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

$$\text{var}(Y_i) = \sigma^2 E(Y_i), \tag{6.1}$$

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

$$\log \mu_i = \eta_i = \boldsymbol{\beta}^T \mathbf{x}_i; \qquad i = 1, \ldots, n. \tag{6.2}$$

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

$$\text{pr}(Y = y) = e^{-\mu}\mu^y/y!; \qquad y = 0, 1, 2, \ldots,$$

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 $\mu \to \infty$

$$(Y - \mu)/\mu^{1/2} \sim N(0,1) + O_p(\mu^{-1/2}).$$

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.

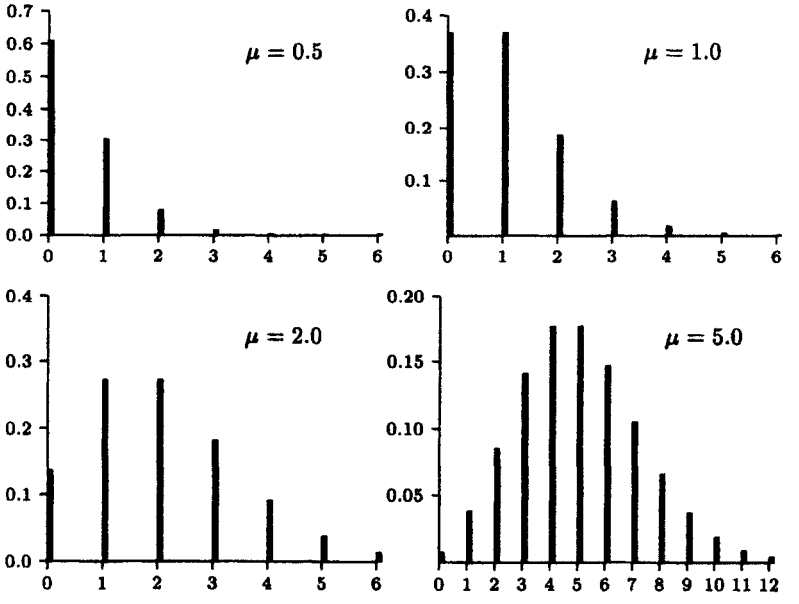

Fig. 6.1. *The Poisson distribution for* $\mu = 0.5, 1.0, 2.0$ *and* 5.0.

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

$$E(Y^{1/2}) \simeq \mu^{1/2} \qquad \text{and} \qquad \text{var}(Y^{1/2}) \simeq 1/4,$$

the error terms being $O(\mu^{-1})$. In fact the subsequent terms in an asymptotic expansion are

$$E(Y^{1/2}) \simeq \mu^{1/2}\{1 - 1/(8\mu)\}$$
$$\text{var}(Y^{1/2}) \simeq \{1 + 3/(8\mu)\}/4,$$

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

$$E(Y^{2/3}) \simeq \mu^{2/3}\{1 - 1/(9\mu)\}$$
$$\text{var}(Y^{2/3}) \simeq \mu^{1/3}\tfrac{4}{9}\{1 + 1/(6\mu)\}$$
$$\kappa_3(Y^{2/3}) \simeq O(\mu^{-1}).$$

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

$$\text{pr}(Y \geq y) \simeq 1 - \Phi(g(y - \tfrac{1}{2})),$$

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

$$l(\boldsymbol{\mu}, \mathbf{y}) = \sum (y_i \log \mu_i - \mu_i), \tag{6.3}$$

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

$$D(\mathbf{y}; \boldsymbol{\mu}) \simeq \textstyle\sum_i (y_i - \mu_i)^2/\mu_i,$$

which is less accurate than the quadratic approximation on the $\mu^{1/3}$ scale. This statistic is due to Pearson (1900).

