

Monographs
on Statistics and
Applied Probability 37

Generalized Linear Models

SECOND EDITION

P. McCullagh and
J.A. Nelder FRS



Chapman and Hall

Generalized Linear Models

SECOND EDITION

P. McCULLAGH

*Department of Statistics,
University of Chicago*

and

J.A. NELDER FRS

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



LONDON NEW YORK
CHAPMAN AND HALL

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

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 2×2 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

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

Preface to the first edition

This monograph deals with a class of statistical models that generalizes classical linear models to include many other models that have been found useful in statistical analysis. These other models include log-linear models for the analysis of data in the form of counts, probit and logit models for data in the form of proportions (ratios of counts), and models for continuous data with constant proportional standard error. In addition, important types of models for survival data are covered by the class.

An important aspect of the generalization is the presence in all the models of a *linear predictor* based on a linear combination of explanatory or stimulus variables. The variables may be continuous or categorical (or indeed a mixture of the two), and the existence of a linear predictor means that the concepts of classical regression and analysis-of-variance models, insofar as they refer to the estimation of parameters in a linear predictor, carry across directly to the wider class of model. In particular, the ideas underlying factorial models, including those of additivity, interaction, polynomial contrasts, aliasing, etc., all appear in the wider context.

Generalized linear models have a common algorithm for the estimation of parameters by maximum likelihood; this uses weighted least squares with an adjusted dependent variate, and does not require preliminary guesses to be made of the parameter values.

The book is aimed at applied statisticians and postgraduate students in statistics, but will be most useful, at least in part, to undergraduates and to numerate biologists. More mathematical sections are marked with asterisks and may be omitted at first reading. Some mathematics has been relegated to the first four appendices, while the fifth contains information on computer software for the fitting of generalized linear models. The book requires the reader to have a knowledge of matrix theory, including generalized inverses, together with basic ideas of probability theory, including orders of

magnitude in probability. As far as possible, however, the development is self-contained, though necessarily fairly condensed because of the constraints on a monograph in this series. Further reading is given in the bibliographic sections of various chapters, and the theory is illustrated with a diverse set of worked examples.

We are grateful to Professor J.V. Zidek of the University of British Columbia and to the Natural Sciences and Engineering Research Council, Canada, for the opportunity to undertake an intensive spell of writing. For permission to use previously unpublished data, we wish to thank Dr Graeme Newell, Lloyds Register of Shipping, and Drs P.M. Morse, K.S. McKinlay and D.T. Spurr. We are grateful to Miss Lilian Robertson for her careful preparation of the manuscript.

London and Harpenden
1983

P. McCullagh
J.A. Nelder

Preface

The subject of *generalized linear models* has undergone vigorous development in the six years since the publication of the first edition of this book. At the same time many of the key ideas, terminology, notation, and so on, have diffused into the statistical mainstream, so there is a need to make the basic material more digestible for advanced undergraduate and graduate students who have some familiarity with linear models. Our chief aims in preparing this second edition have been:

1. to bring the book up to date;
2. to provide a more balanced and extended account of the core material by including examples and exercises.

The book has therefore been extensively revised and enlarged to cover some of the developments of the past six years. For obvious reasons we have had to be selective in our choice of new topics. We have tried to include only those topics that might be directly useful to a research scientist. Within this category, though, our choice of topics reflects our own research interests including, in particular, quasi-likelihood functions and estimating equations, models for dispersion effects, components of dispersion (random-effects models), and conditional likelihoods.

The organization of the basic material in the first six chapters follows that of the first edition, though with greater emphasis on detail and more extensive discussion. Numerous exercises, both theoretical and data-analytic, have been added as a supplement to each chapter. These six chapters should provide sufficient material for a one-quarter introductory course on generalized linear models. The remaining chapters cover more advanced or specialized topics suitable for a second-level course.

We are indebted to a large number of readers who, over the past two years, have contributed to the proof-reading process:

A.C. Atkinson, L. Friedman, M.L. Frigge, E. Iversen, J. Kolassa, M.L. Lam, T.M. Redburg, I.M. Skovgaard, M. Stein, D.L. Wallace and W. Wong. We are especially grateful to D.R. Cox, A.C. Davison, M. Drum, D. Firth, G. Glonek, V.N. Nair, D. Pregibon, N. Reid, D.W. Schafer, S.M. Stigler, R. Tibshirani and S. Zeger for their constructive and detailed comments on a preliminary version of the manuscript.

We wish to thank S. Arnold, J. Streibig, P. Verrell and L. Vleeshouwers for permission to use previously unpublished data.

This edition has been typeset using \TeX . The book includes more than 40 figures and diagrams, which have been drawn using PiCTeX (Wichura, 1986).

Some of the research referred to in parts of this book has been supported in part by National Science Foundation grants over the past three years.

Finally, the efficient secretarial help of B. Brinton, S. Malkani and M. Nakatsuka is gratefully acknowledged.

*Chicago and Harpenden
April 1989*

P. McCullagh
J.A. Nelder

CHAPTER 1

Introduction

1.1 Background

In this book we consider a class of statistical models that is a natural generalization of classical linear models. *Generalized linear models* include as special cases, linear regression and analysis-of-variance models, logit and probit models for quantal responses, log-linear models and multinomial response models for counts and some commonly used models for survival data. It is shown that the above models share a number of properties, such as linearity, that can be exploited to good effect, and that there is a common method for computing parameter estimates. These common properties enable us to study generalized linear models as a single class, rather than as an unrelated collection of special topics.

Classical linear models and least squares began with the work of Gauss and Legendre (Stigler, 1981, 1986) who applied the method to astronomical data. Their data were usually measurements of continuous quantities such as the positions and magnitudes of the heavenly bodies and, at least in the astronomical investigations, the variability in the observations was largely the effect of measurement error. The Normal, or Gaussian, distribution was viewed as a mathematical construct developed to describe the properties of such errors; later in the nineteenth century the same distribution was used to describe the variation between individuals in a biological population in respect of a character such as height, an application quite different in kind from its use for describing measurement error, and leading to the numerous biological applications of linear models.

Gauss introduced the Normal distribution of errors as a device for describing variability, but he showed that many of the important properties of least-squares estimates depend not on Normality but on the assumptions of constant variance and indepen-

dence. A closely related property applies to all generalized linear models. In other words, although we make reference at various points to standard distributions such as the Normal, binomial, Poisson, exponential or gamma, the second-order properties of the parameter estimates are insensitive to the assumed distributional form: the second-order properties depend mainly on the assumed variance-to-mean relationship and on uncorrelatedness or independence. This is fortunate because, in applications, one can rarely be confident that all aspects of the assumed distributional form are correct.

Another strand in the history of statistics is the development of methods for dealing with discrete events rather than with continuously varying quantities. The enumeration of probabilities for various configurations in games of cards and dice was a matter of keen interest for gamblers in the eighteenth century. From their pioneering work grew methods for dealing with data in the form of counts of events. In the context of rare events, the basic distribution is that named after Poisson. This distribution has been applied to diverse kinds of events: a famous example concerns unfortunate soldiers kicked to death by Prussian horses (Bortkewitsch, 1898). The annual number of such incidents during the period 1875–1894 was observed to be consistent with the Poisson distribution having mean about 0.7 per corps per year. There is, however, some variation in this figure between corps and between years. Routine laboratory applications of the Poisson model include the monitoring of radioactive tracers by emission counts, counts of infective organisms as measured by the number of events observed on a slide under a microscope, and so on.

Closely related to the Poisson model are models for the analysis of counted data in the form of proportions or ratios of counts. The Bernoulli distribution is often suitable for modelling the presence or absence of disease in a patient, and the derived binomial distribution may be suitable as a model for the number of diseased patients in a fixed pool of patients under study. In medical and pharmaceutical trials it is usually required to study not primarily the incidence of a particular disease, but how the incidence is affected by factors such as age, social class, housing conditions, exposure to pollutants, and any treatment procedures under study. Generalized linear models permit us to study patterns of systematic variation in much the same way as ordinary linear models are used

to study the joint effects of treatments and covariates.

Some continuous measurements encountered in practice have non-Normal error distributions, and the class of generalized linear models includes distributions useful for the analysis of such data. The simplest examples are perhaps the exponential and gamma distributions, which are often useful for modelling positive data that have positively skewed distributions, such as occur in studies of survival times.

Before looking in more detail at the history of individual instances of generalized linear models, we make some general comments about statistical models and the part they play in the analysis of data, whether experimental or observational.

1.1.1 *The problem of looking at data*

Suppose we have a number of measurements or counts, together with some associated structural or contextual information, such as the order in which the data were collected, which measuring instruments were used, and other differences in the conditions under which the individual measurements were made. To interpret such data, we search for a pattern, for example that one measuring instrument has produced consistently higher readings than another. Such systematic effects are likely to be blurred by other variation of a more haphazard nature. The latter variation is usually described in statistical terms, no attempt being made to model or to predict the actual haphazard contribution to each observation.

Statistical models contain both elements, which we will call *systematic effects* and *random effects*. The value of a model is that often it suggests a simple summary of the data in terms of the major systematic effects together with a summary of the nature and magnitude of the unexplained or random variation. Such a reduction is certainly helpful, for the human mind, while it may be able to encompass say 10 numbers easily enough, finds 100 much more difficult, and will be quite defeated by 1000 unless some reducing process takes place.

Thus the problem of looking intelligently at data demands the formulation of patterns that are thought capable of describing succinctly not only the systematic variation in the data under study, but also for describing patterns in similar data that might

be collected by another investigator at another time and in another place.

1.1.2 *Theory as pattern*

We shall consider theories as generating patterns of numbers, which in some sense can replace the data, and can themselves be described in terms of a small number of quantities. These quantities are called *parameters*. By giving the parameters different values, specific patterns can be generated. Thus the very simple model

$$y = \alpha + \beta x,$$

connecting two quantities y and x via the parameter pair (α, β) , defines a straight-line relationship between y and x . Suppose now that there is some causal relationship between x and y in which x is under control and affects y , and that y can be measured (ideally) without error. Then if we give x the values

$$x_1, x_2, \dots, x_n,$$

y takes the values

$$\alpha + \beta x_1, \alpha + \beta x_2, \dots, \alpha + \beta x_n$$

for the assigned values α and β . Clearly, if we know α and β we can reconstruct the values of y exactly from those of x , so that given x_1, \dots, x_n , the pair (α, β) is an exact summary of y_1, \dots, y_n and we can move between the data and the parameters in either direction.

In practice, of course, we never measure the y s exactly, so that the relationship between y and x is only approximately linear. Despite this lack of exactness, we can still choose values of α and β , *a* and *b* say, that in some suitable sense best describe the now approximately linear relation between y and x . The quantities $a + bx_1, a + bx_2, \dots, a + bx_n$, which we denote by $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ or $\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_n$, are the *theoretical* or *fitted values* generated by the model and the data. They do not reproduce the original data values y_1, \dots, y_n exactly. The pattern that they represent approximates the data values and can be summarized by the pair (a, b) .

1.1.3 Model fitting

The fitting of a simple linear relationship between the y s and the x s requires us to choose from the set of all possible pairs of parameter values a particular pair (a, b) that makes the patterned set $\hat{y}_1, \dots, \hat{y}_n$ closest to the observed data. In order to make this statement precise we need a measure of ‘closeness’ or, alternatively, of distance or discrepancy between the observed y s and the fitted \hat{y} s. Examples of such discrepancy functions include the L_1 -norm

$$S_1(y, \hat{y}) = \sum |y_i - \hat{y}_i|$$

and the L_∞ -norm

$$S_\infty(y, \hat{y}) = \max_i |y_i - \hat{y}_i|.$$

Classical least squares, however, chooses the more convenient L_2 -norm or sum of squared deviations

$$S_2(y, \hat{y}) = \sum (y_i - \hat{y}_i)^2$$

as the measure of discrepancy. These discrepancy formulae have two implications. First, the straightforward summation of individual deviations, either $|y_i - \hat{y}_i|$ or $(y_i - \hat{y}_i)^2$, each depending on only one observation, implies that the observations are all made on the same physical scale and suggests that the observations are independent, or at least that they are in some sense exchangeable, so justifying an even-handed treatment of the components. Second, the use of arithmetic differences $y_i - \hat{y}_i$ implies that a given deviation carries the same weight irrespective of the value of \hat{y} . In statistical terminology, the appropriateness of L_p -norms as measures of discrepancy depends on stochastic independence and also on the assumption that the variance of each observation is independent of its mean value. Such assumptions, while common and often reasonable in practice, are by no means universally applicable.

The discrepancy functions just described can be justified in purely statistical terms. For instance, the classical least squares criterion arises if we regard the x -values as fixed or non-stochastic and the y -values are assumed to have the Normal, or Gaussian, distribution with mean μ , in which

$$\text{frequency of } y \text{ given } \mu \propto \exp\{-(y - \mu)^2/(2\sigma^2)\}, \quad (1.1)$$

where μ is linearly related to x through the coefficients α and β . The scale factor σ , which is the standard deviation of y , describes the ‘width’ of the errors when measured about the mean value. In older statistical texts, 0.67σ is sometimes called the *probable error* in y .

We can look at the function (1.1) in two ways. If we regard it as a function of y for fixed μ , the function specifies the probability density of the observations. On the other hand, for a given observation y , we may regard (1.1) as a function of μ giving the relative plausibility of different values of μ for the particular value of y observed. It was this second interpretation, known as the likelihood function, whose value was first stressed by R.A. Fisher. We notice that the quantity $-2l$, where l is the logarithm of the likelihood function for a sample of n independent values, is equal to

$$\frac{1}{\sigma^2} \sum (y_i - \mu_i)^2.$$

In other words, apart from the factor σ^2 , here assumed known, $-2l$ is identical to the sum-of-squares criterion. As μ varies, $-2l$ takes its minimum value at $\mu = \bar{y}$, the arithmetic mean of the observations. For a more complicated model in which μ varies in a systematic way from observation to observation, we define the closest set $\hat{\mu}$ to be that whose values maximize the likelihood or, equivalently, minimize $-2l$. More generally, we can extend our interest beyond the single point that minimizes $-2l$, to the shape of the likelihood surface in the neighbourhood of the minimum. This shape tells us, in Fisher’s terminology, how much information concerning the parameters there is in the data.

Appendix A gives a concise summary of the principal properties of likelihood functions.

Reverting to our example of a linear relationship, we can plot on a graph with axes α and β , the contours of equal discrepancy $-2l$ for the given data y . In this particular instance, $-2l$ is a quadratic function of (α, β) and hence the contours are ellipses, similar in shape and orientation, with the maximum-likelihood estimate $(\hat{\alpha}, \hat{\beta})$ situated at their centre. The information in the data on the parameters (α, β) is given by the curvature matrix or Hessian matrix of the quadratic. If the axes of the ellipses are not aligned with the (α, β) axes, the estimates are said to be correlated. The information is greatest in the direction for which

the curvature is greatest (see Fig. 3.8). In certain circumstances, the form of the information surface can be determined before an experiment is carried out. In other words, the precision achievable by a given experiment can sometimes be determined in advance and such information can be used to compute the experimental resources needed to estimate parameters with a required accuracy. A similar analysis will also show the parameter combinations that are badly estimated by the data and this information is often valuable in choosing among possible experimental designs. Alas, such calculations are not made nearly often enough!

1.1.4 *What is a good model?*

To summarize, we aim in model fitting to replace our data \mathbf{y} with a set of fitted values $\hat{\boldsymbol{\mu}}$ derived from a model. These fitted values are chosen to minimize some criterion such as the sum-of-squares discrepancy measure $\sum_i (y_i - \hat{\mu}_i)^2$.

At first sight it might seem as though a good model is one that fits the observed data very well, i.e. that makes $\hat{\boldsymbol{\mu}}$ very close to \mathbf{y} . However, by including a sufficient number of parameters in our model, we can make the fit as close as we please, and indeed by using as many parameters as observations we can make the fit perfect. In so doing, however, we have achieved no reduction in complexity – produced no simple theoretical pattern for the ragged data. Thus simplicity, represented by parsimony of parameters, is also a desirable feature of any model; we do not include parameters that we do not need. Not only does a parsimonious model enable the research worker or data analyst to think about his data, but one that is substantially correct gives better predictions than one that includes unnecessary extra parameters.

An important property of a model is its scope, i.e. the range of conditions over which it gives good predictions. Scope is hard to formalize, but easy to recognize, and intuitively it is clear that scope and parsimony are to some extent related. If a model is made to fit very closely to a particular set of data, it will not be able to encompass the inevitable changes that will be found necessary when another set of data relating to the same phenomenon is collected. Both scope and parsimony are related to *parameter invariance*, that is to parameter values that either do not change as some external condition changes or that change in a predictable way.

Modelling in science remains, partly at least, an art. Some principles do exist, however, to guide the modeller. A first, though at first sight, not a very helpful principle, is that all models are wrong; some, though, are more useful than others and we should seek those. At the same time we must recognize that eternal truth is not within our grasp. A second principle (which applies also to artists!) is not to fall in love with one model to the exclusion of alternatives. Data will often point with almost equal emphasis at several possible models and it is important that the statistician recognize and accept this. A third principle recommends thorough checks on the fit of a model to the data, for example by using residuals and other statistics derived from the fit to look for outlying observations and so on. Such diagnostic procedures are not yet fully formalized, and perhaps never will be. Some imagination or introspection is required here in order to determine the aspects of the model that are most important and most suspect. Box (1980) has attempted a formalization of the dual processes of model fitting and model criticism.

1.2 The origins of generalized linear models

1.2.1 Terminology

This section deals with the origin of generalized linear models, describing various special cases that are now included in the class in approximately their chronological order of development. First we need to establish some terminology: data will be represented by a **data matrix**, a two-dimensional array in which the rows are indexed by experimental or survey units. In this context, units are the physical items on which observations are made, for example plots in an agricultural field trial, patients in a medical survey or clinical trial, quadrats in an ecological study and so on. The columns of the data matrix are the **variates** such as measurements or yields, treatments, varieties, plot characteristics, patient's age, weight, sex and so on. Some of the variates are regarded as responses or dependent variates, whose values are believed to be affected by the explanatory variables or covariates. The latter are unfortunately sometimes called independent variates. Tukey (1962) uses the terms **response** and **stimulus** to make this important distinction. Covariates may be quantitative or qualitative. Quantitative

variates take on numerical values: qualitative variates take on non-numerical values or *levels* from a finite set of values or labels. We shall refer to qualitative covariates as *factors*: such covariates include classification variables such as blocks, that serve to group the experimental units, and treatment indicators that may in principle be assigned by the experimenter to any of the experimental units. Dependent variables may be continuous, or discrete (in the form of counts), or they may take the form of factors, where the response is one of a finite set of possible values or classes. For examples of the latter type of response, see Chapter 5.

1.2.2 Classical linear models

In matrix notation the set of observations is denoted by a column vector of observations $\mathbf{y} = \{y_1, \dots, y_n\}^T$. The set of covariates or explanatory variables is arranged as an $n \times p$ matrix \mathbf{X} . Each row of \mathbf{X} refers to a different unit or observation, and each column to a different covariate. Associated with each covariate is a coefficient or parameter, usually unknown. The set of parameters is a vector of dimension p , usually denoted by $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_p\}^T$. For any given value of $\boldsymbol{\beta}$, we can define a vector of residuals

$$\mathbf{e}(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}.$$

In 1805 Legendre first proposed estimating the β s by minimizing $\mathbf{e}^T \mathbf{e} = \sum_i e_i^2$ over values of $\boldsymbol{\beta}$. [Note that both Legendre and Gauss defined the residuals with opposite sign to that in current use, i.e. by $\mathbf{X}\boldsymbol{\beta} - \mathbf{y}$.] In 1809, in a text on astronomy, Gauss introduced the Normal distribution with zero mean and constant variance for the errors. Later in his *Theoria Combinationis* in 1823, he abandoned the Normal distribution in favour of the weaker assumption of constancy of variance alone. He showed that the estimates of $\boldsymbol{\beta}$ obtained by minimizing the least-squares criterion have minimum variance among the class of unbiased estimates. The extension of this weaker assumption to generalized linear models was given by Wedderburn (1974) using the concept of quasi-likelihood. This extension is discussed in Chapter 9.

Most astronomical data analysed using least squares were of the observational kind, i.e. they arose from observing a system, such as the Solar System, without perturbing it or experimenting with

it. The development of the theory of experimental design gave a new stimulus to linear models and is very much associated with R.A. Fisher and his co-workers.

1.2.3 R.A. Fisher and the design of experiments

In 1919, Fisher began work at the agricultural research station at Rothamsted. Within 10 years, he had, among other achievements, laid the foundations of the design of experiments, a subject that was substantially developed by his successor, F. Yates, and others at Rothamsted. In particular, Fisher stressed the value of factorial experiments in which several experimental and classification factors are studied simultaneously instead of being varied one at a time. Thus, with two factors under study, each having two levels, the one-at-a-time design (a) in Fig. 1.1 was replaced with the factorial design (b). In the latter case, all combinations of the two factors are studied.

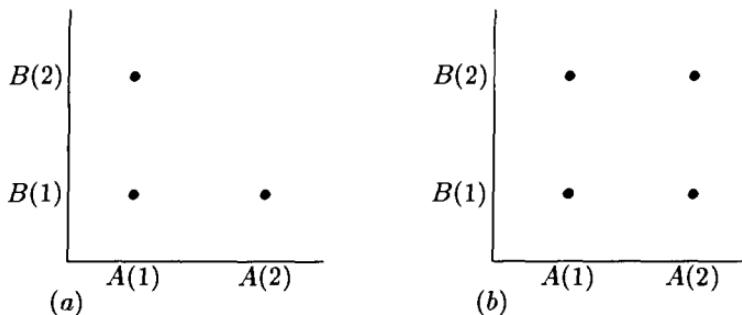


Fig. 1.1. (a) Design for two factors, changing levels one at a time; (b) factorial design.

The use of factorial designs increases the information per observation. Their analysis involves factorial models in which the yield or response is considered to be expressible as the sum of effects due to individual factors acting one at a time (main effects), effects due to pairs of factors above and beyond their separate contributions (two-factor interactions), and so on. Thus, the term 'factorial' refers to a particular class of design matrices or model matrices \mathbf{X} . In the case of factorial models, \mathbf{X} is a matrix of zeros and ones only

and is sometimes called an *incidence matrix* for that particular design. Factorial models are often called analysis-of-variance models to be distinguished and contrasted with linear regression models for which the covariates are continuous and not restricted to the values zero and one. We shall use the terms factorial design and linear regression model as descriptors for different kinds of model matrices \mathbf{X} . However, we shall try not to make a major distinction, but rather to unify the ideas embodied in these two extremes. For instance, we shall include terms in which the slopes defined by regression coefficients are allowed to vary with the level of various indexing factors.

Fisher's influence on the development of generalized linear models extends well beyond models for factorial experiments and includes special models for the analysis of certain kinds of counts and proportions. We now consider some of the non-classical cases of generalized linear models that arose in the period 1922–1960.

1.2.4 *Dilution assay*

The original paper in this context is Fisher (1922), especially section 12.3. A solution containing an infective organism is progressively diluted and, at each dilution, a number of agar plates are 'streaked'. On some of these plates the infective organism produces a growth on the medium: the rest of the plates remain sterile. From the number of sterile plates observed at each dilution, an estimate of the concentration of infective organisms in the original solution is made.

Assuming for simplicity that dilutions are made in powers of two, the argument runs as follows. After x dilutions, the number of infective organisms, ρ_x , per unit volume is

$$\rho_x = \rho_0 / 2^x, \quad x = 0, 1, \dots$$

where ρ_0 , which we wish to estimate, is the density of infective organisms in the original solution. Assuming that each agar plate is streaked using a volume, v , of solution, the expected number of organisms on any plate is $\rho_x v$ and, under suitable mixing conditions, the actual number of organisms follows the Poisson distribution with this parameter. Thus the probability that a plate is infected is just $\pi_x = 1 - \exp\{-\rho_x v\}$, the complement of the first

term in the Poisson series. It follows that at dilution x

$$\log(-\log(1 - \pi_x)) = \log v + \log \rho_x = \log v + \log \rho_0 - x \log 2. \quad (1.2)$$

If at dilution x we have r infected plates out of m , the observed proportion of infected plates $y = r/m$ may be regarded as the realization of a random variable Y satisfying

$$E(Y | x) = \pi_x.$$

However, this time it is not the mean of Y that bears a linear relationship to x , but instead the transformation

$$\eta = \log(-\log(1 - \pi_x))$$

known as the complementary log log transformation. To make the linear relationship explicit, we write

$$\eta = \alpha + \beta x,$$

where $\alpha = \log v + \log \rho_0$ and $\beta = -\log 2$.

In this example, we have a slope β that is known a priori, an intercept α that bears a simple relationship to the quantity ρ_0 that we wish to estimate, and it is not the mean of Y that is linear in x but a known function of $E(Y)$, in this case the complementary log log function. For this dilution assay problem, Fisher showed how to apply maximum likelihood to obtain an estimator. He also used his concept of information to show that another estimator, based solely on the number of sterile plates over all dilutions, contained 87.7% of the information of the maximum-likelihood estimator. Nowadays, we can use a computer to calculate the maximum-likelihood estimate with minimal effort: alternative simpler estimators may still retain a certain appeal, but computational effort is no longer an important criterion for selection. The model just described is a particular instance of a generalized linear model. Fisher's estimation procedure is an early non-trivial application of maximum likelihood to a problem for which no closed-form solution exists.

1.2.5 Probit analysis

The technique known as probit analysis arose in connection with bioassay, and the modern method of analysis dates from Bliss (1935). In toxicology experiments, for example, test animals or insects are divided into sets, usually, but not necessarily of equal sizes. Each set of animals is subjected to a known level x of a toxin, or in other contexts, of a stimulant or dose. The dose varies from set to set but is assumed to be uniform within each set. For the j th set, the number y_j surviving out of the original m_j is recorded, together with the dose x_j administered. It is required to model the proportion surviving, π_x , at dose x as a function of x , which is usually measured in logarithmic units. The probit model is

$$\pi_x = \Phi(\alpha + \beta x), \quad (1.3)$$

where $\Phi(\cdot)$ is the cumulative Normal distribution function, and α and β are unknown parameters to be estimated. This model has the virtue that it respects the property that π_x is a probability and hence must lie between 0 and 1 for all values of x and for all parameter values. For this reason alone, it is not normally sensible to contemplate linear models for probabilities. Note also that if $\beta > 0$, the survival probability is monotonely increasing in the applied dose: otherwise, if $\beta < 0$, the survival probability is monotonely decreasing in the dose.

Because of the occurrence of $y_j = 0$ or $y_j = m_j$ at certain high or low doses, it is not feasible to take $\Phi^{-1}(y_j/m_j)$ as the response variable in order to make the model approximately linear in the parameters. Infinite values can be avoided by using a modified empirical transformation such as $\Phi^{-1}\{(y_j + \frac{1}{2})/(m_j + 1)\}$, but the choice of modification is to a large extent arbitrary.

Linearity in the parameters is an important aspect of the probit model (1.3). Note however, that the linearity does not occur directly in the expression for $E(Y)$ in terms of x nor in $E\{\Phi^{-1}(Y/m)\}$ as a function of x . The linearity in question arises in the expression for $\Phi^{-1}(\pi_x)$, the transformed theoretical proportion surviving at dose x . This is the same sense in which the model for the dilution assay (1.2), is linear, although the transformations required to achieve linearity are different in the two examples.

The probit model exhibits one further feature that distinguishes it from the usual Normal-theory model, namely that the variance of

the observed proportion surviving Y/m , is not constant but varies in a systematic fashion as a function of $\pi = E(Y/m)$. Specifically, under the usual binomial assumption, Y/m has variance $\pi(1 - \pi)/m$, which has a maximum at $\pi = 0.5$. Generalized linear models accommodate unequal variances through the introduction of variance functions that may depend on the mean value through a known function of the mean.

1.2.6 Logit models for proportions

Dyke and Patterson (1952) published an analysis of some cross-classified survey data concerning the proportion of subjects who have a good knowledge of cancer. The recorded explanatory variables were exposures to various information sources, newspapers, radio, solid reading, lectures. All combinations of these explanatory variables occurred in the sample, though some combinations occurred much more frequently than others. A factorial model was postulated in which the logit or log odds of success, $\log\{\pi/(1-\pi)\}$ is expressed linearly as a combination of the four information sources and interactions among them. Success in this context is interpreted as synonymous with 'good knowledge of cancer'. Dyke and Patterson were successful in finding a suitable model of this kind, though the fitting, which was done manually, took several days. Similar computations done today take only a few seconds.

Dyke and Patterson's application of the linear logistic model was to survey data. Linear logistic models had earlier been used in the context of bioassay experiments (see, for example, Berkson, 1944, 1951).

1.2.7 Log-linear models for counts

The analysis of counted data has recently given rise to an extensive literature mainly based on the idea of a log-linear model. In such a model, the two components of the classical linear model are replaced in the following way:

	<i>Classical linear model</i>	<i>Log-linear model</i>
<i>Systematic effects</i>	additive	multiplicative
<i>Nominal error distribution</i>	Normal	Poisson

The Poisson distribution is the nominal distribution for counted data in much the same way that the Normal distribution is the bench-mark for continuous data. Such counts are assumed to take the values $0, 1, 2, \dots$ without an upper limit. The Poisson distribution has only one adjustable parameter, namely the mean μ , which must be positive. Thus the mean alone determines the distribution entirely. By contrast, the Normal distribution has two adjustable parameters, namely the mean and variance, so that the mean alone does not determine the distribution completely.

Since the Poisson mean is required to be positive, an additive model for μ is normally considered to be unsatisfactory. All linear combinations $\eta = \sum \beta_j x_j$ become negative for certain parameter combinations and covariate combinations. Hence, although $\mu = \sum \beta_j x_j$ may be found to be adequate over the range of the data, it is often scientifically dubious and logically unsatisfactory for extrapolation. In the model with multiplicative effects, we set $\mu = \exp(\eta)$ and η rather than μ obeys the linear model. This construction ensures that μ remains positive for all η and hence positive for all parameter and covariate combinations.

The ideas taken from factorial design and regression models carry over directly to log-linear models except that the effects or parameters of interest are contrasts of log frequencies. For the purpose of explanation and exposition, such contrasts are usually best back-transformed to the original frequency scale and expressed as multiplicative effects.

It often happens with counted data that one of the classifying variables, rather than the counts themselves, is best regarded as the response. In this case, the aim usually is to model the way in which the remaining explanatory variables affect the relative proportions falling in the various categories of response. Normally, we would not aim to model the total numbers of respondents as a function of the response variables, but only the way in which these respondents are distributed across the k response categories. In this context, it is natural to consider modelling the errors by the multinomial distribution, which can be regarded as a set of k independent Poisson random variables subject to the constraint that their total is fixed. The relationship between Poisson log-linear models and multinomial response models is discussed further in section 6.4. It is possible, though not always desirable, to handle multinomial response models by using a suitably augmented log-

linear model.

Routine use of log-linear models has had a major impact on the analysis of counted data, particularly in the social sciences. Both log-linear and multinomial response models are special cases of generalized linear models and are discussed further in Chapters 4 to 6.

1.2.8 *Inverse polynomials*

Polynomials are widely used in biological and other work for expressing the shape of response curves, growth curves and so on. The most obvious advantage of polynomials is that they provide an infinite sequence of easily-fitted curves. The main disadvantage is that in most scientific work, the response is bounded, whereas polynomials, when extrapolated, become unbounded. Moreover, responses are often required to be positive, whereas polynomials are liable to become negative in certain ranges. In many applications, for example in the context of growth curves, it is common to find that the response approaches a plateau or asymptote as the stimulus increases. Polynomials do not have asymptotes and hence cannot be consistent with this known form of limiting behaviour.

Hyperbolic response curves of the form

$$x/y = \alpha + \beta x,$$

which do have asymptotes, have been used in a number of contexts such as the Michaelis–Menten equations of enzyme kinetics. The inverse polynomials introduced by Nelder (1966) extend this class of response curve to include inverse quadratic and higher-order inverse polynomial terms. More than one covariate can be included. Details are discussed in Chapter 8, which deals also with the case of continuous response variables in which the coefficient of variation rather than the variance is assumed constant over all observations.

1.2.9 *Survival data*

In the past 15 years or so, great interest has developed in models for survival in the context of clinical and surgical treatments. Similar problems, though with a rather different emphasis, occur in the analysis of failure times of manufactured components. In

medical experiments particularly, the data usually contain censored individuals. Such individuals are known to have survived up to a given time but their subsequent progress is not recorded either because the trial ends before the outcome is known or because the patient can no longer be contacted. In medical trials, such patients are said to be censored or 'lost to follow-up'. Aitkin and Clayton (1980) and Whitehead (1980) have shown how the analysis of censored survival data can be moulded into the framework of generalized linear models. This transformation is simplest to achieve when there are no time-dependent covariates: in more complicated cases, the computations are best done with the assistance of specially written computer programs.

1.3 Scope of the rest of the book

In Chapter 2, we outline the component processes in model fitting, describe the components of a generalized linear model, the definitions of goodness-of-fit of a model to data, a method for fitting generalized linear models and some asymptotic theory concerning the statistical properties of the parameter estimates. Chapter 3 deals with classical models for continuous data, in which the systematic effects are described by a linear model and the error variances are assumed constant and independent of the mean response. Many of the ideas introduced in this classical context carry over with little or no change to the whole class of generalized linear models. In particular, descriptive terms and model formulae that are used to specify design or model matrices are equally appropriate for all generalized linear models. The three subsequent chapters describe models that are relevant for data in the form of counts or proportions. Random variation in this context is often suitably described by the Poisson, binomial or multinomial distributions: systematic effects are assumed to be additive on a suitably chosen scale. The scale is chosen in such a way that the fitted frequencies are positive and the fitted proportions lie between 0 and 1. Where response categories are ordered, models are chosen that respect this order. Chapter 8 introduces generalized linear models for continuous data where, instead of assuming that the variance is constant, it is assumed instead that the coefficient of variation, σ/μ , is constant. In other words, the larger the mean response, the

greater the variability in the response. Examples are drawn from meteorology and the insurance industry.

A major extension of the applicability of generalized linear models was made by Wedderburn (1974) when he introduced the idea of quasi-likelihood. Wedderburn showed that often it is not necessary to make specific detailed assumptions regarding the random variation. Instead, many of the more useful properties of parameter estimates, derived initially from likelihood theory, can be justified on the grounds of weaker assumptions concerning independence and second moments alone. Specifically, it is necessary to know how the variance of each observation changes with its mean value but it is not necessary to specify the distribution in its entirety. Models based on quasi-likelihood are introduced informally, where appropriate, in earlier chapters, while in Chapter 9, a more systematic account is given.

Medical research is much concerned with the analysis of survival times of individual patients. Different patients have different histories and are assigned to one of several treatments. It is required to know how the survival time is affected by the treatment given, making such allowance as may be required for the differing histories of the various patients. There is a close connection between the analysis of survival times and the analysis of, say, 5-year survival rates. The latter problem falls under the rubric of discrete data or binary data. Such connections are exploited in Chapter 13 in order to handle survival times in the context of generalized linear models.

Frequently it happens that a model would fall into the linear category if one or two parameters that enter the model in a non-linear way were known *a priori*. Such models are sometimes said to be *conditionally linear*. A number of extensions to conditionally linear models are discussed in Chapter 11.

Chapter 10 discusses the simultaneous modelling of the mean and dispersion parameters as functions of the covariates, which are typically process settings in an industrial context.

Chapter 14 gives a brief introduction to problems in which there are several variance components, or dispersion components, associated with various sub-groups or populations. In this context it is usually unrealistic to assume that the observations are all independent.

1.4 Bibliographic notes

The historical development of linear models and least squares from Gauss and Legendre to Fisher has previously been sketched. For further historical information concerning the development of probability and statistics up to the beginning of the twentieth century, see the book by Stigler (1986).

The term ‘generalized linear model’ is due to Nelder and Wedderburn (1972), who showed how linearity could be exploited to unify apparently diverse statistical techniques.

For an elementary introduction to the subject, see the book by Dobson (1983).

1.5 Further results and exercises 1

1.1 Suppose that Y_1, \dots, Y_n are independent and satisfy the linear model

$$\mu_i = E(Y_i) = \sum_{j=1}^p x_{ij}\beta_j$$

for given covariates x_{ij} and unknown parameters β . Show that if Y_i has the Laplace distribution or double exponential distribution

$$f_{Y_i}(y_i; \mu_i, \sigma) = \frac{1}{2\sigma} \exp\{-|y_i - \mu_i|/\sigma\}$$

then the maximum-likelihood estimate of β is obtained by minimizing the L_1 -norm

$$S_1(y, \hat{y}) = \sum |y_i - \hat{y}_i|$$

over values of \hat{y} satisfying the linear model.

1.2 In the notation of the previous exercise, show that if Y_i is uniformly distributed over the range $\mu_i \pm \sigma$, maximum-likelihood estimates are obtained by minimizing the L_∞ -norm,

$$S_\infty(y, \hat{y}) = \max_i |y_i - \hat{y}_i|.$$

Show also that linearity of the model is irrelevant for the conclusions in both cases.

1.3 Justify the conclusion of the previous two exercises that the estimates of the regression parameters are unaffected by the value of σ in both cases. Show that the conclusion does not extend to either of the following distributions even though, in both cases, σ is a scale factor.

$$f_Y(y; \mu, \sigma) = \frac{\exp\{(y - \mu)/\sigma\}}{\sigma \{1 + \exp\{(y - \mu)/\sigma\}\}^2}$$

$$f_Y(y; \mu, \sigma) = \frac{1}{\pi \sigma \{1 + (y - \mu)^2/\sigma^2\}}$$

1.4 Find the maximum-likelihood estimate of σ for each model. Show that, for the models in Exercises 1.1 and 1.2, $\hat{\sigma}$ is a function of the minimized norm.

1.5 Suppose that X_1, X_2 are independent unit exponential random variables. Show that the distribution of $Y = \log(X_1/X_2)$ is

$$f_Y(y) = \frac{\exp(y)}{(1 + \exp(y))^2}$$

for $-\infty < y < \infty$.

Find the distribution of Y if the X s have the Weibull density

$$f_X(x) = \tau \rho (\rho x)^{\tau-1} \exp\{-(\rho x)^\tau\}, \quad \rho, \tau, x > 0.$$

[Hint: first find the distribution of $(\rho X)^\tau$.]

1.6 The probable error, τ , of a random variable Y may be defined by

$$\text{pr}(|Y - M| \geq \tau) = 0.5,$$

where M is the median of Y . Find the probable errors of

1. the exponential distribution;
2. the double exponential distribution (Exercise 1.1);
3. the logistic distribution (Exercise 1.3);
4. the Cauchy distribution (Exercise 1.3);
5. the Normal distribution.

Discuss briefly the differences between the probable error and the inter-quartile range.

The historical definition of probable error appears to be vague. Some authors take M to be the mean; others take τ to be a multiple (0.67) of the standard deviation.

CHAPTER 2

An outline of generalized linear models

2.1 Processes in model fitting

In Chapter 1 we considered briefly some of the reasons for model fitting as an aid for interpreting data. Before describing the form of generalized linear models (GLMs) we look first at the processes of model fitting, following closely the ideas of Box and Jenkins (1976), which they applied to time series. Three processes are distinguished: (i) model selection, (ii) parameter estimation and (iii) prediction of future values. Box and Jenkins use ‘model identification’ in place of our ‘model selection’, but we prefer to avoid any implication that a correct model can ever be known with certainty. In distinguishing these three processes, we do not assume that an analysis consists of the successive application of each just once. In practice there are backward steps, false assumptions that have to be retracted, and so on.

We now look briefly at some of the ideas associated with each of the three processes.

2.1.1 *Model selection*

Models that we select to fit to data are usually chosen from a particular class and, if the model-fitting process is to be useful, this class must be broadly relevant to the kind of data under study. An important characteristic of generalized linear models is that they assume independent (or at least uncorrelated) observations. More generally, the observations may be independent in blocks of fixed known sizes. As a consequence, data exhibiting the autocorrelations of time series and spatial processes are expressly excluded. This assumption of independence is characteristic of the

linear models of classical regression analysis, and is carried over without modification to the wider class of generalized linear models. In Chapter 9 we look at the possibility of relaxing this assumption. A second assumption about the error structure is that there is a single error term in the model. This constraint excludes, for instance, models for the analysis of experiments having more than one explicit error term. Perhaps the simplest instance of a model excluded by this criterion is the standard linear model for the split-plot design, which has two error terms, one for between-whole-plot variance and one for within-whole-plot variance. Again, we shall later relax this restriction for certain kinds of GLMs.

In practice, these two restrictions on the form of the error distribution are less restrictive than they might appear at first sight. For instance autoregressive models can easily be fitted using programmes designed expressly for ordinary linear models. Further, certain forms of dependence, such as that occurring in the analysis of contingency tables where a certain marginal total is fixed, can in fact be handled as if the observations were independent. Similarly, though a grouping factor corresponding to a nuisance classification may induce correlations within groups, a within-groups analysis after elimination of the effects of that nuisance factor can proceed as if the observations were independent.

The choice of scale for analysis is an important aspect of model selection. A common choice is between an analysis of Y , i.e. the original scale, or $\log Y$. To the question ‘What characterizes a “good” scale?’ we must answer that it all depends on the purpose for which the scale is to be used. To quote from the preface to the first edition in Jeffreys (1961): ‘It is sometimes considered a paradox that the answer depends not only on the observations but on the question; it should be a platitude’. In classical linear regression analysis a good scale should combine constancy of variance, approximate Normality of errors and additivity of systematic effects. Now there is usually no *a priori* reason to believe that such a scale exists, and it is not difficult to imagine cases in which it does not. For instance, in the analysis of discrete data where the errors are well approximated by the Poisson distribution, the systematic effects are often multiplicative. Here $Y^{1/2}$ gives approximate constancy of variance, $Y^{2/3}$ does better for approximate symmetry or Normality, and $\log Y$ produces additivity of the systematic effects. Evidently, no single scale will simultaneously produce all the desired proper-

ties.

With the introduction of generalized linear models, scaling problems are greatly reduced. Normality and constancy of variance are no longer required, although the way in which the variance depends on the mean must be known. Additivity of effects, while still an important component of all generalized linear models, can be specified to hold on a transformed scale if necessary. In generalized linear models, additivity is, correctly, postulated as a property of the expected responses. Additivity with respect to the data themselves can only ever be a rough approximation.

There remains the problem in model selection of the choice of x -variables (or covariates as we shall call them) to be included in the systematic part of the model. There is a large literature on this topic in linear models. In its simplest form, we are given a number of candidate covariates, x_1, \dots, x_p , and are required to find a subset of these that is in some sense best for constructing the fitted values

$$\hat{\mu} = \sum \mathbf{x}_j \hat{\beta}_j.$$

Implicit in the strategies that have been proposed is that there is a balance to be struck between improving the fit to the observed data by adding an extra term to the model and the usually undesirable increase in complexity implicit in this extra term. Note that even if we could define exactly what is meant by an optimum model in a given context, it is most unlikely that the data would indicate a clear winner among the potentially large number of competing models. We must anticipate that, clustered around the 'best' model will be a set of alternatives almost as good and not statistically distinguishable. Selection of covariates is discussed at various points in the chapters that follow, particularly in section 3.9 and in the various examples.

2.1.2 *Estimation*

Having selected a particular model, it is required to estimate the parameters and to assess the precision of the estimates. In the case of generalized linear models, estimation proceeds by defining a measure of goodness of fit between the observed data and the fitted values generated by the model. The parameter estimates are the values that minimize the goodness-of-fit criterion. We shall

be concerned primarily with estimates obtained by maximizing the likelihood or log likelihood of the parameters for the data observed. If $f(y; \theta)$ is the density function or probability distribution for the observation y given the parameter θ , then the log likelihood, expressed as a function of the mean-value parameter, $\mu = E(Y)$, is just

$$l(\mu; y) = \log f(y; \theta).$$

The log likelihood based on a set of independent observations y_1, \dots, y_n is just the sum of the individual contributions, so that

$$l(\boldsymbol{\mu}; \mathbf{y}) = \sum_i \log f_i(y_i; \theta_i)$$

where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$. Note that the density function $f(y; \theta)$ is considered as a function of y for fixed θ whereas the log likelihood is considered primarily as a function of θ for the particular data y observed. Hence the reversal of the order of the arguments.

There are advantages in using as the goodness-of-fit criterion, not the log likelihood $l(\boldsymbol{\mu}; \mathbf{y})$ but a particular linear function, namely

$$D^*(\mathbf{y}; \boldsymbol{\mu}) = 2l(\mathbf{y}; \boldsymbol{\mu}) - 2l(\boldsymbol{\mu}; \mathbf{y}),$$

which we call the *scaled deviance*. Note that, for the exponential-family models considered here, $l(\mathbf{y}; \mathbf{y})$ is the maximum likelihood achievable for an exact fit in which the fitted values are equal to the observed data. Because $l(\mathbf{y}; \mathbf{y})$ does not depend on the parameters, maximizing $l(\boldsymbol{\mu}; \mathbf{y})$ is equivalent to minimizing $D^*(\mathbf{y}; \boldsymbol{\mu})$ with respect to $\boldsymbol{\mu}$, subject to the constraints imposed by the model.

For Normal-theory linear regression models with known variance σ^2 , we have for a single observation

$$f(y; \mu) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right),$$

so that the log likelihood is

$$l(\mu; y) = -\frac{1}{2} \log(2\pi\sigma^2) - (y - \mu)^2/(2\sigma^2).$$

Setting $\mu = y$ gives the maximum achievable log likelihood, namely

$$l(y; y) = -\frac{1}{2} \log(2\pi\sigma^2),$$

so that the scaled deviance function is

$$D^*(y; \mu) = 2\{l(y; y) - l(\mu; y)\} = (y - \mu)^2/\sigma^2.$$

Apart, therefore, from the known factor σ^2 , the deviance in this instance is identical to the residual sum of squares and minimum deviance is synonymous with least squares.

2.1.3 *Prediction*

Prediction, as interpreted here, is concerned with answers to ‘what-if’ questions of the kind that may be posed following a statistical analysis. In the context of a time series such a question might take the form ‘what is the predicted value of the response at time t in the future, given the past history of the series and the model used in the analysis?’. More generally, prediction is concerned with statements about the likely values of unobserved events, not necessarily those in the future. For example, following an analysis of the incidence of heart disease nationally, the data being classified by region and age-group, a typical ‘what-if’ question is ‘what would be the predicted incidence for a particular city if it had the same age structure as the country as a whole?’. This kind of prediction is an instance of standardization. For another example, consider a quantal response assay in which we measure the proportion of subjects responding to a range of dose levels. We fit a model expressing how this proportion varies with dose, and from the fitted model we predict the dose that gives rise to a 50% response rate, the so-called LD50. This answers the question ‘what would be the predicted dose if the response rate were 50%?’ The word *calibration* is often used here to distinguish inverse prediction problems, in which the response is fixed and we are required to make statements about the likely values of x , from the more usual type in which the roles are reversed.

To be useful, predicted quantities need to be accompanied by measures of precision. These are ordinarily calculated on the assumption that the set-up that produced the data remains constant, and that the model used in the analysis is substantially correct. For an account of prediction as a unifying idea connecting the analysis of covariance and various kinds of standardization, see Lane and Nelder (1982).

2.2 The components of a generalized linear model

Generalized linear models are an extension of classical linear models, so that the latter form a suitable starting point for discussion. A vector of observations \mathbf{y} having n components is assumed to be a realization of a random variable \mathbf{Y} whose components are independently distributed with means $\boldsymbol{\mu}$. The systematic part of the model is a specification for the vector $\boldsymbol{\mu}$ in terms of a small number of unknown parameters β_1, \dots, β_p . In the case of ordinary linear models, this specification takes the form

$$\boldsymbol{\mu} = \sum_1^p \mathbf{x}_j \beta_j, \quad (2.1)$$

where the β s are parameters whose values are usually unknown and have to be estimated from the data. If we let i index the observations then the systematic part of the model may be written

$$E(Y_i) = \mu_i = \sum_1^p x_{ij} \beta_j; \quad i = 1, \dots, n, \quad (2.2)$$

where x_{ij} is the value of the j th covariate for observation i . In matrix notation (where $\boldsymbol{\mu}$ is $n \times 1$, \mathbf{X} is $n \times p$ and $\boldsymbol{\beta}$ is $p \times 1$) we may write

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

where \mathbf{X} is the model matrix and $\boldsymbol{\beta}$ is the vector of parameters. This completes the specification of the systematic part of the model.

For the random part we assume independence and constant variance of errors. These assumptions are strong and need checking, as far as is possible, from the data themselves. We shall consider techniques for doing this in Chapter 12. Similarly, the structure of the systematic part assumes that we know the covariates that influence the mean and can measure them effectively without error; this assumption also needs checking, as far as is possible.

A further specialization of the model involves the stronger assumption that the errors follow a Gaussian or Normal distribution with constant variance σ^2 .

We may thus summarize the classical linear model in the form:

The components of \mathbf{Y} are independent Normal variables with constant variance σ^2 and

$$\mathbf{E}(\mathbf{Y}) = \boldsymbol{\mu} \quad \text{where} \quad \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}. \quad (2.3)$$

2.2.1 The generalization

To simplify the transition to generalized linear models, we shall rearrange (2.3) slightly to produce the following three-part specification:

1. The *random component*: the components of \mathbf{Y} have independent Normal distributions with $E(\mathbf{Y}) = \boldsymbol{\mu}$ and constant variance σ^2 ;
2. The *systematic component*: covariates $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$ produce a *linear predictor* $\boldsymbol{\eta}$ given by

$$\boldsymbol{\eta} = \sum_1^p \mathbf{x}_j \beta_j;$$

3. The *link* between the random and systematic components:

$$\boldsymbol{\mu} = \boldsymbol{\eta}.$$

This generalization introduces a new symbol η for the linear predictor and the third component then specifies that μ and η are in fact identical. If we write

$$\eta_i = g(\mu_i),$$

then $g(\cdot)$ will be called the *link function*. In this formulation, classical linear models have a Normal (or Gaussian) distribution in component 1 and the identity function for the link in component 3. Generalized linear models allow two extensions; first the distribution in component 1 may come from an exponential family other than the Normal, and secondly the link function in component 3 may become any monotonic differentiable function.

We look first at the extended distributional assumption.

2.2.2 Likelihood functions for generalized linear models

We assume that each component of \mathbf{Y} has a distribution in the exponential family, taking the form

$$f_Y(y; \theta, \phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y, \phi)\} \quad (2.4)$$

for some specific functions $a(\cdot), b(\cdot)$ and $c(\cdot)$. If ϕ is known, this is an exponential-family model with canonical parameter θ . It may or may not be a two-parameter exponential family if ϕ is unknown. Thus for the Normal distribution

$$\begin{aligned} f_Y(y; \theta, \phi) &= \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp\{-(y - \mu)^2/2\sigma^2\} \\ &= \exp\{(y\mu - \mu^2/2)/\sigma^2 - \frac{1}{2}(y^2/\sigma^2 + \log(2\pi\sigma^2))\}, \end{aligned}$$

so that $\theta = \mu$, $\phi = \sigma^2$, and

$$a(\phi) = \phi, \quad b(\theta) = \theta^2/2, \quad c(y, \phi) = -\frac{1}{2}\{y^2/\sigma^2 + \log(2\pi\sigma^2)\}.$$

We write $l(\theta, \phi; y) = \log f_Y(y; \theta, \phi)$ for the log-likelihood function considered as a function of θ and ϕ , y being given. The mean and variance of Y can be derived easily from the well known relations

$$E\left(\frac{\partial l}{\partial \theta}\right) = 0 \quad (2.5)$$

and

$$E\left(\frac{\partial^2 l}{\partial \theta^2}\right) + E\left(\frac{\partial l}{\partial \theta}\right)^2 = 0. \quad (2.6)$$

We have from (2.4) that

$$l(\theta; y) = \{y\theta - b(\theta)\}/a(\phi) + c(y, \phi),$$

whence

$$\frac{\partial l}{\partial \theta} = \{y - b'(\theta)\}/a(\phi) \quad (2.7)$$

and

$$\frac{\partial^2 l}{\partial \theta^2} = -b''(\theta)/a(\phi), \quad (2.8)$$

where primes denote differentiation with respect to θ .

From (2.5) and (2.7) we have

$$0 = E\left(\frac{\partial l}{\partial \theta}\right) = \{\mu - b'(\theta)\}/a(\phi),$$

so that

$$E(Y) = \mu = b'(\theta).$$

Similarly from (2.6), (2.7) and (2.8) we have

$$0 = -\frac{b''(\theta)}{a(\phi)} + \frac{\text{var}(Y)}{a^2(\phi)},$$

so that

$$\text{var}(Y) = b''(\theta)a(\phi).$$

Thus the variance of Y is the product of two functions; one, $b''(\theta)$, depends on the canonical parameter (and hence on the mean) only and will be called the *variance function*, while the other is independent of θ and depends only on ϕ . The variance function considered as a function of μ will be written $V(\mu)$.

The function $a(\phi)$ is commonly of the form

$$a(\phi) = \phi/w,$$

where ϕ , also denoted by σ^2 and called the *dispersion parameter*, is constant over observations, and w is a known *prior weight* that varies from observation to observation. Thus for a Normal model in which each observation is the mean of m independent readings we have

$$a(\phi) = \sigma^2/m,$$

so that $w = m$.

The most important distributions of the form (2.4) with which we shall be concerned are summarized in Table 2.1.

Table 2.1 Characteristics of some common univariate distributions in the exponential family[†]

	Normal	Poisson	Binomial	Gamma	Inverse Gaussian
Notation	$N(\mu, \sigma^2)$	$P(\mu)$	$B(m, \pi)/m$	$G(\mu, \nu)$	$IG(\mu, \sigma^2)$
Range of y	$(-\infty, \infty)$	$0(1)\infty$	$\frac{0(1)m}{m}$	$(0, \infty)$	$(0, \infty)$
Dispersion parameter: ϕ	$\phi = \sigma^2$	1	$1/m$	$\phi = \nu^{-1}$	$\phi = \sigma^2$
Cumulant function: $b(\theta)$	$\theta^2/2$	$\exp(\theta)$	$\log(1 + e^\theta)$	$-\log(-\theta)$	$-(-2\theta)^{1/2}$
$c(y; \phi)$	$-\frac{1}{2} \left(\frac{y^2}{\phi} + \log(2\pi\phi) \right)$	$-\log y!$	$\log \left(\frac{m}{my} \right)$ $= \frac{\nu \log(\nu y) - \log y}{-\log \Gamma(\nu)}$	$-\frac{1}{2} \left\{ \log(2\pi\phi y^3) + \frac{1}{\phi y} \right\}$	
$\mu(\theta) = E(Y; \theta)$	θ	$\exp(\theta)$	$e^\theta/(1 + e^\theta)$	$-1/\theta$	$(-2\theta)^{-1/2}$
Canonical link: $\theta(\mu)$	identity	log	logit	reciprocal	$1/\mu^2$
Variance function: $V(\mu)$	1	μ	$\mu(1 - \mu)$	μ^2	μ^3

[†]The mean-value parameter is denoted by μ , or by π for the binomial distribution.

The parameterization of the gamma distribution is such that its variance is μ^2/ν .

The canonical parameter, denoted by θ , is defined by (2.4). The relationship between μ and θ is given in lines 6 and 7 of the Table.

2.2.3 Link functions

The link function relates the linear predictor η to the expected value μ of a datum y . In classical linear models the mean and the linear predictor are identical, and the identity link is plausible in that both η and μ can take any value on the real line. However, when we are dealing with counts and the distribution is Poisson, we must have $\mu > 0$, so that the identity link is less attractive, in part because η may be negative while μ must not be. Models for counts based on independence in cross-classified data lead naturally to multiplicative effects, and this is expressed by the log link, $\eta = \log \mu$, with its inverse $\mu = e^\eta$. Now additive effects contributing to η become multiplicative effects contributing to μ and μ is necessarily positive.

For the binomial distribution we have $0 < \mu < 1$ and a link should satisfy the condition that it maps the interval $(0, 1)$ on to the whole real line. We shall consider three principal functions in subsequent chapters, namely:

1. *logit*

$$\eta = \log\{\mu/(1 - \mu)\};$$

2. *probit*

$$\eta = \Phi^{-1}(\mu);$$

where $\Phi(\cdot)$ is the Normal cumulative distribution function;

3. *complementary log-log*

$$\eta = \log\{-\log(1 - \mu)\}.$$

The power family of links is important, at least for observations with positive mean. This family can be specified either by

$$\eta = (\mu^\lambda - 1)/\lambda \tag{2.9a}$$

with the limiting value

$$\eta = \log \mu; \quad \text{as } \lambda \rightarrow 0, \tag{2.9b}$$

or by

$$\eta = \begin{cases} \mu^\lambda; & \lambda \neq 0, \\ \log \mu; & \lambda = 0. \end{cases} \tag{2.10}$$

The first form has the advantage of a smooth transition as λ passes through zero, but with either form special action has to be taken in any computation with $\lambda = 0$.

2.2.4 Sufficient statistics and canonical links

Each of the distributions in Table 2.1 has a special link function for which there exists a sufficient statistic equal in dimension to β in the linear predictor $\eta = \sum x_j \beta_j$. These *canonical links*, as they will be called, occur when

$$\theta = \eta,$$

where θ is the canonical parameter as defined in (2.4) and shown in Table 2.1. The canonical links for the distributions in that table are thus:

Normal	$\eta = \mu,$
Poisson	$\eta = \log \mu,$
binomial	$\eta = \log\{\pi/(1 - \pi)\},$
gamma	$\eta = \mu^{-1},$
inverse Gaussian	$\eta = \mu^{-2}.$

For the canonical links, the sufficient statistic is $\mathbf{X}^T \mathbf{Y}$ in vector notation, with components

$$\sum_i x_{ij} Y_i, \quad j = 1, \dots, p,$$

summation being over the units. Note however, that, although the canonical links lead to desirable statistical properties of the model, particularly in small samples, there is in general no a priori reason why the systematic effects in a model should be additive on the scale given by that link. It is convenient if they are, but convenience alone must not replace quality of fit as a model selection criterion. In later chapters we shall deal with several models in which non-canonical links are used. We shall find, however, that the canonical links are often eminently sensible on scientific grounds.

2.3 Measuring the goodness of fit

2.3.1 The discrepancy of a fit

The process of fitting a model to data may be regarded as a way of replacing a set of data values \mathbf{y} by a set of fitted values $\hat{\boldsymbol{\mu}}$ derived from a model involving usually a relatively small number of parameters. In general the μ s will not equal the y s exactly, and the question then arises of how discrepant they are, because while a small discrepancy might be tolerable a large discrepancy is not. Measures of discrepancy or goodness of fit may be formed in various ways, but we shall be primarily concerned with that formed from the logarithm of a ratio of likelihoods, to be called the deviance.

Given n observations we can fit models to them containing up to n parameters. The simplest model, the *null model*, has one parameter, representing a common μ for all the y s; the null model thus consigns all the variation between the y s to the random component. At the other extreme the *full model* has n parameters, one per observation, and the μ s derived from it match the data exactly. The full model thus consigns all the variation in the y s to the systematic component leaving none for the random component.

In practice the null model is usually too simple and the full model is uninformative because it does not summarize the data but merely repeats them in full. However, the full model gives us a baseline for measuring the discrepancy for an intermediate model with p parameters.

It is convenient to express the log likelihood in terms of the mean-value parameter $\boldsymbol{\mu}$ rather than the canonical parameter $\boldsymbol{\theta}$. Let $l(\hat{\boldsymbol{\mu}}, \phi; \mathbf{y})$ be the log likelihood maximized over $\boldsymbol{\beta}$ for a fixed value of the dispersion parameter ϕ . The maximum likelihood achievable in a full model with n parameters is $l(\mathbf{y}, \phi; \mathbf{y})$, which is ordinarily finite. The discrepancy of a fit is proportional to twice the difference between the maximum log likelihood achievable and that achieved by the model under investigation. If we denote by $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\boldsymbol{\mu}})$ and $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}(\mathbf{y})$ the estimates of the canonical parameters under the two models, the discrepancy, assuming $a_i(\phi) = \phi/w_i$, can be written

$$\sum 2w_i\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\}/\phi = D(\mathbf{y}; \hat{\boldsymbol{\mu}})/\phi,$$

where $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ is known as the *deviance* for the current model and

is a function of the data only. Note that

$$D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) = D(\mathbf{y}; \hat{\boldsymbol{\mu}})/\phi,$$

so that the scaled deviance $D^*(\mathbf{y}; \boldsymbol{\mu})$ as defined in section 2.1.2 is the deviance expressed as a multiple of the dispersion parameter.

The forms of the deviances for the distributions given in Table 2.1 are as follows, summation being over $i = 1, \dots, n$:

Normal	$\sum(y - \hat{\mu})^2$,
Poisson	$2 \sum\{y \log(y/\hat{\mu}) - (y - \hat{\mu})\}$,
binomial	$2 \sum\{y \log(y/\hat{\mu}) + (m - y) \log[(m-y)/(m-\hat{\mu})]\}$,
gamma	$2 \sum\{-\log(y/\hat{\mu}) + (y - \hat{\mu})/\hat{\mu}\}$,
inverse Gaussian	$\sum(y - \hat{\mu})^2/(\hat{\mu}^2 y)$.

For the Normal distribution the deviance is just the residual sum of squares, while for the Poisson it is the statistic labelled G^2 by Bishop, Fienberg and Holland (1975) and others. The second term in the expressions for the Poisson and gamma deviances is often omitted for brevity. Provided that the fitted model includes a constant term, or intercept, the sum over the units of the second term is identically zero, justifying its omission. For details, see Nelder and Wedderburn, (1972).

The other important measure of discrepancy is the generalized Pearson X^2 statistic, which takes the form

$$X^2 = \sum(y - \hat{\mu})^2/V(\hat{\mu}),$$

where $V(\hat{\mu})$ is the estimated variance function for the distribution concerned. For the Normal distribution, X^2 is again the residual sum of squares, while for the Poisson or binomial distributions it is the original Pearson X^2 statistic.

Both the deviance and the generalized Pearson X^2 have exact χ^2 distributions for Normal-theory linear models (assuming of course that the model is true), and asymptotic results are available for the other distributions. However, asymptotic results may not be specially relevant to statistics calculated from limited amounts of data, and for these either D or X^2 may prove superior in its distributional properties. The deviance has a general advantage as a measure of discrepancy in that it is additive for nested sets of models if maximum-likelihood estimates are used, whereas X^2 in general is not. However, X^2 may sometimes be preferred because of its more direct interpretation.

2.3.2 *The analysis of deviance*

The analysis of variance, particularly when applied to orthogonal data with Normal errors, is a highly useful technique for screening the effects of factors and their interactions. We need some generalization of it applicable to the wider class of generalized linear models. There are two aspects of the generalization that need consideration: first, the terms in the model will, in general, no longer be orthogonal and secondly, sums of squares will, for non-Normal distributions, no longer be appropriate measures of the contribution of a term to the total discrepancy. The second problem is the more easily dealt with, and we consider it first. The terms in the analysis of variance can usefully be thought of as the first differences of the goodness-of-fit statistic for a sequence of models, each including one term more than the previous one. Thus the factorial model for two factors A and B gives rise to an analysis of variance with three terms A , B and the interaction $A.B$. The sums of squares for these are the first differences of the residual sums of squares obtained from fitting successively the models 1, A , $A+B$ and $A+B+A.B$, where 1 stands for the null model containing only the intercept. As an example, consider the following analysis of an unreplicated 4×3 factorial design indexed by A and B :

Model	d.f.	Discrepancy	Analysis of variance		
			s.s.	d.f.	Term
1	11	1000	500	3	A ignoring B
A	8	500	300	2	B eliminating A
$A + B$	6	200	200	6	$A.B$ eliminating A and B
$A + B + A.B$	0	0			

On the left is the sequence of models with their discrepancies, as measured by the residual sums of squares; note that the discrepancy for model 1 is just the total sum of squares about the mean in the analysis-of-variance table, while the last model is the full model, i.e. has as many parameters as observations, so that the degrees of freedom (d.f.) and the discrepancy are both zero. On the right is the analysis-of-variance table, with the sums of squares (s.s.) obtained from the first differences of the discrepancies.

The form of the generalization is now clear. Given a sequence of nested models we can use the deviance as our generalized measure of discrepancy, and form an analysis-of-deviance table by taking the first differences, as before. However, the interpretation of this table is now complicated by the non-orthogonality of the terms. Each number represents the variation accounted for by its corresponding term having eliminated the effects of those terms above it, but ignoring any effects of those terms below it. We may thus need to consider several model sequences, each producing its own analysis-of-deviance table. Note that this problem is present with classical linear models when non-orthogonality occurs. We shall not discuss here the various strategies that have been proposed for generating and looking at the goodness of fit of sets of model sequences. Suffice it to say that the aim of these strategies is to produce parsimonious models for the data in which terms that are not necessary are excluded. Note the use of the plural in 'models'; it is most unlikely with complex data that a single model will be a clear winner, and it can be most misleading to quote only the 'best' model, when several others are very close to it in terms of goodness of fit.

Once we depart from the Normal-theory linear model we generally lack an exact theory for the distribution of the deviance. In certain special cases, for example with observations in a simple design having exponential or inverse Gaussian distributions, exact results can be found. Usually, however, we rely on the χ^2 approximation for differences between deviances for nested models. See appendices A and C. In some circumstances the deviance itself may be approximated by χ^2 , for example in discrete data problems where the counts are large. In general, however, the χ^2 approximations for the deviance are not very good even as $n \rightarrow \infty$. Further work on the asymptotic distribution of $D(\mathbf{Y}; \hat{\boldsymbol{\mu}})$ remains to be done. The analysis-of-deviance table is best regarded as a screening device for picking out obviously important terms, no attempt being made to assign precise significance levels to the raw deviances.

2.4 Residuals

For Normal models we can express the dependent variate in the form

$$y = \hat{\mu} + (y - \hat{\mu}),$$

i.e. datum = fitted value + residual. Residuals can be used to explore the adequacy of fit of a model, in respect of choice of variance function, link function and terms in the linear predictor. Residuals may also indicate the presence of anomalous values requiring further investigation (see Chapter 12). For generalized linear models we require an extended definition of residuals, applicable to all the distributions that may replace the Normal. It is convenient if these residuals can be used for the same purposes as standard Normal residuals.

In the following section, we use the theoretical form, involving μ rather than $\hat{\mu}$, and we define three forms of generalized residual, which we call the Pearson, Anscombe and deviance residuals.

2.4.1 Pearson residual

The Pearson residual, defined by

$$r_P = \frac{y - \mu}{\sqrt{V(\mu)}}, \quad (2.11)$$

is just the raw residual scaled by the estimated standard deviation of Y . The name is taken from the fact that for the Poisson distribution the Pearson residual is just the signed square root of the component of the Pearson X^2 goodness-of-fit statistic, so that

$$\sum r_P^2 = X^2.$$

However Pearson's statistic is used in this book not so much as a goodness-of-fit statistic but as a measure of residual variation.

2.4.2 Anscombe residual

A disadvantage of the Pearson residual is that the distribution of r_P for non-Normal distributions is often markedly skewed, and so it may fail to have properties similar to those of a Normal-theory residual. Anscombe proposed defining a residual using a function $A(y)$ in place of y , where $A(\cdot)$ is chosen to make the distribution of $A(Y)$ ‘as Normal as possible’. Wedderburn (unpublished, but see Barndorff-Nielsen, 1978) showed that, for the likelihood functions occurring in generalized linear models, the function $A(\cdot)$ is given by

$$A(\cdot) = \int \frac{d\mu}{V^{1/3}(\mu)}.$$

Thus for the Poisson distribution we have

$$\int \frac{d\mu}{\mu^{1/3}} = \frac{3}{2}\mu^{2/3},$$

so that we base our residual on $y^{2/3} - \mu^{2/3}$. Now the transformation that ‘Normalizes’ the probability function does not at the same time stabilize the variance, so that we must scale by dividing by the square root of the variance of $A(Y)$, which is, to the first order, $A'(\mu)\sqrt{V(\mu)}$. Thus for the Poisson distribution the Anscombe residual, to be denoted by r_A , is given by

$$r_A = \frac{\frac{3}{2}(y^{2/3} - \mu^{2/3})}{\mu^{1/6}}.$$

See Anscombe (1953) and Cox and Snell (1968) for the definition of the corresponding residual for the binomial distribution. For the gamma distribution the Anscombe residual takes the form

$$r_A = \frac{3(y^{1/3} - \mu^{1/3})}{\mu^{1/3}}.$$

This cube-root transformation was used by Wilson and Hilferty (1931) to normalize variables with a χ^2 distribution. Similarly the inverse Gaussian distribution gives

$$r_A = (\log y - \log \mu)/\mu^{1/2}.$$

Table 2.2 Comparison of Poisson residuals

c	r_A	r_D	r_P
	$\frac{3}{2}(c^{2/3} - 1)$	$\{2(c \log c - c + 1)\}^{1/2}$	$c - 1$
0.0	-1.5	-1.414	-1.0
0.2	-0.987	-0.956	-0.8
0.4	-0.686	-0.683	-0.6
0.6	-0.433	-0.432	-0.2
1.0	0.0	0.0	0.0
1.5	0.466	0.465	0.5
2.0	0.881	0.879	1.0
2.5	1.263	1.258	1.5
3.0	1.620	1.610	2.0
4.0	2.280	2.256	3.0
5.0	2.886	2.845	4.0
10.0	5.462	5.296	9.0

2.4.3 Deviance residual

If the deviance is used as a measure of discrepancy of a generalized linear model, then each unit contributes a quantity d_i to that measure, so that $\sum d_i = D$. Hence if we define

$$r_D = \text{sign}(y - \mu) \sqrt{d_i},$$

we have a quantity that increases with $y_i - \mu_i$ and for which $\sum r_D^2 = D$. Thus for the Poisson distribution we have

$$r_D = \text{sign}(y - \mu) \{2(y \log(y/\mu) - y + \mu)\}^{1/2}.$$

Although the Anscombe and deviance residuals appear to have very different functional forms for non-Normal distributions, the values that they take for given y and μ are often remarkably similar, as is clear from a Taylor series expansion. Consider again the Poisson distribution and set $y = c\mu$, so that

$$r_A = \frac{3}{2}\mu^{1/2}(c^{2/3} - 1)$$

and

$$r_D = \text{sign}(c - 1)\mu^{1/2}[2(c \log c - c + 1)]^{1/2}.$$

Table 2.2 shows that the two functions $\frac{3}{2}(c^{2/3} - 1)$ and $[2(c \log c - c + 1)]^{1/2}$ are numerically very similar for a range of values of c .

Within this range the maximum difference between r_A and r_D is about 6% at $c = 0$, and much less over most of the range. The Pearson residual is considerably greater in the upper part of the range but goes less far in the negative direction.

For a more extensive examination of definitions of residuals in exponential-family models, see Pierce and Schafer (1986).

2.5 An algorithm for fitting generalized linear models

We shall show that the maximum-likelihood estimates of the parameters β in the linear predictor η can be obtained by iterative weighted least squares. In this regression the dependent variable is not y but z , a linearized form of the link function applied to y , and the weights are functions of the fitted values $\hat{\mu}$. The process is iterative because both the *adjusted dependent variable* z and the weight W depend on the fitted values, for which only current estimates are available. The procedure underlying the iteration is as follows. Let $\hat{\eta}_0$ be the current estimate of the linear predictor, with corresponding fitted value $\hat{\mu}_0$ derived from the link function $\eta = g(\mu)$. Form the adjusted dependent variate with typical value

$$z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0) \left(\frac{d\eta}{d\mu} \right)_0,$$

where the derivative of the link is evaluated at $\hat{\mu}_0$. The quadratic weight is defined by

$$W_0^{-1} = \left(\frac{d\eta}{d\mu} \right)_0^2 V_0, \quad (2.12)$$

where V_0 is the variance function evaluated at $\hat{\mu}_0$. Now regress z_0 on the covariates x_1, \dots, x_p with weight W_0 to give new estimates $\hat{\beta}_1$ of the parameters; from these form a new estimate $\hat{\eta}_1$, of the linear predictor. Repeat until changes are sufficiently small.

Note that z is just a linearized form of the link function applied to the data, for, to first order,

$$g(y) \simeq g(\mu) + (y - \mu)g'(\mu)$$

and the right-hand side is

$$\eta + (y - \mu) \frac{d\eta}{d\mu}.$$

The variance of Z is just W^{-1} (ignoring the dispersion parameter), assuming that η and μ are fixed and known. In this formulation the way in which the calculations for the regression are to be done is left open; we discuss some possibilities in section 3.8.

A convenient feature of this algorithm is that it suggests a simple starting procedure to get the iteration under way. This consists of using the data themselves as the first estimate of $\hat{\mu}_0$ and from this deriving $\hat{\eta}_0$, $(d\eta/d\mu)_0$ and V_0 . Adjustments may be required to the data to prevent, for example, our trying to evaluate $\log(0)$ as the starting value for η when the log link is applied to counts whose value is zero. These adjustments are described in the appropriate chapters, as will various complexities sometimes associated with the convergence of the iterative process.

2.5.1 Justification of the fitting procedure

We first show that the maximum-likelihood equations for β_j are given by

$$\sum W(y - \mu) \frac{d\eta}{d\mu} x_j = 0 \quad (2.13)$$

for each covariate x_j , where \sum without a suffix denotes summation over the units, and W is defined in equation (2.12) above.

The log likelihood for a single observation, in canonical form, is given by

$$l = \{y\theta - b(\theta)\}/a(\phi) + c(y, \phi)$$

and we require an expression for $\partial l/\partial\beta_j$. Now, by the chain rule,

$$\frac{\partial l}{\partial\beta_j} = \frac{\partial l}{\partial\theta} \frac{d\theta}{d\mu} \frac{d\mu}{d\eta} \frac{\partial\eta}{\partial\beta_j}.$$

From $b'(\theta) = \mu$ and $b''(\theta) = V$ we derive $d\mu/d\theta = V$, and from $\eta = \sum \beta_j x_j$ we get $\partial\eta/\partial\beta_j = x_j$. Therefore

$$\begin{aligned} \frac{\partial l}{\partial\beta_j} &= \frac{(y - \mu)}{a(\phi)} \frac{1}{V} \frac{d\mu}{d\eta} x_j \\ &= \frac{W}{a(\phi)} (y - \mu) \frac{d\eta}{d\mu} x_j \end{aligned}$$

from (2.12).

With constant dispersion ($a(\phi) = \phi$), the factor $a(\phi)$ disappears and we arrive at (2.13) after summing over the observations. With unequal prior weights, giving a dispersion of the form ϕ/w , an extra factor w enters (2.13).

Fisher's scoring method uses the gradient vector $\partial l/\partial\beta = \mathbf{u}$, say, and minus the expected value of the Hessian matrix

$$-E\left(\frac{\partial^2 l}{\partial\beta_r\partial\beta_s}\right) = \mathbf{A}, \quad \text{say.}$$

Given the current estimate \mathbf{b} of β , we derive an adjustment $\delta\mathbf{b}$ defined as the solution of

$$\mathbf{A}\delta\mathbf{b} = \mathbf{u}.$$

Now the components of \mathbf{u} (omitting the dispersion factor) are

$$u_r = \sum W(y - \mu) \frac{d\eta}{d\mu} x_r,$$

so that

$$\begin{aligned} A_{rs} &= -E \frac{\partial u_r}{\partial \beta_s} \\ &= -E \sum \left[(y - \mu) \frac{\partial}{\partial \beta_s} \left\{ W \frac{d\eta}{d\mu} x_r \right\} + W \frac{d\eta}{d\mu} x_r \frac{\partial}{\partial \beta_s} (y - \mu) \right] \end{aligned} \quad (2.14)$$

The first term vanishes on taking expectations while the second reduces to

$$\sum_i W \frac{d\eta}{d\mu} x_r \frac{\partial \mu}{\partial \beta_s} = \sum_i W x_r x_s.$$

Thus \mathbf{A} is the weighted sums-of-squares-and-products matrix of the covariates with weights W .

The new estimate $\mathbf{b}^* = \mathbf{b} + \delta\mathbf{b}$ of β thus satisfies the equation

$$\mathbf{A}\mathbf{b}^* = \mathbf{A}\mathbf{b} + \mathbf{A}\delta\mathbf{b} = \mathbf{A}\mathbf{b} + \mathbf{u}.$$

Now

$$(\mathbf{A}\mathbf{b})_r = \sum_s A_{rs} b_s = \sum_s W x_r \eta.$$

Thus the new estimate \mathbf{b}^* satisfies

$$(\mathbf{A}\mathbf{b}^*)_r = \sum_i W x_r \{\eta + (y - \mu) d\eta/d\mu\},$$

where the sum is over the n units. These equations have the form of linear weighted least-squares equations with weight

$$W = V^{-1} \left(\frac{d\mu}{d\eta} \right)^2$$

and dependent variate

$$z = \eta + (y - \mu) \frac{d\eta}{d\mu}.$$

Note that simplification occurs for the canonical links where the expected value and the actual value of the Hessian matrix coincide, so that the Fisher scoring method and the Newton-Raphson method reduce to the same algorithm. This comes about because the linear weight function $W d\eta/d\mu$ in the maximum-likelihood equations (2.13) is a constant, so that the first term in the expansion of the Hessian (2.14) is identically zero. Note also that $W = V$ for this case. Finally, if the model is linear on the scale on which Fisher's information is constant, i.e. $g'(\mu) = V^{-\frac{1}{2}}(\mu)$, the vector of weights is constant and need not be recomputed at each iteration.

2.6 Bibliographic notes

The fitting of generalized linear models is accomplished here using a variant of the Newton-Raphson method known as the scoring method. This variation was first introduced in the context of probit analysis by Fisher (1935) in the appendix of a paper by Bliss (1935). Details are given by Finney (1971). For further discussion and extensions see Green (1984) and Jørgensen (1984).

The term 'generalized linear model' is due to Nelder and Wedderburn (1972), who extended the scoring method to deal with maximum-likelihood estimation for exponential-family models. See also Bradley (1973) and Jennrich and Moore (1975).

Linear exponential family models (with canonical link) have been studied by Dempster (1971), Berk (1972), and Haberman (1977). For an extensive rigorous mathematical treatment see Barndorff-Nielsen (1978). Important special cases of linear exponential family models have been considered by Cox (1970) and by Breslow (1976).

2.7 Further results and exercises 2

2.1 Let $f_0(y)$ be an arbitrary density or probability distribution having moment generating function

$$M(\xi) = E\{\exp(\xi Y)\} = \exp\{b(\xi)\},$$

assumed finite for a range of ξ -values that includes 0. Now consider the exponentially weighted density

$$f_Y(y; \theta) \propto \exp(\theta y) f_0(y).$$

Derive the normalization factor for the weighted density and show that $f_Y(y; \theta)$ has the exponential-family form (2.4) with $a(\phi) = 1$.

2.2 Show that the cumulants of the weighted density $f_Y(y; \theta)$ are given by

$$\kappa_r = b^{(r)}(\theta),$$

whereas the cumulants of the initial density are $b^{(r)}(0)$.

2.3 Let Y_1, \dots, Y_ν be ν independent copies of the random variable Y having the weighted density function $f_Y(y; \theta)$. Show that the arithmetic mean $\bar{Y} = (Y_1 + \dots + Y_\nu)/\nu$ has a density of the form (2.4) with $a(\phi) = \nu^{-1}$. Show also that the cumulants of \bar{Y} are

$$\kappa_r(\bar{Y}) = b^{(r)}(\theta)/\nu^{r-1}.$$

Hence establish a central-limit theorem for densities of the form (2.4). [Jørgensen, 1987].

2.4 Discuss the limitations of the averaging operation as a way of generating a two-parameter family of distributions suitable for statistical work. Consider in particular the following points:

1. parameter interpretation,
2. possible non-integer values of ν ,
3. dependence of the support of \bar{Y} or $\nu\bar{Y}$ on the parameters, particularly where $f_0(y)$ is discrete.

2.5 Go through the calculations indicated in the previous four exercises, beginning with the distribution $f_0(y)$, which attaches probability one half to $y = 0$ and $y = 1$. What is the distribution of $\nu\bar{Y}$?

2.6 Beginning with the discrete distribution $f_0(y) \propto 1/y!$ for $y = 1, 2, \dots$, derive the corresponding exponential family by going through the calculations of Exercises 2.1–2.3. Find the cumulant function $b(\theta)$ and hence derive the likelihood equation for $\hat{\theta}$ based on a sample of independent and identically distributed observations.

2.7 For the distribution (2.4), show that the r th cumulant of Y is

$$\kappa_r = b^{(r)}(\theta) \times a^{r-1}(\phi).$$

Hence deduce that

$$\kappa_3 = \kappa_2 \kappa'_2 \quad \text{and} \quad \kappa_4 = \kappa_2 \kappa'_3,$$

where primes denote differentiation with respect to μ .

2.8 Show that

$$f_X(x; \theta, \nu) = \frac{(1 - x^2)^{\nu - 1/2}}{(1 - 2\theta x + \theta^2)^\nu B(\nu + \frac{1}{2}, \frac{1}{2})} \quad -1 \leq x \leq 1,$$

is a probability density on $(-1, 1)$ for all parameter values $\nu > -\frac{1}{2}$, $-1 \leq \theta \leq 1$ (McCullagh, 1989). [If all efforts at integration fail, check that the claim is true for $\theta = \pm 1, 0$ and, by numerical integration using Simpson's rule or other Newton-Cotes formula, for other values of (θ, ν) .]

Sketch the density for $\theta = 0, \pm \frac{1}{2}, \pm 1$, $\nu = 3$.

2.9 For the density given above, show that for all $r > -\nu$,

$$\begin{aligned} E\left(\frac{1 - X^2}{1 - 2\theta X + \theta^2}\right)^r &= \frac{B(\nu + r + \frac{1}{2}, \frac{1}{2})}{B(\nu + \frac{1}{2}, \frac{1}{2})}, \\ E\left(\frac{X - \theta}{1 - 2\theta X + \theta^2}\right) &= 0. \end{aligned}$$

Hence deduce that $T(\theta) = (1 - X^2)/(1 - 2\theta X + \theta^2)$ is a pivotal statistic whose distribution does not depend on θ . Find the distribution of T .

2.10 Show that for fixed θ the density $f_X(x; \theta, \nu)$ given above is of the exponential-family type (2.4) with $\phi = 1$, $y = \log T(\theta)$ and canonical parameter ν . Find the cumulant function $b(\cdot)$.

2.11 Show that $-2\nu \log T(\theta_0)$ is the scaled deviance statistic for testing the hypothesis $H_0 : \theta = \theta_0$ on the basis of a single observation X . Deduce that for large ν and under H_0

$$-(2\nu + \frac{1}{2}) \log T(\theta_0) \sim \chi_1^2$$

approximately.

2.12 Suppose that X_1, \dots, X_n are independent and identically distributed with density $f_X(x; \theta, \nu)$ as given above. Show that $\hat{\theta}_\nu$, the maximum-likelihood estimate of θ for fixed ν , is independent of ν . Calculate the Fisher information for (θ, ν) and show that this matrix is diagonal.

2.13 Using the result given in Exercise 2.8 show that

$$f_X(x; \theta, \nu) = \frac{(1 - x^2)^{\nu - 1/2} |\theta|^{2\nu}}{(1 - 2\theta x + \theta^2)^\nu B(\nu + \frac{1}{2}, \frac{1}{2})} \quad \nu > -\frac{1}{2}, |\theta| \geq 1,$$

is a probability density on the interval $-1 \leq x \leq 1$ for the parameter values indicated. Comment on the behaviour of the likelihood function and the Fisher information near $\theta = \pm 1$. [McCullagh, 1989].

2.14 In order to construct a family of the type (2.4), suppose we begin with the logistic density

$$f_0(x) = \frac{e^x}{(1 + e^x)^2} = \frac{1}{(2 \cosh(x/2))^2} \quad \text{for } -\infty < x < \infty.$$

Show that the associated exponential family, also known as the exponentially tilted family, is

$$f(x; \theta) = \frac{e^{x(1+\theta)}}{(1 + e^x)^2 \Gamma(1 + \theta) \Gamma(1 - \theta)} = \frac{e^{x(1+\theta)} \sin(\pi\theta)}{(1 + e^x)^2 \pi\theta}$$

for $-1 < \theta < 1$. Deduce that $f(x; \theta) = f(-x; -\theta)$. Plot the density for $\theta = 0.25, 0.5$ and 0.75 . Find the cumulant function $b(\theta)$ and show that the mean of the tilted density is

$$E(X; \theta) = b'(\theta) = \frac{1}{\theta} - \pi \cot(\pi\theta).$$

Plot $E(X; \theta)$ against θ to show that the mean-value parameter is a monotone function of the canonical parameter.

For what values of θ does $\exp(X)$ have an F -distribution?

2.15 Discuss the connection between the above exponential family and the family generated by the particular hyperbolic secant density

$$f_2(x; 0) = \frac{x}{2 \sinh(\pi x/2)} \quad \text{for } -\infty < x < \infty,$$

whose cumulant generating function is $-2 \log \cos \theta$ for $|\theta| < \pi/2$.

Find the mean and variance of the tilted density as functions of θ . Plot the exponentially tilted density $f_2(x; \theta)$ for $\theta = 0, \pi/6$ and $\pi/3$. [Morris, 1982, section 5.]

CHAPTER 3

Models for continuous data with constant variance

3.1 Introduction

Generalized linear models are essentially an extension of classical linear models and this chapter presents these classical models in a way that makes the extension appear natural. There is an enormous literature on classical linear models, not all of it helpful to the reader, and no attempt will be made in this chapter to give a comprehensive account of the subject. Rao (1973), Draper and Smith (1981), Seber (1977) and Atkinson (1985) are excellent reference books covering various aspects of classical linear models.

The subject matter of this chapter is linear models, which we shall write in the following form:

$$\begin{array}{lll} Y_i \sim N(\mu_i, \sigma^2), & \boldsymbol{\mu} = \boldsymbol{\eta}, & \boldsymbol{\eta} = \sum_1^p \mathbf{x}_j \beta_j, \\ \text{observations Normally} & \text{identity} & \text{linear predictor} \\ \text{distributed and} & \text{link;} & \text{based on} \\ \text{independent;} & & \text{covariates} \\ & & \mathbf{x}_1, \dots, \mathbf{x}_p. \end{array} \quad (3.1)$$

The data vector \mathbf{y} , the mean vector $\boldsymbol{\mu}$, and the linear predictor, $\boldsymbol{\eta}$, all have n components. The leftmost component of (3.1) is a specification of the random part of the model. The other components describe the systematic parts, which include the construction of the linear predictor $\boldsymbol{\eta}$ from the covariates, and the link between $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$. By suppressing the link, and regarding the \mathbf{x}_j as the p columns of a matrix \mathbf{X} , we recover the standard matrix formulation

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta},$$

where β is the set of parameters written in vector form. Note that we have restricted our attention to the sub-class of linear models in which there is only one error component and in which the errors are independent. Models involving components of variance are therefore excluded.

We now consider in more detail the random and systematic parts of the model (3.1).

3.2 Error structure

In classical linear models, the vector of observations, y , is assumed to be a realization of a random variable, Y , which is Normally distributed with moments

$$E(Y) = \mu \quad \text{and} \quad \text{cov}(Y) = \sigma^2 I. \quad (3.2)$$

Thus the observations are assumed to have equal variances and to be independent.

The assumption of Normality, although important as the basis for an exact small-sample theory, is not so important in large samples. For there the central-limit theorem offers protection from all but the most extreme distributional deviations from Normality. There may, however, be a modest loss of efficiency, which can be recovered if the true distribution is known and used in place of the Normal. For details, see Cox and Hinkley (1968).

The theory of least squares can be developed using only first- and second-moment assumptions in addition to independence, without requiring the additional assumption of Normality. This is fortunate because in applications we can rarely be entirely confident that the assumed distributional form is correct. It is this second-order aspect of linear models that is emphasized here. From the present viewpoint, therefore, the important assumption in (3.2) is that the variance of an observation is the same for all values of μ . This is an assumption that can and should be checked, either by graphical examination of the residuals or by computing an appropriate test statistic. Checks such as these are described in Chapter 12.

The emphasis on second-moment assumptions over fully specified distributional assumptions extends to all generalized linear models and is discussed more fully in Chapter 9.

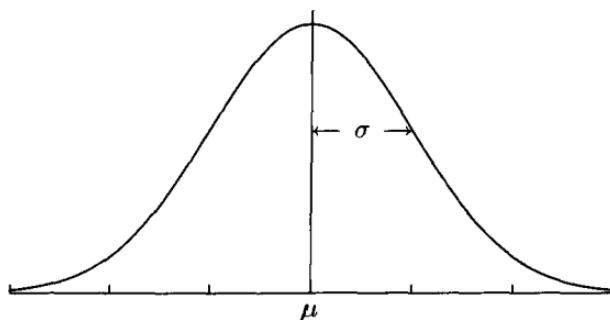


Fig. 3.1. *The Normal (or Gaussian) distribution with mean μ and standard deviation σ .*

The frequency function of the univariate Normal distribution takes the form

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \quad \text{for } -\infty < y < \infty.$$

The distribution is symmetrical with mode, mean and median all at μ . The standard deviation, σ , is the horizontal distance between the mean and the point of inflection of the density. About 68%, 95% and 99.8% of the distribution lies in the ranges $\mu \pm \sigma$, $\mu \pm 2\sigma$ and $\mu \pm 3\sigma$ respectively. The log-likelihood function for a single observation with known variance is a parabola whose maximum is at y and whose second derivative is $-1/\sigma^2$.

The Normal distribution is useful primarily as a model for measurements of continuous quantities, though it can also be used as an approximation for discrete measurements. It is frequently used to model data, such as weights, lengths and time, which, though continuous, are essentially positive, although the distribution itself covers the entire real line. Such usage is acceptable in practice provided that the data values are sufficiently far removed from zero. If, for example, data have a mean of 100 and a standard deviation of 10, the part of the Normal distribution covering the negative half of the real line is negligible for most practical purposes. If data y that are essentially positive approach the origin, then it will often be found that the data themselves contradict the assumption of constant variance independent of μ . When this occurs, a Normal distribution for $\log Y$ will often be found to be a better approximation than a Normal distribution for Y . Alternatively, the gamma distribution (Chapter 8) may be used.

3.3 Systematic component (linear predictor)

We aim in this section to study various aspects of the linear predictor

$$\eta = \sum_1^p \mathbf{x}_j \beta_j,$$

which occurs in all generalized linear models. The covariates, $\mathbf{x}_1, \dots, \mathbf{x}_p$, may be continuous measurements, incidence vectors for qualitative factors of various types, or incidence vectors for interactions among these. Concise description and automatic construction of such vectors is an important aspect of the specification and fitting of generalized linear models.

3.3.1 Continuous covariates

These comprise covariates such as mass, temperature, time, amount of fertilizer or drug, concentration of a solute and so on, which can take values on a continuous scale. Models containing only terms with continuous covariates are often called regression models, to be contrasted with analysis-of-variance models, which have only terms involving qualitative factors. Provided that there is only one component of error variance, we shall not make this distinction. Indeed, by introducing mixed terms in section 3.3.4, we shall deliberately seek to blur the distinction because many interesting models involve terms of both types.

Linearity in the present context means linearity of η in the parameters. Consequently a continuous covariate x in a model term may be replaced by an arbitrary function $g(x)$, such as $\log(\text{dose})$ in a dose-response model, without destroying the linearity of the model. In particular we may use x^2, x^3, \dots in addition to x to build up a polynomial in x , without destroying the linearity. Similarly, the linear model $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ may be expanded to include the product term $\beta_{12} x_1 x_2$, producing a bilinear relationship. If the terms are rearranged in the form

$$(\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_{12} x_2)x_1,$$

they show a linear relationship in x_1 in which both slope and intercept are linear functions of x_2 . The alternative rearrangement

$$(\beta_0 + \beta_1 x_1) + (\beta_2 + \beta_{12} x_1)x_2$$

expresses the bilinearity in complementary form.

A function such as $\exp(\gamma x)$, however, produces a non-linear model unless γ is known a priori. If γ is unknown, the model is not linear, and some non-linear optimization technique is required to minimize the discrepancy function. It may, however, be helpful to fit the model for a few suitably chosen values of γ . Such models that are partly linear and partly non-linear are discussed in Chapter 11.

3.3.2 Qualitative covariates

Sets of observations are frequently indexed by one or more classifying factors, or factors for short. Each factor has an associated index, whose values partition the data into disjoint groups or classes. Thus, in a field experiment, one such factor might define the block into which each unit (plot) falls, while another might define the crop variety to be planted in that plot.

A factor can take only a limited set of possible values, to be called *levels*. The k levels can always be coded using the integers $1, 2, \dots, k$, although the coding $0, 1, \dots, k-1$ is sometimes more convenient. Such a coding defines the *formal levels* of a factor. In practice the levels usually have names or numerical values and these we call *actual levels*. Actual levels may be

1. ordered with numerical formal levels, such as the amount of fertilizer in an agricultural experiment; or
2. ordered but without relative magnitudes for the levels, such as socioeconomic status; or
3. unordered, such as the names of crop varieties in a variety trial.

Factors occurring in a model may be of primary interest, meaning that a principal purpose of the study is to measure their effect. Treatment factors in a designed experiment are obviously of this kind. In surveys, classification factors such as educational status, marital status, religious affiliation and so on are of this type. Factors of secondary interest are those producing effects that must be accommodated in the model, but which are not of primary interest. Examples are blocking factors in a randomized blocks design and, usually, census enumeration district in a survey. The distinction between primary and secondary factors is not absolute, but depends on the aims of the study concerned.

The simplest term in a linear predictor generated by a factor is a component of the intercept. Consider a model with one covariate x and linear predictor

$$\eta = \alpha + \beta x.$$

If A is a factor with index i , then the extended linear predictor might become

$$\eta_i = \alpha_i + \beta x,$$

implying a separate intercept for each level of A , but a common slope β , assumed constant over the levels of the factor. Note that if a factor has numerical levels, we could also treat it as a quantitative covariate having only a few distinct values. If we treat it as a factor, we fit a separate effect for each level in an unstructured way, whereas if we treat it as a quantitative variate, we impose a linear form on the response. Alternatively, and perhaps preferably, we may use polynomials in the actual levels to detect deviations from linearity.

Frequently data are cross-classified by many factors simultaneously. If A , B and C are three such factors with indices i, j, k respectively, the simplest model ordinarily considered has the form

$$\alpha_i + \beta_j + \gamma_k.$$

This is the so-called main-effects model, which implies that if we arrange the data in a rectangular block and then look at cross-sections of the data for each level of A , we shall find that they can be modelled by effects of B and C that are additive and equal in each cross-section. Similarly for the other factors. In order to achieve a satisfactory fit, however, it may be necessary to include terms analogous to $\beta_{12}x_1x_2$ with continuous covariates. Such terms, of the algebraic form $(\alpha\beta)_{ij}$, imply a separate effect for each combination of the indices i and j and are called interactions. We shall refer to $(\alpha\beta)_{ij}$ as a two-factor interaction, but the term ‘first-order interaction’ is also used, the order being one less than the number of factors involved.

The relationships between interactions and main effects have been the subject of much confusion in the literature. We consider them in more detail in section 3.5.

3.3.3 Dummy variates

If i is the index for the levels of factor A with k levels, the term α_i may be written in vector notation as

$$\alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \dots + \alpha_k \mathbf{u}_k,$$

where the \mathbf{u}_j are dummy variates whose components take the value 1 if the unit has factor A at level j , and zero otherwise. The terms *incidence vector* and *indicator vector* are also used. Thus if $k = 3$ and the formal levels for five observations are 1, 2, 2, 3, 3, the dummy variates $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ take values as follows:

<i>Unit</i>	<i>A</i>	\mathbf{u}_1	\mathbf{u}_2	\mathbf{u}_3
1	1	1	0	0
2	2	0	1	0
3	2	0	1	0
4	3	0	0	1
5	3	0	0	1

Note that,

$$\mathbf{u}_1 + \mathbf{u}_2 + \mathbf{u}_3 \equiv \mathbf{1}$$

irrespective of the allocation of levels to units. The constant vector, $\mathbf{1}$, is the dummy variate corresponding to the intercept term, often written as μ , in the linear predictor. The relation between the terms μ and α_j is a simple instance of intrinsic aliasing, to be discussed in section 3.5.

A compound term such as $(\alpha\beta)_{ij}$ has dummy variates, $(\mathbf{uv})_{ij}$, whose values are products of corresponding components of \mathbf{u}_i and \mathbf{v}_j , the dummy variates for A and B as single-factor terms. It follows then that

$$\sum_i (\mathbf{uv})_{ij} = \mathbf{v}_j \quad \text{and} \quad \sum_j (\mathbf{uv})_{ij} = \mathbf{u}_i,$$

again irrespective of the allocation of factor levels to units. Thus main effects are intrinsically aliased with interactions in which they are included.

3.3.4 Mixed terms

In section 3.3.2 we considered a model

$$\eta_i = \alpha_i + \beta x,$$

in which the intercept varies with the factor level, but where the slope is constant over levels. Sometimes, however, the slope may also change with the factor level, requiring the term βx to be replaced by $\beta_i x$. Terms in the linear predictor in which a slope or regression coefficient changes with the level of one or more factors are called *mixed*, because they include aspects of both continuous and qualitative covariates. It is important that any computer program for fitting linear models should allow mixed terms to be specified as easily as continuous and qualitative terms, because the assumption frequently made, that a slope is the same for all levels of a factor, ought to be easily testable. The simplest test is to compare the fit of the model having constant slope with the fit when the slope is allowed to vary from level to level.

Dummy variates for mixed terms take the same form as those for factors except that the 1s are replaced by the corresponding x -values. Using the same factor allocation as in the previous section, and with the covariate x as shown, the dummy variates for the mixed term $\beta_i x$, again written as $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$, take values as follows:

<i>Unit</i>	<i>A</i>	<i>x</i>	\mathbf{u}_1	\mathbf{u}_2	\mathbf{u}_3
1	1	1	1	0	0
2	2	3	0	3	0
3	2	5	0	5	0
4	3	7	0	0	7
5	3	9	0	0	9

Here

$$\mathbf{u}_1 + \mathbf{u}_2 + \mathbf{u}_3 = \mathbf{x},$$

again irrespective of the allocation of levels to units.

3.4 Model formulae for linear predictors

3.4.1 Individual terms

We now describe a notation that is helpful for the specification of linear predictors in generalized linear models. The notation, due to Wilkinson and Rogers (1973), is compact and is easily adapted for use in computer programs. The convention is continued that names beginning with letters from the first half of the alphabet refer to factors, and those from the second half to continuous covariates. The indices associated with the levels of factors A, B, C, \dots are i, j, k, \dots . The Table below lists some kinds of terms that occur in simple model formulae. Algebraic expressions are presented together with the corresponding model formula term. Note the use of λ instead of β for the coefficient of a continuous covariate to avoid confusion with the parameters for factor B .

Type of term	Algebraic	Model formula term
Continuous covariate	λx	X
Factor	α_i	A
Mixed	$\lambda_i x$	$A.X$
Compound	$(\alpha\beta)_{ij}$	$A.B$
Compound mixed	$\lambda_{ij} x$	$A.B.X$

In the model-formula version X stands for itself, a single vector. By contrast, A stands for a set of dummy variates, one variate as indicator for each level of the factor. The remaining types of term also stand for the appropriate set of dummy variates. Thus terms in a model formula represent vector subspaces and do not involve the parameters explicitly. Parameters occur only implicitly, one per basis vector in each subspace.

3.4.2 The dot operator

This operator, already exemplified in the formation of compound terms, implies the formation of all elementwise products of the constituent vectors. For example, if A is the three-level factor and X the covariate vector with values shown at the end of section 3.3.4, then $A.X$ denotes the three vectors $(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$ shown there. By extension, if B is a two-level factor with dummy variates $\mathbf{v}_1, \mathbf{v}_2$, then $A.B$ denotes six dummy vectors corresponding to all

elementwise products of \mathbf{u}_i with \mathbf{v}_j . Note that if A has k levels, $A.A$ comprises k vectors equal to the k dummy vectors for A , and $k(k - 1)$ null vectors with all components zero. Such null vectors may be omitted, effectively making

$$A.A = A.$$

However, in general,

$$X.X \neq X,$$

the left-hand side being a vector with components x_i^2 . (Note that in both the computer programs GLIM and Genstat, where this notation has been adopted, compound terms involving more than one continuous covariate are not permitted in model formulae. They must be computed explicitly, preferably after subtracting column means.)

The dot operator is commutative so that

$$A.B \equiv B.A,$$

and associative, so that

$$(A.B).C \equiv A.(B.C).$$

Thus we may write $A.B.C$ without ambiguity, the order in which the factors are included being unimportant.

3.4.3 The + operator

Terms in a model formula may be joined using the operator $+$, with exactly the same usage as in the algebraic expression for the model formula. Repetitions of terms are ignored, so that

$$A + A \equiv A,$$

it being pointless to specify the same vector subspace twice. In vector-space terminology $A + B$ defines a subspace in R^n spanned by linear combinations of vectors in A and B .

It is convenient to assign lower priority to $+$ than to the dot, so that

$$A.B + C \equiv (A.B) + C.$$

The dot is distributive with respect to $+$, so that

$$A.(B + C) \equiv A.B + A.C.$$

These are the fundamental operators required in the specification of model formulae and the other useful operators that follow are defined in terms of them.

3.4.4 The crossing (*) and nesting (/) operators

The crossing operator, denoted by $*$, is used mainly to simplify the specification of factorial models. Thus

$$A * B \equiv A + B + A.B$$

$$A * B * C \equiv A + B + C + A.B + A.C + B.C + A.B.C,$$

and so on. In these expansions A and B may themselves be replaced by model formulae. The operator $*$ has higher priority than $+$, but lower priority than dot. Thus

$$A * B + C \equiv A + B + C + A.B$$

$$A * B.C \equiv A + B.C + A.B.C.$$

Note the convention followed in expanding expressions, that all simple terms come first, followed by two-component terms and so on. This convention, though not essential, is helpful when it comes to understanding intrinsic aliasing in models (Section 3.5).

The crossing operator is associative, and distributive with respect to $+$, for

$$\begin{aligned} A * (B + C) &\equiv A + (B + C) + A.(B + C) \\ &\equiv A + B + C + A.B + A.C \\ &\equiv A + B + A.B + A + C + A.C \\ &\equiv A * B + A * C. \end{aligned}$$

When a compound term such as $A.B$ is preceded in an expanded model formula by both constituent terms A and B , it is called the *interaction* of A and B . The nature of the interaction term will be discussed further in section 3.5.

The nesting operator $/$ relates to an indexing system, which, in its simplest form, has two indices i and j , but no connection between observations (i, j) and (i', j) , though there is a connection between observations (i, j) and (i, j') . Typically, i defines the levels of a blocking factor and j identifies an element within a block. There is no necessary connection between the first observation in one block and the first observation in another, but two observations in the same block have their block in common and may tend to be similar on that account. For a nested treatment structure, consider

a set of plant varieties categorized as early ($i = 1$), mid-season ($i = 2$), or late ($i = 3$) in cropping. Within each group there is a number of distinct varieties, no variety belonging to more than one group. Two varieties may be connected by being in the same group (i the same), but there is no connection between the first variety in two different cropping groups (j the same, i different). The appropriate linear predictor for nesting is written as

$$A/B \equiv A + A.B,$$

In the expanded formula the compound term $A.B$ is preceded by only one constituent term. The interpretation of $A.B$ is now that of B within A .

As before, A and B may themselves be model formulae, with the rule that if $\text{pt}(A)$ denotes the product term (using dots) of all elements in A , then A/B is defined by

$$A/B \equiv A + \text{pt}(A).B.$$

Thus, for example

$$(A * B)/C \equiv A * B + A.B.C.$$

The nesting operator is associative, so that

$$A/(B/C) \equiv (A/B)/C,$$

and distributive with respect to $+$, since

$$A/(B+C) = A + A.(B+C) = A + A.B + A + A.C = A/B + A/C.$$

Like the crossing operator, the nesting operator is given a priority between $.$ and $+$. By convention we give it higher priority than $*$.

3.4.5 Operators for the removal of terms

The operator $-$ has the obvious meaning as the inverse or opposite of $+$. It is used for the removal of terms in a model formula. Thus

$$A * B - A.B \equiv A + B.$$

Similarly,

$$A * B * C - A.B.C \equiv A + B + C + A.B + A.C + B.C$$

is a concise notation for a model with all main effects and two-factor interactions.

It is sometimes required to remove from a model all those compound terms that include a given factor or factors. Two operators $-/$ and $-*$ cater for this; $-/A$ means ‘remove all compound terms that include A , but excluding A itself’, while $- * A$ means ‘remove all terms that include A ’. Thus

$$A * B * C -/ A \equiv A + B + C + B.C$$

and

$$A * B * C - * A \equiv B + C + B.C.$$

3.4.6 Exponential operator

If M is a model formula and I is an integer, then

$$M^{**I} \equiv M * M * \dots * M,$$

the right side containing I M s. This operator is useful for specifying factorial models that include all terms up to a given level of interaction. For example

$$(A + B + C)^{**2} \equiv A + B + C + A.B + A.C + B.C.$$

This operator has highest priority.

We shall use this notation for the specification of linear predictors wherever possible. Readers should bear in mind that this model-formula notation (strictly speaking, a subset of it) can be used directly for this purpose in the computer systems Genstat and GLIM.

3.5 Aliasing

Each term in a model formula describes a set of covariates to be included in a linear predictor. If such a set is denoted by $\mathbf{x}_1, \dots, \mathbf{x}_p$, the \mathbf{x} s being n -vectors, then the covariates can be thought of as defining p directions in n -dimensional Euclidean space. These p vectors define a subspace of up to p dimensions. The maximum dimension is achieved if the \mathbf{x} s are linearly independent, i.e. if there does not exist a set of coefficients ξ_j , not all zero, such that

$$\sum_1^p \xi_j \mathbf{x}_j = \mathbf{0}.$$

If k independent linear relations exist, then the set of covariates spans a space of dimension $p - k$. Ordinarily the individual terms in an expanded model formula will form subspaces of maximum dimension. Loss of dimension may occur, however, when we consider joint subspaces covered by more than one term.

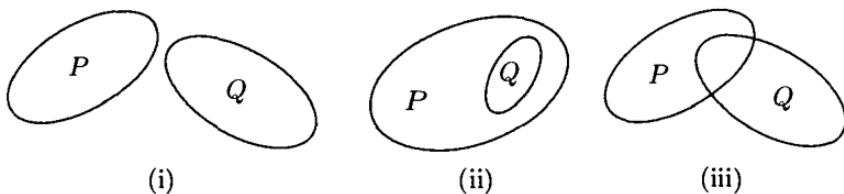


Fig. 3.2. Venn diagrams for relationships between subspaces of terms in a linear model: (i) P and Q linearly independent; (ii) Q entirely aliased with P ; (iii) Q partially aliased with P .

We now consider the possible relationships between the subspaces defined by two terms in a model formula. The terms are denoted by P and Q , their dimensions by p and q , with $p \geq q$. There are three possible relationships between P and Q .

1. All $p + q$ vectors defining P and Q are linearly independent, so that the dimension of the space $P + Q$ is $p + q$.
2. All the vectors of Q are expressible as linear combinations of the p vectors in P , so that the dimension of $P + Q$ is p .
3. k of the q vectors in Q are expressible as linear combinations of those in P .

The corresponding Venn diagrams are shown in Fig. 3.2. Clearly (i) and (ii) are extreme cases of (iii) for which $k = 0$ and $k = q$ respectively. Note the special case of (ii) when $p = q$, so that P and Q span identical subspaces.

The effect on the terms in a generalized linear model of overlapping subspaces is to produce what is called aliasing. Certain combinations of covariates are then identical to other combinations, so that the corresponding combinations of parameters cannot be distinguished. Consider, for example, measurements made on leaves having the property that area = constant \times length \times breadth, with length and breadth being measured as well as area. Suppose that the covariates in the model are

$$\begin{aligned}x_1 &= \log \text{length}, \\x_2 &= \log \text{breadth}, \\x_3 &= \log \text{area},\end{aligned}$$

and that the linear predictor is to be formed as

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

Now, since area = constant \times length \times breadth, we have

$$x_3 = c + x_1 + x_2, \tag{3.3}$$

where c is the logarithm of the constant in the formula for area. Hence η may be expressed in terms of x_1 and x_2 as

$$\begin{aligned}\eta &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3(c + x_1 + x_2) \\&= \beta_0 + \beta_3 c + (\beta_1 + \beta_3)x_1 + (\beta_2 + \beta_3)x_2.\end{aligned}$$

Thus we can distinguish the three combinations of the β s

$$\beta_0 + \beta_3 c, \quad \beta_1 + \beta_3, \quad \text{and} \quad \beta_2 + \beta_3,$$

but not the four parameters $\beta_0, \beta_1, \beta_2, \beta_3$ separately. If we write x_0 for the constant vector, i.e. the dummy vector for the term β_0 , we see from (3.3) that x_3 is a linear combination of x_0, x_1 and x_2 . In other words, the subspace for x_3 , with one dimension, is contained in the sum or span of the subspaces for x_0, x_1 and x_2 .

If in addition the leaves are all of the same shape, in the sense that the ratio of length to breadth is constant, we have

$$x_2 = c' + x_1,$$

where breadth/length = $\exp(c')$. The linear predictor now reduces further to

$$\begin{aligned}\eta &= \beta_0 + \beta_1 x_1 + \beta_2(c' + x_1) + \beta_3(c + x_1 + c' + x_1) \\ &= \beta_0 + \beta_2 c' + \beta_3(c + c') + (\beta_1 + \beta_2 + 2\beta_3)x_1.\end{aligned}$$

Now, only two parameter combinations, namely

$$\beta_0 + \beta_2 c' + \beta_3(c + c') \quad \text{and} \quad \beta_1 + \beta_2 + 2\beta_3,$$

are distinguishable, and the dimension of the space spanned by x_0, x_1, x_2 and x_3 is reduced from four to two.

An important aspect of this example is that the aliasing is intrinsic to the problem. Given that all leaves are the same shape and that all measurements are made without error, aliasing will occur whatever the sizes of the leaves. Such *intrinsic aliasing* is found most commonly, however, where terms involving factors occur in a model.

3.5.1 Intrinsic aliasing with factors

Consider a model formula containing the intercept together with the single factor A , which we write as

$$1 + A,$$

where 1 stands for the dummy vector with all elements 1. An equivalent algebraic expression for the components of the linear predictor is

$$\eta_{ij} = \mu + \alpha_i,$$

where i indexes the groups defined by A and j indexes the units or observations within the groups. The dummy vectors for A add up to the constant vector, or dummy vector for μ , because each observation has factor A at exactly one level. Thus μ is aliased with $\sum \alpha_i$, and further, it is intrinsically aliased because

the relation holds whatever the allocation of units to the groups. The relationship between μ and α_i is not symmetric because the dummy vector for μ lies wholly in the space of the dummy vectors for α_i , but not vice versa. We say that μ is *marginal* to the α s. As a consequence, the terms in the model $\mu + \alpha_i$ are ordered because of the marginality relationship. One effect of this ordering is that it does not make sense to consider the hypothesis that $\mu = 0$ when the α_i are not assumed known.

The linear predictor is clearly unchanged if we add a constant to μ and subtract the same constant from each α_i . This operation leaves unchanged the quantities $\mu + \alpha_i$ and also any contrast $\sum \lambda_i \alpha_i$ with $\sum \lambda_i = 0$. Combinations that are unaffected by this operation are said to be *estimable*. The parameters μ and α_i separately are not estimable because the aliasing pattern makes them indistinguishable from $\mu + c$ and $\alpha_i - c$. When we come to estimate the parameters, this ambiguity can be resolved by imposing a constraint on the estimates to give a unique solution to the least-squares equations. It must be stressed, however, that any such constraint on the estimates $\hat{\mu}, \hat{\alpha}_i$ of μ, α_i is a convention only, and is of no significance in judging the adequacy of the model. Constraints are not to be thought of as part of the model specification: they are merely a convenient way of resolving an ambiguity and they do not affect the meaning or interpretation of the model. In particular, there is no implication that a similar constraint should be imposed on the parameters μ and α_i ; in fact, where intrinsic aliasing occurs, the imposition of constraints on parameters as well as on their estimates is a common source of confusion.

For the above model, three possible constraints, chosen from an infinity of possibilities, are as follows:

1. $\hat{\mu} = 0$, so that the $\hat{\alpha}_i$ give the group means directly;
2. $\hat{\alpha}_1 = 0$, so that the first group mean is $\hat{\mu}$, and $\hat{\alpha}_2, \hat{\alpha}_3, \dots$ measure differences between other group means and the first;
3. $\sum \hat{\alpha}_i = 0$, so that $\hat{\mu}$ is the average of the group means and $\hat{\alpha}_i$ is the deviation of the i th group mean from $\hat{\mu}$.

As an example, consider four groups with means 6, 9, 12 and 13. Then the three constraints produce parameter estimates in the linear predictor with the following values:

<i>Parameter</i>	<i>Estimate with constraint</i>		
	(1.)	(2.)	(3.)
μ	0	6	10
α_1	6	0	-4
α_2	9	3	-1
α_3	12	6	2
α_4	13	7	3

Another constraint that is sometimes used if the group sizes are unequal is

$$\sum w_i \hat{\alpha}_i = 0,$$

where w_i is the i th group size. With this constraint $\hat{\mu}$ is a weighted average of the group means and $\hat{\alpha}_i$ is the deviation of the i th group mean from the weighted average.

3.5.2 Aliasing in a two-way cross-classification

Failure to recognize the aliasing pattern and the arbitrariness of imposed constraints has led to much confusion in the literature, especially in the analysis of models for two-way cross-classifications. The discussion here follows the lines of Nelder (1977).

We are concerned with the linear model

$$1 + A + B + A.B,$$

expressed algebraically by the linear predictor

$$\eta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}.$$

The dummy vectors for the four terms show the following relationships, here written in terms of the parameters rather than the dummy vectors. (The equivalence sign may be read as ‘is indistinguishable from’.)

$$\begin{array}{ll} \sum \alpha_i \equiv \mu, & \sum \beta_j \equiv \mu, \\ \sum_j \gamma_{ij} \equiv \alpha_i, & \sum_i \gamma_{ij} \equiv \beta_j. \end{array}$$

These identities imply that the sum of all the dummy vectors for $A.B$ is the constant vector, or, in terms of the parameters,

$$\sum_{ij} \gamma_{ij} \equiv \mu.$$

Thus the relationships among the terms are as follows:

- μ is marginal to α_i, β_j and γ_{ij} ,
- α_i is marginal to γ_{ij}
- and β_j is marginal to γ_{ij} .

The terms are thus partially ordered as first μ , then α_i and β_j together, and finally γ_{ij} . The estimable parameter combinations are the linear predictor itself,

$$\eta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij},$$

and also the contrasts

$$\begin{aligned} & \sum \lambda_i (\alpha_i + \bar{\gamma}_{i\cdot}); \quad \text{with} \quad \sum \lambda_i = 0, \\ & \sum \lambda_j (\beta_j + \bar{\gamma}_{\cdot j}); \quad \text{with} \quad \sum \lambda_j = 0, \\ & \text{and} \quad \sum \lambda_{ij} \gamma_{ij}; \quad \text{with} \quad \sum_i \lambda_{ij} = \sum_j \lambda_{ij} = 0, \end{aligned}$$

where $\bar{\gamma}_{i\cdot}, \bar{\gamma}_{\cdot j}$ denote averages over the indicated indices.

The ambiguities about the values of the estimates of individual parameters can again be resolved by suitable constraints. Two such constraints are now discussed for a 2×2 array.

The full parameterization has nine parameters, namely $(\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22})$, but only four linearly independent estimable combinations. Thus, five suitably chosen constraints are required to produce unique estimates. The conventional system of symmetric constraints is given by

$$\begin{aligned} \hat{\alpha}_1 + \hat{\alpha}_2 &= 0, & \hat{\beta}_1 + \hat{\beta}_2 &= 0, \\ \hat{\gamma}_{11} + \hat{\gamma}_{12} &= 0, & \hat{\gamma}_{11} + \hat{\gamma}_{21} &= 0, \\ \hat{\gamma}_{21} + \hat{\gamma}_{22} &= 0, & \hat{\gamma}_{12} + \hat{\gamma}_{22} &= 0. \end{aligned} \tag{3.4}$$

Note that only three of the last four constraints are linearly independent, so that only five independent constraints are being applied. With these constraints we can solve the four equations

$$\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} = y_{ij}$$

to give

$$\begin{aligned}\hat{\mu} &= \bar{y}_{..}, \\ \hat{\alpha}_i &= \bar{y}_{i..} - \bar{y}_{..}, \quad \hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..}, \\ \hat{\gamma}_{ij} &= y_{ij} - \bar{y}_{i..} - \bar{y}_{.j} + \bar{y}_{..}.\end{aligned}\tag{3.5}$$

Thus $\hat{\mu}$ is the average of the four observations, $\hat{\alpha}_i$ is the deviation of the i th row mean from the grand mean, and $\hat{\beta}_j$ is a similar deviation for column means. The interaction parameter $\hat{\gamma}_{ij}$ is the deviation of y_{ij} , the linear predictor for that cell, from one based on the addition of main effects, $\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$.

A second set of constraints that lacks symmetry, but is in some ways simpler, is that used by the computer program GLIM, namely

$$\hat{\alpha}_1 = \hat{\beta}_1 = \hat{\gamma}_{11} = \hat{\gamma}_{12} = \hat{\gamma}_{21} = 0.\tag{3.6}$$

More generally, the parameter estimates for the first level of each factor, the first row and column of each two-factor interaction term, and so on, are set equal to zero. With this choice of constraints, the top left-hand corner cell is taken as the baseline, and the estimated parameters are

$$\begin{aligned}\hat{\mu} &= y_{11}, \\ \hat{\alpha}_2 &= y_{21} - y_{11}, \quad \hat{\beta}_2 = y_{12} - y_{11}, \\ \hat{\gamma}_{22} &= y_{22} - y_{12} - y_{21} + y_{11}.\end{aligned}\tag{3.7}$$

The remaining estimates are zero on account of the constraints. Note that if $\hat{\gamma}_{ij} = 0$, but not otherwise, the $\hat{\alpha}$ -contrasts and the $\hat{\beta}$ -contrasts given by formulae (3.5) and (3.7) are identical. If further, $\hat{\alpha}_i = \hat{\beta}_j = 0$, then the $\hat{\mu}$ s also become identical. These properties are consequences of the marginality relations among the terms in the two-factor model.

We stress again that constraints such as (3.4) and (3.6) are not a part of the model, but merely a convention whereby unique values for estimates of the intrinsically aliased parameters can be produced. For fitting, testing and so on, only estimable combinations are relevant, and those combinations are independent of the constraint system imposed.

3.5.3 *Extrinsic aliasing*

The aliasing patterns considered so far have resulted from intrinsic characteristics of the model formula rather than from particular idiosyncrasies of the data observed. However, aliasing can also occur because the particular covariate vectors observed happen to contain linear dependencies. Suppose we have two factors with three levels each, but data are observed for only 5 of the nine possible combinations as shown below.

Factor	level	B		
		1	2	3
A	1	x	x	
	2	x	x	
	3			x

Because of the configuration of the observed factor levels, the dummy vector for α_3 is identical to the dummy vector for β_3 . In a complete design, the main-effect subspaces for factors A and B have only a single dimension in common, but here they have a two-dimensional space in common.

The additional aliasing observed is a consequence of the fact that the table of observed factor levels can be split into two disconnected portions, of sizes 2×2 and 1×1 . If we move one of the occupied cells to produce the following configuration,

Factor	level	B		
		1	2	3
A	1	x	x	
	2	x		
	3		x	x

the aliasing disappears along with the disconnectedness. This example shows how extrinsic aliasing depends on the particular values of covariates in the observed data, in contrast to intrinsic aliasing, which is a property of the model formula alone.

3.5.4 Functional relations among covariates

Covariates may be functionally related without being linearly related. The most familiar example is polynomial regression, in which a linear predictor such as

$$\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3,$$

contains the power terms x , x^2 and x^3 . Provided that more than three distinct x -values are observed, the covariates x , x^2 and x^3 are linearly independent. Thus there is no aliasing of parameters. Nonetheless, there is usually an implied ordering of terms that must be respected in fitting polynomial regression models.

Looking first at the terms β_0 and $\beta_1 x$, we must ask when it makes sense not to use the sequence β_0 , $\beta_0 + \beta_1 x$ in model fitting, but to use instead the reverse sequence in which $\beta_1 x$ is fitted first without the intercept. For the latter procedure to make sense, $x = 0$ must correspond to a special point on the scale at which η must be zero. Though this may sometimes happen, there is usually no strong reason for paying special attention to a particular value of x . In agricultural field experiments with fertilizers, for example, there is invariably some small amount of the relevant nutrient already present in the soil, so that zero fertilizer applied does not mean that no nutrient is available to the plant. Thus, zero is not a special point in this example.

Consider next the relationship between the terms $\beta_1 x$ and $\beta_2 x^2$. To fit the terms β_0 and $\beta_2 x^2$ without including $\beta_1 x$ implies that the maximum (or minimum) of the response occurs at $x = 0$, i.e. exclusion of the linear term implies that $x = 0$ is a special point on the scale. For if the x -scale might equally well be measured by $x + c$ as by x , the response $\beta_0 + \beta_2 x^2$ becomes

$$\beta_0 + \beta_2(x + c)^2 = (\beta_0 + \beta_2c^2) + 2\beta_2cx + \beta_2x^2,$$

and a linear term appears with coefficient $2\beta_2c$. Ordinarily there is no reason to suppose that the turning point of the response is at a specified point on the x -scale, so that the fitting of $\beta_2 x^2$ without the linear term is usually unhelpful.

A further example, involving more than one covariate, concerns the relation between a cross-term such as $\beta_{12}x_1x_2$ and the corresponding linear terms $\beta_1 x_1$ and $\beta_2 x_2$. To include the former in

a model formula without the latter two is equivalent to assuming that the point $(0, 0)$ is a col or saddle-point of the response surface. Again there is usually no reason to postulate such a special property for the origin, so that the linear terms must be included with the cross-term. Likewise, the inclusion of quadratic terms $\beta_{11}x_1^2, \beta_{22}x_2^2$ without the cross-term implies that the elliptical contours of constant response are oriented parallel to the axes. Again, there is usually no reason to expect such behaviour in practice, and all second-degree terms should normally be entered into the model simultaneously, assuming, of course, that the linear terms are already present. Thus the relationships among polynomial terms are very similar to those among factors and interactions. This functional marginality is not a true marginality in the sense of section 3.5.1 because no linear dependencies among covariates are involved. Nevertheless, in a similar way, it does impose constraints on the order in which terms should be introduced into a model.

3.6 Estimation

3.6.1 *The maximum-likelihood equations*

Maximum likelihood is the principal method of estimation used for all generalized linear models. For Normal errors, the log likelihood, l , based on n observations is given by

$$-2l = n \log(2\pi\sigma^2) + \sum_{i=1}^n (y_i - \mu_i)^2 / \sigma^2.$$

For fixed σ^2 , known or unknown, maximization of l is equivalent to minimization of the sum of squares

$$\sum (y - \mu)^2$$

for variation in μ . If, in addition, the model is assumed to be linear, we have

$$\eta_i = \mu_i = \sum_{j=1}^p x_{ij}\beta_j.$$

Differentiating with respect to β_j and equating the derivative to zero gives estimating equations in the form

$$\sum_i x_{ij}(y_i - \hat{\mu}_i) = 0 \quad \text{for } j = 1, \dots, p, \tag{3.8}$$

where the fitted means are given by

$$\hat{\mu}_i = \hat{\eta}_i = \sum x_{ij}\hat{\beta}_j.$$

A useful way of looking at the equations (3.8) is that the p linear combinations of the observations $\sum_i x_{ij}y_i$, $j = 1, \dots, p$ are set equal to the corresponding linear combinations of the fitted values, namely $\sum_i x_{ij}\hat{\mu}_i$. To state the same thing in an equivalent way, the vector of residuals with components $y_i - \hat{\mu}_i$ is orthogonal to the columns of the model matrix \mathbf{X} , so that

$$\mathbf{X}^T(\mathbf{y} - \hat{\boldsymbol{\mu}}) = \mathbf{0}.$$

In particular, if \mathbf{X} is the incidence matrix for the main-effects model in a two-way classification, $\mathbf{X}^T\mathbf{y}$ is the set of observed marginal totals. Maximum-likelihood estimation for this Normal-theory model then corresponds to finding fitted values satisfying the model that have the same marginal totals as those observed.

3.6.2 Geometrical interpretation

Fitting by ordinary least squares has a simple geometrical interpretation. The data vector \mathbf{y} may be regarded as a point in n -dimensional Euclidean space. For any given value of the parameter vector $\boldsymbol{\beta}$, the vector of fitted values $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ is a point in the same space. As $\boldsymbol{\beta}$ varies over all possible values it might take, $\boldsymbol{\mu}$ traces out a linear subspace or hyperplane called the *solution locus*. If \mathbf{y} falls on the solution locus, the observed values can be reproduced exactly by the model. Ordinarily, however, the observed data point \mathbf{y} does not lie on the solution locus and no value of $\boldsymbol{\beta}$ reproduces the data exactly. If $\boldsymbol{\mu}$ represents a point on the solution locus then $\sum(y_i - \mu_i)^2$ is just the squared Euclidean distance between the observed vector \mathbf{y} and $\boldsymbol{\mu}$. Maximizing the likelihood is then equivalent to choosing the point $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ that is nearest to the observed \mathbf{y} in the sense of minimum Euclidean distance.

To illustrate this geometrical construction, consider the model whose components satisfy $\eta_i = x_i\beta$, with only one covariate and one parameter. The solution locus is the set of all vectors $\mathbf{x}\beta$ for $-\infty < \beta < \infty$, i.e. all points on the line through the origin in R^n in the direction \mathbf{x} (Fig. 3.3). The point on the solution locus that

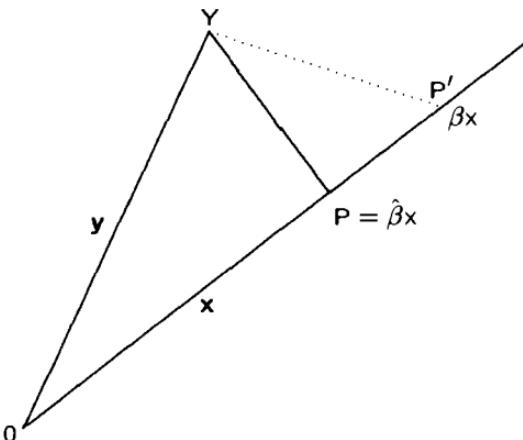


Fig. 3.3. Least squares: the geometry for one parameter.

is nearest to \mathbf{y} is found by dropping a perpendicular YP onto $\mathbf{x}\beta$. The coordinates of P are $\mathbf{x}\hat{\beta}$ where $\hat{\beta}$ is the maximum-likelihood estimate of β . The vector $YP = \mathbf{y} - \mathbf{x}\hat{\beta}$ is called the residual vector. The condition that OP and PY should be orthogonal, expressed algebraically, is

$$\mathbf{x}^T(\mathbf{y} - \mathbf{x}\hat{\beta}) = 0,$$

i.e.

$$\hat{\beta} = \frac{\mathbf{x}^T \mathbf{y}}{\mathbf{x}^T \mathbf{x}}.$$

The *fitted vector*, or vector of *fitted values*, OP , is the *orthogonal projection* of \mathbf{y} on the space \mathbf{x} .

3.6.3 Information

Further insight into the fit can be obtained by considering how the goodness-of-fit statistic, considered as a function of β , varies with β . Let P' be an arbitrary point $\mathbf{x}\beta$ on the solution locus as shown in Fig. 3.3. Then in the triangle YPP' we have

$$(\mathbf{y} - \mathbf{x}\beta)^T(\mathbf{y} - \mathbf{x}\beta) = (\mathbf{y} - \mathbf{x}\hat{\beta})^T(\mathbf{y} - \mathbf{x}\hat{\beta}) + (\hat{\beta} - \beta)\mathbf{x}^T\mathbf{x}(\hat{\beta} - \beta),$$

expressing the Pythagorean relationship among the sides of the triangle YPP' . If we plot the squared length of YP' , which

measures the discrepancy of the data from an arbitrary point of the solution locus, as a function of the parameter β , we obtain a parabola with its minimum at $\beta = \hat{\beta}$, the maximum-likelihood estimate. The minimum discrepancy is $D_{\min} = (\mathbf{y} - \mathbf{x}\hat{\beta})^T(\mathbf{y} - \mathbf{x}\hat{\beta})$, as shown in Fig. 3.4.

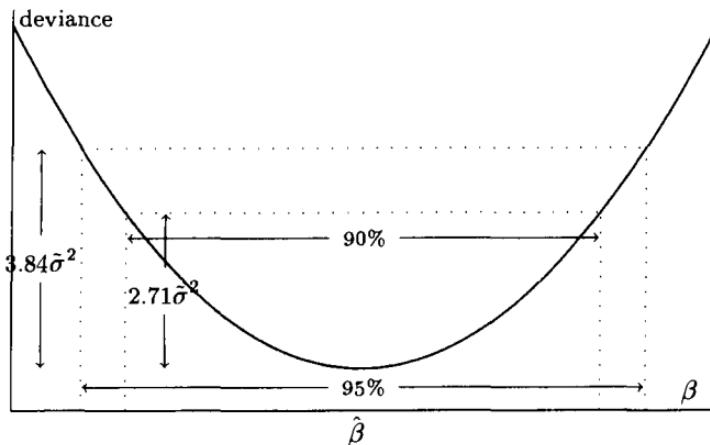


Fig. 3.4. Information curve for one parameter, together with approximate 90% and 95% confidence intervals.

The second derivative at the minimum, as indeed elsewhere for a parabola, is given by $\mathbf{x}^T \mathbf{x}$. If we now restore the dispersion parameter σ^2 , which divided the sum of squares, the second derivative becomes $\mathbf{x}^T \mathbf{x} / \sigma^2$. This is known as the *Fisher information* for β . If the Fisher information, or curvature, is large, the parabola is steep-sided, so that small changes in β away from $\hat{\beta}$ produce large changes in the discrepancy or deviance. In other words, β is well determined by the data. By contrast, if the Fisher information for β is small, the parabola is rather flat and β is not well-determined by the data.

The Fisher information for β is the ratio of two quantities. The numerator depends only on the model matrix, i.e. on the values of the covariates in the model, and not at all on the response values. The denominator depends only on the error variance of the response. The inverse information gives the theoretical sampling variance of the estimate $\hat{\beta}$, i.e. $\text{var } \hat{\beta} = \sigma^2 / (\mathbf{x}^T \mathbf{x})$. Ordinarily, σ^2 is unknown and an estimate is required, either from replicate observations for the same x , or from the residual sum of squares

after fitting an adequate model. The usual unbiased estimate is

$$\tilde{\sigma}^2 = s^2 = D_{\min}/(n - p)$$

where p is the number of covariates in the model and D_{\min} is the minimized discrepancy or deviance.

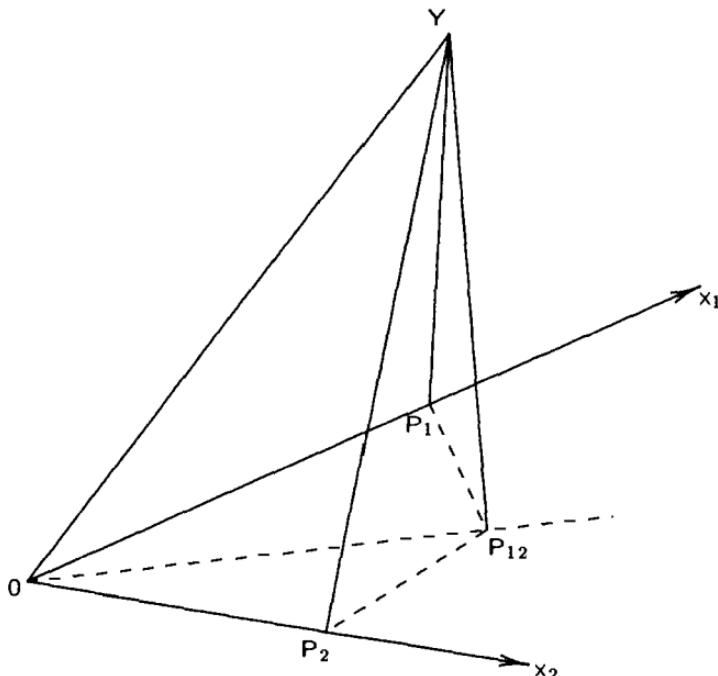


Fig. 3.5. *Least squares: the geometry for two positively correlated covariates.*

3.6.4 A model with two covariates

If there are two covariates \mathbf{x}_1 and \mathbf{x}_2 , say, then the solution locus is the plane in R^n defined by the points $\mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2$ for varying values of β_1 and β_2 . The process of obtaining fitted values for the model

$$\boldsymbol{\eta} = \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2$$

is represented geometrically by the dropping of a perpendicular from the data point \mathbf{y} onto the $(\mathbf{x}_1, \mathbf{x}_2)$ plane. Figures 3.5, 3.6 and

3.7 show the geometry connecting the fits of the single-term models $\mathbf{x}_1\beta_1$ and $\mathbf{x}_2\beta_2$ with the model containing both covariates. In these diagrams, \mathbf{x}_1 and \mathbf{x}_2 are respectively positively correlated (i.e. make an acute angle), negatively correlated (i.e. make an obtuse angle), and uncorrelated (i.e. make a right angle).

