	127
4.1	Introduction	127
4.1.1	Smoothing for VGLMs and VGAMs	127
4.1.2	The Vector Smoothing Problem	129
4.2	Vector Smoothing Methods	129
4.2.1	Vector Splines	130
4.2.2	Local Regression for Vector Responses †	141
4.2.3	On Linear Vector Smoothers	152
4.3	The Vector Additive Model and VGAM Estimation	153
4.3.1	Penalized Likelihood	153
4.3.2	Vector Backfitting	154
4.3.3	Degrees of Freedom and Standard Errors	156
4.3.4	Score Tests for Linearity †	157
4.4	On More Practical Aspects	159
4.4.1	Using the Software	159
4.4.2	<code>vsmooth.spline()</code>	159
4.4.3	Example: Cats and Dogs	160
	Exercises	164
5	Reduced-Rank VGLMs	167
5.1	Introduction	167
5.2	What Are RR-VGLMs?	168
5.2.1	Why Reduced-Rank Regression?	169
5.2.2	Normalizations	170
5.2.3	The Stereotype Model	171
5.3	A Few Details	171
5.3.1	Alternating Algorithm	171
5.3.2	SEs of RR-VGLMs	172
5.4	Other RR-VGLM Topics	172
5.4.1	Summary of RR-VGLM Software	172
5.4.2	Convergence	173
5.4.3	Latent Variable Plots and Biplots	173
5.4.4	Miscellaneous Notes	174
5.5	RR-VGLMs with Two Linear Predictors	174
5.5.1	Two-Parameter Rank-1 RR-VGLMs	174
5.5.2	Some Examples	176
5.6	RR-VGLM Examples	179
5.6.1	RR-Multiple Binomial Model	179
5.6.2	Reduced Rank Regression for Time Series	181
5.7	Row-Column Interaction Models	183
5.7.1	<code>rcim()</code>	184
5.7.2	Examples	187
5.7.3	Quasi-Variations	190
	Exercises	195
6	Constrained Quadratic Ordination	201
6.1	Introduction	201
6.1.1	Ordination	201
6.1.2	Prediction and Calibration	204

6.2	Quadratic RR-VGLMs for CQO	205
6.2.1	An Example: Hunting Spiders Data	207
6.2.2	Normalizations for QRR-VGLMs	211
6.3	Fitting QRR-VGLMs	213
6.3.1	Arguments <code>I.tolerances</code> and <code>eq.tolerances</code>	215
6.3.2	Initial Values and the <code>isd.latvar</code> Argument	217
6.3.3	Estimation	218
6.4	Post-Fitting Analyses	219
6.4.1	Arguments <code>varI.latvar</code> and <code>refResponse</code>	219
6.4.2	Ordination Diagrams	221
6.4.3	Perspective Plots	223
6.4.4	Trajectory Plots	224
6.4.5	Calibration	225
6.5	Some Practical Considerations	227
6.6	A Further Example: Trout Data	230
6.7	Unconstrained Quadratic Ordination	232
6.7.1	RCIMs and UQO	232
	Exercises	236
7	Constrained Additive Ordination	239
7.1	Introduction	239
7.2	Constrained Additive Ordination	239
7.2.1	Controlling Function Flexibility	240
7.2.2	Estimation	241
7.2.3	Practical Advice	241
7.3	Examples	242
7.3.1	Hunting Spiders	242
7.3.2	Trout Data	242
7.3.3	Diseases in a Cross-Sectional Study	245
7.4	Some Afterthoughts	246
	Exercises	247
8	Using the VGAM Package	249
8.1	Introduction	249
8.1.1	Naming Conventions of VGAM Family Functions	249
8.1.2	Naming Conventions of Arguments	249
8.2	Basic Usage of VGAM	252
8.2.1	Some Miscellaneous Arguments	252
8.2.2	Constraints	253
8.2.3	Control Functions	254
8.2.4	Convergence Criteria	255
8.2.5	Smart Prediction	256
8.3	More Advanced Usage of VGAM	258
8.3.1	Initial Values	258
8.3.2	Speeding Up the Computations	259
8.4	Some Details on Selected Methods Functions	260
8.4.1	The <code>fitted()</code> Generic	260
8.4.2	The <code>summary()</code> Generic	263
8.4.3	The <code>simulate()</code> Generic	264