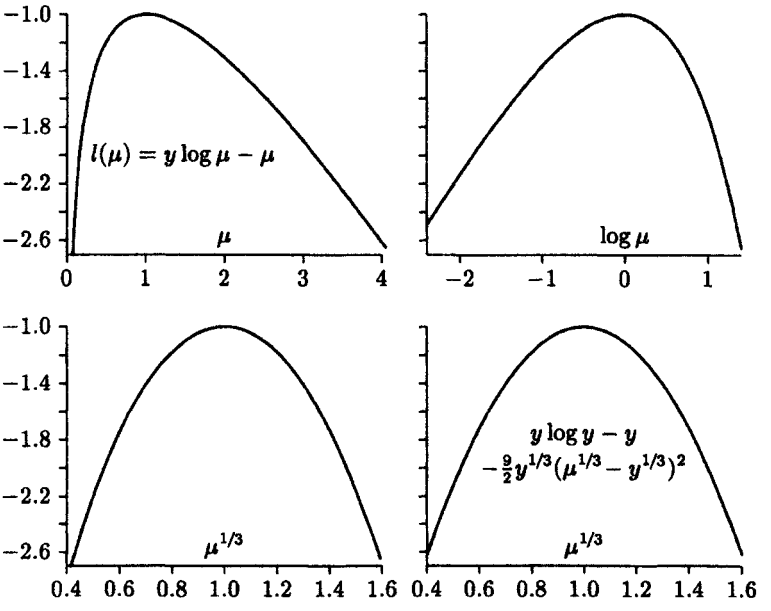

Fig. 6.2. *The Poisson log likelihood function for $y = 1$, using scales μ, $\log \mu$ and $\mu^{1/3}$, together with quadratic approximation on cube root scale.*

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 $Y = Z_1 + Z_2 + \ldots + Z_N$ where the Zs are independent and identically distributed and N has a Poisson distribution independent of $\mathbf{Z}$. We find that

$$E(Y) = E(N)E(Z)$$

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

$$\text{pr}(Y = y; \mu, \phi) = \frac{\Gamma(y + \phi\mu)\phi^{\phi\mu}}{y!\,\Gamma(\phi\mu)(1+\phi)^{y+\phi\mu}}; \qquad y = 0, 1, 2, \ldots$$

(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 $\hat{\boldsymbol{\beta}}$. For details see Chapter 9.

6.2.4 *Asymptotic theory*

The usual asymptotic results concerning consistency and asymptotic Normality of $\hat{\boldsymbol{\beta}}$ are valid provided that the eigenvalues of the information matrix increase without limit. This condition is usually satisfied if p is fixed and $n \to \infty$ or, for fixed n and p, if $\mu_i \to \infty$ for each i. The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is $\sigma^2 \mathbf{i}_\beta^{-1}$ where $\mathbf{i}_\beta$, 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

$$\tilde{\sigma}^2 = X^2/(n-p) = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \Big/ (n-p). \qquad (6.4)$$

The effective degrees of freedom for $\tilde{\sigma}^2$ are given by $f = (n-p)/(1+\frac{1}{2}\bar{\rho}_4)$, where $\bar{\rho}_4$ is the average standardized fourth cumulant of Y_i. Approximate confidence limits for individual components of $\boldsymbol{\beta}$ are then based on the t_f distribution if σ^2 is unknown. This is a minor refinement and for most purposes, unless many of the fitted 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*				
	I	II	III	IV	*Total*
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

$$site + class + volume + tuberculin$$

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(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.

		Equation (6.3)	*Fisher*
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:

Contrast			*Estimate*	*SE*	*Correlation*
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

$$site + class + volume.tuberculin.$$

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*

Ship type	*Year of construction*	*Period of operation*	*Aggregate months service*	*Number of damage incidents*
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:

$$\begin{aligned}\log(&\text{expected number of damage incidents})\\ &= \beta_0 + \log(\text{aggregate months service})\\ &\quad + (\text{effect due to ship type})\\ &\quad + (\text{effect due to year of construction})\\ &\quad + (\text{effect due to service period}).\end{aligned} \tag{6.5}$$

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 $(y - \hat{\mu})/(\tilde{\sigma}\hat{\mu}^{1/2})$, with $\tilde{\sigma}^2 = 1.69$. 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(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*

Parameter		*Estimate*	*Standard error*
Intercept		−6.41	—
Ship type	A	0.00	—
	B	−0.54	0.23
	C	−0.69	0.43
	D	−0.08	0.38
	E	0.33	0.31
Year of construction	1960–64	0.00	—
	1965–69	0.70	0.19
	1970–74	0.82	0.22
	1975–79	0.45	0.30
Service period	1960–74	0.00	—
	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*

Ship type	*Year of construction*			
	1960–64	1965–69	1970–74	1975–79
A	0.0	3.2	4.9	4.9
B	1.1	2.3	2.8	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 $Y_1, \ldots, Y_k$ are independent Poisson random variables with means $\mu_1, \ldots, \mu_k$ and that we require to test the composite null hypothesis $H_0: \mu_1 = \ldots = \mu_k = e^{\beta_0}$. 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 $(k^{-1}, \ldots, k^{-1})$. This conditional distribution is independent of the nuisance parameter β_0 so that the one-sided significance level for alternatives $\beta_1 > 0$, namely $p^+ = \text{pr}(T \geq t_{obs}; H_0)$, 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

$$E(T) = \sum x_j e^{\beta_0} \simeq \sum x_j y_\cdot / k$$
$$\text{var}(T) = \sum x_j^2 e^{\beta_0} \simeq \sum x_j^2 y_\cdot / k,$$

which depend on β_0. The conditional moments on the other hand are

$$E(T \mid Y_\cdot) = \sum x_j y_\cdot / k$$
$$\text{var}(T \mid Y_\cdot) = \sum (x_j - \bar{x})^2 y_\cdot / k.$$