The points P_1 , P_2 and P_{12} are respectively the feet of the perpendiculars from \mathbf{y} onto the \mathbf{x}_1 -line, the \mathbf{x}_2 -line and the $(\mathbf{x}_1, \mathbf{x}_2)$ -plane. The angles $\widehat{OP_1P_{12}}$ and $\widehat{OP_2P_{12}}$, are both right angles because $\widehat{OP_2Y}$, $\widehat{OP_{12}Y}$ and $\widehat{P_2P_{12}Y}$ are all right angles by definition. Consequently P_1 is also the projection of P_{12} onto the \mathbf{x}_1 -line. Similarly, P_2 is the projection of P_{12} onto the \mathbf{x}_2 -line. We can thus express the projection of \mathbf{y} on the $(\mathbf{x}_1, \mathbf{x}_2)$ -plane in the two forms

$$(OP_{12})^2 = (OP_1)^2 + (P_1P_{12})^2 = (OP_2)^2 + (P_2P_{12})^2.$$

The interpretation of these squared lengths is as follows:

$(OP_1)^2$ = sum of squares for x_1 before x_2 ,

$(OP_2)^2$ = sum of squares for x_2 before x_1 ,

$(OP_{12})^2$ = sum of squares for x_1 and x_2 .

In addition,

$(P_1P_{12})^2$ = sum of squares for x_2 adjusted for x_1 ,

$(P_2P_{12})^2$ = sum of squares for x_1 adjusted for x_2 ,

$(OY)^2$ = total sum of squares,

$(YP_{12})^2$ = residual sum of squares after fitting x_1 and x_2 .

The words ‘before’ and ‘after’ are often replaced by ‘ignoring’ and ‘eliminating’ respectively.

Corresponding to the two sequences of fitting we have analyses of variance whose geometrical interpretations are

$$\begin{aligned}(OY)^2 &= (OP_1)^2 + (P_1P_{12})^2 + (P_{12}Y)^2 \\ \text{total} &= (x_1 \text{ before } x_2) + (x_2 \text{ after } x_1) + \text{residual},\end{aligned}$$

and

$$\begin{aligned}(OY)^2 &= (OP_2)^2 + (P_2P_{12})^2 + (P_{12}Y)^2 \\ \text{total} &= (x_2 \text{ before } x_1) + (x_1 \text{ after } x_2) + \text{residual},\end{aligned}$$

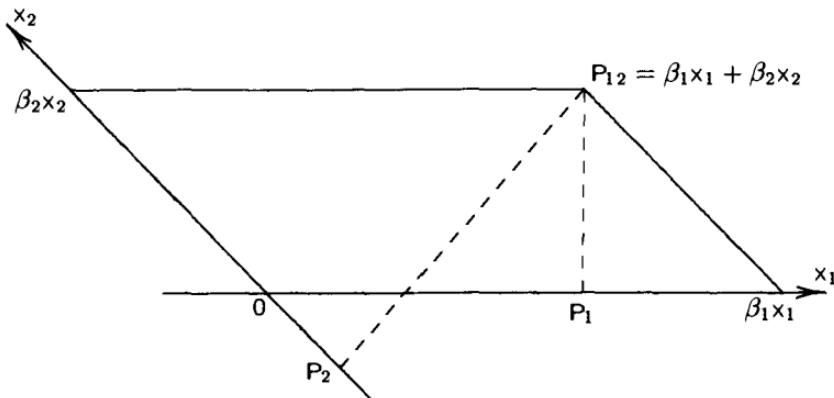


Fig. 3.6. Least squares projections for two negatively correlated covariates: P_1 is the projection on x_1 alone, P_2 is the projection on x_2 alone, and P_{12} is the projection on the joint space.

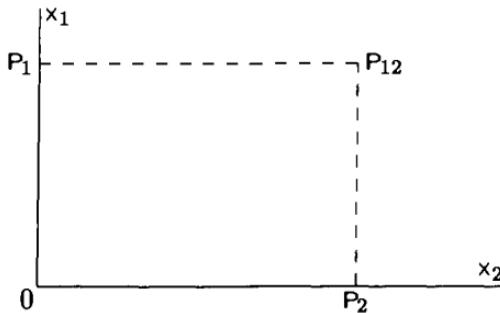


Fig. 3.7. Least squares projections for two orthogonal covariates.

In terms of the parameter estimates we have

$$\begin{aligned}(OP_1) &= \mathbf{x}_1 b_1, \\ (OP_2) &= \mathbf{x}_2 b_2, \\ (OP_{12}) &= \mathbf{x}_1 \hat{\beta}_1 + \mathbf{x}_2 \hat{\beta}_2,\end{aligned}$$

where b_1 and b_2 are the estimates for the single-term models and $\hat{\beta}_1$ and $\hat{\beta}_2$ for the joint model.

There are several important special cases:

1. \mathbf{y} is coplanar with $(\mathbf{x}_1, \mathbf{x}_2)$, so that Y and P_{12} coincide. The residual vector is then null and the joint model gives a perfect fit.

2. \mathbf{x}_1 and \mathbf{x}_2 are orthogonal (Fig. 3.7). The three feet of the perpendiculars, P_1 , P_2 and P_{12} form a rectangle with O , so that $(OP_1) = (P_2P_{12})$ and $(OP_2) = (P_1P_{12})$. The order of fitting the terms in the joint model is then irrelevant, and there is just one analysis of variance. Sums of squares and parameter estimates are unaffected by the order in which terms are entered into the model.
3. \mathbf{y} is orthogonal to \mathbf{x}_1 . Then b_1 is zero for a single-term model, but the estimate $\hat{\beta}_1$ in the joint model is not zero unless \mathbf{x}_1 and \mathbf{x}_2 are orthogonal.

3.6.5 The information surface

If P is an arbitrary point $\mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2$ on the solution locus, then from the relation among the total sum of squares, the residual sum of squares and the regression sum of squares,

$$(YP)^2 = (YP_{12})^2 + (P_{12}P)^2,$$

we obtain

$$\begin{aligned} & (\mathbf{y} - \mathbf{x}_1\beta_1 - \mathbf{x}_2\beta_2)^T(\mathbf{y} - \mathbf{x}_1\beta_1 - \mathbf{x}_2\beta_2) = \\ & (\mathbf{y} - \mathbf{x}_1\hat{\beta}_1 - \mathbf{x}_2\hat{\beta}_2)^T(\mathbf{y} - \mathbf{x}_1\hat{\beta}_1 - \mathbf{x}_2\hat{\beta}_2) \\ & + (\hat{\beta}_1 - \beta_1)^2\mathbf{x}_1^T\mathbf{x}_1 + 2(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2)\mathbf{x}_1^T\mathbf{x}_2 + (\hat{\beta}_2 - \beta_2)^2\mathbf{x}_2^T\mathbf{x}_2. \end{aligned}$$

Note that $\hat{\beta}_1$ and $\hat{\beta}_2$ are the estimates in the joint fit of \mathbf{x}_1 and \mathbf{x}_2 simultaneously. The first term on the right of the above equation is the residual sum of squares from the least squares fit to both covariates. This term does not depend on (β_1, β_2) , but only on \mathbf{y} . The second term measures the squared distance of the arbitrary point P , determined by (β_1, β_2) , from the point of best fit, $(\hat{\beta}_1, \hat{\beta}_2)$. The contours of this latter term, considered as a function of (β_1, β_2) , are similar, similarly situated ellipses centered at $(\hat{\beta}_1, \hat{\beta}_2)$ as shown in Fig. 3.8.

The second derivative matrix of the function with respect to (β_1, β_2) is

$$2 \begin{pmatrix} \mathbf{x}_1^T\mathbf{x}_1 & \mathbf{x}_1^T\mathbf{x}_2 \\ \mathbf{x}_2^T\mathbf{x}_1 & \mathbf{x}_2^T\mathbf{x}_2 \end{pmatrix},$$

which, apart from the factor $\sigma^2/2$, is the Fisher information matrix for (β_1, β_2) .

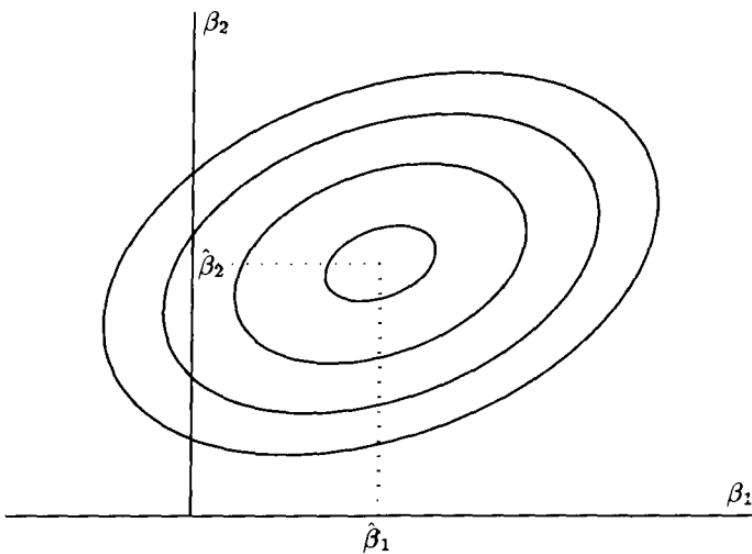


Fig. 3.8. *Least squares: contours of the information surface for two parameters.*

3.6.6 Stability

The point P_{12} depends only on the data vector \mathbf{y} and its relation to the space spanned by $\mathbf{x}_1, \mathbf{x}_2$. Any pair of vectors that spans the same space, for example $\mathbf{x}_1 + \mathbf{x}_2$ and $\mathbf{x}_1 - \mathbf{x}_2$, gives rise to the same projection of \mathbf{y} . However the identification of the point P_{12} by the coefficients $(\hat{\beta}_1, \hat{\beta}_2)$, namely $P_{12} = \mathbf{x}_1\hat{\beta}_1 + \mathbf{x}_2\hat{\beta}_2$, depends heavily on the particular pair of vectors chosen as a basis for the subspace. If θ , the angle between the vectors \mathbf{x}_1 and \mathbf{x}_2 in R^n , is small, then the coefficients $\hat{\beta}_1, \hat{\beta}_2$ are more sensitive to small perturbations of the data than either b_1 or b_2 , the coefficients in the single-term models. That is to say P_{12} itself is unstable in the sense that while small perturbations of \mathbf{x}_1 and \mathbf{x}_2 may produce correspondingly small perturbations of P_1 and P_2 , together they may produce a large perturbation of P_{12} . Thus, a small perturbation of one or both covariates may have a big effect on the space spanned by the two vectors. Consequently, the allocation of variation in Y to \mathbf{x}_1 and \mathbf{x}_2 in the joint regression is sensitive to perturbations of either covariate.

When θ , the angle between \mathbf{x}_1 and \mathbf{x}_2 , is small the information matrix for β_1, β_2 is nearly singular because its determinant is equal

to

$$\|\mathbf{x}_1\|^2 \|\mathbf{x}_2\|^2 \sin^2 \theta.$$

The log-likelihood contours in the (β_1, β_2) plane are now ellipses having one principal axis very long compared to the other. In fact the ratio of the lengths of the principal axes behaves like $1/\sin^2 \theta$. As a consequence, large changes in β_1, β_2 in the direction of the longer axis produce small changes in the likelihood, while similar changes in the perpendicular direction have a large effect.

3.7 Tables as data

It is common to find generalized linear models being fitted not to the original data expressed in data-matrix form, but to data that have already been summarized in the form of a multi-way table. In the process of tabulation, the y -values for units having the same levels of the classifying factors are added together to form a table of totals: parallel tabulation of a vector of 1s gives the associated table of counts showing how many units contribute to each cell total. Division of totals by their associated counts gives a table of means. For continuous data, it is usually this table of means that will be analysed, with the associated counts acting as prior weights. In surveys, however, it is often the counts themselves that are of interest. Suitable methods of analysis for such counts are described in Chapters 4 to 6. Section 5.2.3 emphasizes the duality between models that treat cell averages or scores as the response with the counts acting as weights, and models that treat the observed count as the response, with the cell averages used for generating contrasts.

Broadly speaking, the process of fitting generalized linear models to data in the form of tables is similar to that described previously for data in the form of a data matrix. The following points are not peculiar to tabular data, but they are most often encountered in that context.

3.7.1 Empty cells

When any variate, be it a continuous measurement or an integer-valued variable, is discretized and tabulated as described above, a table of averages and an associated table of counts is formed. The table of averages is different in one important respect from the table

of counts, namely in the significance of zeros. Table 8.1, giving the average value of insurance claims together with the number of claims, is a case in point. For some purposes the value of the claims is of interest, while for others the number of claims might be the most interesting response. It so happens that no teenagers who owned ten-year-old cars of types C or D made claims against the company. These are genuine zeros indicating either that such drivers are unusually careful or few in number or both. As far as the average claim is concerned, however, these are not to be treated as zeros, but rather as 'empty cells' contributing no information whatever about averages. We cannot infer from the absence of claims in these categories that, if and when a claim occurs, its value will be small. Consequently, we use the term 'empty cell' rather than 'structural zero' because there is no suggestion of any value, let alone zero.

It is important to distinguish two varieties of empty cell, namely *necessarily empty cells* and *accidentally empty cells*. Necessarily empty cells occur when some combination of levels of the factors is *a priori* impossible. Simple examples are the class of pregnant males or a self-fertilized cross from a self-sterile variety of plant. When all possible crosses are made between varieties of a self-sterile species, the diagonal cells, corresponding to the selfs, are all necessarily empty. If some varieties are cross-incompatible there may be off-diagonal necessarily empty cells as well. When a model is fitted to a table of associated counts, the necessarily empty cells must not be included as data. For other tables such as Table 8.1, they cannot be included in any analysis because there is no value for that cell. Ordinarily it makes no sense to compute fitted values for necessarily empty cells by extrapolation from the non-empty cells.

An accidentally empty cell is one for which the combination of factor levels is possible, but the combination happens not to occur in the observed data. The empty cells in Table 8.1 are of this type. For this type of empty cell it does usually make sense to compute fitted values by extrapolation from the non-empty cells.

Table 6.2 contains both accidentally empty and necessarily empty cells.

It has been proposed (see, for example, Urquhart and Weeks, 1978) that models fitted to tables should not involve the population means of accidentally empty cells, on the grounds that the data give no information about such means. This proposal would imply that

an additive model of the form

$$A + B$$

for a two-way table should not be fitted if any cells are accidentally empty. The point has been discussed by Nelder (1982), who argues that such a rule is unnecessarily restrictive.

3.7.2 Fused cells

It sometimes happens that the position of a unit in a table is not known uniquely, though it is known to belong to one of a subset of cells. Examples of this phenomenon occur in tables classified by genetic factors when several distinct genotypes produce the same phenotype, which is what is observed. For the analysis we have just the total for the set of cells, fused into a single observable cell. Fused cells may also occur when the individual cells were potentially observable, but for some reason the level of one or more factors was recorded with less precision than intended. The occurrence of fused cells results in an obvious loss of information, and utilization of the data they contain may require prior knowledge of the relative frequency of occurrence in the unobserved component cells. Such knowledge is often available for genetic data.

3.8 Algorithms for least squares

For the linear models discussed in this chapter the estimation procedure requires us to minimize the quadratic form

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

with respect to the components of $\boldsymbol{\beta}$. Equating the derivative to zero produces the *normal equations*

$$(\mathbf{X}^T \mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y}. \quad (3.9)$$

If \mathbf{X} has full rank these equations have a unique solution, namely $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$. If \mathbf{X} is rank-deficient, either because of intrinsic aliasing among factors or for some other reason, we may

replace the inverse of $\mathbf{X}^T \mathbf{X}$ by any generalized inverse. The solution is then not unique, but all estimable contrasts among the β s are independent of the choice of inverse (Pringle and Rayner, 1971). When multiplied by the dispersion parameter σ^2 , the generalized inverse also produces correct variances and covariances for these contrasts.

There are two classes of numerical methods for solving equations (3.9). In the first method $\mathbf{X}^T \mathbf{X}$ is formed explicitly and subsequent computations are performed on this matrix. The second class of methods focuses on the matrix \mathbf{X} and attempts to simplify equations (3.9) by suitably factoring \mathbf{X} . In both cases it is usual to express both \mathbf{y} and the columns of \mathbf{X} about their means. Within each class there are further sub-divisions, which are described below.

In the interests of efficient bookkeeping, both algebraic and numerical, it is convenient in much of the discussion that follows to imagine the observation vector \mathbf{y} appended as an additional column to \mathbf{X} . Thus any row operations applied to \mathbf{X} are considered also to be applied to \mathbf{y} . In addition, the extended information matrix $\mathbf{X}^T \mathbf{X}$ now consists of the sums of squares and products of \mathbf{y} and the p covariates.

3.8.1 Methods based on the information matrix

The two most common methods that operate on the matrix $\mathbf{X}^T \mathbf{X}$ are Gaussian elimination and Choleski decomposition. We discuss these in turn.

A modern form of Gaussian elimination, due to Beaton (1964), uses a symmetric sweep operator. This operator, when applied to the k th row and column of a positive-definite symmetric matrix \mathbf{A} , will be denoted by S_k . The effect of S_k is to transform the components of \mathbf{A} from a_{ij} to

$$\begin{aligned} a_{ij} &\rightarrow a_{ij} - \frac{a_{ik}a_{jk}}{a_{kk}}; & i \neq k, j \neq k, \\ a_{ik} &\rightarrow \frac{a_{ik}}{|a_{kk}|}; & i \neq k, \\ a_{kj} &\rightarrow \frac{a_{kj}}{|a_{kk}|}; & j \neq k, \\ a_{kk} &\rightarrow -\frac{1}{a_{kk}}. \end{aligned}$$

With this definition it is then easily shown that $S_k S_k \mathbf{A} = \mathbf{A}$. In other words, a second application of the symmetric sweep restores the original matrix. The statistical interpretation of the symmetric sweep is as follows. Let $\mathbf{A} = \mathbf{X}^T \mathbf{X}$ be a $p \times p$ matrix of sums of squares and products of the variates $\mathbf{x}_1, \dots, \mathbf{x}_p$. Suppose that the sweeps S_1, \dots, S_k have been applied to the first $k < p$ rows and columns of \mathbf{A} . Following this series of sweeps, \mathbf{A} has been reduced to the form shown in Fig. 3.9, in which only the lower triangle is displayed. The component matrix \mathbf{R} now holds the residual sum-of-squares-and-products matrix for the unswept variates $\mathbf{x}_{k+1}, \dots, \mathbf{x}_p$ after regressing them on $\mathbf{x}_1, \dots, \mathbf{x}_k$. The rows of the matrix \mathbf{B} are the regression coefficients of these formal linear regression equations, while \mathbf{V} is the unscaled covariance matrix for these regressions.

Note that if the final row and column of \mathbf{A} contain the sums of squares and products of the response, then sweeping all but the final row and column gives minus the inverse information matrix $-\mathbf{V}$, bordered by the vector of regression coefficients $\mathbf{B} = \hat{\beta}$, and the residual sum of squares, \mathbf{R} , now a scalar.

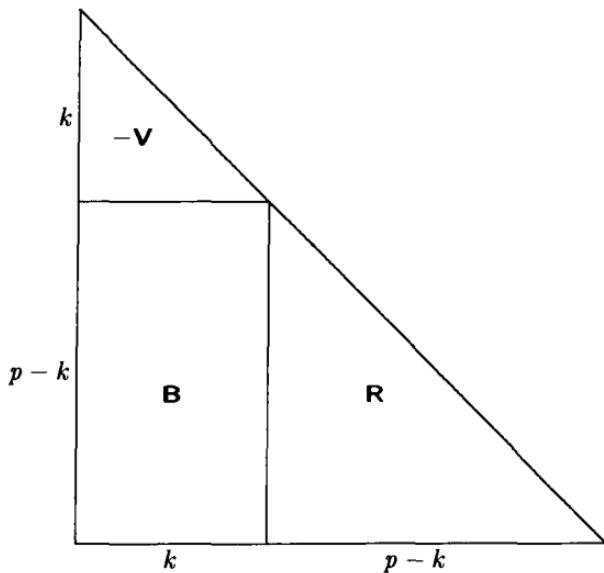


Fig. 3.9. *The matrix of sums and squares and products after the symmetric sweep has been applied to the first k rows and columns.*

If the original $\mathbf{X}^T \mathbf{X}$ is exactly or nearly singular, this will usually

show up during the sweeping process by the appearance of a pivot or diagonal element that is small compared to its original value (Clarke, 1982). For if \mathbf{x}_{k+1} is expressible as a linear combination of $\mathbf{x}_1, \dots, \mathbf{x}_k$, then the residual sum of squares for \mathbf{x}_{k+1} after linear regression on $\mathbf{x}_1, \dots, \mathbf{x}_k$ is zero. If there is a near singularity then the residual sum of squares for \mathbf{x}_{k+1} is small compared to its original total sum of squares. Statistically, this exact collinearity or near singularity means that there is either no information or little information about the corresponding parameter, given that the first k terms are included in the model. If there is an exact singularity, the term may be omitted from the model without affecting the span of \mathbf{X} . Algebraically, this is equivalent to setting the estimate to zero with variance zero. In the algorithm such rows/columns are not swept, but are marked to show their special status. If such a term is subsequently to be removed from the model, then again no sweep is done; however, if other terms involved in the collinearity are subsequently removed, the pivot for the first term may again become substantial. Should this occur, the term can again be included in the model and a reliable estimate of the parameter obtained.

The second method that operates on the information matrix is the Choleski decomposition, which aims to find a lower-triangular $p \times p$ matrix \mathbf{L} that satisfies

$$\mathbf{X}^T \mathbf{X} = \mathbf{L} \mathbf{L}^T.$$

\mathbf{L} is thus a square-root matrix of $\mathbf{X}^T \mathbf{X}$. Details of algorithms for computing \mathbf{L} can be found in the books by Chambers (1977) and Healy (1986). Having computed \mathbf{L} , the inversion of $\mathbf{X}^T \mathbf{X}$ is accomplished via the formula

$$(\mathbf{X}^T \mathbf{X})^{-1} = (\mathbf{L}^{-1})^T \mathbf{L}^{-1}.$$

There is a simple inversion algorithm for triangular matrices, and the inversion can be combined with subsequent multiplication by the transpose. Again, generalized inverses can be obtained by setting any row of \mathbf{L} with a small pivot to zero.

The condition number of a matrix is a measure of closeness to singularity, large values indicating near-singularity; algorithms that use the matrix $\mathbf{X}^T \mathbf{X}$ directly suffer from the disadvantage

that the condition number of $\mathbf{X}^T \mathbf{X}$ is the square of the condition number of \mathbf{X} . Large values of the condition number can give rise to numerical instability from rounding errors in the calculations. For this reason the second class of methods is designed to avoid the formation of $\mathbf{X}^T \mathbf{X}$ altogether.

3.8.2 Direct decomposition methods

Direct decomposition methods operate on the model matrix \mathbf{X} directly. The aim is to decompose \mathbf{X} into the product of an $n \times n$ orthogonal matrix \mathbf{Q} and an $n \times p$ matrix \mathbf{R} of the form

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{pmatrix},$$

where \mathbf{R}_1 is $p \times p$ upper triangular.

The statistical interpretation of the decomposition is as follows. If \mathbf{y} is the observation vector with mean $\mathbf{X}\beta$ and variance $\sigma^2\mathbf{I}$, we may make an orthogonal transformation to new variables \mathbf{u} defined by $\mathbf{u} = \mathbf{Q}^T \mathbf{y}$, where \mathbf{Q} is the $n \times n$ orthogonal matrix described above. The mean and variance of the new variables are

$$\begin{aligned} E(\mathbf{U}) &= \mathbf{Q}^T E(\mathbf{Y}) = \mathbf{Q}^T \mathbf{X}\beta = \mathbf{Q}^T \mathbf{Q}\mathbf{R}\beta \\ &= \mathbf{R}\beta = \begin{pmatrix} \mathbf{R}_1\beta \\ \mathbf{0} \end{pmatrix}, \\ \text{cov}(\mathbf{U}) &= \mathbf{Q}^T \mathbf{I} \mathbf{Q} \sigma^2 = \mathbf{I} \sigma^2. \end{aligned}$$

Thus the last $(n - p)$ components of \mathbf{U} have zero expectation, and so give no information about β . Hence the least-squares solution reduces to equating the first p components of \mathbf{u} , here denoted by \mathbf{u}_1 , to their expectation as a function of $\hat{\beta}$. Thus we arrive at

$$\mathbf{R}_1 \hat{\beta} = \mathbf{u}_1$$

which is easily solved because \mathbf{R}_1 is upper triangular.

It is not necessary to compute \mathbf{Q} explicitly because, if \mathbf{y} is appended to \mathbf{X} , the sequence of operations that takes \mathbf{X} to \mathbf{R} also transforms $\{\mathbf{X} : \mathbf{y}\}$ to $\{\mathbf{R} : \mathbf{u}\}$. The first p rows of this augmented matrix give the coefficients in the above equation for $\hat{\beta}$. The sum of squares of the last $n - p$ components of \mathbf{u} gives the residual sum of squares.

Note that

$$\mathbf{R}_1^T \mathbf{R}_1 = \mathbf{R}^T \mathbf{R} = \mathbf{R}^T \mathbf{Q}^T \mathbf{Q} \mathbf{R} = \mathbf{X}^T \mathbf{X},$$

so that \mathbf{R}_1 is the upper triangular Choleski square-root matrix of $\mathbf{X}^T \mathbf{X}$. In fact \mathbf{R}_1 is the transpose of \mathbf{L} as described in the previous section.

Three methods for finding \mathbf{Q} and \mathbf{R} are associated with the names of

Householder: \mathbf{Q} is a product of reflections,

Givens: \mathbf{Q} is a product of rotations,

and Gram-Schmidt: successive orthogonalization.

A Householder reflection takes the matrix form

$$\mathbf{I} - 2\mathbf{v}\mathbf{v}^T,$$

where \mathbf{v} is an n -vector of unit length ($\mathbf{v}^T \mathbf{v} = 1$). It is possible, given a vector \mathbf{x} , to choose \mathbf{v} so that, after reflection, all components of \mathbf{x} except the first are zero. In the Householder decomposition, \mathbf{v}_1 is chosen to reduce the first column of \mathbf{X} to this form. A second vector \mathbf{v}_2 is chosen to reduce components 3 to n of the second column to zero. Since elements 2 to n of the first column are already zero, this reflection leaves them unaffected. The process continues on components $j+1$ to n of column j for $j = 1, \dots, p-1$. If $\mathbf{Q}_j = \mathbf{I} - 2\mathbf{v}_j\mathbf{v}_j^T$, then the matrix

$$\mathbf{Q} = \mathbf{Q}_{p-1} \cdots \mathbf{Q}_2 \mathbf{Q}_1$$

is the product of $p-1$ reflections. It has the property that

$$\mathbf{Q}^T \mathbf{X} = \mathbf{R},$$

where \mathbf{R} has the form described at the beginning of this section.

Givens rotations are planar rotations through an angle θ . A single rotation is applied to two components of a vector, corresponding to a rotation through an angle θ in the plane of these two components. The angle is chosen to make one of the components equal to zero. We denote by G_{ijk} the rotation that replaces the

*i*th and *j*th rows of \mathbf{X} by linear combinations that make the *k*th element in row *j* equal to zero. Then the sequence

$$((G_{kjk}, j = k + 1 \text{ to } n), k = 1 \text{ to } p)$$

annihilates the elements by columns, like Householder, but setting one element to zero at a time. The sequence

$$((G_{kjk}, k = 1 \text{ to } \min(j-1, p)), j = 2 \text{ to } n)$$

annihilates by rows. The latter sequence is more useful if \mathbf{X} is more easily processed by rows than by columns. The idea of rotations goes back to Jacobi, but the Givens sequence of rotations ensures that previously formed zeros remain zero after subsequent rotations.

The Gram-Schmidt method relies on successive orthogonalization of the columns of \mathbf{X} . The preferred algorithm is due to Björck (1967) and begins by forming

$$\begin{aligned}\mathbf{q}_1 &= \mathbf{x}_1 / \|\mathbf{x}_1\| \\ \mathbf{q}_j &= \mathbf{x}_j - (\mathbf{q}_1^T \mathbf{x}_j) \mathbf{q}_1; \quad j = 2, \dots, p.\end{aligned}$$

The first row of \mathbf{R} is given by

$$r_{1i} = \mathbf{q}_1^T \mathbf{x}_i.$$

This process is then repeated, using at the second stage the vectors $\mathbf{q}_2, \dots, \mathbf{q}_p$ in place of $\mathbf{x}_1, \dots, \mathbf{x}_p$, and so on.

Statistically, we regress columns 2 to *p* of \mathbf{x} on column 1 and replace them by the vectors of residuals. At the second stage, columns 3 to *p* are regressed on column 2, and so on for successive stages. The matrix \mathbf{Q} thus formed is $n \times p$ with orthonormal columns and \mathbf{R} is $p \times p$ upper triangular. The first *j* columns of \mathbf{Q} span the same space as the first *j* columns of \mathbf{X} for *j* = 1, ..., *p*. Calculating the regression of \mathbf{y} on the orthogonalized covariates is easy because of the orthogonality.

The direct decomposition methods are somewhat less convenient for updating models than the symmetric sweep method. For example, if a column of \mathbf{X} is deleted, all the columns of \mathbf{Q} and \mathbf{R} to the right of the deleted column must be recalculated. Details of updating with the Givens algorithm are given by Clarke (1981).

3.8.3 Extension to generalized linear models

In Chapter 2 it was shown that estimation in generalized linear models can be accomplished by iteratively weighted least squares. In order to adapt the algorithms discussed above we must (a) allow iteration and (b) introduce weights and an adjusted dependent variate, both of which ordinarily vary from one iteration to the next. Introduction of weights is straightforward in principle. In the algorithms that use the information matrix we replace $\mathbf{X}^T \mathbf{X}$ by $\mathbf{X}^T \mathbf{W} \mathbf{X}$, where \mathbf{W} is the diagonal matrix of weights, while in the QR algorithms we replace \mathbf{X} by $\mathbf{W}^{1/2} \mathbf{X}$. The attractiveness of QR methods and the Choleski decomposition is then greatly reduced because a new decomposition must be computed at each cycle of the iteration.

There are two new features of the algorithms, the first related to unbounded parameter estimates and the second to *pseudo-aliasing*. Infinite parameter estimates arise most commonly in the fitting of a log-linear model if one or more fitted values are zero. The link function $\hat{\eta} = \log \hat{\mu}$ implies that one or more of the β s contributing to $\hat{\eta}$ must be negatively infinite. Cells in the table for which $\hat{\mu} = 0$ must also have $y = 0$, so that there is no contribution to the deviance. Such cells are best omitted from the fit. Similar effects arise in models for proportions where the fitted proportion is either 0 or 1.

The second feature, pseudo-aliasing, can arise when the changing weights in an iterative fit produce so little information on a parameter that the pivot in the information matrix falls below the tolerance set by the algorithm. However, removal of the covariate will now be found to increase the deviance sharply, showing that it is really required in the model. In this respect, pseudo-aliasing is quite different from true aliasing. The fit obtained immediately before the algorithm detected the apparent aliasing is often a reasonable assessment of the fit of the model.

Numerical iteration requires a definition of effective convergence of the process. If all parameter estimates are finite, straightforward monitoring of the progress of the deviance is sufficient. Convergence is usually rapid, though divergence may occasionally occur for ill-fitting models using non-canonical links. If one or more components are infinite, convergence as measured by the deviance may be slow. It is usually best in these circumstances to halt the

iteration after about 10 cycles and to inspect the estimates at that stage. Usually it is clear which components are tending to $\pm\infty$. If necessary, action can then be taken to omit a subset of the data or to modify the model or both.

3.9 Selection of covariates

Apart from the choice of link function and error distribution, the problem of modelling reduces to finding one or more appropriate parsimonious sets of covariates corresponding to a model matrix \mathbf{X} of order $n \times p$. As elsewhere it is important that the final model or models should make sense physically: at a minimum, this usually means that interactions should not be included without main effects nor higher-degree polynomial terms without their lower-degree relatives. Furthermore, if the model is to be used as a summary of the findings of one out of several studies bearing on the same phenomenon, main effects should usually be included whether significant or not. Strict adherence to this policy makes it easier to compare the results of various studies and helps to avoid the apparent conflicts that occur when different fitted models with different sets of terms are used in each study. The danger is that a term with a coefficient of +1, say, might be rejected in one study because it was insignificant, while in a second study the same term might have a numerically similar coefficient that was highly significant. The fitted models are then different and apparently in conflict, while in reality the two studies are highly concordant.

We discussed in Chapter 1 the justification for seeking a parsimonious model to represent a set of data. Parsimony implies, among other things, that covariates having no detectable effect on the response should ordinarily be excluded from the linear predictor. In a survey concerned with the incidence of a particular disease, large numbers of covariates may be available, describing perhaps age structures of the populations involved, their dietary and smoking habits, aspects of the environment, and so on. The selection of a useful set of covariates from such a large set of possible covariates to form a parsimonious model is then a non-trivial exercise. There are both statistical and computing problems, the latter arising from the ‘combinatorial explosion’ that occurs when all possible subsets of covariates are to be tested for inclusion in the model.

On the statistical side, the problem is that of defining the balance to be struck between two opposing effects of including a new term in the model. The good effect may be a reduction in the discrepancy between the data and the fitted values. The bad effect is that, unless there is good prior knowledge that the covariate has a non-negligible influence on the response, inclusion of the covariate usually complicates the model and statements of conclusions derived from it. At one extreme, if the addition of a single covariate reduces the residual mean square to, say, one third of its original value we have no hesitation in including it in the model, particularly if the number of residual degrees of freedom is large. At the other extreme, if such an addition causes no reduction, by the principle of Occam's razor, parsimony wins and we exclude it. It is the intermediate cases that cause problems. For example, if there is a large number of irrelevant covariates, then statistical accidents will produce a few false positives that appear to influence the response.

The usual F -statistic for the reduction in deviance or sum of squares is the basis of most criteria for selection of covariates. In order to exclude irrelevant terms the significance level for acceptance is set at a low level, but it must not be set so low that important terms are thereby excluded. Another approach is based on the idea of providing the best prediction of response values over a set of covariate values, and yet another uses a criterion based on a measure of information. Atkinson (1981b) points out that all of these procedures can be represented (in our notation) as special cases of minimizing the expression

$$Q = D + \alpha q\phi, \quad (3.10)$$

where D is the deviance function, q is the number of estimable parameters in the linear predictor, ϕ is the dispersion parameter, and α is either constant or a function of n . The idea behind the second term is to penalize the inclusion of unnecessary covariates in the model. Use of Q presumes a knowledge of ϕ . For Poisson and binomial models without over-dispersion, $\phi = 1$, but otherwise ϕ is usually unknown. Even with counted data it is often wise to assume that over-dispersion is present unless the data or prior information indicate otherwise. For details see Chapter 4. When comparing a sequence of models we have the option of replacing ϕ

in (3.10) either by a common estimate for all models in the sequence or by separate estimates $\hat{\phi}_i$ derived from the fit of each model in turn. To make the comparison fair, it seems best in practice to use a single estimate, usually derived from the most complex model in the sequence.

If two models in a nested sequence differ only by the inclusion of one covariate, then the use of the 5% point of the F - or t -distribution as the criterion for model selection is equivalent to setting $\alpha \approx 4$ in (3.10), assuming adequate residual degrees of freedom for estimating ϕ . The most common criteria based on errors of prediction (Akaike, 1969; Mallows, 1973) lead to $\alpha = 2$. For Normal-theory linear models an argument based on maximum posterior probabilities leads to $\alpha(n) = \log(n)$. Atkinson (1981b) suggests that the range $\alpha = 2$ to 6 may provide ‘a set of plausible initial models for further analysis’.

The computing problem, which ignores any relationships that may exist among the covariates, may be specified as follows: ‘find the best s subsets of size r among the covariates’. If k , the total number of covariates available, is small, say $k \leq 12$, the best subsets for each r from 1 to $k-1$ can be found by complete enumeration. For larger k , say up to 35, tree-search methods, using short-cuts, are feasible (Furnival and Wilson, 1974). Approximate methods for generating a single ‘optimum’ subset include:

1. *forward selection*, whereby at each stage the best unselected covariate satisfying the selection criterion is added until no further candidates remain;
2. *backward elimination*, which begins with the full set and eliminates the worst covariates one by one until all remaining covariates are necessary; and
3. *stepwise regression* (Efroymson, 1960), which combines the two previous procedures, following backward elimination by forward selection until both fail to change the model.

In GLIMPSE, (Wolstenholme, O’Brien and Nelder, 1988), a knowledge-based front-end for GLIM (Payne 1986), a model selection strategy is used that results in a tree of candidate models, with the extreme node of each branch forming a possible parsimonious model. The basic step in the algorithm has as input a *kernel*, which contains terms already accepted as necessary, and a set of *free terms*, whose status is currently uncertain. The maximal model

contains the kernel and all the free terms. For each free term two F -statistics are calculated, a forward F -statistic formed by adding it to the kernel, and a backwards F -statistic formed by removing it from the maximal model. The two F -statistics are classified by a decision rule as being either large or small, and action is then taken as shown in the table below.

<i>Forward F</i>	<i>Backward F</i>	<i>Action</i>
<i>large</i>	<i>large</i>	<i>add term to kernel</i>
<i>large</i>	<i>small</i>	<i>leave as free term</i>
<i>small</i>	<i>large</i>	<i>leave as free term</i>
<i>small</i>	<i>small</i>	<i>discard term</i>

The process is begun with a kernel of terms considered necessary a priori and continues until the set of free terms is either null or unchanging. If it is null we have a unique preferred model; if not we add each remaining free term in turn to the kernel, producing a branching in the tree and repeat the basic step. Further branching may then occur, but eventually the final node on each branch will contain a null set of free terms.

Unthinking use of automatic selection procedures has frequently, and rightly, been criticized. Clearly the notion that a particular subset is optimum is hard to sustain when many other subsets of similar size produce almost equally good fits. It may also happen that some covariates are much more expensive to measure than others, and this is not allowed for in a criterion based on purely statistical considerations. Expense may be an important consideration if the goal is to produce good forecasts at reasonable cost. However, if the goal is to understand the mechanism by which the process is generated, cost is largely irrelevant. A further criticism of automatic selection procedures is that they do not take into account the marginality constraints among factors nor functional marginality among polynomial terms (Section 3.5). Further, certain factors, e.g. treatment and block effects, would often be kept in the model whether statistically significant or not.

Further modification of these selection procedures is required for models that require iterative solution, because of the presence of weights and adjusted dependent variates, both of which are functions of the fitted values, and so change as the fitted model changes. The amount of computing is reduced by using an approximate

procedure, which appears to work well in practice; this involves doing the full iterative fit for a large but well-fitting model, and afterwards following the same algorithms as for the non-iterative case. In other words, the weights and adjusted dependent variate are kept fixed throughout. The fully iterated fit may then be recalculated at intervals as a check on the approximation.

3.10 Bibliographic notes

For a history of least squares, see a series of papers by Harter, summarized in Harter (1976).

Among the many texts on linear models, see Atkinson, (1985), Draper and Smith (1981), Mosteller and Tukey (1977), Plackett (1960), Searle (1971), Seber (1977), Sprent (1969) and Williams (1959).

Model formulae for linear predictors were introduced by Nelder (1965a,b) and developed by Wilkinson and Rogers (1973).

Aliasing, marginality and the role of constraints are discussed by Nelder (1977).

For numerical methods for least squares, see Lawson and Hanson (1974), Gentleman (1974a,b), Healy (1986), Chambers (1977), Thisted (1988) and Wampler (1979).

The statistical problems of covariate selection are discussed by Akaike (1973), Mallows (1973), Stone (1977), and summarized by Atkinson (1981b). For a discussion of the computing aspects of covariate selection, see Efroymson (1960), Beale (1970), Stewart (1973), Furnival and Wilson (1974) and Jennrich (1977). Lawless and Singhal (1978) deal explicitly with generalized linear models.

3.11 Further results and exercises 3

3.1 Use the following data to familiarize yourself with a suitable linear regression program (S, GLIM or Minitab should be fine).

1. Plot y against x_1 . Comment on any strong relationships or unusual features of the plot.
2. Plot y against x_2 . Comment on any strong relationships or unusual features of the plot.

x_1	x_2	y	x_1	x_2	y
2.23	9.66	12.37	3.04	7.71	12.86
2.57	8.94	12.66	3.26	5.11	10.84
3.87	4.40	12.00	3.39	5.05	11.20
3.10	6.64	11.93	2.35	8.51	11.56
3.39	4.91	11.06	2.76	6.59	10.83
2.83	8.52	13.03	3.90	4.90	12.63
3.02	8.04	13.13	3.15	6.96	12.46
2.14	9.05	11.44			

3. Plot x_1 against x_2 . Comment on any strong relationships or unusual features of the plot.
4. Regress y on x_1 . Plot the residuals against x_2 .
5. Regress y on x_2 . Plot the residuals against x_1 .
6. Regress y on x_1 and x_2 simultaneously. Compare the coefficients obtained in the joint regression with those in the marginal regressions. Compare the (multiple) correlation coefficients.

[Hamilton, 1987].

3.2 Write out explicitly the model matrix corresponding to a randomized blocks design with three treatments in each of four blocks.

3.3 Suppose that the three treatments mentioned in the previous exercise actually denote increasing concentrations of a chemical used for weed control. Re-parameterize the model using linear and quadratic treatment contrasts. Write out the corresponding model matrix.

3.4 Suppose that two factors A and B with levels i and j respectively have the property that observations are possible only when $i \geq j$. Now define a new factor C with levels $k = i - j + 1$. Assuming that A and B have five levels each, and that each possible combination is observed once, answer the following:

1. Which, if any, of the following models are equivalent:

$$A + B, \quad A + C, \quad B + C \quad \text{and} \quad A + B + C ?$$

2. What are the ranks of the model matrices in part 1?
3. Answer part 1 assuming instead that A , B and C are quantitative covariates taking values i , j and k respectively.

It may be helpful to construct the required factors and variates on the computer and to fit the models to computer-generated data.

For an example of such a triangular arrangement of factors, see Cox and Snell (1981, p. 58).

3.5 Suppose that \mathbf{x}_1 and \mathbf{x}_2 are positively correlated variates in a two-variable linear regression model that includes the intercept. Show that the regression coefficients $\hat{\beta}_1, \hat{\beta}_2$ are negatively correlated. Express the statistical correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$ in terms of the angle between the vectors \mathbf{x}_1 and \mathbf{x}_2 in R^p .

3.6 In section 3.6.6 an expression is given for the determinant of the Fisher information matrix in a two-variable regression model with no intercept. Derive the corresponding expression when the intercept is included.

3.7 Show that the sweep operator, as defined in section 3.8.1, is self-inverse. What advantages accrue from this property?

3.8 The sweep operator has the property that a new variable can be added to an existing regression equation with a single sweep. This produces automatically the updated parameter estimates, the new residual sum of squares and the Fisher information matrix with minimal computational effort. Yet few statistical programs make full use of this property. Discuss briefly the organizational difficulties involved in making full use of the sweep algorithm on an interactive computer system.

3.9 Suppose that A is a factor with 4 levels whose effects are denoted by α_r . Write out explicitly the model matrix \mathbf{X} , of order 6×4 , corresponding to the algebraic expression

$$\eta_{rs} = \alpha_r - \alpha_s, \quad \text{for } r < s.$$

You may assume that all 6 combinations $1 \leq r < s \leq 4$ are observed. What is the rank of \mathbf{X} ? Does the constant vector lie in the column space of \mathbf{X} ? Under what circumstances might such a model formula arise?

3.10 Let A, B, C, D be four factors each with four levels, having the exclusion property that no two factors can simultaneously have the same level. There are thus only $4!$ possible factor combinations

instead of the more usual $4^4 = 256$ unrestricted combinations. What are the ranks of the models

$$A + B + C + D, \quad A + B + C, \quad \text{and} \quad (A + B + C + D)^{**2}?$$

You may assume that all $4!$ permutations are observed.

For what purposes might such models be used?

Table 3.1 *Ascorbic acid concentrations of samples of snap-beans after a period of cold storage.*

Temp. °F	Weeks of storage				Total
	2	4	6	8	
0	45	47	46	46	184
10	45	43	41	37	166
20	34	28	21	16	99
Total	124	118	108	99	449

3.11 The data in Table 3.1, taken from Snedecor and Cochran (1967, p.354), were obtained as part of an experiment to determine the effects of temperature and storage time on the loss of ascorbic acid in snap-beans. The beans were all harvested under uniform conditions at the Iowa Agricultural Experiment Station before eight o'clock one morning. They were prepared and quick-frozen before noon the same day. Three packages were assigned at random to each temperature and storage-time combination. The sum of the three ascorbic acid determinations is shown in the Table.

Suppose for the purpose of model construction that the ascorbic acid concentration decays exponentially fast, with a decay rate that is temperature-dependent. In other words, for a given storage temperature T , the expected concentration after time t (measured in weeks) is $\mu = E(Y) = \exp\{\alpha - \beta_T t\}$. The initial concentration, $\exp(\alpha)$ is assumed in this model to be independent of the storage temperature. Express the above theory as a generalized linear model, treating temperature as a factor and storage time as a variate.

[The above model is unusual in that it contains an interaction between time and temperature, but no main effect of temperature. By design, the concentrations are equal at time zero.]

Estimate the times taken at each of the three temperatures for the ascorbic acid concentration to be reduced to 50% of its original value. Consider carefully how you might construct confidence intervals for this half-life.

Compare your analysis with the factorial decomposition model, using orthogonal polynomial contrasts, as described by Snedecor and Cochran (1967, pp. 354–8).

The mean squared error for individual packets, obtained from the replicates, was 0.706 on 24 degrees of freedom. Is this value consistent with the above analyses?

CHAPTER 4

Binary data

4.1 Introduction

4.1.1 *Binary responses*

Suppose that for each individual or experimental unit, the response, Y , can take only one of two possible values, denoted for convenience by 0 and 1. Observations of this nature arise, for instance, in medical trials where, at the end of the trial period, the patient has either recovered ($Y = 1$) or has not ($Y = 0$). Clearly, we could also have intermediate values associated with different degrees of recovery (see Chapter 5), but for the moment that possibility will be ignored. We may write

$$\text{pr}(Y_i = 0) = 1 - \pi_i; \quad \text{pr}(Y_i = 1) = \pi_i \quad (4.1)$$

for the probabilities of ‘failure’ and ‘success’ respectively.

In most investigations, whether they be designed experiments, surveys or observational studies, we have, associated with each individual or experimental unit, a vector of covariates or explanatory variables (x_1, \dots, x_p) . In a designed experiment, this covariate vector usually comprises a number of indicator variables associated with blocking and treatment factors, together with quantitative information concerning various aspects of the experimental material. In observational studies, the vector of covariates consists of measured variables thought likely to influence the probability of a positive response. The principal objective of a statistical analysis, therefore, is to investigate the relationship between the response probability $\pi = \pi(\mathbf{x})$ and the explanatory variables $\mathbf{x} = (x_1, \dots, x_p)$. Often, of course, a subset of the x s is of primary importance, but due allowance must be made for any effects that might plausibly be attributed to the remaining covariates.

4.1.2 Covariate classes

Suppose that, for the i th combination of experimental conditions characterized by the p -dimensional vector (x_{i1}, \dots, x_{ip}) , observations are available on m_i individuals. In other words, of the $N = m_1 + m_2 + \dots + m_n$ individuals under study, m_i share the covariate vector (x_{i1}, \dots, x_{ip}) . These individuals are said to form a covariate class. If the recorded covariates are factors each having a small number of levels, the number of distinct covariate vectors, n , is often considerably fewer than the number of individuals, N , under study. In these circumstances, it is more convenient and more efficient in terms of storage to list the data by the n covariate classes than by the N individuals.

Table 4.1 Alternative ways of presenting the same data

(a) Data listed by subject No.			(b) Data listed by covariate class		
Subject No.	Covariate (x_1, x_2)	Response Y	Covariate (x_1, x_2)	Class size m	Response Y
1	1, 1	0	1, 1	2	1
2	1, 2	1	1, 2	3	2
3	1, 2	0	2, 1	1	0
4	2, 1	0	2, 2	1	1
5	2, 2	1			
6	1, 2	1			
7	1, 1	1			

To take an over-simplified example, suppose that a clinical trial is undertaken to compare the effectiveness of a newly developed surgical procedure with current standard techniques. In order to recruit sufficient patients in a reasonable period, the trial is conducted at two hospitals ($x_1 = 1, 2$) (with different surgeons and ancillary staff). In each hospital, patients judged by the protocol as suitable for recruitment are assigned at random to one of the two surgical procedures ($x_2 = 1, 2$). One month into the study, seven patients have been recruited. These patients are listed by patient number in Table 4.1a and by covariate class in Table 4.1b. Provided that only these two covariates are recorded, the number of covariate classes remains equal to four however many patients are recruited. Thus the efficiency of Table 4.1b increases as the number of patients grows.

Usually covariate classes are formed for convenience of tabulation and to make the major effects of interest easier to detect by visual scanning. In forming covariate classes from the original data, information concerning the serial order of the subjects is lost, so that we cannot, for example, reconstruct Table 4.1a from Table 4.1b. If serial order of patients is considered irrelevant, no information is lost when the data are grouped by covariate class. On the other hand, the possibility of detecting whether serial order is relevant is also lost in forming covariate classes. Thus, the claim that no information is lost must be regarded either as a tautology or as a self-fulfilling statement. In the example discussed in the previous paragraph, the possibility of a learning effect on the part of the surgeon or his staff should be considered. Such an effect, if present, cannot be detected from an analysis of the grouped data in the form displayed in Table 4.1b, but might possibly be detectable as a serial trend in an analysis of the original data in Table 4.1a.

When binary data are grouped by covariate class, the responses have the form $y_1/m_1, \dots, y_n/m_n$, where $0 \leq y_i \leq m_i$ is the number of successes out of the m_i subjects in the i th covariate class. The vector of covariate class sizes $\mathbf{m} = (m_1, \dots, m_n)$ is called the binomial index vector or binomial denominator vector. Ungrouped data, or data listed by individual subjects, can be considered as a special case for which $m_1 = \dots = m_n = 1$.

The distinction between grouped and ungrouped data is important for at least two reasons.

1. Some methods of analysis appropriate to grouped data, particularly those involving Normal approximation, are not applicable to ungrouped data.
2. Asymptotic approximations for models applied to grouped data can be based on either of two distinct asymptotes, either $\mathbf{m} \rightarrow \infty$ or $N \rightarrow \infty$. Only the latter limit is appropriate for ungrouped data.

4.1.3 Contingency tables

Suppose that the data are indexed by three explanatory factors, A having a levels, B having b levels and C having c levels. Among the subjects observed, therefore, there are at most $a \times b \times c$ covariate classes. There may in fact be fewer covariate classes than this maximum either because, by chance, one or more covariate

classes were not observed or because certain factor combinations are physically or logically impossible (Section 3.7.1). For each covariate class, the number of successes and the number of failures is counted. Such data may be presented as a $2 \times a \times b \times c$ table of counts called a *contingency table*. For instance, the data in Table 4.1 give rise to the $2 \times 2 \times 2$ table

		$y = 0$	$y = 1$			$y = 0$	$y = 1$
		$x_2 = 1$	$x_2 = 2$	$x_1 = 1$	$x_2 = 1$	$x_2 = 2$	$x_1 = 2$
$x_1 = 1$	$x_2 = 1$	1	1	$x_1 = 2$	$x_2 = 1$	1	0
	$x_2 = 2$	1	2		$x_2 = 2$	0	1

In constructing models for such data, one is normally interested in how the response probabilities are affected by the covariates rather than how the individuals are distributed over covariate classes. If the prevalence of the various covariate classes were of interest, it would be appropriate to analyse the marginal table summed over the response. The methods discussed in Chapter 6 may be helpful here. However, if the response probabilities are of interest, it is best to regard the marginal table of covariate class totals, \mathbf{m} , as fixed, whether or not they were predetermined by design. The formal analysis then proceeds conditionally on the observed value of the vector \mathbf{m} .

4.2 Binomial distribution

4.2.1 Genesis

The binomial distribution arises naturally in a number of contexts where the observations Y are non-negative counts bounded above by a fixed value. Two ways in which it can arise are now described.

Suppose that Y_1, Y_2 are independent Poisson random variables with means μ_1, μ_2 . It follows that the total, $Y_1 + Y_2$, has the Poisson distribution with mean $\mu_1 + \mu_2$. The conditional distribution of Y_1 given that $Y_1 + Y_2 = m$ is given by

$$\text{pr}(Y_1 = y | Y_1 + Y_2 = m) = \binom{m}{y} \pi^y (1-\pi)^{m-y}, \quad y = 0, 1, \dots, m \quad (4.2)$$

where $\pi = \mu_1 / (\mu_1 + \mu_2)$. This conditional distribution depends only on the ratio of the Poisson means and not on $\mu_1 + \mu_2$. Details of

the derivation are given in Exercise 4.4. The notation $Y \sim B(m, \pi)$ means that Y has the binomial distribution (4.2) with *index* m and *parameter* π .

The Bernoulli distribution (4.1) is a nearly degenerate case of the binomial distribution for which $m = 1$. A second and more natural way in which the binomial distribution arises in practice is as the sum of independent homogeneous Bernoulli trials. For instance, in the formation of covariate classes as discussed in section 4.1.2, if the individuals so grouped are homogeneous and independent, the totals have the binomial distribution with the same parameter. Details of this and related derivations are given in Exercise 4.2.

4.2.2 Moments and cumulants

The cumulants of the binomial distribution (4.2) are most easily derived using the representation of the binomial as a sum of independent homogeneous Bernoulli random variables whose distribution is given in (4.1). The moment generating function of (4.1) is

$$M_Y(\xi) = E \exp(\xi Y) = 1 - \pi + \pi \exp(\xi). \quad (4.3)$$

Hence, the cumulant generating function is

$$K_Y(\xi) = \log M_Y(\xi) = \log\{1 - \pi + \pi \exp(\xi)\}.$$

It follows that the moment generating function of $Y_1 + \dots + Y_m$, is

$$\{1 - \pi + \pi \exp(\xi)\}^m$$

and that the cumulant generating function is

$$m \log\{1 - \pi + \pi \exp(\xi)\}. \quad (4.4)$$

From the Taylor expansion of (4.4), we find that the first four cumulants are

$$\begin{aligned} \kappa_1 &= m\pi, & \kappa_3 &= m\pi(1 - \pi)(1 - 2\pi), \\ \kappa_2 &= m\pi(1 - \pi), & \kappa_4 &= m\pi(1 - \pi)\{1 - 6\pi(1 - \pi)\}. \end{aligned}$$

All cumulants of Y have the form $m \times$ polynomial in π . The expressions for the moments are more complicated except in the

special case $m = 1$ for which all moments of all orders are equal to π .

It is sometimes of interest in applications to examine what happens to the distribution of the sum when the Bernoulli components lack homogeneity. Suppose, therefore, that $Y = Y_1 + \dots + Y_m$, where $Y_i \sim B(1, \pi_i)$ and the components are independent. From the additive property of cumulants, it is readily seen that the cumulants of Y are

$$\kappa_1 = \sum \pi_i = m\bar{\pi},$$

$$\kappa_2 = \sum \pi_i(1 - \pi_i) = m\bar{\pi}(1 - \bar{\pi}) - (m - 1)k_2(\pi) \leq m\bar{\pi}(1 - \bar{\pi}),$$

$$\kappa_3 = \sum \pi_i(1 - \pi_i)(1 - 2\pi_i),$$

$$\kappa_4 = \sum \pi_i(1 - \pi_i)\{1 - 6\pi_i(1 - \pi_i)\},$$

where $k_2(\pi) = \sum(\pi_i - \bar{\pi})^2/(m - 1)$ is the ‘sample variance’ of the π s. Evidently, the sample variance of Y is deflated relative to the binomial variance. This calculation appears to contradict the common intuition that lack of homogeneity should increase variability rather than decrease it. The reason for the apparent contradiction is that the calculations just given are not relevant to the problem of heterogeneity as usually met. In practice, it is usually known only that there is variability among the π s: the complete set of values π_1, \dots, π_m is rarely known. A more relevant calculation, therefore, is to regard π_1, \dots, π_m as independent random variables with mean $\bar{\pi}$. It is then easily shown that, whatever the distribution of π_i , $Y_i \sim B(1, \bar{\pi})$. Hence, the sum $Y = Y_1 + \dots + Y_m$ is distributed as $B(m, \bar{\pi})$ and the binomial distribution is recovered.

For an extension of these calculations, see Exercises 4.6 and 4.17.

4.2.3 Normal limit

From the cumulant generating function (4.4), we see that, for large m , all cumulants of Y are of order m . Consequently, the cumulants of the standardized random variable

$$Z = \frac{Y - m\pi}{\sqrt{m\pi(1 - \pi)}}$$

are 0, 1, $O(m^{-1/2})$, $O(m^{-1})$ and so on, decreasing in half powers of m . For $r \geq 2$, the r th cumulant of Z is $O(m^{1-r/2})$. As $m \rightarrow \infty$

for any fixed π , the cumulants of Z tend to those of the standard Normal distribution, namely $0, 1, 0, 0, \dots$. Since convergence of the cumulants implies convergence in distribution, approximate tail probabilities may be obtained from

$$\begin{aligned}\text{pr}(Y \geq y) &\simeq 1 - \Phi(z^-) \\ \text{pr}(Y \leq y) &\simeq \Phi(z^+)\end{aligned}\tag{4.5}$$

where $\Phi(\cdot)$ is the cumulative Normal distribution function, y is an integer,

$$z^- = \frac{y - m\pi - \frac{1}{2}}{\sqrt{m\pi(1 - \pi)}} \quad \text{and} \quad z^+ = \frac{y - m\pi + \frac{1}{2}}{\sqrt{m\pi(1 - \pi)}}.$$

The effect on probability calculations of the continuity correction of $\pm \frac{1}{2}$ is of order $O(m^{-1/2})$ and hence asymptotically negligible. In medium-sized samples, however, the effect of the continuity correction is appreciable and almost always improves the approximation.

An improved version of (4.5) utilizing third- and fourth-order cumulants is given in Appendix B.

The rate of convergence to Normality is governed primarily by the third cumulant and is fastest when $\pi = \frac{1}{2}$. The error incurred in using (4.5) is asymptotically $O(m^{-1/2})$ in general: if $\pi = \frac{1}{2}$ the error reduces to $O(m^{-1})$. In practice, the approximation is usually satisfactory if $m\pi(1 - \pi) \geq 2$ and if $|z^-|$ or $|z^+|$ does not exceed 2.5. Note that although the absolute error

$$\epsilon(y) = |\text{pr}(Y \geq y) - 1 + \Phi(z^-)|$$

incurred in using (4.5) is asymptotically small even for large z^- , the relative error,

$$\frac{\epsilon(y)}{\text{pr}(Y \geq y)}$$

may be quite large if z^- is large.

4.2.4 Poisson limit

Suppose that $\pi \rightarrow 0$, $m \rightarrow \infty$ in such a way that $\mu = m\pi$ remains fixed or tends to a constant. From (4.4), the cumulant generating function of Y tends to

$$\frac{\mu}{\pi} \log\{1 + \pi(\exp(\xi) - 1)\} \rightarrow \mu\{\exp(\xi) - 1\},$$

which is the cumulant generating function of a Poisson random variable with mean μ : see section 6.2. In fact, in this limit, all cumulants of Y differ from those of the Poisson distribution, $P(\mu)$, by terms of order $O(m^{-1})$. Probability calculations based on the Poisson distribution are in error by terms of the same order. By contrast, the Normal approximation, with or without the continuity correction, has an error of order $O(m^{-1/2})$.

4.2.5 Transformations

There is a large body of literature concerning transformations of the binomial and other distributions designed to achieve a specified purpose, usually stability of the variance or symmetry of the density. Such transformations are considered in Exercises 4.8–4.11. In this section, we consider two transformations, one connected with achieving approximate additivity in linear logistic models, the other concerned with Normal approximation. We consider the latter first.

Suppose that $Y \sim B(m, \pi)$ and let $\mu = m\pi$ be the mean of Y . It is shown in Appendix C that for large values of m , the cumulants of the signed deviance statistic

$$W = w(Y) = \pm [2Y \log(Y/\mu) + 2(m - Y) \log\{(m - Y)/(m - \mu)\}]^{1/2} + \frac{1 - 2\pi}{6\sqrt{m\pi(1 - \pi)}} \quad (4.6)$$

differ from those of a standard Normal random variable by terms of order $O(m^{-1})$. The sign used in (4.6) is that of $Y - \mu$ and the transformation is monotone increasing in Y . In other words, $w(Y)$ is approximately symmetrically distributed as far as this can be achieved in the discrete case. In fact, the variance of $w(Y)$ is

$$\sigma_W^2 = 1 + \frac{5 - 2\pi(1 - \pi)}{36m\pi(1 - \pi)} + O(m^{-2}).$$

The cumulants of $w(Y)/\sigma_W$ differ from those of $N(0, 1)$ by terms of order $O(m^{-3/2})$, suggesting that a Normal approximation for W ought to give accurate results.

In order to use the discrete Edgeworth approximation as presented in Appendix B, we define the continuity-corrected abscissa and the Sheppard correction as follows:

$$w^+ = w(y + \frac{1}{2})$$

$$\tau = 1 + \frac{1}{24m\pi(1-\pi)}.$$

From equation (B.3), approximate tail probabilities are given by

$$\text{pr}(Y \leq y) \simeq \Phi(w^+\tau/\sigma_W).$$

Note that the ratio of τ to σ_W is

$$\begin{aligned} \tau/\sigma_W &\simeq 1 - \frac{1 - \pi(1 - \pi)}{36m\pi(1 - \pi)} \\ &= 1 + \frac{1}{36m} - \frac{1}{36m\pi(1 - \pi)}. \end{aligned}$$

Analogous approximations are available for the right-hand tail probability. These approximations are more accurate than (4.5).

The *empirical logistic transformation* is a transformation of Y designed to achieve approximate additivity in linear logistic models. These are discussed more fully in the section that follows. Suppose therefore, that $Y \sim B(m, \pi)$ and that we require an approximately unbiased estimate of the log odds,

$$\lambda = \log\left(\frac{\pi}{1 - \pi}\right).$$

It is natural to begin by trying transformations of the form $\log[(Y + c)/(m - Y + c)]$ for some constant $c > 0$. The maximum-likelihood estimator has this form with $c = 0$ and has asymptotic bias of order $O(m^{-1})$. For the particular choice $c = \frac{1}{2}$, we have the transformation

$$Z = \log\left(\frac{Y + \frac{1}{2}}{m - Y + \frac{1}{2}}\right), \quad (4.7)$$

which has the property that

$$E(Z) = \lambda + O(m^{-2}).$$

This is known as the empirical logistic transformation (Cox, 1970). For any other choice of constant, the bias is $O(m^{-1})$: see Exercise 4.15.

Gart and Zweifel's (1967) results support the estimation of $\text{var}(Z)$ by $v = (y + \frac{1}{2})^{-1} + (m - y + \frac{1}{2})^{-1}$. The idea behind transformation is that it may be simpler to use a linear regression model for Z with weights v^{-1} rather than to use a non-linear model for the untransformed responses. This is often a simple and attractive alternative to maximum likelihood. Because the argument is asymptotic in nature, the transformation is useful only if all the binomial indices are fairly large.

4.3 Models for binary responses

4.3.1 Link functions

To investigate the relationship between the response probability π and the covariate vector (x_1, \dots, x_p) , it is convenient, though perhaps not absolutely necessary, to construct a formal model thought capable of describing the effect on π of changes in (x_1, \dots, x_p) . In practice, this formal model usually embodies assumptions such as zero correlation or independence, lack of interaction or additivity, linearity and so on. These assumptions cannot be taken for granted and should, if possible, be checked. Furthermore, the behaviour of the model should, as far as possible, be consistent with known physical, biological or mathematical laws, especially in its limiting behaviour.

Linear models play an important role in both applied and theoretical work — and with good reason. We suppose therefore that the dependence of π on (x_1, \dots, x_p) occurs through the linear combination

$$\eta = \sum_{j=1}^p x_j \beta_j \tag{4.8}$$

for unknown coefficients β_1, \dots, β_p . Unless restrictions are imposed on $\boldsymbol{\beta}$ we have $-\infty < \eta < \infty$. Thus, to express π as the

linear combination (4.8) would be inconsistent with the laws of probability. A simple and effective way of avoiding this difficulty is to use a transformation $g(\pi)$ that maps the unit interval onto the whole real line $(-\infty, \infty)$. This remedy leads to instances of generalized linear models in which the systematic part is

$$g(\pi_i) = \eta_i = \sum_{j=1}^p x_{ij} \beta_j; \quad i = 1, \dots, n. \quad (4.9)$$

A wide choice of link functions $g(\pi)$ is available. Three functions commonly used in practice are

1. the logit or logistic function

$$g_1(\pi) = \log\{\pi/(1 - \pi)\};$$

2. the probit or inverse Normal function

$$g_2(\pi) = \Phi^{-1}(\pi);$$

3. the complementary log-log function

$$g_3(\pi) = \log\{-\log(1 - \pi)\}.$$

A fourth possibility, the log-log function

$$g_4(\pi) = -\log\{-\log(\pi)\},$$

which is the natural counterpart of the complementary log-log function, is seldom used because its behaviour is inappropriate for $\pi < \frac{1}{2}$, the region that is usually of interest. All four functions can be obtained as the inverses of well-known cumulative distribution functions having support on the entire real axis. The corresponding density functions are discussed in Exercises 4.22–4.23. The first two functions are symmetrical in the sense that

$$g_1(\pi) = -g_1(1 - \pi).$$

The latter two functions are not symmetrical in this sense, but are related via

$$g_3(\pi) = -g_4(1 - \pi).$$

All four functions are continuous and increasing on $(0, 1)$.

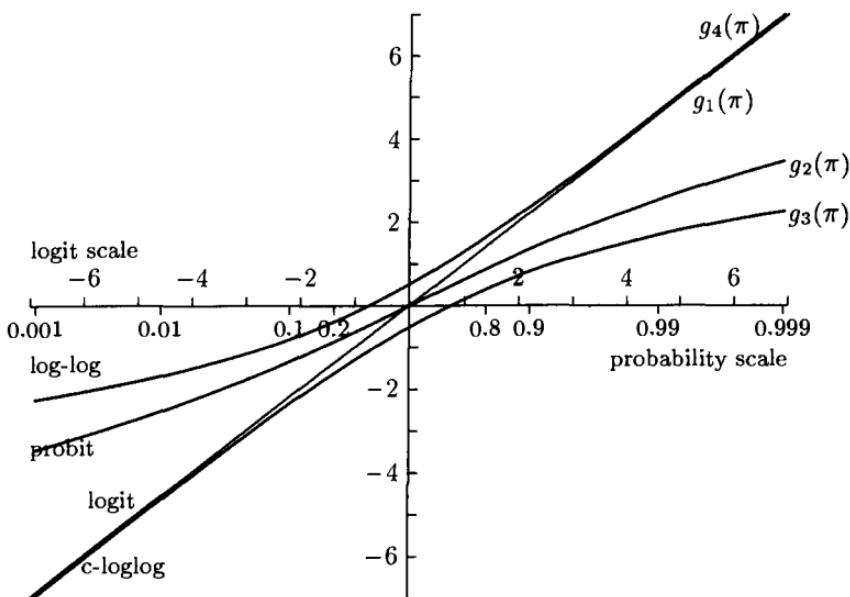


Fig. 4.1. *A graphical comparison of three link functions with the logistic function: the 45° line is the logistic function.*

Figure 4.1 compares the four functions. The logistic function is taken as the standard and $g_2(\pi)$, $g_3(\pi)$, $g_4(\pi)$ are plotted against $g_1(\pi)$ for values of π in the range 0.01 to 0.99.

The logistic and the probit function are almost linearly related over the interval $0.1 \leq \pi \leq 0.9$. For this reason, it is usually difficult to discriminate between these two functions on the grounds of goodness-of-fit; see, for example, Chambers and Cox (1967). For small values of π , the complementary log-log function is close to the logistic, both being close to $\log(\pi)$. As π approaches 1, the complementary log-log function approaches infinity much more slowly than either the logistic or the probit function. Similar comments apply to the log-log function as can be seen from Figure 4.1.

All asymptotic and approximate theory presented in this chapter applies regardless of the choice of link function. However, we shall be concerned mostly with the logistic function, not so much because of its simpler theoretical properties, but because of its simple interpretation as the logarithm of the odds ratio. Apart from this, the logistic function has one important advantage over

all alternative transformations in that it is eminently suited for the analysis of data collected retrospectively. See section 4.3.3.

4.3.2 Parameter interpretation

In order to summarize the conclusions of an analysis in an easily digested form, it is helpful to state the magnitudes of the estimated effects on an easily understood scale. The scale most suitable for this purpose is often different from the scale or link function used to achieve additivity of effects, namely $g(\pi)$. For instance, if a linear logistic model has been used with two covariates x_1 and x_2 , we have the model

$$\log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

for the log odds of a positive response. Equivalently, the model may be written in terms of the odds of a positive response, giving

$$\frac{\pi}{1 - \pi} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2).$$

Finally, the probability of a positive response is

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}.$$

This is the inverse function of $g_1(\pi)$. Assuming that x_1 and x_2 are functionally unrelated, the conclusions based on such a model may be stated as follows. The effect of a unit change in x_2 is to increase the log odds by an amount β_2 . Equivalently, but perhaps preferably, we may say that the effect of a unit change in x_2 is to increase the odds of a positive response multiplicatively by the factor $\exp(\beta_2)$. It is important here that x_1 be held fixed and not be permitted to vary as a consequence of the change in x_2 . These statements are fairly easy to comprehend because the direction and magnitude of the stated effect are unaffected by the values of x_1 and x_2 .

The corresponding statements given on the probability scale are more complicated because the effect on π of a unit change in x_2 depends on the values of x_1 and x_2 . The derivative of π with respect to x_2 is

$$\frac{\partial \pi}{\partial x_2} = \pi(1 - \pi)\beta_2.$$

Thus, a small change in x_2 has a larger effect, as measured on the probability scale, if π is near 0.5 than if π is near 0 or 1. Perhaps the simplest device to assist in the presentation of conclusions is to give the graph of

$$\pi(\eta) = \exp(\eta)/\{1 + \exp(\eta)\}$$

and to state the effect on η of changes in x_2 . The effect on the probability can then be read from the graph. This method works equally well whatever the link function used. The required inverse link functions are

$$\begin{aligned}\pi_2(\eta) &= g_2^{-1}(\eta) = \Phi(\eta), \\ \pi_3(\eta) &= g_3^{-1}(\eta) = 1 - \exp(-e^\eta), \\ \text{and } \pi_4(\eta) &= g_4^{-1}(\eta) = \exp(-e^{-\eta}).\end{aligned}$$

All of these functions are defined for $-\infty < \eta < \infty$ and increase continuously from zero at $-\infty$ to one at ∞ .

4.3.3 Retrospective sampling

One important property of the logistic function not shared by the other link functions is that differences on the logistic scale can be estimated regardless of whether the data are sampled prospectively or retrospectively. To illustrate the difference between these two sampling schemes, suppose that a population is partitioned according to two binary variables, (D, \bar{D}) referring to the presence or absence of disease, and (X, \bar{X}) referring to exposure or non-exposure to the toxin or carcinogen under investigation. Suppose that the proportions of the population in the four categories thus formed are as shown in Table 4.2.

Table 4.2 Hypothetical frequencies of disease and exposure status

		Disease status		Total
		\bar{D}	D	
Exposure status	\bar{X}	$\pi_{00} = 0.70$	$\pi_{01} = 0.02$	$\pi_{0.} = 0.72$
	X	$\pi_{10} = 0.25$	$\pi_{11} = 0.03$	$\pi_{1.} = 0.28$
	Total	$\pi_{.0} = 0.95$	$\pi_{.1} = 0.05$	1.0

In a prospective study, an exposed group of subjects is selected together with a comparable group of non-exposed individuals. The progress of each group is monitored, often over a prolonged period, with a view towards comparing the incidence of disease in the two groups. In this way, the row totals, giving the numbers of subjects in each of the exposure categories, are fixed by design. The column totals are random, reflecting the incidence of disease in the overall population, weighted according to the sizes of exposure groups in the sample.

In a retrospective study, diseased and disease-free individuals are selected — often from hospital records collected over a period of several years. In this design, the column totals are fixed by design and the row totals are random, reflecting the frequency of exposure in the population, weighted according to the sizes of the disease groups in the sample.

Considering the prospective study first, the logits for the two exposure groups are

$$\log(\pi_{01}/\pi_{00}) = -\log(35) = -3.555 \quad \text{and}$$

$$\log(\pi_{11}/\pi_{10}) = -\log(8.3) = -2.120.$$

The difference of logits is thus

$$\Delta = \log(\pi_{11}/\pi_{10}) - \log(\pi_{01}/\pi_{00}) = 1.435.$$

This difference could also be estimated by sampling retrospectively from the two disease groups \bar{D} and D because

$$\Delta = \log(\pi_{11}/\pi_{01}) - \log(\pi_{10}/\pi_{00}).$$

In fact, in the present example, the retrospective design is substantially more efficient than the prospective design. This is because the disease is rare even among those who are exposed to the toxin or carcinogen. Thus, for a prospective study to be effective, a large number of initially healthy subjects must be followed for a prolonged period in order that a sufficiently large number of subjects may eventually fall victim to the disease. In a retrospective study, on the other hand, the investigator has access via hospital records to all cases of the disease recorded over a substantial period of time. In the case of rare diseases, it is common to take a 100% sample of

the diseased individuals and to compare these with a similar sized sample of disease-free subjects. Since exposure is fairly common, ranging from 26% among those who are disease-free to 60% among those with the disease, a substantial number of exposed and non-exposed subjects will be observed both among the cases (D) and among the controls (\bar{D}).

More generally, if there are several exposure groups and other covariates, we may write the linear logistic model in the form

$$\text{pr}(D | \mathbf{x}) = \exp(\alpha + \boldsymbol{\beta}^T \mathbf{x}) / [1 + \exp(\alpha + \boldsymbol{\beta}^T \mathbf{x})] \quad (4.10)$$

for the probability of contracting the disease given that the subject has covariates \mathbf{x} . Included in \mathbf{x} is the information on the exposure category to which the individual belongs, together with other factors considered relevant to the incidence of the disease.

Model (4.10) is specified in a form appropriate for data sampled prospectively. Suppose, however that the data are sampled retrospectively. Introduce the dummy variable Z to define whether an individual is sampled or not, and denote the sampling proportions by

$$\pi_0 = \text{pr}(Z = 1 | D) \quad \text{and} \quad \pi_1 = \text{pr}(Z = 1 | \bar{D}).$$

It is essential here that the sampling proportions depend only on D and not on \mathbf{x} . We may now use Bayes's theorem to compute the disease frequency among sampled individuals who have a specified covariate vector \mathbf{x} .

$$\begin{aligned} \text{pr}(D | Z = 1, \mathbf{x}) &= \frac{\text{pr}(Z = 1 | D, \mathbf{x}) \text{pr}(D | \mathbf{x})}{\text{pr}(Z = 1 | D, \mathbf{x}) \text{pr}(D | \mathbf{x}) + \text{pr}(Z = 1 | \bar{D}, \mathbf{x}) \text{pr}(\bar{D} | \mathbf{x})} \\ &= \frac{\pi_0 \exp(\alpha + \boldsymbol{\beta}^T \mathbf{x})}{\pi_1 + \pi_0 \exp(\alpha + \boldsymbol{\beta}^T \mathbf{x})} \\ &= \frac{\exp(\alpha^* + \boldsymbol{\beta}^T \mathbf{x})}{1 + \exp(\alpha^* + \boldsymbol{\beta}^T \mathbf{x})}, \end{aligned}$$

where $\alpha^* = \alpha + \log(\pi_0/\pi_1)$. In other words, although the data have been sampled retrospectively, the logistic model (4.10) continues to apply with the same coefficients $\boldsymbol{\beta}$ but a different intercept. It follows therefore, that the logistic models described here in the context of prospective studies can be applied to retrospective studies provided that the intercept is treated as a nuisance parameter.

This derivation follows the lines of Armitage (1971) and Breslow and Day (1980). No such simple inversion exists for probit or complementary log-log models.

4.4 Likelihood functions for binary data

4.4.1 Log likelihood for binomial data

The responses y_1, \dots, y_n are assumed to be the observed values of independent random variables Y_1, \dots, Y_n such that Y_i has the binomial distribution with index m_i and parameter π_i . It is convenient initially to consider the log likelihood as a function of the n -vector $\boldsymbol{\pi} = \pi_1, \dots, \pi_n$. Subsequently, when we wish to study specific linear models such as (4.10), the log likelihood is considered as a function of the coefficients appearing in the model. Using (4.2), the log likelihood may be written in the form

$$l(\boldsymbol{\pi}; \mathbf{y}) = \sum_{i=1}^n \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i) \right]. \quad (4.11)$$

The constant function of \mathbf{y} not involving $\boldsymbol{\pi}$, namely

$$\sum \log \binom{m_i}{y_i},$$

has been omitted because it plays no role.

The systematic part of the model specifies the relation between the vector $\boldsymbol{\pi}$ and the experimental or observational conditions as summarized by the model matrix \mathbf{X} of order $n \times p$. For generalized linear models, this relationship takes the form

$$g(\pi_i) = \eta_i = \sum_j x_{ij} \beta_j; \quad i = 1, \dots, n, \quad (4.12)$$

so that the log likelihood (4.11) can be expressed as a function of the unknown parameters β_1, \dots, β_p . It is a good tactical manoeuvre, however, not to make this substitution but to keep the two expressions separate. For instance, we may wish to compare several models by adding or deleting covariates. This operation changes the set of parameters, but leaves expression (4.11) unaltered.

In the case of linear logistic models, we have

$$g(\pi_i) = \eta_i = \log\{\pi_i/(1 - \pi_i)\} = \sum_j x_{ij}\beta_j.$$

Substitution into (4.11) gives

$$l(\boldsymbol{\beta}; \mathbf{y}) = \sum_i \sum_j y_i x_{ij}\beta_j - \sum_i m_i \log\left(1 + \exp\sum_j x_{ij}\beta_j\right), \quad (4.13)$$

where we have written $l(\boldsymbol{\beta}; \mathbf{y})$ instead of $l(\boldsymbol{\pi}(\boldsymbol{\beta}); \mathbf{y})$. The important point to notice here is that, because the logistic link is also the canonical link, the log likelihood depends on \mathbf{y} only through the linear combinations $\mathbf{X}^T \mathbf{y}$. These p combinations are said to be sufficient for $\boldsymbol{\beta}$. In fact, as will be seen shortly, the likelihood equations in this special case amount to setting the observed linear combinations $\mathbf{X}^T \mathbf{y}$ equal to their expectation, namely $E(\mathbf{X}^T \mathbf{Y}; \hat{\boldsymbol{\beta}})$. This may be viewed a special case of the *method of moments*.

Section 2.4.4 and Table 2.1 give the canonical link functions for other distributions.

4.4.2 Parameter estimation

Following the general technique given in section 2.5, we now derive the likelihood equations for the parameters $\boldsymbol{\beta}$ that appear in (4.12). First note that the derivative of the log-likelihood function, in the form given in (4.11), with respect to π_i is

$$\frac{\partial l}{\partial \pi_i} = \frac{y_i - m_i \pi_i}{\pi_i(1 - \pi_i)}.$$

Using the chain rule, the derivative with respect to β_r is

$$\frac{\partial l}{\partial \beta_r} = \sum_{i=1}^n \frac{y_i - m_i \pi_i}{\pi_i(1 - \pi_i)} \frac{\partial \pi_i}{\partial \beta_r}.$$

In the case of generalized linear models, it is convenient to express $\partial \pi_i / \partial \beta_r$ as a product

$$\frac{\partial \pi_i}{\partial \beta_r} = \frac{d\pi_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_r} = \frac{d\pi_i}{d\eta_i} x_{ir}.$$

Thus the derivative with respect to β_r is

$$\frac{\partial l}{\partial \beta_r} = \sum_i \frac{y_i - m_i \pi_i}{\pi_i(1 - \pi_i)} \frac{d\pi_i}{d\eta_i} x_{ir}. \quad (4.14)$$

The Fisher information for β is

$$\begin{aligned} -E\left(\frac{\partial^2 l}{\partial \beta_r \partial \beta_s}\right) &= \sum_i \frac{m_i}{\pi_i(1 - \pi_i)} \frac{\partial \pi_i}{\partial \beta_r} \frac{\partial \pi_i}{\partial \beta_s} \\ &= \sum_i m_i \frac{(d\pi_i/d\eta_i)^2}{\pi_i(1 - \pi_i)} x_{ir} x_{is} \\ &= \{\mathbf{X}^T \mathbf{W} \mathbf{X}\}_{rs}, \end{aligned} \quad (4.15)$$

where \mathbf{W} is a diagonal matrix of weights given by

$$\mathbf{W} = \text{diag}\left\{m_i \left(\frac{d\pi_i}{d\eta_i}\right)^2 / \pi_i(1 - \pi_i)\right\}.$$

In the case of linear logistic models, equation (4.14) reduces to

$$\partial l / \partial \beta = \mathbf{X}^T (\mathbf{Y} - \boldsymbol{\mu})$$

when written in matrix notation. The likelihood equations then amount to equating the sufficient statistic, $\mathbf{X}^T \mathbf{Y}$, to its expectation as a function of β . In addition, the diagonal matrix of weights appearing in the Fisher information reduces to

$$\mathbf{W} = \text{diag}\{m_i \pi_i (1 - \pi_i)\}.$$

Following the lines of the general Newton-Raphson procedure described in Chapter 2, parameter estimates may be obtained in the following way. Given initial estimates $\hat{\beta}_0$, we may compute the vectors $\hat{\pi}_0$ and $\hat{\eta}_0$. Using these values, define the adjusted dependent variate, \mathbf{Z} , with components

$$z_i = \hat{\eta}_i + \frac{y_i - m_i \hat{\pi}_i}{m_i} \frac{d\eta_i}{d\pi_i},$$

all quantities being computed at the initial estimate $\hat{\beta}_0$. Maximum-likelihood estimates satisfy the equation

$$\mathbf{X}^T \mathbf{W} \mathbf{X} \hat{\beta} = \mathbf{X}^T \mathbf{W} \mathbf{Z}, \quad (4.16)$$

which can be solved iteratively using standard least-squares methods. The revised estimate is

$$\hat{\beta}_1 = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Z}$$

where all quantities appearing on the right are computed using the initial estimate.

Failure to converge is rarely a problem unless one or more components of $\hat{\beta}$ are infinite, which usually implies that some of the fitted probabilities are either zero or one. Infinite parameter estimates can occur if the data are sparse and $y_i = 0$ or $y_i = m_i$ for certain components of the response vector. Although the iterative procedure does not converge under these circumstances, nevertheless the sequence of fitted probabilities, $\hat{\pi}^{(j)}$ generally tends quite rapidly towards $\hat{\pi}$ and the deviance towards its limiting value. After a few cycles of (4.16) the fitted values $m_i \hat{\pi}_i$ are normally quite accurate but the parameter estimates and their standard errors may not be. Two criteria ought therefore to be tested to detect abnormal convergence of this type. The primary criterion ought to be based on the change in the fitted probabilities, for instance by using the deviance. A supplementary test for parameter divergence can be based on the change in $\hat{\beta}$ or in the linear predictor, $\hat{\eta}$. Abnormal convergence means that the log likelihood is either very flat or, more likely, has an asymptote. Consequently, the computed parameter estimates and their estimated standard errors are not to be trusted.

Some results concerning the existence and uniqueness of parameter estimates have been given by Wedderburn (1976) and by Haberman (1977). These results show that if the link function is log concave, as it is for the four functions discussed in section 4.3.1, and if $0 < y_i < m_i$ for each i , then $\hat{\beta}$ is finite and the log likelihood has a unique maximum at $\hat{\beta}$.

Starting values $\hat{\beta}^{(0)}$ can be obtained using the method described in Chapter 2, beginning with ‘fitted values’ $\tilde{\mu} = (y + \frac{1}{2})/(m + 1)$. A good choice of starting value usually reduces the number of cycles in (4.16) by about one or perhaps two. Consequently, the choice of initial estimate is usually not critical. A bad choice may, however, result in divergence.

4.4.3 Deviance function

The residual deviance is defined to be twice the difference between the maximum achievable log likelihood and that attained under the fitted model. Under any given model, H_0 , with fitted probabilities $\hat{\pi}$, the log likelihood is

$$l(\hat{\pi}; \mathbf{y}) = \sum_i \left\{ y_i \log \hat{\pi}_i + (m_i - y_i) \log(1 - \hat{\pi}_i) \right\},$$

which is just (4.11) written in a more symmetrical form. The maximum achievable log likelihood is attained at the point $\tilde{\pi}_i = y_i/m_i$, but this point does not usually occur in the model space under H_0 . The deviance function is therefore

$$\begin{aligned} D(\mathbf{y}; \hat{\pi}) &= 2l(\tilde{\pi}; \mathbf{y}) - 2l(\hat{\pi}; \mathbf{y}) \\ &= 2 \sum_i \left\{ y_i \log(y_i/\hat{\mu}_i) + (m_i - y_i) \log\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right) \right\}. \end{aligned}$$

This function behaves in much the same way as the residual sum of squares or weighted residual sum of squares in ordinary linear models. The addition of further covariates has the effect of reducing D .

It is often claimed that the random variable $D(\mathbf{Y}; \hat{\pi})$ is asymptotically or approximately distributed as χ^2_{n-p} , where p is the number of fitted parameters under H_0 . This claim is then used to justify the use of D as a goodness-of-fit statistic for testing the adequacy of the fitted model. Proofs of the limiting χ^2_{n-p} distribution are based on the following assumptions whose relevance in any given application must be open to question.

Assumption 1: The observations are distributed independently according to the binomial distribution. In other words, the possibility of over-dispersion (Section 4.5) is not considered.

Assumption 2: The approximation is based on a limiting operation in which $\dim(\mathbf{Y}) = n$ is fixed, $m_i \rightarrow \infty$ for each i , and in fact $m_i \pi_i(1 - \pi_i) \rightarrow \infty$.

In the limit given by assumption 2, D is approximately independent of the estimated parameters $\hat{\beta}$ and hence approximately independent of the fitted probabilities $\hat{\pi}$. Approximate independence is essential for D to be considered as a goodness-of-fit statistic, but this property alone does not guarantee good power.

If n is large and $m_i\pi_i(1-\pi_i)$ remains bounded the whole theory breaks down in two ways. First, the limiting χ^2 approximation no longer holds. Second, and more importantly, D is not independent of $\hat{\boldsymbol{\pi}}$ even approximately. As a consequence, a large value of D could be obtained with high probability by judicious choice of $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$. In other words, a large value of D cannot necessarily be considered to be evidence of a poor fit. For an extreme instance of this effect, see section 4.4.5.

The deviance function is most directly useful not as an absolute measure of goodness-of-fit but for comparing two nested models. For instance, we may wish to test whether the addition of a further covariate significantly improves the fit. Let H_0 denote the model under test and H_A the extended model containing an additional covariate. The corresponding fitted values are denoted by $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\mu}}_A$ respectively. The reduction in deviance

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\mu}}_A) = 2l(\hat{\boldsymbol{\mu}}_A; \mathbf{y}) - 2l(\hat{\boldsymbol{\mu}}_0; \mathbf{y}) \quad (4.17)$$

is identical to the likelihood-ratio statistic for testing H_0 against H_A . This statistic is distributed approximately like χ^2_1 independently of $\hat{\boldsymbol{\mu}}$ under assumption 1 above provided that either n is large or that assumption 2 is satisfied. In particular, $D(\mathbf{Y}; \hat{\boldsymbol{\mu}}_0)$ need not have an approximate χ^2 distribution nor need it be distributed independently of $\hat{\boldsymbol{\mu}}_0$. The χ^2 approximation is usually quite accurate for differences of deviances even though it is inaccurate for the deviances themselves.

4.4.4 Bias and precision of estimates

To a first order of approximation, maximum-likelihood estimates are unbiased with asymptotic variance equal to the inverse Fisher information matrix (4.15). Specifically, for large n ,

$$\begin{aligned} E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= O(n^{-1}) \\ \text{cov}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \{1 + O(n^{-1})\}. \end{aligned} \quad (4.18)$$

These approximate results are also true for the alternative limit in which n is fixed and $\mathbf{m} \rightarrow \infty$. The errors are then $O(m_i^{-1})$.

It is possible to give an expression for the bias of $\hat{\boldsymbol{\beta}}$ that covers all link functions. However, in order to keep the expressions as simple

as possible, we shall restrict attention to linear logistic models. In that case, the bias of $\hat{\beta}$ involves the 3-way array

$$\kappa_{r,s,t} = \sum_i x_{ir} x_{is} x_{it} m_i \pi_i (1 - \pi_i) (1 - 2\pi_i),$$

which is just the skewness array of the log likelihood derivative $\partial l / \partial \beta$. If we denote by $\kappa_{r,s}$ the elements of the Fisher information matrix $\mathbf{X}^T \mathbf{W} \mathbf{X}$, and by $\kappa^{r,s}$ the elements of the inverse matrix, we have

$$\text{bias}(\hat{\beta}_r) \simeq - \sum_{ijk} \kappa^{r,i} \kappa^{j,k} \kappa_{i,j,k} / 2.$$

The approximate skewness array of $\hat{\beta}$ is

$$\text{cum}(\hat{\beta}_r, \hat{\beta}_s, \hat{\beta}_t) \simeq -2 \sum_{ijk} \kappa^{r,i} \kappa^{s,j} \kappa^{t,k} \kappa_{i,j,k}.$$

Bias and skewness terms represent the major departures of the distribution of $\hat{\beta}$ from the usual Normal approximation. Edgeworth corrections will usually improve the accuracy of the approximation.

The corresponding expressions for other link functions are given by McCullagh (1987, p.209). Computational tactics for generalized linear models are discussed in sections 15.2–15.3.

4.4.5 Sparseness

By sparseness we mean that a sizeable proportion of the observed counts are small. An extreme instance of this phenomenon occurs in Table 4.1a, where data are listed by subject number and hence $m_i = 1$ for each i . More generally, we say that the data are sparse if many components of the binomial index vector are small, say 5 or less. Sparseness does not necessarily imply that there is little information in the data about the values of the parameters. On the contrary, if the data recorded are extensive, (n large), the asymptotic approximation (4.18) is usually quite accurate. The effect of sparseness is noticed mainly on the deviance function and Pearson's statistic, which fail to have the properties required for goodness-of-fit statistics.

To illustrate the nature of the effect, suppose that $Y_i \sim B(1, \pi_i)$ and that a linear logistic model such as (4.10) has been fitted by maximum likelihood, yielding fitted values

$$\hat{\pi}_i = \exp(\mathbf{x}_i^T \hat{\beta}) / [1 + \exp(\mathbf{x}_i^T \hat{\beta})].$$

The residual deviance function is

$$\begin{aligned} D &= 2 \sum \left\{ y_i \log \left(\frac{y_i}{\hat{\pi}_i} \right) + (1 - y_i) \log \left(\frac{1 - y_i}{1 - \hat{\pi}_i} \right) \right\} \\ &= 2 \sum \left\{ y_i \log y_i + (1 - y_i) \log(1 - y_i) \right. \\ &\quad \left. - y_i \log \left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right) - \log(1 - \hat{\pi}_i) \right\}. \end{aligned}$$

Since $y = 0$ or 1 , we have $y \log y = (1 - y) \log(1 - y) = 0$. Further, $\log(\hat{\pi}_i/(1 - \hat{\pi}_i)) = \mathbf{x}_i^T \hat{\beta}$. Thus

$$\begin{aligned} D &= -2 \hat{\beta}^T \mathbf{X}^T \mathbf{Y} - 2 \sum \log(1 - \hat{\pi}_i) \\ &= -2 \hat{\eta}^T \hat{\pi} - 2 \sum \log(1 - \hat{\pi}_i) \end{aligned}$$

since $\mathbf{X}^T \mathbf{Y} = \mathbf{X}^T \hat{\mu}$ is the maximum-likelihood equation. Evidently, therefore, D is a function of $\hat{\beta}$ in this case. In other words, given $\hat{\beta}$, D has a conditionally degenerate distribution and cannot be used to test goodness of fit. Exact degeneracy occurs only for linear logistic models if $m_i = 1$, but near degeneracy occurs for any link function provided that the m_i are small.

The effect of extreme sparseness on Pearson's statistic is less obvious but can be seen from the following example. Suppose that the observations are identically distributed and $Y_i \sim B(1, \pi)$. Then $\bar{\pi} = \bar{y}$ and Pearson's statistic reduces to

$$X^2 = \sum \frac{(y_i - \bar{y})^2}{\bar{y}(1 - \bar{y})} = n.$$

The sample size is not very useful as a test for goodness of fit! The deviance function fares no better, for

$$D = -2n \{ \bar{y} \log \bar{y} + (1 - \bar{y}) \log(1 - \bar{y}) \},$$

is a function of $\hat{\pi}$.

For intermediate cases in which the m_i are small but mostly greater than one, we may use D or X^2 as test statistics. However, in the computation of significance levels it is essential to use the conditional distribution of the statistic given the observed $\hat{\beta}$. Exact conditional moments of X^2 can be computed in some important

special cases: see, for example, the Haldane-Dawson formulae for two-way tables in Exercise 6.16. More generally, however, approximate formulae are available for the conditional mean and variance of X^2 for linear logistic models (McCullagh, 1985). If n is large, it is best to use a Normal approximation for X^2 in which the conditional mean and variance are

$$\begin{aligned} E(X^2 | \hat{\beta}) &\simeq n - p - \frac{1}{2} \sum_i \{1 - 6\hat{\pi}_i(1 - \hat{\pi}_i)\}\hat{V}_{ii} \\ &\quad + \frac{1}{2} \sum_{ij} m_i \hat{\pi}_i(1 - \hat{\pi}_i)(1 - 2\hat{\pi}_i) \hat{V}_{ii} \hat{V}_{ij} (1 - 2\hat{\pi}_j) \\ \text{var}(X^2 | \hat{\beta}) &\simeq (1 - p/n) \left\{ 2 \sum_i \left(\frac{m_i - 1}{m_i} \right) + \sum_{ij} (1 - 2\hat{\pi}_i)(1 - 2\hat{\pi}_j) \hat{V}_{ij} \right\} \end{aligned}$$

where V_{ij} are the elements of $\mathbf{V} = \mathbf{X}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T$, the approximate covariance matrix of $\hat{\eta}$. These expressions are easy to compute following the fitting of a linear logistic model because the matrix \mathbf{V} is readily available. Note that, unlike D , the conditional variance of X^2 is ordinarily not zero for pure binary data.

Similar expressions are available for the conditional cumulants of the deviance statistic but these are too complex for practical use. See, for instance, McCullagh (1986).

It is good statistical practice, however, not to rely on either D or X^2 as an absolute measure of goodness of fit in these circumstances. It is much better to look for specific deviations from the model of a type that is easily understood scientifically. For instance, we may look for interactions among the covariates or non-linear effects by adding suitable terms to the model and observing the reduction in deviance. The reduction in deviance thus induced is usually well approximated by a χ^2 distribution.

4.4.6 Extrapolation

Extrapolation beyond the range of the observed x -values in order to predict the probability of failure at extreme x -values is a hazardous exercise because its success depends heavily on the correctness of the assumed model, particularly on the choice of link function. It is common to find that two models that give similar predictions over the range of observed x -values may give very different predictions

when extrapolated. The need for extreme extrapolation arises most commonly in reliability experiments, where failure is a rare event under naturally-occurring conditions. Experimentation is carried out under an accelerated testing regime using extreme stresses or high doses to increase the observed failure rate. For instance, in certain toxicology experiments, laboratory animals are subjected to unusually high doses of a suspected toxin or carcinogen. On the basis of the observed responses at high dose levels, it is required either to predict the failure rate at much lower dose levels, or to set confidence limits on the dose x_0 that would produce an acceptably low failure rate, π_0 , the so-called maximum safe dose or maximum acceptable dose.

Suppose, by way of example, that the observed dose levels in log units and the responses are as shown in Table 4.3. It is required to predict the failure rate at dose levels equal to 1/50 unit and 1/100 unit, corresponding on the log scale to $x = -3.912$ and -4.605 respectively. On fitting the model

$$g(\pi) = \beta_0 + \beta_1 x$$

for various choices of $g(\pi)$, we find the fitted probabilities as shown in Table 4.3. On treating $(\hat{\beta}_0, \hat{\beta}_1)$ as bivariate Normal with covariance matrix (4.18), we find the predicted failure rates and confidence intervals as shown in Table 4.4. Clearly, the predicted failure probabilities are heavily dependent on the choice of link function, although the fitted probabilities in Table 4.3 are almost identical for the four link functions.

Table 4.3 *Hypothetical responses in a toxicology experiment*

Dose (log units)	Response y/m	Fitted probability			
		logit	probit	c-log log	log-log
0	3/10	0.280	0.281	0.288	0.278
1	5/10	0.540	0.540	0.519	0.558
2	8/10	0.780	0.782	0.793	0.766

The converse problem of setting approximate confidence intervals for the dose x_0 that gives rise to a failure probability π_0 is most easily accomplished using Fieller's method. The linear combination

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 - g(\pi_0)$$

Table 4.4 Failure rates predicted by four models at low doses

x_0	link	$\hat{\pi}(x_0)$	90% Confidence interval	
			g-scale	π -scale
-3.912	logit	0.00513	(-9.47, -1.07)	$(7.7 \times 10^{-5}, 0.26)$
	probit	0.00061	(-5.72, -0.75)	$(5.3 \times 10^{-7}, 0.23)$
	c-log log	0.01684	(-7.04, -1.11)	$(8.8 \times 10^{-4}, 0.28)$
	log log	1.1×10^{-12}	(-6.13, -0.50)	$(10^{-200}, 0.19)$
-4.605	logit	0.00239	(-10.82, -1.25)	$(2.0 \times 10^{-5}, 0.22)$
	probit	0.00011	(-6.54, -0.87)	$(3.1 \times 10^{-9}, 0.19)$
	c-log log	0.00994	(-7.76, -1.26)	$(3.5 \times 10^{-4}, 0.25)$
	log log	2.3×10^{-21}	(-7.09, -0.63)	$(10^{-522}, 0.15)$

is approximately Normally distributed with mean 0 and variance

$$v^2(x_0) = \text{var}(\hat{\beta}_0) + 2x_0 \text{cov}(\hat{\beta}_0, \hat{\beta}_1) + x_0^2 \text{var}(\hat{\beta}_1)$$

The resulting confidence ‘interval’ is the set of all x_0 -values satisfying

$$\left| \frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - g(\pi_0)}{v(x_0)} \right| < k_{\alpha/2}^* \quad (4.19)$$

where $\Phi(k_\alpha^*) = 1-\alpha$. The set (4.19) may be a finite interval, a semi-infinite interval or the complement of an interval. The numerical values produced by (4.19) are again heavily dependent on the choice of link function.

In practice, it is usually a good idea to compute the set (4.19) for a suitable selection of link functions. Only if these are in reasonable agreement can any real confidence be placed in the predictions.

An alternative method for constructing approximate confidence intervals for x_0 using the likelihood function directly is outlined in Exercises 4.19 and 4.20.

4.5 Over-dispersion

4.5.1 Genesis

By the term ‘over-dispersion’, we mean that the variance of the response Y exceeds the nominal variance — in this case the nominal binomial variance, $m\pi(1-\pi)$. Over-dispersion is not uncommon in practice. In fact, some would maintain that over-dispersion is the

norm in practice and nominal dispersion the exception. The incidence and the degree of over-dispersion encountered greatly depend on the field of application. In large-scale epidemiological studies concerning geographical variation in the incidence of disease, the binomial variance is often an almost negligible component of the total variance. Unless there are good external reasons for relying on the binomial assumption, it seems wise to be cautious and to assume that over-dispersion is present to some extent unless and until it is shown to be absent.

Over-dispersion can arise in a number of ways. The simplest, and perhaps the most common mechanism, is clustering in the population, a mechanism previously proposed by Lexis (1979): see Stigler (1986, p. 229–238). Families, households, litters, colonies and neighbourhoods are common instances of naturally-occurring clusters in populations. Clusters usually vary in size, but we shall assume for simplicity that the cluster size, k , is fixed and that the m individuals sampled actually come from m/k clusters. In the i th cluster, the number of positive respondents, Z_i , is assumed to have the binomial distribution with index k and parameter π_i , which varies from cluster to cluster. Thus, the total number of positive respondents is

$$Y = Z_1 + Z_2 + \dots + Z_{m/k}.$$

If we write $E(\pi_i) = \pi$ and $\text{var}(\pi_i) = \tau^2\pi(1 - \pi)$, it may be shown that the unconditional mean and variance of Y are

$$\begin{aligned} E(Y) &= m\pi \\ \text{var}(Y) &= m\pi(1 - \pi)\{1 + (k - 1)\tau^2\} \\ &= \sigma^2 m\pi(1 - \pi). \end{aligned} \tag{4.20}$$

Note that the dispersion parameter $\sigma^2 = 1 + (k - 1)\tau^2$ depends on the cluster size and on the variability of π from cluster to cluster, but not on the sample size, m . This is important because it enables us to proceed as if the observations were binomially distributed and to estimate the dispersion parameter from the residuals.

Over-dispersion can occur only if $m > 1$. If $m = 1$, the mean necessarily determines the variance and all higher-order cumulants. In general, the preceding derivation via cluster sampling forces the dispersion parameter to lie in the interval

$$1 \leq \sigma^2 \leq k \leq m$$

because $0 \leq \tau^2 \leq 1$. It is often desirable, in order to accommodate under-dispersion, to extend the domain of definition to include values of σ^2 in the interval $0 \leq \sigma^2 \leq 1$.

The beta-binomial distribution (Exercise 4.17), is sometimes used as an alternative model for over-dispersion. This distribution has the property that the variance ratio $\text{var}(Y)/\{m\pi(1 - \pi)\}$ is a linear function of m , rather than a constant as in (4.20). By plotting residuals against m it is possible, in principle at least, to discriminate between these two models. The examples that we have examined, however, seem to favour the constant dispersion factor in (4.20) over the beta-binomial model.

4.5.2 Parameter estimation

With specific forms of over-dispersion, such as that described in Exercise 4.17 leading to the beta-binomial model, one can use maximum likelihood to estimate the regression parameters and the dispersion parameter jointly. Though this is an attractive option from a theoretical standpoint, in practice it seems unwise to rely on a specific form of over-dispersion, particularly where the assumed form has been chosen for mathematical convenience rather than scientific plausibility. For that reason, in what follows we assume that the effect of over-dispersion is as shown in (4.20). In other words, the mean is unaffected but the variance is inflated by an unknown factor σ^2 .

With this form of over-dispersion, the models described in section 4.3 may still be fitted using the methods of section 4.4, as if the binomial distribution continued to apply. The only difference occurs in section 4.3.3 where the χ_{n-p}^2 and the χ_1^2 approximations are replaced by $\sigma^2 \chi_{n-p}^2$ and $\sigma^2 \chi_1^2$ respectively. In section 4.4.4, the covariance matrix of $\hat{\beta}$ is replaced by

$$\text{cov}(\hat{\beta}) \simeq \sigma^2 (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}. \quad (4.21)$$

For further details, see Chapter 9.

The expressions for bias and skewness given in section 4.4.4 are not valid without further assumptions concerning the effect of over-dispersion on the higher-order cumulants: see, for instance, Exercise 4.18.

There remains only the problem of estimating the dispersion factor, which is required for setting confidence limits on β and on

components of β . This is exactly analogous to the problem of estimating σ^2 in ordinary Normal-theory linear or non-linear models. Suppose first that there is replication: in other words, for each covariate value \mathbf{x} , several observations $(y_1, m_1), \dots, (y_r, m_r)$ are observed. These observations are independent and essentially identically distributed apart from the fact that the indices m_1, \dots, m_r may be unequal. The estimate of π based on this covariate class alone is

$$\tilde{\pi} = y_{\cdot}/m_{\cdot}.$$

and the expected value of the within-class weighted sum of squares

$$\sum_{j=1}^r (y_j - m_j \tilde{\pi})^2 / m_j$$

is equal to $(r - 1)\sigma^2\pi(1 - \pi)$. In other words,

$$\hat{s}^2 = \frac{1}{r - 1} \sum_j \frac{(y_j - m_j \tilde{\pi})^2}{m_j \tilde{\pi}(1 - \tilde{\pi})} \quad (4.22)$$

is an approximately unbiased estimator of σ^2 on $r - 1$ degrees of freedom. On pooling together these estimators, one for each covariate class in which replication occurs, we obtain the replication estimate of dispersion on $\sum(r - 1)$ degrees of freedom. This estimator has a slight bias of order $O(m_{\cdot}^{-1})$ in the binomial case (for $\sigma^2 = 1$) and has comparable bias otherwise. The value of the replication estimate of σ^2 is independent of the fitted model.

In the absence of replication, or if the number of degrees of freedom for replication is small, an estimate of σ^2 may be based on the residual sum of squares appropriately weighted. If the fitted model is correct,

$$\tilde{\sigma}^2 = \frac{1}{n - p} \sum_i \frac{(y_i - m_i \hat{\pi}_i)^2}{m_i \hat{\pi}_i(1 - \hat{\pi}_i)} = X^2/(n - p) \quad (4.23)$$

is approximately unbiased for σ^2 provided that p is small compared with n . The estimated covariance matrix of $\hat{\beta}$ is then

$$\text{estimated var}(\hat{\beta}) = \tilde{\sigma}^2 (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1}.$$

Note that if $m_i = 1$ for each i we must have $\sigma^2 = 1$. The replication estimator (4.22) has this property to a close approximation but (4.23) based on Pearson's statistic does not.

The alternative estimator of σ^2 based on the normalized residual deviance is approximately equal to $\tilde{\sigma}^2$ in the non-sparse case for which all m_i are large. In the sparse case, however, $\tilde{\sigma}^2$ is consistent for σ^2 whereas $D(\mathbf{y}; \hat{\mu})/(n - p)$ is not. The latter claim is evident from the discussion in section 4.4.5. For instance, if $Y_i \sim B(1, \pi)$ for each i , (4.23) and (4.22) give

$$\tilde{\sigma}^2 = s^2 = n/(n - 1),$$

which tends to unity as n becomes large. By contrast,

$$\frac{D}{n - 1} = -\frac{2n}{n - 1} \{ \hat{\pi} \log \hat{\pi} + (1 - \hat{\pi}) \log(1 - \hat{\pi}) \},$$

whose value ranges from 0 to $1.386 = 2 \log 2$ as $\hat{\pi}$ ranges from 0 to 0.5.

The approximate bias and variance of $\tilde{\sigma}^2$ in the absence of over-dispersion are given in section 4.4.5. In the presence of over-dispersion satisfying (4.20), the bias of $\tilde{\sigma}^2$ is of order $O(n^{-1})$. Both the bias and the variance depend on the effect of over-dispersion on the third and fourth cumulants of Y . If the effect of over-dispersion on these cumulants is as described in Exercise 4.18, explicit expressions can be obtained for the approximate bias and variance of $\tilde{\sigma}^2$. These formulae are moderately complicated and are of limited usefulness in practice because the higher-order dispersion factors must be estimated from the available data. Even for linear models, the estimation of higher-order cumulants is seldom worthwhile unless the data are very extensive.

4.6 Example

4.6.1 Habitat preferences of lizards

The following data are in many ways typical of social-science investigations, although the example concerns the behaviour of lizards rather than humans. The data, taken from Schoener (1970), have subsequently been analysed by Fienberg (1970b) and by

Table 4.5 A comparison of site preferences of two species of lizard, *grahami* and *opalinus*

S	Perch D (in)	H (ft)	T								
			Early			Mid-day			Late		
			G	O	Total	G	O	Total	G	O	Total
Sun	≤ 2	< 5	20	2	22	8	1	9	4	4	8
		≥ 5	13	0	13	8	0	8	12	0	12
	> 2	< 5	8	3	11	4	1	5	5	3	8
		≥ 5	6	0	6	0	0	0	1	1	2
Shade	≤ 2	< 5	34	11	45	69	20	89	18	10	28
		≥ 5	31	5	36	55	4	59	13	3	16
	> 2	< 5	17	15	32	60	32	92	8	8	16
		≥ 5	12	1	13	21	5	26	4	4	8

H, perch height; *D*, perch diameter; *S*, sunny/shady; *T*, time of day; *G*, *grahami*; *O*, *opalinus*.

Bishop *et al.* (1975). Data concerning the daytime habits of two species of lizard, *grahami* and *opalinus*, were collected by observing occupied sites or perches and recording the appropriate description, namely species involved, time of day, height and diameter of perch and whether the site was sunny or shaded. Time of day is recorded here as early, mid-day or late.

As often with such problems, several analyses are possible depending on the purpose of the investigation. We might, for example, wish to compare how preferences for the various perches vary with the time of day regardless of the species involved. We find by inspection of the data (Table 4.5) that shady sites are preferred to sunny sites at all times of day but particularly so at mid-day. Furthermore, again by inspection, low perches are preferred to high ones and small-diameter perches to large ones. There is, of course, the possibility that these conclusions are produced by an artefact of the data-collection process and that, for instance, occupied sites at eye level or below are easier to spot than occupied perches higher up. In fact, selection bias of this type seems inevitable unless some considerable effort is devoted to observing all lizards in a given area.

A similar analysis, but with the same deficiencies, can be made for each species separately.

Suppose instead that an occupied site, regardless of its position, diameter and so on, is equally difficult to spot whether occupied by a *grahami* or an *opalinus* lizard. This assumption would be plausible if the two species were similar in size and colour. Suppose in addition that the purpose of the investigation is to compare the two species with regard to their preferred perches. Thus we see that, of the 22 occupied perches of small diameter low in the tree observed in a sunny location early in the day, only two, or 9%, were occupied by *opalinus* lizards. For similar perches observed later in the day, the proportion is four out of eight, i.e. 50%. On this comparison, therefore, it appears that, relative to *opalinus*, *grahami* lizards prefer to sun themselves early in the day.

To pursue this analysis more formally, we take as fixed the total number m_{ijkl} of occupied sites observed for each combination of i = perch height, j = perch diameter, k = sunny/shady and l = time of day. In the language of statistical theory, these totals or covariate-class sizes are ancillary provided that the purpose of the investigation is to compare preferences or to examine the differences between the site preferences of the two species. The response variable y_{ijkl} gives the observed number or, equivalently, the observed proportion of the m_{ijkl} occupied sites that were occupied by *grahami* lizards. By symmetry, we could equally well work with $m_{ijkl} - y_{ijkl}$, the number of sites occupied by *opalinus* lizards. We take the random variable Y_{ijkl} to be binomially distributed with index m_{ijkl} and parameter π_{ijkl} . Thus π is the probability that an observed occupied site is in fact occupied by a *grahami* lizard. Of course, the possibility of over-dispersion relative to the binomial distribution must be borne in mind.

At the exploratory stage, probably the simplest analysis of these data is obtained by transforming to the logistic scale. Using the empirical logistic transformation (4.7), we have the transformed value for $y_1/m_1 = 20/22$, namely

$$z_1 = \log(20.5/2.5) = 2.1041$$

with approximate variance $1/20.5 + 1/2.5 = 0.4488$. A straightforward linear analysis of the transformed values is usually a satisfactory method of analysis if all the observed counts are moderately large. In this example not all the counts are large and for that reason, we must confirm our findings using a different technique. To

Table 4.6 Computation of logistic factorial standardized contrasts for lizard data

Transformed value	Estimated variance	Raw contrast	Estimated variance	Parameter	Absolute standardized contrast
2.1041	0.4488	30.9092	20.16	<i>I</i>	—
3.2958	2.0741	11.0986	20.16	<i>H</i>	2.47
0.8873	0.4034	-12.4508	20.16	<i>D</i>	2.77
2.5649	2.1538	-5.3629	20.16	<i>HD</i>	1.19
1.0986	0.1159	-5.4698	20.16	<i>S</i>	1.22
1.7451	0.2136	-0.1739	20.16	<i>HS</i>	0.04
0.1214	0.1217	3.9168	20.16	<i>DS</i>	0.87
2.1203	0.7467	5.4014	20.16	<i>HDS</i>	1.20
1.7346	0.7843	-8.3505	11.79	<i>T_L</i>	2.43
2.8332	2.1176	-1.9645	11.79	<i>HT_L</i>	0.57
1.0986	0.8889	-2.1333	11.79	<i>DT_L</i>	0.62
0.0000	4.0000	-6.2926	11.79	<i>HDT_L</i>	1.83
1.2209	0.0632	2.0122	11.79	<i>ST_L</i>	0.59
2.5123	0.2402	-1.7595	11.79	<i>HST_L</i>	0.51
0.6214	0.0473	-0.4950	11.79	<i>DST_L</i>	0.14
1.3633	0.2283	2.0210	11.79	<i>HDST_L</i>	0.59
0.0000	0.4444	-3.2438	45.27	<i>T_Q</i>	0.48
3.2189	2.0900	4.9987	45.27	<i>HT_Q</i>	0.74
0.4520	0.4675	3.2022	45.27	<i>DT_Q</i>	0.48
0.0000	1.3333	2.8773	45.27	<i>HDT_Q</i>	0.43
0.5664	0.1493	-5.6243	45.27	<i>ST_Q</i>	0.84
1.3499	0.3598	-6.2738	45.27	<i>HST_Q</i>	0.93
0.0000	0.2353	-1.2453	45.27	<i>DST_Q</i>	0.19
0.0000	0.4444	0.4582	45.27	<i>HDST_Q</i>	0.07

maintain balance, the observation $(0, 0)$ is transformed to $z_{12} = 0.0$ with 'variance' 4.0.

The first two columns of Table 4.6 give the transformed values and their estimated variances listed in the usual standard order corresponding to the factors *H*, *D*, *S* and *T*. Four steps of Yates's algorithm (not given) produce the raw contrasts, again associated with the four factors in the same standard order. In the case of the factor *T*, which has three ordered levels, linear and quadratic contrasts were used to complete the decomposition. Variances are computed in a similar way, the coefficients being squared. Thus, all main effects and interactions involving *H*, *D* and *S* only have the

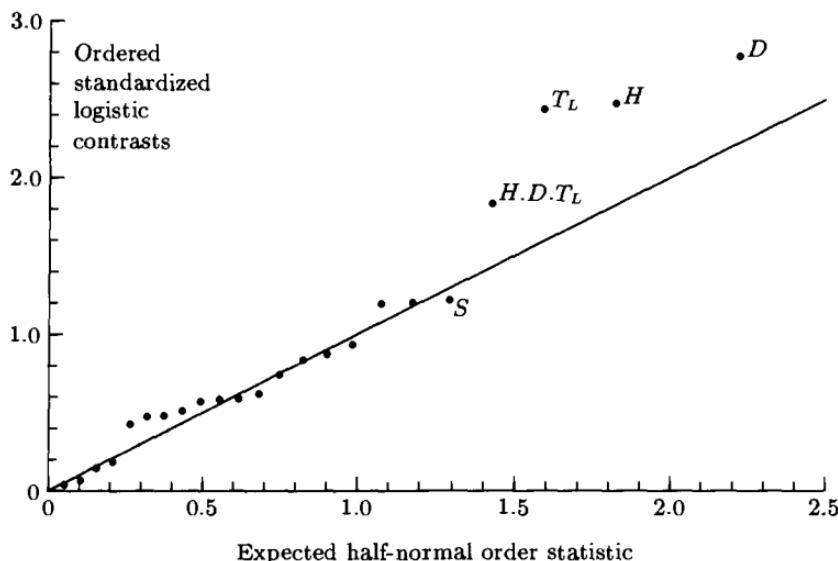


Fig. 4.2. *Half-normal plot of ordered absolute standardized logistic contrasts for the lizard data. The solid line is at 45° .*

same variance, 20.16, which is the total of column 2. Similarly for terms involving T_L and for terms involving T_Q . Finally we compute the standardized contrasts: of these, only the main effects of H and D and the linear effect of time, with standardized contrasts in excess of 2.4, appear to be significant.

A half-Normal plot (Daniel, 1959) of the ordered absolute standardized logistic contrasts against their Normal-theory expected values (Fig. 4.2), suggests that the main effects of height and diameter and the linear effect of time are significant though not overwhelmingly so. The three-factor interaction $H.D.T_L$ also deviates from the theoretical line, but this appears to be an aberration well within the sampling limits especially when due allowance is made for the effect of selection. As a matter of policy, no allowance for selection would normally be made when judging the significance of main effects in a full factorial design. Such effects that are not expected in advance to be null should be excluded from the half-normal plot, though this has not been done in Fig. 4.2.

The unit slope observed in Fig. 4.2 is evidence that $\sigma = 1$ and hence there is no suggestion of over-dispersion.

Because of the numerous small observed counts in this particular example, some caution is required in the interpretation of contrasts

in Table 4.6. It is possible, for example, that the addition of 1/2 to each count before transforming could swamp the data. Indeed such an effect seems to have occurred for the sunny/shady contrast. Here, few *opalinus* lizards were observed in sunny locations, so that the addition of 1/2 to each of these counts has a marked effect on the *S* contrast, reducing it towards zero and so diluting its apparent significance.

We now consider an alternative analysis using a generalized linear model fitted by maximum likelihood, which avoids transformation problems. The preceding analysis in Table 4.6 and Fig. 4.2 suggests that the structure of these data is fairly simple; there appear to be no strong interactions on the logistic scale. We are therefore led initially to consider the linear logistic model including all four main effects. Such a model can be written symbolically as $H + D + S + T$ or, in subscript notation,

$$\text{logit}(\pi_{ijkl}) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l, \quad (4.24)$$

where α , β , γ and δ refer to the four factors H , D , S and T . In fact this model fits the data quite well with no evidence of overdispersion. All main effects including S are significant at the 5% level, illustrating the drawbacks of the previous analysis where S appeared to be insignificant. None of the two-factor interactions appears significant; the relevant statistics are given in Table 4.7. Parameter estimates associated with the model (4.24) are given in Table 4.8, where we use the convention of setting the first level of each factor to zero. It is possible here to replace T by a single contrast corresponding to late afternoon versus earlier in the day without unduly affecting the fit. This replacement reduces the number of parameters by one but does not greatly simplify the model or statements of conclusions.

Finally, an informal examination of the standardized residuals reveals no unexpected features or patterns.

The principal conclusions to be drawn are as follows. An occupied high perch is more likely to be occupied by a *grahami* lizard than is an occupied low perch. The ratio of the odds for high versus low perches is an estimated $3.10 = \exp(1.13)$, and this ratio applies under all conditions of shade, perch diameter and time of day. It would be false to conclude from this analysis that *grahami* lizards prefer high perches to low perches. We may,

Table 4.7 Examination of two-factor interactions for lizard data

Model description*	Degrees of freedom	Deviance	First difference
Main effects only	17†	14.20	
Main + <i>T.S</i>	15	12.93	1.27
Main + <i>T.H</i>	15	13.68	0.52
Main + <i>T.D</i>	15	14.16	0.04
Main + <i>S.H</i>	16	11.98	2.22
Main + <i>S.D</i>	16	14.13	0.07
Main + <i>H.D</i>	16	13.92	0.28

* The factors here are time of day (*T*), sunny/shady (*S*), height (*H*) and diameter (*D*).

† Degrees of freedom are reduced by one because no occupied sites were observed for $(i, j, k, l) = (2, 2, 2, 2)$.

Table 4.8 Parameter estimates for the linear logistic model (4.24)

Parameter	Estimate	S.E.
μ	1.945	0.34
<i>H</i> , height > 5ft	1.13	0.26
<i>D</i> , diameter > 2in	-0.76	0.21
<i>S</i> , shady	-0.85	0.32
<i>T</i> (2), mid-day	0.23	0.25
<i>T</i> (3), late	-0.74	0.30

however, conclude that *grahami* lizards have less aversion to high perches than do *opalinus* lizards, so that an occupied high perch is more likely to contain a *grahami* lizard than an occupied low perch.

Similar conclusions may be drawn regarding the effect of shade, perch diameter and time of day on the probability that an occupied site contains a *grahami* lizard. The odds are largest for small-diameter lofty perches observed in a sunny location at mid-day (or in the morning). In fact, only *grahami* and no *opalinus* lizards were observed under these conditions. Because there is no interaction among the effects, the odds are smallest for the converse factor combinations.

These conclusions differ from those of Fienberg (1970b) and Bishop *et al.* (1975), who found an interaction between *H* and *D* and between *S* and *T* regarding their effect on species' preferences. The principal reason for this difference appears to be the fact

that these authors attempted to consider several unrelated issues simultaneously using only a single model, and did not condition on the totals m_{ijkl} , which are regarded as ancillary in the analysis given here.

4.7 Bibliographic notes

The statistical literature on the analysis of discrete data is very extensive and there is a wealth of excellent text-books treating the subject from a number of different angles. Cox (1970) offers a good introduction to the subject and combines a pleasant blend of theory and application in a slim volume. Plackett (1981) is a good introductory text covering much of the material in this chapter and in the following three chapters, but with a slightly different emphasis. Breslow and Day (1980) concentrate on applications in cancer research. Fleiss (1981) discusses applications in the health sciences generally. Haberman (1978, 1979) concentrates mainly on social-science applications. Engel (1987) gives an extensive discussion of over-dispersion.

There is some overlap with the survival-theory literature, where success is sometimes defined rather arbitrarily as two-year or five-year survival: see, for example, Kalbfleisch and Prentice (1980) or Cox and Oakes (1984).

Other books dealing partially or wholly with binary data include Adena and Wilson (1982), Aickin (1983), Armitage (1971), Ashton (1972), Bishop, Fienberg and Holland (1975), Bock (1975), Everitt (1977), Fienberg (1980), Finney (1971), Gokhale and Kullback (1978), Maxwell (1961), Plackett (1981) and Upton (1978).

4.8 Further results and exercises 4

4.1 Suppose that Y_1, \dots, Y_m are independent Bernoulli random variables for which

$$\text{pr}(Y_i = 0) = 1 - \pi \quad \text{and} \quad \text{pr}(Y_i = 1) = \pi.$$

Show that any fixed sequence comprising y ones and $m - y$ zeros has probability $\pi^y(1 - \pi)^{m-y}$. Hence deduce that the total $Y =$

$Y_1 + \dots + Y_m$ has the binomial distribution (4.2) with index m and parameter π .

4.2 Suppose that $Y_1 \sim B(m_1, \pi)$ and $Y_2 \sim B(m_2, \pi)$ are independent. Deduce from Exercise 4.1 that $Y_* \sim B(m_*, \pi)$.

4.3 Suppose that $Y_1 \sim B(m_1, \pi_1)$ and $Y_2 \sim B(m_2, \pi_2)$ are independent. Show that

$$\text{pr}(Y_* = y_*) = (1 - \pi_1)^{m_1} \pi_2^{y_*} (1 - \pi_2)^{m_2 - y_*} P_0(\psi; m_1, m_2, y_*),$$

where $\psi = \pi_1(1 - \pi_2)/\{\pi_2(1 - \pi_1)\}$ is the odds ratio and $P_0(\psi; \cdot)$ is the polynomial in ψ

$$P_0(\psi; m_1, m_2, y_*) = \sum_{j=a}^b \binom{m_1}{j} \binom{m_2}{y_* - j} \psi^j.$$

The range of summation extends from $a = \max(0, y_* - m_2)$ to $b = \min(m_1, y_*)$. Show also that

$$P_0(1; m_1, m_2, y_*) = \binom{m_*}{y_*},$$

which is consistent with the previous exercise.

4.4 Suppose that Y_1, Y_2 are independent Poisson random variables with means μ and $\rho\mu$ respectively. Show that

$$\begin{aligned} Y_* &= Y_1 + Y_2 \sim P(\mu + \rho\mu) \\ Y_1 | Y_* = m &\sim B(m, 1/(1 + \rho)). \end{aligned}$$

Show how you might use this result to test the composite null hypothesis $H_0: \rho = 1$ against the one-sided alternative $H_A: \rho > 1$.

4.5 Let Y_1, \dots, Y_n be independent random variables such that $Y_i \sim B(m, \pi_i)$ and let $Y = \sum Y_i$ be the sum. Show that, given π_1, \dots, π_n ,

$$\begin{aligned} E(Y) &= m_* \bar{\pi} \\ \text{var}(Y) &= m_* \bar{\pi}(1 - \bar{\pi}) - m(n - 1)k_2(\pi) \end{aligned}$$

where $m_* = nm$. Give the expression for $k_2(\pi)$ in terms of π_1, \dots, π_n .

4.6 In the notation of the previous exercise, assume that π_1, \dots, π_n are independent random variables with common mean π and common variance $\tau^2\pi(1-\pi)$. Show that, unconditionally,

$$\begin{aligned} E(Y) &= m.\pi \\ \text{var}(Y) &= m.\pi(1-\pi)\{1 + (m-1)\tau^2\} \end{aligned}$$

Deduce that $0 \leq \tau^2 \leq 1$, so that $\text{var}(Y) \geq m.\pi(1-\pi)$.

4.7 Define

$$\begin{aligned} B(y) &= \binom{m}{y}\pi^y(1-\pi)^{m-y}, \\ P(y) &= e^{-\mu}\mu^y/y!. \end{aligned}$$

Let $\pi = \mu/m$. Show that, for fixed μ , as $m - y \rightarrow \infty$,

$$\frac{B(y)}{P(y)} \simeq \left(\frac{m}{m-y}\right)^{1/2}.$$

4.8 Suppose that $Y \sim B(m, \pi)$ and that m is large. By expanding in a Taylor series, show that the random variable

$$Z = \arcsin\{(Y/m)^{1/2}\}$$

has approximate first two moments

$$\begin{aligned} E(Z) &\simeq \arcsin(\pi^{1/2}) - \frac{1-2\pi}{8\sqrt{m\pi(1-\pi)}} \\ \text{var}(Z) &\simeq (4m)^{-1}. \end{aligned}$$

4.9 Let $K(\theta)$ be a cumulant function such that the r th cumulant of X is the r th derivative of $mK(\theta)$. Let $\mu = mK'(\theta)$ be the mean of X and let $\kappa_2(\mu), \kappa_3(\mu)$ be the variance and third cumulant respectively of X , expressed in terms of μ rather than in terms of θ . Show that

$$\kappa_3(\mu) = \kappa_2(\mu)\kappa'_2(\mu) \quad \text{and} \quad \frac{\kappa_3}{\kappa_2^2} = \frac{d}{d\mu} \log \kappa_2(\mu).$$

Verify that the binomial cumulants have this form with

$$K(\theta) = \log(1 + e^\theta).$$

4.10 Show that if the cumulants of X are all $O(m)$ for large m , then $Y = g(X)$ is approximately symmetrically distributed if $g(\cdot)$ satisfies the second-order differential equation

$$3\kappa_2^2(\mu)g''(\mu) + g'(\mu)\kappa_3(\mu) = 0.$$

Show that if $\kappa_2(\mu)$ and $\kappa_3(\mu)$ are related as in the previous exercise, then

$$g(x) = \int^x \kappa_2^{-1/3}(\mu) d\mu.$$

[N.B. $\kappa_2(\mu)$ is the variance function denoted by $V(\mu)$ in section 2.2: $\kappa_3(\mu)$ is an obvious extension.]

4.11 Find the corresponding equations that give the variance-stabilizing transformation of X .

4.12 Logistic discrimination: Suppose that a population of individuals is partitioned into two sub-populations or groups, G_1 and G_2 , say. It may be helpful to think of G_1 in a epidemiological context as the carriers of a particular virus, comprising $100\pi_1\%$ of the population, and G_2 as the non-carriers. Measurements Z made on individuals have the following distributions in the two groups:

$$\begin{aligned} G_1: \quad Z &\sim N_p(\boldsymbol{\mu}_1, \Sigma) \\ G_2: \quad Z &\sim N_p(\boldsymbol{\mu}_2, \Sigma). \end{aligned}$$

Let \mathbf{z}^* be an observation made on an individual drawn at random from the combined population. The prior odds that the individual belongs to G_1 are $\pi_1/(1 - \pi_1)$. Show that the posterior odds given \mathbf{z}^* are

$$\text{odds}(Y = 1 | \mathbf{Z}^*) = \frac{\pi_1}{1 - \pi_1} \times \exp(\alpha + \boldsymbol{\beta}^T \mathbf{z}^*)$$

where the logistic regression coefficients are given by

$$\begin{aligned} \alpha &= \frac{1}{2}\boldsymbol{\mu}_2^T \Sigma^{-1} \boldsymbol{\mu}_2 - \frac{1}{2}\boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1 \\ \boldsymbol{\beta} &= \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \end{aligned}$$

Comment briefly on the differences between maximum likelihood estimation of α and $\boldsymbol{\beta}$ via the Normal-theory likelihood and estimation via logistic regression. [Efron, 1975].

4.13 Go through the calculations of the previous exercise, replacing the Normal distributions by exponential distributions having the same means.

4.14 Suppose that $Y \sim B(m, e^\lambda / (1 + e^\lambda))$. Show that $m - Y$ also has the binomial distribution and that the induced parameter is $\lambda' = -\lambda$. Consider

$$\tilde{\lambda} = \log\left(\frac{Y + c_1}{m - Y + c_2}\right)$$

as an estimator of λ . Show that, in order to achieve consistency under the transformation $Y \rightarrow m - Y$, we must have $c_1 = c_2$.

4.15 Using the notation of the previous exercise, write

$$Y = m\pi + \sqrt{m\pi(1 - \pi)} Z,$$

where $Z = O_p(1)$ for large m . Show that

$$E\{\log(Y + c)\} = \log(m\pi) + \frac{c}{m\pi} - \frac{1 - \pi}{2m\pi} + O(m^{-3/2}).$$

Find the corresponding expansion for $E\{\log(m - Y + c)\}$. Hence, if $c_1 = c_2 = c$, deduce that

$$E(\tilde{\lambda}) = \lambda + \frac{(1 - 2\pi)(c - \frac{1}{2})}{m\pi(1 - \pi)} + O(m^{-3/2}).$$

[Cox, 1970, section 3.2].

4.16 Suppose that Y_1, \dots, Y_r are independent and that $Y_i \sim B(m_i, \pi)$. Show that the maximum-likelihood estimate is $\hat{\pi} = Y./m.$, and

$$s^2 = \frac{1}{r-1} \sum_i (Y_i - m_i \hat{\pi})^2 / \{m_i \hat{\pi}(1 - \hat{\pi})\}$$

has expectation

$$E(s^2) = \frac{m.}{m. - 1}.$$

Hence show how the estimator (4.22) may be modified to eliminate bias in the null case of no dispersion. [Haldane, 1937].

4.17 Show that if $Y|P \sim B(m, p)$, where P has the beta distribution

$$f_P(p) = p^{\alpha-1}(1-p)^{\beta-1}/B(\alpha, \beta), \quad (0 \leq p \leq 1),$$

then Y has the beta-binomial distribution

$$\text{pr}(Y = y) = \binom{m}{y} \frac{B(\alpha + y, m + \beta - y)}{B(\alpha, \beta)}$$

for $y = 0, \dots, m$ and $\alpha, \beta > 0$. Show that

$$E(Y) = m\pi \quad \text{and}$$

$$\text{var}(Y) = m\pi(1-\pi)\{1 + (m-1)\tau^2\}$$

and express π and τ^2 in terms of α and β . [Crowder, 1978; Plackett, 1981 p.58; Williams, 1982; Engel, 1987].