8.4.4	The <code>plot()</code> Generic	265
8.4.5	The <code>predict()</code> Generic	267
8.5	Some Suggestions for Fitting VGLMs and VGAMs	268
8.5.1	Doing Things in Style	269
8.5.2	Some Useful Miscellanea	270
8.6	Slots in <code>vgam()</code> Objects	271
8.7	Solutions to Some Specific Problems	272
8.7.1	Obtaining the LM-Type Model Matrix for η_j	272
8.7.2	Predicting at $\bar{\mathbf{x}}$	273
	Exercises	274
9	Other Topics	277
9.1	Introduction	277
9.2	Computing the Working Weights $\dagger$	277
9.2.1	Weight Matrices Not Positive-Definite	278
9.2.2	Simulated Fisher Scoring	278
9.2.3	The BHHH Method	279
9.2.4	Quasi-Newton Updates	280
9.2.5	Numerical Derivatives	280
9.3	Model Selection by AIC and BIC	281
9.4	Bias-Reduction	282
9.4.1	Binary Case	285
9.4.2	Software Implementation	285
	Exercises	286

Part II Some Applications

10	Some LM and GLM Variants	291
10.1	Introduction	291
10.2	LM Variants	291
10.2.1	Varying-Coefficient Models	291
10.2.2	The Tobit Model	294
10.2.3	Seemingly Unrelated Regressions	297
10.2.4	The AR(1) Time Series Model	303
10.3	Binomial Variants	305
10.3.1	Two-Stage Sequential Binomial	305
10.3.2	The Bradley-Terry Model	306
10.3.3	Bivariate Responses: The Bivariate Probit Model	309
10.3.4	Binary Responses: Loglinear Models	310
10.3.5	Double Exponential Models	312
	Exercises	313
11	Univariate Discrete Distributions	317
11.1	Introduction	317
11.1.1	<code>dpqr</code> -Type Functions	318
11.2	Dispersion Models	321
11.3	Negative Binomial Regression	322
11.3.1	Computational Details	323
11.3.2	A Second Parameterization— <code>polyaR()</code>	324
11.3.3	Canonical Link	325

11.3.4	Fitting Other NB Variants	325
11.3.5	Some Practical Suggestions	327
11.4	The Beta-Binomial Model	329
11.5	Lagrangian Probability Distributions	330
	Exercises	332
12	Univariate Continuous Distributions	343
12.1	Introduction	343
12.2	Some Basics	345
12.2.1	Location, Scale and Shape Parameters	345
12.2.2	Initial Values	348
12.2.3	About the Tables	350
12.2.4	Q-Q Plots	351
12.2.5	Scale and Shape Parameters	353
12.3	A Few Groups of Distributions	353
12.3.1	Size Distributions	353
12.3.2	Pearson System	354
12.3.3	Poisson Points in the Plane and Volume	355
	Exercises	355
13	Bivariate Continuous Distributions	371
13.1	Introduction	371
13.1.1	Bivariate Distribution Theory—A Short Summary	371
13.1.2	dpr-Type Functions	373
13.2	Two Bivariate Distributions	373
13.2.1	Bivariate Normal Distribution	373
13.2.2	Plackett's Bivariate Distribution	375
13.3	Copulas	376
13.3.1	Dependency Measures	378
13.3.2	Archimedean Copulas	379
13.3.3	Initial Values	380
	Exercises	380
14	Categorical Data Analysis	385
14.1	Introduction	385
14.1.1	The Multinomial Distribution	385
14.2	Nominal Responses—The Multinomial Logit Model	386
14.2.1	The x_{ij} Argument	388
14.2.2	Marginal Effects and Deviance	390
14.2.3	Stereotype Model	390
14.2.4	Summation Constraints	393
14.3	Using the Software	393
14.3.1	The Poisson Trick	394
14.4	Ordinal Responses	396
14.4.1	Models Involving Cumulative Probabilities	396
14.4.2	Coalminers Example I	401
14.4.3	Coalminers Example II	402
14.4.4	Models Involving Stopping- and Continuation-Ratios	404
14.4.5	Models Involving Adjacent Categories	406
14.4.6	Convergence	406