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 $x_1, \ldots, x_k$. 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

$$\begin{aligned} l_Y(\tau, \beta_1) &= y_\cdot \log \tau - \tau + \beta_1 \sum_j x_j y_j - m \log\{\textstyle\sum \exp(\beta_1 x_j)\} \\ &= l_m(\tau; m) + l_{Y|m}(\beta_1; y). \end{aligned}$$

The first term above is the Poisson log likelihood for τ based on $m = Y_\cdot \sim P(\tau)$. The second component is the multinomial log likelihood for β_1 based on the conditional distribution,

$$Y_1, \ldots, Y_k \,|\, Y_\cdot = m \;\sim\; M(m, \boldsymbol{\pi})$$

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 $Y_\cdot$ depends only on τ whereas the conditional likelihood given $Y_\cdot$ 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 $Y_\cdot$.

The Fisher information matrix for (τ, β_1) is

$$i_{\tau\beta} = \operatorname{diag}\Big\{1/\tau,\ \sum \pi_j (x_j - \bar{x})^2\Big\}$$

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

$$\log \mu_{ij} = \phi_i + \mathbf{x}_{ij}^T \boldsymbol{\beta} \tag{6.6}$$

where $\mu_{ij} = E(Y_{ij})$, $\mathbf{x}_{ij}$ are known p-dimensional vectors, $\boldsymbol{\beta}$ is the parameter of interest and $\phi_1, \ldots, \phi_n$ are incidental parameters. Under this model the dimension of the parameter space, $n + p$, increases as $n \to \infty$ for fixed $\mathbf{p}$. Consequently maximum-likelihood estimates cannot be guaranteed to be efficient or even consistent in the limit as $n \to \infty$. On the other hand the conditional log likelihood derived below depends only on $\boldsymbol{\beta}$ and not on $\boldsymbol{\phi}$ 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 ith 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 $(\boldsymbol{\tau}, \boldsymbol{\beta})$, can be written in the form

$$\begin{aligned} l_Y(\boldsymbol{\tau}, \boldsymbol{\beta}) &= \sum_i (m_i \log \tau_i - \tau_i) \\ &\quad + \sum_i \Big\{ \textstyle\sum_j y_{ij} \mathbf{x}_{ij}^T \boldsymbol{\beta} - m_i \log\Big(\sum_j \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta}) \Big) \Big\} \\ &= l_m(\boldsymbol{\tau}; m) + l_{Y|m}(\boldsymbol{\beta}; y). \end{aligned}$$

The first term above is the Poisson log likelihood for τ based on the row totals $Y_{i\cdot} \sim P(\tau_i)$. The second term is the conditional log likelihood given $\{Y_{i\cdot}\}$, which depends only on $\boldsymbol{\beta}$ and not on the incidental parameters. All the information concerning $\boldsymbol{\beta}$ resides on the second component. In particular, it is apparent that $\hat{\boldsymbol{\beta}}$ and $\operatorname{cov}(\hat{\boldsymbol{\beta}})$ based on $l_{Y|m}(\boldsymbol{\beta}; y)$ 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

$$\pi_{ij} = \frac{\exp(\mathbf{x}_{ij}^T \boldsymbol{\beta})}{\sum_j \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta})} . \tag{6.7}$$

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 $\boldsymbol{\beta}$. 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

$$H.D.S.T \ + \ R.(H + D + S + T). \tag{6.8}$$

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*

	Side-effect	
Target effect	*Severity*	*Duration*
complete cure	none	temporary
partial cure	mild	permanent
no improvement	moderate	
	severe	

In an experiment where several responses $A, B, C, \ldots$ 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 $A, B, C, \ldots$ 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

$$A*B + B*C \qquad \text{and} \qquad A*B + B*C + C*A.$$

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 $A, B, \ldots$ 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 $s_1, s_2, \ldots$ and $t_1, t_2, \ldots$ are available for the response categories of A and B respectively, we may consider the following models, which are intermediate between $A + B$ and $A*B$.