4.18 Suppose, following the cluster-sampling mechanism described in section 4.5.1, that

$$Y = Z_1 + Z_2 + \dots + Z_{m/k}$$

where $Z_i \sim B(k, \pi_i)$ are independent. Assume in addition that the cluster probabilities are independent random variables satisfying

$$E(\pi_i) = \pi, \quad \text{var}(\pi_i) = \tau_2 \pi(1-\pi), \quad \kappa_3(\pi_i) = \tau_3 \pi(1-\pi)(1-2\pi).$$

Show that the marginal cumulants of Y are

$$E(Y) = m\pi$$

$$\text{var}(Y) = m\pi(1-\pi)\{1 + (k-1)\tau_2\}$$

$$\kappa_3(Y) = m\pi(1-\pi)(1-2\pi)\{1 + 3(k-1)\tau_2 + (k-1)(k-2)\tau_3\}.$$

[With obvious extensions, similar calculations for the fourth cumulant give

$$\begin{aligned} \kappa_4(Y) = m\pi(1-\pi) &\{1 + 7(k-1)\tau_2 + 6(k-1)(k-2)\tau_3 \\ &+ (k-1)(k-2)(k-3)\tau_4\} \\ -6m\pi^2(1-\pi)^2 &\{1 + 6(k-1)\tau_2 + 4(k-1)(k-2)\tau_3 \\ &+ (k-1)(k-2)(k-3)\tau_4 + (k-1)(2k-3)\tau_2^2\}. \end{aligned}$$

breaking the early pattern.]

4.19 Consider the dose-response model

$$g(\pi) = \beta_0 + \beta_1 x.$$

Under the hypothesis that the response probability at x_0 is equal to π_0 , show that the model reduces to

$$g(\pi) = \beta_0(1 - x/x_0) + g(\pi_0)x/x_0.$$

How would you fit such a model using your favourite computer program?

4.20 Let $D(x_0)$ be the residual deviance under the reduced model in the previous exercise. How would you use a graph of $D(x_0)$ against x_0 to construct an approximate confidence set for the parameter x_0 ? For the logistic model, compute this interval using the data in Table 4.3 for $\pi_0 = 0.01$. Compare the answer with that given by (4.19).

4.21 Suppose that in a given population, the proportions of individuals in the various categories are as shown in Table 4.2. In a prospective study, 100 subjects in each of the exposure groups are observed over the requisite period. Find the expected numbers in each of the four cells. Show that the estimate of the log odds-ratio has approximate variance

$$\text{var}(\hat{\Delta}_1) \simeq 0.472.$$

In a retrospective study, 100 cases with the disease and 100 disease-free controls are obtained. Their exposure status is subsequently ascertained. Find the expected numbers in the four cells. Show that the estimate of the log odds-ratio has approximate variance

$$\text{var}(\hat{\Delta}_2) \simeq 0.093.$$

Hence compute the relative efficiency of the retrospective design.

4.22 Show that the logistic density

$$f_X(x) = \exp(x)/[1 + \exp(x)]^2$$

is symmetrical about zero. Find the cumulative distribution function and show that the 100p percentile occurs at

$$x_p = \log(p/(1 - p)).$$

Show that the moment generating function of X is

$$M_X(t) = \pi t / \sin(\pi t) = \Gamma(1+t)\Gamma(1-t).$$

for $-1 < t < 1$. Hence deduce that the even cumulants of X are

$$\kappa_2 = \pi^2/3, \quad \kappa_4 = 2\pi^4/15, \quad \kappa_6 = 16\pi^6/63, \quad \kappa_8 = 16\pi^8/15, \dots$$

Deduce that $\kappa_{2r} \sim 2(2r-1)! \{1 + 2^{-2r}\}$ for large r . Check this approximation numerically for the cumulants listed above.

The exact cumulants are given by the series expansion

$$\kappa_{2r} = 2(2r-1)! \zeta(2r) = 2(2r-1)! \{1 + 2^{-2r} + 3^{-2r} + 4^{-2r} + \dots\},$$

where $\zeta(x)$ is the Riemann zeta function.

4.23 Let X be a unit exponential random variable. Show that the density function of $Y = \log X$ is

$$f_Y(y) = \exp(y - e^y) \quad \text{for } -\infty < y < \infty.$$

Plot the density. Find the cumulative distribution function and obtain an expression for the $100p$ percentile.

Show that the moment generating function of Y is

$$M_Y(t) = \Gamma(1+t).$$

Hence deduce that the cumulants of Y are

$$\kappa_{r+1}(Y) = \psi^{(r)}(1) = (-1)^{r+1} r! \zeta(r+1).$$

Show in particular that the first four cumulants are

$$\kappa_1 = -\gamma \simeq -0.57721, \quad \kappa_2 = \pi^2/6, \quad \kappa_3 = -2.40411, \quad \kappa_4 = \pi^4/15.$$

Comment briefly on the relation between the even cumulants of Y and those of the logistic density.

Table 4.9 Number of eggs recovered after 2 days out of 50 of each type[†]

Adult species											
A			B			C					
egg species			egg species			egg species					
a	b	c	a	b	c	a	b	c			
day 1	25	24	15	day 1	25	15	22	day 1	35	21	28
	26	14	26		31	22	33		36	19	34
	26	24	32		24	12	30		33	16	31
day 2	29	14	32	day 2	14	8	13	day 2	24	24	23
	28	13	19		18	12	20		38	24	27
	27	19	16		—	—	—		34	36	27
day 3	26	10	13	day 3	13	6	14				
	20	7	15		18	11	19				
	14	14	23		8	5	12				

[†]Data courtesy of Mr S. Teleky, University of Chicago.

4.24 Show how you would use your friendly computer program to compute the approximate conditional mean and variance of Pearson's statistic using the formulae given at the end of section 4.4.5. [Hint: express $\sum_j V_{ij}(1 - 2\pi_j)$ as the vector of fitted values in a supplementary weighted linear regression problem. This step is unnecessary if your friendly program permits matrix multiplication.]

4.25 Beetles of the genus *Tribolium* are cannibalistic in the sense that adults eat the eggs of their own species as well as those of closely related species. Any species whose adults can recognize and avoid eggs of their own species while foraging has a distinct evolutionary advantage. Table 4.9 presents the results of one experiment conducted at the University of Chicago by Mr S. Teleky of the Department of Evolutionary Biology. The aim of this study was to determine whether any of the three *Tribolium* species, *castaneum* (A), *confusum* (B), or *madens* (C) has evolved such an advantage.

The experimental procedure used was to isolate a number of adult beetles of the same species and to present them with a vial of 150 eggs – 50 of each type – the eggs being thoroughly mixed to ensure a uniform distribution on the vial. The number of eggs of each type remaining after two days was counted and recorded

and is displayed in Table 4.9. Eggs are coded here using the lower-case letters of the adult species. Typically, several such experiments with the same adult species were run in parallel, the adults for each experiment being chosen from a large population of that species. Thus, for adult species *A*, three experiments were run in parallel beginning on each of three days. Days 1, 2 and 3 are not necessarily consecutive, nor is day 1 for species *A* the same as day 1 for species *B* or *C*.

Analyse the data bearing in mind the objective of the experiment and the design of the experiment. Make due allowance for over-dispersion in the computation of standard errors and confidence intervals. Is there any evidence that any of the three adult species has evolved a preference for eggs of the other species?

4.26 The data shown in Table 4.10 were collected by Sir Francis Galton as part of his study of natural inheritance—in this case the study of the inheritance of eye colour in human populations.

1. Set up a six-level factor, P , one level corresponding to each of the distinguishable eye-colour combinations of the two parents.
2. Set up the corresponding factor, G , for the distinguishable eye-colour combinations of the grandparents. How many levels does this factor have?
3. Fit the linear logistic model P , treating the number of light-eyed children as the binomial response. Examine the standardized residuals and set aside any obviously discrepant points.
4. Re-fit the previous model to the remaining data. Compute the fitted probabilities for all eye-colour combinations of the two parents. Arrange these fitted probabilities in a 3×3 table. Comment on any marked trends or other patterns.
5. Add the factor G to the previous model. Look for trends or other patterns among the levels of G . What evidence is there for a grandparent effect above and beyond the parental effect?
6. Outliers are often caused by transposing digits or otherwise misrecording the data. What alternative explanation can you offer for the most discrepant points in this example?
7. Is there any evidence of over-dispersion? Estimate the dispersion parameter.
8. What additional information could be extracted from these data if the eye-colours of the father and mother were separately recorded? Comment on the relevance of this information

Table 4.10 Number of light-eyed children in each of 78 families of not less than six brothers or sisters each, classified by eye-colour of parents and grandparents.[†]

Number of parents			Number of grandparents			Total children	Light-eyed children
Light	Hazel	Dark	Light	Hazel	Dark		
2	0	0	4	0	0	6	6
2	0	0	4	0	0	6	6
2	0	0	4	0	0	6	6
2	0	0	4	0	0	6	5
2	0	0	4	0	0	7	7
2	0	0	4	0	0	7	7
2	0	0	4	0	0	7	7
2	0	0	4	0	0	7	7
2	0	0	4	0	0	7	7
2	0	0	4	0	0	8	8
2	0	0	4	0	0	8	8
2	0	0	4	0	0	8	8
2	0	0	4	0	0	8	8
2	0	0	4	0	0	8	7
2	0	0	4	0	0	8	7
2	0	0	4	0	0	12	12
2	0	0	3	1	0	7	7
2	0	0	3	1	0	10	4
2	0	0	3	1	0	12	12
2	0	0	3	0	1	7	6
2	0	0	3	0	1	8	8
2	0	0	3	0	1	9	9
2	0	0	3	0	1	9	9
2	0	0	3	0	1	9	7
2	0	0	3	0	1	10	10
2	0	0	2	2	0	7	7
2	0	0	2	2	0	10	9
2	0	0	2	1	1	6	6
2	0	0	2	1	1	10	10
0	2	0	2	1	1	7	4
0	2	0	2	0	2	8	5
0	0	2	3	0	1	6	2
0	0	2	2	0	2	9	1
0	0	2	1	0	3	6	1
0	0	2	1	0	3	11	3
0	0	2	1	1	2	6	0
0	0	2	1	1	2	7	4

Continued

Table 4.10 *Continued.*

Number of parents			Number of grandparents			Total	Light-eyed
Light	Hazel	Dark	Light	Hazel	Dark	children	children
1	1	0	3	1	0	6	6
1	1	0	3	1	0	7	6
1	1	0	3	1	0	8	6
1	1	0	3	1	0	9	7
1	1	0	3	1	0	11	10
1	1	0	3	0	1	9	6
1	1	0	3	0	1	11	7
1	1	0	2	2	0	7	6
1	1	0	2	2	0	9	9
1	1	0	2	2	0	11	1
1	1	0	2	0	2	6	6
1	1	0	2	0	2	6	4
1	1	0	2	0	2	8	5
1	1	0	2	0	2	9	7
1	1	0	2	1	1	6	6
1	1	0	2	1	1	10	9
1	1	0	1	3	0	9	4
1	1	0	1	1	2	8	5
1	0	1	4	0	0	7	3
1	0	1	3	0	1	6	4
1	0	1	3	0	1	7	3
1	0	1	3	0	1	8	6
1	0	1	3	0	1	8	5
1	0	1	3	0	1	8	4
1	0	1	3	0	1	9	6
1	0	1	3	0	1	9	5
1	0	1	2	0	2	6	5
1	0	1	2	0	2	6	3
1	0	1	2	0	2	8	4
1	0	1	2	0	2	10	2
1	0	1	2	0	2	14	9
1	0	1	2	1	1	7	5
1	0	1	1	2	1	7	3
1	0	1	1	1	2	7	4
1	0	1	1	0	3	8	4
1	0	1	1	0	3	8	3
1	0	1	0	1	3	6	3
0	1	1	2	0	2	6	3
0	1	1	2	1	1	9	4
0	1	1	1	0	3	13	8
0	1	1	0	4	0	7	2

†Source: Galton (1889, p.216–217).

- (a) from a biological viewpoint and (b) from a sociological viewpoint.
9. Fit the linear logistic model G . Compute the fitted probability for each level of G . Label the levels of G appropriately and comment on any trends or patterns in the fitted probabilities.
10. Examine the residuals from the previous model. Comment briefly on any unusual patterns, particularly in families 32 and 56.

4.27 Using the notation of section 4.4.3 in which $Y_i \sim B(m_i, \pi_i)$, let $H_0 \subset H_1$ denote two nested models for the probability vector $\boldsymbol{\pi}$, with deviances $D(\mathbf{y}, \hat{\boldsymbol{\pi}}_0)$ and $D(\mathbf{y}, \hat{\boldsymbol{\pi}}_1)$ respectively. Show that, in the case of linear logistic models, the deviances satisfy the Pythagorean relationship

$$D(\mathbf{y}, \hat{\boldsymbol{\pi}}_0) = D(\mathbf{y}, \hat{\boldsymbol{\pi}}_1) + D(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\pi}}_0).$$

Hence deduce that for logistic models, but not otherwise, $D(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\pi}}_0)$ is the likelihood-ratio statistic for testing H_0 against H_1 as alternative.

4.28 In the previous exercise, suppose that H_0 and H_1 denote a constant and a single-factor model respectively. Show that the fitted values and the deviance functions are then independent of the link used for model specification. Show also that weighted squared Euclidean distance with weights m_i satisfies the Pythagorean relationship. What other discrepancy functions satisfy the Pythagorean relationship in this special case? [Efron, 1978].

4.29 The asymptotic bias of the components of $\hat{\boldsymbol{\beta}}$ in linear logistic models is given by

$$E(\hat{\boldsymbol{\beta}}^r - \boldsymbol{\beta}^r) \simeq -\frac{1}{2} \boldsymbol{\kappa}^{r,s} \boldsymbol{\kappa}^{t,u} \boldsymbol{\kappa}_{s,t,u},$$

using the index notation of McCullagh (1987, p. 209), in which $\boldsymbol{\kappa}^{r,s}$ is the inverse Fisher information matrix. Express $\boldsymbol{\kappa}_{s,t,u}$ in terms of the components of the model matrix.

For small $\boldsymbol{\beta}$, justify the approximation

$$E(\hat{\boldsymbol{\beta}}) \simeq \boldsymbol{\beta} \times (1 + p/m.),$$

showing that the bias vector is approximately collinear with the parameter vector.

4.30 Let R_i be the unobserved true response for unit i with $\pi_i^* = \text{pr}(R_i = 1)$ satisfying the linear logistic model

$$\text{logit}(\pi_i^*) = \boldsymbol{\beta}^T \mathbf{x}_i.$$

Suppose that the observed response is subject to mis-classification as follows.

$$\text{pr}(Y_i = 1 | R_i = 0) = \delta_i$$

$$\text{pr}(Y_i = 0 | R_i = 1) = \epsilon_i.$$

Show that if the mis-classification errors satisfy

$$\frac{\delta_i}{\epsilon_i} = \frac{\pi_i^*}{1 - \pi_i^*},$$

then the observed response probability $\pi_i = \text{pr}(Y_i = 1)$ satisfies

$$\text{logit}(\pi_i) = \boldsymbol{\beta}^T \mathbf{x}_i.$$

Discuss briefly the plausibility of the assumption concerning the mis-classification probabilities. [Bross, 1954; Ekholm and Palmgren, 1982; Palmgren, 1987; Copas, 1988].

CHAPTER 5

Models for polytomous data

5.1 Introduction

If the response of an individual or item in a study is restricted to one of a fixed set of possible values, we say that the response is polytomous. The k possible values of Y are called the response categories. Often the categories are defined in a qualitative or non-numerical way. A familiar example is the classification of blood types, with unambiguous but qualitative categories O, A, B, AB. Another example is the ILO scale, 0/0, 0/1, ..., 3/3, used for classifying chest X-ray images according to apparent severity of lung disease. These categories, defined rather arbitrarily using 'standard' reproductions, are not devoid of ambiguity. Other instances of the type of response considered in this chapter are rating scales used in food testing, measures of mental and physical well-being, and many variables arising in social science research which are, of necessity, not capable of precise measurement.

We need to develop satisfactory statistical models that distinguish several types of polytomous response or measurement scale. For instance, if the categories are ordered, there is no compelling reason for treating the extreme categories in the same way as the intermediate ones. However, if the categories are simply an unstructured collection of labels, there is no reason a priori to select a subset of the categories for special treatment. Considerations such as these lead us to consider qualitatively different classes of link functions for different types of response scale. Whatever the nature of the scale, we may talk without ambiguity about the response probabilities π_1, \dots, π_k . If the categories are ordered, however, we may prefer to work with the cumulative response probabilities

$$\gamma_1 = \pi_1, \quad \gamma_2 = \pi_1 + \pi_2, \dots, \gamma_k \equiv 1.$$

Obviously, cumulative probabilities are not defined unless the category order is unambiguous. It makes little sense to work with a model specified in terms of γ_j if the response categories are not ordered.

5.2 Measurement scales

5.2.1 General points

Measurement scales can be classified at a number of levels. At one level, we may distinguish between *pure scales* and *compound scales*. Bivariate responses are perhaps the simplest instances of compound measurement scales. One can contemplate bivariate responses in which one response is ordinal and the other binary or even continuous. Other examples of compound scales are discussed in section 5.2.5. Among the spectrum of pure measurement scales, we may identify the following major types:

1. *nominal scales* in which the categories are regarded as exchangeable and totally devoid of structure.
2. *ordinal scales* in which the categories are ordered much like the ordinal numbers, ‘first’, ‘second’,.... In this context it does not ordinarily make sense to talk of ‘distance’ or ‘spacing’ between ‘first’ and ‘second’ nor to compare ‘spacings’ between pairs of response categories.
3. *interval scales* in which the categories are ordered and numerical labels or scores are attached. The scores are treated as category averages, medians or mid-points. Differences between scores are therefore interpreted as a measure of separation of the categories.

Cardinal scales require quite different kinds of models, such as those discussed in Chapters 3, 6 and 8, and are not considered here. Binary measurements are special cases of all of the above in which $k = 2$. The distinction between ordinal, interval and nominal does not then arise.

In applications, the distinction between nominal and ordinal scales is usually but not always clear. For instance responses relating to perception of food quality — excellent, good, ..., bad, appalling — are clearly ordinal. Responses concerning preferences for newspaper or television programme would usually be treated

as nominal, at least initially. Political hue and perceived quality may well be sufficient grounds for the subsequent examination of particular contrasts. Hair colour and eye colour can be ordered to a large extent on the grey-scale from light to dark and are therefore ordinal, although the relevance of the order may well depend on the context. Otherwise, unless there is a clear connection with the electromagnetic spectrum or a grey-scale, colours are best regarded as nominal.

5.2.2 *Models for ordinal scales*

We consider ordinal scales first, mainly because these occur more frequently in applications than the other types. In many of these applications such as food-testing, classification of radiographs, determination of physical or mental well-being and so on, the choice and definition of response categories is either arbitrary or subjective. It is essential, therefore, if we are to arrive at valid conclusions, that the nature of those conclusions should not be affected by the number or choice of response categories. As a consequence, if a new category is formed by combining adjacent categories of the old scale, the form of the conclusions should be unaffected. Of course, the amalgamation of response categories in this way will normally reduce the available information, change the estimate, the attained significance level and so on. The important point is that the same parameter is being measured however many categories are combined. This is an important non-mathematical point that is difficult to make mathematically rigorous: it is not simply a matter of retaining the same Greek letter after category combination.

Such considerations lead fairly directly to models based on the cumulative response probabilities $\gamma_j = \text{pr}(Y \leq j)$ rather than the category probabilities π_j . The two sets of probabilities are equivalent, but simple models for the cumulative probabilities are likely to have better properties for ordinal response scales than equally simple models based on the category probabilities. In particular, linear models using the logistic scale, $\log\{\gamma_j/(1 - \gamma_j)\}$, or the complementary log-log scale, $\log\{-\log(1 - \gamma_j)\}$ are found to work well in practice (McCullagh, 1980).

The simplest models in this class involve parallel regressions on

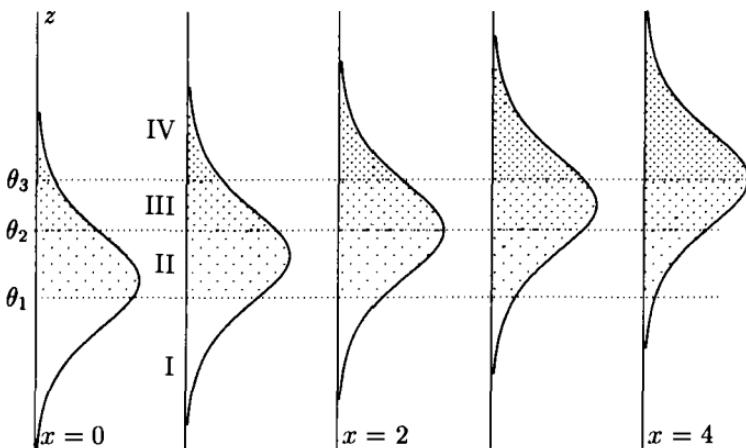


Fig. 5.1a. Diagram showing how the response probabilities for the logistic model (5.1) vary with x when $\beta > 0$. Response categories are represented as four contiguous intervals of the z -axis. Higher-numbered categories have greater shade density.

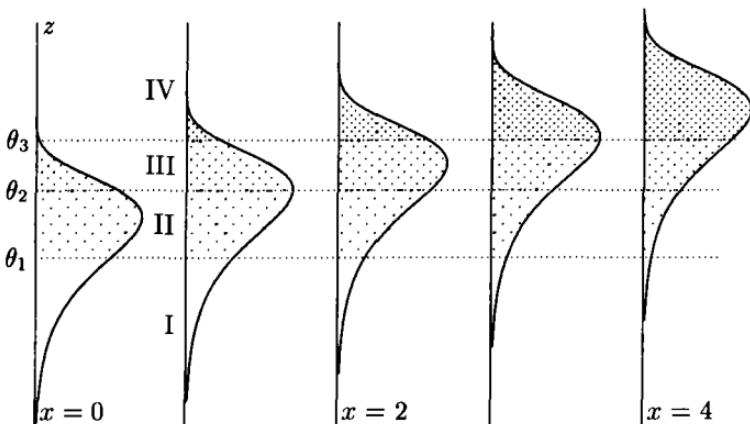


Fig. 5.1b. Diagram showing how the probabilities for the four response categories in the complementary-log-log model (5.3) vary with x when $\beta > 0$. $\pi_1(x)$ and $\pi_4(x)$ each change by a factor of 10 or more, whereas $\pi_3(x)$ is almost constant over $1 \leq x \leq 4$.

the chosen scale, such as

$$\log\{\gamma_j(\mathbf{x})/(1 - \gamma_j(\mathbf{x}))\} = \theta_j - \boldsymbol{\beta}^T \mathbf{x}, \quad j = 1, \dots, k-1 \quad (5.1)$$

where $\gamma_j(\mathbf{x}) = \text{pr}(Y \leq j | \mathbf{x})$ is the cumulative probability up to and including category j , when the covariate vector is \mathbf{x} . Model

(5.1) is known as the proportional-odds model because the ratio of the odds of the event $Y \leq j$ at $\mathbf{x} = \mathbf{x}_1$ and $\mathbf{x} = \mathbf{x}_2$ is

$$\frac{\gamma_j(\mathbf{x}_1)/(1 - \gamma_j(\mathbf{x}_1))}{\gamma_j(\mathbf{x}_2)/(1 - \gamma_j(\mathbf{x}_2))} = \exp\{-\boldsymbol{\beta}^T(\mathbf{x}_1 - \mathbf{x}_2)\}, \quad (5.2)$$

which is independent of the choice of category (j). In particular, if \mathbf{x} is an indicator variable for two treatment groups, T_1 and T_2 , (5.2) may be written as

$$\frac{\text{odds}(Y \leq j | T_1)}{\text{odds}(Y \leq j | T_2)} = \exp(-\Delta), \quad j = 1, \dots, k-1,$$

where Δ measures the treatment effect. The negative sign in (5.1) is a convention ensuring that large values of $\boldsymbol{\beta}^T \mathbf{x}$ lead to an increase of probability in the higher-numbered categories. Both $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ in (5.1) are treated as unknown and $\boldsymbol{\theta}$ must satisfy $\theta_1 \leq \theta_2 \leq \dots \leq \theta_{k-1}$.

For the complementary log-log link, the model corresponding to (5.1) is

$$\log[-\log\{1 - \gamma_j(\mathbf{x})\}] = \theta_j - \boldsymbol{\beta}^T \mathbf{x}, \quad j = 1, \dots, k-1 \quad (5.3)$$

which is known as the proportional-hazards model (Cox, 1972a; McCullagh, 1980). In all of these models we must have $\theta_1 \leq \theta_2 \leq \dots \leq \theta_{k-1}$ to ensure that the probabilities are non-negative.

Both (5.1) and (5.3) correspond to the same model formula and the same response variable, but with different choice of link function. The response is the set of cumulative observed proportions (or totals). Apart from choice of sign, the model formula in both cases is

$$R + \mathbf{x}$$

where R is the response factor having $k-1$ levels and \mathbf{x} is itself a model formula, not involving R , for the covariates used. The observations in this instance are not independent, but that is an aspect of the random part of the model and is considered in section 5.4.

Model (5.1) may be derived from the notion of a tolerance distribution or an underlying unobserved continuous random variable Z , such that $Z - \boldsymbol{\beta}^T \mathbf{x}$ has the standard logistic distribution. If the

unobserved variable lies in the interval $\theta_{j-1} < Z \leq \theta_j$ then $y = j$ is recorded. Thus we find

$$\begin{aligned}\text{pr}(Y \leq j) &= \text{pr}(Z \leq \theta_j) = \text{pr}(Z - \boldsymbol{\beta}^T \mathbf{x} \leq \theta_j - \boldsymbol{\beta}^T \mathbf{x}) \\ &= \frac{\exp(\theta_j - \boldsymbol{\beta}^T \mathbf{x})}{1 + \exp(\theta_j - \boldsymbol{\beta}^T \mathbf{x})}.\end{aligned}$$

Model (5.3) has a similar derivation based on the extreme-value distribution.

Figure 5.1 illustrates the way in which the response probabilities $\pi_j(x)$ vary with x for the single-variable case in which $\beta > 0$. In that case the larger the value of x the greater the probability of falling in the highest-numbered category. The probability for the lowest-numbered category decreases with x . For intermediate categories, the probability increases with x up to a certain point and thereafter decreases. Over certain ranges of x , the probability for some of the intermediate categories is almost constant: over the same range the probabilities for the extreme categories may change quite appreciably.

It is sometimes claimed that (5.1), (5.3) and related models are appropriate only if there exists a latent variable Z . This claim seems to be too strong and, in any case, the existence of Z is usually unverifiable in practice.

Suppose by way of extension that Z has the logistic distribution with mean $\boldsymbol{\beta}^T \mathbf{x}$ and scale parameter $\exp(\boldsymbol{\tau}^T \mathbf{x})$. In other words, $(Z - \boldsymbol{\beta}^T \mathbf{x})/\exp(\boldsymbol{\tau}^T \mathbf{x})$ has the standard logistic distribution. We are then led by the same argument to consider non-linear models of the particular form

$$\text{logit } \gamma_j(\mathbf{x}) = \frac{\theta_j - \boldsymbol{\beta}^T \mathbf{x}}{\exp(\boldsymbol{\tau}^T \mathbf{x})}. \quad (5.4)$$

This model is not of the generalized linear type, but nonetheless it is worthy of serious consideration. In the numerator, $\boldsymbol{\beta}^T \mathbf{x}$ plays the role of linear predictor for the mean and in the denominator $\boldsymbol{\tau}^T \mathbf{x}$ plays the role of linear predictor for the dispersion or variance. Two model formulae are required to specify (5.4) in its most general form. The numerator corresponds to the formula $R + \mathbf{x}$ as in (5.1) and (5.3). The denominator corresponds to an arbitrary model formula not involving R , which may differ from the formula in the numerator.

If, as in (5.2), \mathbf{x} is an indicator variable for treatment, then we have

$$\begin{aligned}\frac{\text{odds}(Y \leq j | T_1)}{\text{odds}(Y \leq j | T_2)} &= \exp\left(\frac{\theta_j - \beta_1}{\sigma_1} - \frac{\theta_j - \beta_2}{\sigma_2}\right) \\ &= \exp\left(\frac{\beta_2 - \beta_1}{\sigma_2}\right) \times \exp\left\{\theta_j\left(\frac{1}{\sigma_1} - \frac{1}{\sigma_2}\right)\right\},\end{aligned}$$

where $\sigma_i = \exp(\tau x_i)$ is the scale parameter for the i th treatment group. Thus the odds ratio is increasing in j if $\sigma_1 < \sigma_2$ and decreasing otherwise. Model (5.4) is useful for testing the proportional-odds assumption against the alternative that the odds ratio is systematically increasing or systematically decreasing in j .

Other link functions can be used in (5.4) in place of the logistic function.

Models in which the $k - 1$ regression lines are not parallel can be specified by writing

$$\theta_j + \boldsymbol{\beta}_j^T \mathbf{x}$$

in place of the right side of (5.1) and (5.3). The corresponding model formula is $R + R.\mathbf{x}$, meaning that the slopes vary, though not necessarily in any systematic way, with the levels of R . The usefulness of non-parallel regression models is limited to some extent by the fact that the lines eventually must intersect. Negative fitted values are then unavoidable for some values of \mathbf{x} , though perhaps not in the observed range. If such intersections occur in a sufficiently remote region of the \mathbf{x} -space, this flaw in the model need not be serious.

5.2.3 Models for interval scales

We now turn our attention to measurement scales of a slightly different type where the categories are ordered, but in a stronger or more rigid sense than that discussed in the previous section. Interval scales are distinguished by the following properties:

1. The categories are of interest in themselves and are not chosen arbitrarily.
2. It does not normally make sense to form a new category by amalgamating adjacent categories.
3. Attached to the j th category is a cardinal number or score, s_j , such that the difference between scores is a measure of distance between or separation of categories.

Property 2 is essential because if we were to combine two categories, we would need an algorithm for calculating the score for the new category. Further, the derived model with the new scores should be consistent with the old model in much the same way that the proportional-odds model (5.1) behaves consistently when categories are combined. In particular, the scores for the remaining categories should be unaffected. These properties are difficult to achieve.

Genuine interval scales having these three properties are rare in practice because, although properties 1 and 2 may be satisfied, it is rare to find a response scale having well-determined cardinal scores attached to the categories. Grouped continuous measurements, on the other hand, may satisfy property 3, but usually not 1 or 2. Nevertheless, it may occasionally be helpful to use artificial scores — usually the first k integers — and to treat these as ~~cardinal~~ rather than ordinal.

At this stage, we have three options for model construction. The first is to work with the cumulative response probabilities and, if necessary, to make suitable adaptations of the proportional-odds and related models. For instance, in (5.1) we might consider modelling the cut-points θ_j as functions of the scores. To begin, we might consider expressing θ_j as

$$\theta_j = \zeta_0 + \zeta_1 \left(\frac{s_j + s_{j+1}}{2} \right)$$

for unknown coefficients ζ_0 and $\zeta_1 > 0$. On balance, this seems unhelpful because the 'cut-points' θ_j are ordinarily considered to be incidental parameters of little interest in themselves. More interesting is the possibility of modelling departures from the proportional-odds assumption by allowing certain systematic deviations from parallelism. The most obvious way to achieve this is to replace $\beta^T x$ in (5.1) by

$$\beta^T x + \zeta^T x(c_j - \bar{c}) \quad (5.5)$$

where c_j is a suitable function of the scores. Two possibilities are

$$c_j = \frac{s_j + s_{j+1}}{2} \quad \text{and} \quad c_j = \text{logit} \left(\frac{s_j + s_{j+1}}{2s} \right).$$

Different choices may be more suitable for other link functions. There is a certain qualitative similarity between (5.5) and the effect achieved in (5.4) without using scores.

The second option is to examine the matrix of probabilities $\{\pi_j(\mathbf{x}_i)\}$ or the matrix of log probabilities

$$\eta_j(\mathbf{x}_i) = \log \pi_j(\mathbf{x}_i), \quad j = 1, \dots, k; i = 1, \dots, n$$

and to decompose $\eta_j(\mathbf{x}_i)$ into a small number of effects or contrasts in much the same way as is done in regression and analysis-of-variance problems. In going from the log probabilities to the probabilities it must be borne in mind that $\sum_j \pi_j(\mathbf{x}_i) = 1$ for each i . The inverse transformation is best written in the form

$$\pi_j(\mathbf{x}_i) = \frac{\exp\{\eta_j(\mathbf{x}_i)\}}{\sum_j \exp\{\eta_j(\mathbf{x}_i)\}}.$$

As a consequence, $\{\eta_j(\mathbf{x}_i)\}$ and $\{\eta_j(\mathbf{x}_i) + \alpha_i\}$ represent the same set of probabilities and fitted values.

The simplest model, that of ‘independence’ or ‘no covariate effect’ may be written as $\pi_j(\mathbf{x}_i) = \pi_j$ or $\eta_j(\mathbf{x}_i) = \eta_j$ or

$$\eta_j(\mathbf{x}_i) = \eta_j + \alpha_i. \quad (5.6)$$

In purely formal terms, (5.6) is equivalent to the model formula

column + row

for the log probabilities, where ‘column’ is a k -level response factor indexed by j , and ‘row’ is a factor with n levels indexed by i .

In order to model departures from (5.6), we may suppose that the effect of the covariate is to increase the probability or log probability in those categories for which the scores are highest. Perhaps the simplest model that achieves this effect is

$$\eta_j(\mathbf{x}_i) = \eta_j + (\boldsymbol{\beta}^T \mathbf{x}_i) s_j + \alpha_i. \quad (5.7)$$

The corresponding model formula is

$$\text{column} + \text{score.x} + \text{row}, \quad (5.8)$$

where `score.x` represents p covariates whose i, j components are $\mathbf{x}_i s_j$. In most applications, `x` is itself a model formula not involving the response factor ‘column’.

One interpretation of (5.7) is that a unit change in $\beta^T \mathbf{x}$ changes the log probabilities from η_j to $\eta_j + s_j$. Consequently, the relative odds for category j over category j' are changed from

$$\frac{\pi_j}{\pi_{j'}} = \exp(\eta_j - \eta_{j'}) \quad \text{to} \quad \exp(\eta_j - \eta_{j'} + s_j - s_{j'}).$$

In other words, the relative odds are increased multiplicatively by the factor

$$\exp(s_j - s_{j'})$$

per unit increase in the combination $\beta^T \mathbf{x}$.

For a two-way table with one response variable and one explanatory factor (5.7) reduces to

$$\eta_{ij} = \eta_j + \alpha_i + \zeta_i s_j.$$

If, further, the explanatory factor has ordered levels, we may select for special consideration the linear contrast $r_i = i - (n+1)/2$, or other suitable contrast. The reduced model is then the same as above with ζ_i replaced by βr_i , namely

$$\eta_{ij} = \eta_j + \alpha_i + \beta r_i s_j.$$

This is also known as the linear \times linear-interaction model, first proposed by Birch (1963).

The third option for model building and significance testing is to reverse the roles of the vector of scores $s = (s_1, \dots, s_k)$ and the vector of counts $y = (y_1, \dots, y_k)$. Instead of regarding y as the response and s as a contrast of special interest, we may regard the observed score as the response and y as the set of observed multiplicities or weights. Thus, if $k = 4$, $y = (5, 7, 10, 3)$ is equivalent to 25 observations on S , namely $S = s_1$ five times, $S = s_2$ seven times, $S = s_3$ ten times and $S = s_4$ three times. On the assumption that the mean observed score is linearly related to the covariates, we have

$$E(S | \mathbf{x}_i) = \sum_j \pi_j(\mathbf{x}_i) s_j = \beta^T \mathbf{x}_i.$$

This is an incompletely specified model because the parameters β determine only the linear combination $\sum_j \pi_j s_j$ and not the individual cell probabilities themselves. Note also the unsatisfactory

property that $E(S|\mathbf{x}_i)$ must lie between s_1 and s_k whereas $\beta^T \mathbf{x}_i$ is not similarly restricted. Despite these drawbacks, useful and interesting conclusions can frequently be drawn from an analysis of the observed mean scores

$$\bar{S}_i = \sum_j s_j y_{ij} / m_i.$$

In particular, if there are only two treatment groups, with observed counts $\{y_{1j}, y_{2j}\}$, we may use as test statistic the standardized difference

$$T = \frac{\bar{S}_1 - \bar{S}_2}{\sqrt{\left(\sum \tilde{\pi}_j s_j^2 - \left(\sum \tilde{\pi}_j s_j\right)^2\right) \left(\frac{1}{m_1} + \frac{1}{m_2}\right)}}$$

where $\tilde{\pi}_j = y_{.j}/m_{..}$. Under the null hypothesis of no treatment effect, and provided that the observations have independent multinomial distributions, T is approximately standard Normal. This statistic is due to Yates (1948) and Armitage (1955).

5.2.4 Models for nominal scales

If the scale is purely nominal, we are forced to work with the category probabilities, π_j , directly. By the same argument used in the previous section, it is more convenient to work with logarithmic probabilities η_j given by

$$\pi_j = \exp(\eta_j) / \sum_j \exp(\eta_j), \quad \text{for } j = 1, \dots, k.$$

The aim is to describe how the vector (η_1, \dots, η_k) is affected by changes in the covariates. In doing so, we must bear in mind that $\boldsymbol{\eta}$ and $\boldsymbol{\eta} + \mathbf{c}$ represent the same probabilities and fitted values.

In the absence of scores, the most general log-linear model has the form

$$\eta_j(\mathbf{x}_i) = \eta_j(\mathbf{x}_0) + \beta_j^T (\mathbf{x}_i - \mathbf{x}_0) + \alpha_i \quad (5.9)$$

for $j = 1, \dots, k$. In this expression, $\eta_j(\mathbf{x}_0)$ is the set of base-line log probabilities and β_j is the change in the j th log probability per unit change in each of the components of \mathbf{x} . To be more precise,

the odds in favour of category j over category j' are increased by the factor

$$\frac{\pi_j(\mathbf{x})}{\pi_{j'}(\mathbf{x})} = \frac{\pi_j(\mathbf{x}_0)}{\pi_{j'}(\mathbf{x}_0)} \times \exp\{(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'})^T(\mathbf{x} - \mathbf{x}_0)\}.$$

Thus contrasts among the vectors $\boldsymbol{\beta}_j$ are of interest rather than the vectors themselves.

Note that in (5.7), the use of response category scores enables us to model the change in the log probabilities for all k response cells using a single covariate vector $\boldsymbol{\beta}$. In the absence of scores, it is necessary in (5.9) to use k covariate vectors, $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_k$. Since only contrasts among the $\boldsymbol{\beta}_j$ are estimable, we may set $\boldsymbol{\beta}_1 = 0$. The net result is that the nominal response model (5.9) contains many more parameters than (5.7) in order to achieve a similar effect.

The model formula for (5.9) is

column + column.x + row,

which is the same as (5.8) with the quantitative variable ‘score’ replaced by the response factor ‘column’. As before, \mathbf{x} may itself be a complicated model formula not involving ‘column’.

5.2.5 Nested or hierarchical response scales

It is difficult to identify precisely the characteristics that distinguish a nested or hierarchical response scale from the types previously discussed. The following examples serve that purpose and show that nested classifications occur in a large number of diverse applications.

Example 1 : A study of mortality due to radiation. Suppose that, in a study of the effects of radiation, exposed and non-exposed individuals are classified at the end of the study period as dead or alive. Further information is available regarding the cause of death, at least to the extent that death can be attributed to a single cause. The nature of the study requires that deaths be classified as ‘due to cancer’ or ‘due to other causes’. At a third stage, cancer deaths are sub-divided into ‘leukaemia deaths’ and ‘deaths from other cancers’. The four mutually exclusive response categories are therefore

1. alive,
2. death from causes other than cancer,
3. death from cancers other than leukaemia,
4. death from leukaemia.

Figure 5.2 emphasizes the nested structure of the responses.

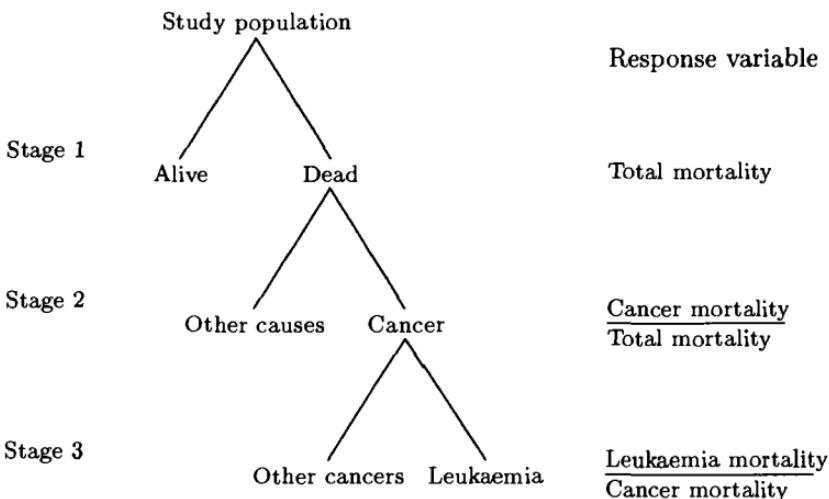


Fig. 5.2. *Hierarchical classification used in the study of radiation effects.*

Here it is probably most appropriate to make a separate study of each of the three response variables, one corresponding to the dichotomy at each level of the hierarchy:

1. total mortality,
2. cancer mortality as a proportion of total mortality, and
3. leukaemia mortality as a percentage of cancer mortality.

Each of these variables may be affected by exposure, but perhaps in quite different ways. For example exposure might have a marked effect on the incidence of leukaemia (or thyroid cancer) without having much effect on total mortality or on the incidence of all cancers.

Example 2 : Fertility of lactating cows. There is some evidence to support the claim that a winter diet containing a high proportion of red clover has the effect of reducing the fertility of milch cows. In order to test this hypothesis, we begin with, say, 80 cows assigned at random to one of the two diets. Most cows become pregnant at the

first insemination but a few require a second or third insemination or occasionally more. After the first insemination, the most fertile cows have become pregnant. The success rate for those that require subsequent insemination is noticeably less than the initial success rate. In this instance there is an indefinite number of stages corresponding to first attempt, second attempt and so on. Three stages are depicted in Fig. 5.3. The variable measured at each stage is the pregnancy success rate. In that respect, this example differs from the previous one, where the variables measured at each stage were scientifically distinct. If indeed, red clover reduces fertility, this reduction should be apparent at all stages, even though the mean fertility of the remaining cows is reduced at each successive stage. Information concerning the treatment effect must be collected from the pregnancy rates observed at each stage.

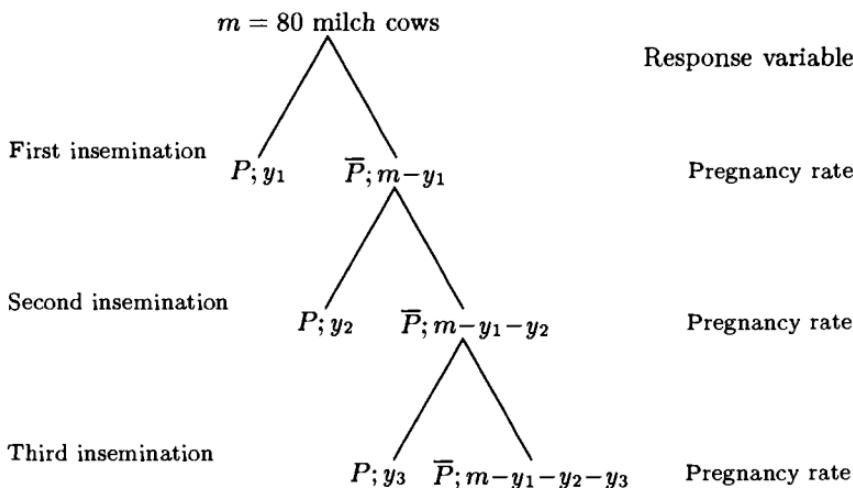


Fig. 5.3. *Hierarchical response in an insemination experiment:*
 P = pregnant, \bar{P} = not pregnant.

There is, of course, the possibility of a more complicated treatment effect whereby only the less fertile cows are affected. The observed treatment effect would then be expected to increase at successive stages. In the analysis, one should be aware that such an interaction might occur and what its symptoms would be.

In order to build up a model for either kind of hierarchical response, it is best to consider separately $k - 1$ responses, one

for each level of the hierarchy. The m subjects available at stage 1 respond positively with probability π_1 , or negatively with probability $1 - \gamma_1$. The observed proportions are necessarily slightly different from the theoretical proportions so that, at stage 2, the experimental set or ‘risk set’ is reduced to $m - y_1$. Among these, the probability of a positive response is $\pi_2/(1 - \gamma_1)$, and the probability of a negative response is $(1 - \gamma_2)/(1 - \gamma_1)$. By the third stage, the risk set is further reduced to $m - y_1 - y_2$. Among these, the probability of a positive response is $\pi_3/(1 - \gamma_2)$, and the probability of a negative response is $(1 - \gamma_3)/(1 - \gamma_2)$. The response is thus broken down into the sequence of conditional factors:

<i>Stage</i>	<i>Response</i>	<i>Probability</i>	<i>Odds</i>
1	$Y_1 m$	π_1	$\pi_1/(1 - \gamma_1)$
2	$Y_2 m - y_1$	$\pi_2/(1 - \gamma_1)$	$\pi_2/(1 - \gamma_2)$
3	$Y_3 m - y_1 - y_2$	$\pi_3/(1 - \gamma_2)$	$\pi_3/(1 - \gamma_3)$

In the particular examples considered, each stage of the hierarchy corresponds to a simple dichotomy. It is natural therefore, to consider binary regression models of the type discussed in Chapter 4. Thus, in the radiation mortality example,

$$g(\pi_1) = \beta_1^T \mathbf{x}$$

relates total mortality to exposure \mathbf{x} via the link function $g(\cdot)$. By extension,

$$g\left(\frac{\pi_2}{1 - \gamma_1}\right) = \beta_2^T \mathbf{x}$$

relates cancer mortality as a proportion of total mortality to exposure. Similarly,

$$g\left(\frac{\pi_3}{1 - \gamma_2}\right) = \beta_3^T \mathbf{x}$$

relates leukaemia cancer mortality as a proportion of total cancer mortality to the exposure variables. There is no good reason here to expect that the coefficients $\beta_1, \beta_2, \beta_3$ might be equal or even comparable. In addition, there is no strong argument for using the same link function in the three regressions. If the identified cancer types at stage three were ‘leukaemia’, ‘thyroid’ and ‘other’, the trichotomy would be regarded as a nominal response scale and

the methods of section 5.2.4 could be used. It would then become impossible to insist on the same link function for each stage.

The insemination example has many of the same features but differs in the important respect that the response is the same at each stage. In constructing a model, however, we must make allowance for the expected decline in fertility at successive stages. A simple sequence of models having a constant treatment effect is as follows:

$$\begin{aligned}g(\pi_1) &= \alpha_1 + \beta^T x, \\g\{\pi_2/(1-\gamma_1)\} &= \alpha_2 + \beta^T x, \\g\{\pi_3/(1-\gamma_2)\} &= \alpha_3 + \beta^T x.\end{aligned}$$

It is essential here to use same link function for each stage. In particular, if the logistic link function is used, we have

$$\log\left(\frac{\pi_j}{1-\gamma_j}\right) = \alpha_j + \beta^T x. \quad (5.10)$$

The incidental parameters $\alpha_1, \dots, \alpha_{k-1}$ make allowance for the expected decline in fertility. If x is an indicator variable for treatment, model (5.10) asserts that treatment increases the odds of success by a factor $\exp(\beta^T x)$ uniformly at each stage of the experiment. Constancy of the effect can be tested in the usual way by the addition of an interaction term between treatment and stage.

5.3 The multinomial distribution

5.3.1 Genesis

The multinomial distribution is in many ways the most natural distribution to consider in the context of a polytomous response variable. It arises in a number of contexts, some apparently artificial, others a consequence of simple random sampling.

Suppose that individuals in some population of interest possess one and only one of the k attributes A_1, \dots, A_k . The attributes might be ‘colour of hair’, ‘socio-economic status’, ‘family size’, ‘cause of death’ and so on depending on the context. If the population is effectively infinitely large and if a simple random sample of size m is taken, how many individuals will be observed

to have attribute A_j ? The answer is given by the multinomial distribution

$$\text{pr}(Y_1 = y_1, \dots, Y_k = y_k; m, \boldsymbol{\pi}) = \binom{m}{\mathbf{y}} \pi_1^{y_1} \dots \pi_k^{y_k}, \quad (5.11)$$

where π_1, \dots, π_k are the attribute frequencies in the infinite population and

$$\binom{m}{\mathbf{y}} = \frac{m!}{y_1! \dots y_k!}.$$

The multinomial distribution arises here simply as a consequence of the method of sampling. A different method of sampling such as cluster sampling or quota sampling would give rise to a different frequency distribution from (5.11).

The sample space or set of all possible values of the vector \mathbf{y} is the set of all integer-valued k -vectors satisfying $0 \leq y_j \leq m$, $\sum y_j = m$ and comprises $\binom{m+k-1}{k-1}$ points. The sample space is a triangular lattice bounded by a regular simplex: see Fig. 5.4 for a diagram of the trinomial distribution.

Another derivation of the multinomial distribution is as follows. Suppose that Y_1, \dots, Y_k are independent Poisson random variables with means μ_1, \dots, μ_k . Then the conditional joint distribution of Y_1, \dots, Y_k given that $\sum Y_j = m$ is given by (5.11) with $\pi_j = \mu_j / \mu_{..}$.

The multinomial distribution for which $\pi_j = 1/k$ is called the uniform multinomial distribution.

5.3.2 Moments and cumulants

The moment generating function of the multinomial distribution, $M(m, \boldsymbol{\pi})$ is

$$M_Y(t) = E \exp \left(\sum t_j Y_j \right) = \left\{ \sum \pi_j \exp(t_j) \right\}^m.$$

Thus the cumulant generating function is

$$K_Y(t) = m \log \left\{ \sum \pi_j \exp(t_j) \right\}.$$

All cumulants have the form $m \times \text{polynomial in } \boldsymbol{\pi}$. In particular, the first four joint cumulants are

$$E(Y_r) = m\pi_r$$

$$\text{cov}(Y_r, Y_s) = \begin{cases} m\pi_r(1 - \pi_r) & r = s \\ -m\pi_r\pi_s & r \neq s \end{cases} \quad (5.12)$$

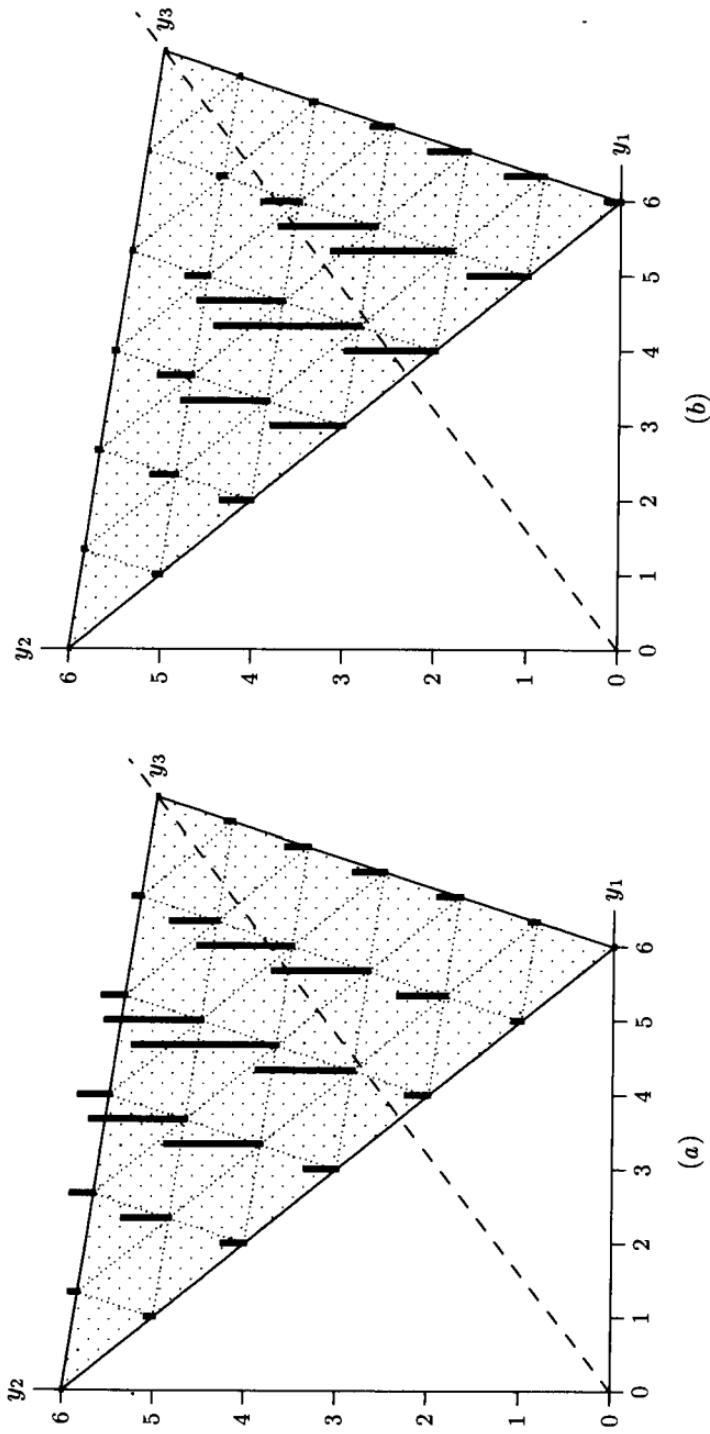


Fig. 5.4. Trinomial sample space and probabilities for $m = 6$. The sample points form a triangular lattice on the equiangular plane $y_1 + y_2 + y_3 = 6$ (shaded) in R^3 ; the y_3 -axis recedes into the plane of the page. In (a), $\pi = (1/3, 1/3, 1/3)$. In (b), $\pi = (0.5, 0.3, 0.2)$.

$$\kappa_3(Y_r, Y_s, Y_t) = \begin{cases} m\pi_r(1 - \pi_r)(1 - 2\pi_r) & r = s = t \\ -m\pi_r\pi_t(1 - 2\pi_r) & r = s \neq t \\ 2m\pi_r\pi_s\pi_t & r, s, t \text{ distinct} \end{cases}$$

$$\kappa_4(Y_r, Y_s, Y_t, Y_u) = \begin{cases} m\pi_r(1 - \pi_r)(1 - 6\pi_r(1 - \pi_r)) & r = s = t = u \\ -m\pi_r\pi_u(1 - 6\pi_r(1 - \pi_r)) & r = s = t \neq u \\ -m\pi_r\pi_t(1 - 2\pi_r - 2\pi_t + 6\pi_r\pi_t) & r = s \neq t = u \\ 2m\pi_r\pi_s\pi_u(1 - 3\pi_r) & r = s \neq t \neq u \\ -6m\pi_r\pi_s\pi_t\pi_u & r, s, t, u \text{ distinct} \end{cases}$$

Frequently, however, it is more convenient to work with the vector of cumulative totals rather than with the cell counts. If we write $\mathbf{Z} = \mathbf{LY}$ where \mathbf{L} is a lower-triangular matrix containing unit values, we see that the vector of cumulative totals is a linear function of \mathbf{Y} . The first four cumulants of \mathbf{Z} are

$$E(Z_r) = m\gamma_r,$$

$$\gamma_{rs} = \text{cov}(Z_r, Z_s) = m\gamma_r(1 - \gamma_s) \quad \text{for } r \leq s, \quad (5.13)$$

$$\kappa_3(Z_r, Z_s, Z_t) = m\gamma_r(1 - 2\gamma_s)(1 - \gamma_t) \quad \text{for } r \leq s \leq t,$$

$$\kappa_4(Z_r, Z_s, Z_t, Z_u) = m\gamma_r(1 - \gamma_u)\{1 - 2(\gamma_t - \gamma_s) - 6\gamma_s(1 - \gamma_t)\} \\ \text{for } r \leq s \leq t \leq u.$$

In other respects as well, the cumulative multinomial vector has simpler properties than the original vector. For instance, it is easily seen that for $r < s < t$, Z_r and Z_t are conditionally independent given Z_s . To be specific, given $Z_s = z_s$,

$$Z_r \sim B(z_s, \gamma_r/\gamma_s)$$

$$Z_t - z_s \sim B\{m - z_s, (\gamma_t - \gamma_s)/(1 - \gamma_s)\}.$$

Linear combinations $\sum s_j Y_j$ with fixed coefficients s_j arise naturally in calculations related to models of the type discussed in section 5.2.3. The cumulants of such a combination are easily obtained either from the expressions given above or by observing that for $m = 1$, $\sum s_j Y_j$ takes the values s_1, \dots, s_k with probabilities π_1, \dots, π_k . Consequently if we write

$$\mu_s = E\{\sum s_j Y_j/m\} = \sum \pi_j s_j$$

we have that

$$\text{var}(\sum s_j Y_j) = m \sum \pi_j (s_j - \mu_s)^2 = m \{ \sum \pi_j s_j^2 - (\sum \pi_j s_j)^2 \}$$

$$\kappa_3(\sum s_j Y_j) = m \sum \pi_j (s_j - \mu_s)^3.$$

Similar expressions may be derived for higher-order cumulants should these be required for Edgeworth approximation.

5.3.3 Generalized inverse matrices

Provided only that the cell probabilities π_j are positive, the multinomial covariance matrix $\Sigma_Y = m\{\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T\}$ has rank $k-1$. The simplest generalized inverse is

$$\Sigma_Y^- = \text{diag}\{1/(m\pi_j)\},$$

which has rank k . This is not the Moore-Penrose inverse, but for most statistical calculations the choice of generalized inverse is immaterial and Σ^- given above is perhaps the simplest such inverse. It is easily verified that

$$\Sigma \Sigma^- \Sigma = \Sigma,$$

which is the defining property of a generalized inverse. In fact all generalized inverses have the form $\Sigma^- - c\mathbf{1}\mathbf{1}^T$ for some c . The Moore-Penrose inverse has $c = 1$.

The vector of cumulative totals, \mathbf{Z} , may be regarded either as a vector having k components, the last of which is fixed, or alternatively as a vector having $k-1$ components, the last component being ignored. In either case the covariance matrix has rank $k-1$. The covariance matrix $\Gamma = \{\gamma_{rs}\}$ in (5.13) is a particular instance of a Green's matrix, whose inverse is a symmetric Jacobi or tri-diagonal matrix. The particular form of inverse for $k=5$ is as follows:

$$\Gamma^- = \frac{1}{m} \begin{pmatrix} \pi_1^{-1} + \pi_2^{-1} & -\pi_2^{-1} & 0 & 0 & 0 \\ -\pi_2^{-1} & \pi_2^{-1} + \pi_3^{-1} & -\pi_3^{-1} & 0 & 0 \\ 0 & -\pi_3^{-1} & \pi_3^{-1} + \pi_4^{-1} & -\pi_4^{-1} & 0 \\ 0 & 0 & -\pi_4^{-1} & \pi_4^{-1} + \pi_5^{-1} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

This is the Moore-Penrose inverse of Γ in (5.13). All generalized inverses have this form, but with arbitrary values in the final row and column.

For a discussion of the geometry of generalized inverse matrices, see Kruskal (1975) or Stone (1987).

5.3.4 Quadratic forms

In order to test the simple null hypothesis $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}^{(0)}$, it is natural to construct, as a test statistic, a quadratic form in the residuals, $R_j = Y_j - m\pi_j^{(0)}$. Since $\sum R_j = 0$, it follows that

$$X^2 = \mathbf{R}^T \boldsymbol{\Sigma}_Y^- \mathbf{R}$$

is independent of the choice of generalized inverse. Taking the particular inverse given in the previous section, we see that

$$X^2 = \sum_j R_j^2 / (m\pi_j^{(0)}) = \sum_j (Y_j - \mu_j^{(0)})^2 / \mu_j^{(0)},$$

which is the familiar statistic due to Pearson (1900).

Equally well, if we choose to work with the cumulative multinomial vector and the corresponding generalized inverse, we obtain

$$\sum_1^{k-1} \frac{(Z_j - m\gamma_j)^2}{m} \left(\frac{1}{\pi_j} + \frac{1}{\pi_{j+1}} \right) - 2 \sum_{j=1}^{k-2} \frac{(Z_j - m\gamma_j)(Z_{j+1} - m\gamma_{j+1})}{m\pi_{j+1}}$$

with $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$ computed under H_0 . It is an elementary if rather tedious exercise to show that the above quadratic form is identical to X^2 . Quadratic forms such as these are invariant under nonsingular linear transformation of the original variables.

The first three null cumulants of X^2 are

$$\begin{aligned} E(X^2) &= k - 1, \\ \text{var}(X^2) &= 2(k-1) \frac{m-1}{m} + (S_{-1} - k^2)/m, \\ \kappa_3(X^2) &= 8(k-1) \frac{m-1}{m} + 4(k-1)(k-6) \frac{m-1}{m^2} \\ &\quad + (S_{-1} - k^2)(22(m-1) - 3k)/m^2 + (S_{-2} - k^3)/m^2, \end{aligned}$$

where $S_r = \sum \pi_j^r$. In the uniform case, $S_{-1} = k^2$, $S_{-2} = k^3$.

If m is large, X^2 is approximately distributed as χ_{k-1}^2 with cumulants $k-1, 2(k-1), 8(k-1), \dots$. The above exact calculations give a measure of the departure in finite samples of X^2 from its limiting distribution.

Similar moment calculations for X^2 for two-way tables are given in Exercise 6.16.

5.3.5 Marginal and conditional distributions

The marginal distribution of each multinomial component of Y is binomial: $Y_j \sim B(m, \pi_j)$. Also, the joint marginal distribution of $(Y_1, Y_2, m - Y_1 - Y_2)$ is multinomial on three categories with index m and parameter $(\pi_1, \pi_2, 1 - \pi_1 - \pi_2)$. This latter property extends to any number of components.

The conditional joint distribution of Y_1, \dots, Y_k , given that $Y_i = y_i$, is multinomial on the reduced set of categories, with reduced index $m \rightarrow m - y_i$ and probabilities renormalized to

$$\pi_j \rightarrow \pi_j / (1 - \pi_i).$$

Analogous results are available for the cumulative multinomial vector, Z . The marginal distribution of Z_j is $B(m, \gamma_j)$. The conditional distribution of Z_i given $Z_j = z_j$, is $B(z_j, \gamma_i / \gamma_j)$ for $i < j$. Also, the conditional distribution of Y_{j+1} given $Z_j = z_j$, is $B(m - z_j, \pi_{j+1} / (1 - \gamma_j))$, which is the basis for the decomposition in section 5.2.5. In fact, the multinomial distribution can be expressed as a product of $k - 1$ binomial factors

$$\text{pr}(Y = y) = p(y_1 | z_0) p(y_2 | z_1) p(y_3 | z_2) \dots p(y_{k-1} | z_{k-2}),$$

where

$$p(y_j | z_{j-1}) = \left(\frac{\pi_j}{1 - \gamma_{j-1}} \right)^{y_j} \left(\frac{1 - \gamma_j}{1 - \gamma_{j-1}} \right)^{m - z_{j-1} - y_j} \binom{m - z_{j-1}}{y_j}$$

and $z_0 = \gamma_0 = 0$.

Evidently, the sequence Z_1, \dots, Z_k has the Markov property, namely that the conditional distribution of Z_j given the entire sequence Z_1, \dots, Z_{j-1} up to $j - 1$, depends only on the most recent value, namely Z_{j-1} . Also, the ‘past’ Z_1, \dots, Z_{j-1} , and the ‘future’ Z_{j+1}, \dots, Z_k are conditionally independent given the ‘present’, Z_j .

5.4 Likelihood functions

5.4.1 Log likelihood for multinomial responses

We suppose that there are available n independent multinomial vectors, each with k categories. These observations are denoted by $\mathbf{y}_1, \dots, \mathbf{y}_n$, where $\mathbf{y}_i = (y_{i1}, \dots, y_{ik})$ and $\sum_j y_{ij} = m_i$ is fixed for each i . As usual, it is more convenient to consider the log likelihood initially as a function of the n probability vectors $\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_n$. Subsequently, when we contemplate a specific model such as (5.1) or (5.5), we may express the probabilities in terms of the parameters that appear in that model.

From the i th observation \mathbf{y}_i , the contribution to the log likelihood is

$$l(\boldsymbol{\pi}_i; \mathbf{y}_i) = \sum_j y_{ij} \log \pi_{ij}.$$

It is understood here that the observations and the probabilities are subject to the linear constraints

$$\sum_j y_{ij} = m_i \quad \text{and} \quad \sum_j \pi_{ij} = 1$$

for each i . Since the n observations are independent by assumption, the total log likelihood is a sum of contributions, one from each of the n observations. Thus,

$$l(\boldsymbol{\pi}; \mathbf{y}) = \sum_{ij} y_{ij} \log \pi_{ij}. \quad (5.14)$$

Differentiation of the log likelihood with respect to π_{ij} subject to the constraint $\sum_j \pi_{ij} = 1$ gives

$$\frac{\partial l(\boldsymbol{\pi}; \mathbf{y})}{\partial \pi_{ij}} = \frac{y_{ij} - m_i \pi_{ij}}{\pi_{ij}}.$$

Equivalently, introducing matrix notation,

$$\begin{aligned} \frac{\partial l(\boldsymbol{\pi}; \mathbf{y})}{\partial \boldsymbol{\pi}_i} &= m_i \boldsymbol{\Sigma}_i^- (\mathbf{y}_i - m_i \boldsymbol{\pi}_i) \\ &= m_i \boldsymbol{\Sigma}_i^- (\mathbf{y}_i - \boldsymbol{\mu}_i). \end{aligned} \quad (5.15)$$

This set of n derivative vectors can be collected into a single matrix equation

$$\frac{\partial l(\boldsymbol{\pi}; \mathbf{y})}{\partial \boldsymbol{\pi}} = \mathbf{M} \boldsymbol{\Sigma}^-(\mathbf{y} - \boldsymbol{\mu}),$$

where $\boldsymbol{\Sigma} = \text{diag}\{\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_n\}$ is $nk \times nk$ of rank $n(k-1)$ and \mathbf{M} is a diagonal matrix of order $nk \times nk$ containing the multinomial indices m_i each repeated k times. The choice of generalized inverse in (5.15) is immaterial because $\mathbf{1}^T(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0$ for each i .

In the above calculations, we have chosen to work with the response vectors \mathbf{y}_i and the probability vectors $\boldsymbol{\pi}_i$. This turns out to be a convenient but arbitrary choice. We could choose to work instead with the cumulative response vectors \mathbf{z}_i together with the cumulative probability vectors $\boldsymbol{\gamma}_i$. Analogous calculations then give

$$\frac{\partial l(\boldsymbol{\gamma}; \mathbf{y})}{\partial \boldsymbol{\gamma}_i} = m_i \boldsymbol{\Gamma}_i^-(\mathbf{z}_i - m_i \boldsymbol{\gamma}_i), \quad (5.16)$$

which can be obtained from (5.15) using the chain rule. In fact,

$$\frac{\partial l}{\partial \gamma_{ij}} = \frac{\partial l}{\partial \pi_{ij}} - \frac{\partial l}{\partial \pi_{i,j-1}} \quad \text{for } 1 < j < k,$$

which is the same as (5.16).

5.4.2 Parameter estimation

The likelihood equations for the parameters are entirely straightforward to obtain at least in principle. We simply multiply (5.15) by the derivative of π_{ij} with respect to each parameter in turn and sum over i and j . Alternatively, and equivalently, we multiply (5.16) by the derivative of γ_{ij} with respect to the parameters and sum over i and j . Obviously, the form of the resulting equations depends heavily on the particular choice of model. We now consider some of the details of two particular choices.

Suppose that the model chosen has the form

$$\text{logit } \gamma_{ij} = \sum_r x_{ijr}^* \beta_r^*$$

for some fixed coefficients x_{ijr}^* and unknown parameters β_r^* . It is helpful here to think of x_{ijr}^* as the components of a matrix \mathbf{X}^*

of order $nk \times p^*$ where p^* is the dimension of β^* . In the case of model (5.1), β^* has dimension $p^* = p + k - 1$ with components

$$\beta^* = (\theta_1, \dots, \theta_{k-1}, \beta_1, \dots, \beta_p).$$

The (i, j) row of \mathbf{X}^* has components $(0, \dots, 1, \dots, 0, \mathbf{x}_i)$, with the unit value in position j . Consequently, the i th block of $k - 1$ rows is

$$[\mathbf{I}_{k-1} : \mathbf{1}\mathbf{x}_i].$$

Differentiation with respect to β^* gives

$$\begin{aligned} \frac{\partial l}{\partial \beta_r^*} &= \sum_{ij} \frac{\partial l}{\partial \gamma_{ij}} \frac{\partial \gamma_{ij}}{\partial \beta_r^*} \\ &= \sum_{ij} x_{ijr}^* \gamma_{ij} (1 - \gamma_{ij}) \frac{\partial l}{\partial \gamma_{ij}}, \end{aligned}$$

where $\partial l / \partial \gamma_{ij}$ is given by (5.16). In fact,

$$\frac{\partial l}{\partial \gamma_{ij}} = \frac{y_{ij} - m_i \pi_{ij}}{\pi_{ij}} - \frac{y_{ij-1} - m_i \pi_{ij-1}}{\pi_{ij-1}} \quad \text{for } 1 < j < k.$$

For the proportional-odds model (5.1), these calculations can be simplified to some extent by exploiting the structure of the array x_{ijr}^* but these details will not be pursued here.

For log-linear models such as (5.5)–(5.9), we have

$$\log \pi_{ij} = \sum_r x_{ijr}^* \beta_r^*$$

for various choices of coefficients x_{ijr}^* dependent on the choice of model. In all of these cases, the likelihood equations take on a particularly simple form, namely

$$\sum_{ij} x_{ijr}^* (y_{ij} - \hat{\mu}_{ij}) = 0 \quad \text{for } r = 1, \dots, p^*.$$

In other words, in this case maximum likelihood is equivalent to the method of moments in which specific linear combinations $\sum x_{ijr}^* y_{ij}$ are equated to their expectations as a function of the parameters. The actual combinations depend on the choice of model. For instance if model (5.8) is used, the combinations are the ‘row’ and ‘column’ totals as well as the ‘interaction combinations’

$$\sum_{ij} x_{ir} s_j y_{ij} \quad \text{for } r = 1, \dots, p.$$

5.4.3 Deviance function

The residual deviance function is twice the difference between the maximum achievable log likelihood and that attained under the fitted model. The maximum achievable log likelihood occurs at the point $\tilde{\pi}_{ij} = y_{ij}/m_i$. The deviance function is therefore

$$\begin{aligned} D(\mathbf{y}; \hat{\boldsymbol{\pi}}) &= 2l(\tilde{\boldsymbol{\pi}}; \mathbf{y}) - 2l(\hat{\boldsymbol{\pi}}; \mathbf{y}) \\ &= 2 \sum y_{ij} \log \tilde{\pi}_{ij} - 2 \sum y_{ij} \log \hat{\pi}_{ij} \\ &= 2 \sum_{ij} y_{ij} \log(y_{ij}/\hat{\mu}_{ij}). \end{aligned}$$

Under the conditions described in section 4.4.3, namely that $\hat{\mu}_{ij}$ are sufficiently large and that there is no over-dispersion, $D(Y; \hat{\boldsymbol{\pi}})$ has an approximate χ^2 distribution. Its use as a goodness-of-fit statistic, however is open to the objections raised in sections 4.4.3, 4.4.5 and 4.5.

5.5 Over-dispersion

Over-dispersion for polytomous responses can occur in exactly the same way as over-dispersion for binary responses. Details are given in section 4.5.1 and will not be repeated here. Under the cluster-sampling model, the covariance matrix of the observed response vector is the sum of the within-cluster covariance matrix and the between-cluster covariance matrix. Provided that these two matrices are proportional, we have

$$\begin{aligned} E(Y) &= m\boldsymbol{\pi}, \\ \text{cov}(Y) &= \sigma^2 \boldsymbol{\Sigma}, \end{aligned} \tag{5.17}$$

where $\boldsymbol{\Sigma}$ is the usual multinomial covariance matrix. The dispersion parameter σ^2 has the same interpretation given in section 4.5.

Parameter estimation (other than σ^2) is unaffected by over-dispersion and proceeds along the lines described in section 5.4.2 as if the multinomial distribution continued to apply. However, the covariance matrix of $\hat{\boldsymbol{\beta}}$, obtained from the multinomial log likelihood, needs to be inflated by the dispersion factor σ^2 . The

only additional step therefore is the estimation of this dispersion factor. For the reasons given in section 4.5.2, we use

$$\begin{aligned}\tilde{\sigma}^2 &= X^2 / \{n(k - 1) - p\} \\ &= X^2 / \{\text{residual d.f.}\},\end{aligned}\quad (5.18)$$

where X^2 is Pearson's statistic. This estimate is approximately unbiased for σ^2 , is consistent for large n regardless of whether the data are sparse and moreover is approximately independent of the estimated $\hat{\beta}$.

For further details see Chapter 9.

5.6 Examples

5.6.1 A cheese-tasting experiment

The following data, kindly provided by Dr Graeme Newell, were obtained from an experiment concerning the effect on taste of various cheese additives. The so-called hedonic scale has nine response categories, ranging from 'strong dislike' (1) to 'excellent taste' (9). In this instance, four additives labelled A, B, C and D were tested. The data are given in Table 5.1.

Here the effects are so great that the qualitative ordering (D, A, C, B) can easily be deduced from visual inspection. Nevertheless it is of some interest to check whether the models described earlier are capable of describing these differences and of evaluating the statistical significance of the differences observed.

Table 5.1 Response frequencies in a cheese-tasting experiment

Cheese	Response category									Total
	I*	II	III	IV	V	VI	VII	VIII	IX†	
A	0	0	1	7	8	8	19	8	1	52
B	6	9	12	11	7	6	1	0	0	52
C	1	1	6	8	23	7	5	1	0	52
D	0	0	0	1	3	7	14	16	11	52
Total	7	10	19	27	41	28	39	25	12	208

*I = strong dislike;

†IX = excellent taste.

Data courtesy of Dr Graeme Newell, Hawkesbury Agricultural College.

The nature of the response is such that a model of the form (5.1) or (5.3) is most obviously appealing. We first try the logistic model with intercept parameters $\theta_1, \dots, \theta_8$ and treatment effects β_1, \dots, β_4 . In this instance, (5.1) can be written in the form

$$\text{logit } \gamma_{ij} = \theta_j - \beta_i$$

for $j = 1, \dots, 8$ and $i = 1, \dots, 4$. As usual, only contrasts among the treatment effects β_i are estimable. We adopt the convention whereby $\hat{\beta}_4 = 0$. The resulting estimates, standard errors and correlation matrix of the $\hat{\beta}$ s are given below.

Logistic treatment effects for cheese-tasting data

Additive	Estimate	SE	Correlations		
A	$\hat{\beta}_1 = -1.613$	0.378	1.0		
B	$\hat{\beta}_2 = -4.965$	0.474	0.525	1.0	
C	$\hat{\beta}_3 = -3.323$	0.425	0.574	0.659	1.0
D	$\hat{\beta}_4 = 0.0$	—	—	—	—

Positive values of β represent a tendency towards the higher-numbered categories relative to the chosen baseline — in this case, the probabilities for cheese D. Negative values indicate the reverse effect. The observed estimates quantify and confirm the ordering (D, A, C, B) from best to worst. The quoted standard errors are based on the assumption that $\sigma^2 = 1$, namely that no over-dispersion is present. The correlations, unlike the covariances, are unaffected by the choice or estimate of σ^2 .

The deviance for these data is reduced from 168.8 on 24 degrees of freedom under the model of zero additive effect ($\beta = 0$) to 20.31 on 21 degrees of freedom under the proportional-odds model. Because of the small numbers in the extreme cells the chi-squared approximation for the deviance is not very good here. Residual analysis is awkward partly for the same reason and partly because row sums are fixed. Using the crude standardization $(y_{ij} - \hat{y}_{ij})/[m_i \hat{\pi}_{ij}(1 - \hat{\pi}_{ij})]^{1/2}$, we find two cell residuals exceeding the value 2.0. The values are 2.23 and 2.30 corresponding to cells (1, 4) and (2, 6) with fitted values 3.16 and 2.47 respectively. However, if residual calculations were based on the cumulative totals $z_{ij} = y_{i1} + \dots + y_{ij}$, arguably a more appropriate procedure here,

the apparently large discrepancies would disappear. At the very least, residuals based on z_{ij} have the strong conceptual advantage that only $k - 1$ of them are defined for each multinomial observation. Correlation among the residuals is a problem regardless of definition but the problem seems more acute for residuals based on z_{ij} . However, these extreme residuals are hardly sufficiently large to refute the model which is, at best, an approximation to reality.

As a further check on the adequacy of the proportional-odds model, we fitted the generalized rational model (5.4) with the treatment factor as the model formula in both numerator and denominator. In other words

$$\text{logit } \gamma_{ij} = \frac{\theta_j - \beta_i}{\exp(\tau_i)}.$$

This gives a reduction in deviance of 3.3 on 3 degrees of freedom, so that there is no evidence that the variability of responses is affected by the cheese additive.

So far we have assumed $\sigma^2 = 1$ without justification. However, the estimate for σ^2 from (5.13) is 20.9/21 or almost exactly unity. Here $p = 11$ is the total number of parameters including the θ s.

A more serious problem that we have so far ignored is that observations corresponding to different treatments are not independent. The same 52 panellists are involved in all four tests. This is likely to induce some positive correlation ρ between the ratings for the different treatments. Variances of contrasts would then be reduced by a factor $1 - \rho^2$ relative to independent measurements. Inferences based on supposing that $\rho = 0$ are therefore conservative. In other words the general qualitative and quantitative conclusions remain valid with the computed variances being regarded as approximate upper limits.

Finally we examine the effect on $\hat{\beta}$ of reducing the number of response categories. Various combinations are possible: here we combine categories 1, ..., 4 and 7, 8, 9, thus reducing the original nine categories to four. This arrangement makes all cell counts positive. The new estimates for β are $(-1.34, -4.57, -3.07, 0)$ corresponding to an average reduction of about 0.7 standard errors compared with the previous analysis. Reduction of the number of categories does not always have this effect. Estimated variances are increased by an average of about 19%. Correlations are virtually unaffected.

The available evidence suggests that, when the data are sparse, the estimate of $\hat{\beta}_j$ may be too large in magnitude. Grouping of the tail categories has the effect of reducing this bias. A very small-scale simulation experiment based on 25 repetitions using the values of $\hat{\theta}$ and $\hat{\beta}$ obtained from the data in Table 5.1 and the same row totals indicates the following:

1. the bias in the estimates $\hat{\beta}_j$ is no more than 5%.
2. the deviance or likelihood-ratio goodness-of-fit statistic is approximately distributed as χ^2_{21} : at least the first two moments do not differ appreciably from those of this reference distribution.
3. the standard errors obtained from the diagonal elements of (5.11) are, if anything, a little too large — by about 10%.

The first claim is buttressed to some extent by the findings in section 7.5.3, where the nuisance parameters are eliminated by suitable conditioning. Because of the small scale of the simulation, the remaining conclusions are tentative. Nonetheless the conclusions are positive and they show that, even with data as sparse as those in Table 5.1 and where the number of parameters (11) is moderately large in comparison to the number of observations (32), the usual asymptotic results are quite reliable at least for the parameters of primary interest.

5.6.2 *Pneumoconiosis among coalminers*

The following example illustrates the use of a quantitative covariate in an ordinal regression model. For comparative purposes we apply both (5.1) and (5.10). Difficulties associated with residual plots are also illustrated.

The data, taken from Ashford (1959), concern the degree of pneumoconiosis in coalface workers as a function of exposure t measured in years. Severity of disease is measured radiologically and is, of necessity, qualitative. A four-category version of the ILO rating scale was used initially, but the two most severe categories were subsequently combined.

A preliminary plot of the transformed variables

$$\log \left(\frac{y_{i1} + \frac{1}{2}}{m_i - y_{i1} + \frac{1}{2}} \right) \quad \text{and} \quad \log \left(\frac{y_{i1} + y_{i2} + \frac{1}{2}}{m_i - y_{i1} - y_{i2} + \frac{1}{2}} \right) \quad (5.19)$$

Table 5.2 *Period of exposure and prevalence of pneumoconiosis amongst a group of coalminers*

Period spent (yr)	Number of men		
	Category I: normal	Category III: severe	
		Category II	pneumoconiosis
5.8	98	0	0
15.0	51	2	1
21.5	34	6	3
27.5	35	5	8
33.5	32	10	9
39.5	23	7	8
46.0	12	6	10
51.5	4	2	5

against t_i reveals approximately parallel but non-linear relationships. Further investigation shows that the transformed variables (5.19) are approximately linear in $\log t_i$. We are thus led to consider the model

$$\log[\gamma_{ij}/(1 - \gamma_{ij})] = \theta_j - \beta \log t_i, \quad j = 1, 2; i = 1, \dots, 8. \quad (5.20)$$

We might expect that the non-linearity of (5.20) in t could have been detected by an appropriate analysis of the residuals after fitting the model linear in t . This is indeed so but some care is required. When the 24 cell residuals, appropriately standardized, are plotted against t_i , no strong curvilinear pattern is discernible. On the other hand, a plot against t_i of the cumulative residuals, $y_{i1} - \hat{y}_{i1}$ and $y_{i1} + y_{i2} - \hat{y}_{i1} - \hat{y}_{i2}$, appropriately standardized, clearly reveals the non-linearity. When $k = 3$ this is equivalent to ignoring the residuals associated with category 2 and changing the sign of the category-3 residuals. The two plots are displayed in Figs. 5.5a and 5.5b respectively. The simplified standardization used here takes no account of the errors involved in using estimated values of the parameters.

The analysis using (5.20) gives a value of $\hat{\beta}$ of 2.60 with standard error 0.38, while the values of $\hat{\theta}_1$ and $\hat{\theta}_2$ are 9.68 and 10.58 respectively. No pattern is discernible among the residuals and the fit is good. The conclusions, therefore, are that for a miner with, say, five years of exposure, the odds of having pneumoconiosis are

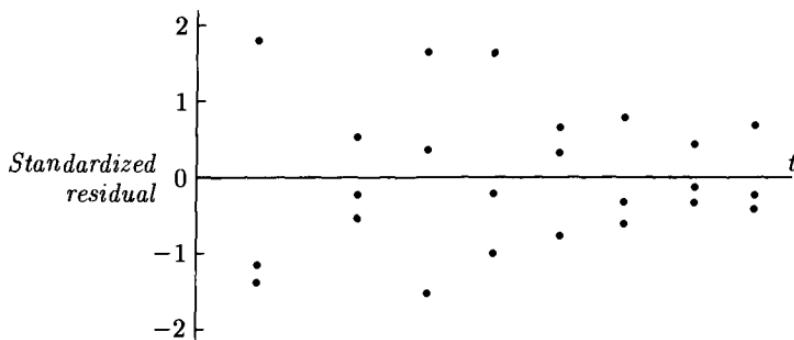


Fig. 5.5a. Plot of cell residuals against t for the pneumoconiosis data.

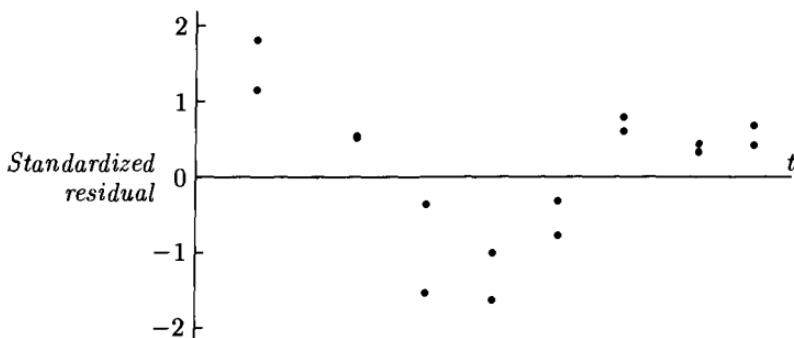


Fig. 5.5b. Plot of cumulative residuals against t for pneumoconiosis data.

one in $\exp(9.68 - 2.60 \log 5)$, i.e. one in 243. Doubling the exposure increases the risk by a factor of $2^{2.60} = 6.0$, so that after 10 years the risk rises to just under one in 40, and after 20 years to just over one in seven. For severe pneumoconiosis the five-year risk is one in $\exp(10.58 + 2.60 \log 5)$, i.e. about one in 600. This risk increases at the same rate as before so that after 10 years exposure the estimated risk is one in 100 and after 20 years is one in 17.

Ashford's analysis of these data proceeds along the same lines as that given here except that he uses the probit function in place of the logit. His conclusions give similar fitted values but his parameter estimates are different, partly because of his use of the probit function and partly because of numerical inaccuracies.

We now proceed to illustrate the use of the alternative model (5.10) in the context of the pneumoconiosis data. Imagine the response categories arranged in the hierarchical format illustrated in Fig. 5.3. The response at stage 1 is the dichotomy between

disease and no disease whereas the response at stage 2 is the dichotomy between the mild form of the disease and the severe form. These responses are very much alike but technically distinct. However, for comparative purposes, we assume here that the effect of continued exposure is comparable for the two responses so that (5.10) can be used. We take model (5.10) in the form

$$\log\left(\frac{\pi_{ij}}{1 - \gamma_{ij}}\right) = \alpha_j - \beta x_i, \quad (5.21)$$

reversing the sign of the coefficient in order to make the results at least qualitatively comparable to those obtained from (5.20).

The odds in favour of category 1 are $\exp(\alpha_1 - \beta x_i)$ so that the odds or risk of having the disease in the first place is $\exp(-\alpha_1 + \beta x_i)$. Here x_i is a general measure of exposure — in this case t_i or $\log t_i$. Thus the risk of disease increases by the factor e^β per unit increase in x . Among those who have the disease, the risk, or odds of having severe symptoms, is $\exp(-\alpha_2 + \beta x_i)$, so that again the risk increases by the factor e^β per unit increase in x . There is clearly the possibility that a different β might be involved in the second expression.

Because of the special structure of the model (5.10) each trinomial observation can be broken into two binomial components. The first component specifies the number of diseased individuals as a proportion of the total number at risk, while the second component gives the number of severely diseased as a proportion of those with the disease. Thus

$$\begin{aligned} y_{11}/m_{11} &= 0/98, & y_{12}/m_{12} &= 0/0, \\ y_{21}/m_{21} &= 3/54, & y_{22}/m_{22} &= 1/3, \\ y_{31}/m_{31} &= 9/43, & y_{32}/m_{32} &= 3/9, \end{aligned}$$

and so on with $m_{ij} = y_{ij} + y_{i,j+1} + \dots + y_{ik}$ giving the total in categories j through k inclusive.

These binomial observations y_{ij}/m_{ij} can be regarded as independent observations with probabilities π_{ij} satisfying

$$\log[\pi_{ij}/(1 - \pi_{ij})] = -\alpha_j + \beta x_i, \quad j = 1, 2; \quad i = 1, \dots, 8, \quad (5.22)$$

If the relationships are not parallel it may be necessary to write β_j instead of β in (5.22). The binomial log likelihood for the logistic

model (5.22) is identical to the multinomial log likelihood for the model (5.21).

For these data, the covariate $\log t_i$ is strongly preferred to t_i ; there is inconclusive evidence on whether $\beta_1 \neq \beta_2$. As measured by the deviance or likelihood-ratio statistic, the fit of (5.22) is comparable to that of (5.20). The goodness-of-fit statistics are 5.1 on 13 d.f. for (5.20) and 7.6 on 12 d.f. for (5.22). One degree of freedom is lost because y_{12} is degenerate or non-random when (5.22) is used. For model (5.22) the residuals give some indication of a faint pattern; for this reason the former model (5.20) might be preferred. In any case the estimate of β is 2.32 with approximate standard error 0.33, these values being similar to those obtained earlier despite the slight difference of interpretation. Thus we are led to the estimate of $2^{2.32} = 5.0$ as the increase in risk associated with doubling the exposure time.

To summarize we can say that (5.1) and (5.10), or equivalently (5.20) and (5.22), are different ways of describing the risk associated with increasing exposure. The conclusions from either analysis support the claim that doubling the exposure increases the risk by an estimated factor of between 5 and 6. Approximate 95% confidence limits for this factor are (3.2, 10.2). It would be of interest to know (i) whether the risk would continue to increase if exposure were to cease, and (ii) whether the risk would increase more slowly if dust levels were reduced. The data given here do not allow us to investigate these questions; indeed as the data stand, such effects would be likely to be confounded with age.

5.7 Bibliographic notes

Many of the methods and models discussed in Chapter 4 for binary data carry over to polytomous responses with only minor alterations. Consequently, most of the references listed in section 4.7 are also relevant here although there is enormous variation in emphasis and coverage. Agresti (1984) concentrates almost entirely on methods for ordinal response variables, including measures of association, which are not covered here. Haberman (1978, 1979) emphasizes methods for fitting a variety of log-linear models, mainly with social science applications in mind. Fienberg (1980) p.110 considers a variety of link functions, all of which are variations on the

logit. Both Haberman and Fienberg devote considerable attention to algorithmic details and the computation of maximum-likelihood estimates for two and three-way tables. Details of the iterative proportional fitting algorithm (Deming and Stephan, 1940; Darroch and Ratcliff, 1972) for log-linear models can be found in the book by Bishop *et al.* (1975). This algorithm forms the core of several log-linear computer packages, but it is not sufficiently general to cover the range of models considered here.

Aickin (1983) makes a distinction between nominal and nested response scales similar to the distinction made in section 5.2, but does not consider proportional-odds or proportional-hazards models for ordinal responses. For further discussion of measurement scales see Stevens (1951, 1958, 1968).

The idea of representing ordered categories as contiguous intervals on a continuous scale goes back at least to Pearson (1901), who investigated coat-colour inheritance in thoroughbred horses. An extension of this idea to two variables led to the development of the tetrachoric and polychoric correlation coefficients and to the quarrel with Yule (Yule, 1912; Pearson and Heron, 1913).

The proportional-odds model described in section 5.2 was previously used by Hewlett and Plackett (1956), Snell (1964), Walker and Duncan (1967), Clayton (1974), Simon (1974), Bock (1975) and others. Ashford (1959), Gurland *et al.* (1960) and Finney (1971) used the probit link in place of the logistic.

Williams and Grizzle (1972) discuss a number of methods including the proportional-odds model as well as scoring methods in the log-linear context. See also Haberman (1974a,b). McCullagh (1980) compares the use of scores in log-linear models with direct application of the proportional-odds model. He concludes that the proportional-odds and related models based on transforming the cumulative proportions are to be preferred to scoring methods because they are invariant under the grouping of adjacent response categories.

Graubard and Korn (1987) discuss the effect of the choice of scores in testing for independence in two-way tables.

Goodhardt, Ehrenberg and Chatfield (1984) use the Dirichlet-multinomial model to accommodate over-dispersion in brand-choice data. This is a natural extension of the beta-binomial model discussed in Chapter 4. For further discussion of specific forms of over-dispersion in this context, see Engel (1987).

Existence and uniqueness of maximum-likelihood estimates for a large subset of the models discussed here has been investigated by Pratt (1981) and by Burridge (1982).

Numerical methods for dealing with composite link functions such as (5.1) are discussed by Thompson and Baker (1981).

5.8 Further results and exercises 5

5.1 Show that

$$\sum_{j=0}^m \binom{j+k-1}{k-1} = \binom{m+k}{k}.$$

Hence deduce that the number of integer-valued sample points \mathbf{y} satisfying $0 \leq y_j \leq m$ and $\sum y_j = m$ is $\binom{m+k-1}{k-1}$. [Hint: consider the series expansion of $(1-x)^{-1}(1-x)^{-m}$.]

5.2 Suppose that $Y \sim M(1, \boldsymbol{\pi})$ and $Z = LY$ is the vector of cumulative totals of Y . Show that

$$E(Z_r Z_s Z_t \dots) = \gamma_r \quad \text{for } r \leq s \leq t \leq \dots.$$

Hence show that

$$\text{cov}(Z_r, Z_s) = \gamma_r(1 - \gamma_s) \quad \text{for } r \leq s.$$

Derive the third- and fourth-order cumulants of Z .

5.3 Show that the following expressions are equivalent:

$$\begin{aligned} & \sum \gamma_j(1 - \gamma_j)(\pi_j + \pi_{j+1}) \\ & \sum \pi_j(1 - \gamma_j - \gamma_{j-1})^2 \\ & \sum \gamma_j \gamma_{j+1} \pi_{j+1} \\ & \sum (1 - \gamma_j)(1 - \gamma_{j-1}) \pi_j \\ & \frac{1}{3} \left\{ 1 - \sum \pi_j^3 \right\}. \end{aligned}$$

All sums run from 1 to k with the convention that $\gamma_k = 1$ and

$$\pi_0 = \gamma_0 = \pi_{k+1} = \gamma_{k+1} = 0.$$

Find the minimum and maximum values for fixed $k \geq 2$.

5.4 By considering the case in which the differences $\theta_i - \theta_j$ are known, show that the asymptotic covariance matrix for $(\hat{\theta}_1, \hat{\beta})$ in (5.1) is given by $(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$. The model matrix \mathbf{X} is $n \times (p+1)$, with the constant vector in column 1, and \mathbf{W} is diagonal with components

$$w_i = \frac{1}{3} m_i \left\{ 1 - \sum_j \pi_{ij}^3 \right\}.$$

Deduce that this approximation gives a lower bound for $\text{cov}(\hat{\beta})$ when no assumptions are made about θ . Show that the approximation is accurate if $k = 2$ or if the log odds-ratios are small.

Hence show that in the two-sample problem following (5.2), the approximate variance of $\hat{\Delta}$ is $1/w_1 + 1/w_2$ (Clayton, 1974). This approximation appears to be satisfactory for $|\Delta| \leq 1$, which is usually the most interesting range in applications.

5.5 The data in Table 5.3, taken from Lowe *et al.* 1971, give the frequencies of anencephalus, spina bifida and other malformations of the central nervous system among live births in various South Wales communities. The principal object of the study was to investigate the effect, if any, of water hardness on the incidence of such diseases. What type of response scale is involved here? Analyse the data paying particular attention to the following points.

1. possible effects of water hardness on the incidence of CNS disorders.
2. differences between manual and non-manual workers in the incidence of CNS disorders.
3. anomalous geographical effects.
4. any systematic differences in the distribution of types of CNS disorders.

Give a brief non-technical summary of your conclusions.

5.6 Show that the complementary log-log model (5.3) is equivalent to the continuation-ratio or nested response model

$$g\{\pi_j(\mathbf{x})/(1 - \gamma_{j-1}(\mathbf{x}))\} = \alpha_j - \beta^T \mathbf{x}$$

if $g(\cdot)$ is the complementary log-log function. Express α_j in terms of the ‘cut-points’ $\theta_1, \dots, \theta_{k-1}$ appearing in (5.3). [Läärä and Matthews, 1985].

Check this claim numerically by replacing the logistic link in (5.20) and (5.21) with the complementary log-log link. Why are

Table 5.3 *Frequencies of central nervous system malformations in live births in eight South Wales communities (1964-66).*

Area	Non-manual				Manual			
	No CNS malformation		CNS malformation		No CNS malformation		CNS malformation	
	An. [†]	Sp. [‡]	Other		An.	Sp.	Other	Water hardness
<i>Cardiff</i>	4091	5	9	5	9424	31	33	14
<i>Newport</i>	1515	1	7	0	4610	3	15	6
<i>Swansea</i>	2394	9	5	0	5526	19	30	4
<i>Glamorgan E.</i>	3163	9	14	3	13217	55	71	19
<i>Glamorgan W.</i>	1979	5	10	1	8195	30	44	10
<i>Glamorgan C.</i>	4838	11	12	2	7803	25	28	12
<i>Mormouth V.</i>	2362	6	8	4	9962	36	37	13
<i>Mormouth other</i>	1604	3	6	0	3172	8	13	3
								122

[†]Anencephalus, the most serious malformation of the central nervous system.

[‡]Spina bifida, without anencephalus.

Where both malformations are present, anencephalus is recorded.

Data taken from Lowe *et al.* (1971) are reconstructed from totals and rates.

Water hardness, which is measured in parts per million, varies to some extent within communities. For details of the within-community variability, see Appendix A of Lowe *et al.* (1971).

the fitted values for category II different for the two models? Show also that the corresponding logit models (5.1) and (5.10) are not equivalent.

5.7 Consider the multinomial response model (5.7) with scores $s = (1, 0, \dots, 0)$. Show that, with these scores, the log-linear model is equivalent to the nested response model

$$\text{logit } \pi_1(\mathbf{x}_i) = \eta_1 + \boldsymbol{\beta}^T \mathbf{x}_i$$

$$\text{logit} \left(\frac{\pi_j(\mathbf{x}_i)}{1 - \gamma_{j-1}(\mathbf{x}_i)} \right) = \eta_j \quad j \geq 2.$$

5.8 Let Y_{ij} be the observations in a two-way table with n independent rows such that $\mathbf{Y}_i \sim M(m_i, \boldsymbol{\pi})$ on k categories. Consider the statistic

$$T = \sum_{ij} r_i s_j Y_{ij},$$

with given scores r_i, s_j , as a possible statistic for testing the hypothesis of independence or no row effect. Show that, under the hypothesis

$$E(T) = m_* \mu_r \mu_s$$

$$\text{var}(T) = \sum_i m_i r_i^2 \sigma_s^2$$

where

$$\mu_r = \sum m_i r_i / m_*,$$

$$\mu_s = \sum \pi_j s_j, \quad \tilde{\mu}_s = \sum y_{*j} s_j / m_*,$$

$$\sigma_s^2 = \sum \pi_j (s_j - \mu_s)^2, \quad \tilde{\sigma}_s^2 = \sum y_{*j} (s_j - \tilde{\mu}_s)^2 / m_*.$$

Explain why, for fixed n, k , the ‘standardized statistic’

$$\frac{T - m_* \mu_r \tilde{\mu}_s}{\sqrt{\tilde{\sigma}_s^2 \sum m_i r_i^2}}$$

is approximately Normally distributed but not with unit variance in the limit as $m_i \rightarrow \infty$. Show that the ‘correct’ standardization is

$$\frac{T - m_* \mu_r \tilde{\mu}_s}{\sigma_r \tilde{\sigma}_s \sqrt{m_*}}$$

where $\sigma_r^2 = \sum m_i (r_i - \mu_r)^2 / m_*$. [Yates, 1948].

5.9 Consider the proportional-odds model

$$\text{logit } \gamma_j(x_i) = \theta_j - \beta x_i$$

with x and β both scalars. Denote by $\hat{\theta}_j$, $\hat{\pi}_j$ the fitted parameters and probabilities under the hypothesis that $\beta = 0$. Show that the derivative of the log likelihood with respect to β at $\beta = 0$, $\theta_j = \hat{\theta}_j$, is given by

$$T = \sum R_{ij} x_i s_j$$

where $R_{ij} = Y_{ij} - m_i \hat{\pi}_j$ is the residual under independence and $s_j = \hat{\gamma}_j + \hat{\gamma}_{j-1}$. [Tests based on the log-likelihood derivative are sometimes called ‘score tests’.]

5.10 Use the results of Exercises 5.2 and 5.7 to find the approximate mean and variance of T in the previous exercise. Hence construct a simple test of the hypothesis that $\beta = 0$. Show that in the two-sample problem T is equivalent to Wilcoxon’s statistic.

5.11 Repeat the calculations of the previous two exercises replacing the proportional-odds model with the complementary log-log model. Which non-parametric test does T now correspond to?

5.12 Show that the score test based on the log-linear model (5.7) is identical to the score test based on the linear logistic model (5.1) provided that ridit scores are used for the response categories in (5.7). [Ridit scores (Bross, 1958) are proportional to the average category rank.]

5.13 Table 5.4, taken from Yates (1948), gives teachers’ ratings for homework, together with an assessment of homework facilities, for 1019 schoolchildren. In both cases, A denotes the highest or best rating and subsequent letters denote lower grades.

1. Which variable is the response?
2. Fit the model of independence and look for patterns among the residuals. Compute X^2 and D and show that these are approximately equal to their degrees of freedom.
3. Using integer-valued scores for both ratings, compute the statistic T as described in Exercise 5.7. Show that the standardized statistic is 1.527, corresponding to an approximate one-sided p -value of 6.3%.
4. Fit the linear complementary log-log model (5.3) using a quantitative integer-valued covariate for homework conditions.

Show that $\hat{\beta} = 0.0476$, $\text{se}(\hat{\beta}) = 0.027$, corresponding to a one-sided p -value of 3.9%. Comment on the direction and magnitude of the estimated effect.

5. Fit the log-linear model (5.7) using integer-valued scores for both rows and columns. Show that $\hat{\beta}/\text{se}(\hat{\beta}) = 1.525$, and that the reduction in deviance is 2.33 on one degree of freedom. Why are these values so remarkably similar to Yates's statistic in part 3. above?

Table 5.4 Relation between conditions under which homework was carried out, and teacher's assessment of homework quality.

Homework conditions	Teacher's rating			Total
	A	B	C	
A	141	131	36	308
B	67	66	14	147
C	114	143	38	295
D	79	72	28	179
E	39	35	16	90

5.14 In Example 2 of section 5.2.5, what modifications to the design of the study would you make if the available test animals comprised 60 milch cows and 20 heifers? What modifications would be required in the analysis of the data from this experiment?

5.15 *Logistic discrimination:* Suppose that a population of individuals is partitioned into k sub-populations or groups, G_1, \dots, G_k , say. It may be helpful to think of the groups as species or distinct populations of the same genus. Measurements Z made on individuals have the following distributions in the k groups:

$$G_j: \quad Z \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}), \quad j = 1, \dots, k.$$

Let \mathbf{z}^* be an observation made on an individual drawn at random from the combined population. The prior odds that the individual belongs to G_j are $\pi_j/(1 - \pi_j)$. Show that the posterior odds for G_j given \mathbf{z}^* are

$$\text{odds}(Y = j | \mathbf{z}^*) = \frac{\pi_j}{1 - \pi_j} \times \frac{\exp(\alpha_j + \boldsymbol{\beta}_j^T \mathbf{z}^*)}{\sum_i \exp(\alpha_i + \boldsymbol{\beta}_i^T \mathbf{z}^*)}.$$

Find expressions for α_j and β_j in terms of μ_j and Σ .

What simplifications can be made if the k Normal means μ_j lie on a straight line in R^p ?

Comment briefly on the differences between maximum likelihood estimation of α_j and β_j via the Normal-theory likelihood and estimation via logistic regression.

5.16 Show that the quadratic form

$$\sum_1^{k-1} \frac{(Z_j - m\gamma_j)^2}{m} \left(\frac{1}{\pi_j} + \frac{1}{\pi_{j+1}} \right) - 2 \sum_{j=1}^{k-2} \frac{(Z_j - m\gamma_j)(Z_{j+1} - m\gamma_{j+1})}{m\pi_{j+1}},$$

given in section 5.3.4, is identical to Pearson's X^2 statistic. In the above expression Z_j are the components of the cumulative multinomial vector and $E(Z_j) = m\gamma_j$.

Table 5.5 *Effect of mother's age and smoking habits on gestation period and perinatal mortality*

Gestation period (days)	Mother's age	Cigarettes smoked	Perinatal mortality mortality/total births
197–260	< 30	≤ 5	50/365
		> 5	9/49
	30+	≤ 5	41/188
		> 5	4/15
261+	< 30	≤ 5	24/4036
		> 5	6/465
	30+	≤ 5	14/1508
		> 5	1/125

Source: Wermuth (1976).

5.17 Table 5.5, taken from Wermuth (1976), gives the gestation period and perinatal mortality rates for a group of German women, many of whom were pregnant for the first time or had complications with previous pregnancies. Which are the response variables? Examine first how the gestation period or probability of premature birth is related to mother's age and smoking habits. Second, examine how the perinatal mortality rate is related to gestation period, mother's age and smoking habits. Summarize your conclusions in non-technical language.

Table 5.6 Frequency table for the 54 numbers used in the Illinois lottery for the 12-month period ending 12 Nov 1988. Six numbers are drawn each week.

Tens	Units										Total
	0	1	2	3	4	5	6	7	8	9	
0+	9	8	5	6	9	4	7	7	6	61	
10+	5	5	6	3	5	4	5	5	3	2	43
20+	6	3	7	10	9	10	9	7	1	7	69
30+	11	6	2	9	10	4	8	4	9	6	69
40+	5	10	3	7	8	3	4	4	1	4	49
50+	3	5	4	4	5						21
<i>Total</i>	30	38	30	38	43	30	30	27	21	25	312

Source: Chicago Tribune, 14 Nov 1988.

5.18 Pick-6 is the weekly Illinois lottery in which the winning ticket comprises six unordered numbers in the range 1–54. The winning numbers are chosen by a physical randomizing device in which 54 numbered ping-pong balls are mixed by a draught of air in a closed transparent container. Six balls carrying the winning numbers are permitted to escape, one at a time, through a hole in the top of the apparatus. The frequency of occurrence of each the 54 numbers in a 12-month period is shown in Table 5.6.

By fitting a log-linear model or otherwise, test the following hypotheses, all of which refer to the uniformity of the numbers generated by the randomizing device.

1. that the variation of the frequencies in Table 5.6 is consistent with the hypothesis of uniformity.
2. that the variation of the column totals in Table 5.6 is consistent with the hypothesis of uniformity.
3. that the variation of the row totals in Table 5.6 is consistent with the hypothesis of uniformity.
4. that the frequency of occurrence of the numbers 45–54 is the same as that of the remaining numbers.

You are now given the additional information that for the first 24 weeks only the numbers 1–44 were used: thereafter all 54 numbers were used. Test hypotheses (1.) and (2.) above making due allowance for this change of regime after 24 weeks.

5.19 Repeat the calculations of the previous exercise making due allowance for the fact that the six balls are chosen each week without replacement from the pool of 44 or 54 balls. Show that for

large m , and provided that the number of balls remains constant from week to week,

$$\frac{(k-1)X^2}{k-n} \sim \chi_{k-1}^2$$

where in this case, $k = 44$ or 54 , $n = 6$, $m = 52$ and X^2 is Pearson's statistic computed in the usual way as if the counts were multinomial variables on k categories. This finite population correction, which is not asymptotically negligible, should also be used to correct deviance statistics derived from Poisson models. Alternatively the counts may be taken as binomial variables with index 52 for numbers 1–44 and 28 for numbers 45–54. It is then necessary to include an offset in all models.

5.20 Under the conditions described in the previous exercise show that

$$E(X^2) = k - n,$$

$$\text{var}(X^2) = 2 \frac{(k-n)^2}{k-1} \frac{m-1}{m}.$$

Check that these calculations are correct for $n = 1$ and $n = k - 1$.

CHAPTER 6

Log-linear models

6.1 Introduction

In this chapter we are concerned mainly with counted data not in the form of proportions. Typical examples involve counts of events in a Poisson or Poisson-like process where the upper limit to the number is infinite or effectively so. One example discussed in section 6.3 deals with the number of incidents involving damage to ships of a specified type over a given period of time. Classical examples involve radiation counts as measured in, say, particles per second by a Geiger counter. In behavioural studies counts of incidents in a time interval of specified length are often recorded.

Under idealized experimental conditions when successive events occur independently and at the same rate, the Poisson model is appropriate for the number of events observed. However, even in well-conducted laboratory experiments, departures from the idealized Poisson model are to be expected for several reasons. Geiger counters experience a ‘dead-time’ following the arrival of a particle. During this short interval the apparatus is incapable of recording further particles. Consequently, when the radioactive decay rate is high, the ‘dead-time’ phenomenon leads to noticeable departures from the Poisson model for the number of events recorded. In behavioural studies involving primates or other animals, incidents usually occur in spurts or clusters. The net effect is that the number of recorded events is more variable than the simple Poisson model would suggest. Similarly with the data on ship damage, inter-ship variability leads to over-dispersion relative to the Poisson model. Here, unless there is strong evidence to the contrary, we avoid the assumption of Poisson variation and assume only that

$$\text{var}(Y_i) = \sigma^2 E(Y_i), \quad (6.1)$$

where σ^2 , the dispersion parameter, is assumed constant over the data. Under-dispersion, a phenomenon less common in practice, is included here by putting $\sigma^2 < 1$ (Chapter 9).

In log-linear models the dependence of $\mu_i = E(Y_i)$ on the covariate vector \mathbf{x}_i is assumed to be multiplicative and is usually written in the logarithmic form

$$\log \mu_i = \eta_i = \boldsymbol{\beta}^T \mathbf{x}_i; \quad i = 1, \dots, n. \quad (6.2)$$

When we use the term log-linear models we mean primarily the log-linear relationship (6.2); often (6.1) is tacitly assumed as a secondary aspect of the model but the choice of variance assumption is usually less important than the choice of link and covariates in (6.2). In applications both components of the log-linear model, but primarily (6.2), require checking.

All log-linear models have the form (6.2). Variety is created by different forms of model matrices; there is an obvious analogy with analysis-of-variance and linear regression models. In the theoretical development it is not usually necessary to specify the form of \mathbf{X} , though in applications, of course, the form of \mathbf{X} is all-important. In section 6.4, which deals with the connection between log-linear and multinomial response models, some aspects of the structure of \mathbf{X} are important. It is shown that, under certain conditions, there is an equivalence between log-linear models and certain multinomial response models dealt with in Chapters 4 and 5.

6.2 Likelihood functions

6.2.1 Poisson distribution

In Chapters 4 and 5 we encountered the binomial and multinomial distributions. These are appropriate as models for proportions where the total is fixed. In the present chapter we concentrate on the Poisson distribution for which the sample space is the set of non-negative integers. In particular there is no finite upper limit on the values that may be observed. The probability distribution is given by

$$\text{pr}(Y = y) = e^{-\mu} \mu^y / y!; \quad y = 0, 1, 2, \dots,$$

from which the cumulant generating function

$$\mu(e^t - 1)$$

may be derived. It follows that the mean, variance and all other cumulants of Y are equal to μ . Any random variable whose cumulants are $O(n)$, where n is some quantity tending to infinity, has the limiting property

$$(Y - \mu)/\kappa_2^{1/2} \sim N(0, 1) + O_p(n^{-1/2}).$$

In particular for the Poisson distribution, as $\mu \rightarrow \infty$

$$(Y - \mu)/\mu^{1/2} \sim N(0, 1) + O_p(\mu^{-1/2}).$$

This proof may also be applied to the binomial and hypergeometric distributions. For the latter distribution, the appropriate limit is approached as the minimum marginal total tends to infinity.

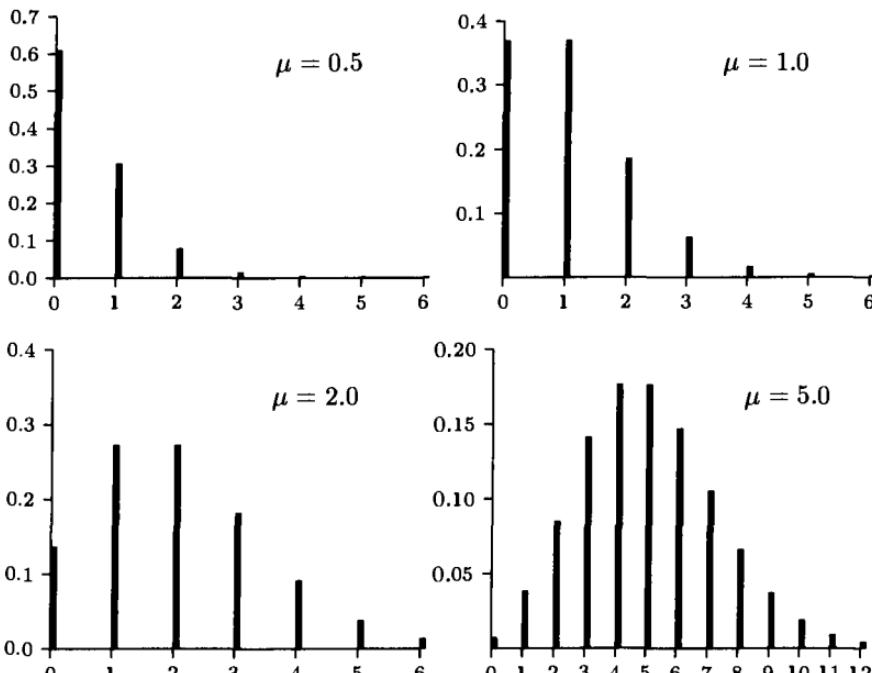


Fig. 6.1. The Poisson distribution for $\mu = 0.5, 1.0, 2.0$ and 5.0 .

Illustrations of the Poisson distribution are given in Fig. 6.1 for $\mu = 0.5, 1.0, 2.0$, and 5.0 . These illustrate the extent of the skewness particularly for small values of μ , and the approach to the Normal limit for large μ . Such a Normal limit refers to the cumulative distribution and not to the density.

The following properties of the Poisson distribution are sometimes useful in applications. The variance-stabilizing transform is $Y^{1/2}$ in the sense that for large μ

$$E(Y^{1/2}) \simeq \mu^{1/2} \quad \text{and} \quad \text{var}(Y^{1/2}) \simeq 1/4,$$

the error terms being $O(\mu^{-1})$. In fact the subsequent terms in an asymptotic expansion are

$$\begin{aligned} E(Y^{1/2}) &\simeq \mu^{1/2}\{1 - 1/(8\mu)\} \\ \text{var}(Y^{1/2}) &\simeq \{1 + 3/(8\mu)\}/4, \end{aligned}$$

showing that the variance is only approximately constant. See Exercises 4.8–4.11 and 6.1.

The power transformation to symmetry is $Y^{2/3}$ (Anscombe, 1953) in the sense that the standardized skewness of $Y^{2/3}$ is $O(\mu^{-1})$ rather than $O(\mu^{-1/2})$ for Y or $Y^{1/2}$. In fact the cumulants of $Y^{2/3}$ are

$$\begin{aligned} E(Y^{2/3}) &\simeq \mu^{2/3}\{1 - 1/(9\mu)\} \\ \text{var}(Y^{2/3}) &\simeq \mu^{1/3}\frac{4}{9}\{1 + 1/(6\mu)\} \\ \kappa_3(Y^{2/3}) &\simeq O(\mu^{-1}). \end{aligned}$$

See Exercise 6.1. Thus the standardized skewness of $Y^{2/3}$ is $O(\mu^{-3/2})$ rather than the $O(\mu^{-1})$ claimed above. Neither of these transformations involves the unknown μ , although the value of μ is required when computing tail probabilities.

An alternative transformation derived as a quadratic approximation to the signed deviance statistic produces both approximate symmetry and stability of variance. This is

$$g(Y) = \begin{cases} 3Y^{1/2} - 3Y^{1/6}\mu^{1/3} + \mu^{-1/2}/6; & Y \neq 0, \\ -(2\mu)^{1/2} + \mu^{-1/2}/6; & Y = 0. \end{cases}$$

Since $g(Y)$ is approximately standard Normal for large μ , tail probabilities may be approximated by

$$\text{pr}(Y \geq y) \simeq 1 - \Phi(g(y - \frac{1}{2})),$$

with an error of order μ^{-1} rather than $O(\mu^{-1/2})$. This approximation is surprisingly accurate even for modest values of μ . For instance with $\mu = 5$ we obtain the following approximation:

y	7	8	9	10	11	12	13
$\text{pr}(Y \geq y)$ (exact)	0.2378	0.1334	0.0681	0.0318	0.0137	0.0055	0.0020
(approx)	0.2373	0.1328	0.0678	0.0318	0.0137	0.0055	0.0021

Non-monotonicity of the function $g(y)$ is not a serious concern because, for discrete y , the effect occurs only if $\mu > 38$ and then in a region of negligibly small probability.

6.2.2 The Poisson log-likelihood function

For a single observation y the contribution to the log likelihood is $y \log \mu - \mu$. Plots of this function versus μ , $\log \mu$ and $\mu^{1/3}$ are given in Fig. 6.2 for $y = 1$. To a close approximation it can be seen that, for $y > 0$,

$$y \log \mu - \mu \simeq y \log y - y - 9y^{1/3}(\mu^{1/3} - y^{1/3})^2/2.$$

For a derivation of this approximation see Exercise 6.2. The signed square root of twice the difference between the function and its maximum value is $3y^{1/6}(y^{1/3} - \mu^{1/3})$, which is the leading term in the transformation $g(y)$ above.

For a vector of independent observations the log likelihood is

$$l(\boldsymbol{\mu}, \mathbf{y}) = \sum (y_i \log \mu_i - \mu_i), \quad (6.3)$$

so that the deviance function is given by

$$\begin{aligned} D(\mathbf{y}; \boldsymbol{\mu}) &= 2l(\mathbf{y}, \mathbf{y}) - 2l(\boldsymbol{\mu}, \mathbf{y}) \\ &= 2 \sum \{y_i \log(y_i/\mu_i) - (y_i - \mu_i)\} \\ &\simeq 9 \sum y_i^{1/3} (y_i^{1/3} - \mu_i^{1/3})^2. \end{aligned}$$

If a constant term is included in the model it can be shown that $\sum(y_i - \hat{\mu}_i) = 0$, so that $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ may then be written in the more usual form $2 \sum y_i \log(y_i/\hat{\mu}_i)$.

Another approximation to $D(\mathbf{y}; \boldsymbol{\mu})$ for large μ is obtained by expanding as a Taylor series in $(y - \mu)/\mu$. We find

$$D(\mathbf{y}; \boldsymbol{\mu}) \simeq \sum_i (y_i - \mu_i)^2 / \mu_i,$$

which is less accurate than the quadratic approximation on the $\mu^{1/3}$ scale. This statistic is due to Pearson (1900).

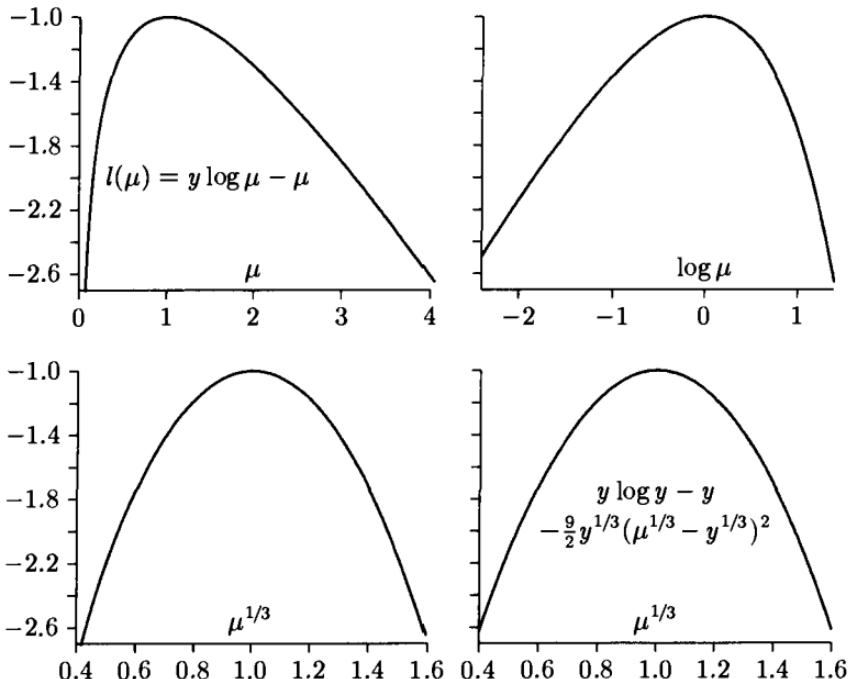


Fig. 6.2. The Poisson log likelihood function for $y = 1$, using scales μ , $\log \mu$ and $\mu^{1/3}$, together with quadratic approximation on cube root scale.

6.2.3 Over-dispersion

Suppose now that the dispersion of the data is greater than that predicted by the Poisson model, i.e. $\text{var}(Y) > E(Y)$. This phenomenon can arise in a number of different ways. We might, for example, observe a Poisson process over an interval whose length is random rather than fixed. Alternatively the data might be produced by a clustered Poisson process where each event contributes a random amount to the total. In other words, we observe $Y = Z_1 + Z_2 + \dots + Z_N$ where the Z s are independent and identically distributed and N has a Poisson distribution independent of Z . We find that

$$E(Y) = E(N)E(Z)$$

$$\text{and } \text{var}(Y) = E(N) \text{var}(Z) + \text{var}(N)\{E(Z)\}^2 = E(N)E(Z^2),$$

so that there is over-dispersion if $E(Z^2) > E(Z)$.

Another way in which over-dispersion may arise is as follows. In behavioural studies and in studies of accident-proneness where there is inter-subject variability, the number of incidents Y for a given individual might be Poisson with mean Z . This mean itself may be regarded as a random variable which we may suppose in the population to have the gamma distribution with mean μ and index $\phi\mu$. In other words $E(Z) = \mu$ and $\text{var}(Z) = \mu/\phi$, mimicking the Poisson distribution itself. This mixture leads to the negative binomial distribution

$$\text{pr}(Y = y; \mu, \phi) = \frac{\Gamma(y + \phi\mu)\phi^{\phi\mu}}{y! \Gamma(\phi\mu)(1 + \phi)^{y + \phi\mu}} ; \quad y = 0, 1, 2, \dots$$

(Plackett, 1981, p. 6). The mean and variance are $E(Y) = \mu$ and $\text{var}(Y) = \mu(1 + \phi)/\phi$. If the regression model is specified in terms of μ , say $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$, and if ϕ is unknown but constant, then the estimating equations for $\boldsymbol{\beta}$ are in general different from those obtained by weighted least squares, and only in very simple cases do the two methods coincide. However, it may be shown that the two sets of parameter estimates, one based on the negative binomial likelihood and the other on the Poisson likelihood, differ by a term that is $O_p(\phi^{-2})$ for large ϕ . For modest amounts of over-dispersion this difference may be neglected (see also section 9.2).

An alternative mixing scheme, in which the variance of Z is proportional to the square of its mean, is obtained by assuming Z to have the gamma distribution with mean μ and constant index ν independent of μ . This mixture again leads to the negative binomial distribution, but now parameterized in such a way that

$$\text{var}(Y) = \mu + \mu^2/\nu.$$

The variance function is now quadratic instead of linear.

If the precise mechanism that produces the over-dispersion or under-dispersion is known (e.g. as with electronic counters), specific methods may be used. In the absence of such knowledge it is convenient to assume as an approximation that $\text{var}(Y) = \sigma^2\mu$ for some constant σ^2 . This assumption can and should be checked, but even relatively substantial errors in the assumed functional form of $\text{var}(Y)$ generally have only a small effect on the conclusions. Parameter estimates may be obtained by maximizing the Poisson log likelihood (6.3) using, for example, the general method of

Chapter 2, with the inverse matrix of second derivatives being multiplied by an estimate of σ^2 in order to obtain an appropriate measure of precision for $\hat{\beta}$. For details see Chapter 9.

6.2.4 Asymptotic theory

The usual asymptotic results concerning consistency and asymptotic Normality of $\hat{\beta}$ are valid provided that the eigenvalues of the information matrix increase without limit. This condition is usually satisfied if p is fixed and $n \rightarrow \infty$ or, for fixed n and p , if $\mu_i \rightarrow \infty$ for each i . The asymptotic covariance matrix of $\hat{\beta}$ is $\sigma^2 \mathbf{i}_\beta^{-1}$ where \mathbf{i}_β , the negative matrix of second derivatives of (6.3), emerges in a very natural way in the iterative weighted least-squares estimation procedure.

The dispersion parameter σ^2 can, if required, be estimated by

$$\tilde{\sigma}^2 = X^2/(n - p) = \sum_i \frac{(y_i - \mu_i)^2}{\mu_i} / (n - p). \quad (6.4)$$

By analogy with (4.14) and (5.13) the effective degrees of freedom for $\tilde{\sigma}^2$ are given by $f = (n - p)/(1 + \frac{1}{2}\bar{\gamma}_2)$ so that approximate confidence limits for individual components of β would be based on the t_f distribution if σ^2 is unknown. This is a minor refinement and for most purposes, unless many of the means are less than 1.0, the Normal or t_{n-p} approximation is adequate.

6.3 Examples

6.3.1 A biological assay of tuberculins

Fisher (1949) published some data concerning a biological assay of two tuberculins, designated Standard and Weybridge, using 'bovine subjects'. The observations are measurements in millimetres of a thickening of the skin observable in a set number of hours after intradermal injection of the tuberculin. The following is a simplified description of the experiment. One hundred and twenty cows were divided into four classes, I, II, III and IV, of 30 cows each. The four tuberculin treatments applied were

- A Standard double,
- B Standard single,

- C Weybridge single,
 D Weybridge half,

where 'single' refers to the amount 0.05 mg. On each cow there were four sites of application, with each cow receiving each of the four tuberculin treatments. The cow classes I, II, III and IV differed only in the sites on the neck, here numbered 1-4, at which the various tuberculins were applied in accordance with the layout in Table 6.1a. In other words, all 30 cows in class IV had the Weybridge half preparation applied at site #1, Weybridge single at site #2 and so on. The observations in Table 6.1b are the totals for each site and cow class of the observed thickenings on 30 cows.

Table 6.1a Latin square design used for tuberculin assay

Sites on neck	Cow class			
	I	II	III	IV
1	A	B	C	D
2	B	A	D	C
3	C	D	A	B
4	D	C	B	A

Table 6.1b Responses in mm. in a biological assay of tuberculins

Sites on neck	Cow class				Total
	I	II	III	IV	
1	454	249	349	249	1301
2	408	322	312	347	1389
3	523	268	411	285	1487
4	364	283	266	290	1203
Total	1749	1122	1338	1171	5380

After extensive preliminary investigation and prior to summarizing the data in the form of Table 6.1, Fisher concluded (a) that the effect of treatment and choice of site were multiplicative and (b) that the variance of any observation was roughly proportional to its expected value. Thus, although the response is a measurement, not a count, the methods of this chapter apply.

The systematic part of the model is thus log-linear where the model matrix \mathbf{X} is the incidence matrix for a non-cyclic 4×4

Latin square with the tuberculin treatments A, B, C and D indexed according to a 2×2 factorial arrangement with no interaction. In this way we can examine the relative potency of the two tuberculin preparations either at the high-dose level or at the low-dose level. The required model formula is thus

$$\text{site} + \text{class} + \text{volume} + \text{tuberculin}$$

where *site* and *class* are factors having four levels each and *tuberculin* is a two-level factor denoting Standard and Weybridge respectively. The remaining variate *volume*, or $\log(\text{volume})$, can be treated either as a quantitative covariate taking values $-1, 0, 1$ for 'half', 'single' and 'double' respectively or as a two-level factor denoting 'low dose' and 'high dose' for each tuberculin. In the latter case 'low dose' for the Standard preparation does not represent the same volume as 'low dose' for Weybridge. If *volume* denotes the quantitative covariate, then the tuberculin effect is the contrast between Weybridge and Standard at equal volumes. By contrast, if *volume* denotes the two-level factor, the tuberculin effect is the contrast between Weybridge half and Standard single, these being the low-dose levels, or between Weybridge single and Standard double at the high-dose level. The choice of parameterization is a matter of taste or convenience. Both parameterizations produce the same fitted values and identical conclusions, but it is important to understand how the parameterization of *volume* affects the tuberculin contrast.

Parameter estimates found by maximizing (6.3) and are similar to those obtained by Fisher (1949) who used a non-iterative method. The values are given in the Table below.

		<i>Equation (6.3)</i>	<i>Fisher</i>
B	Standard single	0.0000	0.0000
A	Standard double	0.2095	0.2089
D	Weybridge half	0.0026	0.0019
C	Weybridge single	0.2121	0.2108

Using (6.4) we find $\tilde{\sigma}^2 = 1.410/7 = 0.2014$ compared with Fisher's value of 0.2018. Taken together with the relevant components of the inverse matrix of second derivatives the standard errors for the treatment contrasts on the log scale are:

<i>Contrast</i>	<i>Estimate</i>	<i>SE</i>	<i>Correlation</i>
High dose vs low dose	0.2095	0.0124	
Weybridge single vs Standard double	0.0026	0.0123	-0.0053

Confidence limits may be constructed based on the t_7 -distribution. These estimates show that the Weybridge single is slightly more potent than the Standard double dose but not significantly so, the ratio of estimated responses being $\exp(0.0026)$. Similarly, doubling the dose increases the response by an estimated factor $\exp(0.2095)$ equal to a 23.3% increase. This factor applies to both Standard and Weybridge.

The relative potency of Weybridge to Standard is the ratio of the volume of Standard to the volume of Weybridge required to produce equal responses. The estimate obtained here is $2^{0.2121/0.2095} = 2.017$ compared with Fisher's estimate of 2.009 (which should apparently have been 2.013).

In the analysis just given it is assumed (a) that the response at one site on the neck is unaffected by the treatment applied at other sites and (b) that the effect on the logarithmic scale of doubling the dose of the Standard preparation is the same as doubling the dose of the Weybridge preparation. This latter assumption can and should be checked by including in the model the interaction term between preparation and volume. This is equivalent to regarding the treatments A, B, C and D as an unstructured four-level factor instead of as two two-level factors having no interaction. The required model formula is

$$\text{site} + \text{class} + \text{volume.tuberculin}.$$

An F -test on 1,6 degrees of freedom, rather than a χ^2 test, is required here because σ^2 is unknown. Alternatively, and perhaps preferably, a t -test based on the parameter estimate may be used.

In the design of this experiment it was recognized that the variability between responses on different animals would be very large but that on different sites on the same animal the variability would be considerably less. It is essential, therefore, in the interests of high precision to make comparisons of the two preparations on the same animal. In the arrangement described in Table 6.1 each

cow is assigned a class I-IV, so that contrasts between sites and between treatments are within the same animal. On the other hand, contrasts between treatment classes are between animals and thus involve an additional component of dispersion. Strictly, the analysis should have been made conditional on the observed column totals but in fact, as we shall see in the following section, this would make no difference to the numerical values of the treatment contrasts or to their estimated precision. However, because of this additional component of variability the standard errors for the treatment-class contrasts in the log-linear model that we have used are inappropriate. This complication does not invalidate the analysis given here because the effects of interest are contrasts within the same animal and do not involve between-animal variation. For further details see section 14.3.

Fisher gives a detailed discussion of the conclusions to be drawn from these data, including a study of the components of dispersion just described. The principal conclusion, that the relative potency is just in excess of 2.0, is complicated by the later discovery, using careful comparative tests with guinea-pigs, that the estimated relative potency was 0.9. Thus it would appear that the two tuberculin preparations must be qualitatively different, though such a difference is unlikely to show up in a study confined to a single species.

Further details of this experiment, including the individual measurements at each site on each cow after 48, 72 and 96 hours, are given in Fisher's paper.

6.3.2 *A study of wave damage to cargo ships*

The data in Table 6.2, kindly provided by J. Crilley and L.N. Heminway of Lloyd's Register of Shipping, concern a type of damage caused by waves to the forward section of certain cargo-carrying vessels. For the purpose of setting standards for hull construction we need to know the risk of damage associated with the three classifying factors shown below.