14.5	Genetic Models	407
14.5.1	The Dirichlet Distribution	408
14.5.2	The Dirichlet-Multinomial Distribution	410
	Exercises	411
15	Quantile and Expectile Regression	415
15.1	Introduction	415
15.1.1	Some Notation and Background	416
15.2	LMS-Type Methods	416
15.2.1	The Box-Cox-Normal Distribution Version	417
15.2.2	Other Versions	419
15.2.3	Example	420
15.2.4	More Advanced Usage of Some Methods Functions	422
15.3	Classical Quantile Regression by Scoring	424
15.3.1	Classical Quantile Regression	425
15.3.2	An Asymmetric Laplace Distribution	427
15.3.3	Example	429
15.3.4	The Onion Method	431
15.4	Expectile Regression	433
15.4.1	Introduction	433
15.4.2	Expectiles for the Linear Model	437
15.4.3	ALS Example	439
15.4.4	Poisson Regression	440
15.4.5	Melbourne Temperatures and the Onion Method	442
15.5	Discussion	442
	Exercises	443
16	Extremes	447
16.1	Introduction	447
16.1.1	Classical EV Theory	448
16.2	GEV	450
16.2.1	The r -Largest Order Statistics	452
16.2.2	The Gumbel and Block-Gumbel Models	452
16.3	GPD	454
16.4	Diagnostics	456
16.4.1	Probability and Quantile Plots	456
16.4.2	Gumbel Plots	457
16.4.3	Mean Excess Plots	457
16.5	Some Software Details	457
16.6	Examples	458
16.6.1	Port Pirie Sea Levels: GEV Model	458
16.6.2	Venice Sea Levels: The Block-Gumbel Model	460
16.6.3	Daily Rainfall Data: The GPD Model	463
	Exercises	466
17	Zero-Inflated, Zero-Altered and Positive Discrete Distributions	469
17.1	Introduction	469
17.1.1	Software Details	472
17.1.2	A Zero-Inflated Poisson Example	472

17.2	The Positive-Bernoulli Distribution	474
17.2.1	Further Software Details	480
17.2.2	Deermice Example	480
17.2.3	Prinia Example	482
17.2.4	A $\mathcal{M}_{tbh}$ Example	484
17.2.5	Using <code>Select()</code>	486
17.2.6	Ephemeral and Enduring Memory Example	487
17.3	The Zero-Inflated Binomial Distribution	489
17.4	RR-ZIPs and RR-ZA Models	489
17.4.1	RR-ZAPs and RR-ZABs and Other Variants	490
	Exercises	491
18	On VGAM Family Functions	499
18.1	Introduction	499
18.2	Link Functions	501
18.2.1	Chain Rule Formulas	502
18.3	Family Function Basics	504
18.3.1	A Simple VGAM Family Function	504
18.3.2	Initial Values	509
18.3.3	Arguments in VGAM Family Functions	509
18.3.4	Extending the Exponential Distribution Family	510
18.3.5	The <code>wz</code> Data Structure	514
18.3.6	Implementing Constraints Within Family Functions	517
18.4	Some Other Topics	519
18.4.1	Writing R Packages and Documentation	519
18.4.2	Some S4 Issues	519
18.5	Examples	520
18.5.1	The Kumaraswamy Distribution Family	520
18.5.2	Simulated Fisher Scoring	524
18.6	Writing Smart Functions $\dagger$	526
	Exercises	530
A	Background Material	533
A.1	Some Classical Likelihood Theory	533
A.1.1	Likelihood Functions	533
A.1.2	Maximum Likelihood Estimation	534
A.1.3	Properties of Maximum Likelihood Estimators	542
A.1.4	Inference	544
A.2	Some Useful Formulas	549
A.2.1	Change of Variable Technique	549
A.2.2	Series Expansions	549
A.2.3	Order Notation	549
A.2.4	Conditional Expectations	551
A.2.5	Random Vectors	551
A.3	Some Linear Algebra	551
A.3.1	Cholesky Decomposition	552
A.3.2	Sherman-Morrison Formulas	553
A.3.3	QR Method	553
A.3.4	Singular Value Decomposition	554

A.4	Some Special Functions	554
A.4.1	Gamma, Digamma and Trigamma Functions	555
A.4.2	Beta Function	556
A.4.3	The Riemann Zeta Function	556
A.4.4	Erf and Erfc	556
A.4.5	The Marcum Q-Function	557
A.4.6	Exponential Integral, Debye Function	557
A.4.7	Bessel Functions	557
	Exercises	559
	Glossary	561
	References	567
	Index	585