Model formula	*Algebraic expression*
$A + B + s.t$	$\alpha_i + \beta_j + \gamma(s_i t_j)$
$A + B + A.t$	$\alpha_i + \beta_j + \gamma_i t_j$
$A + B + A.t + B.s$	$\alpha_i + \beta_j + \gamma_i t_j + \delta_j s_i$

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

$$\eta_{ij} = \log \mu_{ij} = \alpha_i + \beta_j + \rho \epsilon_i \delta_j, \tag{6.9}$$

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

$$A_\epsilon = \hat{\epsilon}_i \qquad \text{if} \qquad A = i$$

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 $A, B, C, \ldots$, 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 linear transformation from π_{ijk} to new probabilities γ_{ijk} given by

$$\boldsymbol{\gamma} = \mathbf{L}\boldsymbol{\pi}, \tag{6.10}$$

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

$$\boldsymbol{\gamma} = \begin{pmatrix} \pi_{i\cdot\cdot} \\ \pi_{\cdot j\cdot} \\ \pi_{\cdot\cdot k} \\ \pi_{ij\cdot} \\ \pi_{i\cdot k} \\ \pi_{\cdot jk} \\ \pi_{ijk} \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right\} \text{univariate marginal probabilities} \\ \left.\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right\} \text{bivariate marginal probabilities} \\ \text{trivariate marginal probabilities.} \end{matrix}$$

Thus in the $2\times2\times2$ 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 $\boldsymbol{\gamma}$ 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

$$\boldsymbol{\eta} = \mathbf{C} \log \boldsymbol{\gamma}, \tag{6.11}$$

for an appropriately chosen contrast matrix $\mathbf{C}$. For example, in the $2\times2\times2$ example discussed earlier, we may take $\boldsymbol{\eta}$ to be the vector of logistic factorial contrasts, namely

$$\begin{aligned}
\eta_a &= \log\pi_{1\cdot\cdot} - \log\pi_{2\cdot\cdot} \\
\eta_b &= \log\pi_{\cdot1\cdot} - \log\pi_{\cdot2\cdot} \\
\eta_c &= \log\pi_{\cdot\cdot1} - \log\pi_{\cdot\cdot2} \\
\eta_{ab} &= \log\pi_{11\cdot} - \log\pi_{12\cdot} - \log\pi_{21\cdot} + \log\pi_{22\cdot} \\
\eta_{ac} &= \log\pi_{1\cdot1} - \log\pi_{1\cdot2} - \log\pi_{2\cdot1} + \log\pi_{2\cdot2} \\
\eta_{bc} &= \log\pi_{\cdot11} - \log\pi_{\cdot12} - \log\pi_{\cdot21} + \log\pi_{\cdot22} \\
\eta_{abc} &= \log\pi_{111} - \log\pi_{121} - \log\pi_{211} + \log\pi_{221} \\
&\quad - \log\pi_{112} + \log\pi_{122} + \log\pi_{212} - \log\pi_{222}
\end{aligned}$$

univariate contrasts; bivariate contrasts; trivariate contrast

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:

$$\eta_a(\mathbf{x}) = \boldsymbol{\beta}_a^T\mathbf{x}, \qquad \eta_b(\mathbf{x}) = \boldsymbol{\beta}_b^T\mathbf{x}, \qquad \eta_c(\mathbf{x}) = \boldsymbol{\beta}_c^T\mathbf{x} \tag{6.12}$$

$$\eta_{ab}(\mathbf{x}) = \eta_{ac}(\mathbf{x}) = \eta_{bc}(\mathbf{x}) = \eta_{abc}(\mathbf{x}) = 0. \tag{6.13}$$

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

$$\eta_{ab}(\mathbf{x}) = \eta_{ab} \qquad \eta_{ac}(\mathbf{x}) = \eta_{ac} \qquad \eta_{bc}(\mathbf{x}) = \eta_{bc}$$
$$\eta_{abc}(\mathbf{x}) = \eta_{abc}. \tag{6.14}$$

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

$$\eta_{ai} = \operatorname{logit} \gamma_{i\cdot} = \operatorname{logit} \operatorname{pr}(A \le i)$$
$$\eta_{bj} = \operatorname{logit} \gamma_{\cdot j} = \operatorname{logit} \operatorname{pr}(B \le j)$$
$$\eta_{abij} = \log \gamma_{ij} - \log(\gamma_{i\cdot} - \gamma_{ij}) - \log(\gamma_{\cdot j} - \gamma_{ij}) + \log \bar{\gamma}_{ij}$$

where

$$\gamma_{i\cdot} = \operatorname{pr}(A \le i), \qquad \gamma_{ij} = \operatorname{pr}(A \le i, B \le j),$$
$$\gamma_{\cdot j} = \operatorname{pr}(B \le j), \qquad \bar{\gamma}_{ij} = \operatorname{pr}(A > i, B > j).$$

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 $k_1, \ldots, k_r$ levels respectively, we have defined $k_1 \ldots k_r - 1$ 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

$$A: \mathbf{x}; \qquad B: \mathbf{x}; \qquad C: \mathbf{x}.$$

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

$$(A; B; C): \mathbf{x}; \qquad (A*B*C): 1.$$

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

$$\begin{aligned} \eta_{a\,i} &= \operatorname{logit} \gamma_{i\cdot} = \theta_i + \beta_a x \\ \eta_{b\,j} &= \operatorname{logit} \gamma_{\cdot j} = \phi_j + \zeta_j z \\ \eta_{ab\,ij} &= \eta_{ab} \end{aligned}$$