Ship type: A-E

Year of construction: 1960-64, 1965-69, 1970-74, 1975-79

Period of operation: 1960-74, 1975-79

Table 6.2 *Number of reported damage incidents and aggregate months service by ship type, year of construction and period of operation*

<i>Ship type</i>	<i>Year of construction</i>	<i>Period of operation</i>	<i>Aggregate months service</i>	<i>Number of damage incidents</i>
A	1960-64	1960-74	127	0
A	1960-64	1975-79	63	0
A	1965-69	1960-74	1095	3
A	1965-69	1975-79	1095	4
A	1970-74	1960-74	1512	6
A	1970-74	1975-79	3353	18
A	1975-79	1960-74	0	0*
A	1975-79	1975-79	2244	11
B	1960-64	1960-74	44882	39
B	1960-64	1975-79	17176	29
B	1965-69	1960-74	28609	58
B	1965-69	1975-79	20370	53
B	1970-74	1960-74	7064	12
B	1970-74	1975-79	13099	44
B	1975-79	1960-74	0	0*
B	1975-79	1975-79	7117	18
C	1960-64	1960-74	1179	1
C	1960-64	1975-79	552	1
C	1965-69	1960-74	781	0
C	1965-69	1975-79	676	1
C	1970-74	1960-74	783	6
C	1970-74	1975-79	1948	2
C	1975-79	1960-74	0	0*
C	1975-79	1975-79	274	1
D	1960-64	1960-74	251	0
D	1960-64	1975-79	105	0
D	1965-69	1960-74	288	0
D	1965-69	1975-79	192	0
D	1970-74	1960-74	349	2
D	1970-74	1975-79	1208	11
D	1975-79	1960-74	0	0*
D	1975-79	1975-79	2051	4
E	1960-64	1960-74	45	0
E	1960-64	1975-79	0	0**
E	1965-69	1960-74	789	7
E	1965-69	1975-79	437	7
E	1970-74	1960-74	1157	5
E	1970-74	1975-79	2161	12
E	1975-79	1960-74	0	0*
E	1975-79	1975-79	542	1

*Necessarily empty cells.

**Accidentally empty cell

Data courtesy of J. Crilley and L.N. Heminway, Lloyd's Register of Shipping.

The data give the number of damage incidents (as distinct from the number of ships damaged), the aggregate number of months service or total period at risk and the three classifying factors. Note that a single ship may be damaged more than once and furthermore that some ships will have been operating in both periods. No ships constructed after 1975 could have operated before 1974, explaining five of the six (necessarily) empty cells.

It seems reasonable to suppose that the number of damage incidents is directly proportional to the aggregate months service or total period of risk. This assumption can be checked later. Furthermore, multiplicative effects seem more plausible here than additive effects. These considerations lead to the initial very simple model:

$$\begin{aligned} \log(\text{expected number of damage incidents}) \\ = \beta_0 + \log(\text{aggregate months service}) \\ + (\text{effect due to ship type}) \\ + (\text{effect due to year of construction}) \\ + (\text{effect due to service period}). \end{aligned} \quad (6.5)$$

The last three terms in this model are qualitative factors. The first term following the intercept is a quantitative variate whose regression coefficient is known to be 1. Such a term is sometimes called an *offset*.

For the random variation in the model, the Poisson distribution might be thought appropriate as a first approximation, but there is undoubtedly some inter-ship variability in accident-proneness. This would lead to over-dispersion as described in section 6.2.2. For these reasons we assume simply that $\text{var}(Y) = \sigma^2 E(Y)$ and expect to find $\sigma^2 > 1$. Parameter estimates are computed using the Poisson log likelihood.

The main-effects model (6.5) fits these data reasonably well but some large residuals remain, particularly observation 21 for which the observed value is 6 and the fitted value is 1.47, giving a standardized residual of 2.87. Here we use the standardization $(y - \hat{\mu})/(\hat{\sigma}\hat{\mu}^{1/2})$, with $\hat{\sigma}^2 = 1.69$. By way of comparison, the deviance residual is 2.15.

As part of the standard procedure for model checking we note the following:

1. All of the main effects are highly significant.
2. The coefficient of $\log(\text{aggregate months service})$, when estimated, is 0.903 with approximate standard error 0.13, confirming the assumed prior value of unity.
3. Neither of the two-factor interactions involving service period is significant.
4. There is inconclusive evidence of an interaction between ship type and year of construction, the deviance being reduced from 38.7 with 25 degrees of freedom to 14.6 with 10. This reduction would have some significance if the Poisson model were appropriate but, with over-dispersion present, the significance of the approximate F -ratio $(38.7 - 14.6)/(15 \times 1.74) = 0.92$ vanishes completely. Here, 1.74 is the estimate of σ^2 with the interaction term included.
5. Even with the interaction term included, the standardized residual for observation 21 remains high at 2.48.

We may summarize the conclusions as follows: the number of damage incidents is roughly proportional to the length of the period at risk and there is evidence of inter-ship variability ($\tilde{\sigma}^2 = 1.69$); the estimate for the effect due to service period (after vs before the 1974 oil crisis) is 0.385 with standard error 0.154. These values are virtually unaffected by the inclusion of the interaction term. Thus, on taking exponents, we see that the rate of damage incidents increased by an estimated 47% with approximate 95% confidence limits (8%, 100%) after 1974. This percentage increase applies uniformly to all ship types regardless of when they were constructed.

Ships of types B and C have the lowest risk, type E the highest. Similarly the oldest ships appear to be the safest, with those built between 1965 and 1974 having the highest risk. Parameter estimates from the main-effects model on which these conclusions are based are given in Table 6.3. Table 6.4 gives the observed rate of damage incidents by ship type and year of construction. The reason for the suggested interaction is that the risk for ships of types A, B and C is increasing over time while the risk for type E appears to be decreasing. The above conclusions would be somewhat modified if observation 21 were set aside.

One final technical point concerns the computation of residual degrees of freedom for the model containing the interaction term. The usual method of calculation used in some computing packages

Table 6.3 *Estimates for the main effects in the ship damage example*

<i>Parameter</i>	<i>Estimate</i>	<i>Standard error</i>
Intercept	-6.41	—
Ship type A	0.00	—
Ship type B	-0.54	0.23
Ship type C	-0.69	0.43
Ship type D	-0.08	0.38
Ship type E	0.33	0.31
Year of construction 1960–64	0.00	—
Year of construction 1965–69	0.70	0.19
Year of construction 1970–74	0.82	0.22
Year of construction 1975–79	0.45	0.30
Service period 1960–74	0.00	—
Service period 1975–79	0.38	0.15

Table 6.4 *Observed rate of damage incidents ($\times 10^3$ per ship month at risk) by ship type and year of construction*

<i>Ship type</i>	<i>Year of construction</i>			
	1960–64	1965–69	1970–74	1975–79
A	0.0	3.2	4.9	4.9
B	1.1	2.3	2.3	2.5
C	1.2	0.7	2.9	3.6
D	0.0	0.0	8.3	2.0
E	0.0	11.4	5.1	1.8

gives 13 instead of 10. However, the appropriate reference set for the computation of significance levels is conditional on the observed value of the sufficient statistic for the model containing the interaction term. One component of the sufficient statistic is the two-way marginal summary given in Table 6.4. The first three columns of this table involve sums of two observations. Apart from the four zeros which give degenerate distributions, each remaining cell in the first three columns contributes one degree of freedom, giving 11 in all. One further degree of freedom is lost because of the effect due to service period. The entries in the '75–'79 column involve only one observation each and therefore contribute only a constant to the value of the statistic.

6.4 Log-linear models and multinomial response models

The following sections deal with the connection between log-linear models for frequencies and multinomial response models for proportions. The connection between the two stems from the fact that the binomial and multinomial distributions can be derived from a set of independent Poisson random variables conditionally on their total being fixed.

6.4.1 Comparison of two or more Poisson means

Suppose that Y_1, \dots, Y_k are independent Poisson random variables with means μ_1, \dots, μ_k and that we require to test the composite null hypothesis $H_0: \mu_1 = \dots = \mu_k = e^{\beta_0}$. The alternative hypothesis under consideration is that for some unknown β_1

$$\log \mu_j = \beta_0 + \beta_1 x_j,$$

where x_j are given constants. Standard theory of significance testing (Lehmann, 1986, section 4.3) leads to consideration of the test statistic $T = \sum x_j Y_j$ conditionally on the observed value of $m = \sum y_j$, which is the sufficient statistic for β_0 . In other words, in the calculation of significance levels we regard the data as having the multinomial distribution with index m and parameter vector (k^{-1}, \dots, k^{-1}) . This conditional distribution is independent of the nuisance parameter β_0 so that the one-sided significance level for alternatives $\beta_1 > 0$, namely $p^+ = \text{pr}(T \geq t_{\text{obs}}; H_0)$, can be computed from the multinomial distribution. Conditioning on the observed total $m = \sum y_i$ has the effect of eliminating the nuisance parameter from all probability calculations.

Note that under H_0 the unconditional moments of T are

$$\begin{aligned} E(T) &= \sum x_j e^{\beta_0} \simeq \sum x_j y_./k \\ \text{var}(T) &= \sum x_j^2 e^{\beta_0} \simeq \sum x_j^2 y_./k, \end{aligned}$$

which depend on β_0 . The conditional moments on the other hand are

$$\begin{aligned} E(T | Y.) &= \sum x_j y_./k \\ \text{var}(T | Y.) &= \sum (x_j - \bar{x})^2 y_./k. \end{aligned}$$

Note that the estimate of the unconditional variance of T is quite different from the exact conditional variance. The conditional variance is unaffected by the addition of a constant to each component of x .

The statistic T can equivalently be regarded as the total of a random sample of size m taken with replacement from the finite population x_1, \dots, x_k . Under H_0 , the k values are selected with equal probability: under H_A the probabilities are exponentially weighted in favour of the larger xs if $\beta_1 > 0$ or the smaller xs if $\beta_1 < 0$.

The Poisson log-likelihood function for (β_0, β_1) in this problem is

$$l_y(\beta_0, \beta_1) = \beta_0 \sum y_j + \beta_1 \sum x_j y_j - \sum \exp(\beta_0 + \beta_1 x_j).$$

In order to see how this is transformed into a multinomial response model we make the parameter transformation

$$\tau = \sum \exp(\beta_0 + \beta_1 x_j).$$

The log likelihood for (τ, β_1) becomes

$$\begin{aligned} l_Y(\tau, \beta_1) &= y_* \log \tau - \tau + \beta_1 \sum_j x_j y_j - m \log \{\sum \exp(\beta_1 x_j)\} \\ &= l_m(\tau; m) + l_{Y|m}(\beta_1; y). \end{aligned}$$

The first term above is the Poisson log likelihood for τ based on $m = Y_* \sim P(\tau)$. The second component is the multinomial log likelihood for β_1 based on the conditional distribution,

$$Y_1, \dots, Y_k | Y_* = m \sim M(m, \boldsymbol{\pi})$$

with $\pi_j = \exp(\beta_1 x_j) / \sum_i \exp(\beta_1 x_i)$. The important point here is that the marginal likelihood based on Y_* depends only on τ whereas the conditional likelihood given Y_* depends only on β_1 . Provided that no information is available concerning the value of β_0 and consequently of τ , we must conclude that all of the information concerning β_1 resides in the conditional likelihood given Y_* .

The Fisher information matrix for (τ, β_1) is

$$i_{\tau\beta} = \text{diag} \left\{ 1/\tau, \sum \pi_j (x_j - \bar{x})^2 \right\}$$

and these parameters are said to be orthogonal. It follows under suitable limiting conditions that the estimates $\hat{\tau}, \hat{\beta}_1$ must be approximately independent. The relevance of this result in the present circumstances is unclear because the precision of $\hat{\beta}_1$ is most naturally assessed from the conditional distribution given Y , whereas the precision of $\hat{\tau} = Y_*$ is based on the marginal distribution of Y_* .

6.4.2 Multinomial response models

The results given in the previous section may readily be extended to show that certain log-linear models are equivalent to multinomial response models of the kind discussed in section 5.2. The following discussion is based largely on Palmgren (1981).

It is convenient to arrange the observations Y_{ij} in a two-way table with n rows and k columns. Thus i runs from 1 to n and j from 1 to k . In practice i is often a compound index generated by the levels of two or more factors, but this complication is ignored in the algebra that follows. Consider the log-linear model

$$\log \mu_{ij} = \phi_i + \mathbf{x}_{ij}^T \boldsymbol{\beta} \quad (6.6)$$

where $\mu_{ij} = E(Y_{ij})$, \mathbf{x}_{ij} are known p -dimensional vectors, $\boldsymbol{\beta}$ is the parameter of interest and ϕ_1, \dots, ϕ_n are incidental parameters. Under this model the dimension of the parameter space, $n + p$, increases as $n \rightarrow \infty$ for fixed p . Consequently maximum-likelihood estimates cannot be guaranteed to be efficient or even consistent in the limit as $n \rightarrow \infty$. On the other hand the conditional log likelihood derived below depends only on $\boldsymbol{\beta}$ and not on ϕ and standard asymptotic theory applies directly to the conditional likelihood.

The log likelihood is

$$\begin{aligned} l_Y(\boldsymbol{\phi}, \boldsymbol{\beta}) &= \sum_{ij} \{y_{ij}(\phi_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}) - \exp(\phi_i + \mathbf{x}_{ij}^T \boldsymbol{\beta})\} \\ &= \sum_i \phi_i y_{i*} + \sum_{ij} y_{ij} \mathbf{x}_{ij}^T \boldsymbol{\beta} - \sum_{ij} \exp(\phi_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}). \end{aligned}$$

Now write $m_i = y_{i*}$ for the i th row total and make the parameter transformation

$$\tau_i = \sum_j \mu_{ij} = \sum_j \exp(\phi_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}).$$

The log likelihood, now considered as a function of (τ, β) , can be written in the form

$$\begin{aligned} l_Y(\tau, \beta) &= \sum_i (m_i \log \tau_i - \tau_i) \\ &\quad + \sum_i \left\{ \sum_j y_{ij} \mathbf{x}_{ij}^T \beta - m_i \log \left(\sum_j \exp(\mathbf{x}_{ij}^T \beta) \right) \right\} \\ &= l_m(\tau; m) + l_{Y|m}(\beta; y). \end{aligned}$$

The first term above is the Poisson log likelihood for τ based on the row totals $Y_{\cdot i} \sim P(\tau_i)$. The second term is the conditional log likelihood given $\{Y_{\cdot i}\}$, which depends only on β and not on the incidental parameters. All the information concerning β resides on the second component. In particular, it is apparent that $\hat{\beta}$ and $\text{cov}(\hat{\beta})$ based on $l_{Y|m}(\beta; y)$ are identical to those based on the full log likelihood. In other words, the log-linear model (6.6) is equivalent to the multinomial response model in which the probabilities are

$$\pi_{ij} = \frac{\exp(\mathbf{x}_{ij}^T \beta)}{\sum_j \exp(\mathbf{x}_{ij}^T \beta)}. \quad (6.7)$$

The equivalence described above between the log-linear model (6.6) and the multinomial response model (6.7) depends heavily on the assumption that the parameters τ_i and hence ϕ_i are unrestricted apart from the necessary inequalities $\tau_i \geq 0$. In particular, the log-linear model (6.6) has the property that the row totals convey no information concerning β . This property makes good sense in the context of multinomial response models because the row totals by themselves cannot provide any information concerning the ratios of the components.

To take a specific example, consider the lizard data analysed in section 4.6. In that section, species was regarded as the binary response and the remaining factors, H , D , S , and T , were regarded as explanatory. For each combination of H , D , S and T , we conditioned on the total number of lizards observed, treating the proportion of *opalinus* lizards as the response. The linear logistic model with R (= species) as response and containing the main effects of H , D , S , and T is equivalent to the log-linear model with model formula

$$H.D.S.T + R.(H + D + S + T). \quad (6.8)$$

It is essential here that the full four-factor interaction $H.D.S.T$ be included even though, by the usual criteria of significance testing, a component of it might be formally statistically insignificant.

Note that the fitted values for the four-way margin $H.D.S.T$ in the log-linear model (6.8) are set equal to the observed values, which are the multinomial or binomial totals. Inclusion of this term in all log-linear models is essential in order that the log-linear models should correspond to the various binomial response models fitted in section 4.6. Note, however, that the inclusion of an arbitrary term or terms in a log-linear model is not equivalent to conditioning on the corresponding sets of marginal totals.

6.4.3 *Summary*

When the parameter of interest is the ratio of Poisson means or, equivalently, the value of a Poisson mean as a fraction of the total, it is usually appropriate to condition on the observed total. Conditioning on the total leads to multinomial or binomial response models of the log-linear type.

Not all log-linear models are equivalent to multinomial response models and, conversely, not all multinomial response models can be generated from log-linear models. For instance, the proportional-odds and related models discussed in section 5.2.2 cannot be derived by conditioning in a log-linear model without extending the accepted definition of a log-linear model.

The derivations given in the preceding sections show that, as far as parameter estimates and the matrix of second derivatives is concerned, it makes no difference numerically whether we condition on the row totals or not, provided that appropriate nuisance parameters are included in the log-linear model. In this respect conditioning appears almost optional, by contrast with Chapter 7 where conditioning affects the entire likelihood, the position of the maximum and the estimate of precision. Exact significance tests are possible only by conditioning on the required totals.

One important consequence of these results is that certain multinomial response models can be fitted using computer packages designed primarily for log-linear models. Such log-linear models invariably contain a large number of incidental parameters relating to the multinomial totals. Thus numerical methods that rely on solving systems of linear equations, where the number of equations

is equal to the number of parameters, may grind to a halt for numerical reasons. The alternative method of iterative proportional scaling (see e.g. Bishop *et al.*, 1975, p. 83) may be used instead. If computational facilities permit, however, it is best to fit the multinomial response model directly.

6.5 Multiple responses

6.5.1 *Introduction*

Suppose that several responses each having two or more categories are observed. In a pharmaceutical trial for instance, a drug is designed with a particular target effect in mind but invariably there are side-effects of varying duration and severity. By their nature, side-effects are difficult to predict but in simple cases might be classifiable according to severity and duration as shown below.

Table 6.5 *Classification of target response and supplementary responses in a pharmaceutical trial*

<i>Target effect</i>	<i>Side-effect</i>	
	<i>Severity</i>	<i>Duration</i>
complete cure	none	temporary
partial cure	mild	permanent
no improvement	moderate	
	severe	

In an experiment where several responses A, B, C, \dots are observed, the following lines of inquiry would often be considered worth pursuing.

1. Model construction for the dependence of each response marginally on covariates \mathbf{x} .
2. Model construction for the joint distribution of all responses.
3. Model construction for the joint dependence of all response variables on covariates \mathbf{x} .

For instance, in the pharmaceutical example it would be of interest to know whether the primary response was independent of the nature and severity of side-effects. Duration and severity as shown in Table 6.5 are not variation independent and hence cannot be

statistically independent. However it is possible that duration might be independent of severity conditionally on there being a detectable side-effect. A model for the joint distribution of all responses provides a description of the complete effect of the drug.

More realistically, however, pharmaceutical trials are usually designed as comparative experiments comparing the effects of two or more drugs. The aim then is to compare the joint response probabilities for one group of subjects with the corresponding probabilities for another group and to find a succinct description of any systematic differences. Problems of this nature are considered in section 6.5.4.

6.5.2 *Independence and conditional independence*

Suppose that we have a single sample of subjects and that several polytomous responses A, B, C, \dots are recorded for each subject. No external variables or covariates are available and we require a purely internal analysis of the joint dependence of the several responses.

Mutual independence of the three responses A, B, C corresponds to the log-linear model $A + B + C$ where A, B, C are factors having the requisite number of levels. The next simplest model, involving one interaction, namely $A*B+C$, means that the joint distribution of A and B is the same at each level of C . In other words, C is independent of A and B jointly. In subscript notation, $A*B+C$ corresponds to

$$\log \mu_{ijk} = (\alpha\beta)_{ij} + \gamma_k.$$

Estimability constraints are a convention and not part of the model. See section 3.5. In the above model it suffices to choose $\hat{\gamma}_1 = 0$.

Path models (Goodman, 1973) involve two or more interactions. For instance $A*B + B*C$ means that conditionally on B , A and C are independent. Note that if B is deleted from the model formula we are left with $A + C$ implying that A and C are independent at each level of B . This conditional independence model can be interpreted in terms of the causal path or chain

$$A \longrightarrow B \longrightarrow C$$

in which A influences B and B subsequently influences C but there is no direct link between A and C . In the context of time-series,

this phenomenon is also called the Markov property, meaning that the future and the past are conditionally independent given the present. In the present context where there is no fixed temporal sequence, the conditional independence model is equally consistent with the ‘time-reversed’ chain

$$C \longrightarrow B \longrightarrow A$$

and with the alternative diagram

$$A \longleftarrow B \longrightarrow C$$

in which B is depicted as the cause of both A and C . In other words, the direction of the hypothesized causal chain cannot be inferred from the model formula alone.

In order to test such a path model it is natural to test whether A has an effect on C above and beyond that transmitted via B . Thus we compare the fits of the models

$$A*B + B*C \quad \text{and} \quad A*B + B*C + C*A.$$

A significant reduction in deviance is evidence against conditional independence and hence evidence against the lineal path models $A \longrightarrow B \longrightarrow C$, $C \longrightarrow B \longrightarrow A$ and $A \longleftarrow B \longrightarrow C$.

In the theory of log-linear models an important distinction is drawn between models such as $A*B + B*C$ and $A*B + B*C*D + C*E$, which are interpretable in terms of conditional independence, and models such as $A*B + B*C + C*A$ and $A*B*C + B*D + C*D$, which are not. The former models are said to be decomposable (Haberman, 1974a) and have closed-form maximum-likelihood estimates for the parameters and fitted values. The latter models are not decomposable and no closed-form estimates exist for the maximum-likelihood estimates.

The definition and rationale for decomposability is concerned with the prohibition of cycles of the form

$$A \longrightarrow B \longrightarrow C \longrightarrow A$$

without the corresponding full interaction term $A.B.C$. To make this notion precise, we say that a model formula \mathcal{M} with response factors A, B, \dots is singular if either

1. there exists a subset of the response variables, A, B, C, D say, such that all the lower-order interactions are in \mathcal{M} but $A*B*C*D$ is not in \mathcal{M} or
2. there exists a closed loop, $ADBCA$ say, such that all adjacent pairs are in \mathcal{M} but none of the possible three-way interactions is in \mathcal{M} .

Such a set of response factors is said to constitute a singularity. For example,

$$\mathcal{M} = A*B*C + B*C*D + A*C*D$$

contains the singularities $ABDA$ and $ABCD$. With the aid of these definitions, a decomposable model formula may be defined as one that contains no singularities. Haberman (1974a) gives a recursive definition that is equivalent to the absence of singularities.

6.5.3 Canonical correlation models

In the log-linear framework there is an unfortunate gap between the model for independence of two responses, $A + B$, and the saturated model with interaction $A*B$. The former contains $k_A + k_B - 1$ parameters whereas the latter is saturated with $k_A k_B$ parameters, where k_A, k_B are the numbers of levels of A and B . It is natural to explore the intermediate ground where the nature of the interaction is described by a small number of parameters.

If scores s_1, s_2, \dots and t_1, t_2, \dots are available for the response categories of A and B respectively, we may consider the following models, which are intermediate between $A + B$ and $A*B$.

<i>Model formula</i>	<i>Algebraic expression</i>
$A + B + s.t$	$\alpha_i + \beta_j + \gamma(s_i t_j)$
$A + B + A.t$	$\alpha_i + \beta_j + \gamma_i t_j$
$A + B + A.t + B.s$	$\alpha_i + \beta_j + \gamma_i t_j + \delta_j s_i$

There is a close similarity here with the multinomial response models described in section 5.2.3. For further discussion of the use and interpretation of scores, see Agresti (1984, Chapter 5) or Goodman (1981, 1986).

In the absence of scores, there appears to be no log-linear model that is intermediate between the two extremes of complete independence and arbitrary dependence. However, if we are prepared to consider models not of the log-linear type, we may entertain the single-root canonical covariance model

$$\eta_{ij} = \log \mu_{ij} = \alpha_i + \beta_j + \rho \epsilon_i \delta_j, \quad (6.9)$$

where ϵ and δ are unknown unit vectors satisfying $\sum \epsilon_i = \sum \delta_j = 0$ and $\rho \geq 0$ is unknown. Note that the likelihood equation for ρ satisfies

$$\sum_{ij} \hat{\epsilon}_i \hat{\delta}_j y_{ij} = \sum_{ij} \hat{\epsilon}_i \hat{\delta}_j \hat{\mu}_{ij}.$$

The left-hand member of this equation is the sample estimate of $E(A_\epsilon B_\delta)$, where

$$A_\epsilon = \hat{\epsilon}_i \quad \text{if} \quad A = i$$

and similarly for B_δ . The right member is the fitted value of the same moment. Consequently, since $y_{i\cdot} = \hat{\mu}_{i\cdot}$ and $y_{\cdot j} = \hat{\mu}_{\cdot j}$, it follows that the fitted correlation between A_ϵ and B_δ is equal to the observed sample correlation. Further, this canonical correlation is independent of the estimability constraints imposed on ϵ and δ .

Model (6.9) is not of the generalized linear type and does not satisfy the usual regularity conditions because the sub-model of independence ($\rho = 0$) is a boundary point of the parameter space. The likelihood-ratio statistic for testing independence against (6.9) does not have an asymptotic χ^2 distribution. The correct asymptotic distribution is the distribution of the largest root of a certain Wishart matrix (Haberman, 1981).

Goodman (1986) refers to (6.9) as a log-bilinear model. Evidently, if ρ is small, we may approximate (6.9) by

$$\mu_{ij} = \alpha'_i \beta'_j \{1 + \rho \epsilon_i \delta_j\},$$

showing that, to this order of approximation, the array of fitted values has rank two. Under independence, the rank is one. The above approximation to (6.9) is in fact the leading term in the singular-value decomposition of the array μ_{ij} . Correspondence analysis is the term used to describe a body of multivariate statistical methods, mainly graphical, based on the first few singular values and vectors (Hill, 1974). For details, see Fisher (1958, sections 49.2–3), Williams (1952), Benzécri (1976), Greenacre (1984), Gilula and Haberman (1986) and the discussion paper of Goodman (1986).

6.5.4 Multivariate regression models

In practice, where several responses are of interest, it is good policy to examine the dependence of each response marginally on the covariates \mathbf{x} . In the pharmaceutical example, for instance, one would normally examine how the cure rate — the principal response — is affected by treatment and other incidental variables. Since the response in this case is polytomous with three ordered categories, the proportional odds model (5.1) is an obvious place to start. Subsequently a regression model for the side-effects reveals which covariates affect the severity of side-effects and in what direction. To complete the analysis it is necessary to examine interactions among the responses. These can be of substantial importance. For instance if the interaction is such that those who are cured of the disease are largely incapacitated by side-effects, the value of treatment would be greatly diminished.

More formally, if there are several responses A, B, C, \dots , we may proceed as follows. For any given value of the covariate \mathbf{x} , we may write $\pi_{ijk}(\mathbf{x})$ for the probability that $A = i, B = j, C = k$. The object then is to construct a model for the way in which changes in \mathbf{x} affect $\boldsymbol{\pi}$. This must be done bearing in mind that some of the responses may be nominal, others ordinal and others nested as discussed in section 5.2. Primary interest usually is focussed on the marginal dependence of each response on \mathbf{x} . Consequently we first make a non-singular linear transformation from π_{ijk} to new probabilities γ_{ijk} ,

$$\boldsymbol{\gamma} = \mathbf{L}\boldsymbol{\pi} \quad (6.10)$$

where \mathbf{L} is a matrix of zeros and ones only. For instance if there are three response factors, it would often be sensible to choose

$$\boldsymbol{\gamma} = \left(\begin{array}{c} \pi_{i..} \\ \pi_{..j} \\ \pi_{...k} \\ \pi_{ij.} \\ \pi_{i..k} \\ \pi_{...jk} \\ \pi_{ijk} \end{array} \right) \quad \begin{array}{l} \text{univariate marginal probabilities} \\ \text{bivariate marginal probabilities} \\ \text{trivariate marginal probabilities.} \end{array}$$

Thus in the $2 \times 2 \times 2$ case $\boldsymbol{\gamma}$ contains six univariate marginal probabilities, twelve bivariate marginal probabilities and eight trivariate marginal probabilities. There is substantial redundancy among

these 26 values: in fact there are only seven linearly independent probabilities but the redundancy is helpful to maintain symmetry in the notation.

If A , B and C were each ordinal we would replace γ with the cumulative univariate, bivariate and trivariate marginal probabilities, as well as the reverse-cumulative probabilities obtained by replacing \leq by $>$. Obvious adjustments must be made if the responses are of mixed types.

The second step in model construction is the formulation of the logarithmic contrasts of interest, namely

$$\boldsymbol{\eta} = \mathbf{C} \log \boldsymbol{\gamma}, \quad (6.11)$$

for an appropriately chosen contrast matrix \mathbf{C} . For example, in the $2 \times 2 \times 2$ example discussed earlier, we may take $\boldsymbol{\eta}$ to be the vector of logistic factorial contrasts, namely

$$\begin{aligned} \eta_a &= \log \pi_{1..} - \log \pi_{2..} \\ \eta_b &= \log \pi_{.1.} - \log \pi_{.2.} \\ \eta_c &= \log \pi_{..1} - \log \pi_{..2} \\ \eta_{ab} &= \log \pi_{11.} - \log \pi_{12.} - \log \pi_{21.} + \log \pi_{22.} \\ \eta_{ac} &= \log \pi_{1.1} - \log \pi_{1.2} - \log \pi_{2.1} + \log \pi_{2.2} \\ \eta_{bc} &= \log \pi_{.11} - \log \pi_{.12} - \log \pi_{.21} + \log \pi_{.22} \\ \eta_{abc} &= \log \pi_{111} - \log \pi_{121} - \log \pi_{211} + \log \pi_{221} \\ &\quad - \log \pi_{112} + \log \pi_{122} + \log \pi_{212} - \log \pi_{222} \end{aligned} \quad \left. \begin{array}{l} \text{univariate} \\ \text{contrasts} \\ \text{bivariate} \\ \text{contrasts} \\ \text{trivariate} \\ \text{contrast} \end{array} \right\}$$

It is important here that the multivariate logit link transformation from $\boldsymbol{\pi}$ to $\boldsymbol{\eta}$ be invertible.

Obvious adjustments are necessary if some of the responses are polytomous. The nature of the adjustment depends on whether the response categories are ordinal or nominal.

Having defined these logarithmic or logistic factorial contrasts, model construction is quite straightforward provided that the factorial nature of the response contrasts is recognized. Perhaps the simplest non-trivial model in this class is as follows:

$$\eta_a(\mathbf{x}) = \boldsymbol{\beta}_a^T \mathbf{x}, \quad \eta_b(\mathbf{x}) = \boldsymbol{\beta}_b^T \mathbf{x}, \quad \eta_c(\mathbf{x}) = \boldsymbol{\beta}_c^T \mathbf{x} \quad (6.12)$$

$$\eta_{ab}(\mathbf{x}) = \eta_{ac}(\mathbf{x}) = \eta_{bc}(\mathbf{x}) = \eta_{abc}(\mathbf{x}) = 0. \quad (6.13)$$

This model asserts that each response has a linear logistic regression on \mathbf{x} and that the three responses are mutually independent.

Obvious extensions are obtained by retaining the marginal regression models (6.12) and replacing (6.13) by

$$\begin{aligned}\eta_{ab}(\mathbf{x}) &= \eta_{ab} & \eta_{ac}(\mathbf{x}) &= \eta_{ac} & \eta_{bc}(\mathbf{x}) &= \eta_{bc} \\ \eta_{abc}(\mathbf{x}) &= \eta_{abc}.\end{aligned}\quad (6.14)$$

The latter model asserts that the interactions among the responses are independent of \mathbf{x} . One could fit a model in which $\eta_{abc} = 0$, but the model formula for the responses is then not decomposable and there is perhaps something objectionable about this.

One would normally include in the $\eta_{ab}(\mathbf{x})$ regression model only those covariates common to the $\eta_a(\mathbf{x})$ regression model and the $\eta_b(\mathbf{x})$ regression model. By extension, the $\eta_{abc}(\mathbf{x})$ regression model should include only those covariates common to the $\eta_{ab}(\mathbf{x})$, $\eta_{ac}(\mathbf{x})$ and $\eta_{bc}(\mathbf{x})$ models. In (6.12) it appears that the same set of covariates \mathbf{x} has been included in each of the marginal regression models. This choice may often be reasonable but it is not necessary and there may well be circumstances in which it is reasonable to exclude a particular covariate from one marginal regression model and include it in another.

In the case of a bivariate ordinal response it is most natural to define the logarithmic contrasts as follows

$$\eta_{ai} = \text{logit } \gamma_{i\cdot} = \text{logit } \text{pr}(A \leq i)$$

$$\eta_{bj} = \text{logit } \gamma_{\cdot j} = \text{logit } \text{pr}(B \leq j)$$

$$\eta_{abij} = \log \gamma_{ij} - \log(\gamma_{i\cdot} - \gamma_{ij}) - \log(\gamma_{\cdot j} - \gamma_{ij}) + \log \bar{\gamma}_{ij}$$

where

$$\gamma_{i\cdot} = \text{pr}(A \leq i), \quad \gamma_{ij} = \text{pr}(A \leq i, B \leq j),$$

$$\gamma_{\cdot j} = \text{pr}(B \leq j), \quad \bar{\gamma}_{ij} = \text{pr}(A > i, B > j).$$

Parallel linear regression models may be used for the marginal logits along the lines of (5.1). In the case of the interaction logits the following linear models are among the options available

$$\eta_{abij} = 0, \quad \eta_{abij} = \theta, \quad \eta_{abij} = \theta_i, \quad \eta_{abij} = \theta_i + \phi_j,$$

although dependence on covariates is also possible.

The Pearson-Plackett family of distributions for a single bivariate response is a special case of the above corresponding to $\eta_{abij} = \eta_{ab}$, a constant for all i and j . For details see Pearson (1913), Plackett (1965), Wahrendorf (1980), Anscombe (1981), Chapter 12 and Dale (1984, 1986).

6.5.5 Multivariate model formulae

A multivariate regression model cannot ordinarily be specified by means of a single model formula. In the multivariate linear model, for example, it is usually necessary to specify a different set of covariates for each of the responses. Further, if it is required to model dispersion effects in addition to regression effects, as in (5.4), two model formulae are required, one for the regression effects and one for the dispersion effects. In the present context, where there are r response factors having k_1, \dots, k_r levels respectively, we have defined $k_1 \dots k_r - 1$ contrasts grouped into $2^r - 1$ factorial-contrast classes. Thus there are $k_1 - 1$ main-effect contrasts for factor A , $k_2 - 1$ for factor B , $(k_1 - 1)(k_3 - 1)$ for the interaction of A with C , and so on. In general, therefore, $2^r - 1$ model formulae are required, one for each factorial-contrast class. Of course, many of these model formulae may be null or empty and these need not be stated explicitly.

The complete specification of a multivariate regression model comprises a list of each response contrast followed by the required model formula. For example model (6.12), (6.13) becomes

$$A: \mathbf{x}; \quad B: \mathbf{x}; \quad C: \mathbf{x}.$$

It is possible and sometimes desirable to use abbreviations such as

$$(A; B; C): \mathbf{x}.$$

By obvious extension, the model (6.12), (6.14) may be abbreviated to

$$(A; B; C): \mathbf{x}; \quad (A * B * C): 1.$$

In this context, where $A * B * C$ precedes a colon, the letters represent response factor contrasts, and the expression is to be expanded using ; in place of +. Thus $(A * B * C): 1$ is the same as

$$(A; B; C; A.B; A.C; B.C; A.B.C): 1.$$

It is perfectly possible to have the same letter appear on both sides of a given colon. For example, if we have a bivariate ordinal response, the model

$$\eta_{a i} = \text{logit } \gamma_{i \cdot} = \theta_i + \beta_a x$$

$$\eta_{b j} = \text{logit } \gamma_{\cdot j} = \phi_j + \zeta_b z$$

$$\eta_{ab ij} = \eta_{ab}$$

corresponds to the model formulae

$$A: A + x; \quad B: B + B.z; \quad A.B: 1.$$

More generally, we may consider factorial models of the type

$$\begin{aligned} A: \mathbf{x}_A; \quad B: \mathbf{x}_B; \quad C: \mathbf{x}_C; \\ A*B: \mathbf{x}_{AB}; \quad A*C: \mathbf{x}_{AC}; \quad B*C: \mathbf{x}_{BC}; \quad A*B*C: \mathbf{x}_{ABC}, \end{aligned}$$

where $\mathbf{x}_A, \dots, \mathbf{x}_{AB}, \dots, \mathbf{x}_{ABC}$ are ordinary model formulae. Note that the model formula is unaffected by replacing \mathbf{x}_A by

$$\mathbf{x}_A + \mathbf{x}_{AB} + \mathbf{x}_{AC} + \mathbf{x}_{ABC}.$$

In other words this factorial model formula notation ensures that covariates that affect interaction contrasts are also included in the models for marginal responses.

Log-linear models for multivariate discrete responses are exceptional in the sense that they can be specified either by means of a single model formula or via the more cumbersome but more explicit notation just described. The single model formula is obtained by replacing all colons by asterisks and all semicolons by '+'.

6.5.6 Log-linear regression models

An important special case of the models considered in the previous section is obtained by taking $\mathbf{L} = \mathbf{I}$ in (6.10) and (6.11). The particular choice of contrast matrix \mathbf{C} is then not vitally important because \mathbf{C} is non-singular and can be absorbed into the model formula. The simplest choice, $\mathbf{C} = \mathbf{I}$, leads to log-linear models in which the log probabilities

$$\eta_{ijk}(\mathbf{x}) = \log \pi_{ijk}(\mathbf{x})$$

are expressed as linear functions in the covariates \mathbf{x} . For instance, if (A, B) is a bivariate response, the model formula

$$A*B + (A + B).\mathbf{x} \tag{6.15}$$

is equivalent to the algebraic expression

$$\log \pi_{ij}(\mathbf{x}) = (\alpha\beta)_{ij} + \boldsymbol{\alpha}_i^T \mathbf{x} + \boldsymbol{\beta}_j^T \mathbf{x}. \tag{6.16}$$

The same model can be specified by taking \mathbf{C} to be the usual matrix of factorial contrasts. Thus in the bivariate binary case, we have

$$\begin{aligned}\eta_a^* &= \log \pi_{11} + \log \pi_{12} - \log \pi_{21} - \log \pi_{22} \\ \eta_b^* &= \log \pi_{11} - \log \pi_{12} + \log \pi_{21} - \log \pi_{22} \\ \eta_{ab}^* &= \log \pi_{11} - \log \pi_{12} - \log \pi_{21} + \log \pi_{22}\end{aligned}$$

Since the transformation from π_{ij} or $\log \pi_{ij}$ to $\boldsymbol{\eta}^*$ is a transformation from factor levels to factor contrasts, model formula (6.15) now implies that

$$\begin{aligned}\eta_a^*(\mathbf{x}) &= \alpha_0^* + \boldsymbol{\alpha}^{*T} \mathbf{x} \\ \eta_b^*(\mathbf{x}) &= \beta_0^* + \boldsymbol{\beta}^{*T} \mathbf{x} \\ \eta_{ab}^*(\mathbf{x}) &= (\alpha\beta)^*\end{aligned}\tag{6.17}$$

where the starred parameters are the factorial contrasts of the unmarked parameters in (6.16). For instance,

$$\begin{aligned}\alpha_0^* &= (\alpha\beta)_{11} + (\alpha\beta)_{12} - (\alpha\beta)_{21} - (\alpha\beta)_{22} \\ \boldsymbol{\beta}^* &= 2\beta_1 - 2\beta_2 \\ (\alpha\beta)^* &= (\alpha\beta)_{11} - (\alpha\beta)_{12} - (\alpha\beta)_{21} + (\alpha\beta)_{22}.\end{aligned}\tag{6.18}$$

It should be emphasized here that (6.16) and (6.17) are entirely equivalent ways of expressing the same model through the model formula (6.15). Both expressions produce the same fitted values and the same deviance. The coefficients are related through the factorial contrast matrix \mathbf{C} as shown above.

This discussion should be contrasted with the interpretation of the same model formula in the context of the bivariate logit transformation following (6.11) for a bivariate binary response. In that context the interpretation of (6.15) is

$$\begin{aligned}\eta_a &= \text{logit } \pi_{1.}(\mathbf{x}) = \alpha_0 + \boldsymbol{\alpha}^T \mathbf{x} \\ \eta_b &= \text{logit } \pi_{.1}(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x} \\ \eta_{ab} &= \log \left\{ \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}(\mathbf{x}) \right\} = (\alpha\beta)\end{aligned}\tag{6.19}$$

The key difference here is that (6.12), (6.13) expresses the logistic factorial contrasts of the response probabilities linearly in \mathbf{x} ,

whereas (6.16), and more explicitly (6.17), may be viewed as a linear model for the factorial logarithmic contrasts. In the former case logistic contrasts are defined in terms of marginal probabilities. In the latter case factorial contrasts are defined in terms of log probabilities. The order in which these two operations, namely marginalization and transformation, is performed is the essence of the distinction.

Unfortunately the log-linear model (6.16) is incompatible with the bivariate logit model (6.19) in the sense that (6.16) implies that $\text{logit } \pi_{1.}(\mathbf{x})$ is non-linear in \mathbf{x} . In other words, apart from a few exceptional cases, the maximum-likelihood fitted values under (6.16) are different from those under (6.19). In addition, the regression parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$ appearing in (6.16) have a different interpretation from those in (6.19). In the bivariate case, however, $(\alpha\beta)$ is the same in both models although the estimates may differ: the final equations in (6.17) and (6.19) are identical.

6.5.7 Likelihood equations

It is instructive at this stage to derive the likelihood equations for multivariate logit regression models in which the composite link transformation $\boldsymbol{\eta}$ in (6.11) is linearly related to known covariates. To keep matters simple we consider only the bivariate binary case in which $\boldsymbol{\eta}$ has components $(\eta_a, \eta_b, \eta_{ab})$ as defined in the equation following (6.11). Since there are four response probabilities, it is sometimes convenient to complete the transformation by defining $\eta_0 = \log \pi_{..}$ as the leading component of $\boldsymbol{\eta}$: this device helps to maintain symmetry in the notation.

With these conventions, we find that the derivative matrix of $\boldsymbol{\eta}$ with respect to the components of $\boldsymbol{\pi}$ is

$$\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\pi}} = \begin{pmatrix} \pi_{11} & \pi_{12} & \pi_{21} & \pi_{22} \\ \eta_0 & \eta_a & \eta_b & \eta_{ab} \\ \begin{matrix} 1 \\ \pi_{1.}^{-1} \\ \pi_{.1}^{-1} \\ \pi_{11}^{-1} \end{matrix} & \begin{matrix} 1 \\ \pi_{1.}^{-1} \\ -\pi_{.2}^{-1} \\ -\pi_{12}^{-1} \end{matrix} & \begin{matrix} 1 \\ -\pi_{2.}^{-1} \\ \pi_{.1}^{-1} \\ -\pi_{21}^{-1} \end{matrix} & \begin{matrix} 1 \\ -\pi_{2.}^{-1} \\ -\pi_{.2}^{-1} \\ \pi_{22}^{-1} \end{matrix} \end{pmatrix}$$

For our present purposes it is the inverse matrix, $\partial \boldsymbol{\pi} / \partial \boldsymbol{\eta}$, that is most directly useful. Fortunately the inverse can be obtained

without much difficulty, particularly if a suitable computerized algebra system is readily available. In this instance we find after some simplification that the inverse matrix is

$$\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta}} = \begin{pmatrix} \eta_0 & \eta_a & \eta_b & \eta_{ab} \\ \pi_{11} & \begin{pmatrix} \pi_{11}\pi_{21} \\ \pi_{1.}\Delta \end{pmatrix} & \begin{pmatrix} \pi_{11}\pi_{12} \\ \pi_{1.}\Delta \end{pmatrix} & V_{ab} \\ \pi_{12} & \begin{pmatrix} \pi_{12}\pi_{22} \\ \pi_{.2}\Delta \end{pmatrix} & -\begin{pmatrix} \pi_{11}\pi_{12} \\ \pi_{1.}\Delta \end{pmatrix} & -V_{ab} \\ \pi_{21} & \begin{pmatrix} \pi_{21}\pi_{21} \\ \pi_{1.}\Delta \end{pmatrix} & \begin{pmatrix} \pi_{21}\pi_{22} \\ \pi_{2.}\Delta \end{pmatrix} & -V_{ab} \\ \pi_{22} & \begin{pmatrix} \pi_{22}\pi_{22} \\ \pi_{.2}\Delta \end{pmatrix} & -\begin{pmatrix} \pi_{21}\pi_{22} \\ \pi_{2.}\Delta \end{pmatrix} & V_{ab} \end{pmatrix}$$

In the above matrix we have used the following quantities for future convenience:

$$\begin{aligned} V_a &= \pi_{1.}\pi_{2..}, & V_b &= \pi_{.1}\pi_{.2..}, \\ V_{ab} &= \left(\frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}} \right)^{-1}, \\ \Delta &= \pi_{11}\pi_{12}\pi_{21}\pi_{22}/(V_a V_b V_{ab}) \end{aligned}$$

and $-1/(V_a V_b V_{ab})$ is the determinant of $\partial \boldsymbol{\eta} / \partial \boldsymbol{\pi}$. Note that V_a is the ‘harmonic total’ of the marginal row probabilities in the sense that

$$V_a = \left(\frac{1}{\pi_{1.}} + \frac{1}{\pi_{2.}} \right)^{-1}.$$

Similarly for V_b and V_{ab} , justifying the notation.

Under independence, but not otherwise, we have

$$V_{ab} = V_a V_b = \pi_{1.}\pi_{2.}\pi_{.1}\pi_{.2} \quad \text{and} \quad \Delta = 1.$$

In general, $0 \leq V_a, V_b \leq 1/4$; $0 \leq V_{ab} \leq 1/16$, although it is possible for V_{ab} to exceed $V_a V_b$. Further, for all $\boldsymbol{\pi}$, $\Delta \leq 1$, with equality only under independence.

The contribution to the log-likelihood function from a single bivariate response is

$$l = \sum_{(ij)} y_{ij} \log \pi_{ij}$$

where $\sum_{(ij)}$ denotes summation over the four response categories. The contribution to the log-likelihood derivative is

$$\frac{\partial l}{\partial \beta} = \sum_{(rs)} \sum_{(ij)} \frac{y_{ij} - m\pi_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial \eta_{rs}} \frac{\partial \eta_{rs}}{\partial \beta}. \quad (6.20)$$

The indices r, s refer to contrasts rather than factor levels. Thus r and s take the values 0, a and 0, b respectively and η_{a0} is understood to be the same as η_a .

For instance, if we take the particular model (6.12), (6.14) and focus on the parameter β_a for the marginal regression of A on \mathbf{x}_a , the contribution to the log-likelihood derivative is

$$\begin{aligned} & \sum_{(ij)} \frac{y_{ij} - m\pi_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial \eta_a} \mathbf{x}_a \\ &= \left(\frac{\pi_{21}}{\pi_{.1}} \epsilon_{11} + \frac{\pi_{22}}{\pi_{.2}} \epsilon_{12} - \frac{\pi_{11}}{\pi_{.1}} \epsilon_{21} - \frac{\pi_{12}}{\pi_{.2}} \epsilon_{22} \right) \frac{\mathbf{x}_a}{\Delta} \\ &= \left(\epsilon_{1.} - \left(\frac{\pi_{11}}{\pi_{.1}} - \frac{\pi_{12}}{\pi_{.2}} \right) \epsilon_{.1} \right) \frac{\mathbf{x}_a}{\Delta} \end{aligned} \quad (6.21)$$

where $\epsilon_{ij} = y_{ij} - m\pi_{ij}$. The second line above comes from the second column of the matrix $\partial\boldsymbol{\pi}/\partial\boldsymbol{\eta}$. The interesting point here is that the derivatives with respect to β_a and β_b depend only on the marginal totals, $y_{i.}$ and $y_{.j}$, and not on the joint composition of the two responses. Thus if the odds ratios for each response are given constants, the marginal totals are sufficient for (β_a, β_b) . Note that although $\epsilon_{..} \equiv 0$, $\epsilon_{1.}$ and $\epsilon_{.1}$ are not necessarily zero.

In the case of the parameter η_{ab} in (6.14), the contribution to the log-likelihood derivative is

$$\begin{aligned} & \sum_{(ij)} \frac{y_{ij} - m\pi_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial \eta_{ab}} 1 \\ &= V_{ab} \left(\frac{\epsilon_{11}}{\pi_{11}} - \frac{\epsilon_{12}}{\pi_{12}} - \frac{\epsilon_{21}}{\pi_{21}} + \frac{\epsilon_{22}}{\pi_{22}} \right). \end{aligned} \quad (6.22)$$

The overall likelihood equations are obtained by summing contributions such as (6.20)–(6.22) over all responses and equating the sum to zero.

Note that the likelihood equation for the marginal regression coefficient β_a , based on the marginal variable A alone, is not the

same as (6.21), which is based on the bivariate response (A, B) . Straightforward linear logistic regression based on the marginal variable $(Y_{1.}, Y_{2.})$, where $Y_{1.} \sim B(m, \pi_{1.})$ gives the following contribution to the log-likelihood derivative in place of (6.21):

$$(y_{1.} - m\pi_{1.})\mathbf{x}_a = \epsilon_{1.}\mathbf{x}_a. \quad (6.23)$$

Under independence of A and B , this is the same as (6.21), but otherwise the two contributions are not the same and the estimated coefficients are different.

Reverting now to the bivariate logit regression model, the Fisher information for $(\eta_0, \eta_a, \eta_b, \eta_{ab})$, again based on a single bivariate response, is

$$\begin{aligned} \mathbf{i}_\eta &= m \left(\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta}} \right)^T \text{diag}(1/\boldsymbol{\pi}) \left(\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta}} \right) \\ &= m \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & V_a/\Delta & \Delta_\pi/\Delta & 0 \\ 0 & \Delta_\pi/\Delta & V_b/\Delta & 0 \\ 0 & 0 & 0 & V_{ab} \end{pmatrix}. \end{aligned}$$

The determinant $\Delta_\pi = \pi_{12}\pi_{21} - \pi_{11}\pi_{22}$ is a measure of departure from independence for the particular bivariate response under consideration. When several bivariate responses are involved, as in a regression context, the quantities V_a , V_b , V_{ab} , Δ and Δ_π are functions of the fitted response probabilities and normally vary from one response to the next. Thus the complete Fisher information matrix for the regression coefficients $(\beta_a, \beta_b, \beta_{ab})$ in the model

$$\eta_a(\mathbf{x}) = \boldsymbol{\beta}_a^T \mathbf{x}, \quad \eta_b(\mathbf{x}) = \boldsymbol{\beta}_b^T \mathbf{x}, \quad \eta_{ab}(\mathbf{x}) = \boldsymbol{\beta}_{ab}^T \mathbf{x}$$

is as follows

$$\mathbf{I}_\beta = \begin{pmatrix} \mathbf{X}^T \mathbf{D}_1 \mathbf{X} & \mathbf{X}^T \mathbf{D}_{12} \mathbf{X} & 0 \\ \mathbf{X}^T \mathbf{D}_{12} \mathbf{X} & \mathbf{X}^T \mathbf{D}_2 \mathbf{X} & 0 \\ 0 & 0 & \mathbf{X}^T \mathbf{D}_3 \mathbf{X} \end{pmatrix} \quad (6.24)$$

where \mathbf{D}_1, \dots are diagonal matrices given by $\mathbf{D}_1 = \text{diag}\{mV_a/\Delta\}$, $\mathbf{D}_2 = \text{diag}\{mV_b/\Delta\}$, $\mathbf{D}_{12} = \text{diag}\{m\Delta_\pi/\Delta\}$ and $\mathbf{D}_3 = \text{diag}\{mV_{ab}\}$.

Note that β_{ab} is orthogonal to the marginal regression parameters and $\hat{\beta}_{ab}$ has asymptotic covariance matrix $(\mathbf{X}^T \mathbf{D}_3 \mathbf{X})^{-1}$ independently of $(\hat{\beta}_a, \hat{\beta}_b)$. This conclusion is correct even if the covariates included in the $\eta_{ab}(\mathbf{x})$ regression model are different from those in the marginal regression models.

Unfortunately it is not possible to invert the Fisher information matrix \mathbf{I}_β algebraically to obtain the asymptotic covariance matrix of $(\hat{\beta}_a, \hat{\beta}_b)$ explicitly. However, it is clear from other considerations that the covariance matrix of $\hat{\beta}_a$ is smaller in the usual matrix sense than the covariance matrix

$$(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}, \quad \text{where } \mathbf{W} = \text{diag}\{m\pi_1, \pi_2.\},$$

derived from the marginal likelihood (6.23) of the response A alone. In the example analysed in the following section, the apparent gain in efficiency is about 3–4%. For longitudinal data in which the same response is observed at different points in time, it may be appropriate to take $\alpha = \beta$ in (6.9). The gain in efficiency from using the full likelihood might then be substantial.

6.6 Example

6.6.1 Respiratory ailments of coalminers

In 1970 Ashford and Sowden published the data shown in Table 6.6, which concerns two respiratory ailments of working coalminers who were smokers without radiological evidence of pneumoconiosis, aged between 20 and 64 at the time of examination. On the basis of a short questionnaire, each respondent was classified as suffering from breathlessness (A), and wheeze (B). In this instance each response factor has two levels and all four combinations are possible. One aim of the investigation was to study how breathlessness and wheeze and their interaction are related to age.

Before proceeding to fit models to these data, it is essential to clarify a number of points concerning the study design and the selection of respondents. First, the study involves only smokers. Second, the study is restricted to those smokers without radiological evidence of pneumoconiosis. Third, the study is restricted to working miners at a ‘representative sample’ of UK collieries. Miners who had retired for health or other reasons are entirely

excluded: miners who were on sick leave at the time of the study are also apparently excluded. Any conclusions drawn from these data must necessarily apply only to that population of coalminers from which this particular group of miners could be considered a random sample. The selection pressures in this example are very strong and are likely to have a substantial effect on the apparent rate at which miners contract these respiratory ailments. For instance, miners who are incapacitated by shortness of breath are excluded if they are no longer active workers. It is difficult to see what useful epidemiological conclusions could be drawn from data such as these where selection pressures inevitably have an appreciable effect on the regression coefficients. These and related points are discussed by Mantel and Brown (1973).

Despite these very important reservations, we shall use the data to illustrate the techniques described in the previous section.

Table 6.6 Coalminers who are smokers without radiological pneumococcosis, classified by age, breathlessness and wheeze

Age-group in years	Breathlessness		No breathlessness		Total
	Wheeze	No wheeze	Wheeze	No wheeze	
20-24	9	7	95	1841	1952
25-29	23	9	105	1654	1791
30-34	54	19	177	1863	2113
35-39	121	48	257	2357	2783
40-44	169	54	273	1778	2274
45-49	269	88	324	1712	2393
50-54	404	117	245	1324	2090
55-59	406	152	225	967	1750
60-64	372	106	132	526	1136

An initial plot of the empirical logistic transformation,

$$Z_a = \log\{(y_{1.} + \frac{1}{2})/(y_{2.} + \frac{1}{2})\},$$

for breathlessness against age shows a strong monotone increasing relationship with the suggestion of a slight quadratic component. The corresponding plot for wheeze is very similar, and again there is a suggestion of a small quadratic component. Similarly, a plot of the empirical odds-ratio

$$Z_{ab} = \log\left\{\frac{(y_{11} + \frac{1}{2})(y_{22} + \frac{1}{2})}{(y_{12} + \frac{1}{2})(y_{21} + \frac{1}{2})}\right\}$$

Table 6.8 Fitted values for the bivariate logistic model $A*B:x$

Age-group in years	Breathlessness		No breathlessness	
	Wheeze	No wheeze	Wheeze	No wheeze
20–24	16.559	9.049	96.446	1829.946
25–29	26.467	12.493	113.972	1638.068
30–34	53.602	22.179	169.100	1868.118
35–39	119.021	44.010	271.298	2348.671
40–44	160.134	54.257	258.869	1800.740
45–49	268.822	86.072	301.303	1736.803
50–54	359.096	112.363	272.491	1346.050
55–59	436.191	137.156	219.814	956.839
60–64	386.823	123.321	128.511	497.345

against age is approximately linear, but decreasing. For these reasons, we focus our attention primarily on the bivariate logistic models in which the transformed parameters satisfy

$$\begin{aligned}\eta_a &= \beta_0^{(a)} + \beta_1^{(a)}x + \beta_2^{(a)}z, \\ \eta_b &= \beta_0^{(b)} + \beta_1^{(b)}x + \beta_2^{(b)}z, \\ \eta_{ab} &= \beta_0^{(ab)} + \beta_1^{(ab)}x + \beta_2^{(ab)}z,\end{aligned}\tag{6.25}$$

where $x = (\text{age} - 42)/5$ and $z = x^2$. The graphical evidence suggests that all three linear coefficients should be large and statistically highly significant. In what follows, this aspect is taken for granted and we focus on testing whether the quadratic coefficients are significant.

Table 6.7 Parameter estimates and standard errors for the bivariate logistic model $A*B:x^\dagger$, using marginal likelihoods and the joint likelihood

Parameter	Marginal likelihood		Joint likelihood	
	Estimate	SE	Estimate	SE
$A: 1$	-2.2597	0.0301	-2.2625	0.0299
$A: x$	0.5125	0.0123	0.5145	0.0121
$B: 1$	-1.4875	0.0206	-1.4878	0.0206
$B: x$	0.3259	0.0089	0.3254	0.0089
$A.B: 1$	3.0230	0.0715	3.0219	0.0697
$A.B: x$	-0.1306	0.0295	-0.1314	0.0284

[†] $x = (\text{age} - 42)/5$.

Parameter estimates and fitted values for the bivariate logistic model $A*B:x$, in which quadratic terms are omitted, are shown in Tables 6.7 and 6.8 respectively. The model formula notation used in Table 6.7 is such that $A.B:x$ is the same as the coefficient $\beta_1^{(ab)}$ in (6.25). By way of comparison, the estimates and standard errors obtained from the marginal logistic regressions for A and B separately are also shown in Table 6.7. The 'marginal-likelihood' estimates for $A.B:1$ and $A.B:x$ were obtained using an unconditional version of the iterative procedure described in section 7.4, but taking the fitted margins from the marginal logistic regressions of A on x and B on x .

The estimates obtained from the joint likelihood are slightly different and apparently slightly more efficient than the estimates obtained from the separate marginal regressions. The increase in efficiency ranges from 0% to 7.5% and averages out to about 3%. If, however, the efficiency calculations are made using the fitted values from the joint likelihood, the maximum gain in efficiency is 3.6%, a truly worthless gain in view of the effort expended!

Table 6.9 *Residual deviances for selected models fitted to the breathlessness/wheeze data in Table 6.6.*

Model formula	Link function		
	Bivariate logit (η)	Log-linear (η^*)	d.f.
$A*B:x$	30.39	41.46	21
$A*B:x; (A+B):z^\dagger$	17.12	18.04	19
$A*B:x+z$	16.96	17.66	18
$A*B:x; (A+B):R^\ddagger$	6.80	6.80	7

[†] $x = (\text{age} - 42)/5$; $z = x^2$:

[‡] R , (= row), treats age as a 9-level factor.

Table 6.9 shows the deviance for the linear model $A*B:x$ and for selected quadratic models. Both quadratic coefficients $\beta_2^{(a)}$ and $\beta_2^{(b)}$ are significant as can be seen by examining the coefficients and their standard errors in the model $A*B:x+(A+B):z$. Inclusion of both quadratic terms has the effect of reducing the residual deviance from 30.41 on 21 degrees of freedom to 17.12 on 19. Despite the overwhelming statistical significance, the quadratic coefficients are numerically very small and would normally have very little effect on the conclusions to be drawn. There is no evidence of a quadratic

effect for the log odds-ratio.

By way of contrast, Table 6.9 also gives the deviances for the corresponding log-linear models. Evidently the choice of link does affect the fitted values for the first three model formulae considered. In fact, the fitted values for the log-linear model $A*B*x$, given by Mantel and Brown (1973), Table 3b, are quite different from those in Table 6.8. In neither case, however, is there any suggestion that the log odds-ratio is non-linear in age. The fitted values for the final model, in which age is treated as a 9-level qualitative factor, are the same for the two links considered. Note as usual that the residual degrees of freedom are related to the rank of the model formula and are independent of the choice of link function.

These data have been analysed by Ashford and Sowden (1970) who fitted a bivariate probit model with constant correlation and used linear regressions for the marginal probits. Subsequently Mantel and Brown (1973) fitted a variety of log-linear models and log-quadratic models, including the first and third in Table 6.9 under the 'Log-linear' column. Similar models were fitted by Grizzle (1971) using a non-iterative method. Mantel and Brown also discuss ways in which various selection pressures could lead to a declining odds-ratio.

6.6.2 Parameter interpretation

It is instructive at this stage to contrast the interpretation of the parameters $A: x$ and $B: x$ in the bivariate logistic model $A*B: x$ with those in the corresponding log-linear model. Under the bivariate logit model the fitted marginal logits are

$$\begin{aligned}\log(\hat{\pi}_{1.}/\hat{\pi}_{2.}) &= -2.261 + 0.515x \\ \log(\hat{\pi}_{.1}/\hat{\pi}_{.2}) &= -1.487 + 0.326x\end{aligned}$$

and the fitted odds ratio is

$$\log(\hat{\pi}_{11}\hat{\pi}_{22}/(\hat{\pi}_{12}\hat{\pi}_{21})) = 3.022 - 0.131x.$$

These coefficients are given in Table 6.7. Thus, for the population in question, the estimated odds of contracting breathlessness increases by a factor of $\exp(0.515) = 1.674$ per unit increase in x : this translates into an annual factor of $\exp(0.103) = 1.108$. Stated in another way, the odds increases exponentially at an annual rate of just under 11%. The corresponding estimated annual rate of

increase for the odds of contracting wheeze is 6.7%. The observed decline in odds-ratio with increasing age is a curious feature of these data that may be attributable to censoring. For a discussion of this point, see Mantel and Brown (1973), p. 653.

Consider now the corresponding log-linear model in the form

$$\log \pi_{ij}(x) = \alpha_{ij} + \beta_{ij}x.$$

The conditional logits for A given $B = 1, 2$ are

$$\begin{aligned}\log\{\pi_{11}/\pi_{21}\} &= \alpha_{11} - \alpha_{21} + (\beta_{11} - \beta_{21})x \\ \log\{\pi_{12}/\pi_{22}\} &= \alpha_{12} - \alpha_{22} + (\beta_{12} - \beta_{22})x.\end{aligned}\quad (6.26)$$

These are linear in x , though not parallel. The corresponding logits for B given $A = 1, 2$ are

$$\begin{aligned}\log\{\pi_{11}/\pi_{12}\} &= \alpha_{11} - \alpha_{12} + (\beta_{11} - \beta_{12})x \\ \log\{\pi_{21}/\pi_{22}\} &= \alpha_{21} - \alpha_{22} + (\beta_{21} - \beta_{22})x.\end{aligned}\quad (6.27)$$

In both cases the difference between conditional logits is

$$(\alpha_{11} - \alpha_{12} - \alpha_{21} + \alpha_{22}) + (\beta_{11} - \beta_{12} - \beta_{21} + \beta_{22})x. \quad (6.28)$$

The maximum-likelihood fitted values for equations (6.26)–(6.28) are

$$\begin{aligned}\text{logit pr}(A = 1 | B=1, x) &= -0.418 + 0.349x \\ \text{logit pr}(B = 1 | A=1, x) &= 1.051 + 0.034x \\ \text{log odds-ratio} &= 3.059 - 0.166x.\end{aligned}$$

Evidently, the fitted odds-ratios in the log-linear model are different from those in the bivariate logit model. The difference between the regression coefficients, $0.166 - 0.131$, corresponds to about 1.25 standard errors.

Note also that the fitted logistic regression coefficient of B on x in the log-linear model is 0.034 for $A = 1$ and 0.201 for $A = 2$. The marginal regression coefficient, at 0.326, is considerably larger than both conditional coefficients. The same effect occurs for A on x , though the difference is less striking.

On balance, where two responses are observed more-or-less simultaneously, it is hard to see why one would be interested in the conditional distributions of each given the values of the other or how these conditional distributions are affected by covariates. On the other hand, the marginal distributions are of interest however many responses are observed and the mere recording of an additional, and possibly irrelevant, response should not deflect the focus of investigation. In the present example, if an additional response, say frequency or severity of stomach problems, C , is observed, the parameters appearing in the trivariate log-linear model bear no simple relation to those in the bivariate model just fitted. In fact, the log-linear models $A*B*x$ and $A*B*C*x$ are mutually contradictory except in degenerate cases. On the other hand, the trivariate logistic model $A*B*C:x$ implies the bivariate logit model $A*B:x$ and the univariate logit model $A:x$. For this reason alone the multivariate logit models seem preferable to log-linear models for multiple responses that are to be treated symmetrically. This argument, if accepted, would seem to outweigh all considerations of goodness-of-fit as a basis for model choice.

6.7 Bibliographic notes

For the most part, the books listed at the end of Chapters 4 and 5 deal also with log-linear models. The books by Agresti (1984) Bishop, Fienberg and Holland (1975), Bock (1975), Fienberg (1980), Goodman (1978), Haberman (1974a), Plackett (1981), and Upton (1978) are especially relevant. Haberman gives a thorough mathematical treatment of log-linear models and also introduces the notion of decomposability as the condition for the existence of closed-form maximum likelihood estimates. Plackett's book contains a very extensive bibliography and a large number of numerical examples. For additional bibliographic material, the reader is referred to Killion and Zahn (1976).

The connection between decomposable models and the larger class of graphical models is discussed by Darroch, Lauritzen and Speed (1980).

Chapter 12 of Anscombe (1981) is refreshingly nonconformist in its treatment of models for contingency tables.

There is now an extensive but fragmented literature on multi-

plicative interaction models, not just for contingency tables, but for factorial designs in general. Mandel (1959, 1971) has considered the uses of multiplicative interaction models for Latin square and other designs. Correspondence analysis (Greenacre, 1984; Benzécri, 1976), uses similar techniques based on the singular-value decomposition for the analysis of contingency tables. For further discussion of this topic see Gilula and Haberman (1986).

Canonical correlation models of the type discussed in section 6.5.3 have been considered previously by Goodman (1979, 1981) and by Haberman (1981).

Cox (1972b) notes the drawback of log-linear models for multivariate binary responses, that the marginal logits are not simply related to the log-linear parameters. He proposes a list of alternatives to the log-linear model which, however, does not include the multivariate logit transformation in section 6.5.

The application of the multivariate logit link function to bivariate and multivariate responses has been studied by Dale (1986).

6.8 Further results and exercises 6

6.1 By writing $Y = \mu(1 + \epsilon)$ and expanding in a Taylor series as far as the fourth degree, show that

$$\begin{aligned} E(Y^{1/2}) &\simeq \mu^{1/2} \left\{ 1 - \frac{1}{8\mu} - \frac{7}{128\mu^2} + O(\mu^{-3}) \right\} \\ \text{var}(Y^{1/2}) &\simeq \frac{1}{4} \left\{ 1 + \frac{3}{8\mu} + O(\mu^{-2}) \right\} \\ \kappa_3(Y^{1/2}) &\simeq -\mu^{-1/2}/16 \left\{ 1 + O(\mu^{-1}) \right\} \end{aligned}$$

where $Y \sim P(\mu)$. Show also that

$$\begin{aligned} E(Y^{2/3}) &\simeq \mu^{2/3} \left\{ 1 - \frac{1}{9\mu} - \frac{1}{27\mu^2} + O(\mu^{-3}) \right\} \\ \text{var}(Y^{2/3}) &\simeq \frac{4\mu^{1/3}}{9} \left\{ 1 + \frac{1}{6\mu} + O(\mu^{-2}) \right\} \\ \kappa_3(Y^{2/3}) &\simeq -68/(729\mu) + O(\mu^{-2}) \end{aligned}$$

Comment briefly on the possible applications of these transformations.

6.2 By expanding in a Taylor series for small $\epsilon = (y - \mu)/\mu$, show that

$$Y \log(Y/\mu) - (Y - \mu) \simeq \mu\{\epsilon^2/2 - \epsilon^3/6 + \epsilon^4/12 - \epsilon^5/20 + \dots\}$$

whereas

$$\frac{9}{2}Y^{1/3}(\mu^{1/3} - Y^{1/3})^2 \simeq \mu\{\epsilon^2/2 - \epsilon^3/6 + 2\epsilon^4/27 - \epsilon^5/27 + \dots\}.$$

Hence, using the result given in Appendix C, show that for large μ ,

$$3Y^{1/6}(Y^{1/3} - \mu^{1/3}) + \mu^{-1/2}/6 \sim N(0, 1) + O_p(\mu^{-1}),$$

at least as far as moments are concerned.

6.3 Suppose conditionally on $Z = z$, that $Y \sim P(z)$ and that Z has the density function

$$f_Z(z; \mu, \phi) dz = \frac{(\phi z)^{\phi\mu} \exp(-\phi z)}{\Gamma(\phi\mu)} d \log z.$$

Show that the marginal distribution of Y is

$$\text{pr}(Y = y; \mu, \phi) = \frac{\Gamma(y + \phi\mu)\phi^{\phi\mu}}{y! \Gamma(\phi\mu)(1 + \phi)^{y + \phi\mu}} \quad y = 0, 1, 2, \dots.$$

Find the unconditional mean and variance of Y .

6.4 Fit the model

$$\text{site + class + volume + tuberculin}$$

to the data in Table 6.1b, treating site and class as four-level factors and tuberculin as a two-level factor. Treat volume as a quantitative variable taking values $-1, 0, 1$ for half, single and double respectively. Estimate the relative potency of the two preparations.

Now treat volume as a two-level factor with levels denoting low dose and high dose respectively. Leave the remaining factors as they stand. Show that the fitted values are identical to those produced by the previous analysis, but that the tuberculin contrast is now nearly zero. Explain why the tuberculin contrast is affected by the parameterization chosen for volume.

6.5 For the data in Table 6.1b, test the hypothesis that the effect on the response of doubling the volume administered is the same for each tuberculin. Use the method described towards the end of section 6.3.1. Compute an approximate p -value.

6.6 Show that for any 2×2 table of probabilities, the quantity Δ defined in section 6.5.6 satisfies

$$\Delta = S_3 / (S_3 + \Delta_\pi^2),$$

where S_3 is a particular symmetric function of the four probabilities. Find expressions for S_3 and Δ_π and deduce that $0 \leq \Delta \leq 1$, with equality only under independence.

6.7 Let (A, B) be a bivariate binary response and let $\eta_a, \eta_b, \eta_{ab}$ be defined as in sections 6.5.4 and 6.5.6. Consider the multivariate regression model

$$\eta_a(\mathbf{x}) = \boldsymbol{\beta}_a^T \mathbf{x}_a, \quad \eta_b(\mathbf{x}) = \boldsymbol{\beta}_b^T \mathbf{x}_b, \quad \eta_{ab}(\mathbf{x}) = \eta_{ab},$$

in which the model matrices $\mathbf{X}_a, \mathbf{X}_b$ each have rank n equal to the number of bivariate responses observed. By considering the log-likelihood derivative (6.21), show that the likelihood equations reduce to

$$y_{i\cdot} = m\hat{\pi}_{i\cdot}, \quad y_{\cdot j} = m\hat{\pi}_{\cdot j}, \quad \text{for each bivariate response,}$$

$$\sum_1^n (y_{11} - m\hat{\pi}_{11}) = 0.$$

The final sum extends over the $(1, 1)$ -components of all n responses.

6.8 Show that the inverse of the multivariate logit transformation $\boldsymbol{\eta} \rightarrow \boldsymbol{\pi}$, following (6.11) can be broken down into the following sequence of steps:

1. exponentiation;
2. iterative proportional scaling;
3. linear transformation, (\mathbf{L}^{-1}) .

Show how Yates's algorithm (McCullagh, 1987, p. 15) can be used in step 3. to exploit the direct product nature of \mathbf{L} .

Table 6.10 *Distribution of four binary responses in two groups*[†]

y_1	y_2	y_3	y_4	<i>Low I.Q. group</i>	<i>High I.Q. group</i>
1	1	1	1	62	122
1	1	1	0	70	68
1	1	0	1	31	33
1	1	0	0	41	25
1	0	1	1	283	329
1	0	1	0	253	247
1	0	0	1	200	172
1	0	0	0	305	217
0	1	1	1	14	20
0	1	1	0	11	10
0	1	0	1	11	11
0	1	0	0	14	9
0	0	1	0	31	56
0	0	1	0	46	55
0	0	0	1	37	64
0	0	0	0	82	53

[†]Source: Solomon (1961).

6.9 The data in Table 6.10, taken from Solomon (1961), lists the responses, (agree/disagree), given by 2982 New Jersey high-school seniors in a 1957 attitude survey, in response to the following four propositions:

1. The development of new ideas is the scientist's greatest source of satisfaction.
2. Scientists and engineers should be eliminated from the military draft.
3. The scientist will make his maximum contribution to society when he has freedom to work on problems that interest him.
4. The monetary compensation of a Nobel Prize-winner in physics should be at least equal to that given to popular entertainers.

Examine how each response marginally depends on the I.Q. group.

Examine the six bivariate distributions to see whether the odds-ratios are different in the two groups.

Give a brief summary of your conclusions in non-technical language.

6.10 Show that the redundancy in the transformation following (6.10) can be avoided by eliminating all components having an

index whose value is 1.

6.11 Show that for a bivariate response (A, B) , in which A is a nominal response with three levels and B is ordinal with four levels, the natural analogue of the bivariate logit transformation is

$$\begin{aligned}\eta_{ai} &= \log(\gamma_{i\cdot}/\gamma_{3\cdot}) \quad i = 1, 2 \\ \eta_{bj} &= \log\{\gamma_{\cdot j}/(1 - \gamma_{\cdot j})\} \quad j = 1, 2, 3 \\ \eta_{abij} &= \log\left(\frac{\gamma_{ij}}{\gamma_{i\cdot} - \gamma_{ij}}\right) - \log\left(\frac{\gamma_{3j}}{\gamma_{3\cdot} - \gamma_{3j}}\right) \\ &= \text{logit}(\gamma_{ij}/\gamma_{i\cdot}) - \text{logit}(\gamma_{3j}/\gamma_{3\cdot}).\end{aligned}$$

In these expressions γ_{ij} is defined as

$$\begin{aligned}\gamma_{ij} &= \text{pr}(A = i, B \leq j) \\ \gamma_{i\cdot} &= \text{pr}(A = i) \\ \gamma_{\cdot j} &= \text{pr}(B \leq j).\end{aligned}$$

6.12 The data listed in Table 6.11, taken from Diaconis (1988), give partial results for the 1980 American Psychological Association presidential election in which there were five candidates, here labelled A,B,C,D and E. For each of the 120 complete rankings of the five candidates, Table 6.11 gives the number of voters who cast their ballots in that way. Thus of the 5738 voters who cast complete ballots, the modal group of 186 voters cast their ballots as '23154' in which candidate A was placed second, candidate B third, candidate C first and so on. Incomplete ballots are not included here.

1. Create five factors, A, B, C, D, E , each with five levels, such that A at level 4 means that candidate A was ranked in fourth position, and so on for the remaining factors. Fit the log-linear models 1 and $A + B + C + D + E$. What is the rank of the latter model matrix. Explain why E can be dropped without affecting the fit.
2. Create linear and quadratic contrasts for each of the five factors such that B_L takes values $-2, -1, 0, 1, 2$ and B_Q takes values $2, -1, -2, -1, 2$ for the five levels of B . Compute the sum of the model vectors $A_L + B_L + C_L + D_L + E_L$. Fit the model

$$A_L + B_L + C_L + D_L + E_L.$$

Table 6.11 Number of voters in the 1980 APA presidential election ranking five candidates in the specified order[†]

Candidates' ranks and number of votes cast																							
Candidates		Candidates		Candidates		Candidates		Candidates		Candidates													
A	B	C	D	E	No.	A	B	C	D	E	No.	A	B	C	D	E	No.	A	B	C	D	E	No.
5	4	3	2	1	29	5	4	3	1	2	67	5	4	2	3	1	37	5	4	2	1	3	24
5	4	1	3	2	43	5	4	1	2	3	28	5	3	4	2	1	57	5	3	4	1	2	49
5	3	2	4	1	22	5	3	2	1	4	22	5	3	1	4	2	34	5	3	1	2	4	26
5	2	4	3	1	54	5	2	4	1	3	44	5	2	3	4	1	26	5	2	3	1	4	24
5	2	1	4	3	35	5	2	1	3	4	50	5	1	4	3	2	50	5	1	4	2	3	46
5	1	3	4	2	25	5	1	3	2	4	19	5	1	2	4	3	11	5	1	2	3	4	29
4	5	3	2	1	31	4	5	3	1	2	54	4	5	2	3	1	34	4	5	2	1	3	24
4	5	1	3	2	38	4	5	1	2	3	30	4	3	5	2	1	91	4	3	5	1	2	84
4	3	2	5	1	30	4	3	2	1	5	35	4	3	1	5	2	38	4	3	1	2	5	35
4	2	5	3	1	58	4	2	5	1	3	66	4	2	3	5	1	24	4	2	3	1	5	51
4	2	1	5	3	52	4	2	1	3	5	40	4	1	5	3	2	50	4	1	5	2	3	45
4	1	3	5	2	31	4	1	3	2	5	23	4	1	2	5	3	22	4	1	2	3	5	16
3	5	4	2	1	71	3	5	4	1	2	61	3	5	2	4	1	41	3	5	2	1	4	27
3	5	1	4	2	45	3	5	1	2	4	36	3	4	5	2	1	107	3	4	5	1	2	133
3	4	2	5	1	62	3	4	2	1	5	28	3	4	1	5	2	87	3	4	1	2	5	35
3	2	5	4	1	41	3	2	5	1	4	64	3	2	4	5	1	34	3	2	4	1	5	75
3	2	1	5	4	82	3	2	1	4	5	74	3	1	5	4	2	30	3	1	5	2	4	34
3	1	4	5	2	40	3	1	4	2	5	42	3	1	2	5	4	30	3	1	2	4	5	34
2	5	4	3	1	35	2	5	4	1	3	34	2	5	3	4	1	40	2	5	3	1	4	21
2	5	1	4	3	106	2	5	1	3	4	79	2	4	5	3	1	63	2	4	5	1	3	53
2	4	3	5	1	44	2	4	3	1	5	28	2	4	1	5	3	162	2	4	1	3	5	96
2	3	5	4	1	45	2	3	5	1	4	52	2	3	4	5	1	53	2	3	4	1	5	52
2	3	1	5	4	186	2	3	1	4	5	172	2	1	5	4	3	36	2	1	5	3	4	42
2	1	4	5	3	24	2	1	4	3	5	26	2	1	3	5	4	30	2	1	3	4	5	40
1	5	4	3	2	40	1	5	4	2	3	35	1	5	3	4	2	36	1	5	3	2	4	17
1	5	2	4	3	70	1	5	2	3	4	50	1	4	5	3	2	52	1	4	5	2	3	48
1	4	3	5	2	51	1	4	3	2	5	24	1	4	2	5	3	70	1	4	2	3	5	45
1	3	5	4	2	35	1	3	5	2	4	28	1	3	4	5	2	37	1	3	4	2	5	35
1	3	2	5	4	95	1	3	2	4	5	102	1	2	5	4	3	34	1	2	5	3	4	35
1	2	4	5	3	29	1	2	4	3	5	27	1	2	3	5	4	28	1	2	3	4	5	30

[†]Source: Diaconis (1988).

Which candidate has the smallest coefficient? Interpret the sizes of these coefficients.

3. Add the terms

$$A_Q + B_Q + C_Q + D_Q + E_Q$$

to the previous model. Which candidate has the largest quad-

ratio coefficient? Interpret the sizes of the quadratic coefficients in terms of heterogeneity among voters and negative voting. Examine the two-way table of total votes indexed by candidate and position. Compute the fitted values for this table under the quadratic model just fitted.

Show that E_L and E_Q can be dropped from the model without affecting the fit.

4. Create all 10 linear \times linear interaction contrasts of the type $A_L B_L = A_L \times B_L$. Add these to the previous model. Show that the fit is substantially improved, even though the residual deviance is still considerably larger than the degrees of freedom. Interpret the coefficients obtained.
5. What additional systematic effects might be present here? State these effects qualitatively. How would you incorporate these into your model?
6. Which candidate has the most first-place votes? Which candidate is least disliked? Which candidate ought to be declared the winner?
7. What modifications of this procedure would you use to analyse the incomplete ballots in which three or fewer candidates are ranked by some voters?
8. Negative voting can be accomplished effectively only with a complete ballot. What implications does this have for comparisons of complete ballots with incomplete ballots?

6.13 Let AB be a two-level factor taking the level 1 if A precedes B in the permutation, and level 2 otherwise. Nine other factors AC, \dots, DE are defined likewise. Thus the model matrix \mathbf{X} corresponding to the model formula

$$1 + AB + AC + AD + AE + BC + BD + BE + CD + CE + DE$$

is the incidence matrix for the set of inversions required to transform $\boldsymbol{\pi}$ to standard order. What is the rank of \mathbf{X} ? Fit this model to the data in Table 6.11.

By extension, let ABC be a six-level factor, one level for each of the possible orders of A, B, C in the permutation $\boldsymbol{\pi}$. Nine other factors ABD, \dots, CDE are defined in like manner. Show that AB , AC and BC are marginal to ABC . What is the rank of the model

matrix \mathbf{Z} corresponding to the model formula

$$1 + ABC + ABD + ABE + ACD + ACE \\ + ADE + BCD + BCE + BDE + CDE?$$

Show that the column space of \mathbf{X} is included in the column space of \mathbf{Z} and that

$$\text{rank}(\mathbf{Z}) = 1 + \binom{k}{2} + 2\binom{k}{3},$$

where k is the number of candidates.

Fit the model \mathbf{Z} to the data in Table 6.11. Give a substantive explanation for the improvement in fit. Show that these models are unaffected by re-labelling candidates. [Babington-Smith, 1950; Mallows, 1957].

6.14 Use the data in Table 4.10 to test the hypothesis that mating occurs at random, at least as regards eye-colour. Take Y to be the 6×1 vector of counts for the various eye-colour combinations of the parents. Formulate the hypothesis of random mating as a log-linear model for Y as response. State what assumptions you have made and indicate whether these assumptions are reasonable in this context. Fit the model and use it to estimate the proportions of light-eyed, hazel-eyed and dark-eyed individuals in the population.

6.15 *The butler effect:* It can safely be assumed in Table 4.10 that a small proportion ϵ of the children are not the biological children of the putative fathers. Consider how you might estimate ϵ under the following assumptions:

1. If both biological parents are light-eyed the children are invariably light-eyed.
2. The distribution of eye-colour is the same for both sexes.
3. The population of butlers is comparable, at least as regards eye-colour, to the population of putative fathers.
4. There are no recording errors in the data.

Other assumptions that might be reasonable include the following: (a) If both biological parents are dark-eyed, the children are light-eyed with probability $1/4$. (b) If one parent is light-eyed and one dark-eyed the children are light-eyed with probability $1/2$. Consider how these additional assumptions might be used to improve the estimate of ϵ .

6.16 In his 1898 monograph L. von Bortkewitsch gives the now famous record of deaths by horse-kicks of soldiers in the Prussian army from 1875 to 1894. The data for $c = 14$ army corps over $r = 20$ years are given by Andrews and Herzberg (1985, pp.17-18). Fit the log-linear model $\text{corps} + \text{year}$. Is the fit adequate? The data are sparse, which suggests that the usual χ^2 approximation may not be accurate.

The Haldane-Dawson formulae (Haldane, 1939; Dawson, 1954) for the exact mean and variance of X^2 for the model of independence in a two-way table are

$$E(X^2) = (r - 1)(c - 1)N/(N - 1),$$

$$\text{var}(X^2) = 2N(\nu - \sigma)(\mu - \tau)/(N - 3) + N^2\sigma\tau/(N - 1),$$

where

$$\nu = (N - r)(r - 1)/(N - 1), \quad \sigma = N \left\{ \sum_i s_i^{-1} - r^2/N \right\} / (N - 2),$$

$$\mu = (N - c)(c - 1)/(N - 1), \quad \tau = N \left\{ \sum_j t_j^{-1} - c^2/N \right\} / (N - 2),$$

The row and column totals are s_i, t_j , and $N = \sum s_i = \sum t_j$.

Use these formulae for the horse-kick data to show that the mean and variance of X^2 are 248.27 and 419.81 respectively. What is the variance of the usual χ^2 approximation?

Is there any evidence of variation in the accident rate over years or between corps? For further details see Quine and Seneta (1987) or Preece, Ross and Kirby (1988).

CHAPTER 7

Conditional likelihoods*

7.1 Introduction

In many applications the likelihood function involves several parameters, only a few of which are of interest to the investigator. The remaining parameters, often pejoratively called incidental or nuisance parameters, are necessary in order that the model make sense physically, but their values are largely irrelevant to the experiment and to the conclusions that are to be drawn. To take a simple example, consider a comparative experiment with two treatment groups in which the response is ordinal with k categories. The proportional-odds model,

$$\log\{\gamma_{ij}/(1 - \gamma_{ij})\} = \theta_j - \beta x_i, \quad j = 1, \dots, k-1; i = 1, 2,$$

in which x_i is an indicator variable for treatment group, involves one parameter of interest, β , and $k-1$ nuisance parameters, $\theta_1, \dots, \theta_{k-1}$. There are applications in which both components $\boldsymbol{\theta}$ and β are of equal interest to the investigator, but usually, in comparative experiments, the focus is on change or rate of change of the response as the stimulus is increased. Thus, when we use the terms ‘nuisance parameter’ or ‘incidental parameter’ it is with certain common applications in mind. The status of a parameter depends on the context.

Two difficulties arise in dealing with likelihood functions that depend on a large number of incidental parameters in addition to the effects of interest. First, from a purely mathematical point of view, there is no guarantee of consistency or optimality in the limit as the number of parameters increases in proportion to the data

*This chapter contains more mathematical material and may be omitted on first reading.

accumulated. Whether this large-sample mathematical difficulty has any relevance in the finite samples actually observed is another matter, and will not be discussed here. The second difficulty is the purely numerical one of maximizing a function of many variables and of obtaining the inverse matrix of second derivatives, but this is a subsidiary consideration in view of the first difficulty. For these reasons, we seek a modified likelihood function that depends on as few of the incidental parameters as possible while, at the same time, sacrificing as little information as possible. Inferences are then based on this modified likelihood function, particularly on its shape in the vicinity of its maximum.

7.2 Marginal and conditional likelihoods

7.2.1 Marginal likelihood

One way of eliminating unwanted nuisance parameters is to work with the marginal likelihood for a suitably chosen subset of the complete data vector. This method does not always work satisfactorily, but when it does, it is clearly desirable to choose as large a subset of the original data as possible so that the information loss is minimized.

In the context of the bivariate logistic model (6.25), if $\beta^{(a)}$ is the parameter of interest and $(\beta^{(b)}, \beta^{(ab)})$ are nuisance parameters, we may eliminate the nuisance parameters by working with the log likelihood for the marginal variable A alone, *i.e.* with the marginal totals $(Y_{1.}, Y_{2.})$. Evidently, from the analysis in section 6.6.1, some small loss of information is thereby incurred, but this loss might well be judged acceptable in view of the simplicity achieved. In this example, the loss of efficiency is balanced by a gain in robustness, because the marginal likelihood estimates are consistent whether or not the assumed models for η_a and η_{ab} in (6.19) are correct, whereas the estimates derived from the full likelihood are not protected against such mis-specifications.

To take an unrelated example, suppose Y_1, \dots, Y_n are observations taken at spatial locations s_1, \dots, s_n , and that the n -vector \mathbf{Y} has the multivariate Normal distribution with cumulants

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}; \quad \text{cov}(\mathbf{Y}) = \boldsymbol{\Sigma}(\boldsymbol{\theta}),$$

where $\Sigma(\boldsymbol{\theta})$ is a known covariance function parameterized by $\boldsymbol{\theta}$. For example, we might have

$$\sigma_{ij}(\boldsymbol{\theta}) = \theta_1^2 \exp\{-|s_i - s_j|/\theta_2\},$$

where θ_2 has the physical dimension of length and θ_1 has the same physical dimension as y . If $\boldsymbol{\theta}$ is the parameter of interest and β is regarded as a nuisance parameter, we may eliminate β from the likelihood by working with the set of contrasts,

$$\mathbf{R} = (\mathbf{I} - \mathbf{P}_X)\mathbf{Y} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T)\mathbf{Y},$$

whose mean is zero and whose distribution does not depend on β . Any complete set of $n - p$ linearly independent contrasts with zero mean is a linear transformation of \mathbf{R} , so that the choice of projection matrix, \mathbf{P}_X , is immaterial. In other words, we could replace \mathbf{P}_X by $\mathbf{X}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}$, for any positive definite matrix \mathbf{W} , without affecting the likelihood. See Exercises 7.10–7.13. In this example there appears to be no loss of information on $\boldsymbol{\theta}$ by using \mathbf{R} in place of \mathbf{Y} , though it is difficult to give a totally satisfactory justification of this claim.

Although \mathbf{R} has a rank-deficient covariance matrix whose determinant is zero, it is still possible to write down explicitly the log likelihood for $\boldsymbol{\theta}$ based on \mathbf{R} . The usual method (Kitanidis, 1987), which is to choose an arbitrary full-rank sub-vector, introduces unnecessary and undesirable asymmetry into the formulae. Assuming that Σ has rank n and that \mathbf{X} has rank p , the marginal log likelihood for $\boldsymbol{\theta}$ based on \mathbf{R} is

$$l(\boldsymbol{\theta}; \mathbf{R}) = -\frac{1}{2} \log \det \Sigma - \frac{1}{2} \log \det(\mathbf{X}^T \Sigma^{-1} \mathbf{X}) - \frac{1}{2} Q_2(\mathbf{R}),$$

where

$$Q_2(\mathbf{R}) = \mathbf{R}^T (\Sigma^{-1} - \Sigma^{-1} \mathbf{X} (\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Sigma^{-1}) \mathbf{R},$$

the weighted residual sum of squares for \mathbf{R} , is unaffected by the choice of projection matrix. An equivalent expression in terms of the eigenvalues of the matrix $(\mathbf{I} - \mathbf{P}_X)\Sigma(\mathbf{I} - \mathbf{P}_X)$ is given by Patterson and Thompson (1971). Yet another equivalent expression is given by Harville (1974, 1977).

Note that in the important special case where $\Sigma = \sigma^2 \mathbf{I}$, this marginal log likelihood becomes

$$-\frac{1}{2}(n-p)\log\sigma^2 - \frac{1}{2}(\text{RSS})/\sigma^2$$

where RSS is the residual sum of squares. This is just the marginal log likelihood derived from $\text{RSS} \sim \sigma^2 \chi_{n-p}^2$.

For a derivation of the marginal log likelihood, also sometimes called the *restricted log likelihood* (Corbeil and Searle, 1976), see Exercises 7.8–7.13.

7.2.2 Conditional likelihood

Suppose that the parameter vector $\boldsymbol{\theta}$ can be partitioned into components $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$, in which $\boldsymbol{\psi}$ is the parameter vector of interest and $\boldsymbol{\lambda}$ is a vector of nuisance parameters. Suppose in addition that for each fixed value $\boldsymbol{\psi}_0$ of $\boldsymbol{\psi}$, the statistic $S_{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0)$ is sufficient for $\boldsymbol{\lambda}$ and complete. For definitions of completeness and sufficiency, see Cox and Hinkley (1974), section 2.2 or Lehmann (1986), sections 1.9 and 4.3. It is essential to distinguish two cases, (i) where $S_{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0)$ depends on $\boldsymbol{\psi}_0$, and (ii) where $S_{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0)$ is independent of $\boldsymbol{\psi}_0$ so that the sufficient statistic for $\boldsymbol{\lambda}$ is the same for all $\boldsymbol{\psi}_0$. In (i) the conditional distribution of \mathbf{Y} given $S_{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0)$ is independent of $\boldsymbol{\lambda}$ only under $\boldsymbol{\psi} = \boldsymbol{\psi}_0$. Thus we write

$$f_{Y|S_{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0)}(y | S_{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0); \boldsymbol{\psi}, \boldsymbol{\lambda})$$

for the conditional distribution.

In (ii), where $S_{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0) \equiv S_{\boldsymbol{\lambda}}$, we write the conditional density in the form

$$f_{Y|S_{\boldsymbol{\lambda}}}(y | S_{\boldsymbol{\lambda}}; \boldsymbol{\psi})$$

emphasizing that the conditional distribution is independent of $\boldsymbol{\lambda}$. Here there is no conceptual difficulty in using

$$l_c(\boldsymbol{\psi}) = \log f_{Y|S_{\boldsymbol{\lambda}}}(y | S_{\boldsymbol{\lambda}}; \boldsymbol{\psi}) = \log f_Y(y; \boldsymbol{\psi}, \boldsymbol{\lambda}) - \log f_{S_{\boldsymbol{\lambda}}}(s_{\boldsymbol{\lambda}}; \boldsymbol{\psi}, \boldsymbol{\lambda}) \quad (7.1)$$

as the conditional log likelihood for $\boldsymbol{\psi}$. The maximizing value $\hat{\boldsymbol{\psi}}_c$, and the Fisher information $i_{\boldsymbol{\psi}|S_{\boldsymbol{\lambda}}}$ based on the conditional log likelihood (7.1) are, in general, different from those derived from

the full likelihood. For a simple but important example of this conditioning argument, see sections 7.2.1 and 7.2.2.

When $S_\lambda(\psi_0)$ depends on ψ_0 , however, it is most important in writing the conditional density to distinguish between the two values of ψ . Thus

$$l_c(\psi, \lambda; \psi_0) = \log f_{Y|S_\lambda(\psi_0)}(y | S_\lambda(\psi_0); \psi, \lambda),$$

considered as a function of ψ and λ for fixed ψ_0 , is a log-likelihood function. For instance, under the usual regularity conditions, the conditional score statistic

$$\frac{\partial l_c(\psi, \lambda; \psi_0)}{\partial \psi}$$

has zero expectation given $S_\lambda(\psi_0)$ and conditional covariance matrix equal to

$$-E\left(\frac{\partial^2 l_c(\psi, \lambda; \psi_0)}{\partial \psi^2}\right).$$

It follows that the score statistic is uncorrelated with $S_\lambda(\psi_0)$ for all parameter values.

By contrast, the reduced function

$$l^*(\psi) = l_c(\psi, \lambda; \psi)$$

is not the logarithm of a density with respect to any fixed measure, and hence does not ordinarily have the properties of a log-likelihood function. The reason for this is that the transformation from the original variables \mathbf{Y} to the pair $(S_\lambda(\psi_0), \bar{S})$, where \bar{S} is a complementary statistic, involves a Jacobian that depends on ψ_0 . So long as ψ_0 is regarded as fixed, the Jacobian has no effect on the log likelihood and can be ignored. But if the Jacobian depends on ψ an extra term must be included in the log likelihood. Ordinarily, therefore, when computing likelihood functions it does not make sense to condition on parameter-dependent statistics. See Exercises 7.1–7.6 for several examples in which these differences are important.

Ideally we would like to choose a value ψ_0 for the conditioning statistic that is as near as possible to the conditional maximum-likelihood estimate, $\hat{\psi}_c$. For purposes of estimation, this effect is

achieved by solving the conditional likelihood equation $U_{\psi} = \mathbf{0}$, where

$$U_{\psi} = \frac{\partial l_c(\psi, \lambda; \psi_0)}{\partial \psi} \Big|_{\psi_0=\psi} \quad (7.2)$$

evaluated at $\hat{\lambda}(\psi)$. This is not the same as finding the roots of $\partial l^*(\psi)/\partial \psi = 0$ because the log likelihood is differentiated only with respect to the first argument. Another way of deriving (7.2) is to bias-correct the unconditional log-likelihood derivatives as follows:

$$\frac{\partial l(\psi, \lambda)}{\partial \psi} - E\left(\frac{\partial l(\psi, \lambda)}{\partial \psi} \mid S_{\lambda}(\psi)\right).$$

This bias-corrected derivative is identical to (7.2) with λ replaced by $\hat{\lambda}(\psi)$. This interpretation via estimating functions has been emphasized by Godambe (1976) and by Lindsay (1982).

The asymptotic variance of $\hat{\psi}_c$ is the inverse of

$$-E\left(\frac{\partial^2 l_c(\psi, \lambda; \psi_0)}{\partial \psi^2}\right) \Big|_{\psi_0=\psi}$$

evaluated at $\hat{\lambda}(\psi)$.

The conditional score function as defined above is unaffected by the parameterization chosen for the nuisance parameters because

$$\frac{\partial l_c(\psi, \lambda; \psi)}{\partial \lambda} \equiv \mathbf{0}$$

for all parameter values. Consequently there is no ambiguity regarding the meaning of (7.2).

This line of argument produces a usable score statistic having zero mean at the true parameter point, a ‘conditional likelihood’ estimator and an approximate standard error, but it does not produce a likelihood function for ψ directly. For this purpose the modified profile likelihood of Barndorff-Nielsen (1985, 1986) may be used. For a related derivation via orthogonal parameters, see Cox and Reid (1987).

The following example illustrates several aspects of the conditioning argument. Suppose that

$$Y_1 \sim N(\mu_1, 1) \quad \text{and} \quad Y_2 \sim N(\mu_2, 1)$$

are independent and that the ratio $\psi = \mu_2/\mu_1$ is the parameter of interest. To complete the parameterization, we may take $\lambda_1 = \mu_1$, $\lambda_2 = \mu_2$ or any other suitable complementary parameter such as $\lambda_3 = \mu_1 + \mu_2$ or the orthogonal parameter $(\mu_1^2 + \mu_2^2)^{1/2}$. A sufficient statistic for the nuisance parameter given $\psi = \psi_0$ is

$$S_\lambda(\psi_0) = Y_1 + \psi_0 Y_2.$$

Other equivalent forms of the sufficient statistic are

$$\begin{aligned}\hat{\mu}_1(\psi_0) &= \hat{\lambda}_1 = (Y_1 + \psi_0 Y_2)/(1 + \psi_0^2), \\ \hat{\mu}_2(\psi_0) &= \hat{\lambda}_2 = \psi_0(Y_1 + \psi_0 Y_2)/(1 + \psi_0^2), \\ \text{and } \hat{\lambda}_3 &= (1 + \psi_0)(Y_1 + \psi_0 Y_2)/(1 + \psi_0^2).\end{aligned}$$

The conditional log likelihood given $S_\lambda(\psi_0)$ is

$$l_c(\psi, \mu_1; \psi_0) = -\frac{1}{2} \frac{(y_2 - \psi_0 y_1 - (\psi - \psi_0)\mu_1)^2}{1 + \psi_0^2} - \frac{1}{2} \log(1 + \psi_0^2).$$

Differentiation with respect to ψ followed by setting $\psi_0 = \psi$ gives

$$U_\psi = \frac{\partial l_c(\psi, \mu_1; \psi_0)}{\partial \psi} \Big|_{\psi_0=\psi} = \frac{\mu_1(y_2 - \psi y_1)}{1 + \psi^2}. \quad (7.3)$$

Note that the Jacobian term vanishes on differentiation. Further,

$$\begin{aligned}E(U_\psi) &= 0, \\ \text{var}(U_\psi) &= \mu_1^2/(1 + \psi^2), \\ -E\left(\frac{\partial^2 l_c(\psi, \mu_1; \psi_0)}{\partial \psi^2}\right) \Big|_{\psi_0=\psi} &= \frac{\mu_1^2}{1 + \psi^2},\end{aligned}$$

so that the usual likelihood properties are satisfied. Also $\hat{\psi}_c = y_2/y_1$ with ‘asymptotic’ variance $(1 + \psi^2)/\mu_1^2$, the usual approximation for the variance of a ratio estimator. In all of these expressions μ_1 is to be replaced by $\hat{\mu}_1(\psi)$.

Normal approximation for the distribution of the ratio is unsatisfactory unless μ_1 is large compared to the standard deviation of Y_1 . Fieller confidence intervals generated via the score statistic U_ψ by

$$\{\psi : U_\psi^2 / \text{var}(U_\psi) \leq k_{\alpha/2}^2\}$$

are exact and are preferred over Normal approximations.

By contrast, differentiation of the reduced function $l^*(\cdot)$ gives

$$\frac{\partial l_c(\psi, \mu_1; \psi)}{\partial \psi} = \frac{y_1(y_2 - \psi y_1)}{1 + \psi^2} + \frac{\psi(y_2 - \psi y_1)^2}{(1 + \psi^2)^2} - \frac{\psi}{1 + \psi^2}.$$

The latter derivative has mean $-\psi/(1 + \psi^2)$. For further discussion of this point see Exercise 7.19.

7.2.3 Exponential-family models

Suppose that the log likelihood for $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$ can be written in the exponential-family form

$$l(\boldsymbol{\theta}; y) = \boldsymbol{\theta}^T \mathbf{s} - b(\boldsymbol{\theta})$$

and admits a decomposition of the form

$$l(\boldsymbol{\theta}; y) = \boldsymbol{\psi}^T \mathbf{s}_1 + \boldsymbol{\lambda}^T \mathbf{s}_2 - b(\boldsymbol{\psi}, \boldsymbol{\lambda}), \quad (7.4)$$

where $\mathbf{s} = (\mathbf{s}_1(y), \mathbf{s}_2(y))$ are functions of the data. Likelihood functions of this type occur most commonly when the observations are independent and the model considered is linear in the canonical parameter. Then the sufficient statistic is a linear function of the data, namely $\mathbf{s} = \mathbf{X}^T \mathbf{y}$. A decomposition of the form (7.4) can be achieved only if the parameter of interest, $\boldsymbol{\psi}$, is a linear function of $\boldsymbol{\theta}$. The choice of nuisance parameter, $\boldsymbol{\lambda}$, is to a large extent arbitrary and the inferences regarding $\boldsymbol{\psi}$ should be unaffected by the parameterization chosen for $\boldsymbol{\lambda}$.

It may be helpful at this stage to consider a simple example. Suppose that Y_1, Y_2 are independent Poisson random variables with means μ_1, μ_2 and that we are interested in the ratio $\psi' = \mu_1/\mu_2$. Here $\theta_i = \log \mu_i$ are the canonical parameters, and the parameter of interest $\psi = \log \psi' = \theta_1 - \theta_2$ is a linear contrast of the canonical parameters. For the nuisance parameter, we may choose any one of a variety of complementary parameters such as

$$\lambda'_1 = \mu_1, \quad \lambda'_2 = \mu_2, \quad \lambda'_3 = \mu_1 + \mu_2 \quad \text{or} \quad \lambda'_4 = \mu_1 \mu_2.$$

The log likelihood expressed initially in terms of $\boldsymbol{\theta}$ is

$$\begin{aligned} l(\boldsymbol{\theta}; \mathbf{y}) &= y_1 \theta_1 + y_2 \theta_2 - \exp(\theta_1) - \exp(\theta_2) \\ &= (y_1 + y_2) \lambda_1 - y_2 \psi - \exp(\lambda_1)(1 + e^{-\psi}) \\ &= (y_1 + y_2) \lambda_2 + y_1 \psi - \exp(\lambda_2)(1 + e^\psi) \\ &= \frac{1}{2}(y_1 + y_2) \lambda_4 + \frac{1}{2}(y_1 - y_2) \psi - 2 \exp(\lambda_4/2) \cosh(\frac{1}{2}\psi), \end{aligned}$$

where $\psi = \log \psi'$ and $\lambda_j = \log \lambda'_j$.

It follows from (7.4) that for any given value of ψ , s_2 is sufficient for the nuisance parameter and that s_2 is the same whatever parameterization is chosen for the nuisance parameter. In the above example, $s_2 = Y_\cdot$ for each of the parameterizations considered. The conditional distribution of the data \mathbf{Y} given s_2 does not depend on λ and hence the conditional log likelihood

$$l(\psi | s_2) = \psi^T s_1 - b^*(\psi; s_2) \quad (7.5)$$

may be used for inferences regarding ψ . Given the sufficient statistic s_2 , the value of the nuisance parameter is irrelevant in subsequent calculations.

Note that the conditional log likelihood retains the exponential-family form in which s_1 is conditionally sufficient for ψ given the value of s_2 . In the Poisson example, the required conditional distribution is that of the pair (Y_1, Y_2) given that $Y_\cdot = m$, and this is well known to yield the binomial distribution. It is immaterial here whether we work with the conditional distribution of $Y_1 | Y_\cdot$, $Y_2 | Y_\cdot$ or $Y_1 - Y_2 | Y_\cdot$ because these conditional distributions differ by a fixed linear transformation.

For a second example, suppose that $Y_1 \sim B(1, \pi_1)$ and $Y_2 \sim B(1, \pi_2)$ are independent and that the odds ratio

$$\psi' = \frac{\pi_1}{1 - \pi_1} / \frac{\pi_2}{1 - \pi_2}$$

is the parameter of interest. The log-likelihood function is

$$\begin{aligned} & y_1 \log\left(\frac{\pi_1}{1 - \pi_1}\right) + y_2 \log\left(\frac{\pi_2}{1 - \pi_2}\right) + \log(1 - \pi_1) + \log(1 - \pi_2) \\ &= y_1 \psi + (y_1 + y_2) \log\left(\frac{\pi_2}{1 - \pi_2}\right) + \log(1 - \pi_1) + \log(1 - \pi_2) \end{aligned}$$

where $\psi = \log \psi'$. By the argument just given, we are led to consider the conditional distribution of (Y_1, Y_2) given that $Y_\cdot = y_\cdot$. If $y_\cdot = 0$ or $y_\cdot = 2$ the conditional distribution is degenerate. Otherwise if $y_\cdot = 1$, we have

$$\begin{aligned} \text{pr}(Y_1 = 0 | Y_\cdot = 1) &= 1/(1 + \psi') \\ \text{pr}(Y_1 = 1 | Y_\cdot = 1) &= \psi'/(1 + \psi') \end{aligned}$$

This is a particular instance of the hypergeometric distribution studied in section 7.3.

7.2.4 Profile likelihood

In those instances where they exist, marginal and conditional likelihoods work well, often with little sacrifice of information. However, marginal and conditional likelihoods are available only in very special problems. The profile log likelihood, while less satisfactory from several points of view, does have the important virtue that it can be used in all circumstances.

Let $\hat{\lambda}_\psi$ be the maximum-likelihood estimate of λ for fixed ψ . This maximum is assumed here to be unique, as it is for most generalized linear models. The partially maximized log-likelihood function,

$$l^\dagger(\psi; y) = l(\psi, \hat{\lambda}_\psi; y) = \sup_{\lambda} l(\psi, \lambda; y)$$

is called the profile log likelihood for ψ . Under certain conditions the profile log likelihood may be used just like any other log likelihood. In particular, the maximum of $l^\dagger(\psi; y)$ coincides with the overall maximum-likelihood estimate. Further, approximate confidence sets for ψ may be obtained in the usual way, namely

$$\{\psi : 2l^\dagger(\hat{\psi}; y) - 2l^\dagger(\psi; y) \leq \chi^2_{p, 1-\alpha}\}$$

where $p = \dim(\psi)$. Alternatively, though usually less accurately, intervals may be based on $\hat{\psi}$ together with the second derivatives of $l^\dagger(\psi; y)$ at the maximum. Such confidence intervals are often satisfactory if $\dim(\lambda)$ is small in relation to the total Fisher information, but are liable to be misleading otherwise.

Unfortunately $l^\dagger(\psi; y)$ is not a log likelihood function in the usual sense. Most obviously, its derivative does not have zero mean, a property that is essential for estimating equations. In fact the derivative of l^\dagger may be written in terms of the partial derivatives of l as follows:

$$\begin{aligned} \frac{\partial l^\dagger}{\partial \psi} &= \frac{\partial}{\partial \psi} l(\psi, \hat{\lambda}_\psi; y) \\ &= \frac{\partial l}{\partial \psi} + \frac{\partial^2 l}{\partial \psi \partial \lambda} (\hat{\lambda}_\psi - \lambda) + \frac{1}{2} \frac{\partial^3 l}{\partial \psi \partial \lambda^2} (\hat{\lambda}_\psi - \lambda)^2 + \dots \\ &\quad + \left\{ \frac{\partial l}{\partial \lambda} + \frac{\partial^2 l}{\partial \lambda^2} (\hat{\lambda}_\psi - \lambda) + \frac{1}{2} \frac{\partial^3 l}{\partial \lambda^3} (\hat{\lambda}_\psi - \lambda)^2 + \dots \right\} \frac{\partial \hat{\lambda}_\psi}{\partial \psi} \end{aligned}$$

The expression in parentheses is just $\partial l(\psi, \lambda)/\partial \lambda$ evaluated at $\hat{\lambda}_\psi$, and hence is identically zero. Under the usual regularity conditions

for large n , the remaining three terms are $O_p(n^{1/2})$, $O_p(n^{1/2})$ and $O_p(1)$ respectively. The first term has zero mean but the remaining two have mean $O(1)$ if $\hat{\lambda}_\psi$ is a consistent estimate of λ . Their expectations may be inflated if $\hat{\lambda}_\psi$ is not consistent.

A simple expression for the approximate mean of $\partial l^\dagger / \partial \psi$ in terms of cumulants of the derivatives of l is given by McCullagh and Tibshirani (1988).

In general, if the dimension of λ is a substantial fraction of n , the mean of $\partial l^\dagger / \partial \psi$ is not negligible and the profile log likelihood can be misleading if interpreted as an ordinary log likelihood.

It is interesting to compare the profile log likelihood with the marginal log likelihood in a model for which both can be calculated explicitly. The covariance-estimation model, considered briefly at the end of section 7.2.1, is such an example. The profile log likelihood for the covariance parameters θ in that problem is

$$l^\dagger(\theta; y) = -\frac{1}{2} \log \det \Sigma - \frac{1}{2} Q_2(\mathbf{R}),$$

which differs from the marginal log likelihood given at the end of section 7.2.1 by the term $\frac{1}{2} \log \det(\mathbf{X}^T \Sigma^{-1} \mathbf{X})$. Both the marginal and profile log likelihoods depend on the data only through the contrasts or residuals, \mathbf{R} . The marginal log likelihood is clearly preferable to l^\dagger in this example, because l^\dagger is not a log likelihood. The derivatives of l^\dagger , unlike those of the marginal log likelihood, do not have zero mean.

The use of profile likelihoods for the estimation of covariance functions has been studied by Mardia and Marshall (1984).

7.3 Hypergeometric distributions

7.3.1 Central hypergeometric distribution

Suppose that a simple random sample of size m_1 is taken from a population of size m_* . The population is known to comprise s_1 individuals who have attribute A and $s_2 = m_* - s_1$ who do not. In the sample, Y individuals have attribute A and the remainder, $m_1 - Y$, do not. The following table gives the numbers of sampled and non-sampled subjects who possess the attribute in question.

	<i>Attribute</i>		
	<i>A</i>	\bar{A}	<i>Total</i>
<i>sampled</i>	$Y \equiv Y_{11}$	$m_1 - Y \equiv Y_{12}$	m_1
<i>non-sampled</i>	$s_1 - Y \equiv Y_{21}$	$m_2 - s_1 + Y \equiv Y_{22}$	m_2
<i>Total</i>	s_1	s_2	$m_* \equiv s_*$

Under the simple random sampling model, the distribution of Y conditionally on the marginal totals \mathbf{m}, \mathbf{s} is

$$\text{pr}(Y = y | \mathbf{m}, \mathbf{s}) = \frac{\binom{m_1}{y} \binom{m_2}{s_1 - y}}{\binom{m_*}{s_1}} = \frac{\binom{s_1}{y} \binom{s_2}{m_1 - y}}{\binom{s_*}{m_1}} \quad (7.6)$$

The range of possible values for y is the set of integers satisfying

$$a = \max(0, s_1 - m_2) \leq y \leq \min(m_1, s_1) = b. \quad (7.7)$$

There are $\min(m_1, m_2, s_1, s_2) + 1$ points in the sample space. If $a = b$, the conditional distribution puts all its mass at the single point a . Degeneracy occurs only if one of the four marginal totals is zero.

The central hypergeometric distribution (7.6) is denoted by $Y \sim H(\mathbf{m}, \mathbf{s})$ or by $Y \sim H(\mathbf{s}, \mathbf{m})$.

An alternative derivation of the hypergeometric distribution is as follows. Suppose that $Y_1 \sim B(m_1, \pi)$ and $Y_2 \sim B(m_2, \pi)$ are independent binomial random variables. Then the conditional distribution of $Y \equiv Y_1$ conditionally on $Y_1 + Y_2 = s_1$ is given by (7.6).

The descending factorial moments of Y are easily obtained from (7.6) as follows:

$$\mu_{[r]} = E\{Y^{(r)}\} = m_1^{(r)} s_1^{(r)} / m_*^{(r)},$$

where $Y^{(r)} = Y(Y - 1)\dots(Y - r + 1)$, provided that $r \leq \min(m_1, s_1)$. From these factorial moments we may compute the cumulants of Y as follows. First, define the following functions of the marginal frequencies in terms of the sampling fraction $\tau = m_1/m_*$.

$$K_1 = s_1/m_*,$$

$$K_2 = s_1 s_s / m_*^{(2)},$$

$$\lambda_1 = m_* \tau_1 = m_1,$$

$$\lambda_2 = m_* \tau_1 (1 - \tau_1) = m_1 m_2 / m_*,$$

$$\begin{aligned}
K_3 &= s_1 s_2 (s_2 - s_1) / m_*^{(3)}, & \lambda_3 &= m_* \tau_1 (1 - \tau_1) (1 - 2\tau_1) \\
&&&= m_1 m_2 (m_2 - m_1) / m_*^2, \\
K_4 &= s_1 s_2 \{m_* (m_* + 1) - 6s_1 s_2\} / m_*^{(4)}, \\
K_{22} &= s_1^{(2)} s_2^{(2)} / m_*^{(4)}, & \lambda_4 &= m_* \tau_1 (1 - \tau_1) (1 - 6\tau_1 (1 - \tau_1)).
\end{aligned}$$

The first four cumulants of Y are

$$\begin{aligned}
E(Y) &= K_1 \lambda_1, & \text{var}(Y) &= K_2 \lambda_2, \\
\kappa_3(Y) &= K_3 \lambda_3, & \kappa_4(Y) &= K_4 \lambda_4 - 6K_{22} \lambda_2^2 / (m_* - 1).
\end{aligned} \tag{7.8}$$

Note that λ_r is the r th cumulant of the $B(m_*, \tau_1)$ distribution associated with the sampling fraction, whereas K_1, \dots, K_4, K_{22} are the population k -statistics and polykay up to order four. Details of these symmetric functions are given in McCullagh (1987), Chapter 4, especially section 4.6. For large m_* and for fixed sampling fraction, the λ s are $O(m_*)$, whereas the K s are $O(1)$ for fixed attribute ratio, s_1/s_2 .

Note that the third cumulant of Y is zero if either $K_3 = 0$ or $\lambda_3 = 0$. In fact all odd-order cumulants are zero under these conditions and the distribution of Y is symmetric.

7.3.2 Non-central hypergeometric distribution

The non-central hypergeometric distribution with odds ratio ψ is an exponentially weighted version of the central hypergeometric distribution (7.6). Thus

$$\text{pr}(Y = y; \psi) = \frac{\binom{m_1}{y} \binom{m_2}{s_1 - y} \psi^y}{P_0(\psi)} \tag{7.9}$$

where $P_0(\psi)$ is the polynomial in ψ ,

$$P_0(\psi) = \sum_{j=a}^b \binom{m_1}{j} \binom{m_2}{s_1 - j} \psi^j.$$

The range of summation is given by (7.7). This distribution arises in the exponentially weighted sampling scheme in which each of the $\binom{m_*}{m_1}$ possible samples is weighted proportionally to ψ^y , where

y is a particular function of the sample. Here y is the number of individuals in the sample who possess attribute A , but in principle any function of the sample could be chosen.

Alternatively, the non-central hypergeometric distribution may be derived as follows. Suppose that $Y_1 \sim B(m_1, \pi_1)$, $Y_2 \sim B(m_2, \pi_2)$ are independent binomial random variables and that $\psi = \pi_1(1-\pi_2)/\{\pi_2(1-\pi_1)\}$ is the odds ratio. Then the conditional distribution of Y_1 given that $Y_+ = s_1$ is non-central hypergeometric with parameter ψ . For conciseness, we write $Y \sim H(\mathbf{m}, \mathbf{s}; \psi)$ to denote the conditional distribution (7.9). Note that $P_0(1) = \binom{m_1}{s_1}$, so that $H(\mathbf{m}, \mathbf{s}; 1)$ is identical to $H(\mathbf{m}, \mathbf{s})$.

An ‘observation’ from the distribution (7.9) is often presented as a 2×2 table in which the marginal totals are \mathbf{m} and \mathbf{s} . The contribution of such an observation to the conditional log likelihood is

$$y \log \psi - \log P_0(\psi),$$

where the dependence on \mathbf{m} and \mathbf{s} has been suppressed in the notation for the polynomial $P_0(\psi)$. This log likelihood has the standard exponential-family form with canonical parameter $\theta = \log \psi$ and cumulant function

$$K(\theta) = \log P_0(e^\theta).$$

The mean and variance of Y are therefore

$$\begin{aligned}\kappa_1(\theta) &= E(Y; \theta) = K'(\theta) = P_1(\psi)/P_0(\psi) \\ \kappa_2(\theta) &= \text{var}(Y; \theta) = K''(\theta) = P_2(\psi)/P_0(\psi) - \{P_1(\psi)/P_0(\psi)\}^2,\end{aligned}$$

where $P_r(\psi)$ is the polynomial

$$P_r(\psi) = \sum_{j=a}^b j^r \psi^j \binom{m_1}{j} \binom{m_2}{s_1 - j}. \quad (7.10)$$

More generally, the moments about the origin are expressible as rational functions in ψ , namely

$$\mu_r(\psi) = P_r(\psi)/P_0(\psi).$$

Unfortunately the functions $\kappa_1(\theta)$ and $\kappa_2(\theta)$ are awkward to compute particularly if the range of summation in (7.10) is extensive. The following approximations are often useful. First, it is easily shown that, conditionally on the marginal totals,

$$E(Y_{11}Y_{22}) = \psi E(Y_{12}Y_{21})$$

and, more generally, that

$$E(Y_{11}^{(r)}Y_{22}^{(r)}) = \psi^r E(Y_{12}^{(r)}Y_{21}^{(r)}).$$

Hence, since $E(Y_{11}Y_{22}) = \mu_{11}\mu_{22} + \kappa_2$, we have

$$\psi = \frac{\mu_{11}\mu_{22} + \kappa_2}{\mu_{12}\mu_{21} + \kappa_2},$$

where $\mu_{11} = E(Y_{11}; \theta), \dots$ are the conditional means for the four cells, and κ_2 is the conditional variance of each cell. Consequently we have the following exact relationship between $\kappa_1 \equiv \mu_{11}$ and κ_2 :

$$\kappa_1(m_2 - s_1 + \kappa_1) + \kappa_2 = \psi \{(s_1 - \kappa_1)(m_1 - \kappa_1) + \kappa_2\}. \quad (7.11)$$

In addition, the following approximate relationship may be derived from asymptotic considerations of the type discussed in section 6.5.6:

$$\kappa_2 \simeq \frac{m_*}{m_* - 1} \left(\frac{1}{\mu_{11}} + \frac{1}{\mu_{12}} + \frac{1}{\mu_{21}} + \frac{1}{\mu_{22}} \right)^{-1}. \quad (7.12)$$

In addition to being asymptotically correct for large $b - a$, this expression is exact for $m_* = 2$, the smallest non-degenerate value, and also for $\psi = 1$, whatever the marginal configuration.

The simultaneous solution to (7.11) and (7.12) gives a very accurate approximation to the conditional mean and variance provided that either $|\theta| < 2$ or the marginal totals are large: see Breslow and Cologne (1986). An equally accurate but slightly more complicated approximation is given by Barndorff-Nielsen and Cox (1979).

7.3.3 Multivariate hypergeometric distribution

Suppose that $\mathbf{Y}_1 \sim M(m_1, \boldsymbol{\pi})$ and $\mathbf{Y}_2 \sim M(m_2, \boldsymbol{\pi})$ are independent multinomial vectors, each on k categories. Then the conditional distribution of the vector $\mathbf{Y} \equiv \mathbf{Y}_1$ given that $\mathbf{Y}_1 + \mathbf{Y}_2 = \mathbf{s}$ is as follows.

$$\text{pr}(\mathbf{Y} = \mathbf{y} | \mathbf{s}) = \frac{\binom{m_1}{\mathbf{y}} \binom{m_2}{\mathbf{s} - \mathbf{y}}}{\binom{m_{\cdot}}{\mathbf{s}}} = \frac{\binom{s_1}{y_1} \dots \binom{s_k}{y_k}}{\binom{s_{\cdot}}{y_{\cdot}}} \quad (7.13)$$

where $s_{\cdot} \equiv m_{\cdot}$ and $y_{\cdot} \equiv m_1$. From a statistical perspective, one important aspect of this conditional distribution is that it does not depend on the multinomial probability vector $\boldsymbol{\pi}$.

An alternative derivation of (7.13) based on simple random sampling from a finite population of size m_{\cdot} is as follows. Suppose that attribute G has k levels and that the k levels of G are mutually exclusive and exhaustive. For instance G might denote a particular genetic marker such as blood group with levels O, A, B, AB. In a different context, G might denote the species of salmon in Lake Michigan, with levels whose labels are *coho*, *chinook*, ... Under simple random sampling, the distribution of species in a sample of size m_1 is given by (7.13), where s_1, s_2, \dots are the numbers of *coho*, *chinook*, ... in the lake.

If the sampled and non-sampled individuals are arranged in a two-way table, the entries appear as follows.

	<i>Attribute level</i>				
	G_1	G_2	\dots	G_k	<i>Total</i>
<i>Sampled</i>	Y_{11}	Y_{12}	\dots	Y_{1k}	m_1
<i>Not sampled</i>	Y_{21}	Y_{22}	\dots	Y_{2k}	m_2
<i>Total</i>	s_1	s_2	\dots	s_k	$m_{\cdot} \equiv s_{\cdot}$

Evidently, in this table we may reverse the roles played by the rows and the columns. Thus, suppose that $Y_1 \sim B(s_1, \tau), \dots, Y_k \sim B(s_k, \tau)$ are independent binomial random variables. Then the joint conditional distribution of $\mathbf{Y} = (Y_1, \dots, Y_k)$, conditional on the event $Y_{\cdot} = m_1$, is again given by (7.13). Note that in the first derivation, conditioning reduces the dimension of the sample space from $2(k-1)$ to $k-1$, in the process eliminating $k-1$ parameters $\boldsymbol{\pi}$.

In the latter derivation, conditioning reduces the dimension from k to $k - 1$, in the process eliminating the single parameter τ .

The joint conditional mean vector and covariance matrix of \mathbf{Y} are given by

$$\begin{aligned} E(Y_j) &= \tilde{\pi}_j m_1 \\ \text{var}(Y_j) &= \tilde{\pi}_j(1 - \tilde{\pi}_j)m_1m_2/(m_* - 1) \\ \text{cov}(Y_i, Y_j) &= -\tilde{\pi}_i\tilde{\pi}_j m_1m_2/(m_* - 1), \end{aligned}$$

where $\tilde{\pi}_j = s_j/s_*$ is the proportion in the population who have attribute j .

7.3.4 Multivariate non-central hypergeometric distribution

Suppose that $\mathbf{Y}_1 \sim M(m_1, \boldsymbol{\pi}_1)$ and $\mathbf{Y}_2 \sim M(m_2, \boldsymbol{\pi}_2)$ are independent multinomial random variables on k categories each. Then the conditional distribution of $\mathbf{Y} \equiv \mathbf{Y}_1$ given that $\mathbf{Y}_1 + \mathbf{Y}_2 = \mathbf{s}$ is as follows:

$$\text{pr}(\mathbf{Y} = \mathbf{y} | \mathbf{s}, \boldsymbol{\psi}) = \frac{\binom{m_1}{\mathbf{y}} \binom{m_2}{\mathbf{s} - \mathbf{y}} \psi_1^{y_1} \dots \psi_k^{y_k}}{\sum_{\mathbf{j}} \binom{m_1}{\mathbf{j}} \binom{m_2}{\mathbf{s} - \mathbf{j}} \psi_1^{j_1} \dots \psi_k^{j_k}}. \quad (7.14)$$

The sum in the denominator runs over the entire conditional sample space, which comprises all non-negative integer-valued vectors \mathbf{y} satisfying the positivity conditions

$$0 \leq y_j \leq s_j, \quad \sum y_j = m_1.$$

The odds-ratio parameters ψ_j are defined as contrasts relative to category k by

$$\psi_j = \frac{\pi_{1j}\pi_{2k}}{\pi_{2j}\pi_{1k}}$$

so that $\psi_k \equiv 1$.

The exact non-central moments and cumulants are complicated functions of $\boldsymbol{\psi}$ and the marginal frequencies \mathbf{s} . Direct computation is awkward because of the nature of the sample space and because of the large number of points it contains. The following equation,

however, gives a simple exact relationship between the conditional mean vector μ_1 of Y and the conditional covariance matrix Σ .

$$\frac{E(Y_{1j}Y_{2k})}{E(Y_{2j}Y_{1k})} = \psi_j = \frac{\mu_{1j}\mu_{2k} - \sigma_{jk}}{\mu_{2j}\mu_{1k} - \sigma_{jk}}. \quad (7.15)$$

Note that

$$\sigma_{jk} = \text{cov}(Y_{1j}, Y_{1k}) = -\text{cov}(Y_{1j}, Y_{2k})$$

is negative for $j < k$.

The covariance matrix Σ of Y_{11}, \dots, Y_{1k} may be approximated quite accurately as follows. Define the vector ζ with components ζ_j given by

$$\frac{1}{\zeta_j} = \frac{1}{\mu_{1j}} + \frac{1}{\mu_{2j}}.$$

The approximate covariance matrix $\tilde{\Sigma}$ is then given in terms of ζ by

$$\tilde{\Sigma} = \frac{m_*}{m_* - 1} \{ \text{diag}(\zeta) - \zeta\zeta^T/\zeta_* \}. \quad (7.16)$$

This matrix has rank $k - 1$. The simultaneous solution of equations (7.15) and (7.16) gives the approximate mean and covariance matrix of Y as a function of ψ .

7.4 Some applications involving binary data

7.4.1 Comparison of two binomial probabilities

Suppose that a clinical trial is undertaken to compare the effect of a new drug or other therapy with the current standard drug or therapy. Ignoring side-effects and other complications, the response for each patient is assumed to be simply ‘success’ or ‘failure’. In order to highlight the differences between the conditional log likelihood and the unconditional log likelihood, it is assumed that the observed data are as shown in Table 7.1. For a single stand-alone experiment, the numbers in this Table are unrealistically small, except perhaps as the information available at an early stage in the experiment when few patients have been recruited. In the context of a large-scale multi-centre clinical trial, however, Table 7.1 might represent the contribution of one of the smaller

Table 7.1 *Hypothetical responses in one segment of a clinical trial*

		<i>Response</i>		
		<i>Success</i>	<i>Failure</i>	<i>Total</i>
<i>Treatment</i>		$Y_1 = 2$	1	$m_1 = 3$
<i>Control</i>		$Y_2 = 1$	3	$m_2 = 4$
<i>Total</i>		$Y_{\cdot} = 3$	4	$m_{\cdot} = 7$

centres to the study. It is in the latter context that the methods described here have greatest impact.

We begin with the usual assumption that responses are independent and homogeneous within each of the two groups. Allowance can be made for the differential effect of covariates measured on individuals, but to introduce such effects at this stage would only complicate the argument. Strict adherence to protocol, together with randomization and concealment, are essential to ensure comparability, internal homogeneity and independence. With these assumptions, the numbers of successes in each treatment group may be regarded as independent binomial variables $Y_i \sim B(m_i, \pi_i)$, where

$$\begin{aligned}\text{logit } \pi_1 &= \lambda + \Delta \\ \text{logit } \pi_2 &= \lambda.\end{aligned}\tag{7.17}$$

For a single experiment or 2×2 table, (7.17) is simply a re-parameterization from the original probability scale to the more convenient logistic scale. Implicit in the re-parameterization, however, is the assumption that the logistic difference, Δ is a good and useful measure of the treatment effect. In particular, when it is required to pool information gathered at several participating sites or hospitals, it is often assumed that λ may vary from site to site but that Δ remains constant over all sites regardless of the success rate for the controls.

In order to set approximate confidence limits for Δ , there are two principal ways in which we may proceed. The simplest way is to fit the linear logistic model (7.17) using the methods described in Chapter 4. Approximate confidence limits may be based on $\hat{\Delta}$ and its large-sample standard error. For the present example this gives

$$\hat{\Delta} = \log \left(\frac{2 \times 3}{1 \times 1} \right) = 1.792, \quad \text{s.e.}(\hat{\Delta}) \simeq 1.683.$$

Note that the large-sample variance of $\hat{\Delta}$ is

$$\text{var } \hat{\Delta} = 1/2 + 1/1 + 1/1 + 1/3 = 17/6.$$

More accurate intervals are obtained by working with the profile deviance,

$$D(y; \Delta) = 2l(\hat{\Delta}, \hat{\lambda}) - 2l(\Delta, \hat{\lambda}_\Delta)$$

where $\hat{\lambda}_\Delta$ is the maximum-likelihood estimate of λ for given Δ . This statistic is easy to compute using standard computer packages. For the data in Table 7.1, the profile deviance is plotted in Fig. 7.1. The nominal 90% large-sample confidence interval, determined graphically, is

$$\{\Delta : D(y; \Delta) - D(y; \hat{\Delta}) < 2.71\} = (-0.80, 4.95).$$

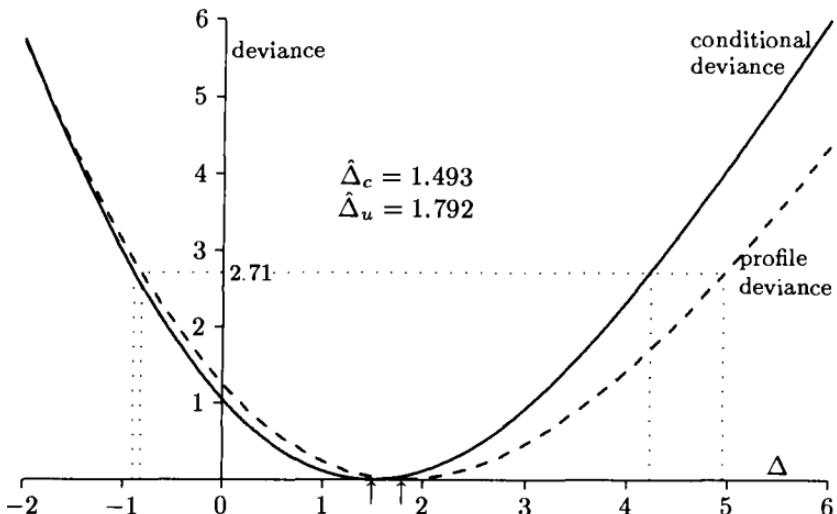


Fig. 7.1 Graphical comparison of hypergeometric and binomial deviance functions for the data in Table 7.1. Nominal 90% intervals for the log odds ratio, Δ , are indicated.

The alternative approach advocated here is to eliminate λ by using the conditional likelihood given Y_* . The hypergeometric log likelihood is

$$l_c(\Delta) = y_1 \Delta - \log P_0(e^\Delta),$$

where, for Table 7.1, $P_0(\psi)$ is equal to the cubic polynomial

$$P_0(\psi) = 4 + 18\psi + 12\psi^2 + \psi^3.$$

The hypergeometric likelihood has its maximum at a point $\hat{\Delta}_c$ different from the unconditional maximum $\hat{\Delta}$. In general $|\hat{\Delta}_c| \leq |\hat{\Delta}|$, with equality only at the origin. More precisely $\hat{\Delta}_c$ satisfies the standard exponential-family condition

$$y_1 = e^{\hat{\Delta}_c} P'_0(e^{\hat{\Delta}_c}) / P_0(e^{\hat{\Delta}_c}) = E(Y_1 | Y.; \hat{\Delta}_c).$$

In the example under discussion we find

$$\hat{\Delta}_c = 1.493, \quad \text{s.e.}(\hat{\Delta}_c) \simeq 1.492,$$

where the standard error is computed in the usual way, namely

$$\text{var}(\hat{\Delta}_c) \simeq 1 / \text{var}(Y; \hat{\Delta}_c) = 1/0.4495.$$

The conditional deviance function

$$2l_c(\hat{\Delta}_c) - 2l_c(\Delta)$$

is plotted as the solid line in Fig 7.1 and departs markedly from the profile deviance for large values of Δ .

