

Generalized Linear Models

Second Edition

P. McCullagh FRS

*Department of Statistics,
University of Chicago*

and

J.A. Nelder FRS

*Department of Mathematics,
Imperial College of Science and Technology,
London*

CHAPMAN & HALL/CRC

Boca Raton London New York Washington, D.C.

Contents

Preface to the first edition	xvi
Preface	xviii
1 Introduction	1
1.1 Background	1
1.1.1 The problem of looking at data	3
1.1.2 Theory as pattern	4
1.1.3 Model fitting	5
1.1.4 What is a good model?	7
1.2 The origins of generalized linear models	8
1.2.1 Terminology	8
1.2.2 Classical linear models	9
1.2.3 R.A. Fisher and the design of experiments	10
1.2.4 Dilution assay	11
1.2.5 Probit analysis	13
1.2.6 Logit models for proportions	14
1.2.7 Log-linear models for counts	14
1.2.8 Inverse polynomials	16
1.2.9 Survival data	16
1.3 Scope of the rest of the book	17
1.4 Bibliographic notes	19
1.5 Further results and exercises 1	19
2 An outline of generalized linear models	21
2.1 Processes in model fitting	21
2.1.1 Model selection	21
2.1.2 Estimation	23
2.1.3 Prediction	25

2.2	The components of a generalized linear model	26
2.2.1	The generalization	27
2.2.2	Likelihood functions	28
2.2.3	Link functions	30
2.2.4	Sufficient statistics and canonical links	32
2.3	Measuring the goodness of fit	33
2.3.1	The discrepancy of a fit	33
2.3.2	The analysis of deviance	35
2.4	Residuals	37
2.4.1	Pearson residual	37
2.4.2	Anscombe residual	38
2.4.3	Deviance residual	39
2.5	An algorithm for fitting generalized linear models	40
2.5.1	Justification of the fitting procedure	41
2.6	Bibliographic notes	43
2.7	Further results and exercises 2	44
3	Models for continuous data with constant variance	48
3.1	Introduction	48
3.2	Error structure	49
3.3	Systematic component (linear predictor)	51
3.3.1	Continuous covariates	51
3.3.2	Qualitative covariates	52
3.3.3	Dummy variates	54
3.3.4	Mixed terms	55
3.4	Model formulae for linear predictors	56
3.4.1	Individual terms	56
3.4.2	The dot operator	56
3.4.3	The + operator	57
3.4.4	The crossing (*) and nesting (/) operators	58
3.4.5	Operators for the removal of terms	59
3.4.6	Exponential operator	60
3.5	Aliasing	61
3.5.1	Intrinsic aliasing with factors	63
3.5.2	Aliasing in a two-way cross-classification	65
3.5.3	Extrinsic aliasing	68
3.5.4	Functional relations among covariates	69
3.6	Estimation	70
3.6.1	The maximum-likelihood equations	70
3.6.2	Geometrical interpretation	71

CONTENTS

ix

3.6.3	Information	72
3.6.4	A model with two covariates	74
3.6.5	The information surface	77
3.6.6	Stability	78
3.7	Tables as data	79
3.7.1	Empty cells	79
3.7.2	Fused cells	81
3.8	Algorithms for least squares	81
3.8.1	Methods based on the information matrix	82
3.8.2	Direct decomposition methods	85
3.8.3	Extension to generalized linear models	88
3.9	Selection of covariates	89
3.10	Bibliographic notes	93
3.11	Further results and exercises 3	93
4	Binary data	98
4.1	Introduction	98
4.1.1	Binary responses	98
4.1.2	Covariate classes	99
4.1.3	Contingency tables	100
4.2	Binomial distribution	101
4.2.1	Genesis	101
4.2.2	Moments and cumulants	102
4.2.3	Normal limit	103
4.2.4	Poisson limit	105
4.2.5	Transformations	105
4.3	Models for binary responses	107
4.3.1	Link functions	107
4.3.2	Parameter interpretation	110
4.3.3	Retrospective sampling	111
4.4	Likelihood functions for binary data	114
4.4.1	Log likelihood for binomial data	114
4.4.2	Parameter estimation	115
4.4.3	Deviance function	118
4.4.4	Bias and precision of estimates	119
4.4.5	Sparseness	120
4.4.6	Extrapolation	122
4.5	Over-dispersion	124
4.5.1	Genesis	124
4.5.2	Parameter estimation	126