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 $\mathbf{x}_A, \ldots, \mathbf{x}_{AB}, \ldots, \mathbf{x}_{ABC}$ 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,

$$\begin{aligned}\alpha_0^* &= (\alpha\beta)_{11} + (\alpha\beta)_{12} - (\alpha\beta)_{21} - (\alpha\beta)_{22}\\ \boldsymbol{\beta}^* &= 2\boldsymbol{\beta}_1 - 2\boldsymbol{\beta}_2\\ (\alpha\beta)^* &= (\alpha\beta)_{11} - (\alpha\beta)_{12} - (\alpha\beta)_{21} + (\alpha\beta)_{22}.\end{aligned} \tag{6.18}$$

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

$$\begin{aligned}\eta_a &= \operatorname{logit}\pi_{1\cdot}(\mathbf{x}) = \alpha_0 + \boldsymbol{\alpha}^T\mathbf{x}\\ \eta_b &= \operatorname{logit}\pi_{\cdot 1}(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}^T\mathbf{x}\\ \eta_{ab} &= \log\left\{\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}(\mathbf{x})\right\} = (\alpha\beta)\end{aligned} \tag{6.19}$$

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 $\text{logit}\,\pi_{1\cdot}(\mathbf{x})$ 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 $\eta_0 = \log \pi_{\cdot\cdot}$ 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

$$\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\pi}} = \begin{array}{c} \\ \eta_0 \\ \eta_a \\ \eta_b \\ \eta_{ab} \end{array} \begin{array}{c} \begin{array}{cccc} \pi_{11} & \pi_{12} & \pi_{21} & \pi_{22} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 \\ \pi_{1\cdot}^{-1} & \pi_{1\cdot}^{-1} & -\pi_{2\cdot}^{-1} & -\pi_{2\cdot}^{-1} \\ \pi_{\cdot 1}^{-1} & -\pi_{\cdot 2}^{-1} & \pi_{\cdot 1}^{-1} & -\pi_{\cdot 2}^{-1} \\ \pi_{11}^{-1} & -\pi_{12}^{-1} & -\pi_{21}^{-1} & \pi_{22}^{-1} \end{pmatrix} \end{array}$$

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

$$\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta}} = \begin{array}{c} \\ \pi_{11} \\ \pi_{12} \\ \pi_{21} \\ \pi_{22} \end{array} \begin{array}{c} \begin{array}{cccc} \eta_0 & \eta_a & \eta_b & \eta_{ab} \end{array} \\ \begin{pmatrix} \pi_{11} & \dfrac{\pi_{11}\pi_{21}}{\pi_{.1}\Delta} & \dfrac{\pi_{11}\pi_{12}}{\pi_{1.}\Delta} & V_{ab} \\ \pi_{12} & \dfrac{\pi_{12}\pi_{22}}{\pi_{.2}\Delta} & -\dfrac{\pi_{11}\pi_{12}}{\pi_{1.}\Delta} & -V_{ab} \\ \pi_{21} & -\dfrac{\pi_{11}\pi_{21}}{\pi_{.1}\Delta} & \dfrac{\pi_{21}\pi_{22}}{\pi_{2.}\Delta} & -V_{ab} \\ \pi_{22} & -\dfrac{\pi_{12}\pi_{22}}{\pi_{.2}\Delta} & -\dfrac{\pi_{21}\pi_{22}}{\pi_{2.}\Delta} & V_{ab} \end{pmatrix} \end{array}$$

In the above matrix we have used the following quantities for future convenience:

$$V_a = \pi_{1.}\pi_{2.}, \qquad V_b = \pi_{.1}\pi_{.2},$$
$$V_{ab} = \left(\frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}}\right)^{-1},$$
$$\Delta = \pi_{11}\pi_{12}\pi_{21}\pi_{22}/(V_a V_b V_{ab})$$

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, $0 \le V_a, V_b \le 1/4$; $0 \le V_{ab} \le 1/16$, although it is possible for V_{ab} to exceed $V_a V_b$. Further, for all $\boldsymbol{\pi}$, $\Delta \le 1$, 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

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{(rs)} \sum_{(ij)} \frac{y_{ij} - m\pi_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial \eta_{rs}} \frac{\partial \eta_{rs}}{\partial \boldsymbol{\beta}}. \tag{6.20}$$