7.4.2 Combination of information from several 2×2 tables

Suppose that data in the form of Table 7.1 are available from several sources, centres or strata, all cooperating in the same investigation. In the context of a multi-centre clinical trial, the strata are the medical centres participating in the trial. In some trials there may be many such centres, each contributing only a small proportion of the total patients enrolled. At each centre, one would expect that the pool of patients suitable for inclusion in the trial would differ in important respects that are difficult to measure. For instance, pollution levels, water hardness, rainfall, noise levels and other less tangible variables might have an effect on the response. In addition, nursing care and staff morale could have an appreciable effect on patients who are required to remain in hospital. Consequently, one

would expect the success rate for any medical treatment to vary appreciably from centre to centre.

Consequently, if we write

$$\pi_{1i} = \text{pr}(\text{success} | \text{treatment})$$

$$\pi_{2i} = \text{pr}(\text{success} | \text{control})$$

for the success probabilities at centre i , we may consider the linear logistic model

$$\begin{aligned}\text{logit } \pi_{1i} &= \lambda_i + \Delta \\ \text{logit } \pi_{2i} &= \lambda_i, \quad i = 1, \dots, n.\end{aligned}\tag{7.18}$$

The idea behind this parameterization is that $\Delta > 0$ implies that treatment is uniformly beneficial at all centres regardless of the control success rate: $\Delta < 0$ implies that the new treatment is uniformly poorer than the standard procedure. There is, of course, the possibility that Δ varies from centre to centre, even to the extent that $\Delta > 0$ for some centres and $\Delta < 0$ for others. Such interactions require careful investigation and detailed plausible explanation.

One obvious difficulty with the linear logistic model (7.18) is that it contains $n+1$ parameters to be estimated on the basis of $2n$ observed binomial proportions. In such circumstances, maximum likelihood need not be consistent or efficient for large n . However, following the general argument outlined in section 7.2.2, if we condition on the observed success totals, $Y_{\cdot i}$, at each of the centres, we have

$$Y_{1i} | Y_{\cdot i} \sim H(\mathbf{m}_i, y_{\cdot i}; \psi).\tag{7.19}$$

The hypergeometric log likelihood is thus the sum of n conditionally independent terms and depends on only one parameter, namely $\psi = e^\Delta$. Provided that the total conditional Fisher information is sufficiently large, standard large-sample likelihood theory applies to the conditional likelihood.

The conditional log likelihood for Δ is

$$l_c(\Delta) = \sum_i \{y_{1i}\Delta - \log P_0(e^\Delta; m_{1i}, m_{2i}, y_{\cdot i})\},$$

where additional arguments have been appended to the polynomial $P_0(\cdot)$ to emphasize its dependence on the marginal totals for stratum i .

The score statistic for no treatment effect is

$$U = \partial l_c / \partial \Delta|_{\Delta=0} = \sum_i \{Y_{1i} - E(Y_{1i})\} = \sum_i \{Y_{1i} - m_{1i}y_{\cdot i}/m_{\cdot i}\}.$$

The exact null variance of U is the sum of hypergeometric variances, namely

$$\text{var}(U) = \sum_i m_{1i}m_{2i}y_{\cdot i}(m_{\cdot i} - y_{\cdot i})/\{m_{\cdot i}^2(m_{\cdot i} - 1)\}.$$

The approximate one-sided significance level for the hypothesis of no treatment effect is $1 - \Phi(z^-)$, where

$$Z^- = (U - \frac{1}{2})/\sigma_U$$

is the continuity-corrected value. This test, first proposed by Mantel and Haenszel (1959), is known as the Mantel-Haenszel test. The Mantel-Haenszel estimator, which is different from the conditional likelihood estimator, is derived in Exercise 9.10.

7.4.3 Example: Ille-et-Vilaine study of oesophageal cancer

The data shown in Table 7.2 is a summary of the Ille-et-Vilaine retrospective study of the effect of alcohol consumption on the incidence of oesophageal cancer. A more complete list of the data, including information on tobacco consumption, is given in Appendix 1 of Breslow and Day (1980). In a retrospective study the numbers of cases (subjects with cancer) and the number of controls is to be regarded as fixed by the study design. The alcohol consumption rate (high/low) is the effective response. However, for the reasons given in section 4.4.3, the roles of these two variables can be reversed. We may, therefore, regard alcohol consumption rate as the explanatory covariate and outcome (cancer/no cancer) as the response even though such a view is not in accord with the sampling scheme. Since the analysis that follows is conditional on both sets of marginal totals, this role-reversal presents no conceptual difficulty.

It is common to find that the incidence of cancer increases with age. The cases in this study are older on average than the controls. If age were ignored in the analysis, the apparent effect of alcohol

Table 7.2 *Ille-et-Vilaine retrospective study of the relationship between alcohol consumption and the incidence of oesophageal cancer*

Age	Cancer		No cancer		$\tilde{\psi}_c$	Fitted values under model (ii)		
	Alcohol consumption					$\hat{\mu}_{11}$	Residual	
	80+	80-	80+	80-				
25-34	1	0	9	106	∞	0.33	1.42	
35-44	4	5	26	164	4.98	4.11	-0.07	
45-54	25	21	29	138	5.61	24.49	0.18	
55-64	42	34	27	139	6.30	40.09	0.59	
65-74	19	36	18	88	2.56	23.74	-1.89	
75+	5	8	0	31	∞	3.24	1.75	
Total	96	104	109	666		96.01	$X^2 = 9.04$	

consumption would be inflated. For that reason it is advisable to stratify the data by age. In other words, cases are matched with controls of a similar age. The treatment effect is therefore a comparison of cancer incidence rates between subjects of similar age.

Three models are considered.

1. a model in which the log odds-ratio is zero, meaning that alcohol consumption has no effect on the incidence of oesophageal cancer.
2. a model in which the log odds-ratio is constant, meaning that increased alcohol consumption increases the odds for oesophageal cancer by the factor e^ψ uniformly over all age groups.
3. a model in which the log odds-ratio increases or decreases linearly with increasing age.

Algebraically, these models may be written in the form

$$\begin{aligned}
 (i) \quad & \log \psi_i = 0, \\
 (ii) \quad & \log \psi_i = \beta_0, \\
 (iii) \quad & \log \psi_i = \beta_0 + \beta_1(i - 3.5),
 \end{aligned} \tag{7.20}$$

where $i = 1, \dots, 6$ indexes the age strata. The residual deviances for these three models are 89.83, 10.73 and 10.29 on 6, 5 and 4 degrees of freedom respectively.

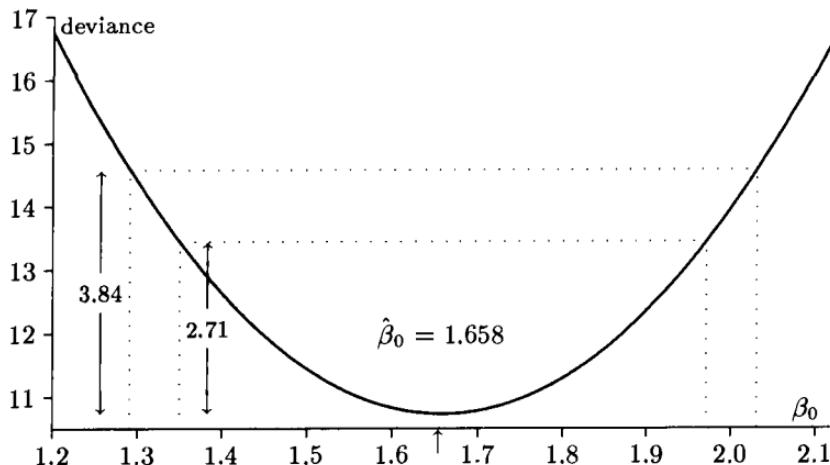


Fig. 7.2 Hypergeometric deviance for model (7.20). Nominal 90% and 95% intervals for the log odds ratio, β_0 , are indicated.

The model formula for (i) is unusual in that it is entirely empty, excluding even the intercept.

The estimate of β_0 for the model of constant odds-ratio is 1.658 with standard error 0.189. Fitted values and residuals under this model are shown in the final two columns of Table 7.2. The residuals, calculated by the formula

$$(y_{11} - \hat{\mu}_{11}) / \sqrt{V(\hat{\mu}_{11})},$$

exhibit no patterns that would suggest systematic deviation from constancy of the odds-ratio. The fact that we have chosen the (1,1) cell is immaterial because the residuals are equal in magnitude for the four cells of the response.

For the third model, the estimates are

$$\begin{aligned}\hat{\beta}_0 &= 1.7026 & \text{s.e.}(\hat{\beta}_0) &\simeq 0.2000 \\ \hat{\beta}_1 &= -0.1255 & \text{s.e.}(\hat{\beta}_1) &\simeq 0.1879\end{aligned}$$

confirming that there is no evidence of a linear trend in the log odds-ratios.

Both Pearson's statistic and the residual deviance statistic are a little on the large side, though of borderline statistical significance when compared to the nominal χ^2_5 distribution. This inflation may be due to factors that have been ignored in the present analysis.

The unconditional analysis for these data, in which each row of Table 7.2 is treated as a pair of independent binomial variables, gives very similar, though not identical, answers in this example. The unconditional residual deviances for the three models (7.20) are 90.56, 11.04 and 10.61. The unconditional maximum-likelihood estimate of β_0 in the second model is 1.670 with asymptotic standard error 0.190. As usual, the unconditional estimate is larger in magnitude than the conditional estimate. The unconditional estimate is biased away from the origin, though in this example the bias is small because the counts are, for the most part, moderately large. There are similar slight differences between the unconditional and conditional estimates for the third model. None of these differences is of sufficient magnitude to affect the conclusions reached.

Thus it appears that the habitual tippler will find no comfort in these data. The odds for oesophageal cancer are higher by an estimated factor of $5.251 = \exp(1.6584)$ in the high alcohol-consumption group than in the low alcohol group. This odds factor applies to all age groups even though the incidence of cancer increases with age. Approximate 95% confidence limits for the odds-ratio are

$$\exp(1.658 \pm 1.96 \times 0.189) = \exp(1.288, 2.028) = (3.624, 7.602),$$

which is almost identical to the interval (3.636, 7.622) obtained from the deviance plot in Fig. 7.2. Normal approximations tend to be more accurate when used on the $\log \hat{\psi}$ -scale rather than the $\hat{\psi}$ -scale.

7.5 Some applications involving polytomous data

7.5.1 Matched pairs: nominal response

Suppose that subjects in a study are matched in pairs and that a single polytomous response is observed for each subject. Following the usual procedure for matched pairs, we shall suppose that the logarithmic response probabilities for the control member of the i th pair are

$$\boldsymbol{\lambda}_i = (\lambda_{i1}, \dots, \lambda_{ik}),$$

which are free to vary in any haphazard or other way from pair to pair. We shall suppose in addition that the treatment effect as measured on the logarithmic scale is the same for all pairs. The logarithmic response probabilities for the treated member of the i th pair are therefore

$$\lambda_i + \Delta = (\lambda_{i1} + \Delta_1, \dots, \lambda_{ik} + \Delta_k).$$

The probability of observing response category j for control subject i is

$$\exp(\lambda_{ij}) / \sum_r \exp(\lambda_{ir}),$$

while the probabilities for the treated subject are

$$\exp(\lambda_{ij} + \Delta_j) / \sum_r \exp(\lambda_{ir} + \Delta_j).$$

Each response can be represented either as an integer R in the range $(1, k)$, or as an indicator vector Z having k components. The components of Z are

$$Z_j = \begin{cases} 1 & \text{if } R = j \\ 0 & \text{otherwise.} \end{cases}$$

Consider now a given pair having logarithmic response probabilities λ and $\lambda + \Delta$, for which the observed categories are r_1 and r_2 respectively. For any given value of Δ , the sufficient statistic for λ is the vector sum, $Z_* = Z_1 + Z_2$, of the observed responses. If $Z_* = (0, \dots, 2, \dots, 0)$, both R_1 and R_2 are determined by Z_* and the conditional distribution given Z_* is degenerate. However, if

$$Z_* = (0, \dots, 1, \dots, 1, \dots, 0),$$

with non-zero values in positions i and j , we must have

$$(R_1, R_2) = (i, j) \quad \text{or} \quad (j, i).$$

For $i \neq j$, the required conditional distribution is

$$\begin{aligned} \text{pr}(R_1 = i | Z_*) &= \frac{e^{\lambda_i} e^{\lambda_j + \Delta_j}}{e^{\lambda_i} e^{\lambda_j + \Delta_j} + e^{\lambda_j} e^{\lambda_i + \Delta_i}} \\ &= e^{\Delta_j} / (e^{\Delta_i} + e^{\Delta_j}), \end{aligned} \tag{7.21}$$

which is independent of λ as required. Every ordered pair of responses (i, j) with $i \neq j$ contributes a factor (7.21) to the conditional likelihood. Identical pairs, for which the control response is the same as the treatment response, contribute a factor of unity and can be ignored in the conditional likelihood. Thus, if Y_{ij} is the number of ordered pairs responding (i, j) , the symmetric total $m_{ij} = Y_{ij} + Y_{ji}$ is just the number of vector sums Z that have values in positions i and j . Hence, conditionally

$$\begin{aligned} Y_{ij} &\sim B(m_{ij}, \pi_{ij}) \quad i < j \\ \text{logit}(\pi_{ij}) &= \Delta_j - \Delta_i. \end{aligned} \tag{7.22}$$

The conditional log likelihood is therefore the product of $k(k-1)/2$ independent binomial factors satisfying the model shown above. As usual, the levels of the treatment factor Δ can be chosen to satisfy $\Delta_1 = 0$ or $\Delta_k = 0$ or $\sum \Delta_j = 0$. Evidently, from (7.22) only the differences are relevant.

Model (7.22), known as the model of quasi-symmetry, was first suggested by Caussinus (1965), though no derivation was given. The same model occurs in studies of population migrations where the term *gravity model* is used. In that context the k categories are k geographical locations and Y_{ij} is the number of families or individuals who migrate from area i to area j in the time period under study. For further details see Scholten and van Wissen (1985) or Upton (1985).

Model (7.22) is formally identical to the Bradley-Terry (1952) model used for ranking individuals in a paired competition. If the probability π_{ij} that subject i beats subject j satisfies (7.22), then the Δ s give a linear ranking of subjects.

A curious and unusual feature of the linear logistic model (7.22) is that the model matrix \mathbf{X} corresponding to the formula $\Delta_j - \Delta_i$ does not include the intercept nor does the constant vector lie in the column space of \mathbf{X} . For instance if $k = 3$ the model formula may be stated explicitly using matrix notation as

$$\begin{pmatrix} \text{logit } \pi_{12} \\ \text{logit } \pi_{13} \\ \text{logit } \pi_{23} \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \Delta_1 \\ \Delta_2 \\ \Delta_3 \end{pmatrix}$$

In this example \mathbf{X} is 3×3 with rank 2 and the sum of the three columns is $\mathbf{0}$.

The adequacy of the linear logistic model (7.22) can be tested by using the residual deviance or Pearson's statistic, each of which has $(k-1)(k-2)/2$ degrees of freedom.

7.5.2 *Ordinal responses*

Consider the application of the proportional-odds model (5.1) for the comparison of two multinomial responses in which the categories are ordered. The observations comprise two independent multinomial vectors

$$\mathbf{Y}_1 \sim B(m_1, \boldsymbol{\pi}_1), \quad \mathbf{Y}_2 \sim B(m_2, \boldsymbol{\pi}_2),$$

in which the cumulative probabilities γ_{1j}, γ_{2j} satisfy

$$\begin{aligned} \log\{\gamma_{1j}/(1 - \gamma_{1j})\} &= \theta_j; \\ \log\{\gamma_{2j}/(1 - \gamma_{2j})\} &= \theta_j - \Delta; \end{aligned} \quad j = 1, \dots, k-1. \quad (7.23)$$

In this model there is a single parameter of interest, Δ measuring the effect of treatment, and $k-1$ base-line parameters that determine the response probabilities for the control. Since the categories are ordered it is helpful to form cumulative totals not just for the probabilities but for the responses themselves. Thus we write $s_j = y_{1j} + y_{2j}$ for the response category totals and

$$\begin{aligned} Z_{1j} &= Y_{11} + \dots + Y_{1j} \\ Z_{2j} &= Y_{21} + \dots + Y_{2j} \\ S_j &= s_1 + \dots + s_j = Z_{\cdot j} \end{aligned}$$

for the cumulative responses in each group and the cumulative totals respectively. With this notation we have $Z_{ij} \sim B(m_i, \gamma_{ij})$ with γ_{ij} satisfying the linear logistic model (7.23). The nuisance parameters may be eliminated by conditioning on S_j . The resulting hypergeometric distribution is

$$\text{pr}(Z_{1j} = z_{1j} | S_j) = \binom{m_1}{z_{1j}} \binom{m_2}{S_j - z_{1j}} \psi^{z_{1j}} / \sum_r \binom{m_1}{r} \binom{m_2}{S_j - r} \psi^r, \quad (7.24)$$

where $\psi = e^\Delta$.

Thus, for each $j = 1, \dots, k - 1$, we have the conditional distribution

$$Z_{1j} | S_j \sim H(\mathbf{m}, S_j; \psi)$$

independently of the nuisance parameters. The aim in the discussion that follows is to construct an efficient estimate of ψ on the basis of these $k - 1$ conditional distributions. Unfortunately there is no joint conditional distribution of $\{Z_{1j}\}$ that depends only on ψ because the conditional distribution of Z_{1j} , given the vector $\mathbf{S} = (S_1, \dots, S_k)$, depends on both θ and ψ .

The following method of estimation seems to work well and has much in common with quasi-likelihood as discussed in Chapter 9. The argument runs as follows. Let $\chi_{1j}(\psi)$ be the conditional mean of Z_{1j} given S_j as derived from the hypergeometric distribution (7.24). For each $j = 1, \dots, k - 1$, the difference

$$Z_{1j} - \chi_{1j}(\psi)$$

has zero mean conditionally on S_j , and hence unconditionally. Note that $\chi_{1j}(\psi)$ depends also on S_j . For all generalized linear models the likelihood equations take the form $U(\hat{\psi}; y) = 0$ where $U(\psi; y)$ is a linear function of the data. By analogy, therefore, we seek to construct a function

$$U(\psi; Z) = \sum w_j^* \{Z_{1j} - \chi_{1j}(\psi)\}, \quad (7.25)$$

where $w_j^* = w_j^*(\psi, S_j)$, such that $U(\psi; Z)$ behaves ‘like’ a log-likelihood derivative. By construction, $U(\psi; Z)$ has zero mean whatever the choice of weights. The weights are chosen so that the mean of $-\partial U / \partial \Delta$ is equal to the variance of U .

The choice of weights is normally not critically important. At worst, a poor choice of weights may lead to a small loss of efficiency. With this in mind, we may choose $w_j^* = 1$ or $w_j^* = S_j(m_j - S_j)$. These are safe choices that lead to consistent estimates of ψ , but the variance of the resulting estimate is not simply related to $\partial U / \partial \Delta$.

More formally, however, standard theory for linear estimating equations shows that the optimal weights are given in vector form by

$$\mathbf{w}^* = \mathbf{V}^{-1} \mathbf{d},$$

where \mathbf{V} is the covariance matrix of Z_{11}, \dots, Z_{1k-1} and $d_j = \partial \chi_{1j}(\psi) / \partial \Delta$. By a common property of exponential-family models,

$$d_j = \text{var}(Z_{1j} | S_j; \psi),$$

which is easily computed either exactly using (7.24) or approximately using (7.12). However, it is unclear what is meant by the covariance matrix \mathbf{V} because the random variables Z_{1j}, \dots, Z_{1j-1} , whose probability distributions are given by (7.24), are defined on different sample spaces. To use the unconditional covariance matrix would be to violate the spirit of the exercise. Yet it does not make sense to talk of the covariance of two random variables unless they can be defined on a common sample space.

A pragmatic solution that has the merit of simplicity is to use the approximate hypergeometric covariance matrix $\tilde{\Sigma}$ as defined in (7.16). Since \mathbf{Z} is the vector of cumulative totals, we have

$$\tilde{\mathbf{V}} = \mathbf{L} \tilde{\Sigma} \mathbf{L}^T,$$

where \mathbf{L} is the lower triangular matrix forming cumulative totals. The matrix $\tilde{\mathbf{V}}$ thus constructed is a Green's matrix similar to the cumulative multinomial covariance matrix (5.13). With this choice for \mathbf{V} , we find

$$\begin{aligned} U(\psi; Z) &= \mathbf{d}^T \mathbf{D}^T \tilde{\Sigma}^{-1} \mathbf{D} (\mathbf{Z} - \boldsymbol{\chi}) \\ &= \mathbf{d}^T \mathbf{D}^T \tilde{\Sigma}^{-1} (\mathbf{Y}_1 - \boldsymbol{\mu}_1) \end{aligned}$$

where $\mathbf{D} = \mathbf{L}^{-1}$, $\mu_{1j} = \chi_{1j} - \chi_{1j-1}$ and $\mathbf{Y}_1^T = (Y_{11}, \dots, Y_{1k})$. We are free to choose the simplest generalized inverse of $\tilde{\Sigma}$, namely

$$\tilde{\Sigma}^{-1} = \frac{m_* - 1}{m_*} \text{diag}\{\zeta^{-1}\} = \frac{m_* - 1}{m_*} \text{diag}\{\mu_{1j}^{-1} + \mu_{2j}^{-1}\}.$$

Thus

$$\begin{aligned} U(\psi; Z) &= \frac{m_* - 1}{m_*} \sum_{j=1}^k (d_j - d_{j-1}) \left(\frac{1}{\mu_{1j}} + \frac{1}{\mu_{2j}} \right) (y_{1j} - \mu_{1j}) \\ &\simeq \sum_{j=1}^{k-1} \left(\frac{\zeta_j + \zeta_{j+1}}{\zeta_*} \right) (Z_{1j} - \chi_{1j}). \end{aligned} \tag{7.26}$$

The latter approximation comes from replacing \mathbf{d} by the diagonal elements of $\mathbf{\hat{V}}$. Note that the weights in the first expression above are not all positive.

The 'conditional likelihood' estimate, $\hat{\Delta}_c$, defined as the solution to the equation $U(\hat{\psi}_c; Z) = 0$, has asymptotic variance

$$\begin{aligned}\text{var}(\hat{\Delta}_c) &\simeq \zeta_{\cdot} \left\{ \sum_{j=1}^{k-1} (\zeta_j + \zeta_{j+1}) d_j \right\}^{-1} \\ &\simeq \frac{3(m_{\cdot} - 1)}{m_{\cdot} \zeta_{\cdot}} \left\{ 1 - \sum \{\zeta_j / \zeta_{\cdot}\}^3 \right\}^{-1},\end{aligned}\quad (7.27)$$

which is essentially the same as the variance of the unconditional maximum-likelihood estimate. See Exercise 5.3.

7.5.3 Example

It is only for very sparse tables that there is any appreciable difference between the 'conditional' likelihood estimator described in the previous section, and the unconditional maximum likelihood estimate as described in Chapter 5. By way of example, we consider here the first two rows of Table 5.1, involving the comparison of two cheese additives, A and B, ignoring C and D. The unconditional maximum-likelihood estimate of Δ in the proportional odds model (7.23) is -3.028 with asymptotic standard error 0.455. Table 7.3 shows the steps involved in one cycle of the iteration, beginning from $\Delta_0 = -3.028$.

Table 7.3 Steps required in one cycle of the iteration to compute $\hat{\Delta}_c$

S	Z_1	d	χ_1	μ_1	ζ	w^*
6	0	0.2842	0.3021	0.3021	0.2869	0.0615
15	0	0.8220	0.8997	0.5976	0.5579	0.1307
28	1	1.8891	2.2860	1.3863	1.2385	0.3288
46	8	3.6281	6.5994	4.3134	3.2798	0.5105
61	16	3.4019	14.5594	7.9600	3.7359	0.4485
75	24	1.9867	25.4325	10.8731	2.4285	0.3050
95	43	0.4461	43.4791	18.0466	1.7626	0.1580
103	51	0.0440	51.0462	7.5671	0.4095	0.0330
104	52	—	52.0	0.9538	0.0441	—
				52.0	13.7437	

The first column gives the cumulative category totals, S_j , and the second column gives the cumulative observations Z_{1j} for the first group. The third and fourth columns give the variances and means

$$d_j = \text{var}(Z_j | S_j) \quad \text{and} \quad \chi_{1j} = E(Z_{1j} | S_j).$$

both computed from the hypergeometric distribution with odds ratio $\psi_0 = \exp(-3.028)$. The fifth column gives the cell means for the first group, $\mu_{1j} = \chi_{1j} - \chi_{1,j-1}$. Finally, the last two columns give

$$\zeta_j = \left(\frac{1}{\mu_{1j}} + \frac{1}{\mu_{2j}} \right)^{-1}$$

where $\mu_{2j} = s_j - \mu_{1j}$ are the cell means for the second group, and $w_j^* = (\zeta_j + \zeta_{j+1})/\zeta_*$ are the required weights.

The score statistic $U = \sum w_j^*(Z_{1j} - \chi_{1j})$ is equal to 0.2881, while $\sum w_j^* d_j = 4.8016$. Thus the updated estimate of Δ is

$$\hat{\Delta}_c = -3.028 + 0.2881/4.8016 = -2.9680.$$

One further cycle gives $\hat{\Delta}_c = -2.9743$ with asymptotic variance

$$(\sum w_j^* d_j)^{-1} = (4.9047)^{-1} = 0.2039$$

and standard error 0.4515.

The difference between the conditional and unconditional estimates is only 12% of a standard error and is unlikely to have much effect on the conclusions reached. As usual, the conditional estimate is smaller in magnitude than the unconditional estimate.

7.6 Bibliographic notes

Conditional and marginal likelihoods for the elimination of nuisance parameters have been in use since the early part of this century. It can be argued that Student's usage of degrees of freedom rather than sample size as divisor in the estimation of σ^2 is an application of marginal likelihood, as shown in section 7.2.1. Bartlett (1936, 1937) made further important contributions, particularly to the problem of estimating a common mean when the sample variances are unequal and unknown.

Neyman and Scott (1948), in an important and influential paper, pointed out that when the number of nuisance parameters grows in proportion to n , maximum-likelihood estimates need not be consistent. Even if they are consistent, they need not be efficient.

The use of marginal likelihoods based on error contrasts, for the estimation of variance components, has been recommended by Patterson and Thompson (1971) and further studied by Harville (1974, 1977), Fraser (1968, 1979) and Corbeil and Searle (1976). The method is known as restricted maximum likelihood (REML). The application of the same technique to spatial covariance estimation is due to Kitanidis (1983, 1987), who points out that the marginal likelihood is superior in this context to full, or profile, likelihood.

The matched pairs design is an extreme case where the number of parameters grows in proportion to n , and consequently this design is used as a testing ground for procedures that purport to handle large numbers of nuisance parameters. For binary responses, Cox (1958b) showed that the conditional likelihood ignores all pairs for which the responses are equal. The test for no treatment effect is a simple comparison of the number of pairs responding (0, 1) with those responding (1, 0). This test had been proposed earlier by McNemar (1947). For further details, see Andersen (1973).

Kalbfleisch and Sprott (1970) give an excellent review of various modifications to the likelihood when there is a large number of nuisance parameters. In particular, they discuss the thorny problem of conditioning, when the conditioning statistic for the removal of λ , $S_\lambda(\psi)$, depends on ψ . See also Godambe (1976) and Lindsay (1982) for a discussion of the same topic from the vantage of optimal estimating equations.

Regression models for the log odds-ratio, based on the non-central hypergeometric distribution, are now widely used in retrospective studies of disease. This line of work can be traced back to Mantel and Haenszel (1959). For more recent work, see Breslow (1976, 1981) and Breslow and Day (1980). There are close connections here with Cox's (1972a, 1975) partial likelihood, which is also designed for the removal of nuisance parameters.

The formulae in section 7.3.2 for non-central hypergeometric moments were given by Harkness (1965). They were subsequently discussed by Mantel and Hankey (1975) and used by Mantel (1977)

for approximating the non-central hypergeometric mean.

Saddlepoint methods for approximating conditional likelihoods for generalized linear models with canonical links are discussed by Davison (1988).

Section 7.5 is based on McCullagh (1982, 1984c).

7.7 Further results and exercises 7

7.1 Suppose that Y_1, \dots, Y_n are *i.i.d.* $N(\mu, \sigma^2)$. Show that if $\mu = \mu_0$ is given, then

$$S_0 \equiv S(\mu_0) = \sum (Y_j - \mu_0)^2$$

is a complete sufficient statistic for σ^2 .

7.2 Show that the log likelihood for (μ, σ^2) in the previous exercise is

$$l(\mu, \sigma^2) = -\frac{1}{2\sigma^2} S(\mu) - \frac{n}{2} \log \sigma^2.$$

Show also that the statistic $S(\mu_0)$ has the non-central χ^2 distribution on n degrees of freedom given by

$$\frac{\exp\{-(S_0 + \lambda)/(2\sigma^2)\} (S_0/(2\sigma^2))^{n/2-1}}{2\sigma^2 \pi^{1/2} \Gamma(\frac{n-1}{2})} \sum_{r=0}^{\infty} \left(\frac{\lambda}{2\sigma^2}\right)^r \frac{r^r}{r!} B\left(\frac{n-1}{2}, r + \frac{1}{2}\right)$$

where $\lambda = n(\mu - \mu_0)^2$. Hence show that the conditional log likelihood given S_0 is

$$\begin{aligned} l_c(\mu, \sigma^2; \mu_0) &= -\frac{1}{2\sigma^2} \{S(\mu) - S(\mu_0)\} - \left(\frac{n}{2} - 1\right) \log S(\mu_0) \\ &\quad + \frac{\lambda}{2\sigma^2} - \log \sum_{r=0}^{\infty} \left(\frac{\lambda}{2\sigma^2}\right)^r \frac{r^r}{r!} B\left(\frac{n-1}{2}, r + \frac{1}{2}\right), \end{aligned}$$

whereas the reduced function $l^*(\mu, \sigma^2)$ is

$$l^*(\mu, \sigma^2) = l_c(\mu, \sigma^2; \mu) = -\frac{1}{2}(n-2) \log S(\mu).$$

7.3 For the function l^* defined in the previous exercise, show that

$$U^* = \frac{\partial l^*}{\partial \mu} = \frac{(n-2)}{S(\mu)} \sum (y - \mu) = \frac{n(n-2)(\bar{y} - \mu)}{(n-1)s^2 + n(\bar{y} - \mu)^2}$$

whereas the derivative of $l_c(\mu, \sigma^2; \mu_0)$ with respect to μ , evaluated at $\mu_0 = \mu$ is

$$\frac{\partial l_c}{\partial \mu} \Big|_{\mu_0=\mu} = \frac{1}{\sigma^2} \sum (y - \mu),$$

which is monotonely increasing in \bar{y} . Comment briefly on the differences between the two derivatives for $n = 1, 2$.

7.4 For the problem discussed in the previous three exercises, show that Y_i and Y_j are conditionally uncorrelated with variance $S(\mu)/n$ and that

$$\text{var}(\bar{Y} | S(\mu)) = S(\mu)/n^2.$$

Hence deduce that the conditional moments of the derivatives of l^* are

$$\begin{aligned} E(U^* | S(\mu)) &= 0 \\ \text{var}(U^* | S(\mu)) &= (n-2)^2/S(\mu) \\ -E\{\partial^2 l^*/\partial \mu^2 | S(\mu)\} &= (n-2)^2/S(\mu). \end{aligned}$$

[Bartlett, 1936].

7.5 Suppose that Y_1, \dots, Y_n are observations taken at spatial locations s_1, \dots, s_n and that the vector \mathbf{Y} may be taken as multivariate Normal with mean and variance given by

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}, \quad \text{cov}(\mathbf{Y}) = \boldsymbol{\Sigma}(\boldsymbol{\theta}; \mathbf{s}),$$

where \mathbf{X} is given and $\boldsymbol{\beta}$ is a nuisance parameter. The parameters $\boldsymbol{\theta}$ appearing in the covariance function, $\boldsymbol{\Sigma}(\boldsymbol{\theta}; \mathbf{s})$, are the focus of investigation. Show that for any given value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, the statistic

$$S_0 = \mathbf{X}^T \mathbf{W}_0 \mathbf{Y}, \quad \text{where} \quad \mathbf{W}_0^{-1} = \boldsymbol{\Sigma}(\boldsymbol{\theta}_0; \mathbf{s}),$$

is sufficient for β . Show that the conditional log likelihood, $l_{Y|S_0}(\beta, \theta; \theta_0)$, for (β, θ) given S_0 is

$$\begin{aligned} & -\frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)^T \{\Sigma^{-1} - \mathbf{W}_0 \mathbf{X} (\mathbf{X}^T \mathbf{W}_0 \Sigma \mathbf{W}_0 \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_0\} (\mathbf{Y} - \mathbf{X}\beta) \\ & -\frac{1}{2} \log \det \Sigma + \frac{1}{2} \log \det (\mathbf{X}^T \mathbf{W}_0 \Sigma \mathbf{W}_0 \mathbf{X}), \end{aligned}$$

and hence that the reduced function $l^*(\theta) = l_{Y|S}(\beta, \theta; \theta)$ satisfies

$$\begin{aligned} l^*(\beta, \theta) = & -\frac{1}{2} \mathbf{Y}^T \Sigma^{-1} (\mathbf{I} - \mathbf{X} (\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Sigma^{-1}) \mathbf{Y} \\ & -\frac{1}{2} \log \det \Sigma + \frac{1}{2} \log \det (\mathbf{X}^T \Sigma^{-1} \mathbf{X}), \end{aligned}$$

which is independent of β . By considering the special case $\mathbf{X} = \mathbf{I}$ or otherwise, show that $l^*(\beta, \theta)$ is not a log-likelihood function.

7.6 Suppose that Y_1, Y_2 are independent, Normally distributed with means μ_1, μ_2 and unit variances. Let $\psi = \mu_2/\mu_1$. Find the conditional distributions of Y_1, Y_2 and $\psi Y_1 - Y_2$ given $Y_1 + \psi Y_2 = C$. Show that these conditional distributions lead to different ‘likelihoods’.

7.7 Find the conditional maximum-likelihood equations for θ in the previous exercise using equation (7.2). Compare this with the marginal maximum-likelihood estimate based on the residuals as described in section 7.2.1.

7.8 Let $\mathbf{H} = [\mathbf{H}_1 : \mathbf{H}_2]$ be an orthogonal matrix partitioned into \mathbf{H}_1 of order $n \times n-p$ and \mathbf{H}_2 of order $n \times p$. Let Σ_0 be an arbitrary symmetric matrix of rank n . Show that

$$\det \Sigma_0 = \det(\mathbf{H}_1^T \Sigma_0 \mathbf{H}_1) / \det(\mathbf{H}_2^T \Sigma_0^{-1} \mathbf{H}_2).$$

provided that $\det(\mathbf{H}_2^T \Sigma_0^{-1} \mathbf{H}_2) \neq 0$.

7.9 Let $\mathbf{P}_X = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ be a projection matrix of order n and rank p . By writing $\mathbf{I} - \mathbf{P}_X = \mathbf{H} \mathbf{I}_{[n-p]} \mathbf{H}^T$, where \mathbf{H} is orthogonal, show that the matrix

$$\Sigma^* = (\mathbf{I} - \mathbf{P}_X) \Sigma_0 (\mathbf{I} - \mathbf{P}_X)$$

satisfies

$$\log \text{DET}(\Sigma^*) = \log \det \Sigma_0 + \log \det(\mathbf{X}^T \Sigma_0^{-1} \mathbf{X}) - 2 \log \text{DET}(\mathbf{X}),$$

where $\text{DET}(A)$ is defined as the product of the singular values of A .

$$\text{DET}(\mathbf{A}) = \prod_{\lambda \neq 0} \lambda_j(\mathbf{A}).$$

It is assumed here that \mathbf{X} is $n \times p$ with rank p and that Σ_0 is positive definite and symmetric. Thus $\det(\Sigma_0) = \text{DET}(\Sigma_0)$.

7.10 Suppose that $\epsilon_1, \dots, \epsilon_n$ are independent and identically distributed with density $\frac{1}{\sigma} f\left(\frac{\epsilon}{\sigma}\right)$, depending on the unknown parameter σ . Suppose also that the observed values y_1, \dots, y_n satisfy

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \boldsymbol{\epsilon}$$

for fixed known matrices \mathbf{X}, \mathbf{Z} and unknown parameters β, γ . Show that the distribution of $\mathbf{R} = (\mathbf{I} - \mathbf{P}_X)\mathbf{Y}$ does not depend on β . [$\mathbf{P}_X = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$.]

Show also that if the ϵ s are Normally distributed, the marginal likelihood based on \mathbf{R} is identical to the conditional likelihood given $\mathbf{P}_X \mathbf{Y}$.

7.11 Define the projection matrices \mathbf{P} and \mathbf{P}_W by

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad \text{and} \quad \mathbf{P}_W = \mathbf{X}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}$$

where $\mathbf{W} = \Sigma^{-1}$. The corresponding residual vectors are

$$\mathbf{R} = (\mathbf{I} - \mathbf{P})\mathbf{Y} \quad \text{and} \quad \mathbf{R}_W = (\mathbf{I} - \mathbf{P}_W)\mathbf{Y}.$$

Assuming that $\text{cov}(\mathbf{Y}) = \Sigma$, show that

$$\text{cov}(\mathbf{R}_W) = (\mathbf{I} - \mathbf{P}_W)\Sigma \quad (= \Sigma_W),$$

and that

$$\Sigma_W^- = \Sigma^{-1}(\mathbf{I} - \mathbf{P}_W)$$

is the Moore-Penrose inverse of Σ_W . Hence deduce that

$$\Sigma_I^- = (\mathbf{I} - \mathbf{P})\Sigma^{-1}(\mathbf{I} - \mathbf{P}_W) \quad \text{and} \quad \Sigma_W^-$$

are both generalized inverses of $(\mathbf{I} - \mathbf{P})\Sigma(\mathbf{I} - \mathbf{P})$.

7.12 Show, using the results of the previous exercise, that

$$\mathbf{R}^T \boldsymbol{\Sigma}_I^- \mathbf{R} = \mathbf{R}_W^T \boldsymbol{\Sigma}_W^- \mathbf{R}_W = \mathbf{Y}^T \mathbf{W} \mathbf{Y} - \mathbf{Y}^T \boldsymbol{\Sigma}^{-1} \mathbf{P}_W \mathbf{Y}.$$

7.13 Show that $\mathbf{R}_W = (\mathbf{I} - \mathbf{P}_W)\mathbf{R}$. Hence deduce that the log likelihood based on \mathbf{R} is identical to that based on \mathbf{R}_W for fixed \mathbf{W} .

7.14 Show that the simultaneous solution to equations (7.11) and (7.12) can be obtained by iteration using the following steps, beginning with initial estimates $\mu_{ij}^{(1)}$ for the conditional means.

- (i) $\kappa_2^{(r)} = \frac{m_*}{m_* - 1} \left(\frac{1}{\mu_{11}^{(r)}} + \frac{1}{\mu_{12}^{(r)}} + \frac{1}{\mu_{21}^{(r)}} + \frac{1}{\mu_{22}^{(r)}} \right)^{-1}$
- (ii) $\psi^{(r)} = \frac{\mu_{11}^{(r)} \mu_{22}^{(r)} + \kappa_2^{(r)}}{\mu_{12}^{(r)} \mu_{21}^{(r)} + \kappa_2^{(r)}},$
- (iii) $\mu_{11}^{(r+1)} = \mu_{11}^{(r)} + \kappa_2^{(r)} \times \{\log \psi - \log \psi^{(r)}\}.$

Hint: for part (iii) use the property of exponential families that $\partial \mu / \partial \theta = \kappa_2$. If the initial estimates are poorly chosen, care must be taken in step (iii) to ensure that μ_{11} does not get out of range. Otherwise the algorithm seems to converge rapidly. [Liao, 1988].

7.15 Show that the two asymptotic variance formulae (7.27) are not identical but are numerically very similar. Compute both expressions for the data in Table 7.3 and show that the difference is approximately one half of 1%.

7.16 Show that if a particular response category is not used, that category may be deleted without affecting (7.26).

7.17 The estimating equation (7.26) does not have the form specified in (7.25) because the weight $(\zeta_j + \zeta_{j+1})/\zeta_j$ depends on the whole vector \mathbf{S} and not just on S_j . Discuss the possible implications of this.

7.18 Fit a model to the data in Table 7.2 in which the odds ratio is constant up to age 75, but different in the 75+ age-group. Show that this model fits better than the linear regression model in (7.20). Give an approximate significance level for the observed

difference, making due allowance for selection effects. Use the deviance reduction rather than the parameter estimate as your test statistic.

7.19 For the Fieller-Creasy problem discussed at the end of section 7.2.2, in which the parameter of interest is the ratio of Normal means, show that the bias-corrected derivative of $l^*(\psi)$ is

$$\frac{\partial l^*(\psi)}{\partial \psi} - E\left(\frac{\partial l^*(\psi)}{\partial \psi} \mid S_\lambda(\psi), \psi\right) \frac{y_2 - \psi y_1}{1 + \psi^2} \frac{s_\lambda(\psi)}{1 + \psi^2}.$$

Discuss briefly the use of this unbiased estimating equation as an alternative to (7.3). Under what circumstances do the two methods produce identical estimating equations?

CHAPTER 8

Models for data with constant coefficient of variation

8.1 Introduction

The classical linear models introduced in Chapter 3 assume that the variance of the response is constant over the entire range of parameter values. This property is required to ensure that the regression parameters are estimated with maximum precision by ordinary least squares and that consistent estimates are obtained for the variance of $\hat{\beta}$. It is common, however, to find data in the form of continuous measurements where the variance increases with the mean. In Chapter 6 we studied models for common types of data in which $\text{var}(Y) \propto E(Y)$, including continuous measurements as well as discrete data. Here we assume that the coefficient of variation is constant, i.e. that

$$\text{var}(Y) = \sigma^2 \{E(Y)\}^2 = \sigma^2 \mu^2.$$

Note that σ is now the coefficient of variation of Y and not the standard deviation.

For small σ , the variance-stabilizing transformation, $\log(Y)$, has approximate moments

$$E(\log(Y)) = \log(\mu) - \sigma^2/2 \quad \text{and} \quad \text{var}(\log(Y)) \simeq \sigma^2.$$

Further, if the systematic part of the model is multiplicative on the original scale, and hence additive on the log scale, then

$$\eta_i = \log\{E(Y_i)\} = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Then, with the exception of the intercept or constant term in the linear model, consistent estimates of the parameters and of their

precision may be obtained by transforming to the log scale and applying ordinary least squares. The intercept is then biased by approximately $-\sigma^2/2$.

For a number of reasons, and particularly if it is required to present conclusions on the original scale of measurement, it is preferable to retain that scale and not to transform the response. Then we have

$$\mu = E(Y) = \exp(\mathbf{x}^T \boldsymbol{\beta})$$

referring directly to the original scale of measurement. The log link function achieves linearity without abandoning the preferred scale, and a quadratic variance function describes the relationship between $\text{var}(Y)$ and $E(Y)$. With this combination of link and variance function, iteratively weighted non-linear least squares may be used to obtain estimates for $\boldsymbol{\beta}$ using the algorithm described in Chapter 1. This method of estimation is equivalent to assuming that Y has the gamma distribution with constant index $\nu = 1/\sigma^2$ independent of the mean in the same sense that ordinary least squares arises as maximum likelihood for the Normal distribution.

In comparing the two methods of estimation described above, we assume that the precise distribution of Y is not specified, for if it is the comparison can be uniquely resolved by standard efficiency calculations and by considerations such as sufficiency. For example, if Y has the log-normal distribution the first method is preferred, while if it has the gamma distribution the second method is preferred. More generally, however, if Y is a variable with a physical dimension or if it is an extensive variable (Cox and Snell, 1981, p. 14) such that a sum of Y 's has some well-defined physical meaning, the method of analysis based on transforming to $\log Y$ is unsatisfactory on scientific grounds and the second method of analysis would be preferred. However, if the analysis is exploratory or if only graphical presentation is required, transformation of the data is convenient and indeed desirable.

Firth (1988) gives a comparison of the efficiencies of the gamma model when the errors are in fact log-Normal with the log-Normal model when the errors have a gamma distribution. He concludes that the gamma model performs slightly better under reciprocal misspecification.

8.2 The gamma distribution

For our present purposes it is most convenient to write the gamma density in the form

$$\frac{1}{\Gamma(\nu)} \left(\frac{\nu y}{\mu} \right)^\nu \exp\left(-\frac{\nu y}{\mu}\right) d(\log y); \quad y \geq 0, \nu > 0, \mu > 0.$$

For brevity we write $Y \sim G(\mu, \nu)$. From its cumulant generating function, $-\nu \log(1 - \mu t/\nu)$, the first four cumulants are easily found as

$$\begin{aligned}\kappa_1 &= E(Y) = \mu, \\ \kappa_2 &= \text{var}(Y) = \mu^2/\nu, \\ \kappa_3 &= E(Y - \mu)^3 = 2\mu^3/\nu^2, \\ \kappa_4 &= 6\mu^4/\nu^3.\end{aligned}$$

More generally $\kappa_r = (r-1)! \mu^r / \nu^{r-1}$. The value of ν determines the shape of the distribution. If $0 < \nu < 1$ the density has a pole at the origin and decreases monotonically as $y \rightarrow \infty$. The special case $\nu = 1$ corresponds to the exponential distribution. If $\nu > 1$ the density is zero at the origin and has a single mode at $y = \mu - \mu/\nu$; however, the density with respect to the differential element $d(\log y)$ has a maximum at $y = \mu$ for all ν . Fig. 8.1 shows the form of the distribution for $\nu = 0.5, 1.0, 2.0$ and 5.0 with $\mu = 1$ held constant. It can be seen from the graphs that the densities are all positively skewed. The standardized skewness coefficient is $\kappa_3/\kappa_2^{3/2} = 2\nu^{-1/2}$, and a Normal limit is attained as $\nu \rightarrow \infty$.

In this chapter we are concerned mostly with models for which the index or precision parameter $\nu = \sigma^{-2}$ is assumed constant for all observations, so that the densities all have the same shape. However, by analogy with weighted linear least squares, where the variances are proportional to known constants, we may, in the context of the gamma distribution, allow ν to vary in a similar manner from one observation to another. In other words we may have $\nu_i = \text{constant} \times w_i$, where w_i are known weights and ν_i is the index or precision parameter of Y_i . Problems of this form occur in estimating variance components where the observations are sums of squares of Normal variables, the weights are one half of their degrees of freedom and the proportionality constant is 1.

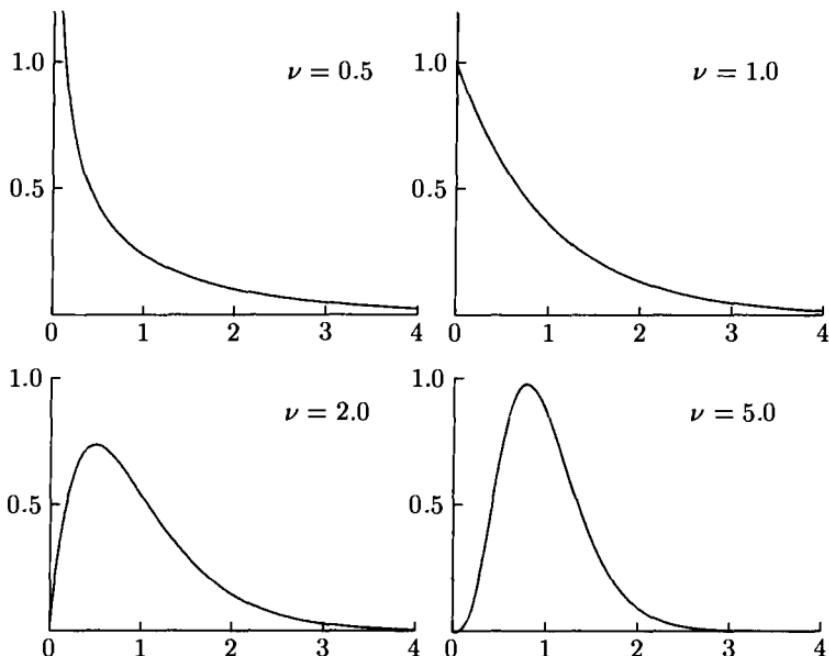


Fig. 8.1. The gamma distribution for $\nu = 0.5, 1.0, 2.0$ and 5.0 , $\mu = 1$.

The densities then have different shapes. For further details see section 8.3.5.

The gamma family, and indeed all distributions of the type discussed in section 2.2, are closed under convolutions. Thus if Y_1, \dots, Y_n are independent and identically distributed in the gamma distribution with index ν , then the arithmetic mean \bar{Y} is distributed in the same family with index $n\nu$. Thus the gamma distribution with integer index, sometimes also called the Erlangian distribution (Cox, 1962), arises in a fairly natural way as the time to the ν th event in a Poisson process.

The log-likelihood function corresponding to a single observation is shown in Fig. 8.2 where we plot the log likelihood against μ , $\log \mu$, $\mu^{-1/3}$ and μ^{-1} . It can be seen that the log-likelihood function is nearly quadratic on the inverse cube-root scale; the log likelihood at μ differs from the value at the maximum by an amount closely approximated by

$$9y^{\frac{2}{3}}(y^{-\frac{1}{3}} - \mu^{-\frac{1}{3}})^2/2.$$

Now it is known that the square root of twice the log-likelihood-ratio statistic is approximately Normally distributed. Thus an

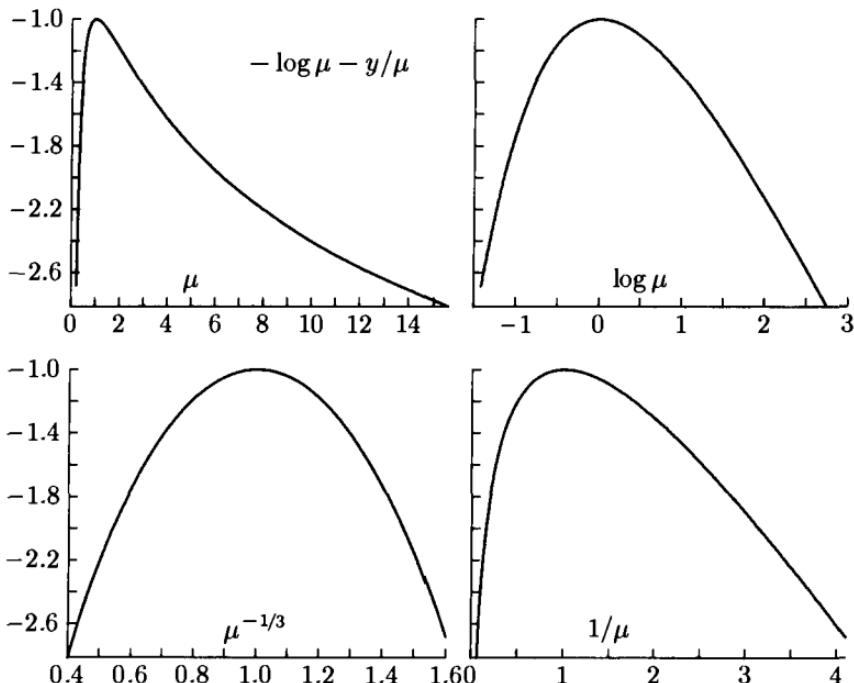


Fig. 8.2. The gamma log likelihood for $y = 1$, plotted against μ , $\log \mu$, $\mu^{-1/3}$ and $1/\mu$.

accurate Normalizing transformation for Y is

$$3\{(Y/\mu)^{\frac{1}{3}} - 1\}.$$

The cube-root transform was originally derived in this context by Wilson and Hilferty (1931) (see also Hougaard, 1982).

8.3 Models with gamma-distributed observations

8.3.1 The variance function

We have already noted that, with the parameterization of the gamma distribution used here, the variance function is quadratic. This result can be obtained directly by writing the log likelihood as a function of both ν and μ in the standard form

$$\nu(-y/\mu - \log \mu) + \nu \log y + \nu \log \nu - \log \Gamma(\nu).$$

It follows in terms of the parameterization used in Chapter 2 that $\theta = -1/\mu$ is the canonical parameter, and $b(\theta) = -\log(-\theta)$ is the cumulant function. From these the mean $b'(\theta) = \mu$ and variance function $b''(\theta) = \mu^2$ may be derived.

8.3.2 The deviance

Taking ν to be a known constant, the log likelihood may be written as

$$\sum_i \nu(-y_i/\mu_i - \log \mu_i)$$

for independent observations. If the index is not constant but is proportional to known weights, $\nu_i = \nu w_i$, the log likelihood is equal to

$$\nu \sum w_i(-y_i/\mu_i - \log \mu_i).$$

The maximum attainable log likelihood occurs at $\boldsymbol{\mu} = \mathbf{y}$, and the value attained is $-\nu \sum w_i(1 + \log y_i)$, which is finite unless $y_i = 0$ for some i . The deviance, which is proportional to twice the difference between the log likelihood achieved under the model and the maximum attainable value, is

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = -2 \sum w_i \{ \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)/\hat{\mu}_i \}.$$

This statistic is defined only if all the observations are strictly positive. More generally, if some components of \mathbf{y} are zero we may replace $D(\mathbf{y}; \boldsymbol{\mu})$ by

$$D^+(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2C(\mathbf{y}) + 2 \sum w_i \log \hat{\mu}_i + 2 \sum w_i y_i / \hat{\mu}_i,$$

where $C(\mathbf{y})$ is an arbitrary bounded function of \mathbf{y} . The only advantage of $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ over $D^+(\mathbf{y}; \hat{\boldsymbol{\mu}})$ is that the former function is always positive and behaves like a residual sum of squares. Note, however, that the maximum-likelihood estimate of ν is a function of $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ and not of $D^+(\mathbf{y}; \hat{\boldsymbol{\mu}})$. Furthermore, if any component of \mathbf{y} is zero then $\hat{\nu} = 0$. This is clearly not a desirable feature of the maximum-likelihood estimator in most applications if only because rounding errors may produce spurious zeros. Alternative estimators are given in section 8.3.6.

The final term in the expression for $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ is identically zero provided that the model formula contains an intercept term. In such cases the final term can be ignored (Nelder and Wedderburn, 1972). Under the same conditions, the final term in $D^+(\mathbf{y}; \hat{\boldsymbol{\mu}})$ is equal to $\sum w_i$ and can be absorbed into $C(\mathbf{y})$.

8.3.3 The canonical link

The canonical link function yields sufficient statistics which are linear functions of the data and it is given by

$$\eta = \mu^{-1}.$$

Unlike the canonical links for the Poisson and binomial distributions, the reciprocal transformation, which is often interpretable as the rate of a process, does not map the range of μ onto the whole real line. Thus the requirement that $\eta > 0$ implies restrictions on the β s in any linear model. Suitable precautions must be taken in computing $\hat{\boldsymbol{\beta}}$ so that negative values of $\hat{\mu}$ are avoided.

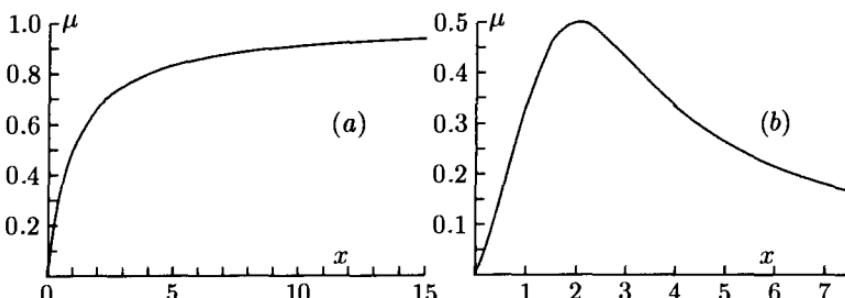


Fig. 8.3. Inverse polynomials: (a) the inverse linear, $\mu^{-1} = 1 + x^{-1}$; (b) the inverse quadratic, $\mu^{-1} = x - 2 + 4/x$.

An example of the canonical link is given by the inverse polynomial response surfaces discussed by Nelder (1966). The simplest case, that of the inverse linear response, is given by

$$\eta = \beta_0 + \beta_1/x \quad \text{with } x > 0.$$

In plant density experiments it is commonly observed that the yield per plant varies inversely with plant density x , so that the mean

yield per plant has the form $1/(\beta_1 + \beta_0 x)$. The yield per unit area is then given by

$$\eta^{-1} = \mu = \frac{x}{\beta_0 x + \beta_1},$$

giving a hyperbolic form for μ against x , with a slope at the origin of $1/\beta_1$ and an asymptote at $\mu = 1/\beta_0$. Inclusion of a linear term in x , gives

$$\eta = \beta_1/x + \beta_0 + \gamma_1 x,$$

which is called the inverse quadratic. Both curves have a slope at the origin of $1/\beta_1$. The inverse quadratic response reaches a maximum of $\mu = \beta_0 + 2\sqrt{(\beta_1 \gamma_1)}$ at $x = \sqrt{(\beta_1 / \gamma_1)}$ corresponding to the optimum plant density. At higher plant densities μ tends to zero like $1/(\gamma_1 x)$ as shown in Fig. 8.3.

The surfaces can be extended to include more than one covariate and by the inclusion of cross-terms in $1/(x_1 x_2)$, x_1/x_2 , and so on. For positive values of the parameters the surfaces have the desirable property that the ordinate η is everywhere positive and bounded; this is in contrast to ordinary polynomials where the ordinate is unbounded at the extremes and often takes negative values.

In practice we often require to fit origins for the covariates, i.e. to make x enter the inverse polynomial in the form $x_0 + x$, where x_0 has to be estimated. The baseline value x_0 is non-linear in a general sense and its estimation requires special treatment—see Chapter 11 for details.

Two other link functions are important for generalized linear models with gamma errors, the log and the identity, and we now consider their uses.

8.3.4 Multiplicative models: log link

By combining the log link with terms linear in x and $1/x$ a large variety of qualitatively distinct response functions can be generated. Four of these are shown in Fig. 8.4, where we have shown $\eta = \log \mu = 1 \pm x \pm 1/x$ for $x > 0$. These curves are sometimes useful for describing response functions that have horizontal or vertical asymptotes, or functions that have turning points but are noticeably asymmetric about that point.

We noted in section 8.1 the close connection between linear models with constant variance for $\log Y$ and multiplicative models

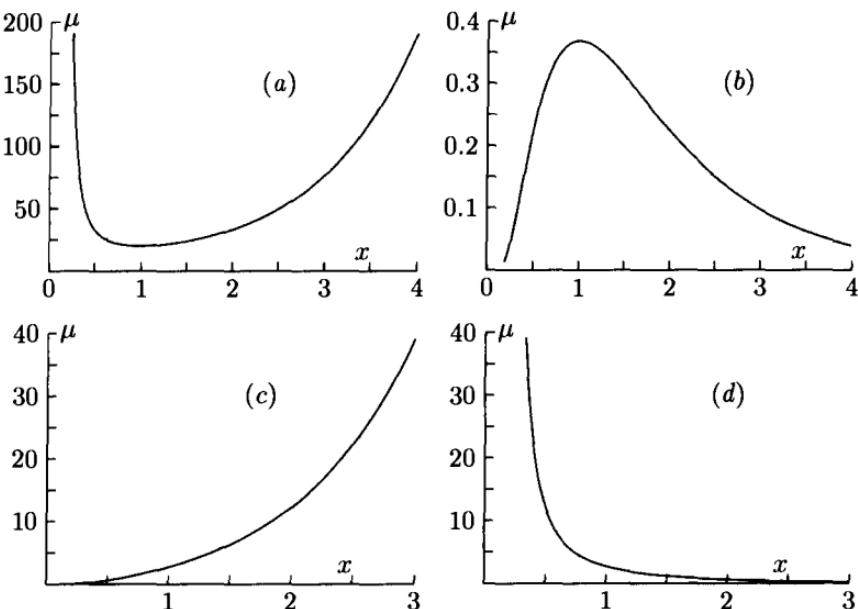


Fig. 8.4. Plots of various logarithmic functions having asymptotes:
 (a) $\log(\mu) = 1 + x + 1/x$, (b) $\log(\mu) = 1 - x - 1/x$,
 (c) $\log(\mu) = 1 + x - 1/x$, (d) $\log(\mu) = 1 - x + 1/x$.

with constant coefficient of variation for Y . Suppose that σ^2 is sufficiently small so that $\text{var}(\log Y) = \sigma^2 = \text{var}(Y)/\mu^2$. In a linear model for $\log Y$ the covariance matrix of the parameter estimates is $\sigma^2(\mathbf{X}^T \mathbf{X})^{-1}$, where \mathbf{X} is the model matrix. For the corresponding multiplicative model the quadratic weight function is exactly unity, giving $\text{cov}(\hat{\boldsymbol{\beta}}) \simeq \sigma^2(\mathbf{X}^T \mathbf{X})^{-1}$ as before. In particular, if \mathbf{X} is the incidence matrix corresponding to an orthogonal design, so that parameter estimates in the Normal-theory linear model are independent, then the corresponding parameter estimates in the gamma-theory multiplicative model are independent asymptotically. This property of approximate independence holds for all generalized linear models whenever the link function is the same as the variance-stabilizing transform.

The preceding analysis and the discussion in section 8.1 indicate that for small σ^2 it is likely to be difficult to discriminate between Normal-theory linear models for $\log Y$ and gamma-theory multiplicative models for Y . Atkinson's (1982) work confirms this assertion even for σ^2 as large as 0.6.

8.3.5 Linear models: identity link

Sums of squares of independent Normal random variables have the chi-squared or, equivalently, the gamma distribution with known index $w = (\text{degrees of freedom})/2$. One method of estimating variance components is to equate the observed mean squares y_i to their expectations which are linear functions of the unknown variance components. Thus

$$\mu_i = E(Y_i) = \sum x_{ij}\beta_j,$$

where x_{ij} are known coefficients and β_j are the variance components. Furthermore if the original data were Normally distributed,

$$\text{var}(Y_i) = \mu_i^2/w_i,$$

where w_i are known weights equal to one-half the degrees of freedom of Y_i . The preceding analysis can equally well be based on sums of squares rather than on mean squares; the coefficients x_{ij} would then be replaced by $2w_i x_{ij}$ and weights would be w_i because the coefficient of variation is unaffected by multiplication of the data by a constant.

If the number of variance components is the same as the number of mean squares, which is commonly the case, the estimating equations may be solved directly by inverting the set of linear equations. The method described above, based on the gamma likelihood, is required only when the number of independent mean squares exceeds the number of variance components. A further advantage of this procedure is that approximate asymptotic variances can be obtained for the estimated variance components. Unfortunately, the sizes of some of these variances often shows that the corresponding estimates are almost worthless. Normal-theory approximations for the distribution of $\hat{\beta}$ s are usually very poor.

The analysis given above is based on the assumption that the mean-square variables are independent and that the original data were Normally distributed. Furthermore, negative estimates of variance components are not explicitly ruled out; these may, however, sometimes be interpretable (Nelder, 1977). In this respect weighted least squares is technically different from maximum likelihood, which does not permit negative variance components. If the weighted least-squares estimates turn out to be negative the

likelihood function attains its maximum on the boundary of the parameter space corresponding to a zero variance component. The two methods coincide only if the weighted least-squares estimates are non-negative.

8.3.6 Estimation of the dispersion parameter

The approximate covariance matrix of the parameter estimates is $\text{cov}(\hat{\beta}) \simeq \sigma^2(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$, where

$$\mathbf{W} = \text{diag}\{(d\mu_i/d\eta_i)^2/V(\mu_i)\}$$

is the $n \times n$ diagonal matrix of weights, \mathbf{X} is the $n \times p$ model matrix and σ is the coefficient of variation. If σ^2 is known, the covariance matrix of $\hat{\beta}$ may be computed directly; usually, however, it must be estimated from the residuals. Under the gamma model the maximum-likelihood estimate of $\nu = \sigma^{-2}$ is given by

$$2n\{\log \hat{\nu} - \psi(\hat{\nu})\} = D(\mathbf{y}; \hat{\mu}), \quad (8.1)$$

where $\psi(\nu) = \Gamma'(\nu)/\Gamma(\nu)$. A suggested improvement to take account of the fact that p parameters have been estimated is to replace the l.h.s. of the above equation by

$$2n\{\log \hat{\nu} - \psi(\hat{\nu})\} - p\hat{\nu}^{-1}, \quad (8.2)$$

the correction being the $O(1)$ term in an asymptotic expansion for $E(D(\mathbf{Y}; \hat{\mu}))$. There is a clear analogue here with the Normal-theory estimates of variance, $\hat{\sigma}^2$ and s^2 . If ν is sufficiently large, implying σ^2 sufficiently small, we may expand (8.1) and (8.2) ignoring terms of order ν^{-2} or smaller. The maximum-likelihood estimate is then approximately

$$\hat{\nu}^{-1} \simeq \frac{\bar{D}(6 + \bar{D})}{6 + 2\bar{D}}$$

where $\bar{D} = D(\mathbf{y}; \hat{\mu})/n$. A similar approximation can be made for the bias-corrected estimate: see Exercises 8.11 and 8.12. For further approximations, see the paper by Greenwood and Durand (1960) and a series of papers by Bain and Engelhardt (1975, 1977).

The principal problem with the maximum-likelihood estimator, and in fact with any estimator based on $D(\mathbf{y}; \hat{\mu})$, is that it is

extremely sensitive to rounding errors in very small observations and in fact $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ is infinite if any component of \mathbf{y} is zero. Equally important is the fact that if the gamma assumption is false, $\hat{\nu}^{-1}$ does not consistently estimate the coefficient of variation. For these reasons we prefer the moment estimator

$$\tilde{\sigma}^2 = \sum \{(y - \hat{\mu})/\hat{\mu}\}^2/(n - p) = X^2/(n - p), \quad (8.3)$$

which is consistent for σ^2 , provided of course that $\boldsymbol{\beta}$ has been consistently estimated. This estimator for σ^2 may be used in the formula $\sigma^2(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$ to obtain an estimate of $\text{cov}(\hat{\boldsymbol{\beta}})$. Note that, unlike the usual Normal-theory estimator of variance s^2 , the bias of $\tilde{\sigma}^2$ is $O(n^{-1})$ even if the data are distributed according to the gamma distribution. The divisor $n - p$ is preferable to n but is not sufficient to remove the $O(n^{-1})$ bias. For a single gamma sample the expected value of $\tilde{\sigma}^2$ is

$$\sigma^2[1 - \sigma^2/n + O(n^{-2})].$$

The negative bias is a consequence of the fact that $V''(\mu) > 0$.

8.4 Examples

8.4.1 Car insurance claims

The data given in Table 8.1, taken from Baxter *et al.* (1980, Table 1), give the average claims for damage to the owner's car for privately owned and comprehensively insured vehicles. Averages are given in pounds sterling adjusted for inflation. The number of claims on which each average is based is given in parallel. Three factors thought likely to affect the average claim are:

1. policyholder's age (PA), with eight levels, 17–20, 21–24, 25–29, 30–34, 35–39, 40–49, 50–59, 60+;
2. car group (CG), with four levels, A, B, C and D;
3. vehicle age (VA), with four levels, 0–3, 4–7, 8–9, 10+.

The numbers of claims m_{ijk} on which each average is based vary widely from zero to a maximum of 434. Since the precision of each average Y_{ijk} , whether measured by the variance or by the squared coefficient of variation, is proportional to the corresponding m_{ijk} ,

these numbers appear as weights in the analysis. This means that the five accidentally empty cells, (1,3,4), (1,4,3), (1,4,4), (2,4,4) and (5,4,4) for which $m = 0$, are effectively left out of the analysis. For computational purposes, however, it is usually more convenient to retain these cells as observations with zero weight, so that they make no contribution to the likelihood. The structure of the factors is then formally that of a complete crossed design.

Baxter *et al.* analyse the data using a weighted Normal-theory linear model with weights m_{ijk} and the three main effects PA + CG + VA. Here we reanalyse the data, making the assumption that the coefficient of variation rather than the variance is constant across cells. In addition, we make the assumption that the systematic effects are linear on the reciprocal scale rather than on the untransformed scale. Justification for these choices is given in Chapters 10 and 11. The model containing main effects only may be written

$$\mu_{ijk} = E(Y_{ijk}) = (\mu_0 + \alpha_i + \beta_j + \gamma_k)^{-1},$$

$$\text{var}(Y_{ijk}) = \sigma^2 \mu_{ijk}^2 / m_{ijk},$$

where α_i , β_j and γ_k are the parameters corresponding to the three classifying factors PA, CG and VA. One way of interpreting the reciprocal transform is to think of $\eta_{ijk} = 1/\mu_{ijk}$ as the rate at which instalments of £1 must be paid to service an average claim over a fixed period of one time unit. In other words, η_{ijk} is the time interval between instalments or the time purchased by an instalment of £1 in servicing an average claim in cell (i, j, k) .

One sequence of models yielded the goodness-of-fit statistics shown in Table 8.2. Using the result that first differences of the deviance have, under the appropriate hypothesis, an approximate scaled chi-squared distribution, it is clear that the model with main effects only provides a reasonable fit and that the addition of two-factor interactions yields no further explanatory power. The estimate of $\tilde{\sigma}^2$ based on the residuals from the main-effects model is

$$\tilde{\sigma}^2 = \frac{1}{109} \sum m(y - \hat{\mu})^2 / \hat{\mu}^2 = 1.21,$$

so that the estimated coefficient of variation of the individual claims is $\tilde{\sigma} = 1.1$. Estimates based on the deviance give very similar values. An examination of approximately standardized residuals

Table 8.1 *Average cost of claims for own damage (adjusted for inflation) for privately owned, comprehensively insured cars in 1975*

Policy-holder's age	Car group	Vehicle age							
		0-3		4-7		8-9		10+	
		£	No.	£	No.	£	No.	£	No.
17-20	A	289	8	282	8	133	4	160	1
	B	372	10	249	28	288	1	11	1
	C	189	9	288	13	179	1	—	0
	D	763	3	850	2	—	0	—	0
21-24	A	302	18	194	31	135	10	166	4
	B	420	59	243	96	196	13	135	3
	C	268	44	343	39	293	7	104	2
	D	407	24	320	18	205	2	—	0
25-29	A	268	56	285	55	181	17	110	12
	B	275	125	243	172	179	36	264	10
	C	334	163	274	129	208	18	150	8
	D	383	72	305	50	116	6	636	1
30-34	A	236	43	270	53	160	15	110	12
	B	259	179	226	211	161	39	107	19
	C	340	197	260	125	189	30	104	9
	D	400	104	349	55	147	8	65	2
35-39	A	207	43	129	73	157	21	113	14
	B	208	191	214	219	149	46	137	23
	C	251	210	232	131	204	32	141	8
	D	233	119	325	43	207	4	—	0
40-49	A	254	90	213	98	149	35	98	22
	B	218	380	209	434	172	97	110	59
	C	239	401	250	253	174	50	129	15
	D	387	199	299	88	325	8	137	9
50-59	A	251	69	227	120	172	42	98	35
	B	196	366	229	353	164	95	132	45
	C	268	310	250	148	175	33	152	13
	D	391	105	228	46	346	10	167	1
60+	A	264	64	198	100	167	43	114	53
	B	224	228	193	233	178	73	101	44
	C	269	183	258	103	227	20	119	6
	D	385	62	324	22	192	6	123	6

Table 8.2 Goodness-of-fit statistics for a sequence of models fitted to the car insurance data (error gamma; link reciprocal)

Model	Deviance	First difference	d.f.	Mean deviance
1	649.9			
PA	567.7	82.2	7	11.7
PA + CG	339.4	228.3	3	76.1
PA + CG + VA	124.8	214.7	3	71.6
+ PA · CG	90.7	34.0	21	1.62
+ PA · VA	71.0	19.7	21	0.94
+ CG · VA	65.6	5.4	9	0.60
Complete	0.0	65.6	58	1.13

using the formula $\sqrt{m}(y - \hat{\mu})/(\hat{\sigma}\hat{\mu})$ shows the six most extreme residuals as corresponding to observations (2,2,1), (3,2,4), (3,4,4), (5,1,2), (5,4,1) and (7,2,1) with values 3.4, 3.2, 2.8, -2.6, -2.2 and -2.5. The corresponding standardized deviance residuals are 3.0, 2.4, 1.8, -3.0, -2.3 and -2.7. The positions of these cells do not show any obvious pattern, and the magnitudes of the most extreme residuals are only moderately large in view of the sample size, which is effectively 109. With a Normal sample of this size one expects the most extreme standardized residuals to be about ± 2.5 .

Table 8.3 Parameter estimates and standard errors ($\times 10^6$) on reciprocal scale for main effects in car insurance example

Level	Age group (PA)	Car group (CG)	Vehicle age (VA)
1	0 (-)	0 (-)	0 (-)
2	101 (436)	38 (169)	336 (101)
3	350 (412)	-614 (170)	1651 (227)
4	462 (410)	-1421 (181)	4154 (442)
5	1370 (419)		
6	970 (405)		
7	916 (408)		
8	920 (416)		

Parameter estimates for the main-effects model are given in

Table 8.3. Standard errors are based on the estimate $\tilde{\sigma} = 1.1$. The estimate for the intercept corresponding to all factors at their lowest level is 3410×10^{-6} . Bearing in mind that the analysis is performed here on the reciprocal scale and that a large positive parameter corresponds to a small claim, we may deduce the following. The largest average claims are made by policyholders in the youngest four age groups, i.e. up to age 34, the smallest average claims by those aged 35–39, and intermediate claims by those aged 40 and over. These effects are in addition to effects due to type of vehicle and vehicle age. The value of claims decreases with car age, although not linearly. There are also marked differences between the four car groups, group D being the most expensive and group C intermediate. No significant difference is discernible between car groups A and B.

It should be pointed out that the parameter estimates given here are contrasts with level 1. In a balanced design the three sets of estimates corresponding to the three factors would be uncorrelated while the correlations within a factor would be 0.5. Even where, as here, there is considerable lack of balance, the correlations do not deviate markedly from these values.

It is possible to test and quantify the assertions made above by fusing levels 1–4, levels 6–8 of PA and levels 1 and 2 of CG. The deviance then increases to 129.8 on 116 d.f., which is a statistically insignificant increase.

The preceding analysis is not the only one possible for these data. In fact a multiplicative model corresponding to a logarithmic link function would lead to similar qualitative conclusions. As is shown in Chapter 10, the data themselves support the reciprocal model better but only marginally so, and it might be argued that quantitative conclusions for these data would be more readily stated and understood for a multiplicative model.

8.4.2 Clotting times of blood

Hurn *et al.* (1945) published data on the clotting time of blood, giving clotting times in seconds (y) for normal plasma diluted to nine different percentage concentrations with prothrombin-free plasma (u); clotting was induced by two lots of thromboplastin. The data are shown in Table 8.4. A hyperbolic model for lot 1 was fitted by Bliss (1970), using an inverse transformation of the data,

and for both lots 1 and 2 using untransformed data. We analyse both lots using the inverse link and gamma errors.

Initial plots suggest that a log scale for u is needed to produce inverse linearity, and that both intercepts and slopes are different for the two lots. This claim is confirmed by fitting the following model sequence:

<i>Model</i>	<i>Deviance</i>	<i>d.f.</i>
1	7.709	17
X	1.018	16
$L + X$	0.300	15
$L + L.X$	0.0294	14

Here $x = \log u$ and L is the factor defining the lots. Clearly all the terms are necessary and the final model produces a mean deviance whose square root is 0.0458, implying approximately a 4.6% standard error on the y -scale. The two fitted lines, with standard errors for the parameters shown in parentheses, are

$$\text{lot 1: } \hat{\mu}^{-1} = -0.01655(\pm 0.00086) + 0.01534(\pm 0.00143)x$$

$$\text{lot 2: } \hat{\mu}^{-1} = -0.02391(\pm 0.00038) + 0.02360(\pm 0.00062)x$$

The plot of the Pearson residuals $(y - \hat{\mu})/\hat{\mu}$ against the linear predictor $\hat{\eta}$ is satisfactory, and certainly better than either (i) the use of constant variance for Y where the residual range decreases with $\hat{\eta}$ or (ii) the use of constant variance for $1/Y$ where the analogous plot against $\hat{\mu}$ shows the range increasing with $\hat{\mu}$. Note that constant variance for $1/Y$ implies (to the first order) $\text{var}(Y) \propto \mu^4$. Thus the assumption of gamma errors (with $\text{var}(Y) \propto \mu^2$) is ‘half-way’ between assuming $\text{var}(Y)$ constant and $\text{var}(1/Y)$ constant.

The estimates suggest that the parameters for lot 2 are a constant multiple (about 1.6) of those for lot 1. If true this would mean that $\boldsymbol{\mu}_2 = k\boldsymbol{\mu}_1$, where the suffix denotes the lot. This model, though not a generalized linear model, has simple maximum-likelihood equations for estimating α, β and k where

$$\begin{aligned}\boldsymbol{\mu}_1 &= 1/\boldsymbol{\eta}_1, & \boldsymbol{\eta}_1 &= \alpha + \beta \mathbf{x}, \\ \boldsymbol{\mu}_2 &= k\boldsymbol{\mu}_1.\end{aligned}$$

Table 8.4 *Mean clotting times in seconds (y) of blood for nine percentage concentrations of normal plasma (u) and two lots of clotting agent*

u	Clotting time	
	Lot 1	Lot 2
5	118	69
10	58	35
15	42	26
20	35	21
30	27	18
40	25	16
60	21	13
80	19	12
100	18	12

These are equivalent to fitting α and β to data y_1 and y_2/k , combined with the equation $\sum(y_2/\mu_1 - k) = 0$. The resulting fit gives $\hat{k} = 0.625$ with deviance = 0.0332 and having 15 d.f. Comparing this with the fit of separate lines gives a difference of deviance of 0.0038 on one degree of freedom against a mean deviance of 0.0021 for the more complex model. The simpler model of proportionality is not discounted, with lot 2 giving times about five-eighths those of lot 1.

8.4.3 Modelling rainfall data using two generalized linear models

Histograms of daily rainfall data are usually skewed to the right with a ‘spike’ at the origin. This form of distribution suggests that such data might be modelled in two stages, one stage being concerned with the pattern of occurrence of wet and dry days, and the other with the amount of rain falling on wet days. The first stage involves discrete data and can often be modelled by a stochastic process in which the probability of rain on day t depends on the history of the process up to day $t - 1$. Often, first-order dependence corresponding to a Markov chain provides a satisfactory model. In the second stage we require a family of densities on the positive line for the quantity of rainfall. To be realistic, this family of densities should be positively skewed and should have variance increasing with μ . The gamma distribution

has been found appropriate in this context, although the log-Normal distribution is also widely used.

(a) *Modelling the frequency of wet days.* Coe and Stern (1982, 1984) describe the application of generalized linear models and give references to earlier work. The data for n years form an $n \times 365$ table of rainfall amounts. (We ignore the complications introduced by leap years.) Considering the years as replicates, each of the n observations of day t is classified by the double dichotomy dry/wet and previous day dry/previous day wet. Combining over replicates we obtain, for each of the 365 days, a 2×2 table of frequencies having the form of Table 8.5.

Table 8.5 *The 2×2 table of frequencies for rainfall data on day t*

		<i>Today</i>		<i>Total</i>
		<i>Wet</i>	<i>Dry</i>	
<i>Yesterday</i>	<i>Wet</i>	y_0	$n_0 - y_0$	n_0
	<i>Dry</i>	y_1	$n_1 - y_1$	n_1
	<i>Total</i>	y_\cdot	$n_\cdot - y_\cdot$	$n_\cdot = n$

Let $\pi_0(t)$ be the probability that day t is wet given that day $t - 1$ was wet: $\pi_1(t)$ is the corresponding probability given that day $t - 1$ was dry. Ignoring end effects, the likelihood for the first-order Markov model is the product over t of terms having the form

$$\pi_0(t)^{y_0} [1 - \pi_0(t)]^{n_0 - y_0} \pi_1(t)^{y_1} [1 - \pi_1(t)]^{n_1 - y_1}.$$

In other words each 2×2 table corresponds to two independent binomial observations in which the row totals are regarded as fixed.

Note that in the above 2×2 table for day t , n_0 is the number of occasions on which rain fell on day $t - 1$ in the years $1, \dots, n$, whereas y_\cdot is the number of occasions on which rain fell on day t . Evidently therefore, in an obvious extension of the notation, $n_0(t+1) = y_\cdot(t)$.

If the target parameter were the difference between $\pi_0(t)$ and $\pi_1(t)$, say

$$\psi(t) = \pi_0(t)[1 - \pi_1(t)] / \{\pi_1(t)(1 - \pi_0(t))\},$$

it would often be preferable to construct a likelihood function depending on $\psi(t)$ alone. The hypergeometric likelihood described in section 7.4 can then be used. In this application, however, we would usually be interested in models for $\pi_0(t)$ and $\pi_1(t)$ themselves and not just in the difference between them.

Coe and Stern use linear logistic models with various explanatory terms. Obvious choices for cyclical terms are the harmonics $\sin(2\pi t/365)$, $\cos(2\pi t/365)$, $\sin(4\pi t/365)$, $\cos(4\pi t/365)$, and so on. The simplest model corresponding to the first harmonic would be

$$\begin{aligned}\text{logit}(\pi_0(t)) &= \alpha_0 + \alpha_{01} \sin(2\pi t/365) + \beta_{01} \cos(2\pi t/365), \\ \text{logit}(\pi_1(t)) &= \alpha_1 + \alpha_{11} \sin(2\pi t/365) + \beta_{11} \cos(2\pi t/365),\end{aligned}$$

which involves six parameters. Note that if the coefficients of the harmonic terms are equal ($\alpha_{01} = \alpha_{11}$, $\beta_{01} = \beta_{11}$), the odds ratio in favour of wet days is constant over t .

If there is a well defined dry season, a different scale corresponding to some fraction of the year might be more appropriate.

Various extensions of these models are possible: a second-order model would take account of the state of the two previous days, producing four probabilities to be modelled. If it is suspected that secular trends over the years are present it is important not to regard the years as replicates. Instead, we would regard the data as $356n$ Bernoulli observations indexed by day, year and previous day wet/dry. The computational burden is increased but no new theoretical problems are involved.

(b) *Modelling the rainfall on wet days.* Coe and Stern use a multiplicative model with gamma-distributed observations to model the rainfall on wet days. The idea is to express $\log[\mu(t)]$ as a linear function involving harmonic components. Here $\mu(t)$ is the mean rainfall on day t conditional on day t being wet. If it is assumed that no secular trends are involved, the analysis requires only the means for each day of the period with the sample sizes entering the analysis as weights. The introduction of secular trends involves the use of individual daily values in the analysis. The assumption of constant coefficient of variation requires checking. The simplest way is to group the data into intervals based on the value of $\hat{\mu}$ and to estimate the coefficient of variation in each interval. Plots against $\hat{\mu}$ should reveal any systematic departure from constancy.

(c) *Some results.* Coe and Stern present the results of fitting the models described above to data from Zinder in Niger spanning 30 years. The rainy season lasts about four months so that data are restricted to 120 days of the year. Table 8.6 shows the results of fitting separate Fourier series to $\pi_0(t)$ and $\pi_1(t)$ in a first-order Markov chain. Each new term adds four parameters to the model, a sine and cosine term for each π .

Table 8.6 *Analysis of deviance for a first-order Markov chain. Rainy days in rainfall data from Niger*

<i>Model</i>	<i>Deviance</i>	<i>First difference</i>	<i>d.f.</i>	<i>Mean deviance</i>
Intercept	483.1	—	238	
+1st harmonic	260.9	222.2	4	55.6
+2nd harmonic	235.6	25.3	4	6.3
+3rd harmonic	231.4	4.2	4	1.05
+4th harmonic	227.7	3.7	4	0.9

By including a sufficient number of harmonic terms in the model, the mean deviance is reduced to a value close to unity, indicating a satisfactory fit. The reductions for the third and fourth harmonic are clearly insignificant. Thus a first-order non-stationary Markov chain, with two harmonic terms for the time-dependence of the transition probabilities, is adequate for these data. This model contains a total of 10 parameters for the transition probabilities.

Table 8.7 *Analysis of deviance of rainfall amounts. Data from Niger*

<i>Model</i>	<i>Deviance</i>	<i>d.f.</i>	<i>Mean deviance</i>
Constant	224.6	119	
+1st harmonic	154.5	117	
+2nd harmonic	147.0	115	1.28
Within days	1205.0	946	1.27

The results of fitting models with gamma errors and log link for the rainfall amounts are shown in Table 8.7. Again two harmonics suffice in the sense that their inclusion reduces the between-day deviance to that within days. The mean deviance within days over years constitutes a baseline for the analysis between days, and its

size (little more than 1) indicates a distribution of rainfall amounts on wet days that is close to exponential.

This analysis is based on the simplifying assumption that the probability of rain occurring on day t depends only on whether rain fell on day $t - 1$, but not otherwise on the amount of rain. The decomposition into two independent generalized linear models depends heavily on this assumption. It is at least plausible, however, that the occurrence of rain on day t depends on whether or not it rained heavily on the previous day. Dependence of this nature can be checked by including in the logistic model for $\pi_0(t)$ the amount of rainfall (log units) on the previous day.

8.4.4 Developmental rate of *Drosophila melanogaster*

The data shown in Table 8.8 were collected by Powsner (1935) as part of an experiment to determine accurately the effect of temperature on the duration of the developmental stages of the fruit fly *Drosophila melanogaster*. Powsner studied four stages in its development, namely the embryonic, egg-larval, larval and pupal stages: only the first of these is considered here.

In all cases the eggs were laid at approximately 25°C and remained at that temperature for 20–30 minutes as indicated in the final column. Subsequently the eggs were brought to the experimental temperature, which was kept constant over the period of the experiment. Column 3 gives the average duration of the embryonic period measured from the time at which the eggs were laid. The number of eggs in each batch, together with the sample standard deviation of each batch, are shown in columns 4 and 5.

Figure 8.5 shows the batch standard deviations plotted against the batch means. Evidently, with the exception of the point at 32°C, the standard deviations are roughly proportional to the mean. A more formal weighted log-linear regression of the sample variances on the log of the sample means, with weights equal to the degrees of freedom, gives the fitted equation

$$\log(\text{sample variance}) \simeq -9.29 + 2.58 \log(\text{sample mean}),$$

suggesting a power relationship for the variance function with index in the range 2–3. In what follows, we assume that $V(\mu) = \mu^2$, implying that the coefficient of variation is constant. In other

Table 8.8 *Mean duration of embryonic period in the development of Drosophila melanogaster*

Temp. °C	Exp. No.	Duration (hours)	Batch size	Std. dev.	Eggs laid at	
					Temp °C	Duration (hours)
14.95	25	67.5 ± 0.33	54	2.41	25.1	0.33
16.16	44	57.1 ± 0.12	182	2.28	25.0	0.50
16.19	26	56.0 ± 0.12	153	1.46	25.1	0.33
17.15	28	48.4 ± 0.12	129	1.40	25.1	0.50
18.20	25	41.2 ± 0.16	64	1.30	25.1	0.33
19.08	33	37.80 ± 0.059	94	0.57	25.1	0.50
20.07	28	33.33 ± 0.080	82	0.73	25.1	0.33
22.14	25	26.50 ± 0.083	57	0.63	25.1	0.33
23.27	28	24.24 ± 0.038	135	0.44	25.1	0.50
24.09	33	22.44 ± 0.029	188	0.40	25.1	0.50
24.81	42	21.13 ± 0.017	217	0.36	25.0	0.50
24.84	40	21.05 ± 0.027	141	0.46	25.0	0.50
25.06	27	20.39 ± 0.064	37	0.38	25.1	0.50
25.06	27	20.41 ± 0.037	84	0.34	25.1	0.50
25.80	26	19.45 ± 0.026	196	0.36	25.1	0.33
26.92	33	18.77 ± 0.029	104	0.30	25.1	0.50
27.68	26	17.79 ± 0.041	148	0.49	25.1	0.33
28.89	29	17.38 ± 0.043	83	0.39	25.1	0.25
28.96	40	17.26 ± 0.031	95	0.43	25.0	0.50
29.00	44	17.18 ± 0.023	232	0.50	25.0	0.50
30.05	26	16.81 ± 0.032	148	0.39	25.1	0.33
30.80	26	16.97 ± 0.028	195	0.39	25.1	0.33
32.00	33	18.20 ± 0.290	58	2.23	25.1	0.50

Source: Powsner (1935).

words, the squared coefficient of variation of the batch means is assumed to be inversely proportional to the batch size. A similar analysis gives 0.0267 as a combined estimate of the coefficient of variation of the individual egg durations, ignoring the batch at 32°C.

The greatly increased variance for the batch of eggs maintained at 32°C suggests either that the tight experimental control was relaxed for this batch or, more plausibly, that the biochemistry of development at such an elevated temperature differs in important ways from the biochemistry at lower temperatures. One possibility is that the smaller eggs may suffer stress from dehydration at such

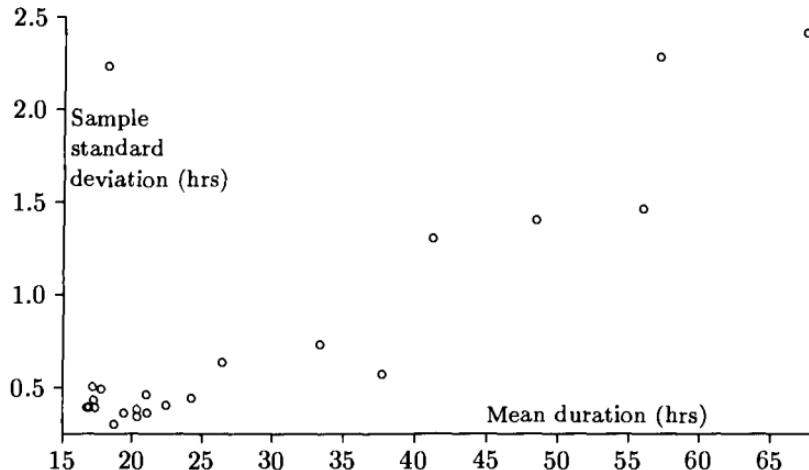


Fig. 8.5. Plot of standard deviations against sample means for 23 batches of eggs. The outlying point corresponds to the highest temperature.

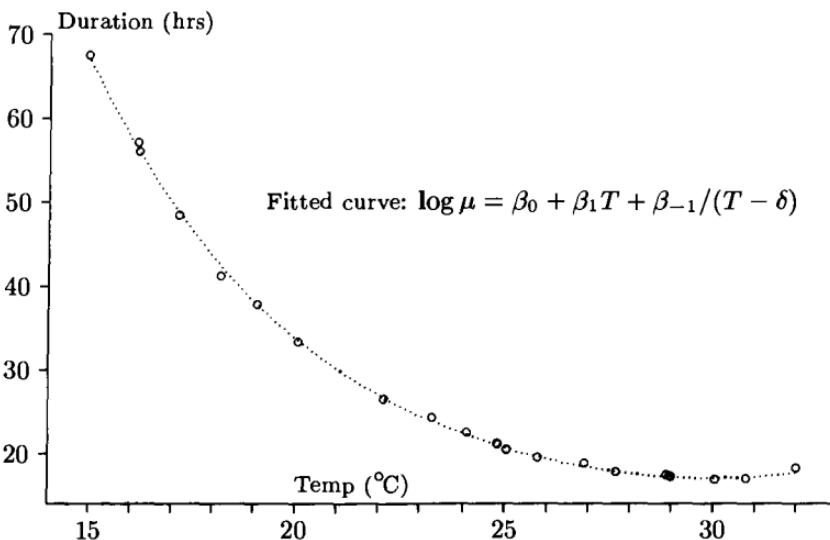


Fig. 8.6. Observed average duration of embryonic period plotted against temperature (circles). The curve was fitted using gamma errors, log link and weighted by sample size.

temperatures.

Figure 8.6 shows the observed mean durations plotted against temperature. Evidently the observed points lie very close to a smooth curve, which may have a minimum at around 29–31°C. To a

large extent the evidence for a minimum rather than an asymptote rests on the observation at 32°C. However, it seems clear on general grounds that if the temperature is sufficiently high the eggs must begin to suffer, so that an eventual increase in duration is to be expected.

One of Powsner's objectives was to test whether the rates of complex biochemical reactions obey the laws that are known to govern simple chemical reactions. Arrhenius's law for the rate of simple chemical reactions is

$$\log \text{rate} = -\mu/(RT),$$

where μ is the 'critical chemical increment' for the reaction, R is the gas constant, and T is the absolute temperature. Since the reaction rate is inversely proportional to its duration, the Arrhenius model predicts a linear relationship in the graph of $\log(\text{duration})$ against the reciprocal of absolute temperature. The observed graph for these data, however, is manifestly non-linear even when the point at 32°C is excluded. Consequently the simple Arrhenius model is unsatisfactory, even over small temperature ranges, as a description of the rate of development of *Drosophila melanogaster*.

Since the Arrhenius model is not even a close approximation to what is observed experimentally, it is necessary to proceed empirically without the support of tested theories. In what follows we treat the observed duration as the response having a squared coefficient of variation inversely proportional to the batch size. In all regression equations, therefore, each batch mean is weighted according to the batch size. This choice of response is not the only possibility: we could, for example, work with the observed duration minus the duration of the egg-laying period. The latter adjustment is rather small and is likely to have a very minor effect on the conclusions.

It is possible to obtain a reasonably good fit to the observed data by using cubic or quartic polynomials for the logarithmic reaction rate, i.e. using gamma errors with log link. However it seems worthwhile in this example to consider functions of temperature that have asymptotes. It is known for example that no development takes place below a certain critical temperature. There is undoubtedly a corresponding upper limit. Thus we are led to consider rational functions of temperature, the simplest of which is

$$\beta_0 + \beta_1 T + \beta_{-1}/(T - \delta). \quad (8.4)$$

This can be expressed as a rational function in T whose denominator is linear and numerator quadratic. It is immaterial in (8.4) whether T is measured in °C, °F or °K.

So far we have not stated whether (8.4) is to be considered as a model for the rate, the log rate, or the duration of the embryonic period. These choices amount to different choices of link functions, namely the reciprocal, logarithm and identity respectively. The model is linear in β for each fixed δ . For a simple comparison among the three link functions, therefore, we take $\delta = 0$. The deviances are 5.97 for the reciprocal, 2.77 for the logarithm and 0.473 for the identity. Among these three choices, the identity link is strongly preferred and the fit is surprisingly good. It is visually indistinguishable from the curve plotted in Fig. 8.6.

Choosing $\delta = 0$ amounts to stating a preference for the Celsius scale over the Kelvin and Fahrenheit scales. Treating δ as a free parameter leaves the choice of scale in the hands of the data. The best-fitting linear model has $\hat{\delta} \approx 0.6^\circ\text{C}$. The best-fitting log-linear model has $\hat{\delta} \approx 58.6^\circ\text{C}$, as can be seen from Fig. 8.8a, while the best-fitting inverse-linear model has $\hat{\delta} \approx 33.5^\circ\text{C}$. The corresponding deviances are 0.47, 0.32 and 1.41 respectively. The log-rational fitted curve is shown in Fig. 8.6. Parameter estimates and nominal standard errors are shown in the table below.

Parameter estimates in the log-rational model (8.4)

Parameter	Estimate	s.e.
β_0	3.201	1.594
β_1	-0.265	0.0355
β_{-1}	-217.08	125.21
δ	58.644	6.48

The estimated minimum duration or maximum rate of embryonic development occurs at

$$\hat{T} = \hat{\delta} - (\hat{\beta}_{-1}/\hat{\beta}_1)^{1/2} = 30.01^\circ\text{C}.$$

The residual coefficient of variation is estimated as

$$\tilde{\sigma} = (0.32/19)^{1/2} = (0.0168)^{1/2} = 0.13,$$

or 13%. This is the estimated coefficient of variation for the duration of the embryonic period of individual eggs. The estimated

coefficient of variation of the batch means is then $0.13/\sqrt{m_i}$, where m_i is the batch size. Despite the exceptionally good fit obtained using this class of functions, the between-batch residual variation, as measured by the coefficient of variation, is substantially larger than the within-batches coefficient of variation. From columns 3 and 5 in Table 8.8, the within-batches coefficient of variation of the individual egg durations is estimated as approximately 2.7%. Thus the ratio of the between- to within-batch squared coefficient of variations is about 23:1. If a reasonable allowance were made for model selection, this ratio would be even larger.

It would appear, therefore, that apart from the temperature differences there must have been other unrecorded differences in experimental conditions from batch to batch, for example differences in humidity, lighting conditions, temperature ranges, ventilation and so on.

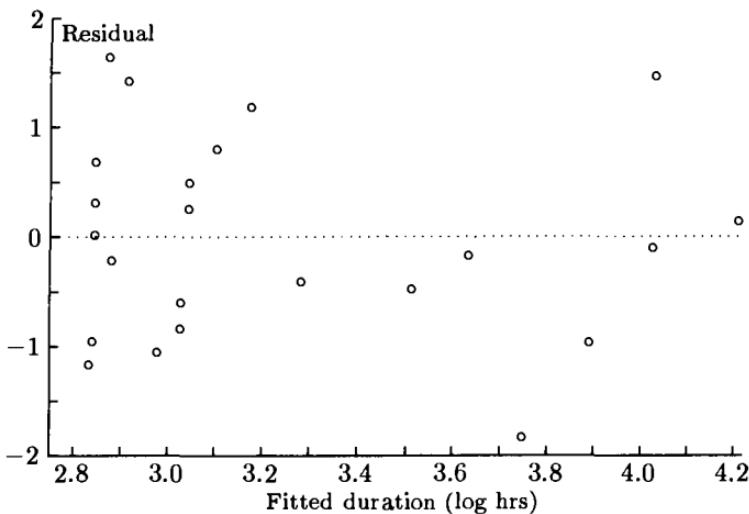


Fig. 8.7. Plot of deviance residuals for model (8.4) against log fitted values for 23 batches of eggs.

So far as the gamma assumption is concerned it is the constancy of the between-batches coefficient of variation and not the within-batch coefficient of variation that is relevant for model-fitting purposes. From that point of view, the diagram in Fig. 8.5 is irrelevant in deciding whether the variance function is quadratic. In order to test whether the between-batches coefficient of variation is constant, we examine the plot of the standardized deviance

residuals plotted against fitted values (Fig. 8.7). These deviance residuals, including the weights, are given by

$$\pm \left[-2w_i \{ \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)/\hat{\mu}_i \} \right]^{1/2} / \tilde{\sigma},$$

where $\tilde{\sigma} = 0.13$ is the estimated between-batches coefficient of variation.

If the between-batches variance function is indeed quadratic, this residual plot should look like an ordinary Normal-theory residual plot. In fact all five residuals corresponding to fitted values in the range 25–50 are small but negative, so there is evidence of a small but systematic departure from the fitted model. However there is no evidence that the dispersion of the residuals increases or decreases systematically with the fitted values. Thus the between-batches coefficient of variation appears to be constant or nearly so.

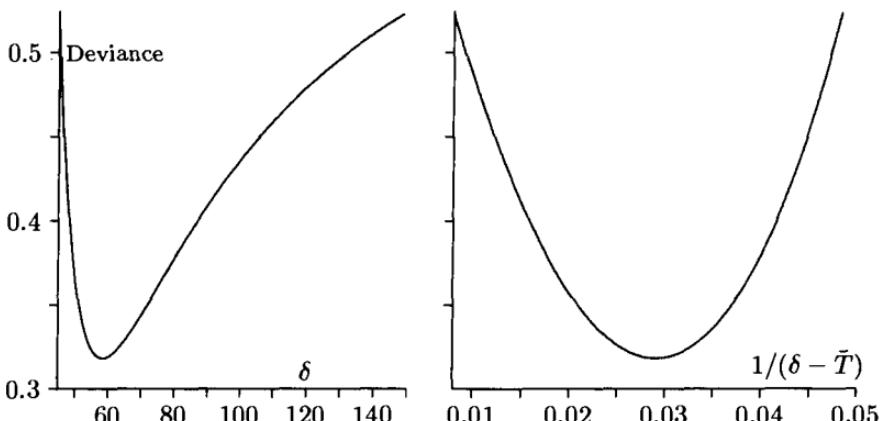


Fig. 8.8. The deviance function for model (8.4) plotted against δ and against $1/(\delta - \bar{T})$.

An awkward aspect of intrinsically non-linear models such as (8.4) is that Normal-theory approximations for the distribution of maximum-likelihood estimators may not be very accurate unless care is taken to make an appropriate transformation of the parameter. In particular, the distribution of $\hat{\delta}$ in (8.4) is noticeably non-Normal and accurate confidence intervals based on the deviance function are noticeably asymmetric. Figure 8.8a shows the residual deviance plotted against δ for values of δ in the range

45–150. Figure 8.8b shows the residual deviance plotted against $\zeta = 1/(\delta - \bar{T})$ over the equivalent range. Evidently, likelihood-based confidence limits for ζ are nearly symmetrically located about $\hat{\zeta}$, so that Normal-theory approximations for $\hat{\zeta}$ are more accurate than Normal-theory approximations for δ . Note that the transformation $\delta \rightarrow \zeta$ takes $\delta = \pm\infty$ to $\zeta = 0$. The likelihood function is continuous in ζ at this point.

For more complicated non-linear models such simplifying parameter transformations may not be easy to construct. In such cases it is necessary to plot the deviance function or log-likelihood function in order to obtain reasonably accurate confidence intervals.

The foregoing discussion presupposes the correctness of the assumed model. In this example, however an equally good fit to the observed points can be obtained using a polynomial model of degree four in place of (8.2), retaining the log link and the assumption of constant coefficient of variation. These two models are equally effective over the range of temperatures observed but they exhibit rather different behaviour on extrapolation. Since the coefficient of variation for these data is so small, an equally effective analysis could be based on the logarithmic transformation of the observed durations.

From the point of view of gaining insight into the biochemistry, neither model (8.4) nor the polynomial model is very helpful. The biochemical mechanism of egg development appears to be rather complicated: a sequence or network of dependent biochemical reactions is likely to be involved. Furthermore, this experiment gives no information on the likely duration of development if the temperature were changed after, say, 5 hours. Powsner (p. 506) discusses the effects of such changes, which are quite complicated.

For a more recent review of the role of *Drosophila melanogaster* as an experimental organism, see the review article by Rubin (1988).

8.5 Bibliographic notes

There is an extensive literature on models for exponentially distributed observations. Such models are widely used for the distribution of lifetimes in industrial reliability experiments. See, for example, the books by Barlow and Proschan (1965, 1975) and Nelson (1982).

Similar models are used for survival analysis: details and further references are provided in Chapter 13.

The family of inverse linear models was introduced by Nelder (1966).

The gamma family, as parameterized here, has many properties in common with the Normal family. In particular, for a single sample, it is possible to construct exact similar regions for composite hypotheses specifying μ . Exact confidence intervals for μ can thereby be constructed, at least in principle. In practice the exact computations are excessively complicated for samples larger than about 3 or 4. For further details and references see Exercise 8.16.

8.6 Further results and exercises 8

8.1 Show that the standard deviation of $\log(Y)$ is approximately equal to the coefficient of variation of Y . Check numerically the adequacy of the approximation in the two cases

$$Y \sim G(\mu, \nu) \quad \text{and} \quad \log(Y) \sim N(\mu, \sigma^2)$$

for $\nu = 1, 2, \dots$ and for various values of σ^2 .

8.2 Show that the gamma distribution has cumulant generating function

$$K(t) = -\nu \log(1 - \mu t / \nu).$$

Hence deduce that for large ν , the standardized random variable $\nu^{1/2}(Y - \mu)/\mu$ is approximately distributed as $N(0, 1)$.

8.3 Assuming that Y has the gamma distribution, calculate the exact mean and variance of $\log(Y)$. Use the Tables in Abramowitz and Stegun (1970) to compare numerically these exact calculations with the approximate formulae in section 8.1.

8.4 Suppose that Y_1, \dots, Y_n are independent and identically distributed with the gamma density $G(\mu, \nu)$. Show that $\bar{Y} = Y./n$ is independent of $T = (Y_1/Y., \dots, Y_n/Y.)$, and that the latter statistic has the symmetric Dirichlet distribution with index ν .

8.5 Show that for a simple random sample from the gamma distribution, the maximum-likelihood estimates of μ and ν are

independent. Derive the conditional maximum-likelihood estimate of ν given $Y_i = y_{ij}$. Compare this estimate with (8.2).

8.6 Fit the log-linear model

$$\text{PA} + \text{CG} + \text{VA}$$

to the insurance-claims data in Table 8.1. Use gamma errors and weight the averages according to the sample sizes. Examine the parameter estimates and state the conclusions to be drawn from your analysis as concisely as you can. Compare and contrast your conclusions with those presented in section 8.4.1.

8.7 For the data in Table 8.7 plot the log duration against the reciprocal of absolute temperature. Hence verify that the simple Arrhenius model does not fit these data.

8.8 Re-analyse the data in Table 8.7 using polynomials in place of (8.4). Try a number of link functions. Plot the residuals against temperature as a check on your model. Estimate the temperature at which the rate of development is a maximum.

8.9 Check whether the model described in section 8.4.4 might be improved by taking the response to be the time spent at the experimental temperature as opposed to the total duration of the embryonic period.

8.10 The data shown in Tables 8.9 and 8.10 were collected by Powsner (1935) in his study of the effect of temperature on the duration of the developmental stages of the fruit fly *Drosophila melanogaster*. In the light of the analyses suggested in section 8.4.4 for the embryonic period, examine carefully how the rates of development for the egg-larval, larval and pupal periods depend on temperature. What evidence is there of a maximum rate of development? Do the maximum developmental rates occur at the same temperature for each developmental stage? Check carefully whether there is any difference between the developmental rates for males and females, and if so, whether the difference is temperature-dependent.

8.11 By using the asymptotic expansion for $\psi(\nu)$, show that the maximum-likelihood estimate $\hat{\nu}$ in (8.1) is given approximately by

$$\frac{1}{\hat{\nu}} \simeq \frac{\bar{D}(6 + \bar{D})}{6 + 2\bar{D}}$$

Table 8.9 Mean duration of egg-larval and larval periods in the development of *Drosophila melanogaster*

Temp. °C	Egg-larval period				Larval period			
	Male		Female		Male		Female	
	Hours ± s.e.	No.	Hours ± s.e.	No.	Hours ± s.e.	No.	Hours ± s.e.	No.
14.86	421.4 ± 1.68	97	16.3	423.8 ± 1.56	114	16.4	356.8 ± 1.6	356.2 ± 1.5
15.24	394.6 ± 0.89	227	13.4	399.1 ± 0.82	227	12.2	330.6 ± 0.88	335.2 ± 0.79
16.06	328.6 ± 0.91	148	11.0	342.2 ± 0.62	187	8.3	261.8 ± 0.90	285.5 ± 0.60
18.04	250.1 ± 0.70	99	7.0	362.7 ± 0.76†	76	6.6	208.3 ± 0.69	222.0 ± 0.74
18.05	241.3 ± 0.32	423	6.8	259.1 ± 0.32	407	6.6	199.5 ± 0.10	217.4 ± 0.10
18.05	243.2 ± 0.53	178	7.1	262.4 ± 0.50	178	6.7	201.5 ± 0.52	220.7 ± 0.20
18.21	242.2 ± 0.73	81	6.6	258.7 ± 0.60	84	5.5	201.3 ± 0.72	217.8 ± 0.58
19.32	207.7 ± 0.36	122	3.9	222.2 ± 0.26	127	3.0	171.8 ± 0.11	186.4 ± 0.26
19.97	188.0 ± 0.71	125	8.0	202.6 ± 0.82	138	9.7	154.5 ± 0.71	169.2 ± 0.82
22.00	141.7 ± 0.18	128	2.0	149.9 ± 0.16	129	1.8	114.9 ± 0.17	123.2 ± 0.15
22.00	141.6 ± 0.21	195	3.0	149.9 ± 0.17	183	2.3	114.8 ± 0.21	123.2 ± 0.16
22.21	146.1 ± 0.31	139	3.6	153.1 ± 0.21	152	2.6	119.8 ± 0.30	126.9 ± 0.20
22.99	130.5 ± 0.19	220	2.9	139.8 ± 0.22	216	3.2	105.9 ± 0.19	115.2 ± 0.22
24.17	118.5 ± 0.29	140	3.4	120.7 ± 0.21	185	2.9	96.2 ± 0.29	98.4 ± 0.21
24.93	113.3 ± 0.14	242	2.2	113.6 ± 0.16	229	2.4	92.4 ± 0.14	92.7 ± 0.16
24.93	112.2 ± 0.13	341	2.4	113.4 ± 0.14	337	2.5	91.3 ± 0.13	92.5 ± 0.13
25.14	113.7 ± 0.15	1004	4.6	114.2 ± 0.16	1039	5.2	93.1 ± 0.14	93.6 ± 0.16
25.56	108.3 ± 0.14	357	2.6	108.3 ± 0.15	359	2.8	88.3 ± 0.14	88.3 ± 0.15
25.99	105.5 ± 0.14	344	2.6	105.8 ± 0.15	298	2.8	86.1 ± 0.14	86.4 ± 0.16
26.89	100.1 ± 0.15	185	2.1	99.9 ± 0.18	203	2.6	81.5 ± 0.12	81.3 ± 0.18
27.77	94.8 ± 0.14	128	1.5	94.3 ± 0.16	123	1.8	76.9 ± 0.13	76.4 ± 0.16
27.77	94.4 ± 0.13	192	1.3	93.5 ± 0.12	207	1.7	76.5 ± 0.09	75.6 ± 0.12
28.07	103.0 ± 0.34	74	2.9	102.9 ± 0.31	63	2.4	85.3 ± 0.32	85.3 ± 0.30
28.99	98.4 ± 0.25	98	2.4	98.4 ± 0.28	96	2.7	81.1 ± 0.24	81.1 ± 0.27
29.47	99.2 ± 0.30	82	2.7	98.4 ± 0.32	88	3.0	82.1 ± 0.30	81.3 ± 0.32
29.98	105.1 ± 0.25	226	3.7	105.1 ± 0.31	242	4.7	88.3 ± 0.24	88.3 ± 0.30
31.04	121.4 ± 0.56	157	7.1	118.5 ± 0.55	188	7.5	104.4 ± 0.57	101.4 ± 0.55

† apparently misrecorded: should perhaps read 262.7 ± 0.76.
 Source: Pownser (1935).

Table 8.10 *Mean duration of pupal period in the development of the fruit-fly Drosophila melanogaster*

Temp. °C	Male				Female				Diff. $M. - F.$
	Hours ± s.e.	No.	σ	Hours ± s.e.	No.	σ			
15.24	320.5 ± 0.45	228	6.72	309.5 ± 0.42	227	6.30	+11.0		
16.17	266.7 ± 0.30	76	2.62	259.0 ± 0.24	186	3.16	+7.7		
18.01	204.4 ± 0.28	97	2.78	195.3 ± 0.26	76	2.26	+9.1		
18.05	204.4 ± 0.15	178	2.06	197.0 ± 0.14	174	1.81	+7.4		
18.21	199.2 ± 0.22	83	2.01	192.2 ± 0.29	84	2.67	+7.0		
19.32	170.6 ± 0.17	120	1.87	164.7 ± 0.10	126	1.12	+5.9		
19.97	160.1 ± 0.21	125	2.40	152.2 ± 0.33	138	3.86	+7.9		
22.00	126.5 ± 0.13	195	1.79	121.4 ± 0.14	182	1.91	+5.1		
22.21	124.04 ± 0.089	138	1.05	120.70 ± 0.082	152	1.01	+3.34		
22.99	115.62 ± 0.089	153	1.09	113.08 ± 0.074	215	1.10	+2.58		
24.17	102.4 ± 0.14	140	1.64	98.65 ± 0.074	185	1.02	+3.75		
24.57	100.51 ± 0.080	238	1.25	96.78 ± 0.065	229	1.00	+3.73		
25.14	96.51 ± 0.044	967	1.37	93.55 ± 0.042	1018	1.35	+2.96		
25.29	96.40 ± 0.096	99	0.96	92.23 ± 0.15	118	1.59	+4.17		
25.99	92.24 ± 0.060	342	1.12	90.20 ± 0.058	298	1.01	+2.04		
26.89	87.23 ± 0.075	185	1.02	82.56 ± 0.069	203	0.99	+4.67		
27.77				80.0 ± 0.11	120	1.21			
28.07	83.2 ± 0.16	72	1.38	78.6 ± 0.13	64	1.07	+4.6		
28.99	80.9 ± 0.11	85	0.99	76.2 ± 0.12	82	1.08	+4.7		
29.47	80.6 ± 0.12	77	1.05	75.8 ± 0.11	70	0.92	+4.8		
29.98	81.3 ± 0.11	233	1.68	77.1 ± 0.10	246	1.60	+4.2		
30.24	82.0 ± 0.13	141	1.50	78.5 ± 0.12	161	1.50	+3.5		
31.04	82.7 ± 0.17	73	1.4	79.3	22		+3.4		

Source: Powsner (1935).

where $\bar{D} = D(\mathbf{y}; \hat{\boldsymbol{\mu}})/n$. Find the expected value of \bar{D} for an i.i.d. sample from the exponential distribution. Solve the above equation to find the approximate expected value of $\hat{\nu}$ when $\nu = 1$.

8.12 Show that the corresponding approximation for the bias-corrected estimate is

$$\frac{1}{\hat{\nu}} \simeq \bar{D} \frac{6(n-p) + n\bar{D}}{6(n-p) + 2n\bar{D}}$$

where $\bar{D} = D(\mathbf{y}; \hat{\boldsymbol{\mu}})/(n-p)$.

8.13 The data in Table 8.11 were obtained by Drs Streibig and Vleeshouwers in an experiment designed to study how the yields of various crops are affected by competition with weeds and by plant density. Taking the fresh weights as response, examine the relationship between the monoculture yields and seed density. (In

Table 8.11 *Yields of barley and the weed Sinapis alba grown in a competition experiment[†]*

<i>Seeds sown</i>	<i>Plants harvested</i>	<i>Fresh weight</i>	<i>Dry weight</i>						
<i>Pot</i>	<i>Barley</i>	<i>Sinapis</i>	<i>Barley</i>	<i>Sinapis</i>	<i>Barley</i>	<i>Sinapis</i>	<i>Barley</i>	<i>Sinapis</i>	
1	3	0	1	3	0	33.7	0.0	2.07	0.00
2	5	0	1	5	0	120.5	0.0	10.57	0.00
3	7	0	1	7	0	187.3	0.0	20.87	0.00
4	10	0	1	10	0	110.1	0.0	6.59	0.00
5	15	0	1	15	0	122.7	0.0	8.08	0.00
6	23	0	1	23	0	214.9	0.0	16.70	0.00
7	34	0	1	33	0	198.6	0.0	21.22	0.00
8	51	0	1	48	0	263.6	0.0	26.57	0.00
9	77	0	1	60	0	254.1	0.0	23.71	0.00
10	115	0	1	70	0	230.4	0.0	20.46	0.00
11	0	5	1	0	5	0.0	254.0	0.00	34.85
12	3	5	1	3	5	14.8	167.6	1.49	29.49
13	7	5	1	6	5	38.1	240.5	2.26	19.75
14	15	5	1	15	5	93.1	132.6	11.08	23.09
15	34	5	1	34	5	120.8	166.9	12.85	25.83
16	77	5	1	50	5	214.5	53.2	24.94	8.76
17	0	7	1	0	7	0.0	228.3	0.00	38.98
18	0	10	1	0	10	0.0	209.8	0.00	28.14
19	3	10	1	3	11	15.2	220.1	1.63	35.43
20	7	10	1	7	7	37.6	203.0	2.80	29.05
21	15	10	1	15	10	93.3	130.5	6.29	17.36
22	34	10	1	31	9	98.6	178.5	7.81	23.30
23	77	10	1	62	10	203.9	81.5	19.51	12.45
24	0	15	1	0	16	0.0	214.4	0.00	36.02
25	0	23	1	0	22	0.0	269.3	0.00	47.24
26	3	23	1	3	23	7.5	272.2	1.06	49.14
27	7	23	1	6	23	18.8	220.1	1.83	35.85
28	15	23	1	14	23	64.7	175.8	9.35	30.05
29	34	23	1	28	25	84.3	240.3	9.75	36.46
30	77	23	1	53	26	125.8	135.5	14.29	19.87
31	0	34	1	0	33	0.0	267.4	0.00	38.23
32	0	51	1	0	53	0.0	244.6	0.00	31.75
33	3	51	1	1	58	3.3	332.0	0.34	35.68
34	7	51	1	7	54	21.5	264.9	2.11	37.95
35	15	51	1	11	53	26.4	221.5	1.89	25.78
36	34	51	1	23	50	32.8	230.1	3.97	39.97
37	77	51	1	61	52	76.9	184.0	7.60	24.42
38	0	77	1	0	81	0.0	291.9	0.00	45.56
39	0	115	1	0	115	0.0	300.3	0.00	43.94
40	3	115	1	3	108	1.3	284.9	0.13	29.39
41	7	115	1	6	109	5.8	243.7	0.55	33.44
42	15	115	1	14	111	12.1	287.9	0.95	35.68
43	34	115	1	26	107	26.4	233.0	2.07	21.53
44	77	115	1	57	115	95.9	189.2	10.14	24.02
45	0	173	1	0	158	0.0	326.4	0.00	35.24

(continued)

Table 8.11 *Continued*

<i>Seeds sown</i>			<i>Plants harvested</i>			<i>Fresh weight</i>	<i>Dry weight</i>		
<i>Pot</i>	<i>Barley</i>	<i>Sinapis</i>	<i>Block</i>	<i>Barley</i>	<i>Sinapis</i>	<i>Barley</i>	<i>Sinapis</i>	<i>Barley</i>	<i>Sinapis</i>
46	3	0	2	3	0	73.1	0.0	5.32	0.00
47	5	0	2	5	0	152.7	0.0	13.59	0.00
48	7	0	2	7	0	125.4	0.0	9.97	0.00
49	10	0	2	10	0	208.9	0.0	21.40	0.00
50	15	0	2	15	0	171.5	0.0	11.07	0.00
51	23	0	2	19	0	98.7	0.0	6.66	0.00
52	34	0	2	27	0	191.8	0.0	14.25	0.00
53	51	0	2	41	0	238.7	0.0	39.37	0.00
54	77	0	2	49	0	197.2	0.0	21.44	0.00
55	115	0	2	72	0	256.4	0.0	30.92	0.00
56	0	5	2	0	5	0.0	227.3	0.00	32.61
57	3	5	2	3	5	28.9	246.3	1.66	34.18
58	7	5	2	8	5	42.3	230.0	3.62	33.63
59	15	5	2	15	5	82.9	156.1	10.41	27.06
60	34	5	2	28	5	116.7	125.9	10.46	19.99
61	77	5	2	57	5	187.7	55.8	23.10	9.01
62	0	7	2	0	7	0.0	231.5	0.00	34.20
63	0	10	2	0	10	0.0	258.8	0.00	44.47
64	3	10	2	3	10	25.8	245.1	2.88	38.13
65	7	10	2	7	11	41.6	185.8	5.32	33.31
66	15	10	2	14	11	67.7	174.3	11.10	32.49
67	34	10	2	33	11	86.0	177.9	9.16	29.20
68	77	10	2	65	10	162.3	75.0	23.18	12.52
69	0	15	2	0	15	0.0	237.6	0.00	40.32
70	0	23	2	0	29	0.0	225.9	0.00	37.46
71	3	23	2	3	29	9.8	274.0	0.76	46.68
72	7	23	2	6	27	27.7	221.4	1.96	34.40
73	15	23	2	13	23	30.2	246.0	4.19	45.07
74	34	23	2	24	25	110.0	147.9	11.34	22.75
75	77	23	2	56	24	85.7	185.4	10.39	30.38
76	0	34	2	0	34	0.0	281.0	0.00	43.76
77	0	51	2	0	51	0.0	318.9	0.00	40.25
78	3	51	2	5	54	6.8	309.5	0.12	45.80
79	7	51	2	7	51	12.1	254.2	0.26	29.29
80	15	51	2	13	51	28.5	226.8	2.32	31.28
81	34	51	2	25	52	55.4	195.3	5.73	29.61
82	77	51	2	49	50	109.7	179.0	8.61	20.14
83	0	77	2	0	76	0.0	304.7	0.00	38.94
84	0	115	2	0	101	0.0	313.5	0.00	43.36
85	3	115	2	3	99	10.6	249.2	0.97	36.54
86	7	115	2	10	109	7.4	255.9	0.01	31.80
87	15	115	2	16	97	17.9	170.9	1.05	24.96
88	34	115	2	22	105	38.1	270.9	3.65	38.52
89	77	115	2	47	97	52.5	266.6	6.40	38.91
90	0	173	2	0	148	0.0	279.2	0.00	39.35

(continued)

Table 8.11 *Continued*

Pot	Seeds sown			Plants harvested			Fresh weight		Dry weight	
	Barley	Sinapis	Block	Barley	Sinapis	Barley	Sinapis	Barley	Sinapis	
91	3	0	3	3	0	42.9	0.0	3.14	0.00	
92	5	0	3	5	0	165.9	0.0	14.69	0.00	
93	7	0	3	7	0	81.4	0.0	5.45	0.00	
94	10	0	3	9	0	223.3	0.0	23.12	0.00	
95	15	0	3	17	0	116.3	0.0	8.28	0.00	
96	23	0	3	20	0	193.7	0.0	19.48	0.00	
97	34	0	3	29	0	237.1	0.0	38.11	0.00	
98	51	0	3	42	0	264.0	0.0	25.53	0.00	
99	77	0	3	47	0	241.0	0.0	19.72	0.00	
100	115	0	3	73	0	269.0	0.0	41.02	0.00	
101	0	5	3	0	5	0.0	184.2	0.00	36.18	
102	3	5	3	3	5	22.9	142.2	1.86	23.39	
103	7	5	3	7	5	58.0	166.6	8.37	31.13	
104	15	5	3	16	10	77.4	181.6	7.97	29.64	
105	34	5	3	32	8	114.8	141.6	14.14	24.57	
106	77	5	3	53	8	124.7	86.4	16.37	15.46	
107	0	7	3	0	11	0.0	235.9	0.00	35.44	
108	0	10	3	0	14	0.0	200.8	0.00	33.60	
109	3	10	3	3	15	12.6	197.7	1.50	37.66	
110	7	10	3	9	14	46.7	231.1	5.61	39.01	
111	15	10	3	16	14	44.5	198.2	4.05	29.21	
112	34	10	3	27	15	73.4	122.1	8.33	20.35	
113	77	10	3	53	12	132.4	121.9	20.59	23.64	
114	0	15	3	0	15	0.0	251.8	0.00	32.16	
115	0	23	3	0	26	0.0	229.3	0.00	28.46	
116	3	23	3	3	24	9.0	244.8	0.30	33.31	
117	7	23	3	7	28	27.1	203.8	3.39	35.58	
118	15	23	3	14	24	37.0	218.2	2.31	30.36	
119	34	23	3	22	35	71.2	152.3	7.35	21.86	
120	77	23	3	57	22	91.9	158.1	13.45	27.69	
121	0	34	3	0	42	0.0	250.9	0.00	41.58	
122	0	51	3	0	53	0.0	275.6	0.00	46.00	
123	3	51	3	3	52	5.1	298.8	0.41	44.17	
124	7	51	3	7	51	15.4	257.2	1.45	38.54	
125	15	51	3	13	53	32.1	232.0	2.50	31.47	
126	34	51	3	27	51	37.1	191.0	2.35	19.96	
127	77	51	3	60	51	107.6	179.8	10.04	22.79	
128	0	77	3	0	79	0.0	286.6	0.00	44.71	
129	0	115	3	0	111	0.0	289.6	0.00	38.34	
130	3	115	3	3	103	3.5	236.9	0.06	27.80	
131	7	115	3	8	113	13.8	239.0	1.25	34.50	
132	15	115	3	14	110	11.6	289.7	0.59	28.42	
133	34	115	3	26	90	38.7	246.6	2.80	27.79	
134	77	115	3	56	82	78.0	194.6	8.67	26.64	
135	0	173	3	0	152	0.0	252.7	0.00	24.41	

[†]Data courtesy of Drs J.C. Streibig and L. Vleeshouwers, Dept. of Crop Science, Royal Veterinary and Agricultural University, Copenhagen.

a few cases there was some doubt concerning the temperature of the drying process, so that the fresh weights may be more reliable than the dry weights.) Show that, for both barley and *Sinapis*, the relation between monoculture yield and seed density is approximately inverse quadratic. The monoculture observations are that subset of Table 8.11 in which only one variety was planted.

8.14 For those plots in which both varieties were sown, examine how the barley proportion of the total yield depends on the barley proportion of the seeds sown, the seed density and the experimental block. Take the log ratio of fresh weights, $\log(Y_B/Y_S)$, as the response and consider models of the form

$$\log(Y_B/Y_S) = \alpha_{BS} + \beta \log(N_B/N_S) + \gamma x + \text{block effect},$$

where N_B, N_S are the numbers of seeds sown, and x is a measure of seed density, say $x = \log(N_B + N_S)$.

What would be the interpretation of the following parameter values?

1. $\beta = 1, \gamma = 0$;
2. $\beta < 1, \gamma = 0$;
3. $\beta = 1, \gamma > 0$;
4. $\beta < 1, \gamma > 0$.

For further information concerning competition experiments, see Williams (1962), Breese and Hill (1973), Mead and Curnow (1983) or Skovgaard (1986).

8.15 Suppose that $Y_i \sim G(\mu_i, \nu)$ independently for each i , with μ_i satisfying the log-linear model

$$\log(\mu_i) = \alpha + \mathbf{x}_i^T \boldsymbol{\beta}$$

and ν an unknown constant. Show that the transformed responses satisfy

$$E(\log(Y_i)) = \alpha^* + \mathbf{x}_i^T \boldsymbol{\beta},$$

$$\text{var}(\log Y_i) = \psi'(\nu).$$

where $\alpha^* = \alpha + \psi(\nu) - \log(\nu)$ and $\psi(\nu) = \Gamma'(\nu)/\Gamma(\nu)$.

Let $\tilde{\boldsymbol{\beta}}$ be the least squares estimator of $\boldsymbol{\beta}$ obtained by fitting a linear regression model to the transformed data. Show that $\tilde{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ and that the asymptotic efficiency of $\tilde{\boldsymbol{\beta}}$ relative to $\hat{\boldsymbol{\beta}}$ is $1/\{\nu\psi'(\nu)\}$. [Bartlett and Kendall 1946; Cox and Hinkley 1966].

8.16 Suppose that Y_1, \dots, Y_n are independent and identically distributed with the gamma distribution $G(\mu, \nu)$, both parameters being taken as unknown. Let \bar{Y} be the arithmetic mean of the observations, and \dot{Y} the geometric mean. Show that under the composite hull hypothesis $H_0: \mu = \mu_0$, $S_0 = \log(\dot{Y}/\mu_0) - \bar{Y}/\mu_0$ is a complete sufficient statistic for ν . Show, again under H_0 , that the conditional joint distribution of $Z_i = \log(Y_i/\mu_0)$ given S_0 is uniform over the surface

$$\sum (z_i - \exp(z_i)) = nS_0.$$

Discuss briefly how you might use this result (i) to construct an exact test of H_0 , and (ii) to construct an exact confidence interval for μ .

Quasi-likelihood functions

9.1 Introduction

One of the few points on which theoretical statisticians of all persuasions are agreed is the importance of the role played by the likelihood function in statistical inference. This role has been illustrated, chiefly from the frequentist viewpoint, in the developments of Chapters 2–8. In order to construct a likelihood function it is usually necessary to posit a probabilistic mechanism specifying, for a range of parameter values, the probabilities of all relevant samples that might possibly have been observed. Such a specification implies either knowledge of the mechanism by which the data were generated or substantial experience of similar data from previous experiments.

Often there is no theory available on the random mechanism by which the data were generated. We may, however, be able to specify the range of possible response values (discrete, continuous, positive, ...), and past experience with similar data is usually sufficient to specify, in a qualitative fashion, a few additional characteristic features of the data, such as

1. how the mean or median response is affected by external stimuli or treatments;
2. how the variability of the response changes with the average response;
3. whether the observations are statistically independent;
4. whether the response distribution under fixed treatment conditions is skewed positively, negatively or is symmetric.

Often interest attaches to how the mean response or other simple functional is affected by one or more covariates. Usually there is substantial prior information on the likely nature of this relationship, but rather little about the pattern of higher-order cumulants

or moments.

The purpose of this chapter is to show how inferences can be drawn from experiments in which there is insufficient information to construct a likelihood function. We concentrate mainly on the case in which the observations are independent and where the effects of interest can be described by a model for $E(Y)$.

9.2 Independent observations

9.2.1 Covariance functions

Suppose that the components of the response vector \mathbf{Y} are independent with mean vector $\boldsymbol{\mu}$ and covariance matrix $\sigma^2 \mathbf{V}(\boldsymbol{\mu})$, where σ^2 may be unknown and $\mathbf{V}(\boldsymbol{\mu})$ is a matrix of known functions. It is assumed throughout this section that the parameters of interest, β , relate to the dependence of $\boldsymbol{\mu}$ on covariates \mathbf{x} . The nature of this relationship need not concern us for the moment, so we write $\boldsymbol{\mu}(\beta)$, thereby absorbing the covariates into the regression function. An important point is that σ^2 is assumed constant—in particular that σ^2 does not depend on β .

Since the components of \mathbf{Y} are independent by assumption the matrix $\mathbf{V}(\boldsymbol{\mu})$ must be diagonal. Thus we write

$$\mathbf{V}(\boldsymbol{\mu}) = \text{diag}\{V_1(\boldsymbol{\mu}), \dots, V_n(\boldsymbol{\mu})\}.$$

One further assumption is required concerning the functions $V_i(\boldsymbol{\mu})$, namely that $V_i(\boldsymbol{\mu})$ must depend only on the i th component of $\boldsymbol{\mu}$. In principle it is possible, even under independence, for $V_1(\boldsymbol{\mu})$ to depend on several components of $\boldsymbol{\mu}$. However it is difficult to imagine a plausible physical mechanism that would produce such dependence in the variance function, while at the same time keeping the random variables statistically independent.

The above assumption of functional independence, namely that

$$\mathbf{V}(\boldsymbol{\mu}) = \text{diag}\{V_1(\mu_1), \dots, V_n(\mu_n)\}, \quad (9.1)$$

although sensible physically, has been made for technical mathematical reasons that will become apparent in section 9.3. It is no more than a happy accident that such a technical mathematical requirement should coincide with what is sensible on external physical or scientific grounds.

In the majority of applications the functions $V_1(\cdot), \dots, V_n(\cdot)$ may be taken to be identical, though their arguments, and hence their values, are different. However, this assumption is not required in the algebra that follows.

9.2.2 Construction of the quasi-likelihood function

Consider first a single component of the response vector \mathbf{Y} , which we write as Y or y without subscripts. Under the conditions listed above, the function

$$U = u(\mu; Y) = \frac{Y - \mu}{\sigma^2 V(\mu)}$$

has the following properties in common with a log-likelihood derivative:

$$\begin{aligned} E(U) &= 0, \\ \text{var}(U) &= 1/\{\sigma^2 V(\mu)\}, \\ -E\left(\frac{\partial U}{\partial \mu}\right) &= 1/\{\sigma^2 V(\mu)\}. \end{aligned} \tag{9.2}$$

Since most first-order asymptotic theory connected with likelihood functions is founded on these three properties, it is not surprising that, to some extent, the integral

$$Q(\mu; y) = \int_y^\mu \frac{y - t}{\sigma^2 V(t)} dt \tag{9.3}$$

if it exists, should behave like a log-likelihood function for μ under the very mild assumptions stated in the previous two sections. Some examples of such quasi-likelihoods for a number of common variance functions are given in Table 9.1. Many, but not all, of these quasi-likelihoods correspond to real log likelihoods for known distributions.