4.6	Example	128
4.6.1	Habitat preferences of lizards	128
4.7	Bibliographic notes	135
4.8	Further results and exercises 4	135
5	Models for polytomous data	149
5.1	Introduction	149
5.2	Measurement scales	150
5.2.1	General points	150
5.2.2	Models for ordinal scales	151
5.2.3	Models for interval scales	155
5.2.4	Models for nominal scales	159
5.2.5	Nested or hierarchical response scales	160
5.3	The multinomial distribution	164
5.3.1	Genesis	164
5.3.2	Moments and cumulants	165
5.3.3	Generalized inverse matrices	168
5.3.4	Quadratic forms	169
5.3.5	Marginal and conditional distributions	170
5.4	Likelihood functions	171
5.4.1	Log likelihood for multinomial responses	171
5.4.2	Parameter estimation	172
5.4.3	Deviance function	174
5.5	Over-dispersion	174
5.6	Examples	175
5.6.1	A cheese-tasting experiment	175
5.6.2	Pneumoconiosis among coalminers	178
5.7	Bibliographic notes	182
5.8	Further results and exercises 5	184
6	Log-linear models	193
6.1	Introduction	193
6.2	Likelihood functions	194
6.2.1	Poisson distribution	194
6.2.2	The Poisson log-likelihood function	197
6.2.3	Over-dispersion	198
6.2.4	Asymptotic theory	200
6.3	Examples	200
6.3.1	A biological assay of tuberculins	200
6.3.2	A study of wave damage to cargo ships	204

6.4	Log-linear models and multinomial response models	209
6.4.1	Comparison of two or more Poisson means	209
6.4.2	Multinomial response models	211
6.4.3	Summary	213
6.5	Multiple responses	214
6.5.1	Introduction	214
6.5.2	Independence and conditional independence	215
6.5.3	Canonical correlation models	217
6.5.4	Multivariate regression models	219
6.5.5	Multivariate model formulae	222
6.5.6	Log-linear regression models	223
6.5.7	Likelihood equations	225
6.6	Example	229
6.6.1	Respiratory ailments of coalminers	229
6.6.2	Parameter interpretation	233
6.7	Bibliographic notes	235
6.8	Further results and exercises 6	236
7	Conditional likelihoods	245
7.1	Introduction	245
7.2	Marginal and conditional likelihoods	246
7.2.1	Marginal likelihood	246
7.2.2	Conditional likelihood	248
7.2.3	Exponential-family models	252
7.2.4	Profile likelihood	254
7.3	Hypergeometric distributions	255
7.3.1	Central hypergeometric distribution	255
7.3.2	Non-central hypergeometric distribution	257
7.3.3	Multivariate hypergeometric distribution	260
7.3.4	Multivariate non-central distribution	261
7.4	Some applications involving binary data	262
7.4.1	Comparison of two binomial probabilities	262
7.4.2	Combination of information from 2x2 tables	265
7.4.3	Ille-et-Vilaine study of oesophageal cancer	267
7.5	Some applications involving polytomous data	270
7.5.1	Matched pairs: nominal response	270
7.5.2	Ordinal responses	273
7.5.3	Example	276
7.6	Bibliographic notes	277
7.7	Further results and exercises 7	279

8 Models with constant coefficient of variation	285
8.1 Introduction	285
8.2 The gamma distribution	287
8.3 Models with gamma-distributed observations	289
8.3.1 The variance function	289
8.3.2 The deviance	290
8.3.3 The canonical link	291
8.3.4 Multiplicative models: log link	292
8.3.5 Linear models: identity link	294
8.3.6 Estimation of the dispersion parameter	295
8.4 Examples	296
8.4.1 Car insurance claims	296
8.4.2 Clotting times of blood	300
8.4.3 Modelling rainfall data using two generalized linear models	302
8.4.4 Developmental rate of <i>Drosophila melanogaster</i>	306
8.5 Bibliographic notes	313
8.6 Further results and exercises 8	314
9 Quasi-likelihood functions	323
9.1 Introduction	323
9.2 Independent observations	324
9.2.1 Covariance functions	324
9.2.2 Construction of the quasi-likelihood function	325
9.2.3 Parameter estimation	327
9.2.4 Example: incidence of leaf-blotch on barley	328
9.3 Dependent observations	332
9.3.1 Quasi-likelihood estimating equations	332
9.3.2 Quasi-likelihood function	333
9.3.3 Example: estimation of probabilities from marginal frequencies	336
9.4 Optimal estimating functions	339
9.4.1 Introduction	339
9.4.2 Combination of estimating functions	340
9.4.3 Example: estimation for megalithic stone rings	343
9.5 Optimality criteria	347
9.6 Extended quasi-likelihood	349
9.7 Bibliographic notes	352
9.8 Further results and exercises 9	352