The indices r, s refer to contrasts, whereas i and j refer to 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) together with (6.14), and focus on the parameter $\boldsymbol{\beta}_a$ for the marginal regression of A on $\mathbf{x}_a$, the contribution to the log-likelihood derivative may be written as follows:

$$\begin{aligned} \sum_{(ij)} \frac{y_{ij} - m\pi_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial \eta_a} \mathbf{x}_a & \\ &= \left(\frac{\pi_{21}}{\pi_{\cdot 1}}\epsilon_{11} + \frac{\pi_{22}}{\pi_{\cdot 2}}\epsilon_{12} - \frac{\pi_{11}}{\pi_{\cdot 1}}\epsilon_{21} - \frac{\pi_{12}}{\pi_{\cdot 2}}\epsilon_{22}\right)\frac{\mathbf{x}_a}{\Delta} \\ &= \left(\epsilon_{1\cdot} - \left(\frac{\pi_{11}}{\pi_{\cdot 1}} - \frac{\pi_{12}}{\pi_{\cdot 2}}\right)\epsilon_{\cdot 1}\right)\frac{\mathbf{x}_a}{\Delta}, \end{aligned} \tag{6.21}$$

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 $\boldsymbol{\beta}_a$ and $\boldsymbol{\beta}_b$ depend only on the marginal totals, $y_{i\cdot}$ and $y_{\cdot j}$, and not otherwise on the joint composition of the two responses. Thus if the odds ratios for each bivariate response are given constants, the marginal totals are sufficient for $(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b)$. Note that although $\epsilon_{\cdot\cdot} \equiv 0$, $\epsilon_{1\cdot}$ and $\epsilon_{\cdot 1}$ are not necessarily zero.

In the case of the parameter η_{ab} in (6.14), the contribution to the log-likelihood derivative is

$$\sum_{(ij)} \frac{y_{ij} - m\pi_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial \eta_{ab}} 1 = V_{ab}\left(\frac{\epsilon_{11}}{\pi_{11}} - \frac{\epsilon_{12}}{\pi_{12}} - \frac{\epsilon_{21}}{\pi_{21}} + \frac{\epsilon_{22}}{\pi_{22}}\right). \tag{6.22}$$

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 $\boldsymbol{\beta}_a$, 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 $(Y_{1\cdot}, Y_{2\cdot})$, where $Y_{1\cdot} \sim B(m, \pi_{1\cdot})$ gives the following contribution to the log-likelihood derivative in place of (6.21):

$$(y_{1\cdot} - m\pi_{1\cdot})\mathbf{x}_a = \epsilon_{1\cdot}\mathbf{x}_a. \tag{6.23}$$

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

$$\begin{aligned} \boldsymbol{i}_\eta &= m\left(\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta}}\right)^T \operatorname{diag}(1/\boldsymbol{\pi})\left(\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\eta}}\right) \\ &= m\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & V_a/\Delta & \Delta_\pi/\Delta & 0 \\ 0 & \Delta_\pi/\Delta & V_b/\Delta & 0 \\ 0 & 0 & 0 & V_{ab} \end{pmatrix}. \end{aligned}$$

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 $(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b, \boldsymbol{\beta}_{ab})$ in the model

$$\eta_a(\mathbf{x}) = \boldsymbol{\beta}_a^T\mathbf{x}, \qquad \eta_b(\mathbf{x}) = \boldsymbol{\beta}_b^T\mathbf{x}, \qquad \eta_{ab}(\mathbf{x}) = \boldsymbol{\beta}_{ab}^T\mathbf{x}$$

is as follows

$$\mathbf{I}_\beta = \begin{pmatrix} \mathbf{X}^T\mathbf{D}_1\mathbf{X} & \mathbf{X}^T\mathbf{D}_{12}\mathbf{X} & \mathbf{0} \\ \mathbf{X}^T\mathbf{D}_{12}\mathbf{X} & \mathbf{X}^T\mathbf{D}_2\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X}^T\mathbf{D}_3\mathbf{X} \end{pmatrix} \tag{6.24}$$

where $\mathbf{D}_1, \ldots$ are diagonal matrices given by $\mathbf{D}_1 = \operatorname{diag}\{mV_a/\Delta\}$, $\mathbf{D}_2 = \operatorname{diag}\{mV_b/\Delta\}$, $\mathbf{D}_{12} = \operatorname{diag}\{m\Delta_\pi/\Delta\}$ and $\mathbf{D}_3 = \operatorname{diag}\{mV_{ab}\}$.