We refer to $Q(\mu; y)$ as the quasi-likelihood, or more correctly, as the log quasi-likelihood for μ based on data y . Since the components of \mathbf{Y} are independent by assumption, the quasi-likelihood for the complete data is the sum of the individual contributions

$$Q(\boldsymbol{\mu}; \mathbf{y}) = \sum Q_i(\mu_i; y_i).$$

Table 9.1. Quasi-likelihoods associated with some simple variance functions

Variance function	Quasi-likelihood	Canonical parameter	Distribution name	Range restrictions
$V(\mu)$	$Q(\mu; y)$	θ		
1	$-(y - \mu)^2/2$	μ	Normal	—
μ	$y \log \mu - \mu$	$\log \mu$	Poisson	$\mu > 0, y \geq 0$
μ^2	$-y/\mu - \log \mu$	$-1/\mu$	Gamma	$\mu > 0, y \geq 0$
μ^3	$-y/(2\mu^2) + 1/\mu$	$-1/(2\mu^2)$	Inverse Gaussian	$\mu > 0, y \geq 0$
μ^ζ	$\mu^{-\zeta} \left(\frac{\mu y}{1-\zeta} - \frac{\mu^2}{2-\zeta} \right)$	$\frac{1}{(1-\zeta)\mu^{\zeta-1}}$	—	$\mu > 0, \zeta \neq 0, 1, 2$
$\mu(1-\mu)$	$y \log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)$	$\log\left(\frac{\mu}{1-\mu}\right)$	Binomial/ m	$0 < \mu < 1, 0 \leq y \leq 1$
$\mu^2(1-\mu)^2$	$(2y-1) \log\left(\frac{\mu}{1-\mu}\right) - \frac{y}{\mu} - \frac{1-y}{1-\mu}$	—	—	$0 < \mu < 1, 0 \leq y \leq 1$
$\mu + \mu^2/k$	$y \log\left(\frac{\mu}{k+\mu}\right) + k \log\left(\frac{k}{k+\mu}\right)$	$\log\left(\frac{\mu}{k+\mu}\right)$	Negative binomial	$\mu > 0, y \geq 0$

By analogy, the quasi-deviance function corresponding to a single observation is

$$D(y; \mu) = -2\sigma^2 Q(\mu; y) = 2 \int_{\mu}^y \frac{y-t}{V(t)} dt, \quad (9.4)$$

which is evidently strictly positive except at $y = \mu$. The total deviance $D(\mathbf{y}; \boldsymbol{\mu})$, obtained by adding over the components, is a computable function depending on \mathbf{y} and $\boldsymbol{\mu}$ alone: it does not depend on σ^2 .

9.2.3 Parameter estimation

The quasi-likelihood estimating equations for the regression parameters $\boldsymbol{\beta}$, obtained by differentiating $Q(\boldsymbol{\mu}; \mathbf{y})$, may be written in the form $\mathbf{U}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, where

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{D}^T \mathbf{V}^{-1} (\mathbf{Y} - \boldsymbol{\mu}) / \sigma^2 \quad (9.5)$$

is called the quasi-score function. In this expression the components of \mathbf{D} , of order $n \times p$, are $D_{ir} = \partial \mu_i / \partial \beta_r$, the derivatives of $\boldsymbol{\mu}(\boldsymbol{\beta})$ with respect to the parameters.

The covariance matrix of $\mathbf{U}(\boldsymbol{\beta})$, which is also the negative expected value of $\partial \mathbf{U}(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$, is

$$\mathbf{i}_{\boldsymbol{\beta}} = \mathbf{D}^T \mathbf{V}^{-1} \mathbf{D} / \sigma^2. \quad (9.6)$$

For quasi-likelihood functions, this matrix plays the same role as the Fisher information for ordinary likelihood functions. In particular, under the usual limiting conditions on the eigenvalues of $\mathbf{i}_{\boldsymbol{\beta}}$, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\text{cov}(\hat{\boldsymbol{\beta}}) \simeq \mathbf{i}_{\boldsymbol{\beta}}^{-1} = \sigma^2 (\mathbf{D}^T \mathbf{V}^{-1} \mathbf{D})^{-1},$$

as can be seen from the argument given below.

Beginning with an arbitrary value $\hat{\boldsymbol{\beta}}_0$ sufficiently close to $\boldsymbol{\beta}$, the sequence of parameter estimates generated by the Newton-Raphson method with Fisher scoring is

$$\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_0 + (\hat{\mathbf{D}}_0^T \hat{\mathbf{V}}_0^{-1} \hat{\mathbf{D}}_0)^{-1} \hat{\mathbf{D}}_0^T \hat{\mathbf{V}}_0^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}}_0).$$

The quasi-likelihood estimate $\hat{\beta}$ may be obtained by iterating until convergence occurs. An important property of the sequence so generated is that it does not depend on the value of σ^2 .

For theoretical calculations it is helpful to imagine the iterations starting at the true value, β . Thus we find

$$\hat{\beta}_1 = \beta + (\mathbf{D}^T \mathbf{V}^{-1} \mathbf{D})^{-1} \mathbf{D}^T \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}), \quad (9.7)$$

showing that the one-step estimate is a linear function of the data. Provided that the eigenvalues of i_β are sufficiently large, subsequent iterations produce asymptotically negligible adjustments to $\hat{\beta}_1$. Thus, for a first-order asymptotic theory, we may take $\hat{\beta} = \hat{\beta}_1$, even though $\hat{\beta}_1$ is not a computable statistic. Approximate unbiasedness and asymptotic Normality of $\hat{\beta}$ follow directly from (9.7) under the second-moment assumptions made in this chapter.

In all of the above respects the quasi-likelihood behaves just like an ordinary log likelihood. For the estimation of σ^2 , however, $Q(\cdot; y)$ does not behave like a log likelihood. The conventional estimate of σ^2 is a moment estimator based on the residual vector $\mathbf{Y} - \hat{\boldsymbol{\mu}}$, namely

$$\tilde{\sigma}^2 = \frac{1}{n-p} \sum_i (Y_i - \hat{\mu}_i)^2 / V_i(\hat{\mu}_i) = X^2 / (n-p),$$

where X^2 is the generalized Pearson statistic.

9.2.4 Example: incidence of leaf-blotch on barley

The data in Table 9.2, taken from Wedderburn (1974), concerns the incidence of *Rhynchosporium secalis*, or leaf blotch, on 10 varieties of barley grown at nine sites in 1965. The response, which is the percentage leaf area affected, is a continuous proportion in the interval $[0, 1]$. For convenience of discussion we take Y to be a proportion in $[0, 1]$ rather than a percentage. Following the precedent set in section 6.3.1, we might attempt an analysis, treating the data as pseudo-binomial observations, taking the variances, at least initially to be $\sigma^2\mu(1-\mu)$. A linear logistic model with main effects appears to describe adequately the site and variety effects.

This analysis is certainly reasonable as a first step. The usual residual plots and additivity tests indicate no significant departures

Table 9.2. Incidence of *R. secalis* on the leaves of ten varieties of barley grown at nine sites: response is the percentage of leaf affected

Site	Variety										Mean
	1	2	3	4	5	6	7	8	9	10	
1	0.05	0.00	0.00	0.10	0.25	0.05	0.50	1.30	1.50	1.50	0.52
2	0.00	0.05	0.05	0.30	0.75	0.30	3.00	7.50	1.00	12.70	2.56
3	1.25	1.25	2.50	16.60	2.50	2.50	0.00	20.00	37.50	26.25	11.03
4	2.50	0.50	0.01	3.00	2.50	0.01	25.00	55.00	5.00	40.00	13.35
5	5.50	1.00	6.00	1.10	2.50	8.00	16.50	29.50	20.00	43.50	13.36
6	1.00	5.00	5.00	5.00	5.00	5.00	10.00	5.00	50.00	75.00	16.60
7	5.00	0.10	5.00	5.00	50.00	10.00	50.00	25.00	50.00	75.00	27.51
8	5.00	10.00	5.00	5.00	25.00	75.00	50.00	75.00	75.00	40.00	
9	17.50	25.00	42.50	50.00	37.50	95.00	62.50	95.00	95.00	95.00	61.50
Mean	4.20	4.77	7.34	9.57	14.00	21.76	24.17	34.81	37.22	49.33	20.72

Source: Wedderburn (1974) taken from an unpublished Aberystwyth Ph.D thesis by J.F. Jenkyn.

from the linear logistic model. The residual deviance is 6.13 on 72 degrees of freedom and Pearson's statistic is equal to 6.39. Thus the estimate of σ^2 is $\tilde{\sigma}^2 = 6.39/72 = 0.089$. Since the data do not involve counts there is no reason to expect σ^2 to be near 1.0.

The estimated variety effects together with their standard errors, are shown below:

<i>Variety</i>										
1	2	3	4	5	6	7	8	9	10	
0.00	0.15	0.69	1.05	1.62	2.37	2.57	3.34	3.50	4.25	
(0.00)	(0.72)	(0.67)	(0.65)	(0.63)	(0.61)	(0.61)	(0.60)	(0.60)	(0.60)	

Since these are simple contrasts with variety 1, the correlations among the estimates are approximately equal to 1/2. The actual correlations range from 0.68 to 0.83, which are larger than expected because variety 1 has a larger variance on the logistic scale than the other varieties. Evidently varieties 1-3 are most resistant to leaf blotch and varieties 8-10 least resistant.

In fact, however, as is shown in Fig. 9.1, the variance function $\mu(1 - \mu)$ is not a satisfactory description of the variability in these data for very small or very large proportions. The variability observed in these plots is smaller at the extreme proportions than that predicted by the binomial variance function. Following Wedderburn's suggestion, we try an alternative variance function of the form $\mu^2(1 - \mu)^2$ to mimic this effect. The resulting quasi-likelihood function can be obtained in closed form as

$$Q(\mu; y) = (2y - 1) \log\left(\frac{\mu}{1-\mu}\right) - \frac{y}{\mu} - \frac{1-y}{1-\mu}.$$

Unfortunately this function is not defined for $\mu = 0$ or $\mu = 1$. A deviance function cannot be defined in the usual way for the data in Table 9.1 because some of the observed proportions are zero.

A linear logistic model with the new variance function gives the following estimated variety effects and standard errors:

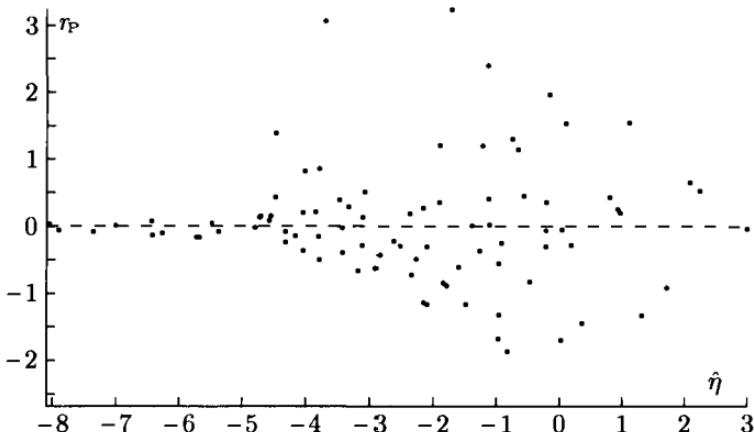


Fig. 9.1a. Pearson residuals plotted against the linear predictor $\hat{\eta} = \log(\hat{\pi}/(1 - \hat{\pi}))$ for the 'binomial' fit to the leaf-blotch data.

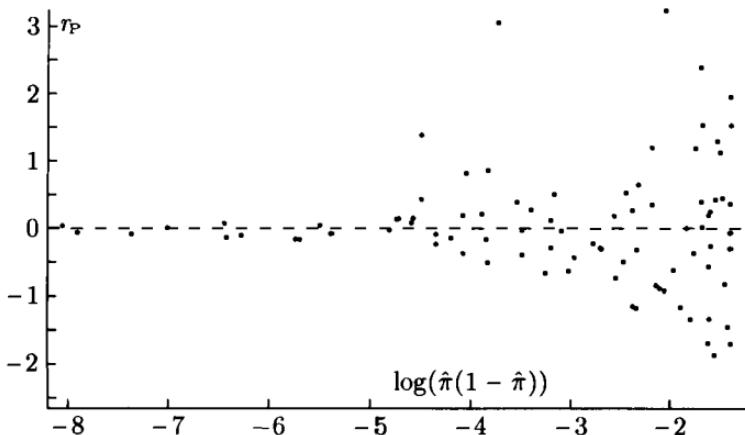


Fig. 9.1b. Pearson residuals plotted against the logarithm of the variance function $\hat{\pi}(1 - \hat{\pi})$ for the 'binomial' fit to the leaf-blotch data.

The estimated dispersion parameter, obtained from the residual weighted mean square, is now $\tilde{\sigma}^2 = 71.2/72 = 0.99$, which differs very slightly from Wedderburn's value. The correlations among these estimates are exactly $1/2$ because the iterative weights are exactly unity, and the analysis is effectively orthogonal. All variety contrasts in this model have equal standard error. Note that the ordering of varieties in the revised analysis differs slightly from the previous ordering. The principal difference between the two analyses, however, is that variety contrasts for which the incidence

is low are now estimated with greater apparent precision than in the previous model. Variety contrasts for which the incidence is high have reduced apparent precision.

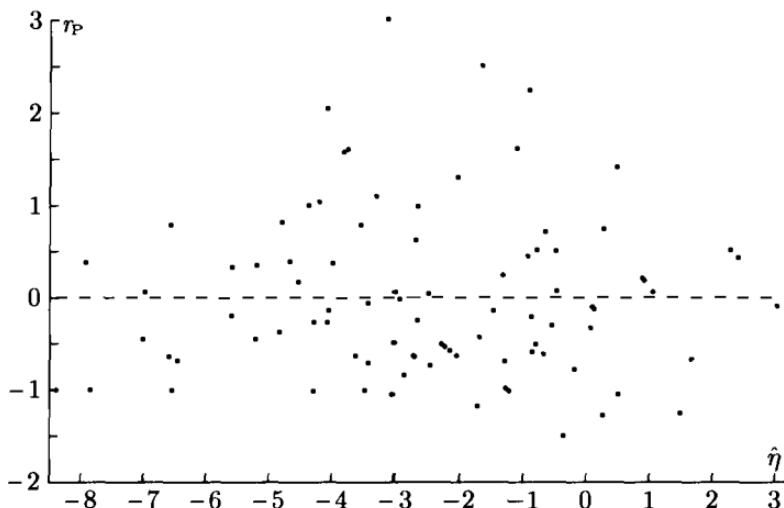


Fig. 9.2. Pearson residuals using the variance function $\pi^2(1-\pi)^2$ plotted against the linear predictor $\hat{\eta}$ for the leaf-blotch data.

The residuals are shown in Fig. 9.2 plotted against the linear predictor. To some extent the characteristic shape of Fig. 9.1a remains, though the effect is substantially diminished. Examination of individual residuals reveals three that are large and positive. These correspond, in decreasing order, to variety 4 at site 3 (3.01), variety 5 at site 7 (2.51), and variety 6 at site 8 (2.24). These residuals are computed by the formula $(y - \hat{\mu})/(\hat{\sigma}\hat{\mu}(1 - \hat{\mu}))$. There is no further evidence of systematic departures from the model.

9.3 Dependent observations

9.3.1 Quasi-likelihood estimating equations

Suppose now that $\text{cov}(\mathbf{Y}) = \sigma^2 \mathbf{V}(\boldsymbol{\mu})$ where $\mathbf{V}(\boldsymbol{\mu})$ is a symmetric positive-definite $n \times n$ matrix of known functions $V_{ij}(\boldsymbol{\mu})$, no longer diagonal. The score function (9.5), with components $U_r(\boldsymbol{\beta})$, has the following properties:

$$E\{U_r(\boldsymbol{\beta})\} = 0,$$

$$\text{cov}\{\mathbf{U}(\boldsymbol{\beta})\} = \mathbf{D}^T \mathbf{V}^{-1} \mathbf{D} / \sigma^2 = \mathbf{i}_{\boldsymbol{\beta}}, \quad (9.8)$$

$$-E\left(\frac{\partial U_r(\boldsymbol{\beta})}{\partial \beta_s}\right) = \mathbf{D}^T \mathbf{V}^{-1} \mathbf{D} / \sigma^2.$$

Thus, for the reasons given in section 9.2.2, we may treat $\mathbf{U}(\boldsymbol{\beta})$ as if it were the derivative with respect to $\boldsymbol{\beta}$ of a log-likelihood function. Under suitable limiting conditions, the root $\hat{\boldsymbol{\beta}}$ of the estimating equation

$$\mathbf{U}(\hat{\boldsymbol{\beta}}) = \hat{\mathbf{D}}^T \hat{\mathbf{V}}^{-1} (\mathbf{Y} - \hat{\boldsymbol{\mu}}) = \mathbf{0}$$

is approximately unbiased for $\boldsymbol{\beta}$ and asymptotically Normally distributed with limiting variance

$$\text{cov}(\hat{\boldsymbol{\beta}}) \simeq \sigma^2 (\mathbf{D}^T \mathbf{V}^{-1} \mathbf{D})^{-1} = \mathbf{i}_{\boldsymbol{\beta}}^{-1}.$$

The exact conditions required for consistency and asymptotic Normality of $\hat{\boldsymbol{\beta}}$ are rather complicated. Roughly speaking, however, it is necessary that as $n \rightarrow \infty$, $\mathbf{U}(\boldsymbol{\beta})$ should be asymptotically Normal, and the eigenvalues of $\mathbf{i}_{\boldsymbol{\beta}}$ should tend to infinity for all $\boldsymbol{\beta}$ in an open neighbourhood of the true parameter point.

Block-diagonal covariance matrices arise most commonly in longitudinal studies, in which repeat measurements made on the same subject are usually positively correlated. Such applications are discussed by Liang and Zeger (1986) and Zeger and Liang (1986). These authors exploit the property that the quasi-likelihood estimate $\hat{\boldsymbol{\beta}}$ is often consistent even if the covariance matrix is misspecified. For a second example in which \mathbf{V} is not block-diagonal, see section 14.5.

9.3.2 Quasi-likelihood function

Thus far, there is no essential difference between the discussion in section 9.2, for independent observations, and the more general case considered here. There is, however, one curious difference whose importance for inference is not entirely clear. If the score vector $\mathbf{U}(\boldsymbol{\beta})$ is to be the gradient vector of a log likelihood or quasi-likelihood it is necessary and sufficient that the derivative matrix of $\mathbf{U}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ be symmetric. In general, however, for $r \neq s$,

$$\frac{\partial U_r(\boldsymbol{\beta})}{\partial \beta_s} \neq \frac{\partial U_s(\boldsymbol{\beta})}{\partial \beta_r},$$

even though these matrices are equal in expectation to $-\mathbf{i}_\beta$. Consequently, unless some further conditions are imposed on the form of the matrix $\mathbf{V}(\boldsymbol{\mu})$, there can be no scalar function whose gradient vector is equal to $\mathbf{U}(\boldsymbol{\beta})$.

To state the same conclusion in a slightly more constructive way, the line integral

$$Q(\boldsymbol{\mu}; \mathbf{y}, \mathbf{t}(s)) = \sigma^{-2} \int_{\mathbf{t}(s)=\mathbf{y}}^{\mathbf{t}(s)=\boldsymbol{\mu}} (\mathbf{y} - \mathbf{t})^T \{\mathbf{V}(\mathbf{t})\}^{-1} d\mathbf{t}(s)$$

along a smooth path $\mathbf{t}(s)$ in R^n from $\mathbf{t}(s_0) = \mathbf{y}$ to $\mathbf{t}(s_1) = \boldsymbol{\mu}$, ordinarily depends on the particular path chosen. Evidently, if the integral is path-independent, the gradient vector of $Q(\boldsymbol{\mu}; \mathbf{y}, \cdot)$ with respect to $\boldsymbol{\mu}$ is $\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu})/\sigma^2$, and the gradient vector with respect to $\boldsymbol{\beta}$ is given by (9.5). The derivative matrix of $\mathbf{U}(\boldsymbol{\beta})$ is then symmetrical. Conversely it can be shown that if the derivative matrix of $\mathbf{U}(\boldsymbol{\beta})$ is symmetrical, the integral is path-independent for all paths of the form $\mathbf{t}(s) = \boldsymbol{\mu}(\boldsymbol{\beta}(s))$. Only if the line integral is independent of the path of integration does it make sense to use this function as a quasi-likelihood. Then, and only then, does quasi-likelihood estimation correspond to the maximization on the solution locus $\boldsymbol{\mu}(\boldsymbol{\beta})$ of a function defined pointwise for each $\boldsymbol{\mu} \in R^n$. We now investigate briefly the conditions on the covariance function that are required to make the integral path-independent.

The integral can be shown to be path-independent if the partial derivatives of the components of $\mathbf{V}^{-1}(\boldsymbol{\mu})$ with respect to $\boldsymbol{\mu}$ form an array that is symmetrical in all three directions—i.e. under permutations of the three indices. In other words, if $\mathbf{W} = \mathbf{V}^{-1}$, we require

$$\partial W_{ij}/\partial \mu_k = \partial W_{ik}/\partial \mu_j = \partial W_{jk}/\partial \mu_i.$$

A necessary and sufficient condition for this to hold is that $\mathbf{V}^{-1}(\boldsymbol{\mu})$ should be the second derivative matrix with respect to $\boldsymbol{\mu}$ of a scalar function $b^*(\boldsymbol{\mu})$, which is necessarily convex. The existence of a convex function $b^*(\boldsymbol{\mu})$ implies the existence of a canonical parameter $\boldsymbol{\theta}(\boldsymbol{\mu})$ and a cumulant function $b(\boldsymbol{\theta})$ defined on the dual space, such that

$$\boldsymbol{\theta} = b^{**}(\boldsymbol{\mu}), \quad \boldsymbol{\mu} = b'(\boldsymbol{\theta}) \quad \text{and} \quad \mathbf{V}(\boldsymbol{\mu}) = b''(\boldsymbol{\theta}). \quad (9.9)$$

These conditions exclude from consideration a large class of covariance functions that are physically unappealing for reasons discussed in section 9.2.1. However, some apparently physically sensible covariance functions are also excluded by the criterion.

In general it is not easy to construct covariance functions satisfying the above property even though the property itself is easy to verify. We now describe ways in which non-diagonal covariance functions satisfying (9.9) can be constructed. The integral can be shown to be path-independent if \mathbf{V}^{-1} can be written as the sum of matrices, each having the form $\mathbf{A}^T \mathbf{W}(\boldsymbol{\gamma}) \mathbf{A}$, in which \mathbf{A} is independent of $\boldsymbol{\mu}$, $\boldsymbol{\gamma} = \mathbf{A}\boldsymbol{\mu}$, and \mathbf{W} is diagonal of the form

$$\mathbf{W} = \text{diag}\{W_1(\gamma_1), \dots, W_n(\gamma_n)\}.$$

In other words we require a decomposition

$$\mathbf{V}^{-1}(\boldsymbol{\mu}) = \sum_{j=1}^k \mathbf{A}_j^T \mathbf{W}_j(\mathbf{A}_j \boldsymbol{\mu}) \mathbf{A}_j \quad (9.10)$$

in which each \mathbf{W}_j is a diagonal matrix of the required form as a function of its argument $\boldsymbol{\gamma}_j = \mathbf{A}_j \boldsymbol{\mu}$. In particular if $k = 1$ the covariance matrix can be diagonalized into the form (9.1). The latter condition can sometimes be verified directly for certain covariance functions—e.g. the multinomial covariance matrix.

In order to construct the quasi-likelihood function explicitly we may consider the straight-line path

$$\mathbf{t}(s) = \mathbf{y} + (\boldsymbol{\mu} - \mathbf{y})s$$

for $0 \leq s \leq 1$, so that $\mathbf{t}(0) = \mathbf{y}$ and $\mathbf{t}(1) = \boldsymbol{\mu}$. Provided that it exists, the quasi-likelihood function is given by

$$Q(\boldsymbol{\mu}; \mathbf{y}) = -(\mathbf{y} - \boldsymbol{\mu})^T \left\{ \sigma^{-2} \int_0^1 s \{ \mathbf{V}(\mathbf{t}(s)) \}^{-1} ds \right\} (\mathbf{y} - \boldsymbol{\mu}). \quad (9.11)$$

This integral is expressed directly in terms of the mean-value parameter and is sometimes useful for purposes of approximation. For example, if $\mathbf{V}^{-1}(\mathbf{t})$ is approximately linear in \mathbf{t} over the straight-line path from $\mathbf{t} = \mathbf{y}$ to $\mathbf{t} = \boldsymbol{\mu}$, the integral may be approximated by

$$\begin{aligned} Q(\boldsymbol{\mu}; \mathbf{y}) \simeq & -\frac{1}{3}(\mathbf{y} - \boldsymbol{\mu})^T \mathbf{V}^{-1}(\boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})/\sigma^2 \\ & -\frac{1}{6}(\mathbf{y} - \boldsymbol{\mu})^T \mathbf{V}^{-1}(\mathbf{y})(\mathbf{y} - \boldsymbol{\mu})/\sigma^2. \end{aligned} \quad (9.12)$$

The alternative, and more familiar expression,

$$Q(\boldsymbol{\mu}; \mathbf{y}) = \sigma^{-2} \{ \mathbf{y}^T \boldsymbol{\theta} - b(\boldsymbol{\theta}) - b^*(\mathbf{y}) \}$$

presupposes the existence of the required functions, and is less useful for computational purposes unless those functions are available in a computable form.

The integral (9.11) can be evaluated whether or not $\mathbf{V}(\boldsymbol{\mu})$ satisfies the conditions in (9.9). However the use of $Q(\boldsymbol{\mu}; \mathbf{y})$ as a test statistic or as a quasi-likelihood function is then less compelling because an additional argument is required to justify the choice of the path of integration.

9.3.3 Example: estimation of probabilities from marginal frequencies

The following rather simplified example concerns the estimation of voter transition probabilities based only on the vote totals for each of two parties, C and L , say, in two successive elections. Suppose for simplicity of exposition that the same electorate votes in two successive elections and that we observe only the vote totals for C and L at each election. For each constituency, the ‘complete data’, which we do not observe, may be displayed as follows:

		Election 2			
		Party	C		Total
Election 1		C	X_1	$m_1 - X_1$	
		L	X_2	$m_2 - X_2$	m_2
	Total		$Y = X.$	$m. - X.$	$m.$

Only the row and column totals of the above Table are observed.

Since interest is focused on transition probabilities, we condition on the observed vote totals at election 1, regarding the entries in the body of the table as random variables. The simplest binomial model takes $X_1 \sim B(m_1, \pi_1)$ and $X_2 \sim B(m_2, \pi_2)$ as independent random variables. Thus π_1 is the probability that a voter who votes for party C in the first election also votes for C in the second election. Similarly π_2 is the probability that a voter who previously voted L subsequently switches to C .

Taking $Y = X.$, together with m_1, m_2 as the observed response in each constituency, we have that

$$\begin{aligned} E(Y) &= m_1\pi_1 + m_2\pi_2 = \mu, \text{ say,} \\ \text{var}(Y) &= m_1\pi_1(1 - \pi_1) + m_2\pi_2(1 - \pi_2). \end{aligned}$$

Evidently $\text{var}(Y)$ is not a function of $E(Y)$ alone, and hence it is not possible to construct a quasi-likelihood function along the lines described in sections 9.2.2 or 9.3.2. Nevertheless, given sufficient data from several constituencies, we may still use the score function (9.5) to estimate the parameters, which in this case are π_1 and π_2 .

Suppose that data are available from each of n constituencies for which the transition probabilities may be assumed constant. Thus we have

$$\begin{aligned} E(\mathbf{Y}) &= \mathbf{M}\boldsymbol{\pi}, \\ \text{cov}(\mathbf{Y}) &= \text{diag}\{m_{i1}\pi_1(1 - \pi_1) + m_{i2}\pi_2(1 - \pi_2)\} = \mathbf{V}(\boldsymbol{\pi}), \end{aligned}$$

where \mathbf{M} is an $n \times 2$ matrix giving the total votes cast for each party in the first election. The quasi-likelihood score function (9.5) is

$$\mathbf{U}(\boldsymbol{\pi}) = \mathbf{M}^T \mathbf{V}^{-1}(\boldsymbol{\pi})(\mathbf{Y} - \mathbf{M}\boldsymbol{\pi}). \quad (9.13)$$

The two components of this vector are

$$\begin{aligned} U_1(\boldsymbol{\pi}) &= \sum_i m_{i1}(y_i - m_{i1}\pi_1 - m_{i2}\pi_2)/V_i(\boldsymbol{\pi}) \\ U_2(\boldsymbol{\pi}) &= \sum_i m_{i2}(y_i - m_{i1}\pi_1 - m_{i2}\pi_2)/V_i(\boldsymbol{\pi}). \end{aligned}$$

Using these expressions it may be verified that

$$\frac{\partial U_1}{\partial \pi_2} \neq \frac{\partial U_2}{\partial \pi_1},$$

showing that (9.13) cannot be the gradient vector of any scalar function $Q(\boldsymbol{\pi})$. The information matrix $\mathbf{i}_{\boldsymbol{\pi}} = \mathbf{M}^T \mathbf{V}^{-1} \mathbf{M}$ has rank 2 provided that the vote ratios m_{i1}/m_{i2} are not all equal.

In order to compare the quasi-likelihood estimates with possible alternatives, we suppose that the following meagre vote totals are observed in three constituencies.

Y	m_1	m_2
7	5	5
5	6	4
6	4	6

After iteration in the usual way, the quasi-likelihood estimate obtained is $\hat{\pi} = (0.3629, 0.8371)$. Thus the fitted values and the information matrix are

$$\hat{\mu} = M\hat{\pi} = \begin{pmatrix} 6.000 \\ 5.526 \\ 6.474 \end{pmatrix} \quad \text{and} \quad i_{\pi} = \begin{pmatrix} 41.4096 & 39.7904 \\ 39.7904 & 42.5357 \end{pmatrix}.$$

It is readily verified that these values satisfy the vector equation $U(\hat{\pi}) = 0$. The approximate standard errors of $\hat{\pi}_1$ and $\hat{\pi}_2$ are 0.489 and 0.482. The correlation, however, is given as -0.948, showing that the sum $\pi_1 + \pi_2$ is tolerably well estimated but there is little information concerning measures of difference such as $\pi_1 - \pi_2$, π_1/π_2 or $\psi = \pi_1(1 - \pi_2)/\{\pi_2(1 - \pi_1)\}$.

If all values in the above Table were increased by a factor of 100 the same parameter estimate and correlation matrix would be obtained. Standard errors of $\hat{\pi}$ would be reduced by a factor of 10.

The likelihood function in this problem for a single observation y is

$$\sum_j \binom{m_1}{j} \binom{m_2}{y-j} \pi_1^j (1 - \pi_1)^{m_1-j} \pi_2^{y-j} (1 - \pi_2)^{m_2-y+j}.$$

The log likelihood for the full data, which is the sum of the logarithms of such factors, is numerically and algebraically unpleasant. It can, however, be maximized using the EM algorithm as described by Dempster *et al.* (1977). A simpler direct method is described in Exercise 9.2. We find $\hat{\pi}_{ML} = (0.2, 1.0)$, on the boundary of the parameter space and rather different from the quasi-likelihood estimate. In both cases, however, $\hat{\pi}_1 + \hat{\pi}_2 = 1.2$, a consequence of the identity $\sum y_i = \sum \hat{\mu}_i$.

The Fisher information matrix and its inverse, evaluated at the maximum quasi-likelihood estimate $\hat{\pi} = (0.363, 0.837)$, are

$$I_{\pi} = \begin{pmatrix} 62.18 & 4.57 \\ 4.57 & 102.24 \end{pmatrix} \quad \text{and} \quad I_{\pi}^{-1} = \begin{pmatrix} 0.0161 & -0.0007 \\ -0.0007 & 0.0098 \end{pmatrix}.$$

Evidently in this case the maximum-likelihood estimator is considerably more efficient than the quasi-likelihood estimator, particularly for the estimation of differences. It is a curious fact that

the Fisher information matrix has rank 2 even when the matrix M has rank 1. Thus it is possible, in principle at least, for the quasi-likelihood estimates of certain contrasts to have negligible efficiency compared with the maximum-likelihood estimate. In both cases the estimated standard error of $\hat{\pi}_1 + \hat{\pi}_2$ is given as 0.1565. For further details see Exercises 9.2–9.3.

If all of the row totals m_{i1} and m_{i2} are equal across constituents then Y_1, \dots, Y_n are identically distributed and Y is essentially sufficient for π . It is still possible to estimate the odds ratio ψ because the variability of Y_i depends on ψ . It is this information, available in the likelihood function, that is discarded by the quasi-likelihood and accounts for the reduction in efficiency.

If all the observed values are increased by a factor of 100 the maximum-likelihood estimate changes to $\hat{\pi}_{ML} = (0.467, 0.733)$. The Fisher information matrix also changes in a moderately complicated way. Evidently the maximum-likelihood estimate is not a simple linear function of the data. This observation makes sense in the light of the comments in the previous paragraph.

9.4 Optimal estimating functions

9.4.1 Introduction

The quasi-score function (9.5) is a rather special case of what is known in Statistics as an estimating function. An estimating function $g(\mathbf{Y}; \boldsymbol{\theta})$ is a function of the data \mathbf{Y} and parameter $\boldsymbol{\theta}$ having zero mean for all parameter values. Higher-order cumulants of $g(\cdot; \cdot)$ need not be independent of $\boldsymbol{\theta}$, so that $g(\mathbf{Y}; \boldsymbol{\theta})$ need not be a pivotal statistic. Provided that there are as many equations as parameters, estimates are obtained as the root of the equation $g(\hat{\boldsymbol{\theta}}; \mathbf{Y}) = \mathbf{0}$.

Usually it is fairly straightforward to construct estimating functions. For example, taking $\boldsymbol{\beta}$ to be the parameter of interest in the context described in sections 9.2 and 9.3, $\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})$ is a vector-valued estimating function. The difficult part is to reduce this n -vector to a p -vector with minimal sacrifice of information. The theory of optimal estimating equations can be used to demonstrate that the score function (9.5) is in fact the optimal combination within the class of linear estimating functions.

If the observations form a time-ordered sequence Y_1, \dots, Y_n , it may be helpful to consider a corresponding sequence of elementary estimating functions of the form $g_t(Y_{(t)}; \boldsymbol{\theta})$, where $Y_{(t)} = (Y_1, \dots, Y_t)$ is the process up to time t . If the conditional expectation of $g_t(\cdot; \cdot)$ given the past history of the process satisfies

$$E\{g_t(Y_{(t)}; \boldsymbol{\theta}) | Y_{(t-1)}\} = 0,$$

the cumulative sequence is said to form a martingale. Evidently $g_t(\cdot; \cdot)$ and all linear combinations of the g_t s are estimating functions. Thus there is a close connection between the theory of martingales and the theory of estimating functions (Godambe and Heyde, 1987).

In the linear regression and related contexts we usually require the elementary estimating functions to have zero mean conditionally on the values of the design points or covariates. This is a stronger condition than simply requiring zero unconditional mean. To underline the role played by such conditioning variables we write

$$E\{g_i(\mathbf{Y}; \boldsymbol{\theta}) | A_i\} = 0,$$

where $A_i \equiv A_i(\mathbf{y}; \boldsymbol{\theta})$, to cover both regression and time-series problems. In the regression context $A_i = A$, the set of covariates. More generally, however, the sequence A_i must be nested in the sense that $A_{i-1} \subseteq A_i$. Usually it is desirable to choose A to have maximum dimension.

A useful property of estimating functions is that very often they are rather simple functions of the data. For example (9.5) is linear in \mathbf{Y} . Statistical properties of the estimate, $\hat{\boldsymbol{\theta}}$, which is a non-linear function of \mathbf{Y} , can frequently be deduced from the properties of the estimating function. We now give a very brief outline of the theory of non-linear estimating functions, concentrating on ways of combining elementary estimating functions.

9.4.2 Combination of estimating functions

Suppose that the observed random variables \mathbf{Y} have a distribution that depends on $\boldsymbol{\theta}$ and that, for each $\boldsymbol{\theta}$, $g_i(\mathbf{Y}; \boldsymbol{\theta})$ is a sequence of independent random variables having zero mean for all $\boldsymbol{\theta}$. For example if the Y s are generated by the autoregressive process

$$Y_t = \theta Y_{t-1} + \epsilon_t, \quad Y_0 = \epsilon_0,$$

where ϵ_t are *i.i.d.* $N(0, 1)$, we could take

$$g_t = Y_t - \theta Y_{t-1}$$

or, if $\theta \neq 0$,

$$g_t^* = Y_t/\theta - Y_{t-1}.$$

A second example in which the g_i are non-linear functions of \mathbf{Y} is given in the following section.

In order to combine the n elementary estimating functions into a single p -dimensional optimal estimating equation for $\boldsymbol{\theta}$, we define the following $n \times p$ matrix, \mathbf{D} , with components D_{ir} , which depend on both $\boldsymbol{\theta}$ and \mathbf{y} .

$$D_{ir} = -E\left\{ \frac{\partial g_i(\mathbf{Y}; \boldsymbol{\theta})}{\partial \theta_r} \mid A_i \right\}. \quad (9.14)$$

If \mathbf{V} is the (diagonal) conditional covariance matrix of g_i given A_i , we take as our estimating function for $\boldsymbol{\theta}$

$$\mathbf{U}(\boldsymbol{\theta}; \mathbf{y}) = \mathbf{D}^T \mathbf{V}^{-1} \mathbf{g}. \quad (9.15)$$

Since D_{ir} and \mathbf{V} are functions of the conditioning variables, and g_i has zero mean conditionally, it follows that $\mathbf{U}(\boldsymbol{\theta}; \mathbf{y})$ also has zero conditional mean.

The above estimating function is unaffected by linear transformation of the elementary estimating functions \mathbf{g} to $\mathbf{g}^* = \mathbf{B}\mathbf{g}$, where $\mathbf{B} \equiv \mathbf{B}(\boldsymbol{\theta})$ is a full-rank $n \times n$ matrix whose components may depend on A . Under this transformation we have

$$\mathbf{g}^* = \mathbf{B}\mathbf{g}, \quad \mathbf{V}^* = \mathbf{B}\mathbf{V}\mathbf{B}^T, \quad \mathbf{D}^* = \mathbf{B}\mathbf{D},$$

so that $\mathbf{D}^{*T} \mathbf{V}^{*-1} \mathbf{g}^* = \mathbf{D}^T \mathbf{V}^{-1} \mathbf{g}$ as claimed. Note that the components of \mathbf{g}^* are not independent.

In the autoregressive process described above, this procedure applied to the sequence g_t gives

$$U(\boldsymbol{\theta}; \mathbf{y}) = \sum_t Y_{t-1} g_t = \sum_t Y_{t-1} (Y_t - \theta Y_{t-1}).$$

When applied to the sequence g_t^* the same estimating function, which is also the log-likelihood derivative, is obtained by a slightly

more circuitous route. Note that, although g_t is linear in \mathbf{Y} , the final estimating function is quadratic in \mathbf{Y} .

By a modification of the argument given in section 9.2.3, the asymptotic variance of $\hat{\theta}$ is

$$\mathbf{i}_\theta^{-1} = \text{cov}(\hat{\theta}) = (\mathbf{D}^T \mathbf{V}^{-1} \mathbf{D})^{-1}.$$

Usually, and particularly so in the autoregressive model, it is better to use the observed value of this matrix rather than its expected value. The same recommendation applies in the example below.

Example: Fieller-Creasy problem

The following rather simple example is chosen to illustrate the fact that elementary estimating functions can frequently be chosen in such a way that incidental parameters are eliminated. Suppose that the observations come in pairs $\mathbf{Y}_i = (Y_{i1}, Y_{i2})$, which are independent with means $(\mu_i, \mu_i/\theta)$ and variances σ^2 . It is assumed that $\theta = E(Y_{i1})/E(Y_{i2})$ is the parameter of interest. From the stated assumptions it follows that

$$g_i(\mathbf{Y}_i; \theta) = Y_{i1} - \theta Y_{i2}$$

has mean zero and variance $\sigma^2(1 + \theta^2)$. The derivative of g_i with respect to θ is $-Y_{i2}$. Application of (9.14), taking $A_i(\theta) = Y_{i2} + \theta Y_{i1}$, gives the residual derivative

$$\begin{aligned} D_i &= Y_{i2} + (Y_{i1} - \theta Y_{i2})\theta/(1 + \theta^2) \\ &= (Y_{i2} + \theta Y_{i1})/(1 + \theta^2). \end{aligned}$$

The estimating function for θ is therefore

$$U(\theta) = \sum_i \frac{(Y_{i2} + \theta Y_{i1})(Y_{i1} - \theta Y_{i2})}{\sigma^2(1 + \theta^2)^2},$$

which is identical to the conditional log likelihood score statistic (7.3).

This score function ordinarily has two roots, one at $\hat{\theta}$, the other at $-1/\hat{\theta}$. One of these corresponds to what would be considered a maximum of the log likelihood: the other corresponds to a minimum. In any case, Normal approximation for $\hat{\theta}$ may be unsatisfactory unless the information is large. The alternative method

of generating confidence sets directly from the score statistic is preferable. In other words we take the set

$$\{\theta : |U(\theta)/i^{1/2}(\theta)| \leq k_{\alpha/2}^*\},$$

where the observed information for θ may be taken to be

$$i(\theta) = \sum \frac{(y_{i2} + \theta y_{i1})^2}{\sigma^2(1 + \theta^2)^3},$$

and $\Phi(k_{\alpha}^*) = 1 - \alpha$.

For an alternative argument leading to the same result see section 7.2.2.

9.4.3 Example: estimation for megalithic stone rings

Suppose that $\mathbf{Y}_i = (Y_{i1}, Y_{i2})$, $i = 1, \dots, n$, are the observed Cartesian coordinates of points in a plane, assumed to lie on or near the circumference of a circle with centre at (ω_1, ω_2) and radius ρ . We have in mind archaeological applications in which the ‘points’ are standing stones forming a curve that is assumed here to be an arc of a circle rather than a complete circle. For a statistical model we assume that the points are generated in the following way:

$$\begin{aligned} Y_{i1} &= \omega_1 + R_i \cos \epsilon_i \\ Y_{i2} &= \omega_2 + R_i \sin \epsilon_i \end{aligned} \tag{9.16}$$

in which R_1, \dots, R_n are positive independent and identically distributed random variables with mean ρ . The quantities $\epsilon_1, \dots, \epsilon_n$ may be regarded either as fixed (non-random) nuisance parameters or as random variables having a joint distribution independent of \mathbf{R} . It is undesirable and unnecessary in the archaeological context to assume that $\epsilon_1, \dots, \epsilon_n$ are identically distributed or mutually independent. Usually the stones are roughly equally spaced around the circle, or what remains of the circle, so the angles cannot be regarded as independent.

In order to construct an estimating equation for the parameters $(\omega_1, \omega_2, \rho)$ we observe first that, under the assumptions stated above,

$$\begin{aligned} g_i &= \{(Y_{i1} - \omega_1)^2 + (Y_{i2} - \omega_2)^2\}^{1/2} - \rho \\ &= R_i(\omega_1, \omega_2) - \rho \end{aligned}$$

are independent and identically distributed with mean zero conditionally on $A = (\epsilon_1, \dots, \epsilon_n)$.

The derivative vector of g_i with respect to $(\omega_1, \omega_2, \rho)$ is equal to $(\cos \epsilon_i, \sin \epsilon_i, 1)$, which is independent of g_i by assumption. Consequently the estimating functions for the three parameters are

$$\begin{aligned} \sum \frac{Y_{i1} - \omega_1}{\rho^2 R_i} (R_i - \rho) &= \cos \epsilon_i \times (R_i - \rho) / \rho^2 \\ \sum \frac{Y_{i2} - \omega_2}{\rho^2 R_i} (R_i - \rho) &= \sin \epsilon_i \times (R_i - \rho) / \rho^2 \\ \sum (R_i - \rho) / \rho^2, \end{aligned} \quad (9.17)$$

where we have taken $\text{var}(R_i) = \sigma^2 \rho^2$. Under the assumption that the ϵ s are independent of \mathbf{R} , or at least that $\cos \epsilon_i$ and $\sin \epsilon_i$ are uncorrelated with \mathbf{R} , these three functions all have zero mean. These equations are in fact the ordinary least-squares equations obtained by minimizing the sum of squares of the radial errors ignoring the angular errors.

We take as our estimate of $\rho^2 \sigma^2$ the mean squared radial error, namely

$$\hat{\rho}^2 \hat{\sigma}^2 = \sum (\hat{R}_i - \hat{\rho})^2 / (n - 3).$$

Note in this case that it would be a fatal mistake to use the unconditional expected value of the derivatives of g_i in the definition of the coefficient matrix (9.14). If the angles are identically distributed, not necessarily uniformly, around the circle then $E(\cos \epsilon)$ and $E(\sin \epsilon)$ both constant. The resulting estimating equations then have rank 1.

To illustrate this method of estimation we consider the data in Table 9.3, taken from Angell and Barber (1977). The Avebury ring has been studied by a number of authors including Thom (1967) and Thom, Thom and Foord (1976) who have divided the stones into four major groups, labelled A, B, C and W. From the diagram in Fig. 9.3 it can be seen that each of the individual arcs is quite shallow and that arc W is not immediately distinguishable from a straight line.

Table 9.3 gives the fitted centres and radii for each of the arcs considered separately. The final line gives the residual mean squared radial error, using degrees of freedom rather than sample

Table 9.3 Stone number and position in the Avebury ring[†]

Arc C			Arc W			Arc A			Arc B		
No.	x	y	No.	x	y	No.	x	y	No.	x	y
1	733.7	44.0	9	445.3	23.4	30	19.3	624.4	40	146.8	936.9
3	659.7	28.0	10	413.8	46.2	31	24.9	663.0	41	175.2	962.4
4	624.2	19.3	11	377.9	74.1	32	33.3	698.3	42	206.7	984.7
5	588.4	13.9	12	357.1	94.1	33	43.7	731.3	43	237.6	1002.9
6	551.6	12.3	13	327.7	112.4	34	55.5	764.4	44	270.3	1022.5
7	515.1	9.5	14	300.6	136.2	35	62.9	790.1	45	292.5	1031.2
8	478.0	16.6	15	272.0	158.8	36	69.2	815.0	46	315.8	1042.0
			16	243.5	183.0	37	85.0	849.8			
			17	216.3	205.0	38	98.5	884.6			
			18	188.9	229.8	39	123.6	910.5			
			19	163.5	255.5						
			20	140.0	285.0						
			21	120.6	305.7						
			22	103.1	323.1						
			23	85.9	344.0						
			24	61.8	371.3						
$\hat{\omega}_1 = 530.8$			$\hat{\omega}_1 = 1472.0$			$\hat{\omega}_1 = 795.0$			$\hat{\omega}_1 = 512.7$		
$\hat{\omega}_2 = 651.0$			$\hat{\omega}_2 = 1553.4$			$\hat{\omega}_2 = 516.5$			$\hat{\omega}_2 = 533.1$		
$\hat{\rho} = 638.8$			$\hat{\rho} = 1840.4$			$\hat{\rho} = 782.8$			$\hat{\rho} = 545.4$		
$\hat{\rho}^2 \tilde{\sigma}^2 = 5.60$			$\hat{\rho}^2 \tilde{\sigma}^2 = 3.78$			$\hat{\rho}^2 \tilde{\sigma}^2 = 9.00$			$\hat{\rho}^2 \tilde{\sigma}^2 = 0.72$		

[†]Data taken from Angell and Barber (1977).

size as divisor. In order to test whether the data are consistent with a single circle for arcs A, B and C, we fitted a circle to these three arcs together. The position of the fitted centre is shown in Fig. 9.3. The residual sum of squared radial errors is 878.8 on 21 degrees of freedom, whereas the pooled residual sum of squares from the separate fits is

$$4 \times 5.60 + 7 \times 9.00 + 4 \times 0.72 = 88.3$$

on 15 degrees of freedom. The increase in residual sum of squares is clearly statistically highly significant, showing that the three arcs are not homogeneous.

When models are fitted to shallow arcs, the parameter estimates tend to be highly correlated. For example the standard errors and

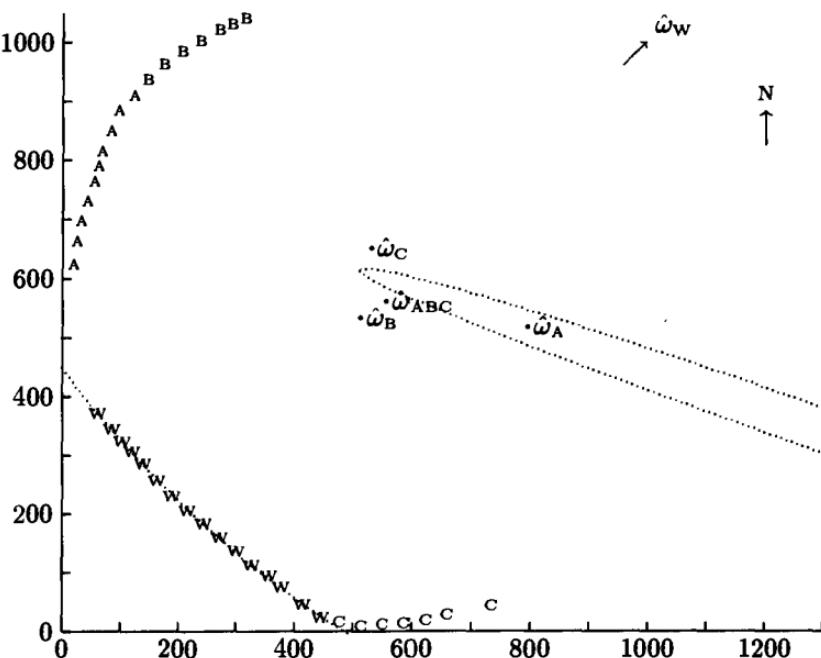


Fig. 9.3 Diagram of the Avebury ring showing the fitted centres of arcs A, B and C, together with the joint fit and the fitted arc W. An approximate 99% confidence set is shown for the centre of arc A. Distances in feet.

correlation matrix for arc C are

$$\begin{aligned} s.e.(\hat{\rho}) &= 108.4 \\ s.e.(\hat{\omega}_1) &= 14.57 \quad \text{and} \quad \begin{pmatrix} 1.0 & & \\ -0.87883 & 1.0 & \\ 0.99995 & -0.87584 & 1.0 \end{pmatrix} \\ s.e.(\hat{\omega}_2) &= 108.6 \end{aligned}$$

A confidence set for the position of the centre would be approximately elliptical with major axis almost vertical and very much larger than the minor axis. Such a 99% confidence set for ω_A , corresponding to $RSS \leq 181$, is shown in Fig. 9.3. The correlations are considerably smaller when the three arcs are combined. If the arc were a complete circle the correlations would be nearly zero.

This example has been chosen by way of illustration. In the archaeological context it is difficult to take the standard errors literally, but they may be useful in a qualitative way. The main statistical assumption, that the radial errors are independent, is not supported by residual plots of fitted residuals against stone

number. There is a noticeable but complicated residual pattern of positive serial dependence for arc W.

A version of model (9.16) for fitting ellipses is described in Exercise 9.5.

The estimates given here are the same as those obtained by Freeman (1977), but different from the two sets of estimates given by Thom, Thom and Foord (1976) and Angell and Barber (1977). For a comparison of various alternative methods of estimation from the viewpoint of consistency, efficiency and so on, see Berman and Griffiths (1985), Berman and Culpin (1986) and Berman (1987). A related, but less tractable model for fitting circles is discussed by Anderson (1981).

9.5 Optimality criteria

In order to justify the optimality claims made on behalf of the estimating functions (9.5) and (9.15) it is essential to state clearly the criterion used for making comparisons and also to state the class of estimators within which (9.5) and (9.15) are optimal. Evidently from the discussion in section 9.3.3 the quasi-likelihood estimate is sometimes less efficient than maximum likelihood, so that claims for optimality, even asymptotic optimality, cannot be global.

In keeping with the major theme of this chapter we consider first the class of linear estimating functions

$$\mathbf{h}(\mathbf{y}; \boldsymbol{\beta}) = \mathbf{H}^T(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) \quad (9.18)$$

where \mathbf{H} , of order $n \times p$, may be a function of $\boldsymbol{\beta}$ but not of \mathbf{y} . Clearly $\mathbf{h}(\mathbf{y}; \boldsymbol{\beta})$ is linear in \mathbf{y} for each $\boldsymbol{\beta}$. However the estimate $\tilde{\boldsymbol{\beta}}$, here assumed unique, satisfying $\mathbf{h}(\mathbf{y}; \tilde{\boldsymbol{\beta}}) = \mathbf{0}$, is ordinarily nonlinear in \mathbf{y} . We now demonstrate that, asymptotically at least, all linear functions of $\tilde{\boldsymbol{\beta}}$ have variance at least as great as the variance of the same linear function of $\hat{\boldsymbol{\beta}}$, i.e. $\text{var}(\mathbf{a}^T \tilde{\boldsymbol{\beta}}) \geq \text{var}(\mathbf{a}^T \hat{\boldsymbol{\beta}})$ where $\hat{\boldsymbol{\beta}}$ is the root of (9.5).

Under the usual asymptotic regularity conditions we may expand the estimating function in a Taylor series about the true parameter point giving

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \simeq (\mathbf{H}^T \mathbf{D})^{-1} \mathbf{h}(\mathbf{y}; \boldsymbol{\beta}),$$

where \mathbf{H} and \mathbf{D} are evaluated at the true parameter point. Thus the asymptotic covariance matrix of $\tilde{\boldsymbol{\beta}}$ is

$$\text{cov}(\tilde{\boldsymbol{\beta}}) \simeq \sigma^2 (\mathbf{H}^T \mathbf{D})^{-1} \mathbf{H}^T \mathbf{V} \mathbf{H} (\mathbf{D}^T \mathbf{H})^{-1},$$

where $D_{ir} = \partial\mu_i/\partial\beta_r$. The claim for asymptotic optimality of the quasi-likelihood estimate rests on the fact that the matrix

$$\text{cov}(\tilde{\boldsymbol{\beta}}) - \text{cov}(\hat{\boldsymbol{\beta}}) \simeq \sigma^2 (\mathbf{H}^T \mathbf{D})^{-1} \mathbf{H}^T \mathbf{V} \mathbf{H} (\mathbf{D}^T \mathbf{H})^{-1} - i_{\boldsymbol{\beta}}^{-1}$$

is non-negative definite for all \mathbf{H} . In order to demonstrate this claim we need only show that the difference between the precision matrices

$$\{\text{cov}(\hat{\boldsymbol{\beta}})\}^{-1} - \{\text{cov}(\tilde{\boldsymbol{\beta}})\}^{-1}$$

is non-negative definite (Exercise 9.7). This difference is equal to

$$\mathbf{D}^T (\mathbf{V}^{-1} - \mathbf{H}(\mathbf{H}^T \mathbf{V} \mathbf{H})^{-1} \mathbf{H}^T) \mathbf{D},$$

which is the residual covariance matrix of $\mathbf{D}^T \mathbf{V}^{-1} \mathbf{Y}$ after linear regression on $\mathbf{H}^T \mathbf{Y}$, and hence is non-negative definite. This completes the proof, that for large n , $\text{cov}(\tilde{\boldsymbol{\beta}}) \geq \text{cov}(\hat{\boldsymbol{\beta}})$ with the usual strong partial ordering on positive-definite matrices. The covariance matrices are equal only if \mathbf{H} is expressible as a linear combination of the columns of $\mathbf{V}^{-1} \mathbf{D}$.

The proof just sketched is formally identical to a proof of the Gauss-Markov theorem for linear estimators. The strength of this proof is that it applies to a considerably wider class of estimators than does the Gauss-Markov theorem: its weakness is that it is an asymptotic result, focusing on $\hat{\boldsymbol{\beta}}$ rather than on the score function directly.

The proof just given applies equally to the so-called non-linear estimating function (9.15) provided that we agree to make all probability calculations appropriately conditional. Provided that $A_i = A$, the same for each i , (9.15) is conditionally linear in \mathbf{g} . In other words $\hat{\boldsymbol{\beta}}$ is asymptotically conditionally optimal given A within the class of estimating functions that are conditionally linear in \mathbf{g} . One difficulty with this criterion is that there may well be some ambiguity regarding the best choice of A . The theory offers little guidance in this respect.

A variety of optimality conditions, including both fixed-sample and asymptotic criteria, are discussed by Godambe and Heyde (1987).

9.6 Extended quasi-likelihood

The discussion in sections 9.2, 9.3 has been concerned entirely with the fitting and comparison of regression models in which the variance function is known. The quasi-likelihood function (9.3) or, if it exists, (9.11), cannot be used for the formal comparison of models having different variance functions or different dispersion parameters. The properties listed in section 9.3.1 refer only to derivatives with respect to β and not with respect to σ^2 .

We have seen in section 9.2.4 how different variance functions can be compared graphically. The purpose of this section is to supplement and to formalize such comparisons via an extended quasi-likelihood. Then pairs of residual plots such as Figs 9.1a and 9.2 may be compared numerically via the extended quasi-likelihood as well as visually. The introduction of such an extended quasi-likelihood also opens up the possibility of modelling the dispersion parameter as a function of covariates (see Chapter 10).

In order to avoid difficulties of the kind encountered in section 9.3, we assume here that the observations are independent. The extended quasi-likelihood is then a sum over the n components of \mathbf{y} . For a single observation y we seek to construct a function $Q^+(\mu, \sigma^2; y)$ that, for known σ^2 , is essentially the same as $Q(\mu; y)$, but which exhibits the properties of a log likelihood with respect to σ^2 -derivatives. Thus we must have

$$\begin{aligned} Q^+(\mu, \sigma^2; y) &= Q(\mu; y) + h(\sigma^2; y) \\ &= -\frac{D(y; \mu)}{2\sigma^2} + h(\sigma^2; y) \end{aligned}$$

for some function $h(\sigma^2; y)$, which we take to be of the form

$$h(\sigma^2; y) = -\frac{1}{2}h_1(\sigma^2) - h_2(y).$$

In order for Q^+ to behave like a log likelihood with respect to σ^2 we must have $E(\partial Q^+ / \partial \sigma^2) = 0$. Thus

$$0 = \frac{1}{2\sigma^4} E\{D(Y; \mu)\} - \frac{1}{2}h'_1(\sigma^2),$$

implying that

$$\sigma^4 h'_1(\sigma^2) = E\{D(Y; \mu)\}. \quad (9.19)$$

To a rough first-order of approximation we have $E(D(Y; \mu)) = \sigma^2$, giving $h_1(\sigma^2) = \log(\sigma^2) + \text{const}$. Thus the extended quasi-likelihood function is given by

$$Q^+(\mu, \sigma^2; y) = -\frac{1}{2}D(y; \mu)/\sigma^2 - \frac{1}{2}\log\sigma^2. \quad (9.20)$$

This expression can be justified to some extent as a saddlepoint approximation for the log density provided that σ^2 is small and all higher-order cumulants are sufficiently small. To be explicit, suppose that the higher-order cumulants of Y are given by

$$\kappa_{r+1} = \kappa'_r \kappa_2, \quad \text{for } r \geq 2, \quad (9.21)$$

where $\kappa_2 = \sigma^2 V(\mu)$, and differentiation is with respect to μ . As shown in Exercise 2.1, (9.21) is a property of exponential-family distributions and of averages from such distributions in which σ^{-2} is an effective sample size. Thus $\kappa_3 = O(\sigma^4)$, $\kappa_4 = O(\sigma^6)$ and so on. The saddlepoint approximation for the log density is then

$$-\frac{1}{2}D(y; \mu)/\sigma^2 - \frac{1}{2}\log(2\pi\sigma^2 V(y)),$$

which differs from Q^+ by an additive function of y . See, for example, Barndorff-Nielsen and Cox (1979), Nelder and Pregibon (1987), Efron (1986), Jørgensen (1987) or McCullagh (1987, Chapter 6). Note that saddlepoint approximations depend on the entire set of cumulants, and not just on the low-order cumulants.

More accurate approximations can be obtained for $E(D(Y; \mu))$ provided that information is available concerning higher-order cumulants of Y . To a certain extent, however, this requirement violates the spirit of least squares, which is based on first and second moment assumptions only. Using the representation (9.4) it can be shown that

$$E(D(Y; \mu)) \simeq \sigma^2 + \frac{1}{12V^2} \{6\sigma^4 VV'^2 - 3\sigma^4 V^2 V'' - 4V' \kappa_3\}. \quad (9.22)$$

If (9.21) can be justified up to order 4, this expression may be reduced to

$$\begin{aligned} E(D(Y; \mu)) &\simeq \sigma^2 \{1 + (5\rho_3^2 - 3\rho_4)/12\}, \\ &= \sigma^2 \{1 + \sigma^2(2V'^2/V - 3V'')/12\}, \end{aligned}$$

where the standardized cumulants $\rho_3^2 = \kappa_3^2/\kappa_2^3$ and $\rho_4 = \kappa_4/\kappa_2^2$ are both $O(\sigma^2)$. By the same argument we find

$$\begin{aligned}\text{var}(D) &\simeq 2\kappa_2^2/V^2 = 2\sigma^4 \\ \text{cov}(D, Y) &\simeq (\kappa_3 - \kappa_2\kappa'_2)/V.\end{aligned}$$

The approximate covariance reduces to zero under the simplifying assumption (9.21) but not otherwise.

In what follows we assume that σ^2 is sufficiently small to justify the approximation $E(D) \simeq \sigma^2$. It follows then that the derivatives

$$\begin{aligned}\frac{\partial Q^+}{\partial \mu} &= \frac{Y - \mu}{\sigma^2 V(\mu)} \\ \frac{\partial Q^+}{\partial \sigma^2} &= \frac{D(Y; \mu)}{2\sigma^4} - \frac{1}{2\sigma^2}\end{aligned}$$

have zero mean and approximate covariance matrix

$$\begin{pmatrix} 1 & \frac{\kappa_3 - \kappa_2\kappa'_2}{2\sigma^6 V^2} \\ \frac{\kappa_3 - \kappa_2\kappa'_2}{2\sigma^6 V^2} & \frac{1}{2\sigma^4} \end{pmatrix}.$$

The expected value of the negative second derivative matrix is the same as the above except that the off-diagonal elements are zero. Note that if

$$\kappa_3 - \kappa_2\kappa'_2 = O(\sigma^4)$$

for small σ , the correlation of the two derivatives is $O(\sigma)$, and hence negligible. Consequently, with this entirely reasonable condition, Q^+ has the properties of a quasi-likelihood with respect to both mean parameter and dispersion parameter. Further, the Fisher information matrix for (μ, σ^2) is diagonal, a property that simplifies some calculations.

The argument just given is a partial justification for the use of the extended quasi-likelihood for the joint parameter (μ, σ^2) . The assumptions required are that σ^2 be small and that $\kappa_r(Y) = O(\sigma^{2(r-1)})$. Efron (1986) and Jørgensen (1987), using the stronger assumption (9.21), reach similar conclusions.

9.7 Bibliographic notes

The term quasi-likelihood seems first to have been used in this context by Wedderburn (1974) although calculations similar to those in section 9.2 appear also in unpublished work by Jarrett (1973).

Questions of efficiency and optimality, and to a lesser extent robustness, are addressed by Godambe (1960), Bhapkar (1972), Cox and Hinkley (1968), Morton (1981), Cox (1983), McCullagh (1983, 1984), Firth (1987), Godambe and Heyde (1987) and Hill and Tsai (1988).

Estimates having increased efficiency can sometimes be obtained by considering non-linear estimating functions or by combining two or more estimating functions. This subject has been studied by Jarrett (1973) and subsequently by Crowder (1987), Firth (1987) and Heyde (1987).

The problem discussed in section 9.3.3 has previously been studied by Firth (1982).

9.8 Further results and exercises 9

9.1 Suppose, conditionally on $M = m$ that $Y \sim P(m)$, the Poisson distribution with parameter m , and that M in turn has the gamma distribution $M \sim G(\alpha\nu, \nu)$ with mean $\mu = E(M) = \alpha\nu$ and coefficient of variation $\nu^{-1/2}$. Show that the unconditional mean and variance are $E(Y) = \mu = \alpha\nu$, and

$$\text{var}(Y) = \alpha\nu + \alpha^2\nu.$$

Suppose now that \mathbf{Y} has independent components generated in the above way with $\mu_i = E(Y_i)$ not all equal. Show that if $\nu_i = \nu$, a known constant, then the distribution of \mathbf{Y} has the natural exponential-family form with variance function $V(\mu) = \mu + \mu^2/\nu$, which is quadratic in μ . On the other hand if $\alpha_i = \alpha$, a constant, show that the variance function has the standard over-dispersed Poisson form $V(\mu) = \phi\mu$ with $\phi = 1 + \alpha$, but that \mathbf{Y} does not then have the linear exponential-family form.

More generally, if both α and ν vary according to the relations

$$\alpha_i = \theta + \psi\mu_i; \quad \nu_i^{-1} = \psi + \theta\mu_i^{-1},$$

show that $V(\mu) = \mu + \theta\mu + \psi\mu^2$ and that the distribution of \mathbf{Y} again does not have the linear exponential-family form. Compare the exact likelihood with the corresponding quasi-likelihood in the second and third cases.

9.2 Show that the likelihood function given at the end of section 9.3.3. can be written in the form

$$(1 - \pi_1)^{m_1} \pi_2^y (1 - \pi_2)^{m_2-y} P_0(\psi; m_1, m_2, y),$$

where $P_0(\psi)$ is defined in section 7.3.2. Hence show that the log-likelihood derivatives with respect to $\lambda_i = \text{logit}(\pi_i)$ are

$$\begin{aligned}\frac{\partial l}{\partial \lambda_1} &= \sum_i \{\kappa_1(\psi) - m_1 \pi_1\}, \\ \frac{\partial l}{\partial \lambda_2} &= \sum_i \{y - \kappa_1(\psi) - m_2 \pi_2\},\end{aligned}$$

where $\kappa_1(\psi)$ is the non-central hypergeometric mean, and summation runs over constituencies. Interpret these equations in terms of the EM algorithm.

9.3 Deduce that the maximum-likelihood estimates in the previous exercise satisfy $\sum y = \sum m_1 \hat{\pi}_1 + m_2 \hat{\pi}_2$. Show that the Fisher information matrix for $\boldsymbol{\theta} = (\pi_., \lambda_1 - \lambda_2)$ is a sum over constituencies of matrices of the form

$$\mathbf{I}_{\boldsymbol{\theta}} = \begin{pmatrix} m_1 V_1 + m_2 V_2 & (m_1 - m_2) V_1 V_2 \\ (m_1 - m_2) V_1 V_2 & (m_1 V_1 + m_2 V_2) V_1 V_2 - \kappa_2 V_.^2 \end{pmatrix} \frac{1}{V_.^2},$$

where $V_1 = \pi_1(1 - \pi_1)$, $V_2 = \pi_2(1 - \pi_2)$ and κ_2 is the hypergeometric variance, which depends on ψ . Under what conditions are these parameters orthogonal? Deduce that the Fisher information matrix has rank 2 even if $m_{i1} = m_{i2}$ for each i , but that $\lambda_1 - \lambda_2$ is not consistently estimated unless \mathbf{X} has rank 2.

9.4 Show that the integral along the straight-line path $\mathbf{t}(s)$ from $\mathbf{t}(s_0) = \mathbf{b}$ to $\mathbf{t}(s_1) = \mathbf{c}$ of the tangential component of the vector $\mathbf{A}^T \mathbf{t}$ is given by

$$\int_b^c \mathbf{t}^T \mathbf{A} d\mathbf{t}(s) = \frac{1}{2} (\mathbf{c} + \mathbf{b})^T \mathbf{A} (\mathbf{c} - \mathbf{b}).$$

Find the value of the integral along the path from \mathbf{b} to \mathbf{c} to \mathbf{d} and back to \mathbf{b} . Hence deduce that if $\mathbf{A} = \mathbf{A}^T$ the integral around the loop is zero. Conversely, deduce that if the integral around every closed loop is zero then \mathbf{A} must be symmetric.

9.5 Consider the model

$$\begin{aligned} Y_{i1} &= \omega_1 + \rho R_i \cos \epsilon_i \cos \phi - \lambda R_i \sin \epsilon_i \sin \phi \\ Y_{i2} &= \omega_2 + \rho R_i \cos \epsilon_i \sin \phi + \lambda R_i \sin \epsilon_i \cos \phi \end{aligned}$$

for an ellipse centered at (ω_1, ω_2) with semi-axes of length ρ, λ inclined at an angle ϕ to the x -axis. Assume that R_i are independent and identically distributed with mean 1, and independently of the ϵ_i s. Using the method described in section 9.4.3, construct an unbiased estimating function for the parameters $(\omega_1, \omega_2, \rho, \lambda, \phi)$.

Take as the elementary estimating functions

$$R_i - 1 = \left(\frac{X_{i1}^2}{\rho^2} + \frac{X_{i2}^2}{\lambda^2} \right)^{1/2} - 1,$$

where

$$\begin{aligned} X_{i1} &= (Y_{i1} - \omega_1) \cos \phi + (Y_{i2} - \omega_2) \sin \phi = \rho R_i \cos \epsilon_i \\ X_{i2} &= -(Y_{i1} - \omega_1) \sin \phi + (Y_{i2} - \omega_2) \cos \phi = \lambda R_i \sin \epsilon_i. \end{aligned}$$

Show that the required coefficients (9.14) are

$$\begin{aligned} D_{i1} &= \cos \epsilon_i \cos \phi / \rho - \sin \epsilon_i \sin \phi / \lambda, \\ D_{i2} &= \cos \epsilon_i \sin \phi / \rho + \sin \epsilon_i \cos \phi / \lambda, \\ D_{i3} &= \cos^2 \epsilon_i / \rho, \\ D_{i4} &= \sin^2 \epsilon_i / \lambda, \\ D_{i5} &= (\rho - \lambda) \cos \epsilon_i \sin \epsilon_i. \end{aligned}$$

Hence compute the information matrix for the five parameters.

9.6 Suppose that the covariance matrix \mathbf{V} can be written in the form

$$\mathbf{V} = \mathbf{D} \mathbf{R} \mathbf{D},$$

where \mathbf{R} is independent of $\boldsymbol{\mu}$ and

$$\mathbf{D} = \text{diag}\{D_1(\mu_1), \dots, D_n(\mu_n)\}.$$

Show that the quasi-likelihood function exists only if V_{ij} is independent of $\boldsymbol{\mu}$ for all $i \neq j$. In other words \mathbf{R} must be diagonal or \mathbf{D} must be independent of $\boldsymbol{\mu}$. [Section 9.3.2].

9.7 Let \mathbf{A} and \mathbf{B} be any two positive-definite matrices of the same order. Prove that

$$\mathbf{A} - \mathbf{B} \geq 0 \quad \text{implies} \quad \mathbf{A}^{-1} - \mathbf{B}^{-1} \leq 0,$$

where ≥ 0 means non-negative definite.

9.8 Suppose that the random variables Y_1, \dots, Y_n are independent with variance $\text{var}(Y_i) = \sigma^2 \mu_i^2$, where the coefficient of variation, σ , is unknown. Suppose that inference is required for β_1 , where

$$\log(\mu_i) = \beta_0 + \beta_1(x_i - \bar{x}_i).$$

Show that the quasi-likelihood estimates of β_0, β_1 are uncorrelated with asymptotic variances

$$\text{var}(\hat{\beta}_0) = \sigma^2/n, \quad \text{and} \quad \text{var}(\hat{\beta}_1) = \sigma^2 / \sum(x_i - \bar{x}_i)^2.$$

9.9 Suppose, for the problem described above, that a Normal-theory likelihood is used even though the data may not be Normal. Show that the ‘maximum likelihood’ estimates $\tilde{\beta}_0, \tilde{\beta}_1$ thus obtained are consistent under the assumptions stated. Show also that the true asymptotic variance of $\tilde{\beta}_1$, as opposed to the apparent value given by the Normal-theory log likelihood, is

$$\text{var}(\tilde{\beta}_1) = \frac{\sigma^2 \{1 + 2\sigma\rho_3 + \sigma^2(\rho_4 + 2)\} (1 + 2\sigma^2)^{-2}}{\sum(x_i - \bar{x})^2},$$

where $\rho_3 = \kappa_3/\kappa_2^{3/2}$ and $\rho_4 = \kappa_4/\kappa_2^2$ are the standardized third and fourth cumulants, assumed constant over i . For a range of plausible values of σ^2, ρ_3, ρ_4 , compare the efficiencies of the two methods of estimation.

Derive the asymptotic relative efficiency of $\tilde{\beta}_1$ to $\hat{\beta}_1$ under the assumption that the data are in fact Normal. [McCullagh, 1984b].

9.10 Suppose, conditionally on $Y_{i.} = m_i$, $Y_{.j} = s_j$, that $\mathbf{Y} = (Y_{11}, Y_{12}, Y_{21}, Y_{22})$ has the non-central hypergeometric distribution (7.9) with odds ratio ψ . Deduce that

$$g(\mathbf{Y}; \psi) = Y_{11}Y_{22} - \psi Y_{12}Y_{21}$$

is an unbiased estimating function for ψ . Show also that if $\psi = 1$

$$\text{var}(g(\mathbf{Y}; 1)) = \frac{m_1 m_2 s_1 s_2}{m_\cdot - 1}.$$

Hence deduce that for n independent 2×2 tables $\mathbf{Y}^{(i)}$ with common odds-ratio ψ

$$\sum_i \frac{Y_{11}^{(i)} Y_{22}^{(i)} - \psi Y_{12}^{(i)} Y_{21}^{(i)}}{m_\cdot^{(i)}}$$

is an unbiased estimating equation for ψ , but is not optimal unless $\psi = 1$.

The estimator produced by this function

$$\hat{\psi}_{\text{MH}} = \frac{\sum Y_{11}^{(i)} Y_{22}^{(i)} / m_\cdot^{(i)}}{\sum Y_{12}^{(i)} Y_{21}^{(i)} / m_\cdot^{(i)}}$$

is known as the Mantel-Haenszel estimator. Find an expression for the asymptotic variance as $n \rightarrow \infty$ of $\hat{\psi}_{\text{MH}}$ when $\psi = 1$. [Mantel and Haenszel (1959); Mantel and Hankey (1975); Breslow and Day (1980, p.240); Breslow (1981); Breslow and Liang (1982)].

9.11 Use the above estimating equation to estimate the common odds-ratio for the data in Table 7.2. Compare your estimate and its estimated variance with the values obtained in section 7.4.3.

Joint modelling of mean and dispersion

10.1 Introduction

In the models so far discussed the variance of a response has been assumed to take the form

$$\text{var}(Y_i) = \phi V(\mu_i),$$

in which $V(\mu)$ is a known variance function. The choice of variance function determines the interpretation of ϕ . For example if $V(\mu) = 1$, ϕ is the response variance: if $V(\mu) = \mu^2$, ϕ is the squared coefficient of variation (noise-to-signal ratio) of the response. Similarly for other variance functions. In the simplest of generalized linear models the dispersion parameter ϕ is a constant, usually unknown, but in circumstances where Y_i is the average of m_i elementary observations it may be appropriate to assume that ϕ_i is proportional to known ‘weights’ $w_i = 1/m_i$. For one example of this, see the discussion of the insurance-claims data in section 8.4.1. More generally it may be the case that ϕ_i varies in a systematic way with other measured covariates in addition to the weights. In this chapter, therefore, we explore the consequences of constructing and fitting formal models for the dependence of both μ_i and ϕ_i on several covariates, following a suggestion made by Pregibon (1984).

To a large extent the impetus for studying this extended class of models derives from the recent surge of interest in industrial quality-improvement experiments in which both the mean response and the signal-to-noise ratio are of substantive interest. For economy of effort, fractional factorial and related experimental designs are often used for this purpose. The aim very often is

to select that combination of factor levels that keeps the mean at a pre-determined ‘ideal’ value, while at the same time keeping the variability in the product at a minimum. It is thus necessary to study not just how the mean response is affected by factors under study, but also how the variance, or other suitable measure of variability such as the noise-to-signal ratio, is affected by these factors.

For grouped data or ordinal responses, model (5.4) is designed to achieve a similar purpose.

10.2 Model specification

The joint model is specified here in terms of the dependence on covariates of the first two moments. For the mean we have the usual specification

$$\begin{aligned} E(Y_i) &= \mu_i; \quad \eta_i = g(\mu_i) = \sum_j x_{ij}\beta_j, \\ \text{var}(Y_i) &= \phi_i V(\mu_i), \end{aligned} \tag{10.1}$$

in which the observations are assumed independent. The dispersion parameter is no longer assumed constant but instead is assumed to vary in the following systematic way.

$$\begin{aligned} E(d_i) &= \phi_i; \quad \zeta_i = h(\phi_i) = \sum_j u_{ij}\gamma_j \\ \text{var}(d_i) &= \tau V_D(\phi_i). \end{aligned} \tag{10.2}$$

In this specification $d_i \equiv d_i(Y_i; \mu_i)$ is a suitable statistic chosen as a measure of dispersion; $h(\cdot)$ is the dispersion link function; ζ is the dispersion linear predictor, and $V_D(\phi)$ is the dispersion variance function. The dispersion covariates u_i are commonly, but not necessarily, a subset of the regression covariates x_i .

Two possible choices for the dispersion statistic d_i are

1. the generalized Pearson contribution

$$d_i = r_p^2 = (Y_i - \mu_i)^2 / V(\mu_i)$$

2. the contribution to the deviance of unit i : $d_i = r_D^2$.

For Normal-theory models, but not otherwise, the two forms are equivalent. Note that, when evaluated at the true μ , $E(r_p^2) = \phi$ exactly whereas $E(r_D^2) \approx \phi$ only approximately.

To fit the extended model we must choose suitable dispersion variance and link functions. If Y is Normal d_i has the $\phi_i \chi_1^2$ distribution, so that a gamma model with $V_D(\phi) = 2\phi^2$ would be chosen. The most natural link functions include the identity, corresponding to additive variance components, and the log, corresponding to multiplicative effects of the covariates. If Y is non-Normal, adjustments to the dispersion model may be necessary to account for the bias in r_D^2 or for the excess variability of r_p^2 . Some possibilities are discussed in section 10.5.

The two models (10.1) and (10.2) are interlinked; that for the mean requires an estimate of $1/\phi_i$ to be used as prior weight, while the dispersion model requires an estimate of μ_i in order to form the dispersion response variable d_i . The form of the interlinking suggests an obvious algorithm for fitting these models, whereby we alternate between fitting the model for the means for given weights $1/\hat{\phi}_i$, and fitting the model for the dispersion using the response variable $d_i = d_i(Y_i, \hat{\mu}_i)$.

10.3 Interaction between mean and dispersion effects

If the data contain replicate observations for each combination of covariate values for the mean, then an estimate of the variance can be formed for each distinct point in the covariate space of the model for $E(Y)$. Suppose now that we fit p parameters in the model for the mean response. With a total of n' distinct x -values this leaves $n' - p$ contrasts having zero mean that contain information about the dispersion. The information from these $n' - p$ contrasts can then be combined with the replicate estimates to improve the model for the dispersion. The practical difficulty is that use of supposedly null contrasts presupposes that the model for the mean be substantially correct. For suppose, for example, that a continuous covariate contributes a term βx to the linear predictor for the mean, but β is small so that the effect is judged insignificant. Nonetheless its omission from the model for the mean may produce relatively large values of $(y - \hat{\mu})^2$ at the two ends of the x -scale and small values at the centre. This characteristic

pattern will appear as a quadratic effect of x in the dispersion model. Likewise, omission of an interaction between two factors in the linear predictor for the mean will result in the inflation of supposedly null contrasts used to model the dispersion.

The correct choice of variance function for the mean is also important if distortion of the dispersion model is to be avoided. Thus, in designing an experiment for modelling both mean and dispersion, it is advisable to have estimates of dispersion based on pure replicates. Information from null contrasts can then be combined with the information from replicate contrasts if they prove compatible.

10.4 Extended quasi-likelihood as a criterion

The extended quasi-likelihood, Q^+ , developed in section 9.6, provides a possible criterion to be maximized for the estimation of μ and ϕ and for measuring the goodness of fit. We write

$$-2Q^+ = \sum_1^n \frac{d_i}{\phi_i} + \sum_1^n \log(2\pi\phi_i V(y_i)), \quad (10.3)$$

where d_i are the deviance components in the model for the means, i.e.

$$d_i = 2 \int_{\mu_i}^{y_i} \frac{y_i - t}{V(t)} dt.$$

Suppose now that the two parts of the model are parameterized as $\mu = \mu(\beta)$ and $\phi = \phi(\gamma)$. Then, from equation (10.3) we see that the estimating equations for β are the Wedderburn quasi-likelihood equations

$$\sum_{i=1}^n \frac{y_i - \mu_i}{\phi_i V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0, \quad (10.4)$$

except that $1/\phi_i$ must now be included as a weight, the dispersion being non-constant.

The estimating equations for γ are given by

$$\sum_{i=1}^n \frac{d_i - \phi_i}{\phi_i^2} \frac{\partial \phi_i}{\partial \gamma_j} = 0. \quad (10.5)$$

These are the Wedderburn quasi-likelihood equations for $V(\mu) = \mu^2$ with the deviance component as response variable. Thus, so far as estimation is concerned, the use of Q^+ as an optimizing criterion is equivalent to assuming that the deviance component has a variance function of the form $V_D(\phi) = \phi^2$, regardless of the variance function for Y . This can only be approximately correct, so we now consider some adjustments to the estimating equations for the dispersion to allow for non-Normal Y and other factors.

10.5 Adjustments of the estimating equations

10.5.1 Adjustment for kurtosis

The estimating equations for the dispersion parameters obtained from Q^+ are the same as those that would be obtained by assuming that d_i has the $\phi_i \chi_1^2$ distribution, i.e. that we have gamma errors with a scale factor of 2 (gamma index $= \frac{1}{2}$). In fact, however, the variance of d_i often exceeds the nominal value of $2\phi_i^2$, and appropriate allowance should be made for this excess variability. The correct variance of $(Y - \mu)^2$ is

$$\text{var}\{(Y - \mu)^2\} = \kappa_4 + 2\kappa_2^2 = 2\kappa_2^2(1 + \rho_4/2),$$

where

$$\rho_4 = \kappa_4/\kappa_2^2$$

is the standardized fourth cumulant. Thus the variance of $r_p^2 = (Y - \mu)^2/V(\mu)$ is $2\phi^2(1 + \rho_4/2)$. To use this result we need to know the value of ρ_4 . However for over-dispersed Poisson and binomial distributions the adjustment can be made provided that the fourth cumulant of Y has a particular pattern in relation to the second cumulant. If condition (9.21) holds up to fourth order then

$$\kappa_2 = \phi V, \quad \kappa_3 = \phi^2 \frac{\partial V}{\partial \theta}, \quad \text{and} \quad \kappa_4 = \phi^3 \frac{\partial^2 V}{\partial \theta^2}.$$

Consequently ρ_3 and ρ_4 are expressible in terms of ϕ and the derivatives of the variance function as follows:

$$\rho_3 = \phi^{1/2} V'(\mu)/\{V(\mu)\}^{1/2} \quad \text{and} \quad \rho_4 = \phi V''(\mu) + \rho_3^2,$$

where primes denote differentiation with respect to μ .

Under similar assumptions the approximate mean and variance of the deviance contribution, r_D^2 , are

$$E(r_D^2) \simeq \phi(1+b)$$

$$\text{var}(r_D^2) \simeq 2\phi^2(1+b)^2,$$

where $b = b(\phi, \mu) = (5\rho_3^2 - 3\rho_4)/12$ is usually a small adjustment. In general the standardized cumulants $\rho_3^2 = \kappa_3^2/\kappa_2^3$ and ρ_4 depend on both μ and ϕ .

Expressions for these adjustments are shown in Table 10.1.

Table 10.1. Dispersion adjustments for some standard distributions

Distribution	$1 + \rho_4/2$	$b(\phi, \mu)$
Normal	1	0
Poisson [†]	$1 + \phi/(2\mu)$	$\phi/(6\mu)$
Binomial [‡]	$1 + \frac{\phi}{2m} \left(\frac{1 - 6\pi(1 - \pi)}{\pi(1 - \pi)} \right)$	$\frac{\phi}{6m} \left(\frac{1 - \pi(1 - \pi)}{\pi(1 - \pi)} \right)$
Gamma	$1 + 3\phi$	$\phi/6$
Inverse Gaussian	$1 + 15\phi\mu/2$	0

[†]with over-dispersion (Section 6.2.3)

[‡]with over-dispersion (Section 4.5)

The estimating equations for the dispersion parameters may be adjusted by incorporating $(1 + \rho_4/2)^{-1}$ or an estimate thereof as a prior weight. The use of such an adjustment has been proposed by Prentice (1988) in the context of over-dispersed binary data.

10.5.2 Adjustment for degrees of freedom

The dispersion estimating equations derived from Q^+ make no allowance for the fact that p parameters have been fitted to the means. The effect of fitting is to decrease the average size of the dispersion response variables d_i . A simple adjustment is to multiply the second term in Q^+ by ν/n , where $\nu = n - p$ is the residual degrees of freedom for the deviance. Thus, for the purpose of fitting the dispersion response model, we use the modified Q_M^+

defined by

$$-2Q_M^+ = \sum_i \frac{d_i}{\phi_i} + \frac{\nu}{n} \sum_i \log(\phi_i V(y_i)). \quad (10.6)$$

For a model in which the dispersion is constant, ($\phi_i = \phi$), this adjustment gives

$$\hat{\phi} = D/\nu$$

by analogy with the unbiased estimator of variance for Normal-theory linear models. More generally, this adjustment yields approximate restricted maximum likelihood estimates, which are widely preferred to unadjusted maximum-likelihood estimates for the estimation of variance components and covariance functions. See section 7.2. If, as is common, the dispersion link is the logarithm, the modification changes only the intercept, which is often of little interest. However, with a beta-binomial model, for which the dispersion factor is

$$\phi_i = 1 + \theta(m_i - 1),$$

Q_M^+ and the unmodified Q^+ give different estimates of θ .

10.5.3 Summary of estimating equations for the dispersion model

The preceding discussion indicates that there is a variety of minor variations among the possible estimating equations for fitting the dispersion model. There are at least $2^3 = 8$ variations based on the following:

1. choice between $d = r_P^2$ and $d = r_D^2$;
2. choice between prior weight 1 and $(1 + \rho_4/2)^{-1}$;
3. adjustment for degrees of freedom or not.

On balance it appears desirable to make the adjustment for degrees of freedom. Adjustment for kurtosis also seems to be desirable provided that a reasonably accurate estimate of ρ_4 is available. The choice between r_P^2 and r_D^2 is less clear-cut.

Note that Q^+ and Q_M^+ provide an optimizing criterion using extended quasi-likelihood for only two of these forms. For the remainder we must rely on the theory of optimum estimating equations. Further work is required to give guidance for selection among these alternatives. Yet another form is based on work by Godambe and Thompson (1988), which we describe next.

10.6 Joint optimum estimating equations

Beginning with the pair of elementary estimating functions

$$\begin{aligned} g_{1i} &= Y_i - \mu_i, \\ g_{2i} &= (Y_i - \mu_i)^2 - \phi_i V(\mu_i), \end{aligned}$$

both of which have zero mean, optimum estimating equations for both the regression parameters and the dispersion parameters may be derived using the method described in section 9.4. In order to derive these equations we require the covariance matrix of \mathbf{g} together with the expected derivative matrix of \mathbf{g} with respect to the parameters. In carrying out this differentiation it is convenient initially to take the parameters (μ_i, ϕ_i) to be unrestricted.

The covariance matrix of (g_{1i}, g_{2i}) is

$$\mathbf{V}_i = \begin{pmatrix} \kappa_2 & \kappa_3 \\ \kappa_3 & \kappa_4 + 2\kappa_2^2 \end{pmatrix},$$

in which $\kappa_2 \equiv \phi_i V(\mu_i)$ and the subscript i has been omitted from the cumulants. The inverse covariance matrix is

$$\mathbf{V}_i^{-1} = \frac{1}{\Delta} \begin{pmatrix} \kappa_4 + 2\kappa_2^2 & -\kappa_3 \\ -\kappa_3 & \kappa_2 \end{pmatrix},$$

where $\Delta = \det(\mathbf{V}_i) = \kappa_2^3(2 + \rho_4 - \rho_3^2)$.

The negative expected derivative matrix of (g_{1i}, g_{2i}) with respect to (μ_i, ϕ_i) is

$$\mathbf{D}_i = \begin{pmatrix} 1 & 0 \\ \phi V' & V \end{pmatrix}$$

with rows indexed by the components of \mathbf{g} . Thus

$$\mathbf{D}_i^T \mathbf{V}_i^{-1} = \frac{1}{\Delta} \begin{pmatrix} \kappa_4 + 2\kappa_2^2 - \kappa_3 \phi V' & \kappa_2 \phi V' - \kappa_3 \\ -\kappa_3 V & \kappa_2 V \end{pmatrix}.$$

Provided that the regression and dispersion models have no parameters in common, the estimating equations thus obtained for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ are

$$\sum_i \left\{ \frac{\kappa_4 + 2\kappa_2^2 - \kappa_3 \phi V'}{\Delta} g_{1i} + \frac{\kappa_2 \phi V' - \kappa_3}{\Delta} g_{2i} \right\} \frac{\partial \mu_i}{\partial \beta_j} = 0, \quad (10.7)$$

$$\sum_i (g_{2i} - \kappa_3 g_{1i}/\kappa_2) \frac{\kappa_2 V}{\Delta} \frac{\partial \phi_i}{\partial \gamma_r} = 0.$$

The subscript i has been omitted in all coefficients. These are not the same as the extended quasi-likelihood equations (10.4), (10.5), even after (10.5) is adjusted for kurtosis.

Note that if $\kappa_3 = \kappa_2\phi V'$, a property of exponential-family distributions, we have

$$\mathbf{D}^T \mathbf{V}^{-1} = \begin{pmatrix} \kappa_2^{-1} & 0 \\ -\kappa_3 V/\Delta & \kappa_2 V/\Delta \end{pmatrix}.$$

It follows that if the regression and dispersion models have no parameters in common then the estimating equations for the regression parameters are

$$\sum_i \frac{y_i - \mu_i}{\phi_i V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0, \quad j = 1, \dots, p,$$

which is identical to the quasi-likelihood equation (10.4).

10.7 Example: the production of leaf-springs for trucks

These data, taken from Pignatiello and Ramberg (1985), relate to the production of leaf springs for trucks. A heat treatment is to be designed such that the free height y of a spring in an unloaded condition is as close as possible in mean value to eight inches, while having as small a variability as possible. To this end a one half fraction of a 2^5 experiment, with each treatment combination replicated three times, was performed using the factors

- B : furnace temperature,
- C : heating time,
- D : transfer time,
- E : hold-down time,
- O : quench oil temperature.

The data are given in Table 10.2.

Of the five factors O is somewhat different from the others in that it is apparently less easily controlled. We shall nevertheless follow Nair and Pregibon (1988) by treating it in the same way as the others.

Table 10.2. Data for a replicated 2^{5-1} factorial experiment to investigate the free height of leaf springs[†]

Run	Factor levels					Free height		
	B	C	D	E	O			
1	-	-	-	-	-	7.78	7.78	7.81
2	+	-	-	+	-	8.15	8.18	7.88
3	-	+	-	+	-	7.50	7.56	7.50
4	+	+	-	-	-	7.59	7.56	7.75
5	-	-	+	+	-	7.94	8.00	7.88
6	+	-	+	-	-	7.69	8.09	8.06
7	-	+	+	-	-	7.56	7.62	7.44
8	+	+	+	+	-	7.56	7.81	7.69
9	-	-	-	-	+	7.50	7.25	7.12
10	+	-	-	+	+	7.88	7.88	7.44
11	-	+	-	+	+	7.50	7.56	7.50
12	+	+	-	-	+	7.63	7.75	7.56
13	-	-	+	+	+	7.32	7.44	7.44
14	+	-	+	-	+	7.56	7.69	7.62
15	-	+	+	-	+	7.18	7.18	7.25
16	+	+	+	+	+	7.81	7.50	7.59

[†]Source: Pignatiello and Ramberg (1985).

We require models for both the mean and dispersion effects with a view to finding the factor combination that minimizes the dispersion while keeping the mean close to the target value of eight inches. We begin by modelling the mean assuming homogeneity of the dispersion. The range of the response variable is small in relation to the mean response, so we are unlikely to find evidence to cast doubt on the assumptions of Normality or constancy of variance. In Fig. 10.1, where the run variances plotted against the run means, there is little evidence that the variance changes with the mean response.

A main effects model for the mean response shows that D has a negligible effect, and a model with all two-factor interactions shows the effect of E to be independent of the rest. Finally we arrive at a linear predictor of the form

$$M = (B + C).O + E.$$

Note that the defining contrast for this design is

$$I = BCDE.$$

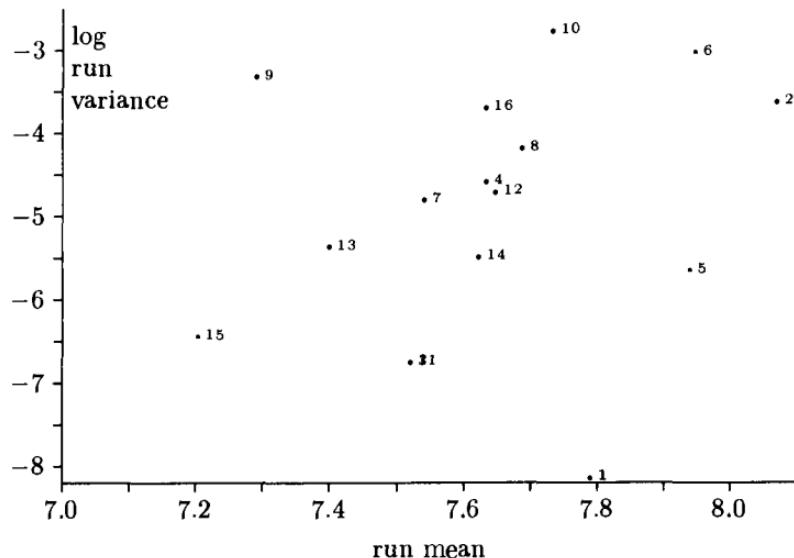


Fig. 10.1 *Run variances (log scale) plotted against run means for leaf-spring data.*

Thus the aliased pairs of two-factor interactions are

$$BD \equiv CE, \quad CD \equiv BE, \quad DE \equiv BC.$$

Fortunately none of these interactions is significant in this analysis.

A shortened analysis of variance table takes the form

Term	s.s.	d.f.	m.s.
$B + C + E + O$	1.898	4	0.4745
$B.O + C.O$	0.414	2	0.2071
Total M	2.312	6	0.3853
Rest of treatments	0.124	9	0.0138
Replicates	0.530	32	0.0166

The table of means for model M shows that C has virtually no effect at the higher level of O , but has a negative effect at the lower level. Increased furnace temperature (B) has a positive effect at all levels of O , but is twice as great at the higher level of O than at the lower level. Hold-down time (E) has a positive effect regardless of the other factors. The combination giving a fitted value closest to the target of eight inches is

$$(B, C, D, E, O) = (+, -, \pm, -, -), \text{ followed by } (+, -, \pm, +, -).$$

The fitted means are 7.953 and 8.057 respectively.

The analysis of dispersion is less clear cut. The original analysis by Pignatiello and Ramberg uses a linear model for the logarithms of the within-replicate sample variances. They selected the largest of the 15 factorial contrasts, which were B , DO , $BCO \equiv DEO$, $CD \equiv BE$ and $CDO \equiv BEO$. However, this selection procedure ignores the marginality conditions discussed in section 3.5, and carries a strong risk of selecting the accidentally large contrasts. A further objection to this analysis is that, with high probability, at least one of the sample variances will be exceptionally small. This may occur because of rounding. In the first run, for example, the sample variance is 0.0003, the next smallest being 0.0012. On the log scale, the sample variance for run 1 is exceptionally large and negative, which explains why such high-order interactions were found in the dispersion model.

If the original data were Normally distributed, the sample variances would be distributed as $s^2 \sim \phi \chi^2_2 / 2$ with $\phi = \sigma^2$. More generally, for samples of size $r = 3$, we have

$$E(s^2) = \kappa_2 = \phi, \quad \text{var}(s^2) = \frac{\kappa_4}{r} + \frac{2\kappa_2^2}{r-1} = \phi^2(1 + \rho_4/3).$$

Thus we treat the replicate variances as the response, using gamma errors and log link. If ρ_4 is taken as zero the distribution of s^2 is effectively exponential, so the scale factor should then be taken to be unity. The resulting fits show that only B and C have any appreciable dispersion effects. The deviances for selected dispersion models are shown together with the extended quasi-likelihood criterion in Table 10.3.

Table 10.3 Deviances for selected log-linear dispersion models fitted to the leaf-spring data

<i>Dispersion model</i>	<i>Gamma deviance</i>	<i>d.f.</i>	<i>Quasi-likelihood</i> $-2Q_M^+$	<i>d.f.</i>
1	26.57	15	106.9	31
B	20.58	14	101.1	30
C	22.99	14	103.4	30
$B + C$	16.08	13	96.4	29
$B + C + D + E + O$	15.00	10	95.4	26
$B.C$	15.89	12	96.3	28

The parameter estimates and standard errors for the log-linear model $B + C$ are

$$\begin{aligned}\hat{b} &= 1.369 \pm 0.514, \\ \hat{c} &= -1.092 \pm 0.514,\end{aligned}\tag{10.8}$$

showing that the variance at the higher furnace temperature (B) exceeds the variance at the lower temperature by an estimated factor of $\exp(1.369) \approx 3.9$. The effect of increased heating time (C) is to decrease the variance by the factor $\exp(-1.092) \approx 0.34$. The standard errors given above are based on the estimate $1 + \tilde{\rho}_4/3 = 1.056 = X^2/13$ obtained from the dispersion model $B + C$. Note that for exponential errors the expected value of the gamma mean deviance is approximately $7/6$, which is almost exactly what is observed!

Table 10.4 Deviances for selected log-linear dispersion models

Dispersion model	$-2Q_M^+$	d.f.
1	135.5	40
B	132.7	39
C	134.7	39
$B + C$	132.4	38
$B + C + D + E + O$	130.8	35
$B.E$	131.5	37

So far we have not used the information in the null contrasts for the means to augment the replicate variance estimates. We therefore repeat the above exercise using model $M = (B+C).O+E$ for the means. The results as shown in Table 10.4 are in conflict with those in Table 10.3. In particular C now has a negligible effect, and the effect of B is much reduced. The joint effect of B and C , which was highly significant in the analysis based on replicates (deviance reduction = 10.5), is now insignificant with a deviance reduction of only 3.1 on two degrees of freedom.

One possible explanation is that the null contrasts and the replicate contrasts are measuring variability of two different types or from different sources. To examine this possibility in more detail we now present an analysis of the null contrasts alone, ignoring the replicate contrasts. To accomplish this we replace all observations by their run means and fit the dispersion models as before. In this

Table 10.5 *Deviances for selected log-linear dispersion models using the between-runs contrasts*

<i>Dispersion model</i>	$-2Q_M^+$	<i>d.f.</i>
1	67.4	8
<i>B</i>	55.1	7
<i>C</i>	8.13	7
<i>B+C</i>	-3.16	6
<i>B.C</i>	-5.63	5

way the replicate variance is eliminated, and the analysis then uses information from the null contrasts alone.

The results of a sequence of fits are as shown in Table 10.5. There are insufficient data available for fitting the dispersion model $B+C+D+E+O$. These results show a very large effect for *C*, and a smaller but substantial effect for *B*. The parameter estimates for the dispersion model *B+C* are

$$\hat{b} = -1.820 \pm 0.943,$$

$$\hat{c} = 4.785 \pm 0.943,$$

and so are of opposite sign to those derived from the replicate-contrasts analysis (10.11). This reversal of sign explains the apparently small effects of *B* and *C* obtained from the combined analysis using all contrasts.

10.8 Bibliographic notes

The idea of using a linked pair of generalized linear models for the simultaneous modelling of mean and dispersion effects was first put forward by Pregibon (1984). For linear models with Normal errors the idea is much older, a simple case being that of heterogeneous variances defined by a grouping factor; see Aitkin (1987) for a general treatment or Cook and Weisberg (1983), who discuss score tests. Smyth (1985) compares different algorithms for the estimation of mean and dispersion effects.

10.9 Further results and exercises 10

10.1 Using expression (15.10) or (C.4) from Appendix C justify the claim following (10.8) that, for exponential observations, the expected value of the mean deviance is approximately 7/6.

10.2 Explain why it is necessary in (10.7) to impose the condition that the regression and dispersion models should have no parameters in common, although they may have covariates and factors in common. Discuss briefly whether this is a reasonable condition in practice.

10.3 Derive the results listed in Table 10.1 using the assumptions of section 10.5.1 for over-dispersed Poisson and binomial distributions.

10.4 Show that the expected Fisher information matrix derived from Q^+ for the parameters (β, γ) is block-diagonal.

CHAPTER 11

Models with additional non-linear parameters

11.1 Introduction

The word ‘non-linear’ in the chapter title is used in a specialized sense because, apart from the subset of classical linear models, all generalized linear models are in a strict sense non-linear. However, their non-linearity is limited in that it enters only through the variance function and the link function, the linearity of terms contributing to the linear predictor being preserved.

So far we have assumed that the variance is a known function of the mean, except possibly for a multiplicative dispersion parameter, and that the link function is also known. In this chapter we describe a number of models in which unknown parameters enter either the variance or the link function or both. In addition we also consider the use of terms in the linear predictor of an intrinsically non-linear type. One example of such a model occurred in section 8.4.4, where

$$\log \mu = \beta_0 + \beta_1 T + \beta_{-1}/(T - \delta)$$

was used as a model for the mean. Parameters such as δ are called non-linear in this specialized sense.

Intrinsically non-linear parameters complicate the fitting algorithm either by introducing an extra level of iteration or by introducing covariates that change at each iteration. Either of these effects may render convergence of the iterative process much less certain, and may also require starting values to be given for the non-linear parameters. In addition asymptotic covariances for the linear terms may be produced by the fitting algorithm conditional on fixed values of the non-linear parameters, and so need adjustment if uncertainties in the non-linear parameters are to be allowed for.

11.2 Parameters in the variance function

In the generalized linear models described in Chapters 3–8, five distributions for error were used. Two of these, the Normal and gamma, contain dispersion parameters explicitly. The discrete distributions in their standard forms do not contain such parameters, although here quasi-likelihood arguments extended the analysis to include an unknown dispersion parameter also. Provided that the dispersion parameter is constant, its value is not required in the solution of the likelihood equations for β . In that sense σ^2 plays a special role and its estimation is treated separately from the regression parameters.

The negative binomial distribution provides an example of a variance function containing an unknown parameter that is not a dispersion parameter. The distribution, which is discrete, can be written in the form

$$\text{pr}(Y = y; \alpha, k) = \frac{(y + k - 1)!}{y! (k - 1)!} \frac{\alpha^y}{(1 + \alpha)^{y+k}}; \quad y = 0, 1, 2 \dots$$

This may be contrasted with the expression in section 6.2.3 in which a different parameterization is used. In the above parameterization, the mean and variance are given by

$$\begin{aligned} E(Y) &= \mu = k\alpha, \\ \text{var}(Y) &= k\alpha + k\alpha^2 = \mu + \mu^2/k. \end{aligned}$$

The log likelihood can be written in the form

$$l = y \log\{\alpha/(1 + \alpha)\} - k \log(1 + \alpha) + (\text{function of } y, k),$$

which, for fixed k , has the form of a generalized linear model with canonical link

$$\eta = \log\left(\frac{\alpha}{1 + \alpha}\right) = \log\left(\frac{\mu}{\mu + k}\right),$$

and variance function

$$V = \mu + \mu^2/k.$$

The term μ can be thought of as the Poisson variance function and μ^2/k as the extra component arising from mixing the Poisson distribution with a gamma distribution for the mean to obtain the negative binomial.

Ordinarily k is not known *a priori*, and is clearly not a dispersion parameter. Estimates of k for single samples and for several samples have been discussed by Anscombe (1949), but we require an estimator for arbitrarily structured data. The maximum-likelihood estimate requires the solution of a non-linear equation involving the digamma function. Alternative estimators are those that make the mean deviance equal to unity or the Pearson X^2 statistic equal to its expectation.

Little use seems to have been made of the negative binomial distribution in applications; in particular the use of the canonical link is problematical because it makes the linear predictor a function of a parameter of the variance function. Note that if μ varies from observation to observation the above formulation assumes that, of the two parameters α and k in the mixing distribution, only α changes with k remaining fixed. See Manton *et al.* (1981) for an analysis of survey data in which both α and k are made to depend upon the classifying covariates; such a model, though of undoubtedly interest, lies outside the present framework.

For another example of additional parameters in the variance function, consider data that are to be modelled with gamma errors, but which have been collected with an absolute measurement (rounding) error, rather than with the desirable proportional error. With proportional rounding error or the retention of a fixed number of digits, the error variance retains the form $V = \sigma^2\mu^2$: with absolute rounding error, or the retention of a fixed number of decimal places, the variance function takes the form $V = \tau^2 + \sigma^2\mu^2$. The first term arises from the constant rounding error and the second from the assumed underlying gamma errors. The effect of this modified variance function is to reduce relatively the weight given to small observations. The quasi-likelihood model with this variance function would require the estimation of σ^2/τ^2 in the same way that k must be estimated for models using the negative binomial distribution.

Note that rounding from Z to Y has the effect of increasing the variance. In fact the rounding error, although numerically a

deterministic function of Z , is essentially statistically independent of Z , and not of Y . See Exercise 11.1.

11.3 Parameters in the link function

While link functions in generalized linear models are usually assumed known, it may be useful on occasion to assume that the link comes from a class of functions, members of the class being indexed by one or more unknown parameters. The goodness of fit expressed as a function of these parameters can then be inspected to see what range of parameter values is consistent with the data. If a particular value is of interest we can perform a goodness-of-link test (Pregibon, 1980) by comparing the deviance for that value with the deviance for the best-fitting value, or by using a score test.

11.3.1 One link parameter

A commonly considered class of link functions is that given by the power function, either in the form

$$\eta = \begin{cases} \mu^\lambda & \text{for } \lambda \neq 0, \\ \log \mu & \text{for } \lambda = 0, \end{cases}$$

or, in the form having continuity at $\lambda = 0$,

$$\eta = \frac{\mu^\lambda - 1}{\lambda}.$$

Exploration of this class of functions, used as transformations of the data rather than of the fitted values, was considered by Box and Cox (1964). For any fixed value of λ the model can be fitted with that power link function, and the deviance obtained in the usual way. When this is done for a range of λ -values, the deviances may be plotted against λ to display the range of λ -values that are most consistent with the observed data (and the model formula used).

If we wish to optimize over λ we can adopt the linearizing strategy proposed by Pregibon (1980), whereby we expand the link

function in a Taylor series about a fixed λ_0 and take only the linear term. Thus for the power family we have

$$\begin{aligned} g(\mu; \lambda) &= \mu^\lambda \simeq g(\mu; \lambda_0) + (\lambda - \lambda_0)g'_\lambda(\mu; \lambda_0) \\ &= \mu^{\lambda_0} + (\lambda - \lambda_0)\mu^{\lambda_0} \log \mu, \end{aligned} \quad (11.1)$$

so that we can approximate the correct link function $\eta = \mu^\lambda$ by

$$\begin{aligned} \eta_0 &= \mu^{\lambda_0} = \mu^\lambda - (\lambda - \lambda_0)\mu^{\lambda_0} \log \mu \\ &= \sum \beta_j x_j - (\lambda - \lambda_0)\mu^{\lambda_0} \log \mu. \end{aligned}$$

Given a first estimate λ_0 of λ , with corresponding fitted values $\hat{\mu}_0$, we then extend the linear predictor to include the covariate $-\hat{\mu}_0^{\lambda_0} \log \hat{\mu}_0$. Its parameter estimate gives the first-order adjustment to λ_0 . The reduction in deviance from its inclusion gives a test of whether λ_0 is an acceptable value for λ . To obtain the maximum-likelihood estimate for λ we repeat the above process forming a new adjusted value for λ at each stage. Convergence is not guaranteed, however, and requires that λ_0 , our starting value, be sufficiently close to $\hat{\lambda}$ for the linear expansion (11.1) to be adequate. To obtain convergence, an inner iteration, whereby the extra covariate's values are refined for fixed λ_0 , may be needed. Pregibon (1980) comments 'that the method is likely to be most useful for determining if a reasonable fit can be improved, rather than for the somewhat more optimistic goal of correcting a hopeless situation'.

Figure 11.1 shows the effect with the car-insurance data of changing the link by varying λ in the power family $\eta = \mu^\lambda$. The linear predictor contains the main effects only and the variance function is taken as $V(\mu) \propto \mu^2$. The minimum deviance of 124.51 occurs near $\lambda = -1$, corresponding to the reciprocal link originally chosen for the analysis, though the 95% limits

$$\left\{ \lambda : \text{dev}(\lambda) - 124.51 < \frac{124.51}{108} \times 3.93 \right\}$$

show an appreciable range of compatible values for λ , including zero, corresponding to the log link.

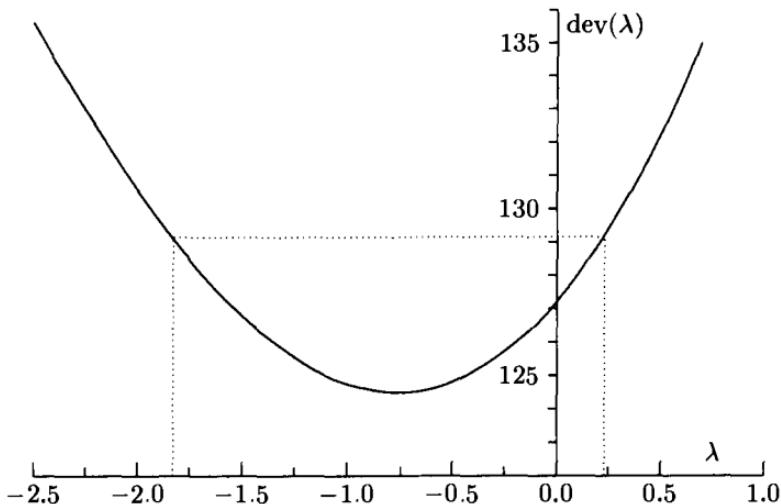


Fig. 11.1 *Car insurance data: deviance for varying λ in the link function $\eta = \mu^\lambda$, with nominal 95% confidence limits. Linear predictor includes main effects, and variance function $\propto \mu^2$.*

11.3.2 More than one link parameter

The above method extends in principle to more than one parameter in the link function. For each parameter λ , we add an extra covariate

$$-\left(\frac{\partial g}{\partial \lambda}\right)_{\lambda=\lambda_0}$$

to the model matrix and its parameter estimate gives the first-order adjustment to the starting value λ_0 . Pregibon (1980) discusses two examples with two parameters; the first is given by

$$g(\mu; \alpha, \lambda) = \{(\mu + \alpha)^\lambda - 1\}/\lambda,$$

i.e. a shifted form of the power family, indexed by λ , but with an added unknown origin α . Note that

$$g(\mu; 1, 1) = \mu,$$

so that the identity link is a member of the family.

The second of Pregibon's examples is useful for models based on tolerance distributions, such as probit analysis. The generalized link function is given by

$$g(\mu; \lambda, \delta) = \frac{\pi^{\lambda-\delta} - 1}{\lambda - \delta} - \frac{(1 - \pi)^{\lambda+\delta} - 1}{\lambda + \delta}$$

when π is the proportion responding, i.e. μ/m . The family contains the logit link as the limiting form

$$\lim_{\lambda, \delta \rightarrow 0} g(\mu; \lambda, \delta).$$

The one-parameter link family for binomial data,

$$g(\mu; \lambda) = \log \left[\{(1/(1 - \pi))^{\lambda} - 1\}/\lambda \right],$$

contains both the logistic ($\lambda = 1$) and complementary log-log link ($\lambda \rightarrow 0$) as special cases. This family may be used to assess the adequacy of an assumed linear logistic model against alternatives in the direction of the complementary log-log link.

11.3.3 Transformation of data vs transformation of fitted values

Transformations of fitted values through link functions (whether with or without unknown parameters) must be distinguished from transformations of the data values. The latter are fully discussed in Box and Cox (1964), who deal in particular with the power family. In using a function $g(y)$, rather than y , in the analysis we usually seek a transformation that yields simultaneously additive effects in the systematic part of the model and constancy of variance for the random part. Such a search may be successful; see, for example, the data set on survival times, given by Box and Cox, where a reciprocal transformation of the data allowed a linear regression model to be applied. However, there is no guarantee that both properties will result from the same transformation. Thus Nelder and Wedderburn (1972) in their reanalysis of the tuberculin-test data of Fisher (1949) (see the example in section 6.3.1) show that while a square-root transformation produces desirable error properties, a log transformation is required for additivity of effects. It is an advantage of generalized linear models over data transformation methods that the transformation to produce additivity can be made through the link function quite independently of any transformation of the data to produce approximate Normality or constancy of variance. Indeed the latter is itself often rendered unnecessary by the possibility of using a variance function other than a constant. Thus with the Fisher data mentioned above analysis of

Y with variance function $V \propto \mu$

and of

$Y^{1/2}$ with variance function $V = \text{constant}$,

using a log link for each, produce effectively identical results (Baker and Nelder, 1978, Appendix D).

11.4 Non-linear parameters in the covariates

As remarked in section 3.3.1 a function of x , such as e^{kx} , is an acceptable covariate in a linear predictor, provided that k is known; we simply use the values of e^{kx} in place of x in the model matrix. However, if k is to be estimated from the data, then non-linearity arises. Box and Tidwell (1962) describe a fitting technique by linearization, which follows closely that for non-linear parameters in the link function described above. If $g(x; \theta)$ is the covariate to be used, with θ unknown, we expand about an initial value θ_0 to give the linear approximation

$$g(x; \theta) \simeq g(x; \theta_0) + (\theta - \theta_0)[\partial g / \partial \theta]_{\theta=\theta_0}.$$

Thus if a non-linear term in the linear predictor is given by

$$\beta g(x; \theta),$$

we replace it by two linear terms

$$\beta u + \gamma v,$$

where

$$u = g(x; \theta_0), \quad v = [\partial g / \partial \theta]_{\theta=\theta_0} \quad \text{and} \quad \gamma = \beta(\theta - \theta_0).$$

An extra level of iteration is again required, and after fitting a model including u and v as covariates we obtain

$$\theta_1 = \theta_0 + \hat{\gamma}/\beta$$

as the improved estimate, and iterate. Convergence is not guaranteed for starting values arbitrarily far from the solution. If the process does converge then the presence of the extra term γv ensures that the asymptotic covariances produced for the remaining

parameters are correctly adjusted for the fitting of θ . If we wish to obtain the asymptotic variance of $\hat{\theta}$ directly, we need a final iteration with $\hat{\beta}v$ in the place of v ; the components of $(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$ corresponding to that covariate then give the approximate variance of $\hat{\theta}$ and its covariances with the other parameters.

While this technique is undoubtedly useful, and indeed probably under-used, it is usually unwise to try to include more than a very few non-linear parameters in this way, especially when the other covariates are themselves appreciably correlated in the data set. It will usually be found that estimates of the non-linear parameters have large sampling errors, and are highly correlated with the linear parameters and perhaps with each other. This is especially likely to be so in models where the systematic part consists of sums of exponentials of the form

$$\beta_0 + \beta_1 e^{k_1 x_1} + \beta_2 e^{k_2 x_2},$$

with the k s in the exponents requiring to be estimated as well as the β s.

One example where non-linear parameters arise in a fairly natural way concerns models for the joint action of a mixture of drugs (see the example in section 11.5.3). Here, apart from a single parameter that enters the model non-linearly, the model is of the generalized linear type with covariate $\log(x_1 + \theta x_2)$, where x_1 and x_2 are the amounts of the two drugs in the mixture. One method of analysis is to use the linearizing technique described above. Alternatively, following Darby and Ellis (1976), we may maximize the likelihood for various values of θ and plot the residual sum of squares $\text{RSS}(\theta)$ against θ , thereby obtaining a profile deviance curve similar to that shown in Fig. 11.1. The minimum, usually unique, gives $\hat{\theta}$ and the residual mean deviance $s^2 = D(\hat{\theta})/(n - p)$, where p is the total number of parameters, including θ . Approximate confidence limits for θ can be found from

$$\{\theta : \text{RSS}(\theta) - \text{RSS}(\hat{\theta}) < s^2 F_{1,n-p,\alpha}^*\},$$

where $F_{1,n-p,\alpha}^*$ is the upper $100(1 - \alpha)$ percentage point of the F -distribution on 1 and $n - p$ degrees of freedom.

Unfortunately this method of analysis does not allow the covariances of the parameter estimates, allowing for the uncertainty in θ , to be calculated easily. If these are required, the linearization technique should be used (see example in section 11.5.3).

11.5 Examples

11.5.1 The effects of fertilizers on coastal Bermuda grass

Welch *et al.* (1963) published the results of a 4^3 factorial experiment with the three major plant nutrients, nitrogen (N), phosphorus (P) and potassium (K), on the yield of coastal Bermuda grass. The experiment was performed to produce a response surface for the effects of the three nutrients, so that an optimal dressing could be predicted. The four levels for the three factors (all in lb/acre) were:

<i>Levels</i>	1	2	3	4
N	0	100	200	400
P	0	22	44	88
K	0	42	84	168

The grass was cut about every five weeks and oven-dried. The yields (in tons/acre) averaged over the three years 1955–57 and three replicates are shown in Table 11.1 with the factor levels coded 0, 1, 2, 3. Inspection of the data shows a hyperbolic type of response to the nutrients but the yield for the (0,0,0) plot shows that it will be necessary to allow for the nutrients already present in the soil if inverse polynomials (Section 7.3.3) are to be used to describe the response surface. We consider, therefore, an inverse linear response surface with

$$1/\mu = \eta = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \beta_3 u_3,$$

where $u_i = 1/(x_i + \alpha_i)$, $i = 1, 2, 3$. Here x_i ($i = 1, 2, 3$) are the applied amounts of N, P and K respectively, while α_i are the (unknown) amounts in the soil. If we assume for the moment that the proportional standard deviation of the yield is constant, we have a model with gamma errors and the reciprocal (canonical) link, but with three non-linear parameters α_1 , α_2 and α_3 to be estimated. The linearizing technique of section 11.4 leads to our adding the extra covariates $v_i = \partial u_i / \partial \alpha_i = -u_i^2$ to the model, giving the corrections

$$\delta a_i = c_i/b_i; \quad i = 1, 2, 3,$$

Table 11.1 *Yields of coastal Bermuda grass as affected by N, P and K*

<i>Nitrogen</i> (N)	<i>Phosphorus</i> (P)	<i>Potassium (K)</i>			
		0	1	2	3
0	0	1.98	2.13	2.19	1.97
0	1	2.38	2.24	2.10	2.60
0	2	2.18	2.56	2.22	2.47
0	3	2.22	2.47	2.94	2.48
1	0	3.88	3.91	3.66	4.07
1	1	4.35	4.59	4.47	4.55
1	2	4.14	4.36	4.55	4.35
1	3	4.26	4.72	4.83	4.85
2	0	4.40	4.91	5.10	5.23
2	1	5.01	5.64	5.68	5.60
2	2	4.77	5.69	5.80	6.07
2	3	5.17	5.45	5.85	6.43
3	0	4.43	5.31	5.15	5.87
3	1	4.95	6.27	6.49	6.54
3	2	5.22	6.27	6.35	6.72
3	3	5.66	6.24	7.11	7.32

Data from Welch *et al.* (1963).

to the current estimates of α_i in each iteration, where a_i is the estimate of α_i , c_i is the coefficient of v_i and b_i that of u_i . Starting values are required for a_i and these can be obtained by taking reciprocals of the data, forming N, P and K margins and plotting these against u_i for various trial values of α_i . The following suggested values are obtained

$$a_1 = 40, \quad a_2 = 22, \quad a_3 = 32.$$

Six iterations refine these to

$$a_1 = 44.60, \quad a_2 = 15.56, \quad a_3 = 32.39,$$

with a final deviance of 0.1965 on 57d.f., corresponding to a percentage standard deviation per observation of 5.9. The X^2 statistic is 0.1986, trivially different from the deviance, as is typical when the coefficient of variation is small. A final iteration with v_i replaced by $b_i v_i$ enables us to obtain the asymptotic standard errors

of the a_i directly. The parameter estimates with their standard errors, are given by:

b_0	0.09746 ± 0.00963
b_1	13.5 ± 1.350
b_2	0.7007 ± 0.457
b_3	1.336 ± 0.956
a_1	44.6 ± 4.18
a_2	15.6 ± 8.44
a_3	32.4 ± 19.1

These agree closely with those given by Nelder (1966) who used an approximate non-iterative method.

The correlation matrix shows high correlations between the a_i and b_i . These are respectively

$$0.9702, \quad 0.9850, \quad 0.9849,$$

and reflect the fact that if the a_i are taken as known the standard errors of the b_i are reduced by factors of from 4 to 6. Note too the large standard errors for a_2 and a_3 , which do not exclude impossible negative values; the assumptions of (asymptotic) Normality must be treated cautiously here.

The inverse linearity of the response may be tested by including the inverse quadratic terms $(x_i + \alpha_i)$ in the model. This gives a deviance of 0.1938 with 54d.f., a negligible reduction. The Pearson residuals $(y - \hat{\mu})/\hat{\mu}$ show one possible outlier, the yield 2.94 for levels (0, 3, 2). The fitted value is 2.43, so that the possibility of a transposition at some stage from 2.49 to 2.94 might be investigated. Omission of this point does not change the fit greatly, the largest effect being on b_2 . A plot of the residuals against fitted values does not contradict the assumption of gamma errors.

As shown in Nelder (1966), the quadratic polynomial with 10 parameters fits less well than the inverse linear surface with unknown origins, which has seven parameters. The latter is also additive for the three nutrients whereas the quadratic polynomial requires all the two-factor interaction terms for an adequate fit.

11.5.2 Assay of an insecticide with a synergist

The data for this example, shown in Table 11.2, are taken from a forthcoming paper by Morse, McKinlay and Spurr on the estimation of lowest-cost mixtures of insecticides and synergists. They relate to assays on a grasshopper *Melanopus sanguinipes* (F.) with the insecticide carbofuran and the synergist piperonyl butoxide (PB), which enhances the toxicity of the insecticide. The first model to be tried is of a type suggested by Hewlett (1969) and having the form of a logit link and binomial error with 'linear' predictor given by

$$\eta = \alpha + \beta_1 x_1 + \frac{\beta_2 x_2}{\delta + x_2},$$

where x_1 is the log dose of insecticide and x_2 is the dose of the synergist PB. The effect of the synergist is thus modelled as affecting the intercept by adding a hyperbolic term tending to β_2 for large x_2 . The slope of β_1 is assumed unaffected by the amount of PB. If δ were known we could set $u = x_2/(\delta + x_2)$ and a generalized linear model would result. To estimate δ we set up u for some starting value of δ and include the derivative $\partial u / \partial \delta = -u^2/x_2$ as a further covariate. Starting with $\delta = 1$ the standard process converges in four iterations to $\hat{\delta} = 1.763$ and a deviance of 53.34 with 11d.f. The fit is poor with a deviance nearly five times the base level of 11.

Inspection of the residuals shows that the major part of the discrepancy comes from the low doses of insecticide where the fitted kills are all considerably greater than those measured. The alternative links, probit and complementary log-log, give very similar results, suggesting that the log dose is not a satisfactory scale for the insecticide. The low kills for the low doses suggest that there may be a threshold value for the insecticide, and we can test this by putting a second non-linear parameter θ in the model to represent the threshold. The model now takes the form

$$\eta = \alpha + \beta_1 \log(z - \theta) + \beta_2 x_2 / (\delta + x_2),$$

where z is the dose of insecticide. Given current values, θ_0 and δ_0 , of θ and δ , the linearized form is given by

$$\eta = \alpha + \beta_1 \log(z - \theta_0) - \gamma_1 \left(\frac{1}{z - \theta_0} \right) + \beta_2 \left(\frac{x_2}{\delta_0 + x_2} \right) - \gamma_2 \frac{x_2}{(\delta_0 + x_2)^2}$$

Table 11.2 Data from assay on insecticide and synergist

<i>Number killed,</i> <i>y</i>	<i>Sample size,</i> <i>m</i>	<i>Dose of</i> <i>insecticide</i>	<i>Dose of</i> <i>synergist</i>
7	100	4	0
59	200	5	0
115	300	8	0
149	300	10	0
178	300	15	0
229	300	20	0
5	100	2	3.9
43	100	5	3.9
76	100	10	3.9
4	100	2	19.5
57	100	5	19.5
83	100	10	19.5
6	100	2	39.0
57	100	5	39.0
84	100	10	39.0

Data courtesy of Drs Morse, McKinlay and Spurr of Agriculture Canada.

With starting values $\delta_0 = 1.76$ from the first model and $\theta_0 = 1.5$ the estimation process again converges quickly to give estimates $\hat{\theta} = 1.67$, $\hat{\delta} = 2.06$, with a deviance of 18.70 with 10d.f., clearly a great improvement on the first model. A test of variation in the slope β_1 with level of x_2 now gives no significant reduction in the deviance, whereas with the first model the deviance was nearly halved by allowing the slope to vary. A final iteration multiplying the two derivative covariates by b_1 and b_2 and using the mean deviance as a heterogeneity factor gives the estimates, standard errors and correlations shown in Table 11.3.

Table 11.3 Results of the analysis of the insecticide-synergist assay

Parameter	Estimate	SE	Correlations			
α	-2.896	0.340				
β_1	1.345	0.143	-0.97			
θ	1.674	0.154	0.78	-0.77		
β_2	1.708	0.241	-0.31	0.26	-0.03	
δ	2.061	1.49	0.09	-0.07	0.06	0.61

Note that the two non-linear parameters are estimated almost

independently (correlation -0.07), and that δ is ill-determined. In particular δ must be positive, so that the standard error is suspect; a plot of the deviance against fixed values of δ near 2 shows a curve that is non-quadratic. The use of a square-root transformation for δ is a great improvement and indicates that confidence limits for δ should be calculated on the square-root scale. The fitting of a separate intercept for each level of synergist in place of the term in $x_2/(\delta+x_2)$ gives a trivial reduction in deviance, indicating that the hyperbolic form is satisfactory. There remain two large residuals, for units 1 and 2, of opposite sign, whose removal from the fit reduces the deviance from 18.70 to 5.69; however, there seems to be no good reason to reject them. The relevant values are:

<i>Unit</i>	<i>y</i>	<i>m</i>	$\hat{\mu}$	r_P
1	7	100	14.67	-2.17
2	59	200	43.51	2.66

This analysis indicates that the effect of doubling the dose of insecticide is to increase the odds of a kill by an estimated factor of $2^{\beta_1} = 2.54$. Since there is apparently no interaction between the insecticide dose and the synergist dose, this odds factor of 2.54 applies at any fixed dose of the synergist. The synergist exhibits decreasing returns of scale, a large dose increasing the odds of a kill by the factor $\exp(\hat{\beta}_2) = 5.52$. A moderate dose of 19.5 units increases the odds of a kill by an estimated $\exp(1.54) = 4.69$. These factors apply at any fixed dose of insecticide.

11.5.3 Mixtures of drugs

If two drugs provoke similar responses, a mixture of both may exhibit either additive or synergistic effects. If the effect is additive one drug can be replaced by a suitable proportion of the other to give the same response. With positive synergism the joint effect is greater than the sum of the effects of the two drugs administered separately: negative synergism is the term used to describe the opposite effect. In an experiment to test for such synergism, Darby and Ellis (1976) quote the data of Table 11.4, where the response y is the conversion of (3-3H)glucose to toluene-extractable lipids in isolated rat fat cells, and the two drugs are insulin in two forms, (1) standard and (2) A1-B29 suberoyl insulin. These are given in

seven different mixtures, each at two total doses; there are four replicate readings for the 14 treatments. Darby and Ellis proposed the model

$$E(Y_{ijk}) = \alpha + \beta \log(x_{1ij} + \theta x_{2ij}) \quad (11.2)$$

with constant-variance errors. Here i indexes the mixtures, j the total dose and k the replicates, while x_{1ij} and x_{2ij} are the amounts of the two drugs given for mixture i with dose j .

Table 11.4 Results of insulin assay

Mixture	Ratio of insulin to A1-B29 suberoyl insulin	Total dose (pmol l ⁻¹)	Responses for four replicates			
1	1:0	20.9	14.0	14.4	14.3	15.2
		41.9	24.6	22.4	22.4	26.7
2	1:1.85	52.9	11.7	15.0	12.9	8.3
		106	20.6	18.0	19.6	20.5
3	1:5.56	101	10.6	13.9	11.5	15.5
		202	23.4	19.6	20.0	17.8
4	1:16.7	181	13.8	12.6	12.3	14.0
		362	15.8	17.4	18.0	17.0
5	1:50.0	261	8.5	9.0	13.4	13.5
		522	20.6	17.5	17.9	16.8
6	1:150	309	12.7	9.5	12.1	8.9
		617	18.6	20.0	19.0	21.1
7	0:1	340	12.3	15.0	10.1	8.8
		681	20.9	17.1	17.2	17.4

Data from Darby and Ellis (1976).

Here θ is the non-linear parameter and we can fit the model by linearizing it, using the two covariates

$$u = \log(x_1 + \theta x_2), \quad v = \frac{\partial u}{\partial \theta} = \frac{x_2}{x_1 + \theta x_2};$$

we fit $\alpha + \beta u + \gamma v$ for some value θ_0 and θ is then updated by $\hat{\theta}_1 = \theta_0 + \gamma/\beta$. The estimate obtained after iteration is $\hat{\theta} = 0.0461 \pm 0.0036$, with a corresponding deviance (residual sum of squares) of 244.0 with 53d.f. Comparing this fit with the

replicate error of 154.8 with 42d.f., we find an F -value for residual treatment variation of $F(11, 42) = 2.20$, just beyond the 5% point. Darby and Ellis are concerned to compare this model with one in which θ is allowed to vary with the mixture, and in doing so to provide possible evidence of synergism or antagonism. Such a model, which requires a set of partial-derivative covariates, one for each mixture, reduces the deviance to 194.6 with 48d.f., giving a residual $F(6, 42) = 1.80$. While no longer significant at 5%, there is still a noticeable lack of fit, which investigation shows to be almost entirely concentrated in the first 'mixture', that for which x_2 is zero. Without this we find $\hat{\theta} = 0.0524$ and a deviance of 191.2 with 51d.f., giving a residual treatment deviance of 36.40 with 9d.f., so that the mean deviance is now close to the replicate error mean square of 3.686.

On this interpretation the interaction between the two drugs is expressed by saying that one unit of drug 2 is equivalent to 0.052 units of drug 1 in mixtures containing ratios by weight of 1:1.85 or more. In the absence of drug 2 the actual response to drug 1 is larger than predicted from the model (11.2), the predicted values for 'mixture' 1 (omitting it from the fit) being 12.9 and 19.8 as against 14.5 and 24.0 actually measured. Plots of residuals from the restricted model show no pattern when plotted against fitted values or mixture number, and there are no obvious outliers. The final analysis of variance is given in Table 11.5; it should be noted that this analysis differs somewhat from that of Darby and Ellis.

Table 11.5 *Analysis of variance for insulin assay*

	<i>s.s.</i>	<i>d.f.</i>	<i>m.s.</i>
<i>Treatments</i>	906.6	13	
<i>Model</i> (11.2)	817.4	2	408.7
<i>Separate</i> θ 's	49.4	5	9.88
<i>Residual</i>	39.8	6	6.63
<i>Alternative subdivision</i>			
<i>Model</i> (11.2)	817.4	2	
<i>Removal of</i> <i>mixture</i> 1	52.8	2	26.4
<i>Residual</i>	36.4	9	4.04
<i>Within treatments</i>	154.8	42	3.686

11.6 Bibliographic notes

The use of linearization methods for the optimization of non-linear functions has a long history going back to Gauss (1826), who gave a non-linear surveying problem to illustrate the technique of least squares. Its use in testing possible non-linear transformation of covariates was stressed by Box and Tidwell (1962) in the context of regression analysis.

Pregibon (1980) introduced goodness-of-link tests involving the estimation of parameters in the link function. Nelder and Pregibon (1987) described methods for the joint estimation of parameters in both link and variance functions.

11.7 Further results and exercises 11

11.1 Rounding errors: Let Z be a continuous random variable whose density $f(z)$ has derivatives satisfying the integrability condition

$$\int_{-\infty}^{\infty} f^{(\nu)}(z) dz < \infty,$$

where $\nu \geq 2$ is an even integer. Suppose that the recorded value is

$$Y = \epsilon \langle Z/\epsilon \rangle,$$

where $\langle x \rangle$ is the nearest integer to x . If $\epsilon = 10^{-d}$ then Y is Z rounded to d decimal places. We can thus write

$$Z = Y + \epsilon R,$$

where $-\frac{1}{2} < R \leq \frac{1}{2}$ is the normalized rounding error.