10 Joint modelling of mean and dispersion	357
10.1 Introduction	357
10.2 Model specification	358
10.3 Interaction between mean and dispersion effects	359
10.4 Extended quasi-likelihood as a criterion	360
10.5 Adjustments of the estimating equations	361
10.5.1 Adjustment for kurtosis	361
10.5.2 Adjustment for degrees of freedom	362
10.5.3 Summary of estimating equations for the dispersion model	363
10.6 Joint optimum estimating equations	364
10.7 Example: the production of leaf-springs for trucks	365
10.8 Bibliographic notes	370
10.9 Further results and exercises 10	371
 11 Models with additional non-linear parameters	 372
11.1 Introduction	372
11.2 Parameters in the variance function	373
11.3 Parameters in the link function	375
11.3.1 One link parameter	375
11.3.2 More than one link parameter	377
11.3.3 Transformation of data vs transformation of fitted values	378
11.4 Non-linear parameters in the covariates	379
11.5 Examples	381
11.5.1 The effects of fertilizers on coastal Bermuda grass	381
11.5.2 Assay of an insecticide with a synergist	384
11.5.3 Mixtures of drugs	386
11.6 Bibliographic notes	389
11.7 Further results and exercises 11	389
 12 Model checking	 391
12.1 Introduction	391
12.2 Techniques in model checking	392
12.3 Score tests for extra parameters	393
12.4 Smoothing as an aid to informal checks	394
12.5 The raw materials of model checking	396

12.6	Checks for systematic departure from model	398
12.6.1	Informal checks using residuals	398
12.6.2	Checking the variance function	400
12.6.3	Checking the link function	401
12.6.4	Checking the scales of covariates	401
12.6.5	Checks for compound discrepancies	403
12.7	Checks for isolated departures from the model	403
12.7.1	Measure of leverage	405
12.7.2	Measure of consistency	406
12.7.3	Measure of influence	406
12.7.4	Informal assessment of extreme values	407
12.7.5	Extreme points and checks for systematic discrepancies	408
12.8	Examples	409
12.8.1	Carrot damage in an insecticide experiment	409
12.8.2	Minitab tree data	410
12.8.3	Insurance claims (continued)	413
12.9	A strategy for model checking?	414
12.10	Bibliographic notes	415
12.11	Further results and exercises 12	416
13	Models for survival data	419
13.1	Introduction	419
13.1.1	Survival functions and hazard functions	419
13.2	Proportional-hazards models	421
13.3	Estimation with a specified survival distribution	422
13.3.1	The exponential distribution	423
13.3.2	The Weibull distribution	423
13.3.3	The extreme-value distribution	424
13.4	Example: remission times for leukaemia	425
13.5	Cox's proportional-hazards model	426
13.5.1	Partial likelihood	426
13.5.2	The treatment of ties	427
13.5.3	Numerical methods	429
13.6	Bibliographic notes	430
13.7	Further results and exercises 13	430
14	Components of dispersion	432
14.1	Introduction	432
14.2	Linear models	433

CONTENTS	xv
14.3 Non-linear models	434
14.4 Parameter estimation	437
14.5 Example: A salamander mating experiment	439
14.5.1 Introduction	439
14.5.2 Experimental procedure	441
14.5.3 A linear logistic model with random effects	444
14.5.4 Estimation of the dispersion parameters	448
14.6 Bibliographic notes	450
14.7 Further results and exercises 14	452
 15 Further topics	 455
15.1 Introduction	455
15.2 Bias adjustment	455
15.2.1 Models with canonical link	455
15.2.2 Non-canonical models	457
15.2.3 Example: Lizard data (continued)	458
15.3 Computation of Bartlett adjustments	459
15.3.1 General theory	459
15.3.2 Computation of the adjustment	460
15.3.3 Example: exponential regression model	463
15.4 Generalized additive models	465
15.4.1 Algorithms for fitting	465
15.4.2 Smoothing methods	466
15.4.3 Conclusions	467
15.5 Bibliographic notes	467
15.6 Further results and exercises 15	467
 Appendices	 469
A Elementary likelihood theory	469
B Edgeworth series	474
C Likelihood-ratio statistics	476
 References	 479
 Index of data sets	 500
 Author index	 501
 Subject index	 506