Note that $\boldsymbol{\beta}_{ab}$ is orthogonal to the marginal regression parameters and $\hat{\boldsymbol{\beta}}_{ab}$ has asymptotic covariance matrix $(\mathbf{X}^T\mathbf{D}_3\mathbf{X})^{-1}$ independently of $(\hat{\boldsymbol{\beta}}_a, \hat{\boldsymbol{\beta}}_b)$. 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 $\mathbf{I}_\beta$ algebraically to obtain the asymptotic covariance matrix of $(\hat{\boldsymbol{\beta}}_a, \hat{\boldsymbol{\beta}}_b)$ explicitly. However, it is clear from Exercise 6.17 that the covariance matrix of $\hat{\boldsymbol{\beta}}_a$ is smaller in the usual matrix sense than the covariance matrix $(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$, where

$$\mathbf{W} = \text{diag}\{mV_a\} = \mathbf{D}_1 - \mathbf{D}_{12}\mathbf{D}_2^{-1}\mathbf{D}_{12},$$

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, is may be appropriate to take $\boldsymbol{\alpha} = \boldsymbol{\beta}$ in (6.19). 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 pneumoconiosis, 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,

$$Z_a = \log\{(y_{1.} + \tfrac{1}{2})/(y_{2.} + \tfrac{1}{2})\},$$

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

$$Z_{ab} = \log\left\{\frac{(y_{11} + \tfrac{1}{2})(y_{22} + \tfrac{1}{2})}{(y_{12} + \tfrac{1}{2})(y_{21} + \tfrac{1}{2})}\right\}$$

Table 6.7 *Parameter estimates and standard errors for the bivariate logistic model $A*B: x^{\dagger}$, using marginal likelihoods and the joint likelihood*

	Marginal likelihood		*Joint likelihood*	
Parameter	*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

$^{\dagger}x = (\text{age} - 42)/5$.

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.

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.

	Link function		
Model formula	*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

$^\dagger x = (\text{age} - 42)/5$; $z = x^2$:

$^\ddagger R$, (= 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

$$\log(\hat{\pi}_{1.}/\hat{\pi}_{2.}) = -2.261 + 0.515x$$
$$\log(\hat{\pi}_{.1}/\hat{\pi}_{.2}) = -1.487 + 0.326x$$

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

$$\begin{aligned} \log\{\pi_{11}/\pi_{21}\} &= \alpha_{11} - \alpha_{21} + (\beta_{11} - \beta_{21})x \\ \log\{\pi_{12}/\pi_{22}\} &= \alpha_{12} - \alpha_{22} + (\beta_{12} - \beta_{22})x. \end{aligned} \tag{6.26}$$

These are linear in x, though not parallel. The corresponding logits for B given $A = 1, 2$ are

$$\begin{aligned} \log\{\pi_{11}/\pi_{12}\} &= \alpha_{11} - \alpha_{12} + (\beta_{11} - \beta_{12})x \\ \log\{\pi_{21}/\pi_{22}\} &= \alpha_{21} - \alpha_{22} + (\beta_{21} - \beta_{22})x. \end{aligned} \tag{6.27}$$

In both cases the difference between conditional logits is

$$(\alpha_{11} - \alpha_{12} - \alpha_{21} + \alpha_{22}) + (\beta_{11} - \beta_{12} - \beta_{21} + \beta_{22})x. \tag{6.28}$$

The maximum-likelihood fitted values for equations (6.26)–(6.28) are

$$\begin{aligned} \text{logit pr}(A = 1 \mid B{=}1, x) &= -0.418 + 0.349x \\ \text{logit pr}(B = 1 \mid A{=}1, x) &= 1.051 + 0.034x \\ \text{log odds-ratio} &= 3.059 - 0.166x. \end{aligned}$$

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

$$E(Y^{1/2}) \simeq \mu^{1/2}\left\{1 - \frac{1}{8\mu} - \frac{7}{128\mu^2} + O(\mu^{-3})\right\}$$
$$\text{var}(Y^{1/2}) \simeq \frac{1}{4}\left\{1 + \frac{3}{8\mu} + O(\mu^{-2})\right\}$$
$$\kappa_3(Y^{1/2}) \simeq -\mu^{-1/2}/16\left\{1 + O(\mu^{-1})\right\}$$

where $Y \sim P(\mu)$. Show also that

$$E(Y^{2/3}) \simeq \mu^{2/3}\left\{1 - \frac{1}{9\mu} - \frac{1}{27\mu^2} + O(\mu^{-3})\right\}$$
$$\text{var}(Y^{2/3}) \simeq \frac{4\mu^{1/3}}{9}\left\{1 + \frac{1}{6\mu} + O(\mu^{-2})\right\}$$
$$\kappa_3(Y^{2/3}) \simeq -68/(729\mu) + O(\mu^{-2})$$

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

$$\text{pr}(Y = y;\mu,\phi) = \frac{\Gamma(y+\phi\mu)\phi^{\phi\mu}}{y!\,\Gamma(\phi\mu)(1+\phi)^{y+\phi\mu}} \qquad y = 0,1,2,\dots .$$

Find the unconditional mean and variance of Y.