Using the Euler-Maclaurin summation formula (Bhattacharya and Rao, 1976, p.256; Jeffreys and Jeffreys, 1956, p.280) or otherwise, show that the joint cumulant generating function of (Z, R) satisfies

$$K_{Z,R}(\xi_1, \xi_2) = K_Z(\xi_1) + K_R(\xi_2) + O(\epsilon^\nu)$$

$$K_R(\xi) = \log\left(\frac{\sinh(\frac{1}{2}\xi)}{\frac{1}{2}\xi}\right) + O(\epsilon^\nu)$$

for small ϵ . Hence deduce that to a close approximation for small ϵ , R is uniformly distributed on $(-\frac{1}{2}, \frac{1}{2}]$, and that the cumulants of the rounded random variable Y are given by

$$\begin{aligned}\kappa_r(Y) &\simeq \kappa_r(Z) \quad \text{for } r \text{ odd,} \\ \kappa_r(Y) &\simeq \kappa_r(Z) + \epsilon^r \kappa_r(R) \quad \text{for } r \text{ even.}\end{aligned}$$

[Kolassa and McCullagh, 1987]. The curious aspect of this result is that even though R is a deterministic function of Z , the joint asymptotic distribution is such that R and Z are statistically independent to a high order of approximation, provided ν is moderately large. By contrast, R and Y are also asymptotically independent, but only to a first order of approximation. In fact $\text{cov}(R, Y) \simeq \epsilon \text{ var}(R) = \epsilon/12$.

11.2 Deduce that if Z has the gamma distribution with index $\nu \geq 2$ and if Y is equal to Z rounded to d decimal places, then

$$\text{var}(Y) \simeq \mu_Z^2 / \nu + \epsilon^2 / 12,$$

where $\epsilon = 10^{-d}$. What would be the effect on the variance function if Z were rounded to d significant decimal digits rather than to d decimal places?

11.3 Show that if U is uniformly distributed on $(0, 1)$ the first four cumulants are $\frac{1}{2}$, $\frac{1}{12}$, 0 and $-\frac{1}{120}$.

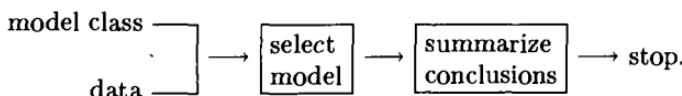
11.4 Use the linearization technique described in section 11.4 to fit the non-linear model (8.4) to the data in Table 8.8. Use gamma errors, log link and take weights equal to the batch sizes. Find the maximum-likelihood estimates of the four parameters, together with their asymptotic standard errors. Plot the residual deviance for fixed δ against δ in the range $50\text{--}100^\circ\text{C}$ to check on the adequacy of the Normal approximation for $\hat{\delta}$.

CHAPTER 12

Model checking

12.1 Introduction

The process of statistical analysis as presented in many textbook examples appears to take the form



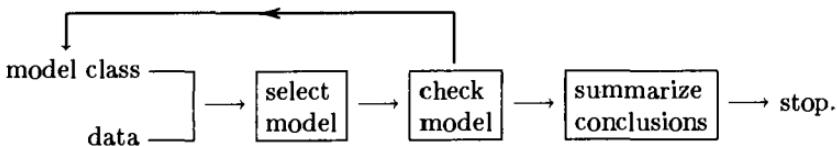
Sometimes the model class has only one member, as, for example, in the standard analysis of a randomized blocks experiment with an unstructured treatment factor. At other times model selection may involve, say, selecting a subset of terms from a full factorial model, using differences in deviance to decide which terms should be included. Whether or not the prior model is unique the process effectively assumes that at least one model from the class is the right one, so that, after fitting, all that remains is to summarize the analysis in terms of parameter estimates, standard errors covariance matrix, and so on.

A good statistician selects his model class carefully, paying attention to the type and structure of the data. Thus in modelling counts, fitted values from a model should be confined to non-negative values, because counts are. Similarly if it is known a priori that a response to a stimulus variable x tails off beyond a certain level that is well within the range of x in the data, then a linear term only in x will not be adequate for the model.

However, even after a careful selection of model class, the data themselves may indicate that the particular model selected is unsuitable. Such indications can take two forms. It may be that the data as a whole show some systematic departure from the fitted values, or it may be that a few data values are discrepant from the

rest. The detection of both systematic and isolated discrepancies is part of the technique of model checking. To exemplify a systematic discrepancy consider a plot of the residuals r against x , one of the covariates in the linear predictor. If the fit is good the pattern expected is null, i.e. no relation between r and x . However, if the plot shows residuals of one sign concentrated at the ends of the x scale and residuals of the other sign at the centre then this may be evidence that x^2 has been omitted as a covariate in the linear predictor, or for wrong choice of link function. By contrast, an isolated discrepancy would occur if a few points have residuals very far from the rest. This latter pattern indicates something unusual about those particular points; they may be at the extremes of the x range in a region where the model does not apply, or, more mundanely, the values may simply be wrong, the result of misrecording or errors in transcription.

The effect of model checking is to introduce a loop into the analysis process as follows:



The introduction of this loop changes profoundly the process of analysis and the reliability of the final models found. In this chapter we extend, where possible, techniques originally devised for regression models to the whole class of generalized linear models, and develop new ones for aspects of the general class that have no analogues in the restricted one.

12.2 Techniques in model checking

Model-checking techniques may be either informal or formal. Informal techniques rely upon the human mind and eye to detect pattern. Such methods take a successful model to be one that, among other things, leaves a patternless set of residuals. The argument is that if we can detect pattern in the residuals we can find a better model; the practical problem is that any finite set of residuals can be made to yield some kind of pattern if we look hard enough,

so that we have to guard against over-interpretation. Nonetheless informal methods are an important component in model checking.

Formal methods rely on embedding the current model in a wider class that includes extra parameters. If θ is such a parameter, and θ_0 its value in the current model, then a formal method would find $\hat{\theta}$, the estimate of θ giving the best fit in the wider class, and compare the fit at $\hat{\theta}$ with that at θ_0 . The current model passes the check if the inclusion of θ as an extra parameter does not markedly improve the fit. Extra parameters might arise from including an additional covariate, from embedding a covariate x in a family $h(x; \theta)$ indexed by θ , from embedding a link function $g(\eta)$ in a similar family $g(\eta; \theta)$, or from including a constructed variate, say $\hat{\eta}^2$, obtained from the original fit. Formal methods thus look for deviations from the fit in certain definite directions thought likely to be important a priori.

Formal methods for dealing with isolated discrepancies include adding dummy variates taking the value 1 for the discrepant unit and zero elsewhere. The change in deviance then measures the effect of that unit on the fit. The addition of such a dummy variate has an effect on the fit equivalent to deleting that unit from the data matrix. In assessing the significance of that change, due allowance must be made for the effect of having picked the most discrepant unit.

12.3 Score tests for extra parameters

Many procedures used in model checking can be shown to be special cases of the class of score tests (Rao, 1973, Chapter 6). Consider two models, one (M_0) with p parameters and a second (extended) model (M_1) with $p + k$ parameters. The deviance test is based on the reduction in deviance for M_1 relative to M_0 . The score test, on the other hand, is based on the log likelihood derivatives with respect to the extra parameters: both the derivatives and the Fisher information are computed under M_0 . For generalized linear models the score statistic can be computed by first fitting M_0 , followed by one step of the iteration for M_1 . The reduction in X^2 in this first step is the score statistic, sometimes also called the quadratic score statistic. Pregibon (1982) gives the details.

The computing advantage of the score test over the deviance

test is that it requires only a single iteration for the extended model, compared with iteration to convergence for the deviance (or likelihood-ratio) test. Note that the two tests give identical results for linear models with constant variance. For $k = 1$ and no nuisance parameters the statistics are interpreted geometrically in Fig. 12.1a, in which the log likelihood derivative is plotted against the extra parameter λ .

Figure 12.1b shows a typical comparison of the two statistics over a range of λ -values. The solid curve gives the minimum deviance for the model M_1 for varying values of the extra parameter λ , M_0 itself being defined by $\lambda = \lambda_0$. The deviance statistic is thus the difference in ordinates $D_0 - D_1$. The score statistic at λ_0 has the form $S(\lambda_0) = U(\lambda_0)^T i^{-1}(\lambda_0 | \cdot) U(\lambda_0)$, where $U(\lambda_0)$ is the log-likelihood derivative with respect to λ at λ_0 , and $i(\lambda_0 | \cdot)$ is the Fisher information for λ as defined in Appendix A, treating the original parameters as nuisance parameters. Its value for varying λ is shown by the dashed line in Fig 12.1b. At $\lambda = \hat{\lambda}$ both statistics are minimized and $S(\hat{\lambda}) = 0$.

Neither the score statistic nor the deviance statistic is affected by re-parameterization of λ . The difference between the two statistics, which is typically small, arises chiefly from two sources, (i) the difference between the observed and expected Fisher information, and (ii) third-order properties of the log-likelihood function. Wald's statistic, which is a quadratic form based on $\hat{\lambda} - \lambda$, is not similarly invariant.

12.4 Smoothing as an aid to informal checks

Some informal checks involve assessing a scatter plot for approximate linearity (or other relation) in the underlying trend. This exercise may be difficult, particularly if the density of points on the x scale varies widely over the range of x values observed. The problem is that where the density of x -values is high, the expected range of y values is larger than in a neighbourhood where the x -values are less dense. The eye finds it hard to make the necessary adjustments, and may be greatly helped if the scatter-plot is augmented by an empirical curve produced by a suitable smoothing algorithm (see, e.g. Cleveland, 1979). Such smoothed curves must be treated with some caution, however, since the algorithm is quite

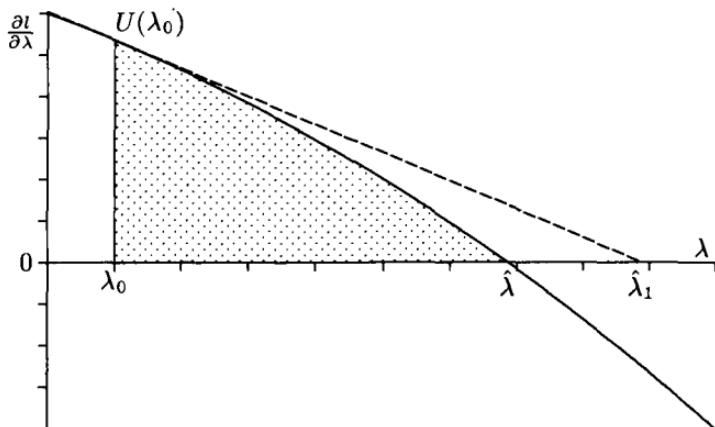


Fig. 12.1a. *The geometry of the score test for one parameter. The solid line is the graph of the log likelihood derivative. The shaded area is one half of the likelihood ratio statistic: the score statistic is twice the area of the triangle λ_0 , $U(\lambda_0)$, $\hat{\lambda}_1$.*

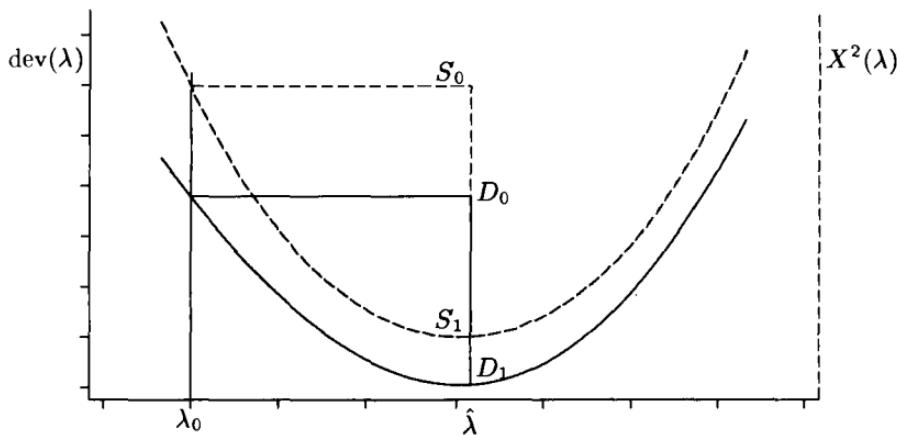


Fig. 12.1b. *The geometry of the score test: dashed line is the curve of X^2 using the adjusted dependent variate and weights from the fit at λ ; solid line is the deviance for varying λ .*

capable of producing convincing-looking curves from entirely random configurations. Nonetheless, smoothing is undoubtedly useful as an aid to informal checks.

12.5 The raw materials of model checking

We consider first linear regression models with supposedly constant variance, where model checking uses mainly the following statistics derived from a fit:

The fitted values $\hat{\mu}$,

The residual variance s^2 ,

The diagonal elements h of the projection ('hat') matrix,

$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$, which maps \mathbf{y} into $\hat{\mu}$.

An important idea is that of case deletion, whereby the fit using all the cases (points) is compared with that obtained when one point is deleted. We shall denote statistics obtained by deleting point i by a suffix in brackets, so that, for example, $s_{(i)}^2$ is the residual variance for the model fitted omitting point i .

A central role is played by the residuals from a fit, and several forms have been proposed in addition to the basic $r = y - \hat{\mu}$. We shall call a residual **standardized** if it has been divided by a factor that makes its variance constant. Standardization produces the form

$$\frac{y_i - \hat{\mu}_i}{\sqrt{1 - h_i}},$$

where $\{h_i\}$ are the components of \mathbf{h} . If in addition the residual is scaled by dividing by s , we call it a Studentized standardized residual, and write

$$r'_i = \frac{y_i - \hat{\mu}_i}{s\sqrt{1 - h_i}}. \quad (12.1)$$

Note that r'^2_i is just the reduction in the residual sum of squares caused by omitting the point i , scaled by the residual mean square for all the points.

Finally there is the important **deletion residual** defined by

$$r_i^* = \frac{y_i - \hat{\mu}_{(i)}}{s_{(i)}\sqrt{1 + h_{(i)}}} = \frac{y_i - \hat{\mu}_i}{s_{(i)}\sqrt{1 - h_i}}. \quad (12.2)$$

in which \mathbf{x}_i^T is the i th row vector, $\mathbf{X}_{(i)}$ is the model matrix with the i th row deleted, and

$$h_{(i)} = \mathbf{x}_i^T (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} \mathbf{x}_i = h_i / (1 - h_i).$$

The two residuals r' and r^* are related by

$$r_i^* = r_i' s / s_{(i)},$$

(see Atkinson, 1985), so that r_i^{*2} is again a reduction in the residual sum of squares, but this time scaled by $s_{(i)}^2$ instead of s^2 . The deletion residual r_i^* measures the deviation of y_i from the value predicted by the model fitted to the remaining points, standardized and Studentized like r' . (The difference in sign in the terms in the denominators of r' and r^* arises from the fact that y_i and $\hat{\mu}_{(i)}$ are independent, whereas y_i and $\hat{\mu}_i$ are positively correlated.)

For generalized linear models some extensions and modifications are needed to the above definitions. First, where checks on linearity are involved the vectors \mathbf{y} and $\hat{\mu}$ are ordinarily replaced by \mathbf{z} , the adjusted dependent variate, and $\hat{\eta}$, the linear predictor. The residual variance is replaced by an estimate of the dispersion parameter ϕ , and the \mathbf{H} matrix becomes

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{\frac{1}{2}}, \quad (12.3)$$

equivalent to replacing \mathbf{X} by $\mathbf{W}^{\frac{1}{2}} \mathbf{X}$ in the regression version. It can be shown that, to a close approximation,

$$\mathbf{V}^{-1/2}(\hat{\mu} - \mu) \simeq \mathbf{H} \mathbf{V}^{-1/2}(\mathbf{Y} - \mu),$$

where $\mathbf{V} = \text{diag}(V(\mu_i))$. Thus \mathbf{H} measures the influence in Studentized units of changes in \mathbf{Y} on $\hat{\mu}$. The corresponding matrix in unstandardized units is $\mathbf{V}^{1/2} \mathbf{H} \mathbf{V}^{-1/2}$, which is an asymmetric projection matrix.

In section 2.4 we defined three forms of residual for generalized linear models, and two of these, the Pearson and deviance residuals, have been widely used in model checking. For the Pearson residual the analogous form of (12.1) is given by

$$r'_P = \frac{y - \hat{\mu}}{\sqrt{\hat{\phi} V(\hat{\mu})(1 - h)}}. \quad (12.4)$$

The calculations of Cox and Snell (1968) support a similar standardization for the deviance residual giving

$$r'_D = \frac{r_D}{\sqrt{\hat{\phi}(1 - h)}}. \quad (12.5)$$

Exact calculations for deletion residuals may become expensive when iteration is required for every fit. It has become common practice to approximate the quantities involved in the deletion residuals by using one-step approximations. This involves doing one cycle of the fit without point i , starting from the fitted values, weights, etc. from the complete fit. Use of one-step approximations allows certain shortcuts to be made, analogous to those used for regression. We write ${}_1r'_P$ and ${}_1r'_D$ for the one-step approximations to (12.4) and (12.5); thus ${}_1r'_P$ is the one-step approximation to r_P^2 , measuring the change in the Pearson χ^2 caused by omitting a point. The analogous approximation for the change in deviance has been shown by Williams (1987) to be given by

$$r_D^2 = h {}_1r'_P^2 + (1 - h) {}_1r'_D^2. \quad (12.6)$$

An equivalent formula is given by Pregibon (1981, p.720). In general the deviance residual, either unstandardized or standardized, is preferred to the Pearson residual for model checking procedures because its distributional properties are closer to the residuals arising in linear regression models (Pierce and Schafer, 1986).

12.6 Checks for systematic departure from model

We consider first checks for systematic departure from the model, beginning with three residual plots.

12.6.1 *Informal checks using residuals*

If the data are extensive, no analysis can be considered complete without inspecting the residuals plotted against some function of the fitted values. Standardized deviance residuals are recommended, plotted either against $\hat{\eta}$ or against the fitted values transformed to the constant-information scale of the error distribution. Thus we use

- $\hat{\mu}$ for Normal errors,
- $2\sqrt{\hat{\mu}}$ for Poisson errors,
- $2 \sin^{-1} \sqrt{\hat{\mu}}$ for binomial errors,
- $2 \log \hat{\mu}$ for gamma errors,
- $-2\hat{\mu}^{-\frac{1}{2}}$ for inverse Gaussian errors.

The argument for the constant-information scale is as follows: for Normal errors if we plot $y - \hat{\mu}$ against $\hat{\mu}$ then the contours of fixed y are parallel straight lines with a slope of -1 . With other distributions the contours are curves but the constant-information scale gives a slope of -1 at $r = 0$ to match the Normal case and makes the curvature generally slight. For data with binomial errors, note that $\hat{\mu}$ is interpreted as $\hat{\pi}$ rather than $m\hat{\pi}$.

The null pattern of this plot is a distribution of residuals for varying $\hat{\mu}$ with mean zero and constant range. Typical systematic deviations are (i) the appearance of curvature in the mean and (ii) a systematic change of range with fitted value. Smoothing may be useful in judging whether curvature is present, but cannot help with assessing variable range (see section 12.6.2). Note that this plot is generally uninformative for binary data because all the points lie on one of two curves according as $y = 0$ or 1 . Furthermore for $\hat{\mu}$ near zero almost all the points have $y = 0$, and conversely for $\hat{\mu}$ near one.

Curvature may arise from several causes, including the wrong choice of link function, wrong choice of scale of one or more covariates, or omission of a quadratic term in a covariate. Ways of distinguishing between these will be discussed further in sections 12.6.3–4.

A second informal check plots the residuals against an explanatory variable in the linear predictor. The null pattern is the same as that for residuals *vs* fitted values. Again the appearance of systematic trend may indicate the wrong choice of link function or scale of the explanatory variable, or point to a missing quadratic term. Such a trend may also be an artefact caused by a faulty scale in another explanatory variable closely correlated with the one under investigation. Smoothing may help in overcoming the effect of variable density of points.

A third residual plot, known as an added-variable plot, gives a check on whether an omitted covariate, \mathbf{u} , say, should be included in the linear predictor. It is not adequate to plot the residuals against \mathbf{u} itself for this purpose. First we must obtain the unstandardized residuals for \mathbf{u} as response, using the same linear predictor and quadratic weights as for \mathbf{y} . The unstandardized residuals for \mathbf{y} are then plotted against the residuals for \mathbf{u} . If \mathbf{u} is correctly omitted no trend should be apparent.

12.6.2 Checking the variance function

A plot of the absolute residuals against fitted values gives an informal check on the adequacy of the assumed variance function. The constant-information scale for the fitted values is usually helpful in spreading out the points on the horizontal scale. The null pattern shows no trend, but an ill-chosen variance function will result in a trend in the mean. Again smoothing may help to see the trend more clearly. A positive trend indicates that the current variance function is increasing too slowly with the mean, so that, for example, an original choice of $V(\mu) \propto \mu$ may need to be replaced by $V(\mu) \propto \mu^2$. A negative trend would indicate the reverse.

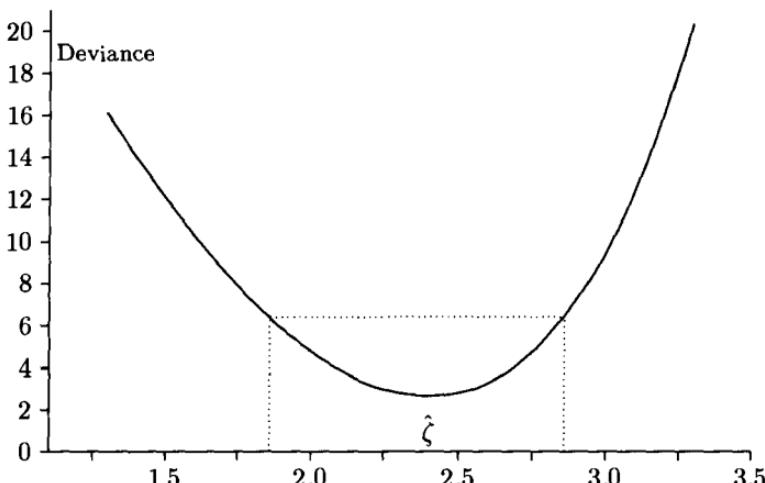


Fig. 12.6. *The profile extended quasi-likelihood curve plotted against ζ for the power-family of variance functions applied to the car-insurance data.*

To check the variance function formally we embed the current one in a suitable family, usually $V(\zeta) = \mu^\zeta$, indexed by a parameter ζ , and observe how the fit improves as ζ varies. For this comparison we need the extended quasi-likelihood discussed in section 9.6, which allows the comparison of different variance functions. We compute the deviance for a range of ζ , producing a profile quasi-likelihood curve; approximate likelihood limits are given by the χ_1^2 values for a chosen significance level and the prior value ζ_0 is evaluated in respect of the interval so produced. Fig. 12.2 shows

the curve obtained for the car insurance data (the example in section 8.4.1); the 95% limits for ζ are about (1.87, 2.85) showing that our original choice of $\zeta_0 = 2$ was satisfactory.

12.6.3 Checking the link function

An informal check involves examining the plot of the adjusted dependent variable z against $\hat{\eta}$, the estimated linear predictor. The null pattern is a straight line. For link functions of the power family an upwards curvature in the plot points to a link with higher power than that used, and downwards curvature to a lower power. Smoothing may be helpful in interpreting the plot. For binary data this plot is uninformative and formal methods must be used.

There are two formal checks in common use. The simpler (Hinkley, 1985) involves adding $\hat{\eta}^2$ as an extra covariate and assessing the fall in deviance. (A score test may be used as an alternative.) The other formal check involves embedding the link function in a family indexed by a parameter λ and testing the prior value λ_0 in the usual way. Uncertainty about the link function is probably commonest with continuous data having gamma errors, and with proportions having binomial errors. For the former the power family $\eta = \mu^\lambda$ is the most useful. section 11.3 describes the techniques for estimating λ and assessing the adequacy of λ_0 .

For binomial data various families have been constructed that include the logistic link (the canonical link) and the complementary log-log link as special cases. Some of these are discussed in section 11.3.2

Checks on the link functions are inevitably affected by failure to establish the correct scales for the explanatory variables in the linear predictor. In particular, if the formal test constructed by adding $\hat{\eta}^2$ to the linear predictor indicates deviation from the model this may point either to a wrong link function, or to wrong scales for explanatory variables or both. The methods described in the next section may help in distinguishing the various alternatives.

12.6.4 Checking the scales of covariates

The partial residual plot is an important tool for checking whether a term βx in the linear predictor might be better expressed as $\beta h(x; \theta)$ for some monotone function $h(\cdot; \theta)$. In its generalized form

the partial residual is defined by

$$u = z - \hat{\eta} + \hat{\gamma}x$$

where z is the adjusted dependent variable, $\hat{\eta}$ the fitted linear predictor and $\hat{\gamma}$ the parameter estimate for the explanatory variable x .

The plot of u against x provides an informal check. If the scale of x is satisfactory the plot should be approximately linear. If not its form may suggest a suitable alternative. The scatter about any trend may not be uniform, in which case smoothing may help interpretation. The partial residual plot, if smoothed, can be remarkably informative even for binary data. However, distortions will occur if the scales of other explanatory variables are wrong, so that iteration may be necessary in looking at the partial residual plots for several xs . This problem may be less severe with the following formal check which allows simultaneous transformation of several xs to be tested.

As usual the formal check involves embedding the current scale x in a family $h(x; \theta)$ indexed by θ ; we then calculate the deviance for a suitable grid of values of θ to find the position of the minimum, which gives $\hat{\theta}$. The fit at $\hat{\theta}$ can then be compared with that at our initial choice of θ_0 , which is usually 1. The method is equivalent to the use of a maximum profile-likelihood estimator. Clearly this procedure can be used for several xs simultaneously. This is particularly useful when several xs have the same physical dimensions, so that a simultaneous transformation is likely to be required. By far the commonest family of transformations is the power family given by

$$h(x; \theta) = \begin{cases} \frac{x^\theta - 1}{\theta} & \text{for } \theta \neq 0, \\ \log(\theta) & \text{for } \theta = 0. \end{cases}$$

An informal check for a single covariate takes the form of a constructed-variable plot for $v = \partial h / \partial \theta_0$; we first fit a model with v as dependent variable, with the linear predictor and quadratic weight as for y , and form the residuals. We then plot the residuals of y against the residuals of v ; a linear trend indicates a value of $\theta \neq \theta_0$, while a null plot would indicate no evidence against $\theta = \theta_0$.

12.6.5 Checks for compound systematic discrepancies

So far we have mainly considered checks for a single cause of discrepancy, e.g. a covariate on the wrong scale, a covariate omitted, or a faulty link function. Each of these discrepancies can be tested formally by including an extra variable in the linear predictor and calculating either the deviance reduction or the score statistic for its inclusion; the process is thus analogous to forward selection (Section 3.9). The danger, as usual, is that correlations between the extra variables can lead to each mimicking the effect of the others. The use of backward-selection, where possible, gives a means of avoiding the danger; we now fit all the extra variables, giving the joint effect of all the causes of discrepancy to be tested, and then find the effect of omitting each one in turn from the joint fit. Again either the deviance reduction or the score statistic may be used. Davison and Tsai (1988) give examples of this technique.

12.7 Checks for isolated departures from the model

In this section we look at model-checking procedures associated with particular points in the data, especially those that appear in some way at variance with the pattern set by the remainder. We deal first with the case of a single possibly discrepant point.

For simplicity we consider first data with a response variable y and one explanatory variable x . We assume that an identity link is relevant. The scatter plot of y against x may show an isolated extreme point, which we define loosely as one well apart from the main cluster. There are three types of configuration that are worth distinguishing, and these are shown in Fig. 12.3 with the extreme point indicated by a circle. In Fig. 12.3(a) the x -value of the extreme point is close to the mean. Exclusion of this point has only a small effect on the estimate of the slope, but it substantially reduces the intercept. Its exclusion also produces a big improvement in the goodness of fit.

In Fig. 12.3(b) the extreme point is consistent with the rest, in that a straight line fitted through the rest passes near the extreme point. Inclusion of the extreme point will increase the accuracy of $\hat{\beta}$ without affecting its estimate greatly.

In Fig. 12.3(c) the straight line fitted through the non-extreme points does not pass close to the extreme point, so that if it is

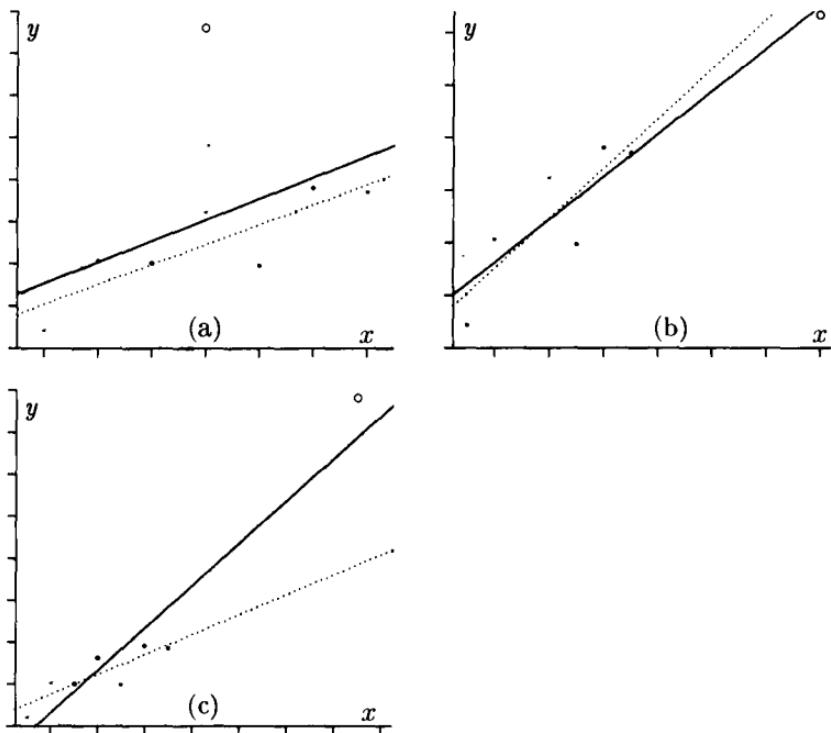


Fig. 12.3. Scatter plots showing the effect of an extreme point in relation to the configuration of the remaining points. The solid lines are the least squares fits with all points included: the dotted lines are the fits with the extreme points excluded.

now included in the fit the value of $\hat{\beta}$ will change sharply, and the deviance also.

Three ideas are useful in thinking about the configurations in Fig. 12.3. The first is that of leverage, which distinguishes (a) from (b) and (c). The inclusion of the extreme point in (b) and (c) greatly increases the information about $\hat{\beta}$, i.e. the point has high leverage whereas in (a) it has low leverage. The second idea is that of consistency which distinguishes (b) from (a) and (c). In (b) the (x, y) values of the extreme point are consistent with the trend suggested by the remainder, while in (a) and (c) they are not. The third idea, termed influence, distinguishes (c) from (a) and (b). The extreme point has high influence if the estimate of the slope is greatly changed by its omission, as in (c), and low influence if it is little changed, as in (a) or (b).

We now develop statistics for measuring leverage, consistency and influence quite generally.

12.7.1 Measure of leverage

For linear regression models the well-known measure of leverage is given by the diagonal elements of the ‘hat’ matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T, \quad (12.7)$$

the i th element of which is

$$h_i = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i.$$

Note that if the columns of \mathbf{X} are orthogonalized with the first column constant, \mathbf{H} is unaffected. The diagonal elements are then expressible in the form

$$h_i = 1/n + \frac{x_{i2}^2}{\sum x_{j2}^2} + \dots + \frac{x_{ip}^2}{\sum x_{jp}^2}.$$

Thus $h_i - 1/n$ is an invariant measure of squared distance between \mathbf{x}_i and the centroid of all n points in the x -space.

The general form of \mathbf{H} , given in (12.3), has \mathbf{X} replaced by $\mathbf{W}^{1/2} \mathbf{X}$, which effectively allows for the change in variance with the mean. It can also be thought of in an informal sense as the ratio of the covariance matrix of $\hat{\boldsymbol{\mu}}$ to that of \mathbf{Y} . Now

$$\sum h_i = \text{trace } \mathbf{H} = p,$$

and there is some advantage in working with a standardized form

$$h'_i = nh_i/p$$

so that $\sum h'_i = n$. Hoaglin and Welsch (1978) suggest using $h' > 2p/n$, i.e. $h' > 2$, to indicate points of high leverage. An isolated point of high leverage may have a value of h approaching unity. An index plot of h' with the limit $h' = 2$ marked is a useful informal tool for looking at leverage.

Note that for GLMs a point at the extreme of the x -range will not necessarily have high leverage if its weight is very small.

12.7.2 Measure of consistency

An inconsistent point is one with a large residual from the curve fitted to the remaining points. Thus the deletion residual introduced in section 12.5 is a natural measure of inconsistency, i.e. small deletion residuals denote consistent points. For generalized linear models the one-step approximation given by (12.6) is appropriate.

12.7.3 Measure of influence

Influence can be measured as a suitably weighted combination of the changes $\hat{\beta}_{(i)} - \hat{\beta}$, where $\hat{\beta}_{(i)}$ denotes the estimates without the extreme point, and $\hat{\beta}$ those with it. Cook (1977) first proposed a statistic, which for regression models takes the form

$$D_i = (\hat{\beta}_{(i)} - \hat{\beta})(\mathbf{X}^T \mathbf{X})(\hat{\beta}_{(i)} - \hat{\beta})/ps^2, \quad (12.8)$$

as a measure of influence of the i th point, when s^2 is an estimate of the dispersion parameter. As shown in Fig. 12.2(a) not all parameters are equally affected by an extreme point, and D_i is intended to provide a suitably weighted combined measure.

From the relation

$$\hat{\beta}_{(i)} - \hat{\beta} = -(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i r_i / (1 - h_i)$$

(Atkinson, 1985, p.21), it follows that

$$D_i = \frac{r_i'^2 h_i}{p(1 - h_i)} \quad (12.9)$$

showing that D_i is a function of the quantities involved in the measurement of leverage and consistency.

Atkinson (1981) suggests modifications to D which have advantages in standardizing it for different configurations of \mathbf{X} and making extreme points stand out more sharply. First he replaces r'^2 in (12.9) by r^{*2} , which is equivalent to the use of $s_{(i)}^2$ in place of s^2 . Secondly he scales by a factor $(n - p)/p$; this has the effect of making the modified D_i equal to r^{*2} when all points have equal leverage. Finally he takes the square root, producing the modified Cook statistic

$$C_i = \left\{ \frac{n-p}{p} \cdot \frac{h_i}{1-h_i} \right\}^{\frac{1}{2}} |r_i^*|. \quad (12.10)$$

To adapt these statistics for use with generalized linear models is straightforward. D_i is now defined by

$$D_i = (\hat{\beta}_{(i)} - \hat{\beta})(\mathbf{X}^T \mathbf{W} \mathbf{X})(\hat{\beta}_{(i)} - \hat{\beta})/p\hat{\phi},$$

where the $\hat{\beta}_{(i)}$ will usually be the one-step approximations discussed in section 12.5. The modified Cook statistic C_i can be adapted by simply replacing r^* by r_D^* , the one-step approximation to the deletion deviance residual.

12.7.4 Informal assessment of extreme values

The three statistics h , r^* and C , introduced above for the measurement of leverage, consistency and influence respectively, each yield a vector of n values. The interesting values for model-checking purposes are the large ones (of either sign in the case of r^*). To interpret these we need plots that allow for the fact that we have chosen the most extreme values to examine. We thus need some measure of how large the extreme values would be in a sample of a given size even if no unusual points were present.

The simple index plot of the statistic against case number does not have this property, but it has value, particularly when a few points are far from the rest. Normal plots, which make allowance for selection effects, come in two forms, the half-Normal plot and the full Normal plot. The former is appropriate for non-negative quantities like h and C ; for a statistic like r^* , there are two options, either a half-Normal plot of $|r^*|$ or a full Normal plot of r^* itself. For either plot the ordered values of the statistic are plotted against the expected order statistics of a Normal sample. The latter may be generated with sufficient accuracy for practical purposes by

$$\Phi^{-1}\left(\frac{i - \frac{3}{8}}{n + \frac{1}{4}}\right) \quad i = 1, \dots, n$$

for the full Normal plot, and by

$$\Phi^{-1}\left(\frac{n + i + \frac{1}{2}}{2n + \frac{9}{8}}\right) \quad i = 1, \dots, n$$

for the half Normal plot. Note that the use of either of these scales is to some extent conventional, for, while the plot for r^* may, in

the absence of unusual points, be approximately linear, there is usually no reason to expect the h or C plots to be so. Extreme points will appear at the extremes of the plot, possibly with values that deviate from the trend indicated by the remainder.

To aid the interpretation of Normal plots Atkinson (1981) developed the useful idea of an envelope constructed by simulation. For generalized linear models with a fully specified error distribution this is constructed as follows: for each simulation form pseudo-data \mathbf{y}^* by generating random variables from the appropriate error distribution with mean $\hat{\mu}$ and dispersion $\hat{\phi}$. Refit the model and calculate the statistic. Order its values. Do k simulations and for each ordered position select the extreme values from the k simulations. Plot these with the original points to give the envelope. More stable envelopes can be obtained, if simulation is cheap, by using larger values of k and less extreme values of the ordered samples.

Atkinson (1985) gives detailed examples on the interpretation of plots with envelopes for regression models. Simulation is particularly simple here, because the variance is independent of the mean, so that $\hat{\mu}$ can be disregarded, and y^* requires just $N(0, 1)$ variables. Note that simulation for models with non-Normal errors can be speeded up using a one or two-step approximation to the full iteration beginning with $\hat{\mu}$ as the initial estimate of the fitted values.

Residuals from data in the form of counts or proportions will show distortions if there are many zeros (counts) or zeros and ones (proportions). These produce a concentration of small residuals near zero, which may appear as a plateau in the Normal plot.

12.7.5 *Extreme points and checks for systematic discrepancies*

Up to now we have divided model-checking techniques into those for systematic and those for isolated discrepancies. However it is possible to formulate questions that involve both kinds of discrepancy. For example we might ask ‘does the evidence for the inclusion of a covariate depend largely on the influence of a few isolated points?’ One way of answering such a question has been given by Williams (1987); see also Davison and Tsai (1988).

Consider a test by backward selection (Section 12.6.5) for a systematic discrepancy as measured by the extra variable \mathbf{u} . Suppose that the full linear predictor gives squared residuals r_{G1}^2 , and that

without \mathbf{u} the residuals are r_{G0}^2 . Then the differences $r_{G0}^2 - r_{Gi}^2$ can be used in an index plot to show the influence of each point on this test for systematic discrepancy. Such a plot might reveal, for example, that most of the evidence for the effect in question comes from one or two points that have previously been identified as possible outliers. If required, the analysis may be repeated with suspected outliers omitted.

12.8 Examples

12.8.1 Damaged carrots in an insecticide experiment

The data shown in Table 12.1, taken from Phelps (1982), are discussed by Williams (1987). They give the proportion of carrots showing insect damage in a trial with three blocks and eight dose levels of insecticide. With a logit link function and simple additive linear predictor $block + x$, where x is the log dose, we find a deviance of 40.0 with 20 d.f., rather too large for binomial variation.

Table 12.1 Proportion of carrots damaged in an insecticide experiment

level j	log dose x_j	Dose			Block		
		1	2	3	1	2	3
1	1.52	10/35	17/38	10/34			
2	1.64	16/42	10/40	10/38			
3	1.76	8/50	8/33	5/36			
4	1.88	6/42	8/39	3/35			
5	2.00	9/35	5/47	2/49			
6	2.12	9/42	17/42	1/40			
7	2.24	1/32	6/35	3/22			
8	2.36	2/28	4/35	2/31			

Source: Phelps (1982).

There may be general over-dispersion or perhaps isolated extreme points. Fig. 12.4 shows an index plot (with the data ordered by columns) of the one-step deletion residual r^* . This plot quickly decides the issue; point 14 (dose level 6 and block 2) is far away from the rest. The fit omitting this point gives a deviance of 25.3 with 19 d.f. Though somewhat above the baseline of 19, it is clearly

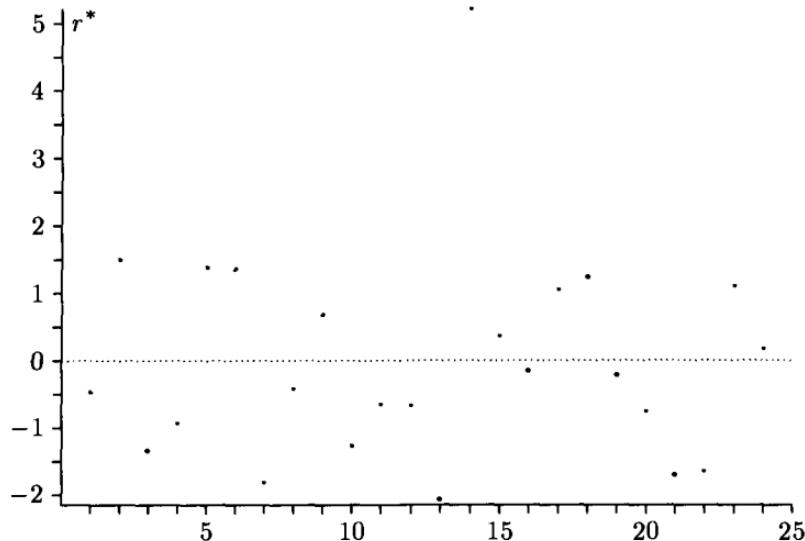


Fig. 12.4 Index plot of one-step deletion residuals r^* for carrot data, showing a single outlier.

a great improvement on the original fit. The fitted value for point 14 is much closer to 7 than to the 17 recorded.

Inclusion of the constructed variable $\hat{\eta}^2$ after omitting point 14 gives an insignificant reduction (0.2) in the deviance, so that our choice of link function is not contradicted. (Phelps used the complementary log-log link, but the difference in fit between it and the logistic is small.) This example thus illustrates the effect of an isolated extreme point having an anomalous y -value.

12.8.2 Minitab tree data

This famous set of data on the volume, diameter (at 4' 6" above ground level), and height of black cherry trees was given by Ryan *et al.* (1976). Interest attaches to deriving a formula to predict tree volume v from measurements of diameter d and height h . If all the trees were the same shape we would expect to find

$$v = c \times d^2 \times h \quad (12.11)$$

for some constant c . Thus we might expect that a successful linear model would involve $\log v$ as response variable with $\log d$ and $\log h$ as explanatory variables.

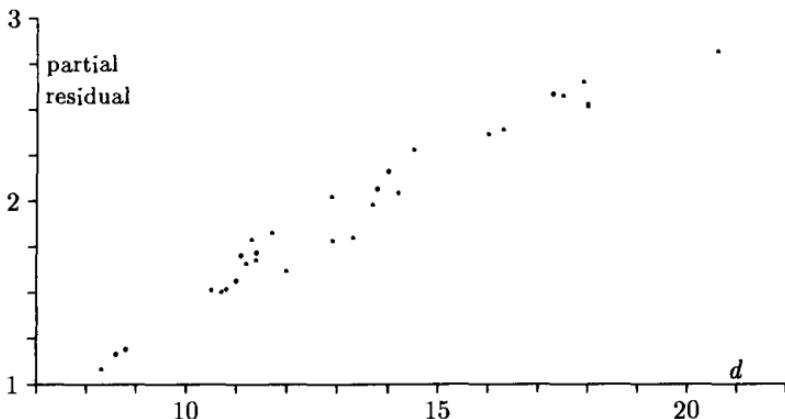


Fig. 12.5a. Partial residual plot for d in the joint fit of d and h to $\log(v)$.

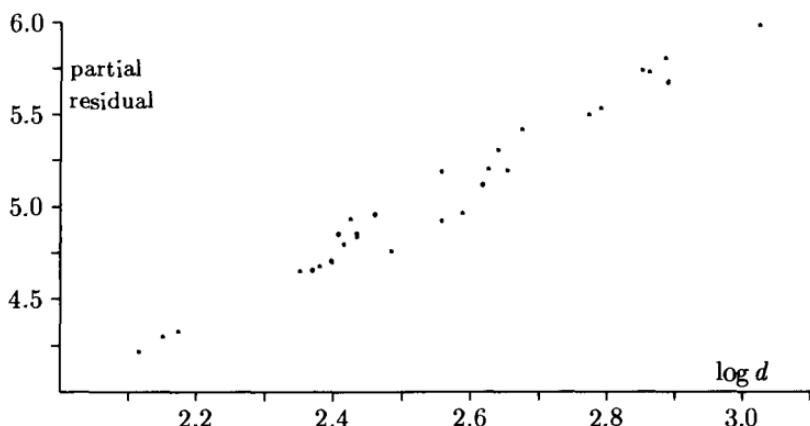


Fig. 12.5b. Partial residual plot for $\log d$ in the joint fit of $\log d$ and h to $\log(v)$.

Suppose, however, that someone with no understanding of dimensionality begins by regressing v linearly on d and h with Normal errors. The deviance is 421.9 with 28 d.f. The $(y, \hat{\mu})$ plot is curved upwards and the $(r, \hat{\mu})$ plot is quadratic. Addition of $\hat{\eta}^2$ to the linear predictor decreases the deviance by an enormous 242.6, about 57.5%.

We can now either transform y to a lower power, or choose an equivalent link function; for simplicity we shall follow the first path, and examine the effect of using $\log v$ in place of v . The result of adding $\hat{\eta}^2$ is still appreciable, the deviance falling from 0.262

to 0.181, though proportionally much less than before, while the absolute residual plot shows no obvious pattern: Remembering that the test based on adding $\hat{\eta}^2$ to the linear predictor may reflect either a faulty link or a faulty covariate scale, we next look at the partial residual plot of d , the more important of the two explanatory variables. This is shown in Fig. 12.5(a); it is curved downwards. That for h is more scattered and does not deviate obviously from linearity. The partial residual plot for d suggests a lower power for d and so we try $\log d$. Given that the dimensions of d and h are identical, external considerations suggest that we transform h to $\log h$ at the same time. The deviance is now 0.185, very similar to that given by the $\hat{\eta}^2$ test with d and h , and the addition of $\hat{\eta}^2$ does not further improve the fit. Both partial residual plots look linear and that for $\log d$ is shown in Fig. 12.5(b). There is no monotone trend in the absolute-residual plot, though all the big residuals are for points in the intermediate range. The formal test for the joint power transformation of d and h to d^θ and h^θ gives the deviance curve shown in Fig. 12.6. The minimum is at about $\hat{\theta} = 0.15$ where the deviance is 0.1829 with 27 d.f. giving $s^2 = 0.006530$. This gives 95% limits for the deviance of $0.1829 + 4s^2 = 0.2088$, corresponding to limits for θ of $(-0.32, 0.63)$. This excludes the original $\theta = 1$ and includes the final $\theta = 0$, corresponding to the log transformation.

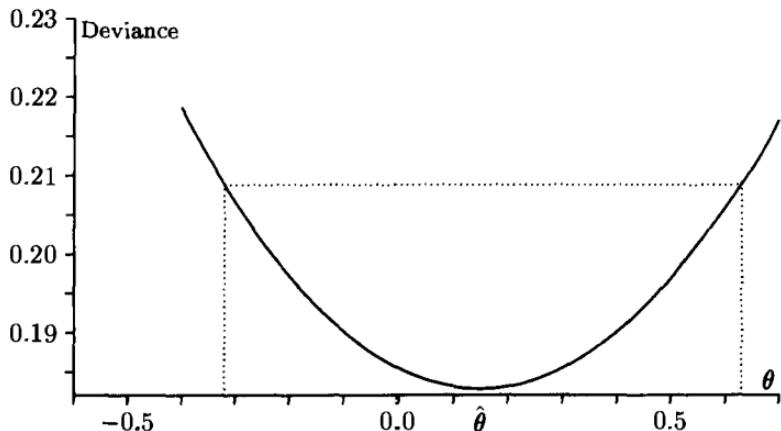


Fig. 12.2. *Minitab tree data: deviance for joint transformation of d and h to d^θ and h^θ .*

Note that our original guess at the relation, equation (12.11)

predicts coefficient for $\log d$ and $\log h$ of 2 and 1 respectively. Use of these in the fit gives a deviance of 0.1877 with 30 d.f., which is trivially different from 0.185 found from the model with the parameters estimated from the data.

Checks on isolated discrepancies show that:

trees 20 and 31 have high leverage;

trees 15 and 18 have the largest negative deletion residuals;

trees 11 and 17 have the largest positive deletion residuals;

tree 18 has a large modified Cook statistic.

The fit omitting point 18 does substantially reduce the deviance (to 0.154), but affects markedly only the intercept among the parameters. There is a curious cluster of extreme residuals for trees 15–18, which may be accidental. On the whole these latter checks do not suggest the rejection of any of the trees from the final model.

12.8.3 Insurance claims (continued)

In the initial analysis of these data in Chapter 8 a model was fitted using gamma errors and the inverse link. Subsequently the link function was embedded in the family $g(\mu; \lambda) = \mu^\lambda$ (Section 11.3.1), and separately the variance function was embedded in the same power family $V(\mu; \zeta) = \mu^\zeta$ (Section 12.6.2). The original choice of $\lambda_0 = -1$ and $\zeta_0 = 2$ was thus compared with the best fitting $\hat{\lambda}$ with $\zeta = \zeta_0$ fixed, and also with $\hat{\zeta}$ for $\lambda = \lambda_0$ held constant. We now consider a formal check on the joint settings (λ_0, ζ_0) against the best fitting $(\hat{\lambda}, \hat{\zeta})$ when both parameters are allowed to vary. The criterion is the extended (quasi) deviance when ϕ is also estimated, namely

$$\sum_i \log(\hat{\phi} V(y_i; \zeta))$$

where $\hat{\phi}$ is estimated by the mean deviance.

The contours are shown in Fig. 12.7 for the χ^2 values for $p = 0.50, 0.80, 0.95$ and 0.99 . The minimum occurs at $\zeta = 2.4$ and $\lambda = 0.75$, with the original choice of $(2, -1)$ lying comfortably inside the 95% contour. The fit for $(2, 0)$, i.e. with a log link, is less good, but again lies within the 95% contour. Note that the axes of the contours are closely aligned to the axes of (ζ, λ) , showing that the parameters are effectively orthogonal. Thus conservative 95% limits can be obtained by projecting onto the axes.

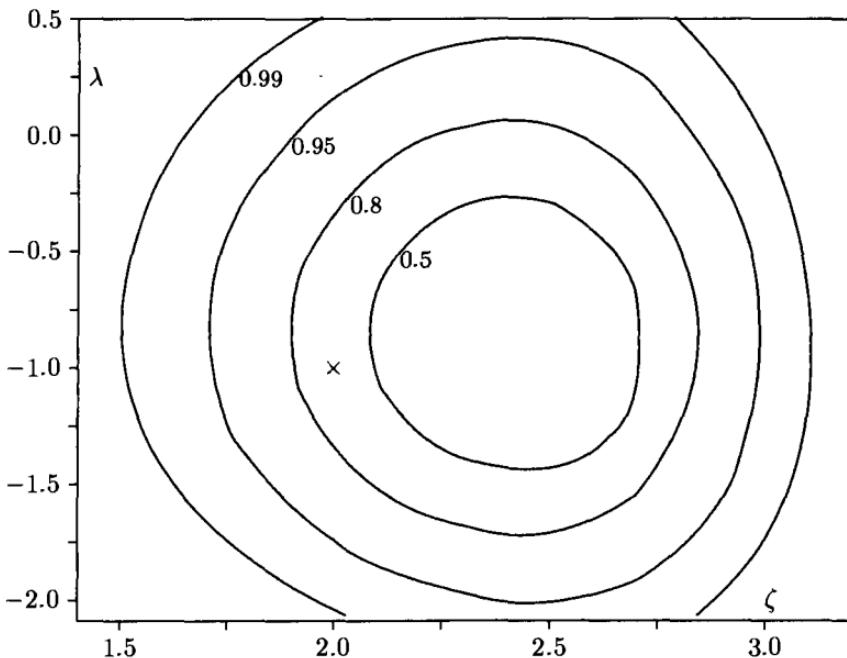


Fig. 12.7 Contour plot of the extended profile quasi-deviance for (ζ, λ) for the car-insurance data. The contours correspond to nominal confidence levels of 0.5, 0.8, 0.95 and 0.99.

12.9 A strategy for model checking?

This chapter has presented a number of techniques, both formal and informal, for checking the internal consistency of models, looking for both systematic deviation and for isolated unusual points. In principle we might hope to develop a strategy in which the various techniques are applied in an algorithmic fashion to give a complete check with accurate diagnoses of any deviations present. In practice such a strategy seems a long way off and model checking remains almost as much art as science. A major problem lies in the complex way that different deviations can interact. Thus a goodness-of-link test may give a significant result because of a faulty choice of link function; but it may also fail because one or more covariates is mis-scaled, or an interaction term is missing, or because of the presence of a few 'bad' points. Similarly an inconsistent point of high leverage may reflect the fact that the chosen model is breaking down at the edge of the treatment space,

or that someone made a recording error or a mis-transcription. ('All interesting points are wrong' is one cynic's view). Again many model-checking methods for extreme points are heavily dependent on the extreme points being fairly isolated. The occurrence of a small clump of such points may be much harder to identify. One promising possibility, for regression models, is the method of least median of squares, in which a very robust fit is used to identify such clumps of extreme points: see Atkinson (1986) and Rousseeuw and Leroy (1988). The action to be taken after identifying an inconsistent point is itself heavily dependent on the context of the problem and the special knowledge of the analyst. These considerations all lead to the presence of the question mark in the section heading.

12.10 Bibliographic notes

For an early account of methods for the examination of residuals, see Anscombe (1961) and Anscombe and Tukey (1963).

Most recent published work is for linear regression models, usually under the heading of regression diagnostics. Atkinson (1985) gives a very readable account, with some references to generalized linear models, though he mostly prefers the route via data transformation rather than the GLM specification involving link functions for the mean μ . See also the books by Cook and Weisberg (1982), Belsley *et al.* (1980) and Hawkins (1980).

Cox and Snell (1968) discuss residuals in a very general context including topics such as standardization and non-linear transformations. Goodness-of-link tests are discussed by Pregibon (1980) and by Atkinson (1982). The technique of using a constructed variable for detecting departures of a specific nature goes back to Tukey (1949). Other useful references are Andrews and Pregibon (1978) and Pregibon (1979).

Cook (1977, 1979) introduced the notion of influential observations in regression. Williams (1987) deals explicitly with methods for GLMs and introduces the modified form of deviance deletion residual. Chatterjee and Hadi (1986) review work in regression, while Kay and Little (1987) and Fowlkes (1987) deal specifically with the awkward case of binary data.

12.11 Further results and exercises 12

12.1 Schreiner Gregoire and Lawrie (1962) conducted an experiment to examine the effect of supposedly inert gases on fungal growth. Their data were presented in graphical form only: this report and the data are taken from Bliss (1967).

The fungus *Neurospora crassa* was grown at 30°C on an agar medium in tubes filled with an inert gas containing approximately 5% oxygen. The following growth rates in millimetres per hour, each the mean of 5 or 6 tests, were thought to be related to a suitable function of the molecular weight (MW) of the inert gas.

Table 12.2 *Rate of growth of the fungus Neurospora crassa*

Gas	He	Ne	N ₂	N ₂	Ar	Ar	Kr	Kr	Xe	Xe
MW	4.0	20.2	28.2	28.2	39.9	39.9	83.8	83.8	131.3	131.3
mm/hr	3.51	3.14	3.03	2.83	2.71	2.76	2.27	2.17	1.88	1.85

Source: Bliss (1967), p.471.

1. Plot the growth rate against MW and (MW)^{1/2}. Comment.
2. Taking the growth rate, R , as the response fit linear models using identity, log and reciprocal links, combined with various power transformations (identity, $x^{2/3}$, $x^{1/2}$, $x^{1/3}$ and log) of MW. For which combinations is the residual deviance smallest?
3. Interpret in biological terms the combination of powers obtained in part 1. What possible physical interpretation could be given to the $x^{2/3}$ transformation?
5. Examine the deviance residuals, first for pattern in the plot against MW, and second for conformity marginally with the Normal distribution.
5. Is the estimate of residual variance consistent with that obtained from the four replicate pairs?
6. Schreiner *et al.* report that

$$R = 3.88 - 0.1785(\text{MW})^{1/2}.$$

Is this summary consistent with your findings?

Table 12.3 *Relation in male cats of heart weight in gm. to body weight in kg.*

<i>Body wt.</i>	<i>Heart weight (gm.)</i>								
1.7	6.5	7.0							
1.8	5.8	7.3	6.1	7.1	7.7	7.4			
1.9	8.1	9.1	8.0	7.2	7.3	8.0			
2.0	6.5	6.5	6.7	7.5	7.8	8.1	8.6	7.7	
2.1	10.1	7.0	7.2	8.1	8.3				
2.2	7.2	7.6	10.7	9.6	9.1	7.9	8.5	9.6	8.9
2.3	9.6	9.6	8.5	8.8	8.2	9.2	8.7	8.9	
2.4	9.3	9.1	7.3	7.9	7.9	9.6	9.1	9.0	10.8 9.6
2.5	8.8	12.7	8.6	12.7	9.3	7.9	11.0	8.8	9.3 8.2 8.7 10.4 9.6
2.6	10.5	8.3	9.4	7.7	11.5	9.4	13.6	10.1	10.9 9.6 9.9
2.7	12.0	10.4	8.0	9.6	9.6	9.8	12.5	9.0	11.1 10.5 11.6 11.9
2.8	10.0	12.0	13.5	13.3	9.1	10.2	11.4	10.1	10.9
2.9	9.4	11.3	10.1	10.6	11.8				
3.0	13.3	10.0	13.8	10.6	12.4	12.7	10.4	11.6	12.2
3.1	9.9	12.1	14.3	12.5	11.5	13.0			
3.2	11.6	13.6	12.3	13.0	13.5	11.9			
3.3	11.5	14.9	14.1	15.4	12.0				
3.4	14.4	12.2	12.8	11.2	12.4				
3.5	15.6	11.7	15.7	12.9	17.2				
3.6	14.8	13.3	15.0	11.8					
3.7	11.0								
3.8	14.8	16.8							
3.9	14.4	20.5							

Source: Chen, Bliss and Robbins (1942)

12.2 Repeat the analysis of the data in Table 12.1, using an index plot of the deviance residuals to detect outliers. Compare this plot with that in Fig. 12.4. Comment on the differences and similarities.

12.3 Chen, Bliss and Robbins (1942) obtained the data shown in Table 12.3 as part of an assay experiment comparing the effect of *calotropin* with other cardiac substances such as uscarin and ouabain. Following the experiment the animals' hearts were weighed to see whether the cardiac effect might be more closely related to heart weight than to body weight. Table 12.3 shows the relationship between heart mass and body mass for 149 male cats used in the experiment.

1. Plot heart weight against body weight. Comment.

2. Fit the regression model of heart weight against body weight. Are the data consistent with a straight-line model passing through the origin? Fit the model passing through the origin.
3. Plot the residuals against body weight for the models fitted in part 2. Comment.
4. Regress $\log(\text{heart weight})$ against $\log(\text{body weight})$. Are the data consistent with the hypothesis that the slope is unity? What is the physiological interpretation of a unit slope? Fit the model in which the slope is unity and examine the residuals graphically.
5. Compute the mean of $\log(\text{heart weight})$ for each of the 23 distinct values of body weight. Regress these sample means on $\log(\text{body weight})$ using a weighted linear regression model. Compare the parameter estimates and standard errors in this weighted regression with those obtained in part 4. Explain the similarities and discrepancies observed.
6. For the model fitted in part 5, plot the residuals against the fitted values, taking care to use an appropriate standardization. Comment on this plot and on its relation to the plot in part 3.
7. Test the adequacy of the linear regression model in part 4 by including a non-linear term in $\log(\text{body weight})$.
8. Give a brief summary of your findings.

CHAPTER 13

Models for survival data

13.1 Introduction

This chapter deals with models for the analysis of data in which the response variate is the lifetime of a component or the survival time of a patient. Survival data usually refers to medical trials, but the ideas are useful also in industrial reliability experiments where the emphasis is on failure times rather than survival times.

Survival data are distinguished from most other types by the widespread occurrence of *censoring*. Censoring occurs when the outcome of a particular unit (patient or component) is unknown at the end of the study. Thus we may know only that a particular patient was still alive six months into the study, but the exact failure time is unknown either because the patient withdrew from the study or because the study ended while the patient was still alive. Censoring is so common in medical experiments that estimation methods must allow for it if they are to be generally useful.

A second characteristic of survival data is the frequent occurrence of *time-dependent covariates*. These arise when the status of a subject changes during a trial. Such a covariate \mathbf{x} , say, cannot be represented by a single value x_i for patient i , but takes values that may change with time.

13.1.1 *Survival functions and hazard functions*

Let the survival time, T , for individuals in a population have a density function $f(t)$. (In practice $f(\cdot)$ usually depends on other parameters, but for the moment we omit reference to these.) The corresponding distribution function

$$F(t) = \int_{-\infty}^t f(s) ds$$

is the fraction of the population dying by time t . The complementary function $1 - F(t)$, often called the *survivor function*, is the fraction still surviving at time t . The *hazard function* $h(t)$ measures the instantaneous risk, in that $h(t)\delta t$ is the probability of dying in the next small interval δt given survival to time t . From the relation

$$\begin{aligned} \text{pr(survival to } t + \delta t) \\ = \text{pr(survival to } t) \text{ pr(survival for } \delta t | \text{ survival to } t) \end{aligned}$$

we have

$$1 - F(t + \delta t) = \{1 - F(t)\} \{1 - h(t) \delta t\},$$

whence

$$\delta t F'(t) = \{1 - F(t)\} h(t) \delta t,$$

so that the hazard function is given by

$$h(t) = f(t)/\{1 - F(t)\}.$$

A distribution for survival times must have a hazard function with suitable properties. Thus for large t a hazard function should not decrease, because beyond a certain point the chance of breakdown or death does not ordinarily decrease with time. For small t various forms can be justified, including one that initially declines with t , for such a distribution could describe the behaviour of a machine part with a settling-in period, where reliability increases once the initial period is over.

The simplest hazard function, a constant, implies an exponential distribution of survival times and hence a Poisson process. For if T has the density

$$f(t) = \lambda e^{-\lambda t}; \quad t \geq 0,$$

then

$$F(t) = 1 - e^{-\lambda t},$$

and so

$$h(t) = \lambda.$$

Other forms of hazard function appear in later sections.

13.2 Proportional-hazards models

The hazard function depends in general both on time and on a set of covariates, some of which may be time-dependent. The proportional-hazards model separates these components by specifying that the hazard at time t for an individual whose covariate vector is \mathbf{x} is given by

$$h(t; \mathbf{x}) = \lambda(t) \exp\{G(\mathbf{x}; \boldsymbol{\beta})\},$$

where the second term is written in exponential form because it must be positive. This model implies that the ratio of the hazards for two individuals is constant over time provided that their covariates do not change. It is conventional, but not necessary (Oakes, 1981) to assume that the effects of the covariates on the hazard are also multiplicative. This additional assumption leads to models that may be written in the form

$$h(t; \mathbf{x}) = \lambda(t) \exp(\boldsymbol{\beta}^T \mathbf{x}), \quad (13.1)$$

where $\eta = \boldsymbol{\beta}^T \mathbf{x}$ is the linear predictor. The model thus implies that the ratio of hazards for two individuals depends on the difference between their linear predictors at any time, and so, with no time-dependent covariates, is a constant independent of time. This is a strong assumption that clearly needs checking in applications. Various assumptions may be made about the $\lambda(t)$ function. If a continuous survival distribution is assumed, $\lambda(t)$ is a smooth function of t , defined for all $t \geq 0$. Cox's model (Cox, 1972a) treats $\lambda(t)$ as analogous to the block factor in a blocked experiment, defined only at points where deaths occur, thus making no assumptions about the trend with time. In practice it frequently makes surprisingly little difference to estimates and inferences whether we put a structure on the base-line hazard function $\lambda(t)$ or not. We consider first estimation of $\boldsymbol{\beta}$ in the linear predictor with an explicit survival distribution, following closely the development in Aitkin and Clayton (1980).

13.3 Estimation with a specified survival distribution

We begin with the proportional-hazards model (13.1), and develop the likelihood for the data, some of which may be censored; for this we need both the density function and the survivor function.

From the definition of the hazard function we have

$$h(t) = F'(t)/\{1 - F(t)\} = \lambda(t)e^\eta,$$

so that

$$-\log\{1 - F(t)\} = \Lambda(t)e^\eta,$$

where

$$\Lambda(t) = \int_{-\infty}^t \lambda(u) du$$

is known as the cumulative hazard function. Thus the survivor function is given by

$$S(t) = 1 - F(t) = \exp\{-\Lambda(t)e^\eta\},$$

and the density function by minus its derivative, i.e.

$$f(t) = \lambda(t) \exp\{\eta - \Lambda(t)e^\eta\}.$$

At the end of the study an individual who died at time t contributes a factor $f(t)$ to the likelihood, while one censored at time t contributes $S(t)$. Suppose now that we define w as a variate taking the value 1 for an uncensored observation and value 0 for a censored one, and let there be n uncensored and m censored observations. Then the log likelihood takes the form

$$\begin{aligned} l &= \sum_{i=1}^{n+m} \{w_i \log f(t_i) + (1 - w_i) \log S(t_i)\} \\ &= \sum_i \{w_i \{\log \lambda(t_i) + \eta_i\} - \Lambda(t_i)e^{\eta_i}\} \\ &= \sum_i \left\{ w_i \{\log \Lambda(t_i) + \eta_i\} - \Lambda(t_i)e^{\eta_i} + w_i \log \left(\frac{\lambda(t_i)}{\Lambda(t_i)} \right) \right\}. \end{aligned}$$

Now if we write $\mu_i = \Lambda(t_i)e^{\eta_i}$, l becomes

$$\sum_i (w_i \log \mu_i - \mu_i) + \sum_i w_i \log \left(\frac{\lambda(t_i)}{\Lambda(t_i)} \right).$$

The first term is identical to the kernel of the likelihood function for $(n + m)$ independent Poisson variates w_i with means μ_i , while the second term does not depend on the unknown β s. Thus, given $\Lambda(t)$, we can obtain estimates of the β s by treating the censoring indicator variate w_i as Poisson distributed with mean $\mu_i = \Lambda(t_i)e^{\eta_i}$. The link function is the same as for log-linear models except that there is a fixed intercept $\log \Lambda(t_i)$ to be included in the linear predictor. Such a quantity is known in GLIM terminology as an offset.

The estimation process is less straightforward if the offset contains parameters of the survival density whose values are not known in advance. First, however, we deal with the exponential distribution for which no such difficulties arise.

13.3.1 *The exponential distribution*

For this distribution $\lambda(t)$ is the constant λ , so that the cumulative hazard function is

$$\Lambda(t) = \int_0^t \lambda(s) ds = \lambda t.$$

Thus $\lambda(t)/\Lambda(t) = 1/t$ and no extra parameters are involved. It follows that

$$\log \mu_i = \log t_i + \eta_i,$$

so that the offset is just $\log t_i$ and the log-linear model can be fitted directly. Two other distributions give particularly simple forms for $\Lambda(t)$ and these we now consider.

13.3.2 *The Weibull distribution*

By setting $\Lambda(t) = t^\alpha$, $\alpha > 0$, we obtain a hazard function proportional to $\alpha t^{\alpha-1}$ and a corresponding density $f(t)$ of the Weibull form:

$$f(t) = \alpha t^{\alpha-1} \{ \exp(\eta - t^\alpha e^\eta) \}; \quad t \geq 0.$$

Now $\lambda(t)/\Lambda(t) = \alpha/t$ and depends on the unknown parameter α , which must be jointly estimated with the β s. The kernel of the log-likelihood function is

$$n \log \alpha + \sum_i (w_i \log \mu_i - \mu_i),$$

Table 13.1 *Times of remission (weeks) of leukaemia patients, treated with drug (sample 1) and placebo (sample 2)*

<i>Sample 1</i>	(6)	6	6	6	7	(9)	(10)
10	(11)	13	16	(17)	(19)	(20)	
22	23	(25)	(32)	(32)	(34)	(35)	
<i>Sample 2</i>	1	1	2	2	3	4	4
	5	5	8	8	8	8	11
	11	12	12	15	17	22	23

Data from Freireich *et al.* (1963).

Figures in parentheses denote censored observations.

and, given α , the likelihood equations for the β s are the same as those for a log-linear model with offset $\alpha \log t_i$. The equation for α given the β s takes the form

$$n/\hat{\alpha} = \sum_i (\hat{\mu}_i - w_i) \log t_i. \quad (13.2)$$

The estimation procedure begins with $\alpha = 1$ (the exponential distribution), uses the log-linear model algorithm to fit the β s, then estimates α from (13.2), oscillating between the two stages until convergence is attained.

Note that the log likelihood for the model differs from that of the log-linear model by the inclusion of the extra term $n \log \hat{\alpha}$, so that the deviance requires adjustment by a term $-2n \log \hat{\alpha}$.

13.3.3 *The extreme-value distribution*

For this distribution $\Lambda(t) = e^{\alpha t}$, giving a hazard function proportional to $\alpha e^{\alpha t}$ and a density $f(t)$ in the extreme-value form

$$f(t) = \alpha e^{\alpha t} \exp(\eta - e^{\alpha t + \eta}). \quad (13.3)$$

Note that the transformation $u = \exp(t)$ transforms the distribution to the Weibull form. It follows that we need only replace t by u in the estimating procedure for the Weibull to obtain the corresponding one for this distribution.

13.4 Example: remission times for leukaemia

The data in Table 13.1 from Freireich *et al.* (1963) have been analysed by Gehan (1965), Aitkin and Clayton (1980) and others. There are two samples of 21 patients each, sample 1 having been given an experimental drug and sample 2 a placebo. The times of remission are given in weeks and figures in parentheses denote censored observations. To fit a survival model of the type discussed in section 13.3, we set up a pseudo-Poisson variable taking values 0 for the censored and 1 for the uncensored observations. To fit the exponential distribution we apply an offset of $\log t$, and models with a single mean and with separate sample means (S) give deviances as follows:

<i>Model</i>	<i>Deviance</i>	<i>d.f.</i>
1	54.50	41
<i>S</i>	38.02	40

The more general Weibull distribution yields for model S the value $\hat{\alpha} = 1.366$ with a deviance of 34.13 on 39 d.f. The reduction in deviance of 3.89 is thus marginally significant at the 5% level. Separate fits of α to the two samples give similar estimates of 1.35 and 1.37; the estimate of the sample difference for $\hat{\alpha} = 1.366$ is $b_1 = 1.731$, corresponding to a hazard ratio of $\exp(1.731) = 5.65$ for sample 2 as compared with sample 1. The standard error of b_1 , for α with a fixed prior value of 1.366, is ± 0.398 ; to adjust this for the simultaneous fitting of α , we must border the information matrix for the two parameters in the linear predictor with the second derivatives that include α and then invert the expanded matrix. The details for this example are given in Aitkin and Clayton (1980) and give $b_1 = 1.73 \pm 0.41$; the validity of this SE may be checked by plotting the deviance for fixed values of β_1 in the neighbourhood of the estimate 1.73. The resulting curve rises slightly more steeply on the lower than on the upper side, but the effect at the $2 \times \text{SE}$ distance is quite small. Thus our 95% limits for the log hazard difference are $1.73 \pm (1.96 \times 0.41) = (0.93, 2.53)$ corresponding to hazard ratios of (2.52, 12.6).

The data of this example have been analysed by Whitehead (1980) using Cox's model and treating ties by both Peto's and Cox's

Table 13.2 Comparison of estimators for the leukaemia data

Model (treatment of ties)	b_1	SE
Exponential	1.53	0.40
Weibull	1.73	0.41
Cox (Peto)	1.51	0.41
Cox (Cox)	1.63	0.43

methods. His results (after correcting observation 6 in sample 1, which was censored), together with those obtained above using parametric survival functions, are summarized in Table 13.2. The estimates all fall within a range of about half a standard error, and the increase in standard error from the Cox model as against the parametric survival functions is quite small. Efron (1977) and Oakes (1977) discuss this phenomenon from a theoretical viewpoint.

13.5 Cox's proportional-hazards model

Cox's (1972a) version of the proportional-hazards model is only partially parametric in the sense that the baseline hazard function $\lambda(t)$ is not modelled as a smooth function of t . Instead, $\lambda(t)$ is permitted to take arbitrary values and is irrelevant in the sense that it does not enter into the estimating equations derived from Cox's partial likelihood (Cox, 1975).

13.5.1 Partial likelihood

The argument used to derive the partial likelihood function is as follows. First observe that we need only consider times at which failures occur because, in principle at least, the hazard could be zero over intervals that are free of failures and no contribution to the likelihood would be made by these intervals. Let $t_1 < t_2 < \dots$ be the distinct failure times and suppose for simplicity that there are no tied failure times. The risk set immediately prior to the j th failure, $R(t_j)$, is the set of individuals any of whom may be found to fail at time t_j . Thus, individuals who have previously failed or who have been censored are excluded from $R(t_j)$. Given that one failure is to occur in the interval $(t_j - \delta t, t_j)$, the relative probabilities of failure for the individuals in $R(t_j)$ are proportional

to the values of their hazard functions. Let \mathbf{x}_j be the value of the covariate vector for the failed individual. The probability under the proportional-hazards model that the individual who fails at time t_j is the one actually observed is

$$\frac{\lambda(t) \exp(\boldsymbol{\beta}^T \mathbf{x}_j)}{\sum \lambda(t) \exp(\boldsymbol{\beta}^T \mathbf{x})} = \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_j)}{\sum \exp(\boldsymbol{\beta}^T \mathbf{x})}, \quad (13.4)$$

where summation extends over the risk set $R(t_j)$.

This conditional probability is the probability of observing \mathbf{x}_j in sampling from the finite population corresponding to the covariate vectors in $R(t_j)$, where the selection probabilities are proportional to $\exp(\boldsymbol{\beta}^T \mathbf{x})$. This is a generalization of the non-central hypergeometric distribution (section 7.3.2). This argument effectively reverses the roles of random failure times and fixed covariates to fixed failure times and covariates selected according to the probability distribution described above.

The partial likelihood for $\boldsymbol{\beta}$ is the product over the failure times of the conditional probabilities (13.4), and so independent of the baseline hazard function $\lambda(t)$. These conditional probabilities have the form of a linear exponential-family model so that $\boldsymbol{\beta}$ can be estimated by equating the vector sum of the covariates of the failed individuals to the sum of their conditional means. Note, however, that the conditioning event changes from one failure time to the next as individuals are removed from the risk set either through failure or through censoring.

13.5.2 The treatment of ties

The occurrence of ties among the failure times complicates the analysis, and several techniques have been proposed for dealing with this complication. One method due to Cox (1972a) is as follows. Suppose for definiteness that two failures occur at time t and that the vector sum of the covariates of these two failed individuals is \mathbf{s}_j . The factor corresponding to (13.4) is then defined to be

$$\exp(\boldsymbol{\beta}^T \mathbf{s}_j) / \sum \exp(\boldsymbol{\beta}^T \mathbf{s}), \quad (13.5)$$

where the sum in the denominator extends over all distinct pairs of individuals in $R(t_j)$. In other words we construct the finite

population consisting of sums of the covariate vectors for all distinct pairs of individuals in the risk set at time t_j . The probability under an exponentially weighted sampling scheme that the failures were those of the pair actually observed is given by (13.5), which again has the exponential-family form. Note however that the number of terms in the denominator of (13.5) quickly becomes exceedingly large for even a moderate number of ties at any failure time.

Any reasonable method for dealing with ties is likely to be satisfactory if the number of failed individuals constitutes only a small fraction of the risk set. In fact the likelihood contribution (13.5) is exact only if failures are thought of as occurring in discrete time. In practice, however, ties occur principally because of grouping. With grouped data the appropriate likelihood (Peto, 1972) involves the sum over all permutations of the failed individuals consistent with the ties observed. Suppose, for example, that two failures are tied and that the failed individuals have covariate vectors \mathbf{x}_1 and \mathbf{x}_2 . The probability for the sequence in time $(\mathbf{x}_1, \mathbf{x}_2)$ or $(\mathbf{x}_2, \mathbf{x}_1)$, either of which is possible given the tie, is

$$\frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_1)}{\sum_R \exp(\boldsymbol{\beta}^T \mathbf{x})} \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_2)}{\sum_{R_1} \exp(\boldsymbol{\beta}^T \mathbf{x})} + \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_2)}{\sum_R \exp(\boldsymbol{\beta}^T \mathbf{x})} \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_1)}{\sum_{R_2} \exp(\boldsymbol{\beta}^T \mathbf{x})}, \quad (13.6)$$

where R_j is the risk set excluding \mathbf{x}_j ($j = 1, 2$). Clearly the likelihood contribution becomes increasingly cumbersome as the number of ties becomes appreciable.

Expressions (13.5) and (13.6) for the contribution to the likelihood can both be derived by arguments involving exponentially weighted sampling from a finite population without replacement. If the number of ties is small we may use the simpler expression

$$\frac{\exp(\boldsymbol{\beta}^T \mathbf{s})}{\{\sum_R \exp(\boldsymbol{\beta}^T \mathbf{x})\}^m}, \quad (13.7)$$

where \mathbf{s} is the sum of the covariate vectors of the m tied individuals (Peto, 1972). This term corresponds to sampling with replacement.

13.5.3 Numerical methods

The likelihood formed by taking the product over failure times of the conditional probabilities (13.4) can, in principle, be maximized directly using the weighted least-squares method discussed in Chapters 2 and 8. Alternatively we can regard the covariate vector of the failed individuals as the response and condition on the set of covariates of all individuals in the risk set at each failure time, these being regarded as fixed. If we write \mathbf{y} for the covariate vector of the failed individual the log likelihood for one failure time takes the form

$$\boldsymbol{\beta}^T \mathbf{y} - \log \left\{ \sum \exp(\boldsymbol{\beta}^T \mathbf{x}) \right\},$$

with summation over the risk set. This has the form of an exponential family model with canonical parameter $\boldsymbol{\beta}$ and $b(\theta)$ (in the notation of section 2.2) equal to $\log \left\{ \sum \exp(\boldsymbol{\beta}^T \mathbf{x}) \right\}$. The (conditional) mean is then given by $b'(\theta)$ and the variance by $b''(\theta)$. However, this formulation is unhelpful computationally because there is no explicit expression for the quadratic weight (here equal to the variance function) as a function of the mean.

The computational difficulty can be avoided by a device similar to that used in section 13.4. Suppose that k_j individuals are at risk immediately prior to t_j and that just one individual is about to fail. If we regard the observation on the failed individual as a multinomial observation with k_j categories, taking the value 1 for the failed observation and 0 for the remainder, then the contribution to the likelihood is again of the form (13.4), but now interpreted as a log-linear model for the cell probabilities. Thus the numerical methods of Chapter 5 may be used provided that the algorithm allows variable numbers of categories for the multinomial observations.

Alternatively (Whitehead, 1980) a Poisson log likelihood may be used provided that a blocking factor associated with failure times is included. The idea here is that at each failure time each individual in the risk set contributes an artificial Poisson response of 1 for failure and 0 for survival. The mean of this response is $\exp(\alpha + \boldsymbol{\beta}^T \mathbf{x})$ for an individual whose covariate value is \mathbf{x} and α represents the blocking factor associated with failure times. Because of the equivalence of the Poisson and multinomial likelihoods discussed in section 6.4, the estimate of $\boldsymbol{\beta}$ and the estimate of its precision are identical to those obtained from the multinomial likelihood and

hence to the partial likelihood.

The computations can be simplified if the number of distinct covariate vectors is small so that individuals in the risk set may be grouped into sets of constant hazard. The adjustment for ties is simple for the third method described above (often called Peto's method). In the multinomial log likelihood we set the multinomial total equal to the observed number of tied failures at that time. No adjustment to the algorithm is required. The corresponding Poisson log likelihood is equivalent to Peto's version of the partial likelihood.

Whitehead (1980) describes the adjustments to the Poisson likelihood required to maximize the likelihood corresponding to Cox's method for dealing with ties.

13.6 Bibliographic notes

The recent literature on the analysis of survival data includes books by Cox and Oakes (1984), Elandt-Johnson and Johnson (1980), Gross and Clark (1975), Lawless (1982), Lee (1980), Kalbfleisch and Prentice (1980) and Miller (1981).

Cox's model was proposed by Cox (1972a), and fitting via GLIM discussed by Whitehead (1980); the pseudo-Poisson model for parametric survival functions was proposed by Aitkin and Clayton (1980), who also discuss the definition of residuals and the necessary adaptation of standard graphical techniques (see also Crowley and Hu 1977). For a comparison of Cox and Weibull models, see Byar (1983).

13.7 Further results and exercises 13

13.1 In medical trials the recruitment of patients frequently continues over a prolonged period, spanning perhaps the entire trial. Consider such a trial to test a new drug that is claimed to benefit patients suffering from angina by reducing the incidence of coronary disease. The protocol specifies eligible patients to be those aged 55–75, showing symptoms of angina who have no previous record of heart attack and are taking no other medication. After being judged eligible and consent has been obtained, a patient

is randomized to one of two groups, either the new drug or the standard treatment.

Discuss how you might analyse the data that have accumulated after two years in such a trial. Consider in particular the following points.

1. What are appropriate definitions of failure:
 - deaths from all causes;
 - deaths from coronary disease only;
 - all heart attacks whether fatal or not.
2. Choice of origin for the time scale:
 - calendar time from the beginning of the study;
 - time from individual patient randomization;
 - time from first appearance of patient's angina symptoms.
3. Non-compliance because of non-fatal side-effects:
4. Who to include in the risk set:
 - all known survivors among those randomized;
 - all survivors excluding those no longer complying.

13.2 Let $X_{(1)} < X_{(2)} < \dots < X_{(n)}$ be an ordered sample of *i.i.d.* exponential random variables of unit mean. Define the normalized differences

$$Y_1 = nX_{(1)}, \quad Y_i = (n - i + 1)(X_{(i)} - X_{(i-1)}), \quad i = 2, \dots, n.$$

Show that Y_1, \dots, Y_n are *i.i.d.* exponential random variables of unit mean.

Components of dispersion

14.1 Introduction

The models so far considered all have the property that the observations are assumed independent and the variability associated with each observation is determined by at most a single dispersion factor, σ^2 or ϕ , to be estimated. In many areas of application, however, the experimental or survey design is such that variation is present at several levels or strata. In the absence of external effects, units in the same cluster tend to be more alike, or positively correlated, than units in different clusters. In an educational-testing context pupils within classrooms within schools form a natural hierarchy of three strata, and it is natural to associate a random component to the units within each of the three strata. Usually these variance components, or dispersion components as we shall call them, are given contrasting descriptive labels such as ‘between-schools variance’ and ‘between-classrooms-within-schools variance’, or ‘between-animals variance’ and ‘within-animals variance’ depending on the context. Usually the between-blocks or between-animals variance is larger than the within-blocks or within-animals variance. For example in the tuberculin-assay problem discussed in section 6.3.1, where four measurements were taken on each cow, it was anticipated that the variability of reaction between different cows would be considerably larger than the variability between sites on the same animal. Consequently, provided that there is no interaction between tuberculins applied at different sites, greater precision can be achieved by comparing tuberculins on the same animal.

14.2 Linear models

Before proceeding to consider more complicated non-linear models it is helpful to consider first the essential ingredients of linear models with so-called ‘random effects’. We take as our example the tuberculin-assay data in Table 6.1. To keep the discussion as simple as possible it is assumed initially that all effects are additive with constant variance on the log scale. This assumption is almost, but not quite, in accord with the analysis in section 6.3.1.

The standard mathematical description is to express the response as an additive function of both fixed (non-stochastic) and random effects, all random effects being assumed independent. Thus we write

$$\log Y_{ij(k)} = \alpha_i + \gamma_j + \tau_k + \epsilon_{ijk}, \quad (14.1)$$

where k denotes the tuberculin type and volume applied at site i to cows in cow class j . Primary interest centres on the tuberculin effects τ_k , and the analysis in section 6.3 focused exclusively on the joint effect of tuberculin type and volume. Of secondary interest, but nonetheless perhaps physiologically important, is the sensitivity to tuberculin of the four sites on the cow’s neck. Thus τ_k and α_i are considered here as fixed effects. Since there can be no lasting scientific interest in the sensitivity of individual unidentified cows, it is appropriate to take γ_j , the cow-class effect corresponding to differences between cows, as independent random variables. Finally, the residual errors ϵ_{ijk} are taken to be mutually independent and independent of the cow-class effects γ_j .

Since $Y_{ij(k)}$ is the sum of observed values for 30 cows, it is appropriate to take

$$\text{var}(\epsilon_{ijk}) = \sigma^2/30, \quad \text{var}(\gamma_j) = \sigma_b^2/30.$$

Thus σ^2 is the ‘within-cow’ variance and σ_b^2 is the ‘between-cow’ variance. Note that $\text{var}(\log Y) \simeq \text{cv}^2(Y)$ (Exercises 14.10–11). Further, if Y_i are independent,

$$\text{cv}^2\{\sum Y\} \simeq \text{cv}^2(Y)/n,$$

where $\text{cv}(Y)$ is the coefficient of variation of Y . The variances σ^2 and σ_b^2 then refer to individual cows and not to averages or totals over 30 cows.

The usual analysis-of-variance decomposition is shown in Table 14.1. From the 'error' line in this table we obtain the estimate $\tilde{\sigma}^2 = 0.0216$ or $\tilde{\sigma} = 0.147$. In other words, the within-cow coefficient of variation is approximately 15%. From the 'cow-class' line together with the 'error' line we find $\tilde{\sigma}_b^2 = 0.8821$ or $\tilde{\sigma}_b = 0.939$, showing that the between-cow coefficient of variation is approximately 94%. In other words there is very substantial variation between animals, and comparatively little variation within animals.

Table 14.1 *Analysis of variance for the tuberculin-assay data*

<i>Source</i>	<i>S.S.</i>	<i>d.f.</i>	<i>M.S.</i>	<i>E(M.S.)</i>
<i>Cow class</i>	0.47232	3	0.11833	$\sigma^2/30 + 4\sigma_b^2/30$
<i>Sites</i>	0.08324	3	0.02775	$\sigma^2/30 + 4 \sum (\alpha_i - \bar{\alpha})^2/3$
<i>Treatments</i>	0.17596	3	0.05865	$\sigma^2/30 + 4 \sum (r_k - \bar{r})^2/3$
<i>Error</i>	0.00433	6	0.00072	$\sigma^2/30$
<i>Total</i>	0.73584	15		

Because of the orthogonality built into the Latin square design the site effects and the tuberculin effects are uncorrelated, and independent under Normality. Further, the variances of the estimated tuberculin effects and site effects are functions only of σ^2 and not of σ_b^2 . For this reason, if interest focuses on the site and tuberculin effects, the same results would be obtained if we were to condition on the cow-class effects and take them as fixed. The estimated site effects are

$$\boldsymbol{\alpha} = (0.000, 0.093, 0.128, -0.053),$$

and the standard errors of simple contrasts are $\tilde{\sigma}/\sqrt{60} = 0.019$. Thus sites 2 and 3 are significantly more sensitive than sites 1 and 4.

14.3 Non-linear models

The discussion in this section is meant to be quite general, but in order to keep the terminology concrete we use the tuberculin-assay example. The analysis that follows is performed on the original scale and avoids explicitly transforming the data. This feature is a characteristic of generalized linear models.

Suppose that the conditional mean and variance of $Y_{ij(k)}$ given the cow-class effects are as follows:

$$\begin{aligned} E(Y_{ij(k)} | \text{assignment of cows to classes}) &= M_{ij(k)}, \\ \text{var}(Y_{ij(k)} | \text{assignment of cows to classes}) &= \sigma^2 V(M_{ij(k)}), \end{aligned}$$

where $V(\cdot)$ is the known conditional variance function. All observations, whether in the same cow class or in different classes, are assumed to be conditionally independent. Suppose in addition that the conditional mean satisfies the log-linear model

$$\log M_{ij(k)} = \alpha_i + \gamma_j + \tau_k \quad (14.2)$$

in which the treatment effects τ_k , and possibly the site effects α_i , are the parameters of interest. Evidently this conditional formulation specifies a generalized linear model, which happens in this case to be log-linear. In this example all parameters, including γ_j , can be estimated in the usual way without further assumptions, and no new issues arise. The analysis in section 6.3 corresponds to the choice $V(M) = M$, whereas the analysis in the previous section via response transformation corresponds roughly to $V(M) = M^2$.

Suppose now that we insert into the model the further, quite reasonable, assumption that the cow-class effects are independent and identically distributed random variables from a particular family of distributions, say

$$\gamma_j \sim N(0, \sigma_b^2). \quad (14.3)$$

This additional assumption does not invalidate the conditional analysis recommended above, but it does open up the possibility that a more efficient analysis could be devised by making use of the random-effects assumption (14.3).

The most illuminating examples of this type occur when the parameters of interest are not estimable in the conditional model (14.2). Estimation is then impossible in the absence of a further assumption such as (14.3). Suppose, by way of example, that the tuberculin-assay design is such that all cows in a given class receive the same treatment at each of the sites. Suppose further that two replicates of the experiment are available, so that there are eight cow classes in all. In other words the experimental arrangement

Table 14.2 Alternative design for tuberculin assay problem

Cow class (j)	I	II	III	IV	V	VI	VII	VIII
Treatment (k)	A	B	C	D	A	B	C	D
1								
2								
3								
4								
					$Y_{ij(k)}$			

is as shown in Table 14.2. Evidently the treatment contrasts are now aliased with a subset of the cow-class contrasts, and are not estimable in the fixed-effects model (14.2).

Taking the log-linear model (14.2) together with the random-effects assumption (14.3), however, we have

$$M_{\cdot j(k)} = \sum_i \exp(\alpha_i) \times \exp(\gamma_j) \times \exp(\tau_k)$$

so that the unconditional means for the cow-class totals are

$$\begin{aligned} \mu_{\cdot j(k)} &= E(M_{\cdot j(k)}) = \sum_i \exp(\alpha_i) \times E(e^\gamma) \times \exp(\tau_k) \\ &= \sum_i \exp(\alpha_i) \times \exp(\sigma_b^2/2) \times \exp(\tau_k). \end{aligned}$$

On the log scale we have

$$\log(\mu_{\cdot j(k)}) = \tau_k + \text{const}, \quad (14.4)$$

which depends only on the treatment applied. Clearly the eight cow-class totals are independent random variables. Their unconditional variances are given by

$$\text{var}(Y_{\cdot j(k)}) = \sigma^2 \sum_i E\{V(M_{ij(k)})\} + \text{var}(M_{\cdot j(k)}). \quad (14.5)$$

If $V(M) = M$ this gives

$$\begin{aligned} \text{var}(Y_{\cdot j(k)}) &= \sigma^2 \mu_{\cdot j(k)} + \mu_{\cdot j(k)}^2 \text{cv}^2(e^\gamma) \\ &= \sigma^2 \mu_{\cdot j(k)} + \mu_{\cdot j(k)}^2 (\exp(\sigma_b^2) - 1). \end{aligned}$$

On the other hand if $V(M) = M^2$ we have

$$\begin{aligned} \text{var}(Y_{\cdot j(k)}) &= \sigma^2 \sum_i E(M_{ij(k)}^2) + \text{var}(M_{\cdot j(k)}) \\ &= \sigma^2 \sum_i \mu_{ij(k)}^2 \{1 + \text{cv}^2(e^\gamma)\} + \mu_{\cdot j(k)}^2 \text{cv}^2(e^\gamma) \\ &= \mu_{\cdot j(k)}^2 \{\text{cv}^2(e^\gamma) + \frac{\sigma^2}{4} (1 + \text{cv}^2(e^\gamma)) (1 + \text{cv}^2(e^\alpha))\}, \end{aligned}$$

where $\text{cv}^2(\alpha') = \sum(\alpha'_i - \bar{\alpha}')^2/(n\bar{\alpha}'^2)$. In the first case the unconditional variance function is approximately quadratic provided that $\sigma_b^2 \bar{\mu} \gg \sigma^2$. In the second case the unconditional variance function is exactly quadratic. Thus the parameters in (14.4) can be estimated from the eight cow-class totals by using a quadratic variance function and a log link.

From two replicates the coefficient of variation of the cow-class totals can be estimated on four degrees of freedom. It is this estimate that must be used for setting confidence limits for treatment effects.

For the estimation of site effects it is unnecessary to use the random effects assumption (14.3). The site effects are estimable in the fixed-effects model (14.2) and their variances do not depend on σ_b^2 . From two replicates the residual variance σ^2 can be estimated on 21 degrees of freedom using the residual deviance from model (14.2).

14.4 Parameter estimation

Parameter estimates are obtained using the quasi-likelihood estimating equation (9.5). We use this method in preference to explicit maximum likelihood chiefly because it is often much simpler and is based only on properties of the unconditional mean and covariance matrix of the observations. Suppose that the unconditional mean and covariance matrix of \mathbf{Y} are

$$E(\mathbf{Y}) = \boldsymbol{\mu}(\boldsymbol{\beta}), \quad \text{cov}(\mathbf{Y}) = \mathbf{V}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2),$$

where the components of $\boldsymbol{\sigma}^2$ are the dispersion components. Then, following the notation established in Chapter 9, the estimating equations for $\boldsymbol{\beta}$ are $\mathbf{U}(\hat{\boldsymbol{\beta}}, \boldsymbol{\sigma}^2) = \mathbf{0}$, where

$$\mathbf{U}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \mathbf{D}^T \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})). \quad (14.6)$$

Ordinarily, but with some important exceptions, the solution to this equation depends on the dispersion components, usually on ratios of dispersion components. In such cases it is necessary to use a supplementary set of estimating equations for the dispersion parameters. Such a system of estimating equations is now derived.

In all of the examples that we have in mind the dispersion components are physical characteristics of identifiable populations and the covariance matrix can then be written in the form

$$\mathbf{V}(\boldsymbol{\mu}, \sigma^2) = \sigma_1^2 \mathbf{V}_1(\boldsymbol{\mu}) + \dots + \sigma_k^2 \mathbf{V}_k(\boldsymbol{\mu}). \quad (14.7)$$

Usually the rank of $\mathbf{V}_j(\boldsymbol{\mu})$ is equal to the number of elements in the sample that are drawn from the population indexed by j .

By way of illustration we consider the example discussed in the previous section, in which there are two dispersion components. The response variances are given by

$$\text{var}(Y_{ij(k)}) = \sigma^2 E(V(M_{ij(k)})) + \text{var}(M_{ij(k)}).$$

If the conditional variance function is $V(M) = M$, this expression reduces to

$$\text{var}(Y_{ij(k)}) = \sigma^2 \mu_{ij(k)} + \{\exp(\sigma_b^2) - 1\} \mu_{ij(k)}^2.$$

Similar calculations show that the only non-zero covariances have the form

$$\begin{aligned} \text{cov}(Y_{ij(k)}, Y_{i'j(k')}) &= \text{cov}(M_{ij(k)}, M_{i'j(k')}) \\ &= \mu_{ij(k)} \mu_{i'j(k')} \{\exp(\sigma_b^2) - 1\} \end{aligned}$$

for $i' \neq i$. On comparing these expressions with the general form (14.7) we find $\sigma_1^2 = \sigma^2$, $\mathbf{V}_1(\boldsymbol{\mu}) = \text{diag}(\boldsymbol{\mu})$, $\sigma_2^2 = \exp(\sigma_b^2) - 1$, and $\mathbf{V}_2(\boldsymbol{\mu}) = \boldsymbol{\mu} \mathbf{J} \boldsymbol{\mu}^T$, where \mathbf{J} is a block-diagonal unit matrix taking the value unity if two observations refer to the same cow class, and zero otherwise. Note that the rank of \mathbf{V}_2 is equal to the number of cow classes.

The most natural way to estimate the dispersion components is to choose k suitable quadratic forms and to equate the observed values of these to their expectations as functions of the parameters. Usually the root of (14.6) is fairly insensitive to the choice of σ^2 , so for that purpose the choice of quadratic forms is not critical. However, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$,

$$\text{cov}(\hat{\boldsymbol{\beta}}) \simeq (\mathbf{D}^T \mathbf{V}^{-1} \mathbf{D})^{-1}$$

does depend heavily on σ^2 .

In many applications there is a natural set of quadratic forms

$$Q_r = (\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{P}_r (\mathbf{Y} - \boldsymbol{\mu}), \quad r = 1, \dots, k, \quad (14.8)$$

in which \mathbf{P}_r is a fixed projection matrix associated with the r th random effect. Usually $\mathbf{P}_r \mathbf{Y}$ is equivalent to a set of marginal totals, in which case Q_r is the sum of squares for that set of totals. Provided that $\mathbf{V}(\boldsymbol{\mu})$ satisfies the additive decomposition (14.7), these quadratic forms have expectations

$$E(Q_r) = \sum_{j=1}^k \text{tr}\{\mathbf{P}_r \mathbf{V}_j\} \sigma_j^2, \quad (14.9)$$

which are linear functions of the dispersion components. In this way we obtain k simultaneous equations for the k dispersion components. As usual with variance components, the estimates obtained need not be positive.

More generally the method of restricted maximum likelihood, as described in section 7.2, can be used here, even though the observations are usually not Normally distributed. However this method gives rise to more complicated estimating equations and requires the inversion of $\mathbf{V}(\boldsymbol{\mu}, \sigma^2)$ for various values of the dispersion parameters. For that reason we use the simpler but less efficient method described above.

In practice, if $p = \dim(\boldsymbol{\beta})$ is positive, it is necessary to use $\hat{\boldsymbol{\mu}}$ in place of $\boldsymbol{\mu}$ in (14.8) and to iterate between (14.6) and (14.8). This method of estimation is illustrated in the following section.

14.5 Example: A salamander mating experiment

14.5.1 Introduction

This section describes a particular problem, involving an experiment with binary responses, for which neither the methods of Chapter 4 nor those of Chapter 9 are directly applicable. In fact straightforward application of linear logistic models could lead to misleading conclusions, so it is essential to recognize the characteristics of a design in which explicit recognition of dispersion components is necessary. The problem described here involves several dispersion components associated with different subgroups or

Table 14.3. *Design used in a salamander mating experiment[†]*

Females	Males					
	June 4	June 8	June 12	June 16	June 20	June 24
1	1	4	5	1	4	5
2	5	5	3	3	1	2
RBF 3	RB 2	WS 1	RB 1	WS 4	RB 3	WS 3
4	4	2	2	5	5	4
5	3	3	4	2	2	1
6	9	9	10	7	6	8
7	8	8	9	9	7	6
RBF 8	WS 6	RB 6	WS 7	RB 10	WS 10	RB 9
9	10	7	8	6	9	10
10	7	10	6	8	8	7
1	9	9	7	10	10	8
2	7	6	9	7	6	10
WSF 3	RB 8	WS 7	RB 6	WS 9	RB 7	WS 6
4	10	10	8	8	9	9
5	6	8	10	6	8	7
6	5	2	3	4	2	1
7	4	1	5	2	1	5
WSF 8	WS 1	RB 4	WS 2	RB 5	WS 5	RB 3
9	3	3	1	1	4	4
10	2	5	4	3	3	2

[†]RBF rough-butt female; WSF whiteside female; RB rough-butt male; WS whiteside male.

Each female is paired with three males of each type: each male is paired with three females of each type.

populations, and is not to be confused with simple over-dispersion in which the observations are independent and there is only one dispersion component.

The analysis presented here is admittedly rather rough and ready. It involves at various stages Taylor approximations whose adequacy, even for applied work, is more dubious than usual. However, we believe that problems of this nature are sufficiently common, and suitable statistical techniques either primitive or excessively complicated, that a protracted discussion of a particular example is worthwhile.

14.5.2 *Experimental procedure*

The purpose of this experiment, conducted by S. Arnold and P. Verrell of the Department of Ecology and Evolution at the University of Chicago, was to study the extent to which mountain dusky salamanders from different populations will interbreed. These populations, all belonging to the same species, are geographically isolated from one another, and are found at high elevations in the southern Appalachian mountains of the eastern United States. Thus the salamanders paired together in the laboratory would never encounter each other in their natural environment—at least not at this moment in geological time. The manner in which mechanisms that prevent interbreeding evolve is of great interest to biologists studying speciation. The question of interest here is whether these barriers to interbreeding can evolve in populations that are isolated from one another.

The data given here refer to two populations called Rough Butt (RB) and Whiteside (WS). Forty animals were used in each of three experiments, one conducted in the summer of 1986 and two in the Fall of the same year. Normal practice is to use fresh animals for each experiment but in this instance the animals used in the first Fall experiment were identical to those used in the Summer experiment. Repeat experimentation using the same animals is potentially important but, because it is the exception to normal practice, this aspect will be ignored in the analyses that follow. The three experiments will be treated as if different animals were used in each.

The forty salamanders available in each of the three experiments comprise

- 10 Rough Butt males numbered 1–5 & 6–10 (RBM)
- 10 Rough Butt females numbered 1–5 & 6–10 (RBF)
- 10 Whiteside males numbered 1–5 & 6–10 (WSM)
- 10 Whiteside females numbered 1–5 & 6–10 (WSF)

According to the design in Table 14.3, which was used in each of the three experiments, Rough Butt females numbered 1–5 (RBF 1–5) were sequestered as heterosexual pairs with Rough Butt males numbered 1–5 on 3 occasions June 4, 12 and 20, and with Whiteside males numbered 1–5 also on 3 occasions, June 8, 16 and 24. These Rough Butt females were never permitted to see RBM 6–10 nor WSM 6–10. Thus, in effect, the design contains two replicates as

Table 14.4. *Observed matings: Summer '86*

Females	Date					
	June 4	June 8	June 12	June 16	June 20	June 24
1	1	1	1	0	1	1
2	1	1	1	1	1	1
RBF 3	RB 1	WS 0	RB 1	WS 1	RB 1	WS 1
4	1	1	1	0	1	1
5	1	1	1	1	1	1
6	1	1	1	0	1	1
7	0	0	0	1	0	0
RBF 8	WS 0	RB 1	WS 0	RB 0	WS 1	RB 1
9	0	0	1	1	1	1
10	0	0	1	0	1	0
1	0	1	1	1	0	1
2	0	0	0	1	0	0
WSF 3	RB 0	WS 0	RB 0	WS 0	RB 0	WS 1
4	0	1	1	1	0	1
5	0	1	0	0	0	0
6	0	0	1	0	0	0
7	1	1	1	0	1	1
WSF 8	WS 1	RB 0	WS 1	RB 0	WS 1	RB 0
9	1	1	1	1	1	0
10	1	0	0	1	1	0

shown in Table 14.8. The non-zero values in Table 14.8 are the actual numbers of matings observed for the various crosses in the Summer '86 experiment. It should be pointed out that, although there are 25 possible crosses between the females RBF 1-5 and the males WSM 1-5, only 15 of these crosses are permitted by the design. For example, RBF 1 was not permitted to see RBM 2 or RBM 3. Conversely RBM 1 did not see RBF 4 or RBF 5.

In order for the design to be complete in this sense it would have been necessary to extend the experiment over 10 nights instead of 6. This was considered impractical because four days of recuperation are required between successive pairings, enabling females to transport sperm, and males to synthesize spermatophores. A complete design would have extended the duration of the experiment from 21 days to 37 days. The design used here comprises eight replicates of an incomplete Latin square, although little use is made of that design in the present analysis.

The design of the experiment permits a comparison of the

Table 14.5. *Observed matings: Fall '86 (re-runs)*

<i>Females</i>	<i>Date</i>					
	Sept. 4	Sept. 8	Sept. 12	Sept. 16	Sept. 20	Sept. 24
1	1	0	1	0	1	0
2	1	1	1	0	0	1
RBF 3	RB 0	WS 1	RB 1	WS 1	RB 1	WS 1
4	1	0	0	0	0	0
5	1	0	1	0	0	0
6	1	1	1	0	0	0
7	1	1	0	1	0	1
RBF 8	WS 1	RB 0	WS 0	RB 0	WS 1	RB 0
9	1	1	1	1	0	1
10	0	0	1	1	1	0
1	0	1	0	0	0	1
2	0	1	0	0	0	1
WSF 3	RB 0	WS 1	RB 0	WS 1	RB 0	WS 1
4	0	0	1	1	0	1
5	0	0	0	0	0	0
6	1	0	1	1	1	1
7	1	0	1	0	0	0
WSF 8	WS 0	RB 0	WS 0	RB 0	WS 0	RB 0
9	1	1	1	1	1	1
10	1	0	1	1	1	0

mating probabilities for the four possible crosses. In the analysis that follows we concentrate on comparing the mixed crosses. In other words we focus on the question of whether the RBF/WSM crosses result in mating more frequently or less frequently than the WSF/RBM crosses. The observed number of matings in the Summer '86 experiment for each of the crosses is shown in Table 14.7. For the comparison in question we observe that the RBF/WSM cross resulted in 20 matings out of 30 encounters: the WSF/RBM cross resulted in only 7 matings out of 30 encounters. While the difference is apparently quite substantial, it is difficult to set confidence limits on the difference or on the odds ratio because the Bernoulli observations in Table 14.4 are not independent. Even if there is no time trend it would be a gross over-simplification to assume that there is no variation among individual males or among females or to assume that the six observations on any one animal are independent. A binomial comparison is inappropriate

Table 14.6. *Observed matings: Fall '86*

Females	Date					
	Oct. 4	Oct. 8	Oct. 12	Oct. 16	Oct. 20	Oct. 24
1	1	1	1	0	1	1
2	0	0	0	1	0	0
RBF 3	RB 1	WS 1	RB 1	WS 0	RB 1	WS 1
4	1	0	1	1	0	0
5	0	1	1	0	1	0
6	0	1	1	1	1	1
7	0	1	0	0	0	0
RBF 8	WS 1	RB 1	WS 1	RB 1	WS 1	RB 0
9	0	0	1	0	1	1
10	0	1	1	1	1	1
1	0	0	0	1	0	1
2	0	1	0	1	0	1
WSF 3	RB 1	WS 0	RB 1	WS 0	RB 0	WS 1
4	1	1	1	0	0	0
5	0	1	1	1	0	1
6	0	0	1	0	0	0
7	1	0	1	0	0	0
WSF 8	WS 0	RB 0	WS 1	RB 0	WS 1	RB 0
9	1	0	0	0	1	0
10	0	0	1	0	1	0

and potentially misleading here because the variability in the totals is undoubtedly in excess of that predicted by the binomial model.

14.5.3 A linear logistic model with random effects

It is convenient to begin by considering a model for the observed data conditionally on the actual animals used in the experiment. From this conditional perspective it is reasonable to suppose that the 120 observations in Table 14.4 are independent but not identically distributed. For the sake of simplicity we assume here that there is no time trend and that all effects are additive on the logistic scale. The conditional fixed-effects model formula is therefore

$$\text{individual female effect} + \text{individual male effect} + \text{cross effect}. \quad (14.10)$$

The male and female effects are factors whose 20 levels identify the animals, and the cross effect has the four levels R/R, R/W, W/R and W/W in the order female/male. The cross effect is partly

aliased with the sum of the male and female effects, and hence not all cross-effect contrasts are estimable in (14.10).

Table 14.7. *Observed number of matings in an incomplete crossed design (Summer '86)*

<i>Females</i>	<i>Males</i>			
	RBM 1-5	WSM 1-5	RBM 6-10	WSM 6-10
RBF 1-5	15	12	—	—
WSF 6-10	5	12	—	—
RBF 6-10	—	—	7	8
WSF 1-5	—	—	2	9

We do not propose to fit the fixed-effects model for a number of reasons. First, most of the contrasts of interest are not estimable. Second, some of the individual effects may be estimated as $\pm\infty$. Rough Butt females 2 and 5 had success rates of 6/6, so their effects are estimated as $+\infty$ on the logistic scale. Third, the individual animal effects are of scientific importance only to the extent that they mimic the populations from which they are drawn. The main advantage of considering (14.10) explicitly is that the random-effects model of interest is a sub-model of (14.10). Consequently the statistic that is sufficient for the fixed-effects model must also be sufficient for the random-effects model.

In order to make further progress or to draw any interesting scientific conclusions it is necessary to make further assumptions regarding the relationship between the experimental animals and the populations that they are supposed to represent. We shall assume here that the experimental animals are, either in fact or in effect, a random sample from their respective populations. The individual effects in (14.10) are therefore regarded as random variables with variances σ_M^2 and σ_F^2 . The male and female variances are assumed constant across populations.

The sufficient statistic, \mathbf{S} , for the fixed-effects model is the set of totals for each animal and cross type. The values for Summer '86 are shown in Table 14.8.

In what follows we apply the method of quasi-likelihood either to the vector \mathbf{S} or to the original observations \mathbf{Y} . We use unconditional expectations and variances, and denote the four probabilities of interest by π_{RR} , π_{RW} , π_{WR} and π_{WW} . If we denote by Y_{ij} the observation corresponding to female i and male j , then $E(Y_{ij})$ is

Table 14.8 Components of the sufficient statistic \mathbf{S} : Summer '86

Animal	Random effects				Systematic effects	
	RBF	WSF	RBM	WSM	Observed	Fitted
1	5	4	5	4	R/R	22
2	6	1	3	5	R/W	20
3	5	1	5	6	W/R	7
4	5	4	3	5	W/W	21
5	6	1	4	4		
6	5	1	2	3		
7	1	5	1	1		
8	3	3	2	5		
9	4	5	3	4		
10	2	3	1	4		
<i>Fitted</i>		4.2	2.8	2.9	4.1	
s^2		2.84	2.84	2.10	1.88	

equal to one of these four probabilities according to the populations from which the male and female are drawn. (At no point in the experiment was the same pair observed twice, so this notation introduces no ambiguity.) Similarly, the components of $E(\mathbf{S})$ are linear functions of the four probabilities.

The unconditional covariance matrix of Y_{ij} is assumed to have the pattern implied by (14.10), namely that observations on non-overlapping pairs of animals are independent. Selected elements of the covariance matrix are assumed to have the following form:

$$V_{ij,kl} = \begin{cases} \pi_{RR}(1 - \pi_{RR}) & \text{if } i = k, j = l, \text{ both of type R;} \\ \sigma_F^2 g^2(\pi_{RR}) & \text{if } i = k [= R]; j \neq l, \text{ both R;} \\ \sigma_M^2 g(\pi_{RW})g(\pi_{WW}) & \text{if } i \neq k [= R, W]; j = l [= W]; \\ 0 & \text{if } i \neq k, j \neq l. \end{cases} \quad (14.11)$$

and so on for some known function $g(\cdot)$. Other non-zero components have the form

$$\pi_{RW}(1 - \pi_{RW}), \quad \sigma_F^2 g^2(\pi_{RW}), \quad \sigma_M^2 g^2(\pi_{RW}), \quad \sigma_F^2 g(\pi_{RR})g(\pi_{RW}), \dots,$$

depending on whether the two observations are identical, have one Rough Butt female in common, one Whiteside male in common, one Rough Butt female in common, and so on. This covariance

matrix with $g(\pi) = \pi(1 - \pi)$ can be obtained via a Taylor approximation based on (14.10) provided that σ_M^2 and σ_F^2 are sufficiently small. For details see Exercise 14.2.

With this covariance matrix it can be shown that the quasi-likelihood estimates of the four probabilities are

$$\begin{aligned}\hat{\pi}_{RR} &= \sum_{R/R} Y_{ij}/30, & \hat{\pi}_{RW} &= \sum_{R/W} Y_{ij}/30, \\ \hat{\pi}_{WR} &= \sum_{W/R} Y_{ij}/30 \quad \text{and} \quad \hat{\pi}_{WW} &= \sum_{W/W} Y_{ij}/30,\end{aligned}\tag{14.12}$$

as indicated in Table 14.8. For a sketch of the derivation see Exercises 14.8–9. The remarkable aspect of this is that the quasi-likelihood estimates do not depend on the values of σ_F^2 or σ_M^2 , though their distributions do. Provided that the covariance matrix has the form (14.11) the quasi-likelihood estimates are the same as if the two dispersion components were zero.

Since the parameter estimates are linear functions of the data, the covariance matrix of $\hat{\boldsymbol{\pi}}$ may be obtained in terms of the dispersion components using (14.11). We find after some considerable algebraic reduction that

$$30 \operatorname{cov}(\hat{\boldsymbol{\pi}}) = \Pi(\mathbf{I} - \Pi)$$

$$+ \mathbf{G} \begin{pmatrix} 2(\sigma_F^2 + \sigma_M^2) & 3\sigma_F^2 & 3\sigma_M^2 & 0 \\ 3\sigma_F^2 & 2(\sigma_F^2 + \sigma_M^2) & 0 & 3\sigma_M^2 \\ 3\sigma_M^2 & 0 & 2(\sigma_F^2 + \sigma_M^2) & 3\sigma_F^2 \\ 0 & 3\sigma_M^2 & 3\sigma_F^2 & 2(\sigma_F^2 + \sigma_M^2) \end{pmatrix} \mathbf{G}\tag{14.13}$$

where $\mathbf{G} = \operatorname{diag}\{g(\pi_{RR}), g(\pi_{RW}), g(\pi_{WR}), g(\pi_{WW})\}$ and $\Pi = \operatorname{diag}\{\boldsymbol{\pi}\}$. A similar but slightly more complicated expression can be obtained if the exact covariance matrix is used in place of (14.11). For example the exact variance of $30\hat{\pi}_{RR}$ is

$$\begin{aligned}\pi_{RR}(1 - \pi_{RR}) &+ \operatorname{cov}(F(\alpha_{RR} + \epsilon_1 + \delta_1), F(\alpha_{RR} + \epsilon_1 + \delta_2)) \\ &+ \operatorname{cov}(F(\alpha_{RR} + \epsilon_1 + \delta_1), F(\alpha_{RR} + \epsilon_2 + \delta_1)),\end{aligned}$$

while the covariance of $30\hat{\pi}_{RR}$ and $30\hat{\pi}_{RW}$ is

$$3 \operatorname{cov}(F(\alpha_{RR} + \epsilon_1 + \delta_1), F(\alpha_{RW} + \epsilon_1 + \delta_2)),\tag{14.14}$$

where $F(\cdot)$ is the cumulative logistic function, and ϵ_i , with variance σ_F^2 , and δ_i , with variance σ_M^2 , are the female and male random effects. Only the first term of the Taylor expansions is included in (14.13).

The quasi-likelihood estimates are exactly the maximum likelihood estimates obtained from the linear logistic model containing only the four-level factor 'cross type'. The covariance matrix obtained from this analysis is the first term in (14.13). The second term arises entirely as a result of covariances among the observations, and could conceivably dominate the first term.

14.5.4 Estimation of the dispersion parameters

From the approximate covariance matrix (14.11) we see that the covariance matrix of $\{Y_{ij}\}$ is expressible in the form

$$\mathbf{V}_0(\boldsymbol{\pi}) + \sigma_F^2 \mathbf{V}_1(\boldsymbol{\pi}) + \sigma_M^2 \mathbf{V}_2(\boldsymbol{\pi}).$$

Consequently the expected value of any quadratic form in the residuals is linear in the dispersion components σ_F^2, σ_M^2 . For purposes of estimation the easiest way to proceed is to choose two suitable quadratic forms and to equate the observed values of these to their expectations as functions of σ_M^2, σ_F^2 . This gives a pair of simultaneous linear equations for the dispersion components. For reasons given in the previous section such quadratic forms should be functions of the sufficient statistic S .

In Table 14.8 the expected value of each entry in column 1 is $3\pi_{RR} + 3\pi_{RW}$, estimated as 4.2, which is the column mean. Similar calculations apply to the next three columns. The variance of each entry in column 1 is

$$\begin{aligned} \text{var}(S_1) &= 3\pi_{RR}(1 - \pi_{RR}) + 3\pi_{RW}(1 - \pi_{RW}) \\ &\quad + 6\sigma_F^2 \{ \pi_{RR}^2(1 - \pi_{RR})^2 + \pi_{RW}^2(1 - \pi_{RW})^2 + 3\pi_{RR}(1 - \pi_{RR})\pi_{RW}(1 - \pi_{RW}) \} \end{aligned}$$

which we denote by $\kappa_2(S)$. Within column 1 some of the covariances are zero: the non-zero covariances are not all equal because some pairs of females have one male in common and others two. The sum of the covariances between S_i and S_j for $i \neq j$ in column 1 is

$$90\bar{\kappa}_{11}(S) = 60\sigma_M^2 \{ \pi_{RR}^2(1 - \pi_{RR})^2 + \pi_{RW}^2(1 - \pi_{RW})^2 \}.$$

It follows from the result established in Exercise 14.6 that the expected mean square for column 1 is

$$E(\text{mean square}) = \kappa_2(S) - \bar{\kappa}_{11}(S),$$

which is estimated as

$$1.2533 + 1.3080\sigma_F^2 - 0.0584\sigma_M^2.$$

Similar calculations for the second column give an expected mean square of

$$1.1667 + 1.1328\sigma_F^2 - 0.0507\sigma_M^2.$$

The estimates used here are based on the pooled mean square for females, and the pooled mean square for males. Details of the calculations are shown in Table 14.9.

Table 14.9 *Estimation of the dispersion components from S*

Source	Mean square	$E(\text{Mean square})$	$\hat{\sigma}^2$
RBF	2.8444	$1.2533 + 1.3080\sigma_F^2 - 0.0584\sigma_M^2$	
WSF	2.8444	$1.1667 + 1.1328\sigma_F^2 - 0.0507\sigma_M^2$	
Total F	5.6888	$2.4200 + 2.4408\sigma_F^2 - 0.1091\sigma_M^2$	$1.3704 = \hat{\sigma}_F^2$
RBM	2.1000	$1.1233 + 1.0511\sigma_M^2 - 0.0468\sigma_F^2$	
WSM	1.8778	$1.2967 + 1.4009\sigma_M^2 - 0.0623\sigma_F^2$	
Total M	3.9778	$2.4200 + 2.4520\sigma_M^2 - 0.1091\sigma_F^2$	$0.6963 = \hat{\sigma}_M^2$

The results of similar calculations for the Fall '86 experiments are summarized in Table 14.10. Evidently the mating probabilities are quite consistent across the three experiments, $\hat{\pi}_{WR}$ being consistently lower than the other three probabilities. The estimated dispersion components are similar for the first two experiments, but the pattern is reversed for the third experiment. It should be borne in mind, however, that the variability of these estimates is appreciable, and negative estimates are not impossible. Because of correlations among the components of S it is difficult to assign degrees of freedom to the sums of squares in Table 14.9.

Using the pooled estimates of all parameters together with (14.13), we find that the estimated covariance matrix of $90\hat{\pi}$ is

$$\begin{pmatrix} 0.2222 & & \\ & 0.2469 & \\ & & 0.1665 \end{pmatrix} + \begin{pmatrix} 0.1772 & 0.1506 & 0.0977 \\ 0.1506 & 0.2188 & 0.1448 \\ 0.0977 & & 0.0995 & 0.1015 \\ & & 0.1448 & 0.1015 & 0.1772 \end{pmatrix}.$$

Table 14.10 *Summary of parameter estimates from three experiments.*

Experiment	Parameter estimate				Dispersion component	
	$\hat{\pi}_{RR}$	$\hat{\pi}_{RW}$	$\hat{\pi}_{WR}$	$\hat{\pi}_{WW}$	$\tilde{\sigma}_F^2$	$\tilde{\sigma}_M^2$
Summer '86	0.7333	0.6667	0.2333	0.7000	1.3704	0.6963
Fall '86 (re-run)	0.6000	0.4667	0.2333	0.6667	0.9787	0.5997
Fall '86	0.6667	0.5333	0.1667	0.6333	0.3954	1.3440
Pooled estimate	0.6667	0.5556	0.2111	0.6667	0.9148	0.8800

The random effects account for about half of the total variability in the parameter estimates, although not all contrasts among the $\hat{\pi}$ s are affected equally (Exercise 14.12). The pooled estimate of the mixed contrast is $\hat{\pi}_{RW} - \hat{\pi}_{WR} = 0.3445$ with estimated standard error 0.0904. Thus the evidence for a non-zero mixed contrast is evidently very strong. From this analysis there is no evidence of differences among the probabilities π_{RR}, π_{RW} and π_{WW} .

Despite the fact that the magnitudes of the random effects as estimated in Table 14.10 are approximately equal for the two sexes it appears from closer inspection that the nature of the effect for males is quite different than for females. A comparison of the two sets of animal totals for the Summer '86 and the first Fall '86 experiment shows a strong correlation for males but no evidence of any correlation for females. Thus it appears that the male random effects persist over several months at least whereas the female effects are short-lived. This conclusion applies to both male populations and both female populations.

14.6 Bibliographic notes

Linear models in which there is more than one variance component have been used for many years going back at least to the work of Fisher and Yates in agricultural experiments. For detailed accounts of specific designs see Kempthorne (1952, Chapter 9) or Cox (1958c, Chapters 7,8). For the estimation of variance components more generally in cases where the design is unbalanced, the method of restricted maximum likelihood is usually preferred to ordinary maximum likelihood. For an account of this see section 7.2, Patterson and Thompson (1971) or Harville (1974, 1977).

There is a parallel but extensive literature on educational test-

ing, which uses ordinary maximum-likelihood estimation for unbalanced Normal-theory linear models. See, for example, Goldstein (1986) or Bock (1989).

Although the need has long been recognized by practising statisticians, the development of analogous models and techniques of estimation for non-Normal data or for non-linear effects has proceeded very slowly. Some notable exceptions are Stiratelli, Laird and Ware (1984), who use a Bayesian argument, with a diffuse prior on the regression coefficients, for the estimation of the dispersion parameters. Gilmour, Anderson and Rae (1985) discuss a random-effects probit model. For Normal-theory linear models the previous two methods reduce to restricted maximum likelihood. Anderson and Aitkin (1985) develop an unmodified maximum-likelihood estimation procedure for nested variance components in linear logistic and probit models. Their likelihood is obtained by direct numerical integration assuming Normal random effects, and the resulting estimates are not the same as those suggested here.

In the previous three papers the effects of interest are defined as linear contrasts of the *conditional* logits. In section 14.5.3, however, the effects of interest are defined in terms of the *unconditional* probabilities or logits: the random-effects model (14.10) is used only to justify the choice of covariance matrix. A qualitatively similar argument is given by Prentice (1988). Zeger, Liang and Self (1985) and Liang and Zeger (1986) also argue in favour of specifying models for the unconditional rather than the conditional probabilities, but the choice must ultimately be governed by what the interesting parameters are in any given context. For further discussion regarding the distinction between subject-specific models such as (14.10) and population-averaged models corresponding to the estimates (14.12), see Zeger, Liang and Albert (1988).

Morton (1987) considers models for counted data in which the random effects are nested and effects are multiplicative. His method of estimation, using quasi-likelihood estimating functions, is very similar to that used here. The salamander example is rather unusual in that the random effects are crossed rather than nested.

Despite the apparent paucity of references, this subject is extremely important with a broad range of applications. At the time of writing it is evident that the topic is a rich source of good research problems.

14.7 Further results and exercises 14

14.1 Consider the vector \mathbf{S} whose 44 components are displayed in Table 14.8. Show that $T_1 = S_1 - S_{\text{RR}}/10 - S_{\text{RW}}/10$ is uncorrelated with S_{RR} , S_{RW} , S_{WR} and S_{WW} , and that $E(T_1) = 0$. Hence deduce that the quasi-likelihood estimates of the parameters are given by (14.12).

14.2 Suppose that ϵ is a Normal random variable with mean zero and variance σ^2 . Define the following random variable and its expectation:

$$P = \frac{\exp(\alpha + \epsilon)}{1 + \exp(\alpha + \epsilon)}, \quad \text{and} \quad \pi = E(P).$$

Justify empirically the approximation $\pi \simeq F(\alpha^*)$, where $F(x) = e^x/(1 + e^x)$, and

$$\alpha^* = \alpha - \frac{1}{2}\sigma^2 \tanh\{\alpha(1 + 2 \exp(-\sigma^2/2))/6\}.$$

For $\sigma^2 < 2$, the maximum error of this approximation is about 0.003 on the probability scale. Show that the approximation has the correct limiting behaviour in the limit as $\alpha \rightarrow \pm\infty$ for fixed σ^2 .

By differentiating the above approximation with respect to α , show that the variance of P is approximately

$$\text{var}(P) \simeq \sigma^2 \pi(1 - \pi)\pi^\dagger(1 - \pi^\dagger) \times \frac{1}{3}(1 + 2 \exp(-\sigma^2/2)),$$

where $\pi^\dagger = F(\alpha(1 + 2 \exp(-\sigma^2/2))/3)$. This approximation is considerably more accurate than the Taylor approximation in (14.11).

14.3 Suppose that ϵ_1, ϵ_2 are Normal random variables with variances σ_1^2, σ_2^2 , and correlation ρ . Show that the correlation ρ' of $F(\epsilon_1)$ and $G(\epsilon_2)$ satisfies

$$\frac{|\rho|}{|\rho'|} \simeq 1 + \frac{1}{4} \left\{ \sigma_1^2 \left(\frac{F''}{F'} \right)^2 + \sigma_2^2 \left(\frac{G''}{G'} \right)^2 - 2\rho\sigma_1\sigma_2 \frac{F''G''}{F'G'} \right\} \geq 1$$

when terms of order σ^4 and smaller are ignored. Under what conditions is ρ' equal to ρ ?

14.4 Using the notation of the previous Exercise, define the random variables P_1, P_2 , together with their expectations as follows:

$$\begin{aligned} P_1 &= \frac{\exp(\alpha_1 + \epsilon_1)}{1 + \exp(\alpha_1 + \epsilon_1)}, & P_2 &= \frac{\exp(\alpha_2 + \epsilon_2)}{1 + \exp(\alpha_2 + \epsilon_2)}, \\ \pi_1 &= E(P_1), & \pi_2 &= E(P_2). \end{aligned}$$

Using the results given in the previous Exercises, find an approximation for the covariance of P_1 and P_2 .

Compute the exact covariance numerically for variances and covariances in the range 0–2. Comment on the adequacy of the approximate formula.

14.5 Compute the components of the sufficient statistic \mathbf{S} for the first fall '86 experiment (data in Table 14.5). Compare the individual animal totals with those in Table 14.8. Compute the correlation coefficients or regression coefficients of the Fall totals on the Summer totals for each of the four groups. Comment on your findings.

14.6 Show that if Y_1, \dots, Y_n have common mean μ and covariance matrix $\kappa_{i,j}$, then the expected value of

$$s^2 = \frac{1}{n-1} \sum (Y_i - \bar{Y})^2$$

is given by

$$E(s^2) = \frac{1}{n} \sum_i \kappa_{i,i} - \frac{1}{n(n-1)} \sum_{i \neq j} \kappa_{i,j}.$$

14.7 The pooled estimate of the dispersion components in Table 14.10 is the average estimate for the three experiments. By pooling together the sums of squares for F and M from the three experiments and equating these totals to their expectations, show that the pooled estimates are $\tilde{\sigma}_F^2 = 0.9035$ and $\tilde{\sigma}_M^2 = 0.8759$.

14.8 Suppose that the random vector \mathbf{Y} satisfies the linear model

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}, \quad \text{cov}(\mathbf{Y}) = \mathbf{V}.$$

Let $\tilde{\beta}$ be the ordinary least-squares estimate $\tilde{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ and let \mathbf{R} be the residual vector

$$\mathbf{R} = (\mathbf{I} - \mathbf{P})\mathbf{Y} = \mathbf{Y} - \mathbf{X}\tilde{\beta}.$$

Let \mathcal{X} denote the column space of \mathbf{X} . Show that if $\mathbf{V}x$ lies in \mathcal{X} for each $x \in \mathcal{X}$ then \mathbf{R} and $\tilde{\beta}$ are uncorrelated and

$$\tilde{\beta} = \hat{\beta} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y},$$

where $\mathbf{W} = \mathbf{V}^{-1}$. Deduce also that $\mathbf{V}^{-1}x \in \mathcal{X}$. [Kruskal, 1968].

14.9 Using the result established in the previous exercise show that the maximum quasi-likelihood estimates of the parameters in the salamander example are given by (14.12) provided that the experiment is suitably balanced. Obtain the required balance condition.

14.10 Show that if Y has the gamma distribution $G(\mu, \nu)$, then

$$\text{var}\{\log(Y)\} = \psi'(\nu),$$

where $\psi(x) = \Gamma'(x)/\Gamma(x)$. Under what conditions is the approximation $\text{var}(\log Y) \simeq \text{cv}^2(Y)$ adequate for the gamma family?

14.11 Repeat the calculations of the previous exercise under the assumption that Y has the log-Normal distribution, $\log(Y) \sim N(\mu, \sigma^2)$.

14.12 Consider the model formula (14.10) written in matrix notation in the form

$$\mathbf{Z}\mathbf{b} + \mathbf{X}\boldsymbol{\pi},$$

in which \mathbf{Z} is a (120×40) incidence matrix identifying the animals paired together in each trial, and \mathbf{X} is a (120×4) incidence matrix for the cross type, R/R, R/W, W/R and W/W. Show that the matrix $(\mathbf{I} - \mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T)\mathbf{X}$ has rank 1, with span corresponding to the contrast R/R - R/W - W/R + W/W. Hence justify the claim made in section 14.5.3 that most of the effects of interest are not estimable in (14.10).

Using (14.12) and (14.13) find an estimate of the above contrast and obtain a numerical estimate of the standard error, (i) under the assumption that the between-animal variances are zero, and (ii) using the estimates given in Table 14.10. Explain briefly why the standard error in (ii) is smaller than in (i).

CHAPTER 15

Further topics

15.1 Introduction

This chapter describes briefly a number of topics related to generalized linear models, some of which are of current research interest.

15.2 Bias adjustment

In large samples the bias of maximum-likelihood estimators is $O(n^{-1})$, and hence negligible compared with standard errors. For samples of more modest size, or for problems in which the number of parameters is appreciable compared with n , the bias may not be entirely negligible. In such cases the usual approximations can often be improved by making a bias adjustment to the maximum-likelihood estimate. In what follows we describe how the leading term in the asymptotic bias can be computed by weighted linear regression.

15.2.1 *Models with canonical link*

In the case of full exponential-family models with canonical link function, such as linear logistic models for binomial data, log-linear models for Poisson data, inverse linear models for exponential data, the approximate bias of the maximum-likelihood estimate can be obtained by a very simple supplementary computation, which we now describe.

Using tensor notation with implicit summation over indices that appear twice, the components of the approximate bias vector $\mathbf{b} = E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ are given as follows:

$$E(\hat{\boldsymbol{\beta}}^r - \boldsymbol{\beta}^r) = b^r \simeq -\frac{1}{2} \kappa^{r,s} \kappa^{t,u} \kappa_{s,t,u}. \quad (15.1)$$

For a derivation of this and related asymptotic formulae for maximum likelihood estimators, see McCullagh (1987, Chapter 7). In this formula $\kappa^{r,s}$ are the components of the inverse Fisher information matrix, elsewhere written using matrix notation as $(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$, where, for canonical-link models, $\mathbf{W} = \text{cov}(\mathbf{Y})$. The expression for the components of the three-way array $\kappa_{s,t,u}$ in terms of the model matrix $\mathbf{X} = \{x_r^i\}$ is

$$\kappa_{s,t,u} = \sum_{i=1}^n x_s^i x_t^i x_u^i \kappa_{3i},$$

where κ_{3i} is the third cumulant of the i th component of the response vector.

As an intermediate step in the derivation it is helpful to consider the contracted array with components

$$b_s = -\frac{1}{2} \kappa_{s,t,u} \kappa^{t,u} = -\frac{1}{2} \sum_i x_s^i \kappa_{3i} x_t^i x_u^i \kappa^{t,u}. \quad (15.2)$$

Using matrix notation, the product $x_s^i x_u^i \kappa^{t,u}$ is written as the $n \times n$ symmetric matrix

$$\mathbf{Q} = \mathbf{X}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T,$$

which is the asymptotic covariance matrix of $\hat{\eta}$. Consequently the final factor appearing in (15.2) is the i th diagonal element of \mathbf{Q} , which we write as Q_{ii} . Thus

$$b_s = -\frac{1}{2} \sum_i x_s^i \kappa_{2i} Q_{ii} \frac{\kappa_{3i}}{\kappa_{2i}}.$$

Using matrix notation the components on the right of the above equation are just $\mathbf{X}^T \mathbf{W} \xi$, where $\xi_i = -\frac{1}{2} Q_{ii} \kappa_{3i} / \kappa_{2i}$.