6.4 Fit the model

$$\text{site} + \text{class} + \text{volume} + \text{tuberculin}$$

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 6.5.7 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

$$\eta_a(\mathbf{x}) = \boldsymbol{\beta}_a^T \mathbf{x}_a, \qquad \eta_b(\mathbf{x}) = \boldsymbol{\beta}_b^T \mathbf{x}_b, \qquad \eta_{ab}(\mathbf{x}) = \eta_{ab},$$

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

$$y_{i\cdot} = m\hat{\pi}_{i\cdot}, \qquad y_{\cdot j} = m\hat{\pi}_{\cdot j}, \quad \text{for each bivariate response,}$$
$$\sum_1^n (y_{11} - m\hat{\pi}_{11}) = 0.$$

The final sum extends over the $(1,1)$-components of all n responses.

6.8 Show that the inverse of the multivariate logit transformation $\boldsymbol{\eta} \to \boldsymbol{\pi}$, 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	*Low I.Q. group*	*High I.Q. group*
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

$$\begin{aligned}
\eta_{ai} &= \log(\gamma_{i\cdot}/\gamma_{3\cdot}) \qquad i = 1, 2 \\
\eta_{bj} &= \log\{\gamma_{\cdot j}/(1-\gamma_{\cdot j})\} \qquad j = 1, 2, 3 \\
\eta_{ab\,ij} &= \log\Big(\frac{\gamma_{ij}}{\gamma_{i\cdot} - \gamma_{ij}}\Big) - \log\Big(\frac{\gamma_{3j}}{\gamma_{3\cdot} - \gamma_{3j}}\Big) \\
&= \text{logit}(\gamma_{ij}/\gamma_{i\cdot}) - \text{logit}(\gamma_{3j}/\gamma_{3\cdot}).
\end{aligned}$$

In these expressions γ_{ij} is defined as

$$\begin{aligned}
\gamma_{ij} &= \text{pr}(A = i, B \le j) \\
\gamma_{i\cdot} &= \text{pr}(A = i) \\
\gamma_{\cdot j} &= \text{pr}(B \le j).
\end{aligned}$$

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*					
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-

ratic 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×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 $AC, \ldots, DE$ 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 $ABD, \ldots, CDE$ are defined in like manner. Show that AB, AC and BC are marginal to ABC. Show that the rank of the model

matrix $\mathbf{Z}$ corresponding to the model formula

$$1 + ABC + ABD + ABE + ACD + ACE + ADE + BCD + BCE + BDE + CDE$$

is given by

$$\text{rank}(\mathbf{Z}) = 1 + \binom{k}{2} + 2\binom{k}{3},$$

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, p. 17–18). Fit the log-linear model *corps + 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

$$\nu = (N-r)(r-1)/(N-1), \quad \sigma = N\Big\{\sum_i s_i^{-1} - r^2/N\Big\}/(N-2),$$
$$\mu = (N-c)(c-1)/(N-1), \quad \tau = N\Big\{\sum_j t_j^{-1} - c^2/N\Big\}/(N-2).$$

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).

6.17 Suppose that the Fisher information matrix for $(\hat{\boldsymbol{\beta}}_a, \hat{\boldsymbol{\beta}}_b)$ has the partitioned form

$$\mathbf{I}_\beta = \begin{pmatrix} \mathbf{X}_a^T\mathbf{D}_1\mathbf{X}_a & \mathbf{X}_a^T\mathbf{D}_{12}\mathbf{X}_b \\ \mathbf{X}_b^T\mathbf{D}_{12}\mathbf{X}_a & \mathbf{X}_b^T\mathbf{D}_2\mathbf{X}_b \end{pmatrix}$$

in which $\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_{12}$ are $n \times n$ diagonal matrices as defined in (6.24), and $\mathbf{X}_a, \mathbf{X}_b$ are model matrices. Show that the inverse asymptotic covariance matrix of the component $\hat{\boldsymbol{\beta}}_a$ is given by

$$\{\text{cov}\,\hat{\boldsymbol{\beta}}_a\}^{-1} = \mathbf{X}_a^T\{\mathbf{D}_1 - \mathbf{D}_{12}\mathbf{X}_b(\mathbf{X}_b^T\mathbf{D}_2\mathbf{X}_b)^{-1}\mathbf{X}_b^T\mathbf{D}_{12}\}\mathbf{X}_a,$$

and that this formula reduces to $\mathbf{X}_a^T\mathbf{W}\mathbf{X}_a$ with $\mathbf{W} = \text{diag}\{mV_a\}$ if either (i) $\mathbf{D}_{12} = \mathbf{0}$ or (ii) $\text{rank}(\mathbf{X}_b) = n$.

Hence justify the claim that $\text{cov}(\hat{\boldsymbol{\beta}}_a) \leq (\mathbf{X}_a^T\mathbf{W}\mathbf{X}_a)^{-1}$.