Evidently from (15.2) the bias vector \mathbf{b} is obtained by premultiplying the components b_s by the inverse Fisher information matrix. This operation gives

$$\mathbf{b} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \xi, \quad (15.3)$$

which is easily obtained as the vector of regression coefficients in the formal linear regression of ξ on \mathbf{X} with \mathbf{W} as weight vector. In other words we retain the weights and the model formula from the log-linear or linear logistic model, but the link function becomes the identity and response vector becomes ξ . The binomial index vector and any prior weights are assumed to be incorporated into \mathbf{W} .

15.2.2 Non-canonical models

For non-canonical models the tensor expression for the first-order asymptotic bias of $\hat{\beta}$ is a little more complicated because it involves the covariance between the vector of first-order derivatives and the matrix of second-order derivatives of the log-likelihood function. However, calculations similar to those given in the previous section show that the bias vector can be obtained by a similar supplementary regression computation. It is only necessary to re-define the formal response vector in this regression and then to use (15.3). We find that the components of ξ are given by

$$\xi_i = -\frac{1}{2} \left(\frac{\mu_i''}{\mu_i'} \right) Q_{ii}, \quad (15.4)$$

where $\mu'_i = \partial \mu_i / \partial \eta_i$ and $\mu''_i = \partial^2 \mu_i / \partial \eta_i^2$ are the derivatives of the inverse link function. Note that the weights in (15.3) are the usual quadratic weights, namely $W_i = \mu_i'^2 / \kappa_{2i}$.

The following Table gives expressions for ξ_i for some common link functions.

Link	ξ_i
identity	0
log	$-Q_{ii}/2$
logit	$Q_{ii}(\pi_i - \frac{1}{2})$
probit	$Q_{ii}\eta_i/2$
c-log-log	$Q_{ii}(\exp(\eta_i) - 1)/2$

For binary regression models ξ_i has the same sign as η_i , though the vectors ξ and η are not co-linear in R^n . However, under conditions of approximate quadratic balance ($Q_{ii} = \text{const}$), and provided that $|\beta|$ is small, it may be shown that the bias vector b and the parameter vector β are approximately co-linear. A very rough approximation for small $|\beta|$ is

$$b \simeq p\beta/m., \quad (15.5)$$

where $m. = \sum m_i$ and $p = \dim(\beta)$. Thus bias adjustment for binary regression models has an effect on the parameter estimates approximately the same as shrinkage towards the origin by the factor $1 - p/m.$

15.2.3 Example: Lizard data (continued)

To illustrate these computations we use the linear logistic model (4.24) applied to the data in Table 4.5. The parameter estimates shown in Table 4.8 lead to the following fitted quantities:

$$\begin{aligned}\hat{\pi} &= (0.8749, 0.8977, 0.7699, 0.9558, 0.9645, 0.9120, \dots), \\ \hat{Q}_{ii} &= (0.1161, 0.1333, 0.1246, 0.1506, 0.1749, 0.1530, \dots), \\ \hat{\xi}_i &= (0.0435, 0.0530, 0.0336, 0.0687, 0.0812, 0.0630, \dots), \\ \hat{w}_i &= (2.4085, 0.8266, 1.4171, 0.5488, 0.2740, 0.9634, \dots).\end{aligned}$$

Only the first six components of the fitted vectors are shown here: these correspond to the first two rows of Table 4.5. Note that \hat{w}_i for linear logistic models is just $m_i \hat{\pi}_i(1 - \hat{\pi}_i)$.

Weighted linear regression of $\hat{\xi}$, using the same model formula, gives the bias vector $\hat{\mathbf{b}}$ shown together with $\hat{\boldsymbol{\beta}}$ in the following Table:

Parameter	Estimate	S.E.	$\hat{\mathbf{b}}$	$\hat{\boldsymbol{\beta}} - \hat{\mathbf{b}}$
μ	1.9447	0.3408	0.0436	1.9011
H	1.1300	0.2568	0.0238	1.1062
D	-0.7626	0.2112	-0.0090	-0.7536
S	-0.8473	0.3217	-0.0302	-0.8171
$T(2)$	0.2271	0.2500	-0.0009	0.2280
$T(3)$	-0.7368	0.2988	-0.0095	-0.7273

The largest biases here are about 10% of a standard error. In cases of marginal statistical significance biases of this magnitude could have a small effect on the conclusions, but they are unlikely to be of any consequence in this example. However an examination of the approximate biases is helpful here as a check on the significance of the factor S , which, though significant in the maximum-likelihood analysis, does not show up as significant in the preliminary analysis in Table 4.6 and Fig. 4.2. The bias adjustment indicated above reduces the significance of S , but not by an amount sufficient to alter the conclusions.

15.3 Computation of Bartlett adjustments

15.3.1 General theory

A simple, or fully specified, null hypothesis $H_0: \boldsymbol{\theta} = \boldsymbol{\theta}_0$ may be tested using the likelihood-ratio statistic, which is twice the difference between the maximum log likelihood and the value attained at $\boldsymbol{\theta}_0$. For generalized linear models in which the dispersion parameter is known, this difference may be written in terms of the deviance as follows:

$$\Lambda = 2l(\hat{\boldsymbol{\theta}}; Y) - 2l(\boldsymbol{\theta}_0; Y) = D(Y; \boldsymbol{\theta}_0) - D(Y; \hat{\boldsymbol{\theta}}).$$

We have assumed here for simplicity of notation that the dispersion parameter is equal to unity.

Under the usual asymptotic regularity conditions for large samples the asymptotic mean of this statistic is

$$\begin{aligned} E\{D(Y; \boldsymbol{\theta}_0) - D(Y; \hat{\boldsymbol{\theta}})\} &= p + \epsilon_p + O(n^{-2}), \\ &= p\{1 + b_p(\boldsymbol{\theta}_0)\} + O(n^{-2}), \end{aligned}$$

where $p = \dim(\boldsymbol{\theta})$ and $b(\boldsymbol{\theta})$ is known as the Bartlett adjustment factor. In fact it is possible to show that all cumulants up to any fixed order, r , are given to the same order of approximation by

$$\kappa_r\{D(Y; \boldsymbol{\theta}_0) - D(Y; \hat{\boldsymbol{\theta}})\} = (r-1)! 2^{r-1} p \{1 + b_p(\boldsymbol{\theta}_0)\}^r + O(n^{-2}). \quad (15.6)$$

For an outline proof in the single-parameter case, see Appendix C. The leading term in this expression is just the r th cumulant of the χ_p^2 distribution. From the multiplicative property of cumulants it can be seen immediately that the cumulants of the adjusted statistic

$$\Lambda' = \frac{\Lambda}{1 + b_p}$$

agree with those of the χ_p^2 distribution when terms of order $O(n^{-2})$ are ignored. Note that b_p and ϵ_p are both $O(n^{-1})$ by assumption.

Although convergence of the cumulants implies convergence in distribution provided that the asymptotic cumulants uniquely determine a distribution, the order of magnitude of the discrepancy in the cumulants is not necessarily the same as the size of the

error in the cumulative distribution function. Nevertheless it seems plausible to conclude that the distribution of the adjusted statistic is given by

$$\Lambda' \sim \chi_p^2 + O(n^{-2}),$$

and in fact this claim is correct at least in the non-lattice case. In the lattice case the error cannot be reduced below $O(n^{-1/2})$ without resorting to discontinuous approximations, which are extremely inconvenient. It is unclear in that case whether the adjustment improves the approximation or not.

For composite null hypotheses, which are more common in applications, a similar adjustment can be made. The statistic can be written in the form

$$\Lambda(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_0) = 2l(\hat{\boldsymbol{\theta}}; Y) - 2l(\hat{\boldsymbol{\theta}}_0; Y) = D(Y; \hat{\boldsymbol{\theta}}_0) - D(Y; \hat{\boldsymbol{\theta}}), \quad (15.7)$$

where $\hat{\boldsymbol{\theta}}_0$ is the estimate of the nuisance parameters under H_0 , and $\hat{\boldsymbol{\theta}}$ is the unrestricted estimate. Assuming that the hypotheses are nested, and that $q < p$ is the dimension of the parameter space under H_0 , the mean of this statistic is

$$\begin{aligned} E\{\Lambda(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_0)\} &= p + \epsilon_p + O(n^{-2}) - (q + \epsilon_q + O(n^{-2})) \\ &= p - q + (\epsilon_p - \epsilon_q) + O(n^{-2}) \\ &= (p - q)\{1 + b_{pq}(\boldsymbol{\theta})\} + O(n^{-2}). \end{aligned}$$

Thus the required adjustment factor is now

$$b_{pq} = (pb_p - qb_q)/(p - q) = (\epsilon_p - \epsilon_q)/(p - q). \quad (15.8)$$

The cumulants of the maximized likelihood-ratio statistic (15.7) obey (15.6) with p replaced by $p - q$. Hence the cumulants of the adjusted statistic agree with those of χ_{p-q}^2 apart from terms of order $O(n^{-2})$.

15.3.2 Computation of the adjustment

We focus here on computing ϵ_p as if the null hypothesis were simple. Differencing is necessary if nuisance parameters are present. All quantities are computed at $\hat{\boldsymbol{\theta}}_0$ rather than the true $\boldsymbol{\theta}$, which is usually unknown.

The calculations that follow use the general expression (29) of McCullagh and Cox (1986). The aim here is to present that expression in a more readily computable form, by exploiting special properties of generalized linear models. Final expressions are presented in matrix notation although intermediate calculations make some use of index notation.

For the present discussion it is convenient to introduce the following diagonal matrices.

$$\begin{aligned}\mathbf{D}^{(1)} &= \text{diag}\{\mu_i'\} = \text{diag}\{d\mu_i/d\eta_i\}, \\ \mathbf{D}^{(2)} &= \text{diag}\{\mu_i'' - \mu_i'^2 d \log V_i / d\mu_i\}, \\ \mathbf{W} &= \text{diag}\{\mu_i'^2/V_i\} = \mathbf{D}^{(1)} \mathbf{V}^{-1} \mathbf{D}^{(1)}.\end{aligned}$$

In addition to these, the following non-diagonal matrices arise naturally as a by-product of the weighted least squares algorithm used to compute parameter estimates.

$$\begin{aligned}\mathbf{Q} &= \mathbf{X}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T, \\ \mathbf{P} &= \mathbf{D}^{(1)} \mathbf{Q} \mathbf{D}^{(1)} \mathbf{V}^{-1}.\end{aligned}$$

Note that $\sigma^2 \mathbf{Q}$ is the asymptotic covariance matrix of $\hat{\eta}$ and $\sigma^2 \mathbf{D}^{(1)} \mathbf{Q} \mathbf{D}^{(1)}$ is the asymptotic covariance matrix of $\hat{\mu}$.

The first and second derivatives of the log likelihood with respect to the regression parameters β_r are

$$\begin{aligned}U_r &= \partial l / \partial \beta_r = \sum_i x_r^i D_i^{(1)} V_i^{-1} (Y_i - \mu_i), \\ U_{rs} &= \partial^2 l / \partial \beta_r \partial \beta_s = \sum_i x_r^i x_s^i D_i^{(2)} V_i^{-1} (Y_i - \mu_i) - x_r^i x_s^i W_i,\end{aligned}$$

where x_r^i are the components of the model matrix \mathbf{X} . The Fisher information matrix is $\mathbf{X}^T \mathbf{W} \mathbf{X}$, which we write as $\kappa_{r,s}$, with inverse $\kappa^{r,s}$.

Evidently, for all generalized linear models, the log likelihood derivatives are linear functions of Y , a property that simplifies much of the subsequent calculations. A key step in deriving a simple expression for the Bartlett factor involves working with the residual second derivative matrix rather than U_{rs} . The covariance of U_{rs} and U_t is a three-way array whose components are

$$\kappa_{rs,t} = \sum_i x_r^i x_s^i D_i^{(2)} V_i^{-1} D_i^{(1)} x_t^i.$$

Since only arithmetic multiplication is involved, the order of terms in the above sum is immaterial. It is a remarkable property of generalized linear models that all such ‘mixed’ cumulants are symmetric under index permutation. In other words, for generalized linear models, but not in general, $\kappa_{rs,t} = \kappa_{rt,s} = \kappa_{st,r}$ and so on.

The residual matrix of second derivatives after linear regression on U_r is

$$V_{rs} = \sum_i x_r^i x_s^i D_i^{(2)} V_i^{-1} \{ \delta_{ij} - P_{ij} \} (Y_j - \mu_j).$$

For purposes of computation we may write

$$V_{rs} \simeq \sum_i x_r^i x_s^i D_i^{(2)} V_i^{-1} (Y_i - \hat{\mu}_i),$$

but the former expression is simpler for theoretical calculations.

The Bartlett factor can now be given as a linear combination of six invariant functions of the joint cumulants of U_r and V_{rs} . The invariant fourth cumulant of U_r is

$$\begin{aligned} \rho_4 &= \sum_i x_r^i x_s^i x_t^i x_u^i \mu_i'^4 V_i^{-4} \kappa_{4i} \kappa^{r,s} \kappa^{t,u}, \\ &= \sum_i (P_{ii})^2 \rho_{4i}, \end{aligned} \tag{a}$$

where P_{ii} are the diagonal elements of the asymmetric projection matrix \mathbf{P} , and $\rho_{4i} = \kappa_{4i}/\kappa_{2i}^2$ is the usual standardized fourth cumulant of Y_i . The two quadratic skewness scalars are

$$\begin{aligned} \rho_{13}^2 &= \sum_{ij} x_r^i Q_{ii} \mu_i'^3 V_i^{-3} \kappa_{3i} \kappa^{r,s} x_r^j Q_{jj} \mu_j'^3 V_j^{-3} \kappa_{3j} \\ &= \sum_{ij} (P_{ii} \kappa_{3i} / \kappa_{2i}) V_i^{-1} P_{ij} (P_{jj} \kappa_{3j} / \kappa_{2j}); \end{aligned} \tag{b}$$

$$\rho_{23}^2 = \sum_{ij} (V_i^{-1} P_{ij})^3 \kappa_{3i} \kappa_{3j}. \tag{c}$$

These scalars can be computed easily using simple matrix operations.

Similar calculations show that the two scalar measures of the variability of V_{rs} can be simplified as follows.

$$\nu_{rs,tu} \kappa^{r,s} \kappa^{t,u} = \mathbf{q}^T \mathbf{D}^{(2)} \mathbf{V}^{-1} (\mathbf{I} - \mathbf{P}) \mathbf{D}^{(2)} \mathbf{q}, \quad (d)$$

$$\nu_{rs,tu} \kappa^{r,t} \kappa^{s,u} = \sum_{ij} Q_{ij} Q_{ij} [\mathbf{D}^{(2)} \mathbf{V}^{-1} (\mathbf{I} - \mathbf{P}) \mathbf{D}^{(2)}]_{ij}, \quad (e)$$

where \mathbf{q} is a vector with components Q_{ii} . Finally we have one further scalar

$$\nu_{r,s,tu} \kappa^{r,s} \kappa^{t,u} = \mathbf{q}^T \mathbf{D}^{(2)} \mathbf{V}^{-1} (\mathbf{I} - \mathbf{P}) \mathbf{q}^*, \quad (f)$$

where $q_j^* = q_j W_j \kappa_{3j} / \kappa_{2j}$. The notation on the left of the preceding three equations is taken from McCullagh and Cox (1986).

Although the tensor expressions on the left of (d)–(f) are algebraically more appealing than the matrix formulae, the latter expressions have the advantage for numerical purposes that they use only simple operations on matrices and vectors. Numerical computation involving higher-order arrays is thereby avoided.

In terms of these scalars the correction may be written as the linear combination

$$\epsilon_p = -\frac{1}{4}(a) + \frac{1}{4}(b) + \frac{1}{6}(c) - \frac{1}{4}(d) + \frac{1}{2}(e) - \frac{1}{2}(f). \quad (15.9)$$

If the canonical link is used only the first three of these terms contribute.

15.3.3 Example: exponential regression model

Suppose that Y_1, \dots, Y_n are independent exponential random variables with $\eta_i = \log \mu_i$ satisfying the linear model $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$. We find that the required matrices are as follows:

$$\begin{aligned} \mathbf{D}^{(1)} &= \text{diag}\{\mu_i\}, & \mathbf{D}^{(2)} &= \text{diag}\{-\mu_i\}, & \mathbf{W} &= \mathbf{I}, \\ \mathbf{Q} &= \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T, & \mathbf{P} &= \text{diag}\{\mu_i\} \mathbf{Q} \text{diag}\{\mu_i^{-1}\}, \end{aligned}$$

so that $P_{ii} = Q_{ii}$ even though \mathbf{P} is not symmetrical. These properties permit simplification of (15.9) giving

$$\epsilon_p = \frac{1}{6} \sum Q_{ij}^3 - \frac{1}{4} \mathbf{q}^T (\mathbf{I} - \mathbf{Q}) \mathbf{q}. \quad (15.10)$$

In this particular case, since the model is of the translation type, ϵ_p does not depend on the value of the parameter.

To take a simple numerical example, consider the data on survival times for 17 leukaemia patients given by Feigl and Zelen (1965). The data are discussed by Cox and Snell (1981, pp.148–150), who consider the model

$$\log \mu_i = \beta_0 + \beta_1 x_i \quad (15.11)$$

in which μ_i is the expected survival time, and x_i is the logarithm of the initial white blood cell count. The likelihood-ratio statistic for testing the hypothesis $H_0: \beta_1 = 0$ comes to 6.826 on one degree of freedom yielding a p -value of 0.89%. Here we compute the Bartlett-adjusted statistic as a check on the adequacy of the χ^2_1 approximation.

Since in this case

$$Q_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum(x_i - \bar{x})^2},$$

it follows that

$$\sum_{ij} Q_{ij}^3 = \frac{4}{n} + \frac{\left(\sum(x_i - \bar{x})^3\right)^2}{\left(\sum(x_i - \bar{x})^2\right)^3}.$$

For the leukaemia data this gives

$$\epsilon_2 = (4/17 + 0.00003849)/6 - 0.0688/4 = 0.0220.$$

Similar calculations under H_0 give $\epsilon_1 = 1/(6n) = 0.00980$. Thus $b_{21} = 0.0122$ giving a corrected statistic of 6.744. The p -value is increased to 0.94%—a trivial increase in this case. The usual χ^2 approximation appears to be quite accurate here even though the sample size is not especially large.

On the other hand if we were interested in testing the adequacy of the model (15.11) with no specific alternative in mind, the likelihood-ratio statistic is equal to 19.46 on 15 degrees of freedom. For the Bartlett adjustment we find $\epsilon_{17} = 17/6$ giving

$$b_{17,2} = (17/6 - 0.0220)/15 = 0.187.$$

Thus the adjusted statistic is $19.46/1.187 = 16.39$. The p -value is thereby increased from 19.4% to 35.6%. The 19% reduction in the value of the statistic is substantial and could in principle weaken the force of the conclusions, although that has not happened in this case. For example if the unadjusted statistic corresponds to a p -value of 1%, the p -value for the adjusted statistic is 4%.

15.4 Generalized additive models

In choosing the set of terms to be included in the linear predictor of a generalized linear model we make the prior assumption that some subset of the terms chosen will give an adequate picture of the response surface generated by the covariates. If the actual shape of the surface is not expressible as a function of the terms and link function chosen then the fitted values will deviate systematically from the actual response surface. Model-checking techniques described in Chapter 12 can be used to detect this state of affairs, and perhaps to remedy the misfit. An alternative method, due to Hastie and Tibshirani (1986, 1987ab), is to fit *generalized additive models*, which, for continuous covariates x , replace the linear predictor $\eta = \sum_j x_j \beta_j$ by the additive model

$$\eta = \alpha + \sum_j f_j(x_j)$$

in which $f_j(x_j)$ are smooth functions estimated from the data. The model remains additive with respect to the covariates, but it is no longer linear in them. The functions $f_j(\cdot)$ are identifiable up to an arbitrary constant, much like the levels of a factor.

15.4.1 Algorithms for fitting

We consider first the fitting procedure for a single covariate x , using a generalized additive model with link function $g(\cdot)$ and variance function $V(\cdot)$. One method associates with each point (y_i, x_i) a set of neighbours on the x -axis, and estimates a value of the response function at x_i for each i by applying the standard GLM algorithm with linear predictor $\alpha + \beta x_i$ in that neighbourhood. Then $\hat{f}(x_i)$ is the fitted value of $\eta - \alpha$ at x_i , i.e. we use linear interpolation within each neighbourhood. This algorithm, also called the local-scoring algorithm, can be thought of as using a weighted running-lines smoother on the adjusted dependent variable z . The algorithm is efficient because the running-lines smoother can be updated easily as we pass from one neighbourhood to the next. Neighbourhoods are usually defined to include a certain fraction, or span, of the points. Commonly, spans of 40–50% are used, with appropriate contraction at the ends. The fit obtained depends on the span used: the shorter the span the rougher the fit.

When more than one covariate is involved the algorithm acquires an additional loop in which each $f_j(\cdot)$ is fitted using the current estimates of the remaining functions. This is known as the back-fitting algorithm, and takes the form

for $j = 1, \dots, p$

form partial residual $r_j = z - \hat{\eta} + \hat{f}_j(x_j)$

smooth r_j with GLM weights W to update $\hat{f}_j(x_j)$

repeat.

The full algorithm may be set out as follows:

Initialize: $f_j^{(0)}(x_j) = 0$, $\hat{\alpha}_0 = g(\bar{y})$

for $i = 0, 1, \dots$

form current estimates of $\hat{\eta}^{(i)}(x_j)$, $\hat{\mu}^{(i)}$, $x^{(i)}$ *and* $W^{(i)}$

perform back-fitting algorithm to obtain $\hat{\alpha}^{(i+1)}$ *and* $\hat{f}_j^{(i+1)}(x_j)$

repeat until deviance stabilizes.

This algorithm depends on the choice of span or fraction of points used for the local linear fit. At one extreme, if the span is 1, the algorithm produces the generalized linear fit. At the other extreme, if the span is equal to $1/n$ and if all the x -values are distinct then $\hat{\mu}_i = y_i$, and no smoothing occurs.

15.4.2 Smoothing methods

In the algorithm just described the running-lines smoother may be replaced by any other smoother, which may in turn be linear or non-linear. Running-lines and cubic-spline smoothers are linear in y , while a running-median smoother is non-linear. Another non-linear method involves maximizing the local likelihood for each neighbourhood. In practice, however, this usually gives results very close to that produced by the linear local scoring method.

The choice of span can be made data-dependent by choosing it to minimize the cross-validation deviance.

The effective number of parameters associated with an estimated smooth function can be calculated from the formula

$$\text{tr}(2\mathbf{S} - \mathbf{S}^T \mathbf{W} \mathbf{S} \mathbf{W}^{-1})$$

in which \mathbf{S} is the smoothing matrix applied to the vector \mathbf{z} to produce $f(\cdot)$, and \mathbf{W} is the weight matrix. For the running-lines smoother this formula simplifies to $\text{tr}(\mathbf{S})$.

15.4.3 Conclusions

Generalized additive models can be used either as a descriptive tool for expressing the joint effect of several explanatory variables as the sum of functions of them individually, or as an exploratory device to suggest a suitable class of transformations of covariates to be included in a generalized linear model. The restriction to additive terms can be relaxed by including product terms of the form $f_{12}(x_1 x_2)$ in addition to $f_1(x_1)$ and $f_2(x_2)$. However not all functions of two variables are expressible in the form

$$f_1(x_1) + f_2(x_2) + f_{12}(x_1 x_2).$$

Generalized partially additive models, in which some covariates enter linearly and others additively, may eventually prove to be the most useful application of these techniques.

15.5 Bibliographic notes

Bartlett (1937, 1954) gave explicit correction factors for a number of likelihood-ratio test statistics, including a number of test statistics that arise in multivariate analysis. Similar expressions for log-linear models were given by Williams (1976), and, for generalized linear models, by Cordeiro (1983, 1987). See also Ross (1987).

The matrix formulae in section 15.3 appear to be new. An alternative scheme for computing Bartlett adjustments is described by Barndorff-Nielsen and Blaesild (1986).

15.6 Further results and exercises 15

15.1 Compute the angle in R^p between the parameter vector $\hat{\beta}$ and the bias vector $\hat{\mathbf{b}}$ for the main-effects model fitted to the lizard data in section 15.2.3. Use the Fisher information as the inner product matrix. Comment briefly on the adequacy of the approximation (15.5).

15.2 Show that the matrix \mathbf{P} defined in section 15.3.2 is a projection matrix. Describe the range space of \mathbf{P} , i.e. the set of vectors \mathbf{x} such that $\mathbf{Px} = \mathbf{x}$. Explain how this space depends on the value of the parameter vector.

15.3 Derive expression (15.10) for the Bartlett adjustment for exponential regression models from the general expression (15.9). Show that the second term in (15.10) vanishes if \mathbf{X} is the design matrix for a one-way layout.

15.4 Suppose $Y_i \sim \sigma_i^2 \chi_{f_i}^2 / f_i$, independently for $i = 1, \dots, k$. Specify the null hypothesis $H_0: \sigma_i^2 = \sigma^2$ and the unrestricted alternative as generalized linear models. Use (15.9) or (15.10) to compute the Bartlett adjustment factor. [Bartlett, 1937].

15.5 Fit the log-linear model (15.11) to the leukaemia data and check the calculations given in section 15.3.3.

APPENDIX A

Elementary likelihood theory

Scalar parameter

This appendix contains a concise summary, without proofs and omitting esoteric details of regularity conditions, of the more important properties of likelihood functions, derivatives and estimates, that are used throughout the book.

Definition: The log likelihood is the logarithm of the joint probability or probability density function of the data, denoted by

$$l(\theta; y) = \log f_Y(y; \theta).$$

If Y is a vector having n independent components, the log likelihood is a sum of n independent terms

$$l(\theta; \mathbf{y}) = \sum \log f_{Y_i}(y_i; \theta).$$

This representation is often used, at least implicitly, in the derivation of asymptotic results for large n .

Derivatives: Under mild regularity conditions the log-likelihood derivatives satisfy the following moment identities:

$$\begin{aligned} E_\theta\left(\frac{\partial l}{\partial \theta}\right) &= 0 \\ E_\theta\left(\frac{\partial^2 l}{\partial \theta^2}\right) + \text{var}_\theta\left(\frac{\partial l}{\partial \theta}\right) &= 0. \end{aligned} \tag{A.1}$$

The notation above is chosen to emphasize the fact that the twin processes of differentiation and averaging take place at the same value of θ . These relations are obtained by differentiating with respect to θ the identity

$$\int f_Y(y; \theta) dy \equiv 1.$$

The necessary regularity conditions are those required to justify interchanging the order of differentiation with respect to the parameter and integration over the sample space. In particular, the sample space must be the same for all values of the parameter, or at least for all θ in an open neighbourhood of the true parameter point.

Further differentiation with respect to θ gives higher-order identities, sometimes called the Bartlett identities after Bartlett (1954). The third-order identity is

$$E_\theta \left(\frac{\partial^3 l}{\partial \theta^3} \right) + 3 \text{cov}_\theta \left(\frac{\partial^2 l}{\partial \theta^2}, \frac{\partial l}{\partial \theta} \right) + E_\theta \left(\frac{\partial l}{\partial \theta} \right)^3 = 0. \quad (A.2)$$

These results, connecting moments of log-likelihood derivatives, are exact for all sample sizes provided, of course, that all the necessary moments are finite.

Terminology: The log-likelihood derivative $U(\theta; y) = \partial l / \partial \theta$ is sometimes called the score statistic. Its variance

$$i(\theta) = \text{var}_\theta \left(\frac{\partial l}{\partial \theta} \right) = -E_\theta \left(\frac{\partial^2 l}{\partial \theta^2} \right)$$

is called the *Fisher information* for θ and plays an important role in much of what follows.

If the components of Y are independent we may write

$$\begin{aligned} U(\theta; y) &= \sum \frac{\partial \log f_{Y_i}(y_i; \theta)}{\partial \theta}, \\ i(\theta) &= \sum E \left(\frac{\partial \log f_{Y_i}(y_i; \theta)}{\partial \theta} \right)^2, \end{aligned}$$

showing explicitly that the score statistic is a sum of n independent contributions and that the Fisher information based on the vector Y is the sum of the Fisher informations from the components.

Asymptotic results: The following results hold under further regularity conditions related to the behaviour of the sequence of observations for large n , or to be more precise, as the amount of information, $i(\theta)$, becomes large. In particular, the first derivative suitably normalized converges in distribution to a standard Normal random variable. Thus

$$i(\theta)^{-1/2} \left(\frac{\partial l}{\partial \theta} \right) \sim N(0, 1) + O_p(n^{-1/2}) \quad (A.3)$$

provided that the assumed model is correct and that the derivative is computed at the true parameter point.

The error term is governed more by the magnitude of $i(\theta)$ than by the number of components of Y . Most commonly, however, $i(\theta)$ is roughly proportional to n and it is then immaterial whether we normalize by n or by $i(\theta)$.

Maximum-likelihood estimation: Ordinarily the likelihood function has a single maximum in the interior of the parameter space. The maximum-likelihood estimate, denoted by $\hat{\theta}$, is then obtained as the solution of the equation $U(\hat{\theta}; y) = 0$. For large $i(\theta)$, the distribution of $\hat{\theta}$ is often adequately approximated by

$$\hat{\theta} - \theta \sim N(0, i(\theta)^{-1}), \quad (A.4)$$

assuming, as always, that the model is correct. Higher-order approximations based on Edgeworth series are given by McCullagh (1987, p.210).

Occasionally, however, it may be found that the maximum occurs at a boundary point of the parameter space, which may be finite or infinite. The above approximation is then inappropriate. Approximate confidence limits can be obtained directly from the likelihood function or the likelihood-ratio statistic.

Likelihood-ratio statistics: For large n , the log likelihood at $\hat{\theta}$ differs from the log likelihood at the true parameter point by a random amount whose approximate distribution is given by

$$2l(\hat{\theta}; Y) - 2l(\theta; Y) \sim \chi_1^2 + O(n^{-1}). \quad (A.5)$$

This approximation is often quite accurate for small values of n even when the Normal approximation (A.4) is unsatisfactory. The set of all θ -values satisfying

$$2l(\hat{\theta}; y) - 2l(\theta, y) \leq \chi_{1,\alpha}^2$$

is an approximate $100(1 - \alpha)\%$ confidence set for the parameter and is usually more accurate in terms of coverage probability than intervals based on (A.4).

The above approximations can be improved further using methods given in Appendix C.

Vector parameter

For vector-valued parameters, the same results apply with suitable minor changes of notation. The score statistic is the gradient vector of the log likelihood at θ and the Fisher information $i(\theta)$ is now to be interpreted as a matrix. With obvious modifications, identities (A.1) apply to the vector case. The vector version of (A.2) is given by McCullagh (1987, p.202).

The asymptotic results (A.3) and (A.4) extend readily to vector-valued parameters provided that the limit $i(\theta) \rightarrow \infty$ is understood to apply to the eigenvalues of the information matrix and not to the components. An important regularity condition is that $i(\theta)$ have constant rank for all θ in the region of interest.

Nuisance parameters: Suppose that θ is partitioned into two components $\theta = (\psi, \lambda)$, both of which may be vector-valued. The first component is to be regarded as the parameter of interest. The joint Fisher information matrix for θ may then be partitioned as follows:

$$i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix}.$$

Its inverse is denoted by

$$i(\theta)^{-1} = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix}$$

so that, from the formulae for the inverse of a partitioned matrix,

$$\{i^{\psi\psi}\}^{-1} = i_{\psi\psi} - i_{\psi\lambda} i_{\lambda\lambda}^{-1} i_{\lambda\psi} \quad (A.6)$$

is the approximate inverse covariance matrix of $\hat{\psi}$. Moreover, if $\hat{\lambda}_\psi$ is the maximum likelihood estimate of λ for fixed ψ , the gradient vector of the log likelihood, calculated at $(\psi, \hat{\lambda}_\psi)$, has approximate covariance matrix $\{i^{\psi\psi}\}^{-1}$ rather than $i_{\psi\psi}$. By contrast, the derivative with respect to ψ at (ψ, λ) has exact covariance matrix $i_{\psi\psi}$. Thus it is reasonable to regard (A.6) rather than $i_{\psi\psi}$ as the Fisher information for ψ when λ is unknown. Note that two matrix inversions are required in order to produce the expression in (A.6).

Likelihood-ratio statistics: For large n , the log likelihood at $\hat{\theta}$ differs from the log likelihood at the true parameter point by a random amount whose approximate distribution is given by

$$2l(\hat{\theta}; Y) - 2l(\theta; Y) \sim \chi_p^2 + O(n^{-1}), \quad (A.7)$$

where p is the dimension of θ or the rank of $i(\theta)$. If there are nuisance parameters in the model, the maximized likelihood-ratio statistic has approximate distribution

$$2l(\hat{\psi}, \hat{\lambda}; Y) - 2l(\psi, \hat{\lambda}_\psi; Y) \sim \chi^2_{p-q} + O(n^{-1}), \quad (A.8)$$

where $p - q$ is the dimension of ψ or the rank of $i^{\psi\psi}$ in (A.6). These approximations are often quite accurate for small values of n even when Normal approximations for parameter estimates are unsatisfactory. The set of all ψ -values satisfying

$$2l(\hat{\psi}, \hat{\lambda}; y) - 2l(\psi, \hat{\lambda}_\psi, y) \leq \chi^2_{p-q, \alpha}$$

is an approximate $100(1 - \alpha)\%$ confidence set for ψ .

APPENDIX B

Edgeworth series

Suppose that Y_1, \dots, Y_n are independent and identically distributed random variables having finite cumulants, $\kappa_1 \equiv \mu$, $\kappa_2 \equiv \sigma^2$, κ_3, κ_4 , up to order four. Define the standardized sum

$$X_n = \frac{Y_1 + \dots + Y_n - n\mu}{\sigma\sqrt{n}}.$$

Denote by $F_n(x)$ the probability $\text{pr}(X_n \leq x)$. By the central limit theorem, X_n is asymptotically standard normal and $F_n(x) \rightarrow \Phi(x)$ as $n \rightarrow \infty$. If the distribution of Y has a continuous component, then $F_n(x)$ may be approximated more accurately for large n by an Edgeworth series as follows:

$$\begin{aligned} E_n(x) = \Phi(x) - \phi(x) & \left\{ \rho_3(x^2 - 1)/(6n^{1/2}) + \rho_4(x^3 - 3x)/(24n) \right. \\ & \left. + \rho_3^2(x^5 - 10x^3 + 15x)/(72n) \right\} \quad (B.1) \end{aligned}$$

where $\rho_3 = \kappa_3/\kappa_2^{3/2}$ and $\rho_4 = \kappa_4/\kappa_2^2$ are the standardized cumulants of Y . The difference $F_n(x) - E_n(x)$ is $o(n^{-1})$ uniformly in x on bounded intervals.

In the case of lattice distributions, this expansion is incorrect because $F_n(x)$ is discontinuous with jumps of order $O(n^{-1/2})$ at the possible values of X_n . The Edgeworth approximation is continuous and hence must involve an error of order $O(n^{-1/2})$ near the discontinuities of $F_n(x)$. However, the Edgeworth series can be adjusted to accommodate these discontinuities in $F_n(x)$ as follows. Suppose that the possible values of Y_i are the integers $0, 1, 2, \dots$. Define the continuity-corrected abscissa and a 'precision

adjustment' or Sheppard correction as follows:

$$z = \frac{y_1 + \dots + y_n - n\mu + \frac{1}{2}}{\sigma\sqrt{n}} \quad (B.2)$$

$$\tau = 1 + \frac{1}{24n\sigma^2}.$$

Then the Edgeworth series with the usual two correction terms may be written

$$F_n(z) = E_n(\tau z) + o(n^{-1}). \quad (B.3)$$

This approximation is valid only when computed at the 'continuity-corrected' points as defined by (B.2). The distribution function is constant over intervals of the form $[z - \frac{1}{2\sigma\sqrt{n}}, z + \frac{1}{2\sigma\sqrt{n}}]$.

Note that the correction terms in (B.3) are identical to the correction terms in (B.1). The only difference is the correction for continuity and the adjustment of the argument. The continuity correction has an effect of order $O(n^{-1/2})$ and the Sheppard correction has an effect of order $O(n^{-1})$.

The discrete Edgeworth approximation may be used for the binomial distribution where $Y \sim B(m, \pi)$, provided that m is sufficiently large. The relevant coefficients are

$$z = (y - m\pi + \frac{1}{2})/\sqrt{m\pi(1 - \pi)}$$

$$\tau = 1 + 1/\{24m\pi(1 - \pi)\}$$

$$\rho_3 = (1 - 2\pi)/\sqrt{m\pi(1 - \pi)}$$

$$\rho_4 = \{1 - 6\pi(1 - \pi)\}/\{m\pi(1 - \pi)\}.$$

The sample size, n , is built into these coefficients through the binomial index, m . Thus we may take $n = 1$ in (B.1) and (B.3). In this case, the approximation seems to be quite accurate if $m\pi(1 - \pi)$ exceeds 2.0.

Approximation (B.3) is a simplified version of a series expansion given by Esseen (1945), who gives the expansion to higher order than that considered here.

APPENDIX C

Likelihood-ratio statistics

The deviance or deviance difference is just a log-likelihood ratio statistic. In this appendix, we derive the approximate distribution of the log likelihood-ratio statistic for testing a simple null hypothesis concerning a scalar parameter. The corresponding derivations when there are several parameters of interest or when there are nuisance parameters follow similar lines but are considerably more complicated than the proofs presented here.

Suppose that the log likelihood for θ based on data y can be written in the exponential family form

$$l(\theta; y) = n\{t\theta - K(\theta)\},$$

where $t \equiv t(y)$ is the sufficient statistic and θ is called the canonical parameter. The cumulants of the random variable $T = t(Y)$ are

$$\kappa_r(T) = K^{(r)}(\theta)/n^{r-1}.$$

The maximum-likelihood estimate of θ satisfies

$$K'(\hat{\theta}) = t \quad \text{or} \quad \hat{\theta} = g(t), \text{ say.}$$

Hence the log likelihood-ratio statistic for testing $H_0 : \theta = \theta_0$ is

$$W^2 = 2l(\hat{\theta}) - 2l(\theta_0) = 2n\{tg(t) - tg(\mu_0) + h(t) - h(\mu_0)\} \quad (C.1)$$

where $h(\cdot) = K(g(\cdot))$ and $\mu_0 = K'(\theta_0)$ is the null mean of T . Thus $h(t) = K(\hat{\theta})$ and $h(\mu) = K(\theta)$.

Under H_0 , the statistic W^2 has an approximate χ_1^2 distribution and hence it is reasonable to expect that the signed version

$$W = \pm\{2l(\hat{\theta}) - 2l(\theta_0)\}^{1/2}$$

might have an approximate Normal distribution. The sign of W is taken to be the same as that of $t - \mu_0$ and, in fact, W is a monotone increasing function of $t - \mu_0$. To the crudest first-order of approximation, W is the standardized version of T , namely

$$X = (T - \mu_0)/\kappa_2^{1/2}.$$

If we expand W as a power series in X and keep together terms that are of the same asymptotic order in n , we find after a little effort that

$$W = X - \frac{1}{6}\rho_3 X^2 + \frac{1}{72}(8\rho_3^2 - 3\rho_4)X^3 + O_p(n^{-3/2}), \quad (C.2)$$

where ρ_3 and ρ_4 are the standardized cumulants of T or the unstandardized cumulants of X . Note that $\rho_3 = O(n^{-1/2})$ and $\rho_4 = O(n^{-1})$.

It is readily verified that the first two moments of W are

$$\begin{aligned} E(W) &= \rho_3/6 + O(n^{-3/2}) \\ \text{var}(W) &= 1 + (14\rho_3^2 - 9\rho_4)/36 + O(n^{-2}) \end{aligned} \quad (C.3)$$

If we define the adjusted statistic or re-standardized statistic

$$W' = \{W + \frac{1}{6}\rho_3\}\{1 + (9\rho_4 - 14\rho_3^2)/72\},$$

it is easily checked that, with error $O_p(n^{-3/2})$,

$$W' = X - \frac{1}{6}\rho_3(X^2 - 1) - \frac{1}{24}\rho_4(X^3 - 3X) + \frac{1}{36}\rho_3^2(4X^3 - 7X).$$

This series can be recognized as the inverse Cornish-Fisher expansion or the polynomial normalizing transformation. See, for example, Kendall and Stuart (1977, (6.54)) or McCullagh (1987, section 5.7 and Exercise 5.15). Provided that X has an Edgeworth expansion, it follows that

$$W' \sim N(0, 1) + O(n^{-3/2})$$

in the Edgeworth sense. This conclusion also follows from the observation that W , as given in (C.2), has third and fourth cumulants of orders $O(n^{-3/2})$ and $O(n^{-2})$ respectively.

In the discrete case, the support points of the distribution of T are usually equally spaced and the discrete Edgeworth approximation given in Appendix A may be used to approximate the distribution of T . The support points of W are only approximately equally spaced. It appears therefore, that the normal approximation with continuity correction for the distribution of W' has an error of order $O(n^{-1})$. The Sheppard correction does not appear to eliminate the $O(n^{-1})$ error term entirely, though it may reduce it substantially.

As a corollary in the continuous case, it follows directly that

$$W^2 \sim (1 + b)\chi_1^2 + O(n^{-3/2})$$

where $1 + b$ is the sum of the variance and the squared bias of W . The adjustment

$$b/n = (5\rho_3^2 - 3\rho_4)/12 \quad (C.4)$$

is known as the Bartlett adjustment factor. Its use greatly improves the accuracy of the chi-squared approximation for the likelihood-ratio statistic. In both the discrete and the continuous case, the cumulants of $W^2/(1 + b)$ differ from those of χ_1^2 by terms of order $O(n^{-2})$.

Although the derivations in general are considerably more complicated than that presented here, these results can be extended in the following ways:

1. to models not in the exponential family provided that the usual regularity conditions are satisfied.
2. to multi-parameter problems.
3. to problems involving nuisance parameters.

Computational details for generalized linear models are discussed in section 15.2. For proofs and additional information, see Lawley (1956), Barndorff-Nielsen and Cox (1984), McCullagh (1984a, 1987 Chapter 7) and McCullagh and Cox (1986).

References

- Abramowitz, M. and Stegun, I.A. (1970) *Handbook of Mathematical Functions*. Dover, New York.
- Adena, M.A. and Wilson, S.R. (1982) *Generalized Linear Models in Epidemiological Research*. INTSTAT, Sydney.
- Agresti, A. (1984) *Analysis of Ordinal Categorical Data*. J. Wiley & Sons, New York.
- Aitkin, M. (1983) *Linear Statistical Analysis of Discrete Data*. J. Wiley & Sons, New York.
- Aitkin, M. (1987) Modelling variance heterogeneity in Normal regression using GLIM. *Appl. Statist.* **36**, 332–9.
- Aitkin, M. and Clayton, D.G. (1980) The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Appl. Statist.* **29**, 156–63.
- Akaike, H. (1969) Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* **21**, 243–7.
- Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory*. (eds. B.N. Petrov and F. Czaki). Akadémiai Kiadó, Budapest, pp. 267–81.
- Andersen, E.B. (1973) *Conditional Inference and Models for Measuring*. Mentalhygiejinsk Forlag, Copenhagen.
- Anderson, D.A. (1981) The circular structural model. *J. R. Statist. Soc. B* **43**, 131–41.
- Anderson, D.A. and Aitkin, M. (1985) Variance component models with binary response. *J. R. Statist. Soc. B* **47**, 203–10.
- Andrews, D.F. and Herzberg, A.M. (1985) *Data*. Springer-Verlag, New York.
- Andrews, D.F. and Pregibon, D. (1978) Finding the outliers that matter. *J. R. Statist. Soc. B* **40**, 85–93.
- Angell, I. and Barber, J. (1977) An algorithm for fitting circles and ellipses to megalithic stone rings. *Science and Archaeology* **20**, 11–16.
- Anscombe, F.J. (1949) The statistical analysis of insect counts based on the negative binomial distribution. *Biometrics* **15**, 165–73.
- Anscombe, F.J. (1953) Contribution to the discussion of H. Hotelling's paper. *J. R. Statist. Soc. B* **15**, 229–30.

- Anscombe, F.J. (1961) Examination of residuals. *Proc. Fourth Berkeley Symposium* **1**, 1-36.
- Anscombe, F.J. (1981) *Computing in Statistical Science through APL*. Springer-Verlag, New York.
- Anscombe, F.J. and Tukey, J.W. (1963) The examination and analysis of residuals. *Technometrics* **5**, 141-60.
- Armitage, P. (1955) Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375-86.
- Armitage, P. (1957) Studies in the variability of pock counts. *J. Hygiene Camb.* **55**, 564-81.
- Armitage, P. (1971) *Statistical Methods in Medical Research*. Blackwell, Oxford.
- Ashford, J.R. (1959) An approach to the analysis of data for semi-quantal responses in biological assay. *Biometrics* **15**, 573-81.
- Ashford, J.R. and Sowden, R.R. (1970) Multi-variate probit analysis. *Biometrics* **26**, 535-46.
- Ashton, W.D. (1972) *The Logit Transformation with Special Reference to its Uses in Bioassay*. Griffin, London.
- Atkinson, A.C. (1981a) Two graphical displays for outlying and influential observations in regression. *Biometrika* **68**, 13-20.
- Atkinson, A.C. (1981b) Likelihood ratios, posterior odds and information criteria. *J. Econometrics* **16**, 15-20.
- Atkinson, A.C. (1982) Regression diagnostics, transformations and constructed variables (with discussion). *J. R. Statist. Soc. B* **44**, 1-36.
- Atkinson, A.C. (1985) *Plots, Transformations and Regression*. Oxford, Clarendon Press.
- Atkinson, A.C. (1986) Masking unmasked. *Biometrika* **73**, 533-41.
- Babington-Smith, B. (1950) Discussion of Professor Ross's paper. *J. R. Statist. Soc. B* **12**, 53-56.
- Bain, L.J. and Engelhardt, M. (1975) A two-moment chi-square approximation for the statistic $\log(\bar{x}/\tilde{x})$. *J. Am. Statist. Assoc.* **70**, 948-50.
- Baker, R.J. and Nelder, J.A. (1978) *The GLIM System*. Release 3, *Generalized Linear Interactive Modelling*. Numerical Algorithms Group, Oxford.
- Barlow, R.E. and Proschan, F. (1965) *Mathematical Theory of Reliability*. J. Wiley & Sons, New York.
- Barlow, R.E. and Proschan, F. (1975) *Statistical Theory of Reliability and Life Testing*. New York: Holt, Rinehart and Winston.
- Barndorff-Nielsen, O.E. (1978) *Information and Exponential Families in Statistical Theory*. Chichester: Wiley.
- Barndorff-Nielsen, O.E. (1985) Confidence limits from $c|\hat{J}|^{1/2}\bar{L}$ in the single parameter case. *Scand. J. Statist.* **12**, 83-7.
- Barndorff-Nielsen, O.E. and Blaesild, P. (1986) A note on the calculation of Bartlett adjustments. *J. R. Statist. Soc. B* **48**, 353-8.
- Barndorff-Nielsen, O.E. and Cox, D.R. (1979) Edgeworth and saddle-point approximations with statistical applications (with discussion).

- J. R. Statist. Soc. B* **41**, 279–312.
- Barndorff-Nielsen, O.E. and Cox, D.R. (1984) Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *J. R. Statist. Soc. B* **46**, 483–95.
- Barnwal, R.K. and Paul, S.R. (1988) Analysis of one-way layout of count data with negative binomial variation. *Biometrika* **75**, 215–22.
- Bartlett, M.S. (1936a) Some notes on insecticide tests in the laboratory and in the field. *J. R. Statist. Soc. Suppl.* **3**, 185–94.
- Bartlett, M.S. (1936b) The information available in small samples *Proc. Camb. Phil. Soc.* **32**, 560–6.
- Bartlett, M.S. (1937) Properties of sufficiency and statistical tests. *Proc. Roy. Soc. A* **160**, 268–82.
- Bartlett, M.S. (1954) A note on some multiplying factors for various χ^2 approximations. *J. R. Statist. Soc. B* **16**, 296–8.
- Bartlett, M.S. and Kendall, M.G. (1946) The statistical analysis of variance-heterogeneity and the logarithmic transformation. *J. R. Statist. Soc. Suppl.* **8**, 123–38.
- Baxter, L.A., Coutts, S.M. and Ross, G.A.F. (1980) Applications of linear models in motor insurance, in *Proceedings of the 21st International Congress of Actuaries*. Zurich. pp. 11–29.
- Beale, E.M.L. (1970) Note on procedures for variable selection in multiple regression. *Technometrics* **12**, 909–14.
- Beaton, A.E. (1964) The use of special matrix operators in statistical calculus. *Research Bulletin RB-64-51, Educational Testing Service*. Princeton, New Jersey, USA.
- Belsley, D.A., Kuh, E. and Welsch, R.E. (1980) *Regression Diagnostics*. J. Wiley & Sons, New York.
- Benzécri, J.-P. (1976) *L'Analyse des Données: 1 La Taxonomie; 2 L'analyse des Correspondances*. Dunod, Paris.
- Berk, R.H. (1972) Consistency and asymptotic normality of MLE's for exponential models. *Ann. Math. Statist.* **43**, 193–204.
- Berkson, J. (1944) Application of the logistic function to bio-assay. *J. Am. Statist. Assoc.* **39**, 357–65.
- Berkson, J. (1951) Why I prefer logits to probits. *Biometrics* **7**, 327–39.
- Berman, M. (1987) The asymptotic statistical behaviour of some estimators for a circular structural model. Unpublished technical report.
- Berman, M. and Culpin, D. (1986) The statistical behaviour of some least squares estimators of the centre and radius of a circle. *J. R. Statist. Soc. B* **48**, 183–96.
- Berman, M. and Griffiths, D. (1985) Incorporating angular information into models for stone circle data. *Appl. Statist.* **34**, 237–45.
- Bhapkar, V.P. (1972) On a measure of efficiency in an estimating equation. *Sankhya A* **34**, 467–72.
- Bhattacharya, R.N. and Rao, R.R. (1976) *Normal Approximation and Asymptotic Expansions*. J. Wiley & Sons, New York.

- Birch, M.W. (1963) Maximum likelihood in three-way contingency tables. *J. R. Statist. Soc. B* **25**, 220–33.
- Birch, M.W. (1965) The detection of partial association II: the general case. *J. R. Statist. Soc. B* **27**, 111–24.
- Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975) *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.
- Björck, Å. (1967) Solving least squares problems by Gram-Schmidt orthogonalization. *BIT* **7**, 257–78.
- Bliss, C.I. (1935) The calculation of the dosage–mortality curve. *Ann. Appl. Biol.* **22**, 134–67.
- Bliss, C.I. (1967) *Statistics in Biology*. vol. I, McGraw-Hill, New York.
- Bliss, C.I. (1970) *Statistics in Biology*. vol. II, McGraw-Hill, New York.
- Bock, R.D. (1975) *Multivariate Statistical Methods in Behavioral Research*. McGraw-Hill, New York.
- Bock, R.D. (ed.) (1989) *Multilevel Analysis of Educational Data*. Academic Press, New York.
- Bortkewitsch, L. von (1898) *Das Gesetz der Kleinen Zahlen*. Teubner, Leipzig.
- Box, G.E.P. (1980) Sampling and Bayes' inference in scientific modelling and robustness. *J. R. Statist. Soc. A* **143**, 383–430.
- Box, G.E.P. (1988) Signal-to-noise ratios, performance criteria, and transformations (with discussion). *Technometrics* **30**, 1–17.
- Box, G.E.P. and Cox, D.R. (1964) An analysis of transformations. *J. R. Statist. Soc. B* **26**, 211–52.
- Box, G.E.P. and Cox, D.R. (1982) An analysis of transformations revisited, rebutted. *J. Am. Statist. Assoc.* **77**, 209–10.
- Box, G.E.P. and Jenkins, G.M. (1976) *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.
- Box, G.E.P. and Tidwell, P.W. (1962) Transformation of the independent variables. *Technometrics* **4**, 531–50.
- Bradley, E.L. (1973) The equivalence of maximum likelihood and weighted least squares in the exponential family. *J. Am. Statist. Assoc.* **68**, 199–200.
- Bradley, R.A. and Terry, M.A. (1952) Rank analysis of incomplete block designs I. *Biometrika* **39**, 324–45.
- Breese, E.L. and Hill, J. (1973) Regression analysis of interactions between competing species. *Heredity* **31**, 181–200.
- Breslow, N.E. (1976) Regression analysis of the log odds ratio: a method for retrospective studies. *Biometrics* **32**, 409–16.
- Breslow, N.E. (1981) Odds ratio estimators when the data are sparse. *Biometrika* **68**, 73–84.
- Breslow, N.E. and Liang, K.Y. (1982) The variance of the Mantel-Haenszel estimator. *Biometrics* **38**, 943–52.
- Breslow, N.E. and Cologne, J. (1986) Methods of estimation in log odds ratio regression models. *Biometrics* **42**, 949–954.

- Breslow, N.E. and Day, N.E. (1980) *Statistical Methods in Cancer Research. 1: The Analysis of Case-Control Studies*. I.A.R.C, Lyon.
- Brillinger, D.R. and Preisler, H.K. (1986) Two examples of quantal data analysis: a) multivariate point process, b) pure death process in an experimental design. *Proc XIII International Biometric Conference Seattle*.
- Bross, I.D.J. (1954) Misclassification in 2×2 tables. *Biometrics* **10**, 478-86.
- Bross, I.D.J. (1958) How to use ridit analysis. *Biometrics* **14**, 18-38.
- Burridge, J. (1982) Some unimodality properties of likelihoods derived from grouped data. *Biometrika* **69**, 145-51.
- Byar, D.T. (1983) Analysis of survival data: Cox and Weibull models with covariates, in *Statistics in Medical Research: Methods and Issues, with Applications in Clinical Oncology*. (eds. V. Mike and K. Stanley), J. Wiley & Sons, New York.
- Caussinus, H. (1965) Contribution à l'analyse statistique des tableaux de corrélation. *Ann. Fac. Sci. Univ. Toulouse* **29**, 77-182.
- Chambers, E.A. and Cox, D.R. (1967) Discrimination between alternative binary response models. *Biometrika* **54**, 573-8.
- Chambers, J.M. (1977) *Computational Methods for Data Analysis*. J. Wiley & Sons, New York.
- Chatterjee, S. and Hadi, A.S. (1986) Influential observations, high-leverage points, and outliers in linear regression. *Statistical Science* **1**, 379-416.
- Chen, K.K., Bliss, C.I. and Robbins, E.B. (1942) The digitalis-like principles of *calotropis* compared with other cardiac substances. *J. Pharmacol. Experimental Therapeutics* **74**, 223-34.
- Clarke, M.R.B. (1981) Algorithm AS163: a Givens algorithm for moving from one linear model to another without going back to the data. *Appl. Statist.* **30**, 198-203.
- Clarke, M.R.B. (1982) Algorithm AS178: the Gauss-Jordan sweep operator with detection of collinearity. *Appl. Statist.* **31**, 166-8.
- Clayton, D.G. (1974) Some odds ratio statistics for the analysis of ordered categorical data. *Biometrika* **61**, 525-31.
- Cleveland, W. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Statist. Assoc.* **74**, 829-836.
- Coe, R. and Stern, R.D. (1982) Fitting models to daily rainfall. *J. Appl. Meteorol.* **21**, 1024-31.
- Coe, R. and Stern, R.D. (1984) A model fitting analysis of daily rainfall data. *J. R. Statist. Soc. A* **147**, 1-34.
- Cook, R.D. (1977) Detection of influential observations in linear regression. *Technometrics* **19**, 15-18.
- Cook, R.D. (1979) Influential observations in linear regression. *J. Am. Statist. Assoc.* **74**, 169-74.
- Cook, R.D. and Weisberg, S. (1982) *Residuals and Influence in Regression*. Chapman and Hall, London.

- Cook, R.D. and Weisberg, S. (1983) Diagnostics for heteroscedasticity in regression. *Biometrika* **70**, 1-10.
- Copas, J.B. (1988) Binary regression models for contaminated data. *J. R. Statist. Soc. B* **50**, 225-65.
- Corbeil, R.R. and Searle, S.R. (1976) Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics* **18**, 31-8.
- Cordeiro, G.M. (1983) Improved likelihood-ratio tests for generalized linear models. *J. R. Statist. Soc. B* **45**, 404-13.
- Cordeiro, G.M. (1987) On the corrections to the likelihood-ratio statistics. *Biometrika* **74**, 265-74.
- Cornish, E.A. and Fisher, R.A. (1937) Moments and cumulants in the specification of distributions. *Revue de l'Institut Internationale de Statistique* **4**, 1-14.
- Cox, D.R. (1958a) The regression analysis of binary sequences (with discussion). *J. R. Statist. Soc. B* **20**, 215-42.
- Cox, D.R. (1958b) Two further applications of a model for binary regression. *Biometrika* **45**, 562-5.
- Cox, D.R. (1958c) *Planning of Experiments*. J. Wiley & Sons, New York.
- Cox, D.R. (1970) *The Analysis of Binary Data*. Chapman and Hall, London.
- Cox, D.R. (1972a) Regression models and life tables (with discussion). *J. R. Statist. Soc. B* **74**, 187-220.
- Cox, D.R. (1972b) The analysis of multivariate binary data. *Appl. Statist.* **21**, 113-20.
- Cox, D.R. (1975) Partial likelihood. *Biometrika* **62**, 269-76.
- Cox, D.R. (1983) Some remarks on over-dispersion. *Biometrika* **70**, 269-74.
- Cox, D.R. and Hinkley, D.V. (1968) A note on the efficiency of least-squares estimates. *J. R. Statist. Soc. B* **30**, 284-9.
- Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics*. Chapman and Hall, London.
- Cox, D.R. and Oakes, D. (1984) *Analysis of Survival Data*. Chapman and Hall, London.
- Cox, D.R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference, (with discussion). *J. R. Statist. Soc. B* **49**, 1-39.
- Cox, D.R. and Snell, E.J. (1968) A general definition of residuals. *J. R. Statist. Soc. B* **30**, 248-75.
- Cox, D.R. and Snell, E.J. (1981) *Applied Statistics: Principles and Examples*. Chapman and Hall, London.
- Critchlow, D.E. (1985) *Metric Methods for Analysing Partially Ranked Data*. Lecture Notes in Statistics **34**, Springer-Verlag, New York.
- Crowder, M.J. (1978) Beta-binomial anova for proportions. *Appl. Statist.* **27**, 34-7.

- Crowder, M.J. (1985) Gaussian estimation for correlated binomial data. *J. R. Statist. Soc. B* **47**, 229-37.
- Crowder, M.J. (1987) On linear and quadratic estimating functions. *Biometrika* **74**, 591-7.
- Crowley, J. and Hu, M. (1977) Covariance analysis of heart transplant data. *J. Am. Statist. Assoc.* **72**, 27-36.
- Dale, J.R. (1984) Local versus global association for bivariate ordered responses. *Biometrika* **71**, 507-14.
- Dale, J.R. (1986) Global cross-ratio models for bivariate discrete ordered responses. *Biometrics* **42**, 909-17.
- Daniel, C. (1959) Use of half-normal plots for interpreting factorial two-level experiments. *Technometrics* **1**, 311-41.
- Darby, S.C. and Ellis, M.J. (1976) A test for synergism between two drugs. *Appl. Statist.* **25**, 296-9.
- Darroch, J.N. and Ratcliff, D. (1972) Generalized iterative scaling for log-linear models. *Ann. Math. Statist.* **43**, 1470-80.
- Darroch, J.N., Lauritzen, S.L. and Speed, T.P. (1980) Markov fields and log-linear interaction models for contingency tables. *Ann. Math. Statist.* **8**, 522-39.
- Davidian, M. and Carroll, R.J. (1988) A note on extended quasi-likelihood. *J. R. Statist. Soc. B* **50**, 74-82.
- Davison, A.C. (1988) Appproximate conditional inference in generalized linear models. *J. R. Statist. Soc. B* **50**, 445-61.
- Davison, A.C. and Tsai, C.-L. (1988) Diagnostics for regression models with a linear part. Unpublished manuscript.
- Dawson, R.B. (1954) A simplified expression for the variance of the χ^2 -function on a contingency table. *Biometrika* **41**, 280.
- Day, N.E. and Byar, D.T. (1979) Testing hypotheses in case-control studies: equivalence of Mantel-Haenszel statistics and logit score tests. *Biometrics* **35**, 623-30.
- Deming, W.E. and Stephan, F.F. (1940) On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* **11**, 427-44.
- Dempster, A.P. (1971) An overview of multivariate data analysis. *J. Mult. Analysis* **1**, 316-46.
- Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B* **39**, 1-38.
- Diaconis, P. (1988) *Group Representations in Probability and Statistics*. I.M.S. Monograph Series **11**, Hayward, CA.
- Dobson, A.J. (1983) *An Introduction to Statistical Modelling*. Chapman and Hall, London.
- Draper, N.R. and Smith, H. (1981) *Applied Regression Analysis*. (2nd edn). J. Wiley & Sons, New York.
- Dyke, G.V. and Patterson, H.D. (1952) Analysis of factorial arrangements when the data are proportions. *Biometrics* **8**, 1-12.

- Efron, B. (1975) The efficiency of logistic regression compared to Normal discriminant analysis. *J. Am. Statist. Assoc.* **70**, 892-8.
- Efron, B. (1977) The efficiency of Cox's likelihood function for censored data. *J. Am. Statist. Assoc.* **72**, 557-65.
- Efron, B. (1978) Regression and ANOVA with zero-one data: measures of residual variation. *J. Am. Statist. Assoc.* **73**, 113-21.
- Efron, B. (1986) Double exponential families and their use in generalized linear regression. *J. Am. Statist. Assoc.* **81**, 709-21.
- Efroymson, M.A. (1960) Multiple regression analysis, in *Mathematical Methods for Digital Computers*. vol. 1 (eds. A. Ralston and H.S. Wilf), J. Wiley & Sons, New York, 191-203.
- Ekholm, A. and Palmgren, J. (1982) A model for a binary response with mis-classification. In *GLIM 82: Proceedings of the International Conference on Generalized Linear Models*. (ed. R. Gilchrist), 128-43, Springer-Verlag, New York.
- Elandt-Johnson, R.C. and Johnson, N.L. (1980) *Survival Models and Data Analysis*. J. Wiley & Sons, New York.
- Engel, J. (1987) *The Analysis of Dependent Count Data*. PhD Thesis, Landbouwuniversiteit, Wageningen.
- Engelhardt, M. and Bain, L.J. (1977) Uniformly most powerful unbiased tests of the scale parameter of a gamma distribution with a nuisance shape parameter. *Technometrics* **19**, 77-81.
- Engelhardt, M. and Bain, L.J. (1978) Construction of optimal unbiased inference procedures for the parameters of the gamma distribution. *Technometrics* **20**, 485-9.
- Esseen, C.G. (1945) Fourier analysis of distribution functions. *Acta Mathematica* **77**, 1-125.
- Everitt, B.S. (1977) *The Analysis of Contingency Tables*. Chapman and Hall, London.
- Feigl, P. and Zelen, M. (1965) Estimation of exponential survival probabilities with concomitant information. *Biometrics* **21**, 826-38.
- Fienberg, S.E. (1970a) An iterative procedure for estimation in contingency tables. *Ann. Math. Statist.* **41**, 907-17.
- Fienberg, S.E. (1970b) The analysis of multidimensional contingency tables. *Ecology* **51**, 419-33.
- Fienberg, S.E. (1980) *The Analysis of Cross-Classified Categorical Data*. (2nd edn). MIT Press, Cambridge, MA.
- Finney, D.J. (1971) *Probit Analysis*. (3rd edn). Cambridge University Press.
- Finney, D.J. (1976) Radioligand assay. *Biometrics* **32**, 721-40.
- Firth, D. (1982) Estimation of voter transition matrices. MSc Thesis, University of London.
- Firth, D. (1987) On the efficiency of quasi-likelihood estimation. *Biometrika* **74**, 233-45.
- Firth, D. (1988) Multiplicative errors: log-normal or gamma? *J. R. Statist. Soc. B* **50**, 266-8.

- Fisher, R.A. (1922) On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc.* **222**, 309–68.
- Fisher, R.A. (1935) The case of zero survivors (Appendix to Bliss, C.I. (1935)). *Ann. Appl. Biol.* **22**, 164–5.
- Fisher, R.A. (1949) A biological assay of tuberculins. *Biometrics* **5**, 300–16.
- Fisher, R.A. (1958) *Statistical Methods for Research Workers*. 13th edition. Oliver & Boyd, Edinburgh.
- Fleiss, J.L. (1981) *Statistical Methods for Rates and Proportions*. J. Wiley & Sons, New York.
- Fowlkes, E.B. (1987) Some diagnostics for binary logistic regression via smoothing. *Biometrika* **74**, 503–15.
- Fraser, D.A.S. (1968) *The Structure of Inference*. J. Wiley & Sons, New York.
- Fraser, D.A.S. (1979) *Inference and Linear Models*. McGraw-Hill, New York.
- Freeman, P.R. (1977) Note: Thom's survey of the Avebury ring. *J. Hist. Astronomy* **8**, 134–6.
- Freireich, E.J. et al. (1963) The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia. *Blood* **21**, 699–716.
- Frome, E.L. (1983) The analysis of rates using Poisson regression models. *Biometrics* **39**, 665–74.
- Furnival, G.M. and Wilson, R.W. (1974) Regression by leaps and bounds. *Technometrics* **16**, 499–511.
- Gail, M.H., Lubin, J.H. and Rubenstein, L.V. (1981) Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika* **68**, 703–7.
- Galton, F. (1889) *Natural Inheritance*. Macmillan, London.
- Gart, J.J. and Zweifel, J.R. (1967) On the bias of various estimators of the logit and its variance with applications to quantal bioassay. *Biometrika* **54**, 181–7.
- Gauss, C.F. (1823) *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*. Göttingen: Dieterich.
- Gehan, E.A. (1965) A generalized Wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika* **52**, 203–23.
- Gentleman, W.M. (1974a) Regression problems and the QR decomposition. *Bull. I.M.A.* **10**, 195–7.
- Gentleman, W.M. (1974b) Basic procedures for large, sparse or weighted linear least squares problems (Algorithm AS75). *Appl. Statist.* **23**, 448–57.
- Gilchrist, R. (ed.) (1982) *GLIM 82: Proceedings of the International Conference on Generalized Linear Models*. Springer-Verlag, New York.
- Gilchrist, R., Francis, B. and Whittaker, J. (eds.) (1985) *Generalized Linear Models*. Lecture Notes in Statistics, **32**, Springer-Verlag, Berlin.

- Gilmour, A.R., Anderson, R.D. and Rae, A.L. (1985) The analysis of binomial data by a generalized linear mixed model. *Biometrika* **72**, 593-9.
- Gilula, Z. and Haberman, S.J. (1986) Canonical analysis of contingency tables by maximum likelihood. *J. Am. Statist. Assoc.* **81**, 780-8.
- Godambe, V.P. (1960) An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31**, 1208-12.
- Godambe, V.P. (1976) Conditional likelihood and optimum estimating equations. *Biometrika* **63**, 277-84.
- Godambe, V.P. and Heyde, C.C. (1987) Quasi-likelihood and optimal estimation. *Int. Statist. Rev.* **55**, 231-44.
- Godambe, V.P. and Thompson, M.E. (1988) An extension of quasi-likelihood estimation. *J. Statist. Planning and Inference* **22**, (to appear).
- Gokhale, D.V. and Kullback, S. (1978) *The Information in Contingency Tables*. Marcel Dekker, New York.
- Goldstein, H. (1986) Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* **73**, 43-56.
- Goodhardt, G.J., Ehrenberg, A.S.C. and Chatfield, C. (1984) The Dirichlet: A comprehensive model of buying behaviour. *J. R. Statist. Soc. A* **147**, 621-55.
- Goodman, L.A. (1973) The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. *Biometrika* **60**, 179-92.
- Goodman, L.A. (1978) *Analysing Qualitative/Categorical Data*. Addison-Wesley, Reading, MA.
- Goodman, L.A. (1979) Simple models for the analysis of association in cross-classifications having ordered categories. *J. Am. Statist. Assoc.* **74**, 537-52.
- Goodman, L.A. (1981) Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *J. Am. Statist. Assoc.* **76**, 320-34.
- Goodman, L.A. (1986) Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach to the analysis of contingency tables (with discussion). *Int. Statist. Rev.* **54**, 243-309.
- Graubard, B.I. and Korn, E.L. (1987) Choice of column scores for testing independence in ordered $2 \times k$ contingency tables. *Biometrics* **43**, 471-6.
- Green, P.J. (1984) Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). *J. R. Statist. Soc. B* **46**; 149-92.
- Green, P.J. (1987) Penalized likelihood for general semi-parametric regression models. *Int. Statist. Rev.* **55**, 245-60.
- Greenacre, M.J. (1984) *Theory and Applications of Correspondence Analysis*. Academic Press, New York.

- Greenwood, J.A. and Durand, D. (1960) Aids for fitting the gamma distribution by maximum likelihood. *Technometrics* **2**, 55–65.
- Grizzle, J.E. (1971) Multivariate logit analysis. *Biometrics* **27**, 1057–62.
- Gross, A.J. and Clark, V.A. (1975) *Survival Distributions: Reliability Applications in the Biomedical Sciences*. J. Wiley & Sons, New York.
- Gurland, J., Lee, I. and Dahm, P.A. (1960) Polychotomous quantal response in biological assay. *Biometrics* **16**, 382–98.
- Haberman, S.J. (1974a) *The Analysis of Frequency Data*. University of Chicago Press, Chicago.
- Haberman, S.J. (1974b) Log-linear models for frequency tables with ordered classifications. *Biometrics* **30**, 589–600.
- Haberman, S.J. (1977) Maximum likelihood estimates in exponential response models. *Ann. Statist.* **5**, 815–41.
- Haberman, S.J. (1978) *Analysis of Qualitative Data, 1: Introductory Topics*. Academic Press, New York.
- Haberman, S.J. (1979) *Analysis of Qualitative Data, 2: New Developments*. Academic Press, New York.
- Haberman, S.J. (1981) Tests for independence in two-way contingency tables based canonical correlation and on linear-by-linear interaction. *Ann. Statist.* **9**, 1178–86.
- Haberman, S.J. (1982) The analysis of dispersion of multinomial responses. *J. Am. Statist. Assoc.* **77**, 568–80.
- Haldane, J.B.S. (1937) The exact value of the moments of the distribution of χ^2 used as a test of goodness of fit when expectations are small. *Biometrika* **29**, 133–43.
- Haldane, J.B.S. (1939) The mean and variance of χ^2 when used as a test of homogeneity when expectations are small. *Biometrika* **31**, 346–65.
- Hamilton, D. (1987) Sometimes $R^2 > r_{yx_1}^2 + r_{yx_2}^2$: correlated variables are not always redundant. *Amer. Statistician* **41**, 129–32.
- Harkness, W.L. (1965) Properties of the extended hypergeometric distribution. *Ann. Math. Statist.* **36**, 938–45.
- Harter, H.L. (1976) The method of least squares and some alternatives. *Int. Statist. Rev.* **44**, 113–59.
- Harville, D.A. (1974) Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383–5.
- Harville, D.A. (1977) Maximum likelihood approaches to variance component estimation and to related problems (with discussion). *J. Am. Statist. Assoc.* **72**, 320–40.
- Hastie, T. and Tibshirani, R. (1986) Generalized additive models (with discussion). *Statistical Science* **1**, 297–318.
- Hastie, T. and Tibshirani, R. (1987a) Generalized additive models: some applications. *J. Am. Statist. Assoc.* **82**, 371–86.
- Hastie, T. and Tibshirani, R. (1987b) Non-parametric logistic and proportional-odds regression. *Appl. Statist.* **36**, 260–76.
- Hawkins, D. (1980) *Identification of Outliers*. Chapman and Hall, London.

- Healy, M.J.R. (1968) Triangular decomposition of a symmetric matrix (Algorithm AS6); Inversion of positive semi-definite symmetric matrix (Algorithm AS7), *Appl. Statist.* **17**, 192–7.
- Healy, M.J.R. (1981) A source of assay heterogeneity. *Biometrics* **37**, 834–5.
- Healy, M.J.R. (1986) *Matrices for Statistics*. Oxford: Clarendon Press.
- Healy, M.J.R. (1988) *GLIM: An Introduction*. Oxford: Clarendon Press.
- Hewlett, P.S. (1969) Measurement of the potencies of drug mixtures. *Biometrics* **25**, 477–87.
- Hewlett, P.S. and Plackett, R.L. (1956) The relation between quantal and graded responses to drugs. *Biometrics* **12**, 72–8.
- Heyde, C.C. (1987) On combining quasi-likelihood estimating functions. *Stochastic Processes and their Applications* **25**, 281–7.
- Hill, J.R. and Tsai, C.-L. (1988) Calculating the efficiency of maximum quasi-likelihood estimation. *Appl. Statist.* **37**, 219–30.
- Hill, M.O. (1974) Correspondence analysis: A neglected multivariate method. *Appl. Statist.* **23**, 340–54.
- Hinkley, D.V. (1985) Transformation diagnostics for linear models. *Biometrika* **72**, 487–96.
- Hoaglin, D.C. and Welsch, R.E. (1978) The hat matrix in regression and ANOVA. *Am. Statistician* **32**, 17–22 (Corr. **32**, 146).
- Hougaard, P. (1982) Parametrizations of non-linear models. *J. R. Statist. Soc. B* **44**, 244–52.
- Hurn, M.W., Barker, N.W. and Magath, T.D. (1945) The determination of prothrombin time following the administration of dicumarol with specific reference to thromboplastin. *J. Lab. Clin. Med.* **30**, 432–47.
- Jarrett, R.G. (1973) *Efficiency and Estimation in Asymptotically Normal Distributions*. University of London PhD Thesis.
- Jeffreys, H. (1961) *Theory of Probability*. (3rd edn). Oxford: Clarendon Press.
- Jeffreys, H. and Jeffreys, B.S. (1956) *Methods of Mathematical Physics*. (3rd edn). Cambridge University Press.
- Jennrich, R.I. (1977) Stepwise regression, in *Statistical Methods for Digital Computers*. (eds. Enslein, Ralston and Wilf), pp. 58–75, J. Wiley & Sons, New York.
- Jennrich, R.I. and Moore, R.H. (1975) Maximum likelihood estimation by means of non-linear least squares. *Amer. Statist. Assoc. Proc. Statist. Computing Section*, 57–65.
- Johansen, S. (1983) Some topics in regression (with discussion). *Scand. J. Statist.* **10**, 161–94.
- Johnson, B.W. and McCulloch, R.E. (1987) Added-variable plots in linear regression. *Technometrics* **29**, 427–33.
- Jørgensen, B. (1984) The delta algorithm and GLIM. *Int. Statist. Rev.* **52**, 283–300.

- Jørgensen, B. (1987) Exponential dispersion models (with discussion). *J. R. Statist. Soc. B* **49**, 127–62.
- Kalbfleisch, J.D. and Prentice, R.L. (1980) *The Statistical Analysis of Failure Time Data*. J. Wiley & Sons, New York.
- Kalbfleisch, J.D. and Sprott, D.A. (1970) Application of likelihood methods to models involving a large number of nuisance parameters (with discussion). *J. R. Statist. Soc. B* **32**, 175–208.
- Kay, R. and Little, S. (1987) Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika* **74**, 503–15.
- Kempthorne, O. (1952) *Design and Analysis of Experiments*. J. Wiley & Sons, New York.
- Kendall, M.G. and Stuart, A. (1977) *The Advanced Theory of Statistics*, 1. (4th edn). Griffin, London.
- Killion, R.A. and Zahn, D.A. (1976) A bibliography of contingency table literature. *Int. Statist. Rev.* **44**, 71–112.
- Kitanidis, P.K. (1983) Statistical estimation of polynomial generalized covariance functions in hydrologic applications. *Water Resources Research* **19**, 909–21.
- Kitanidis, P.K. (1987) Parametric estimation of covariances of regionalized variables. *Water Resources Bulletin* **23**, 557–67.
- Knuth, D.E. (1986) *The TeXbook*. Reading, Mass.: Addison-Wesley.
- Kolassa, J. and McCullagh, P. (1987) Edgeworth series for lattice distributions. Technical Report No. 220, Dept of Statistics, University of Chicago.
- Kruskal, W.H. (1968) When are Gauss-Markov and least squares estimators identical? A coordinate-free approach. *Ann. Math. Statist.* **39**, 70–5.
- Kruskal, W.H. (1975) The geometry of generalized inverses, *J. R. Statist. Soc. B* **37**, 272–83.
- Läärä, E. and Matthews, J.N.S. (1985) The equivalence of two models for ordinal data. *Biometrika* **72**, 206–7.
- Lane, P.W. and Nelder, J.A. (1982) Analysis of covariance and standardization as instances of prediction. *Biometrics* **38**, 613–21.
- Lawless, J.F. (1982) *Statistical Models and Methods for Lifetime Data*. J. Wiley & Sons, New York.
- Lawless, J.F. (1987) Negative binomial and mixed Poisson regression. *Can. J. Statist.* **15**, 209–25.
- Lawless, J.F. and Singhal, K. (1978) Efficient screening of non-normal regression models. *Biometrics* **43**, 318–27.
- Lawley, D.N. (1956) A general method for approximating to the distribution of likelihood-ratio criteria. *Biometrika* **43**, 295–303.
- Lawson, C.L. and Hanson, R.J. (1974) *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, NJ.
- Lee, A.H. (1987) Diagnostic displays for assessing leverage and influence in generalized linear models. *Austral. J. Statist.* **29**, 233–43.

- Lee, A.H. (1988) Assessing partial influence in generalized linear models. *Biometrics* **44**, 71–7.
- Lee, E.T. (1980) *Statistical Methods for Survival Data Analysis*. Lifetime Learning Publications, Belmont, CA.
- Lehmann, E.L. (1986) *Testing Statistical Hypotheses*. J. Wiley & Sons, New York.
- Liang, K.Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.
- Lexis, W. (1879) Über die Theorie der Stabilität statistischer Reichen. *Jahrbücher für Nationalökonomie und Statistik* **32**, 60–98.
- Liao, J. (1988) 2×2 algorithm. *Royal Statistical Society News and Notes* **14**, 9 (May 1988) p.3.
- Lindsay, B. (1982) Conditional score functions: some optimality results. *Biometrika* **69**, 503–12.
- Lowe, C.R., Roberts, C.J. and Lloyd, S. (1971) Malformations of central nervous system and softness of local water supplies. *British Medical Journal* **15**, 357–61.
- McCullagh, P. (1980) Regression models for ordinal data (with discussion). *J. R. Statist. Soc. B* **42**, 109–42.
- McCullagh, P. (1982) Some applications of quasisymmetry. *Biometrika* **69**, 303–8.
- McCullagh, P. (1983) Quasi-likelihood functions. *Ann. Statist.* **11**, 59–67.
- McCullagh, P. (1984a) Local sufficiency. *Biometrika* **71**, 233–44.
- McCullagh, P. (1984b) Generalized linear models. *European J. Operational Research* **16**, 285–92.
- McCullagh, P. (1984c) On the elimination of nuisance parameters in the proportional odds model. *J. R. Statist. Soc. B* **46**, 250–6.
- McCullagh, P. (1985) On the asymptotic distribution of Pearson's statistic in linear exponential-family models. *Int. Statist. Rev.* **53**, 61–7.
- McCullagh, P. (1986) The conditional distribution of goodness-of-fit statistics for discrete data. *J. Am. Statist. Assoc.* **81**, 104–7.
- McCullagh, P. (1987) *Tensor Methods in Statistics*. Chapman and Hall, London.
- McCullagh, P. (1989) Some statistical properties of a new family of continuous univariate distributions. *J. Am. Statist. Assoc.* **84**, 125–9.
- McCullagh, P. and Pregibon, D. (1987) k -statistics and dispersion effects in regression. *Ann. Statist.* **15**, 202–19.
- McCullagh, P. and Tibshirani, R. (1988) A simple method for the adjustment of profile likelihoods. Technical Report No. 238, Dept of Statistics, University of Chicago.
- McLeish, D.L. and Small, C.G. (1988) *The Theory and Application of Statistical Inference Functions*. Lecture Notes in Statistics **44**, Springer-Verlag, New York.

- McNemar, Q. (1947) Note on the sampling error of the difference between two correlated proportions or percentages. *Psychometrika* **12**, 153-7.
- Mallows, C.L. (1957) Non-null ranking models, I. *Biometrika* **44**, 114-30.
- Mallows, C.L. (1973) Some comments on C_p . *Technometrics* **15**, 661-75.
- Mandel, J. (1959) The analysis of Latin squares with a certain type of row-column interaction. *Technometrics* **1**, 379-87.
- Mandel, J. (1971) A new analysis of variance model for non-additive data. *Technometrics* **13**, 1-18.
- Mantel, N. (1966) Models for complex contingency tables and poly-chotomous dosage response curves. *Biometrics* **22**, 83-95.
- Mantel, N. (1977) Tests and limits for the common odds ratio of several 2×2 tables: methods in analogy with the Mantel-Haenszel procedure. *J. Statist. Planning and Inference* **1**, 179-89.
- Mantel, N. and Brown, C. (1973) A logistic reanalysis of Ashford and Sowden's data on respiratory symptoms in British coalminers. *Biometrics* **29**, 649-65.
- Mantel, N. and Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst.* **22**, 719-48.
- Mantel, N. and Hankey, W. (1975) The odds ratio of a 2×2 contingency table. *Am. Statistician* **29**, 143-5.
- Manton, K.G., Woodbury, M.A. and Stallard, E. (1981) A variance components approach to categorical data models with heterogeneous cell populations: analysis of spatial gradients in lung cancer mortality rates in North Carolina counties. *Biometrics* **37**, 259-69.
- Mardia, K.V. and Marshall, R.J. (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135-46.
- Maxwell, A.E. (1961) *Analysing Qualitative Data*. Methuen, London.
- Mead, R. and Curnow, R.N. (1983) *Statistical Methods in Agriculture and Experimental Biology*. Chapman and Hall, London.
- Miller, R.G. (1981) *Survival Analysis*. J. Wiley & Sons, New York.
- Moore, D.F. (1987) Modelling the extraneous variance in the presence of extra-binomial variation. *Appl. Statist.* **36**, 8-14.
- Morris, C.N. (1982) Natural exponential families with quadratic variance functions. *Ann. Statist.* **10**, 65-80.
- Morris, C.N. (1983) Natural exponential families with quadratic variance functions: statistical theory. *Ann. Statist.* **11**, 515-29.
- Morton, R. (1981) Efficiency of estimating equations and the use of pivots. *Biometrika* **68**, 227-33.
- Morton, R. (1987) A generalized linear model with nested strata of extra-Poisson variation. *Biometrika* **74**, 247-57.

- Mosteller, F. and Tukey, J.W. (1977) *Data Analysis and Regression*. Addison-Wesley, New York.
- Muirhead, C.R. and Darby, S.C. (1987) Modelling the relative and absolute risks of radiation-induced cancers. *J. R. Statist. Soc. A* **150**, 83–118.
- Nair, V.N. and Pregibon, D. (1988) Contribution to the discussion of Box (1988). *Technometrics* **30**, 24–29.
- Nelder, J.A. (1965a) The analysis of randomized experiments with orthogonal block structure. I. Block structure and the null analysis of variance. *Proc. Roy. Soc. A* **283**, 147–62.
- Nelder, J.A. (1965b) The analysis of randomized experiments with orthogonal block structure. II. Treatment structure and the general analysis of variance. *Proc. Roy. Soc. A* **283**, 163–78.
- Nelder, J.A. (1966) Inverse polynomials, a useful group of multi-factor response functions. *Biometrics* **22**, 128–41.
- Nelder, J.A. (1974) Log linear models for contingency tables: a generalization of classical least squares. *Appl. Statist.* **23**, 323–9.
- Nelder, J.A. (1977) A reformulation of linear models. *J. R. Statist. Soc. A* **140**, 48–77.
- Nelder, J.A. (1982) Linear models and non-orthogonal data. *Utilitas Mathematica* **21B**, 141–51.
- Nelder, J.A. (1984) Models for rates with Poisson errors. *Biometrics* **40**, 1159–62.
- Nelder, J.A. and Pregibon, D. (1987) An extended quasi-likelihood function. *Biometrika* **74**, 221–32.
- Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized linear models. *J. R. Statist. Soc. A* **135**, 370–84.
- Nelson, W. (1982) *Applied Life Data Analysis*. J. Wiley & Sons, New York.
- Neyman, J. and Scott, E.L. (1948) Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.
- Oakes, D. (1977) The asymptotic information in censored survival data. *Biometrika* **64**, 441–8.
- Oakes, D. (1981) Survival times: aspects of partial likelihood. *Int. Statist. Rev.* **49**, 235–64.
- Palmgren, J. (1981) The Fisher information matrix for log-linear models arguing conditionally in the observed explanatory variables. *Biometrika* **68**, 563–6.
- Palmgren, J. (1987) *Models for categorical data with errors of observation*. Tilastotieteellisiä Tutkimuksia **8**, Finnish Statistical Society: Yliopistopaino, Helsinki.
- Patterson, H.D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–54.
- Payne, C.D. (ed.) (1986) *The GLIM Manual, Release 3.77*. NAG: Oxford.

- Pearson, K. (1900) On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag.* (5) **50**, 157–75. Reprinted in *Karl Pearson's Early Statistical Papers* (ed. E.S. Pearson). Cambridge University Press, Cambridge (1948).
- Pearson, K. (1901) Mathematical contributions to the theory of evolution. *Phil. Trans. R. Soc.* **195**, 79–150.
- Pearson, K. (1913) Note on the surface of constant association. *Biometrika* **9**, 534–7.
- Pearson, K. and Heron, D. (1913) On theories of association. *Biometrika* **9**, 159–315.
- Peto, R. (1972) Contribution to the discussion of Cox (1972): Regression modes and life tables. *J. R. Statist. Soc. B* 205–7.
- Phelps, K. (1982) Use of the complementary log-log function to describe dose-response relationships in insecticide evaluation field trials. In *Lecture Notes in Statistics, No. 14. GLIM.82: Proceedings of the International Conference on Generalized Linear Models*. (ed. R. Gilchrist), Springer-Verlag, New York.
- Pierce, D.A. and Schafer, D.W. (1986) Residuals in generalized linear models. *J. Am. Statist. Assoc.* **81**, 977–86.
- Pignatiello, J.J. and Ramberg, J.S. (1985) Contribution to discussion of off-line quality-control, parameter design, and the Taguchi method. *J. Quality Technology* **17**, 198–206.
- Plackett, R.L. (1960) *Principles of Regression Analysis*. Oxford: Clarendon Press.
- Plackett, R.L. (1965) A class of bivariate distributions. *J. Am. Statist. Assoc.* **60**, 516–22.
- Plackett, R.L. (1981) *The Analysis of Categorical Data*. Griffin, London.
- Powsner, L. (1935) The effects of temperature on the durations of the developmental stages of *Drosophila Melanogaster*. *Physiological Zoölogy* **8**, 474–520.
- Pratt, J.W. (1981) Concavity of the log likelihood. *J. Am. Statist. Assoc.* **76**, 103–6.
- Preece, D.A., Ross, G.J.S. and Kirby, S.P.J. (1988) Bortkewitsch's horse-kicks and the generalised linear model. *The Statistician* **37**, 313–18.
- Pregibon, D. (1979) Data analytic methods for generalized linear models. PhD Thesis, University of Toronto.
- Pregibon, D. (1980) Goodness of link tests for generalized linear models. *Appl. Statist.* **29**, 15–24.
- Pregibon, D. (1981) Logistic regression diagnostics. *Ann. Statist.* **9**, 705–24.
- Pregibon, D. (1982) Score tests in GLIM with applications. In *Lecture Notes in Statistics, no. 14. GLIM.82: Proceedings of the International Conference on Generalized Linear Models*. (ed. R. Gilchrist).

- Springer-Verlag, New York.
- Pregibon, D. (1984) Review of *Generalized Linear Models*, *Ann. Statist.* **12**, 1589–96.
- Prentice, R.L. (1986) Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *J. Am. Statist. Assoc.* **81**, 321–7.
- Prentice, R.L. (1988) Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033–48.
- Prentice, R.L. (1987) Regression on variances and correlations using generalized estimating equations. Unpublished Technical report.
- Pringle, R.M. and Rayner, A.A. (1971) *Generalized Inverse Matrices with Applications to Statistics*. Griffin, London.
- Quine, M.P. and Seneta, E. (1987) Bortkiewicz's data and the law of small numbers. *Int. Statist. Rev.* **55**, 173–81.
- Rao, C.R. (1973) *Linear Statistical Inference and its Applications*. (2nd edn). J. Wiley & Sons, New York.
- Ross, W.H. (1987) The expectation of the likelihood-ratio criterion. *Int. Statist. Rev.* **55**, 315–30.
- Rousseeuw, P.J. and Leroy, A.M. (1988) *Robust Regression and Outlier Detection*. J. Wiley & Sons, New York.
- Rubin, G.M. (1988) *Drosophila melanogaster* as an experimental organism. *Science* **240**, 1453–9.
- Ryan, B.F., Joiner, B.L. and Ryan, T.A. (1985) *Minitab Handbook*. (2nd edn). Duxbury: Boston.
- Schoener, T.W. (1970) Nonsynchronous spatial overlap of lizards in patchy habitats. *Ecology* **51**, 408–18.
- Scholten, H. and van Wissen, L. (1985) A comparison of the loglinear interaction models with other spatial interaction models. In *Measuring the Unmeasurable*. NATO ASI Series D No. 22, (eds. P. Nijkamp, H. Leitner and N. Wrigley) Martinus Nijhoff, Dordrecht. pp. 177–96.
- Schreiner, H.R., Gregoire, R.C. and Lawrie, J.A. (1962) New biological effect of gases in the helium group. *Science* **136**, 653–4.
- Searle, S.R. (1971) *Linear Models*. J. Wiley & Sons, New York.
- Searle, S.R. (1979) Annotated computer output for analysis of variance of unequal-sub-class-numbers data. *Am. Statistician* **33**, 222–3.
- Seber, G.A.F. (1977) *Linear Regression Analysis*. J. Wiley & Sons, New York.
- Simon, G. (1974) Alternate analyses for the singly ordered contingency table. *J. Am. Statist. Assoc.* **69**, 971–6.
- Simpson, E.H. (1951) The interpretation of interaction in contingency tables. *J. R. Statist. Soc. B* **13**, 238–41.
- Skovgaard, I.M. (1986) A statistical model for competition experiments. *Scand. J. Statist.* **13**, 29–38.
- Smyth, G.K. (1985) Coupled and separable iterations in nonlinear estimation. PhD Thesis, Australian National University.

- Snedecor, G.W. and Cochran, W.G. (1967) *Statistical Methods*. (Sixth Edition), Iowa University Press.
- Snell, E.J. (1964) A scaling procedure for ordered categorical data. *Biometrics* **20**, 592-607.
- Solomon, H. (1961) Classification procedures based on dichotomous response vectors. In *Studies in Item Analysis and Prediction*. (ed. H. Solomon) Stanford University Press, pp. 177-86.
- Sprent, P. (1969) *Models in Regression and Related Topics*. Methuen, London.
- Stevens, S.S. (1951) Mathematics, measurement and psychophysics, in *Handbook of Experimental Psychology* (ed. S.S. Stevens) J. Wiley & Sons, New York.
- Stevens, S.S. (1958) Problems and methods of psychophysics, *Psychol. Bull.* **55**, 177-96.
- Stevens, S.S. (1968) Measurement, statistics and the schematic view. *Science* **161**, 849-56.
- Stewart, G.W. (1973) *Introduction to Matrix Computations*. Academic Press, New York.
- Stigler, S.M. (1978) Francis Ysidro Edgeworth, Statistician (with discussion). *J. R. Statist. Soc. A* **141**, 287-322.
- Stigler, S.M. (1981) Gauss and the invention of least squares. *Ann. Statist.* **9**, 465-74.
- Stigler, S.M. (1986) *The History of Statistics*. Belknap Press, Cambridge, Mass.
- Stiratelli, R., Laird, N.M. and Ware, J.H. (1984) Random-effects models for serial observations with binary response. *Biometrics* **40**, 961-71.
- Stone, M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Statist. Soc. B* **39**, 44-7.
- Stone, M. (1987) *Coordinate-Free Multivariable Statistics*. Oxford, Clarendon Press.
- Thisted, R.A. (1988) *Elements of Statistical Computing*. Chapman and Hall, London.
- Thom, A. (1967) *Megalithic Sites in Britain*. Oxford, Clarendon Press.
- Thom, A., Thom, A.S. and Foord, T.R. (1976) Avebury (1): a new assessment of the geometry and metrology of the ring. *J. Hist. Astronomy* **7**, 183-92.
- Thompson, R. and Baker, R.J. (1981) Composite link functions in generalised linear models. *Appl. Statist.* **30**, 125-31.
- Tukey, J.W. (1949) One degree of freedom for non-additivity. *Biometrics* **5**, 232-42.
- Tukey, J.W. (1962) The future of data analysis. *Ann. Math. Statist.* **33**, 1-67.
- Tweedie, M.C.K. (1957a) Statistical properties of inverse Gaussian distributions I. *Ann. Math. Statist.* **28**, 362-77.
- Tweedie, M.C.K. (1957b) Statistical properties of inverse Gaussian distributions II. *Ann. Math. Statist.* **28**, 696-705.

- Tweedie, M.C.K. (1981) An index which distinguishes between some important exponential families. *Proc. Indian Statistical Inst. Golden Jubilee International Conference*, 579–604.
- Upton, G.J.G. (1978) *The Analysis of Cross-Tabulated Data*. J. Wiley & Sons, New York.
- Upton, G.J.G. (1985) Modelling cross-tabulated regional data. In *Measuring the Unmeasurable*. NATO ASI Series D No. 22, (eds. P. Nijkamp, H. Leitner and N. Wrigley), 197–218. Martinus Nijhoff, Dordrecht.
- Urquhart, N.S. and Weeks, D.L. (1978) Linear models in messy data: some problems and alternatives. *Biometrics* **34**, 696–705.
- Wahrendorf, J. (1980) Inference in contingency tables with ordered categories using Plackett's coefficient of association for bivariate distributions. *Biometrika* **67**, 15–21.
- Walker, S.H. and Duncan, D.B. (1967) Estimation of the probability of an event as a function of several independent variables. *Biometrika* **54**, 167–79.
- Wampler, J.H. (1979) Solution to weighted least squares problems by modified Gram-Schmidt with iterative refinement. *ACM Trans. Math. Software* **5**, 494–9.
- Wang, P.C. (1987) Residual plots for detecting nonlinearity in generalized linear models. *Technometrics* **29**, 435–8.
- Wedderburn, R.W.M. (1974) Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* **61**, 439–47.
- Wedderburn, R.W.M. (1976) On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63**, 27–32.
- Welch, L.F., Adams, W.E. and Corman, J.L. (1963) Yield response surfaces, isoquants and economic fertilizer optima for coastal Bermudagrass. *Agron. J.* **55**, 63–7.
- Wermuth, N. (1976) Exploratory analyses of multidimensional contingency tables. *Proc. 9th Int. Biometrics Conference*, V.I, 279–95.
- West, M. (1985) Generalized linear models: scale parameters, outlier accommodation and prior distributions. *Bayesian statistics* **2**, 531–58.
- Whitehead, J. (1980) Fitting Cox's regression model to survival data using GLIM. *Appl. Statist.* **29**, 268–75.
- Wichura, M. (1986) The PiCTeX manual. Technical Report No. 205, Dept. of Statistics, University of Chicago.
- Wilkinson, G.N. and Rogers, C.E. (1973) Symbolic description of factorial models for analysis of variance. *Appl. Statist.* **22**, 392–9.
- Williams, D.A. (1976) Improved likelihood ratio tests for complete contingency tables. *Biometrika* **63**, 33–7.
- Williams, D.A. (1982) Extra-binomial variation in logistic linear models. *Appl. Statist.* **31**, 144–8.

- Williams, D.A. (1987) Generalized linear model diagnostics using the deviance and single-case deletions. *Appl. Statist.* **36**, 181–91.
- Williams, E.J. (1952) Use of scores for the analysis of association in contingency tables. *Biometrika* **39**, 274–80.
- Williams, E.J. (1959) *Regression Analysis*. J. Wiley & Sons, New York.
- Williams, E.J. (1962) The analysis of competition experiments. *Austral. J. Biol. Sci.* **15**, 509–25.
- Williams, O.D. and Grizzle, J.E. (1972) Analysis of contingency tables having ordered response categories. *J. Am. Statist. Assoc.* **67**, 55–63.
- Wilson, E.B. and Hilferty, M.M. (1931). The distribution of chi-square. *Proc. Nat. Acad. Sci.* **17**, 684–8.
- Wixley, R.A.J. (1987) Power transformations and the generalized linear model. PhD Thesis, University of Natal.
- Wolstenholme, D.E., O'Brien, C.M. and Nelder, J.A. (1988) GLIMPSE: A knowledge-based front end for statistical analysis. *Knowledge-based Systems* **1**, 173–8.
- Yandell, B.S. (1986) Algorithms for nonlinear generalized cross-validation. *18th Symposium on the Interface of Computer Science and Statistics*.
- Yates, F. (1948) The analysis of contingency tables with groupings based on quantitative characters. *Biometrika* **35**, 176–81; corr. **35**, 424.
- Yule, G.U. (1912) On methods of measuring association between two attributes (with discussion). *J. R. Statist. Soc.* **75**, 579–652.
- Zeger, S.L., Liang, K.Y. and Self, S.G. (1985) The analysis of binary longitudinal data with time-independent covariates. *Biometrika* **72**, 31–8.
- Zeger, S.L. and Liang, K.Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–30.
- Zeger, S.L. and Qaqish, B. (1986) Markov regression models for time-series: a quasi-likelihood approach. *Biometrics* **44**, 1019–31.
- Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988) Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**, 1049–60.

Index of data sets

Ascorbic acid concentration in snap-beans after cold storage	96
Site preferences of two species of lizard	129
Number of eggs recovered in a cannibalism experiment	143
Eye-colour of children, parents and grandparents	145
Response frequencies in a cheese-tasting experiment	175
Prevalence of pneumoconiosis among coalminers	179
Central nervous system malformations in South Wales	186
Homework conditions and work quality	189
Effect of mother's age and smoking habits on perinatal mortality	190
Frequency table of Illinois lottery numbers	191
A biological assay of two tuberculins	201
Number of damage incidents to cargo-carrying vessels	205
Breathlessness and wheeze among a group of coalminers	230
Attitude survey of New Jersey high-school seniors	239
1980 APA presidential election results (ranked data)	241
Clotting times of blood	302
Duration of the embryonic period of <i>Drosophila melanogaster</i>	307
Duration of the larval period of <i>Drosophila melanogaster</i>	316
Duration of the pupal period of <i>Drosophila melanogaster</i>	317
Yields of barley and <i>Sinapis</i> in a competition experiment	318
Incidence of <i>R. secalis</i> on barley leaves	329
Stone number and position in the Avebury ring	345
Free height of leaf springs for trucks	366
Effect of fertilizers on the yield of coastal Bermuda grass	382
Grasshopper mortality related to insecticide and synergist	385
Responses to mixtures of insulin	387
Proportion of carrots damaged in an insecticide experiment	409
Rate of growth of the fungus <i>Neurospora crassa</i>	416
Heart mass vs body mass of male cats	417
Salamander mating experiment	440

Author index

- Abramowitz, M., 314, 479
Adams, W.E., 381, 382, 498
Adena, M.A., 135, 479
Agresti, A., 182, 217, 235, 479
Aickin, M., 135, 183, 479
Aitkin, M., 17, 370, 421, 425, 430,
 451, 479
Akaike, H., 91, 93, 479
Albert, P.S., 451, 499
Andersen, E.B., 278, 479
Anderson, D.A., 347, 451, 479
Anderson, R.D., 451, 488
Andrews, D.F., 244, 415, 479
Angell, I., 344, 345, 347, 479
Anscombe, F.J., 38, 196, 221, 235,
 374, 415, 479, 480
Armitage, P., 114, 135, 159, 480
Arnold, S., 441
Ashford, J.R., 178, 180, 183, 229,
 233, 480
Ashton, W.D., 135, 480
Atkinson, A.C., 48, 90, 91, 93, 293,
 397, 406, 408, 415, 480
- Babington-Smith, B., 243, 480
Bain, L.J., 295, 480, 486
Baker, R.J., 184, 379, 480, 497
Barber, J., 344, 345, 347, 479
Barker, N.W., 300, 490
Barlow, R.E., 313, 480
Barndorff-Nielsen, O.E., 38, 44, 250,
 259, 350, 467, 478, 480, 481
Bartlett, M.S., 277, 280, 321, 467,
 468, 470, 481
Baxter, L.A., 296, 481
Beale, E.M.L., 93, 481
Beaton, A.E., 82, 481
Belsley, D.A., 415, 481
Benzécri, J.-P., 218, 236, 481
- Berk, R.H., 44, 481
Berkson, J., 14, 481
Berman, M., 347, 481
Bhapkar, V.P., 352, 481
Bhattacharya, R.N., 389, 481
Birch, M.W., 158, 482
Bishop, Y.M.M., 34, 128, 134, 135,
 183, 214, 235, 482
Björck, Å., 87, 482
Blaesild, P., 467, 480
Bliss, C.I., 13, 43, 300, 416, 417, 482,
 483
Bock, R.D., 135, 183, 235, 451, 482
Bortkewitsch, L. von, 2, 244, 482
Box, G.E.P., 8, 21, 375, 378, 379,
 389, 482
Bradley, E.L., 43, 482
Bradley, R.A., 272, 482
Breese, E.L., 321, 482
Breslow, N.E., 44, 114, 135, 259, 267,
 278, 356, 482, 483
Bross, I.D.J., 148, 188, 483
Brown, C., 230, 233, 234, 493
Burridge, J., 184, 483
Byar, D.T., 430, 483, 485
- Carroll, R.J., 485
Caussinus, H., 272, 483
Chambers, E.A., 109, 483
Chambers, J.M., 84, 93, 483
Chatfield, C., 183, 488
Chatterjee, S., 415, 483
Chen, K.K., 417, 483
Clark, V.A., 430, 489
Clarke, M.R.B., 84, 87, 483
Clayton, D.G., 17, 183, 185, 421,
 425, 430, 479, 483
Cleveland, W., 394, 483
Cochran, W.G., 96, 97, 497

- Coe, R., 303, 304, 305, 483
 Cologne, J., 259, 482
 Cook, R.D., 370, 406, 415, 483, 484
 Copas, J.B., 148, 484
 Corbeil, R.R., 248, 278, 484
 Cordeiro, G.M., 467, 484
 Corman, J.L., 381, 382, 498
 Cornish, E.A., 484
 Coutts, S.M., 296, 481
 Cox, D.R., 38, 44, 49, 95, 107, 109,
 135, 139, 153, 236, 248, 250, 259,
 278, 286, 288, 321, 350, 352, 375,
 378, 397, 415, 421, 426, 427, 430,
 450, 461, 463, 464, 478, 480-4
 Crilley, J., 204, 205
 Critchlow, D.E., 484
 Crowder, M.J., 140, 352, 484, 485
 Crowley, J., 430, 485
 Culpin, D., 347, 481
 Curnow, R.N., 321, 493
- Dahm, P.A., 183, 489
 Dale, J.R., 221, 236, 485
 Daniel, C., 132, 485
 Darby, S.C., 380, 386, 387, 485, 494
 Darroch, J.N., 183, 235, 485
 Davidian, M., 485
 Davison, A.C., 279, 403, 408, 485
 Dawson, R.B., 244, 485
 Day, N.E., 114, 135, 267, 278, 356,
 483, 485
 Deming, W.E., 183, 485
 Dempster, A.P., 44, 338, 485
 Diaconis, P., 240, 241, 485
 Dobson, A.J., 19, 485
 Draper, N.R., 48, 93, 485
 Duncan, D.B., 183, 498
 Durand, D., 295, 489
 Dyke, G.V., 14, 485
- Efron, B., 138, 147, 350, 351, 426,
 486
 Efroyimson, M.A., 91, 93, 486
 Ehrenberg, A.S.C., 183, 488
 Ekholm, A., 148, 486
 Elandt-Johnson, R.C., 430, 486
 Ellis, M.J., 380, 386, 387, 485
 Engel, J., 135, 140, 183, 486
 Engelhardt, M., 295, 480, 486
 Esseen, C.G., 475, 486
 Everitt, B.S., 135, 486
- Feigl, P., 464, 486
 Fienberg, S.E., 34, 128, 134, 135,
 182, 183, 214, 235, 482, 486
 Finney, D.J., 43, 135, 183, 486
 Firth, D., 352, 486
 Fisher, R.A., 6, 10, 11, 19, 43, 200,
 202, 204, 218, 378, 450, 484, 487
 Fleiss, J.L., 135, 487
 Foord, T.R., 344, 347, 497
 Fowlkes, E.B., 415, 487
 Francis, B., 487
 Fraser, D.A.S., 278, 487
 Freeman, P.R., 347, 487
 Freireich, E.J., 424, 487
 Furnival, G.M., 93, 487
- Galton, F., 144, 146, 487
 Gart, J.J., 107, 487
 Gauss, C.F., 1, 9, 19, 389, 487
 Gehan, E.A., 425, 487
 Gentleman, W.M., 93, 487
 Gilchrist, R., 487
 Gilmour, A.R., 451, 488
 Gilula, Z., 218, 236, 488
 Givens, W.J., 86
 Godambe, V.P., 250, 278, 340, 348,
 352, 363, 488
 Gokhale, D.V., 135, 488
 Goldstein, H., 451, 488
 Goodhardt, G.J., 183, 488
 Goodman, L.A., 215, 217, 218, 235,
 236, 488
 Gosset, W.S., 277
 Graubard, B.I., 183, 488
 Green, P.J., 43, 488
 Greenacre, M.J., 218, 236, 488
 Greenwood, J.A., 295, 489
 Gregoire, R.C., 416, 496
 Griffiths, D., 347, 481
 Grizzle, J.E., 183, 233, 489, 499
 Gross, A.J., 430, 489
 Gurland, J., 183, 489
- Haberman, S.J., 44, 117, 135, 182,
 183, 216, 217, 218, 235, 236, 488,
 489
 Hadi, A.S., 415, 483
 Haenszel, W., 267, 278, 356, 493
 Haldane, J.B.S., 139, 244, 489
 Hamilton, D., 94, 489
 Hankey, W., 278, 356, 493
 Hanson, R.J., 93, 491
 Harkness, W.L., 278, 489

- Harter, H.L., 93, 489
Harville, D.A., 247, 278, 450, 489
Hastie, T., 465, 489
Hawkins, D., 415, 489
Healy, M.J.R., 84, 93, 490
Heminway, L.N., 204, 205
Heron, D., 183, 495
Herzberg, A.M., 244, 479
Hewlett, P.S., 183, 384, 490
Heyde, C.C., 340, 348, 352, 488, 490
Hilferty, M.M., 38, 289, 499
Hill, J., 321, 482
Hill, J.R., 352, 490
Hill, M.O., 218, 490
Hinkley, D.V., 49, 248, 321, 352, 401,
 484, 490
Hoaglin, D.C., 405, 490
Holland, P.W., 34, 134, 135, 183,
 214, 235, 482
Hougaard, P., 289, 490
Householder, A.S., 86
Hu, M., 430, 485
Hurn, M.W., 300, 490
- Jarrett, R.G., 352, 490
Jeffreys, B.S., 389, 490
Jeffreys, H., 22, 389, 490
Jenkins, G.M., 21, 482
Jenkyn, J.F., 329
Jennrich, R.I., 43, 93, 490
Johansen, S., 490
Johnson, B.W., 490
Johnson, N.L., 430, 486
Joiner, B.L., 410, 496
Jørgensen, B., 43, 44, 350, 351, 490,
 491
- Kalbfleisch, J.D., 135, 278, 430, 491
Kay, R., 415, 491
Kempthorne, O., 450, 491
Kendall, M.G., 321, 477, 481, 491
Killion, R.A., 235, 491
Kirby, S.P.J., 244, 495
Kitanidis, P.K., 247, 278, 491
Knuth, D.E., xix, 491
Kolassa, J., 390, 491
Korn, E.L., 183, 488
Kruskal, W.H., 168, 454, 491
Kuh, E., 415, 481
Kullback, S., 135, 488
- Läärä, E., 185, 491
Laird, N.M., 338, 451, 485, 497
Lane, P.W., 25, 491
Lauritzen, S.L., 235, 485
Lawless, J.F., 93, 430, 491
Lawley, D.N., 478, 491
Lawrie, J.A., 416, 496
Lawson, C.L., 93, 491
Lee, A.H., 491, 492
Lee, E.T., 430, 492
Lee, I., 183, 489
Legendre, A.M., 1, 9, 19
Lehmann, E.L., 209, 248, 492
Leroy, A.M., 415, 496
Lexis, W., 125, 492
Liang, K.Y., 333, 356, 451, 482, 492,
 499
Liao, J., 283, 492
Lindsay, B., 250, 278, 492
Little, S., 415, 491
Lloyd, S., 185, 186, 492
Lowe, C.R., 185, 186, 492
- Magath, T.D., 300, 490
Mallows, C.L., 91, 93, 243, 493
Mandel, J., 236, 493
Mantel, N., 230, 233, 234, 267, 278,
 356, 493
Manton, K.G., 374, 493
Mardia, K.V., 255, 493
Marshall, R.J., 255, 493
Matthews, J.N.S., 185, 491
Maxwell, A.E., 135, 493
McCullagh, P., 45, 46, 120, 122, 151,
 153, 183, 238, 255, 257, 279, 350,
 352, 355, 390, 456, 461, 463, 471,
 472, 477, 478, 491, 492
McCulloch, R.E., 490
McKinlay, K.S., 384, 385
McLeish, D.L., 492
McNemar, Q., 278, 493
Mead, R., 321, 493
Miller, R.G., 430, 493
Moore, D.F., 493
Moore, R.H., 43, 490
Morris, C.N., 47, 493
Morse, P.M., 384, 385
Morton, R., 352, 451, 493
Mosteller, F., 93, 493
Muirhead, C.R., 494

- Nair, V.N., 365, 494
 Nelder, J.A., 16, 19, 25, 34, 43, 65,
 81, 91, 93, 291, 294, 314, 350, 378,
 379, 383, 389, 480, 491, 494, 499
 Nelson, W., 313, 494
 Newell, G., 175
 Neyman, J., 277, 494
 O'Brien, C.M., 91, 499
 Oakes, D., 135, 421, 426, 430, 484,
 494
 Palmgren, J., 148, 211, 486, 494
 Patterson, H.D., 14, 247, 278, 450,
 485, 494
 Payne, C.D., 91, 494
 Pearson, K., 169, 183, 197, 221, 495
 Peto, R., 426, 428, 495
 Phelps, K., 409, 495
 Pierce, D.A., 40, 398, 495
 Pignatiello, J.J., 365, 366, 368, 495
 Plackett, R.L., 93, 135, 140, 183,
 199, 221, 235, 490, 495
 Poisson, S.D., 2
 Powsner, L., 306, 495
 Pratt, J.W., 184, 495
 Preece, D.A., 244, 495
 Pregibon, D., 350, 357, 365, 370,
 375, 376, 377, 389, 393, 398, 415,
 479, 492, 494-6
 Prentice, R.L., 135, 362, 430, 451,
 491, 496
 Pringle, R.M., 82, 496
 Proschan, F., 313, 480
 Qaqish, B., 499
 Quine, M.P., 244, 496
 Rae, A.L., 451, 488
 Ramberg, J.S., 365, 366, 368, 495
 Rao, C.R., 48, 393, 496
 Rao, R.R., 389, 481
 Ratcliff, D., 183, 485
 Rayner, A.A., 82, 496
 Reid, N., 250, 484
 Robbins, E.B., 417, 483
 Roberts, C.J., 185, 186, 492
 Rogers, C.E., 56, 93, 498
 Ross, G.A.F., 296, 481
 Ross, G.J.S., 244, 495
 Ross, W.H., 467, 496
 Rousseeuw, P.J., 415, 496
 Rubin, D.B., 338, 485
 Rubin, G.M., 313, 496
 Ryan, B.F., 410, 496
 Ryan, T.A., 410, 496
 Schafer, D.W., 40, 398, 495
 Schoener, T.W., 128, 496
 Scholten, H., 272, 496
 Schreiner, H.R., 416, 496
 Scott, E.L., 277, 494
 Searle, S.R., 93, 248, 278, 484, 496
 Seber, G.A.F., 48, 93, 496
 Self, S.G., 451, 499
 Seneta, E., 244, 496
 Simon, G., 183, 496
 Simpson, E.H., 496
 Singhal, K., 93, 491
 Skovgaard, I.M., 321, 496
 Small, C.G., 492
 Smith, H., 48, 93, 485
 Smyth, G.K., 496
 Snedecor, G.W., 96, 97, 497
 Snell, E.J., 38, 95, 183, 286, 397, 415,
 464, 484, 497
 Solomon, H., 239, 497
 Sowden, R.R., 229, 233, 480
 Speed, T.P., 235, 485
 Sprent, P., 93, 497
 Sprott, D.A., 278, 491
 Spurr, D.T., 384, 385
 Stallard, E., 374, 493
 Stegun, I.A., 314, 479
 Stephan, F.F., 183, 485
 Stern, R.D., 303, 304, 305, 483
 Stevens, S.S., 183, 497
 Stewart, G.W., 93, 497
 Stigler, S.M., 1, 19, 125, 497
 Stratelli, R., 451, 497
 Stone, M., 93, 168, 497
 Streibig, J.C., 317, 320
 Stuart, A., 477, 491
 Teleky, S., 143
 Terry, M.A., 272, 482
 Thisted, R.A., 93, 497
 Thom, A., 344, 347, 497
 Thom, A.S., 344, 347, 497
 Thompson, M.E., 363, 488
 Thompson, R., 184, 247, 278, 450,
 494, 497
 Tibshirani, R., 255, 465, 489, 492
 Tidwell, P.W., 379, 389, 482

- Tsai, C.-L., 352, 403, 408, 485, 490
Tukey, J.W., 8, 93, 415, 480, 493,
 497
Tweedie, M.C.K., 498
- Upton, G.J.G., 135, 235, 272, 498
Urquhart, N.S., 80, 498
- van Wissen, L., 272
Verrell, P., 441
Vleeshouwers, L., 317, 320
- Wahrendorf, J., 221, 498
Walker, S.H., 183, 498
Wampler, J.H., 93, 498
Wang, P.C., 498
Ware, J.H., 451, 497
Wedderburn, R.W.M., 9, 18, 19, 34,
 38, 43, 117, 291, 328, 329, 352,
 360, 378, 494, 498
Weeks, D.L., 80, 498
Weisberg, S., 370, 415, 483, 484
Welch, L.F., 381, 382, 498
Welsch, R.E., 405, 415, 481, 490
- Wermuth, N., 190, 498
West, M., 498
Whitehead, J., 17, 426, 429, 430, 498
Whittaker, J., 487
Wichura, M., xix, 498
Wilkinson, G.N., 56, 93, 498
Williams, D.A., 93, 140, 398, 408,
 409, 415, 467, 498
Williams, E.J., 218, 321, 498, 499
Williams, O.D., 183, 499
Wilson, E.B., 38, 289, 499
Wilson, R.W., 93, 487
Wilson, S.R., 135, 479
Wixley, R.A.J., 499
Wolstenholme, D.E., 91, 499
Woodbury, M.A., 374, 493
- Yandell, B.S., 499
Yates, F., 10, 159, 187, 188, 450, 499
Yule, G.U., 183, 499
- Zahn, D.A., 235, 491
Zeger, S.L., 333, 451, 492, 499
Zelen, M., 464, 486
Zweifel, J.R., 107, 487

Subject index

- * operator, 58
- ** operator, 60
- + operator, 57
- operator, 59
- / operator, 59
- . operator, 57
- / operator, 58
- G^2 statistic, 34
- L_p -norm, 5
- χ^2 approximation for deviance, 119
- χ^2 distribution, 34

- Absolute error, 104
- Accelerated testing, 123
- Accident-proneness, 206
- Accidentally empty cell, 80
- Acsrine transformation, 137
- Added-variable plot, 399
- Additivity, 22, 23
- Adjusted dependent variable, 40, 116
- Algorithm
 - EM, 338, 353
 - for fitting GLM, 40–3, 81, 465
 - Givens, 86
 - Householder, 86
 - iterative proportional scaling, 183, 214, 238
 - QR, 88
 - Yates's, 131, 238
- Aliasing, 54, 61–8, 157, 367
 - extrinsic, 68
 - intrinsic, 54, 63–7, 157
- All subsets regression, 91
- Analysis of covariance, 25
- Analysis of deviance, 35, 36
- Analysis of variance, 1, 11, 35
- Ancillary statistic, 130
- Angular transformation, 137
- Anomalous value, 37, 144

- Anscombe residual, 37, 38
- Armitage's test for trend, 159
- Arrhenius model, 309, 315
- Asymptote, 16, 309
- Autoregressive model, 21, 22, 340

- Back-fitting algorithm, 466
- Backward elimination, 91, 403
- Bartlett factor, 350, 362, 459, 478
- Bartlett identity, 470
- Bayes's theorem, 113, 138
- Beta distribution, 46
- Beta-binomial distribution, 126, 140, 183
- Bias adjustment, 119, 147, 455
- Bilinear model, 51, 218
- Binary data, 98, 262, 439
- Binomial deviance, 118
- Binomial distribution, 2, 30, 31, 101–7, 475
- Bioassay, 13, 14, 200
- Bivariate binary regression, 238
- Bivariate logit transformation, 240
- Bivariate probit model, 233
- Boundary point, 218
- Box-Cox transformation, 375
- Bradley-Terry model, 272
- Butler effect, 243

- Calibration, 25
- Canonical correlation, 217–8
- Canonical link, 30, 32, 291
- Canonical parameter, 334, 476
- Case deletion, 396
- Case-control study, 267
- Cauchy density, 20
- Causal path, 215
- Censoring, 17, 230, 234, 419

- Checking
 the link function, 401
 the covariates, 401
 the variance function, 400
 other departures, 398, 403
- Choleski decomposition, 82, 84, 88
- Cluster sampling, 125, 140, 165
- Clustered Poisson process, 198
- Coefficient of variation, 16, 17, 285, 355, 433
- Collinearity, 84
- Combination of information, 265, 340-2
- Competition experiment, 292, 317
- Complementary log-log, 12, 31, 108, 151, 153, 188, 410
- Completeness of statistic, 248
- Component of dispersion, 204, 432
- Component of variance, 49, 439
- Composite link, 81, 184, 220, 225
- Conditional
 independence, 215
 inference, 209
 likelihood, 210, 282, 248
 linear model, 18
 mean, 209
 score statistic, 249
 sufficiency, 253
 variance, 209
- Consistency, 404, 406
- Constraint, 64-67
- Constructed variable, 376, 393, 402
- Continuation-ratio model, 163, 180-2
- Continuity correction, 104, 106, 267, 474
- Continuous proportion, 328
- Contrast, 220, 247
- Controls, 267
- Convergence of cumulants, 104
- Convergence of algorithm, 41, 117
- Cook statistic, 406
- Cornish-Fisher expansion, 477
- Correspondence analysis, 218
- Covariance function, 247, 255, 280, 324
- Covariate, 3, 23, 99
- Coverage probability, 471
- Cox model, 421, 426
- Cube root transform, 198, 289
- Cumulant function, 30, 46, 137, 195, 258, 334
- Cumulants
 of binomial, 102
 of gamma distribution, 287
 of logistic density, 142
 of multinomial, 165
- Cumulative hazard function, 422
- Cut-point, 156, 185
- Decomposability, 216, 221, 235
- Defining contrast, 366
- Deletion residual, 396
- Dependent observations, 332
- Deviance function, 33-6, 118, 174
 for binary data, 121
 for binomial, 118
 reduction, 119
 residual, 37, 39
 χ^2 approx for, 119, 122
- Direct decomposition, 85
- Direct product matrix, 238
- Dirichlet distribution, 314
- Discrete Edgeworth series, 106, 474
- Discrimination among models, 293
- Discrimination, logistic, 138, 189
- Dispersion component, 432-50
- Dispersion effect, 222
- Dispersion parameter, 29, 30, 125
 estimation of, 295, 448
- Dot operator, 56
- Double exponential distribution, 19
- Dummy variable, 54
- Edgeworth series, 106, 167, 474
- Effective sample size, 350
- Efficiency, 232
 of linear least squares, 321
- Eigenvalue, 200, 327
- Ellipse fitting, 347
- EM algorithm, 338, 353
- Empirical logistic transformation, 106, 107
- Empty cell, 79, 80, 205
- Empty model formula, 269
- Erlangian distribution, 288
- Estimable function, 64
- Estimating function, 250, 254, 274, 339
- Euler-Maclaurin formula, 389
- Exponential decay model, 96
- Exponential distribution, 423
 relation to logistic, Weibull, 20

- Exponential
 family, 19, 27, 28, 43, 46, 252, 476
 operator, 60
 weighting, 44, 210, 257, 428
- Extended quasi-likelihood, 349, 360, 400
- Extrapolation, 16, 122
- Extreme points, 408
- Extreme-value density, 142, 154, 424
- Extrinsic aliasing, 68
- F-distribution, 20, 47
- Factor level, 9, 52
- Factorial contrast, 224
- Factorial design, 10, 14, 381
- Factorial moment, 256
- False positive, 90
- Fieller-Creasy problem, 123, 250, 284, 342
- Finite population correction, 192
- Fisher information, 6, 43, 73, 77, 119, 327, 470
- Fisher's scoring method, 42
- Fitted value, 4, 23, 71
- Forward selection, 91, 403
- Fourier series, 305
- Frequentist theory, 323
- Fused cells, 81
- Gamma deviance, 290
- Gamma distribution, 2, 3, 30, 199, 237, 287–9, 374
- Gaussian distribution, 50
- Gaussian elimination, 82
- Generalized additive model, 465
- Generalized inverse, 82, 168
- Genstat, 57, 60
- Givens rotation, 86
- Glim, 57, 60, 67, 91, 93
- Goodness of fit, 23, 24, 36, 118
- Goodness-of-link test, 375
- Graphical model, 235
- Gravity model, 272
- Green's matrix, 168, 275
- Growth curve, 16
- Half-Normal plot, 132, 407
- Hat matrix, 405
- Hazard function, 420
- Hedonic scale, 175
- Hessian matrix, 6, 42
- Householder reflection, 86
- Hyperbolic model, 16, 300
- Hyperbolic secant density, 47
- Hypergeometric distribution, 136, 255–8, 283, 353
- Incidental parameter, 245, 342
- Independence in two-way table, 187
- Index plot, 409
- Indicator vector, 54
- Influence, 397, 404, 406
- Information matrix, see Fisher information
- Instantaneous risk, 420
- Interaction, 53, 58
 without main effect, 96
- Interval scale, 150, 155
- Intrinsic aliasing, 54, 63, 157
- Inverse Gaussian distribution, 30, 38
- Inverse linear model, 291, 381
- Inverse polynomial, 16, 291, 321
- Inversion in permutation, 242
- Isolated discrepancy, 392, 408
- Iterative proportional scaling, 183, 214, 238
- Jacobi matrix, 168
- k*-statistic, 257
- Kurtosis, 361
- Laplace distribution, 19
- Large deviation, 137
- Latin square, 202, 434
- Lattice distribution, 474
- Law of small numbers, 244
- LD50, 25
- Least median of squares, 415
- Least squares, 1, 25
- Leverage, 404, 405
- Lexian dispersion theory, 125
- Likelihood, 6, 24, 28, 114, 171, 194, 323
 conditional, 245, 248
 equation, 225
 for binomial data, 114
 for gamma data, 289–296
 for hypergeometric data, 255–267
 for multinomial response, 171
 for multiple responses, 225
 for Poisson response, 194–8
 marginal, 246
 quasi—, 323–7
 restricted 248, 278, 439
- Likelihood-ratio statistic, 471–8

- Line integral, 334, 353
 Linear contrast, 158
 Linear estimating function, 274, 328,
 347
 Linear model, 294, 433
 Linear predictor, 27, 31, 32
 Linear \times linear interaction, 158
 Link function, 27–32, 107–110
 Log concavity of link, 117
 Log link, 31
 Log odds, 14, 106
 Log-bilinear model, 218
 Log-linear model, 1, 14, 193, 209,
 223
 Log-log function, 108
 Log-Normal distribution, 454
 Log-rank test, 188
 Logistic density, 20, 46, 141
 cumulants of, 142
 Logistic discrimination, 138, 189
 Logit model, 1, 14, 31, 106–114
 Longitudinal study, 229, 333
- Mahalanobis's distance, 405
 Mantel-Haenszel
 estimator, 267, 356
 test, 267
 Marginal distribution, 170
 Marginal likelihood, 210, 282, 246
 Marginality constraint, 64, 89, 242
 Marginalization, 225
 Markov model, 302–3
 Markov property, 170, 216
 Martingale, 340
 Matched design, 95
 Matched pair, 270
 Mean-value parameter, 24
 Measurement scale, 150
 Measure of association, 182
 Method of moments, 115, 173
 Misclassification, effect of, 148
 Michaelis–Menten model, 16
 Migration study, 272
 Minitab, 93, 410
 Mixture distribution, 199
 Mixture of drugs, 386
- Model
 checking, 391, 392
 choice, 235
 for binary response, 107
 for circle, 343
 for ellipse, 347
 for interval scale, 155
 for nominal scale, 159
 for ordinal scale, 151
 for polytomous data, 149
 for survival data, 419
 formula, 56, 358
 matrix, 10, 26
 selection, 21, 23
- Modified maximum likelihood, 248
 Modified profile likelihood, 250
 Moment estimator, 296
 Moment generating function, 44
 Moments and cumulants, 102, 165
 Monoculture yield, 317
 Moore-Penrose inverse, 168, 282
 Multinomial
 distribution, 15, 164–7, 209
 covariance matrix, 168
 cumulants of, 165
 deviance function, 174
 likelihood, 171–3
 quadratic form in, 169
 relation to Poisson, 210
 response model, 1, 211
- Multiplicative effect, 31, 292
 Multivariate hypergeometric, 260–1
 Multivariate logit link, 220
 Multivariate regression model,
 219–222
- Negative binomial distribution, 199,
 237, 373
 Negative dispersion component, 449
 Nested response scale, 160
 Nesting operator, 58
 Noise-to-signal ratio, 358
 Nominal scale, 150
 Non-canonical model, 457
 Non-central
 χ^2 distribution, 279
 hypergeometric, 427, 257, 353
 Non-linear model, 154, 434, 379

- Normal**
 distribution, 1, 26, 28
 equation, 81
 limit, 103
 order statistic, 407
 plot, 407
- Normalizing transformation, 38, 477
- Nuisance parameter**, 245, 342, 472, 476
- Observational study**, 3, 98
- Observed information**, 342
- Odds ratio, 136, 356
- Offset, 12, 206, 423
- Operator**
 $\star\star$, 60
 \star , 58
 $+$, 57
 $-$, 59
 $-/-$, 59
 \cdot , 57
 $/$, 58
 nesting, 58
 for removal of terms, 59
- Ordinal response, 17, 150–1, 273
- Orthogonal matrix, 281
- Orthogonal parameter, 46, 211, 229, 250, 351, 413
- Orthogonal polynomial, 53, 94, 97
- Orthogonal projection, 72
- Outlier, 37, 144
- Over-dispersion, 90, 103, 124, 135–7, 174, 198
 for binomial, 124–8, 137
 for multinomial, 174–5
- Paired comparison, 272
- Parameter interpretation, 110, 233
- Partial likelihood, 426, 426
- Partial ordering, 348
- Partitioned matrix, 281, 472
- Path model, 215
- Path-independent integral, 334
- Pearson residual, 37
- Pearson's statistic, 34, 37, 127
 cumulants of, 169, 244
- Pearson-Plackett family, 221
- Permutation, model for, 96, 240–2, 272
- Peto's method for ties, 430
- Pivotal statistic, 46, 84, 339
- Plant density experiment, 291
- Poisson**
 distribution, 2, 11, 22, 30, 101, 194
 deviance, 197
 distribution truncated, 45
 distribution, cumulants of, 195
 limit, 137, 105
 log-likelihood function, 197
 mixture, 237
 process, 193, 288
 relation to multinomial, 210
- Polychoric correlation coefficient, 183
- Polykay**, 257
- Polynomial regression, 69
- Polytomous data, 270
- Posterior odds, 138, 189
- Power family, 31
- Precision matrix, 348
- Prediction, 7, 21, 25
- Prior information, 323
- Prior odds, 138, 189
- Prior weight, 29
- Probit model, 1, 13, 31, 108
- Profile deviance, 264
- Profile likelihood, 254
- Projection matrix, 247, 397
- Proportional-hazards model, 153, 421, 421
- Proportional-odds model, 153, 188, 213
- Prospective study, 111
- Pseudo-aliasing, 88
- Pythagorean relationship for deviances, 147
- QR algorithm, 85–8
- Quadratic contrast, 131
- Quadratic form, 169
- Qualitative covariate, 9, 52
- Quality-improvement, 357
- Quantitative covariate, 9
- Quasi-likelihood, 9, 323, 332–4
 deviance function, 327
 score function, 327
- Quasi-symmetry, 95, 272
- Quota sampling, 165
- REML, 248, 278, 363, 439, 450, 451
- Random effect, 3, 452
- Randomization, 99, 263
- Randomized blocks, 94
- Rank-deficient matrix, 82
- Ranked data, model for, 240–22, 272
- Ratio, confidence interval for, 251–2

- Rational model, 154, 177
Regularity condition, 218
Relative potency, 202–4, 237
Reliability experiment, 123
Residual, 9, 37, 396
Response transformation, 23
Restricted likelihood (REML), 248, 278, 363, 439, 450, 451
Retrospective study, 111, 278
Ridit score, 188
Risk set, 163
Robustness, 246, 352
Rounding error, 85, 296, 374, 389
- S (computer program), 93
Saddlepoint approximation, 350
Safe dose, 123
Sampling fraction, 256
Sampling with replacement, 428
Scale for analysis, 22
Scaled deviance, 24, 34, 46
Score statistic, 267, 470
Score test, 188, 393
Scores in log-linear model, 150–8
Scoring method, 43
Selection bias, 129
Selection effect, 132, 230, 284, 393
Selection of covariates, 23, 89
Self-sterile variety, 80
Sheppard correction, 106, 375, 390, 475, 478
Signal-to-noise ratio, 357
Signed deviance statistic, 105, 237, 476
Simpson's paradox, 234
Simpson's rule, 45
Simulation, 178
Singular information matrix, 84
Singular model formula, 217
Singular-value decomposition, 85, 218, 282
Skewness array, 120
Small numbers, law of, 244
Smoothing of scatterplot, 394, 466
Span of smoother, 465
Sparseness, 120–2, 169, 244
Spatial process, 21, 247, 281
Split-plot design, 22
Square root transformation, 236
Square-root matrix, 86
- Standardized cumulant, 351
Stepwise regression, 91
Structural zero (empty cell), 79–80
Studentized standardized residual, 396
Sufficient statistic, 32, 115–6, 209, 252, 476
Survivor function, 420
Sweep operator, 82–5, 95
Symmetric constraint, 66
Symmetric function, 238
Symmetrizing transformation, 138, 236
Synergism, 386
- Tail probability, 104
Tetrachoric correlation, 183
Tied failure times, 427, 427
Time series, 21, 25
Time-dependent covariate, 17, 419
Time-ordered sequence, 340
Tolerance distribution, 153
Transformation of response, 23, 378
Transformation to symmetry, 105, 196
Transition probability, 336
Tri-diagonal matrix, 168
Triangular arrangement of factors, 95
Truncated Poisson distribution, 45
- Unbiased estimating function, 250, 284
Under-dispersion, 126, 194
Uniform multinomial, 165
Uniqueness of estimate, 117, 184
- Variance component, 18, 294, 359, 432, 439
Variance function, 14, 29, 30, 138, 289
Variance-stabilizing transform, 137, 138, 196, 236, 293
- Weibull density, 20, 423
Wilcoxon's statistic, 188
Wishart matrix, 218
- Yates's algorithm, 131, 238
Yates's statistic, 187–9