CHAPTER 4

Binary data

4.1 Introduction

4.1.1 *Binary responses*

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

$$\mathrm{pr}(Y_i = 0) = 1 - \pi_i; \qquad \mathrm{pr}(Y_i = 1) = \pi_i \tag{4.1}$$

for the probabilities of 'failure' and 'success' respectively.

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

4.1.2 *Covariate classes*

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

Table 4.1 *Alternative ways of presenting the same data*

(a) *Data listed by subject No.*

Subject No.	*Covariate* (x_1, x_2)	*Response* Y
1	1, 1	0
2	1, 2	1
3	1, 2	0
4	2, 1	0
5	2, 2	1
6	1, 2	1
7	1, 1	1

(b) *Data listed by covariate class*

Covariate (x_1, x_2)	*Class size* m	*Response* Y
1, 1	2	1
1, 2	3	2
2, 1	1	0
2, 2	1	1

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

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

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

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

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

4.1.3 *Contingency tables*

Suppose that the data are indexed by three explanatory factors, A having a levels, B having b levels and C having c levels. Among the subjects observed, therefore, there are at most $a \times b \times c$ covariate classes. There may in fact be fewer covariate classes than this maximum either because, by chance, one or more covariate classes were not observed or because certain factor combinations are physically or logically impossible (Section 3.7.1). For each covariate class, the number of successes and the number of failures is counted. Such data may be presented as a $2 \times a \times b \times c$ table of counts called a *contingency table*. For instance, the data in Table 4.1 give rise to the $2 \times 2 \times 2$ table

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

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

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

4.2 Binomial distribution

4.2.1 *Genesis*

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

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

$$\mathrm{pr}(Y_1 = y \mid Y_1 + Y_2 = m) = \binom{m}{y} \pi^y (1-\pi)^{m-y}, \qquad y = 0, 1, \ldots, m \tag{4.2}$$

where $\pi = \mu_1/(\mu_1 + \mu_2)$. This conditional distribution depends only on the ratio of the Poisson means and not on $\mu_1 + \mu_2$. Details of the derivation are given in Exercise 4.4. The notation $Y \sim B(m, \pi)$ means that Y has the binomial distribution (4.2) with *index* m and *parameter* π.

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

4.2.2 *Moments and cumulants*

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

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

Hence, the cumulant generating function is

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

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

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

and that the cumulant generating function is

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

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

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

All cumulants of Y have the form $m\times$polynomial in π. The expressions for the moments are more complicated except in the special case $m = 1$ for which all moments of all orders are equal to π.

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

$$\begin{aligned}
\kappa_1 &= \sum \pi_i = m\bar{\pi},\\
\kappa_2 &= \sum \pi_i(1-\pi_i) = m\bar{\pi}(1-\bar{\pi}) - (m-1)k_2(\pi) \leq m\bar{\pi}(1-\bar{\pi}),\\
\kappa_3 &= \sum \pi_i(1-\pi_i)(1-2\pi_i),\\
\kappa_4 &= \sum \pi_i(1-\pi_i)\{1-6\pi_i(1-\pi_i)\},
\end{aligned}$$

where $k_2(\pi) = \sum(\pi_i - \bar{\pi})^2/(m-1)$ is the 'sample variance' of the πs. Evidently, the sample variance of Y is deflated relative to the binomial variance. This calculation appears to contradict the common intuition that lack of homogeneity should increase variability rather than decrease it. The reason for the apparent contradiction is that the calculations just given are not relevant to the problem of heterogeneity as usually met. In practice, it is usually known only that there is variability among the πs: the complete set of values $\pi_1, \ldots, \pi_m$ is rarely known. A more relevant calculation, therefore, is to regard $\pi_1, \ldots, \pi_m$ as independent random variables with mean $\bar{\pi}$. It is then easily shown that, whatever the distribution of π_i, $Y_i \sim B(1, \bar{\pi})$. Hence, the sum $Y = Y_1 + \ldots + Y_m$ is distributed as $B(m, \bar{\pi})$ and the binomial distribution is recovered.

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

4.2.3 *Normal limit*

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

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

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

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

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

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

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

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

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

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

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

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

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

may be quite large if z^- is large.

4.2.4 *Poisson limit*

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

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

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

4.2.5 *Transformations*

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

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

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

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

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

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

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

$$\begin{aligned} w^+ &= w(y+\tfrac{1}{2}) \\ \tau &= 1 + \frac{1}{24m\pi(1-\pi)}. \end{aligned}$$

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

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

Note that the ratio of τ to σ_W is

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

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

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

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

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

$$Z = \log\Big(\frac{Y+\frac{1}{2}}{m-Y+\frac{1}{2}}\Big), \qquad (4.7)$$

which has the property that

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

This is known as the empirical logistic transformation (Cox, 1970). For any other choice of constant, the bias is $O(m^{-1})$: see Exercise 4.15.

Gart and Zweifel's (1967) results support the estimation of $\mathrm{var}(Z)$ by $v = (y + \frac{1}{2})^{-1} + (m - y + \frac{1}{2})^{-1}$. The idea behind transformation is that it may be simpler to use a linear regression model for Z with weights v^{-1} rather than to use a non-linear model for the untransformed responses. This is often a simple and attractive alternative to maximum likelihood. Because the argument is asymptotic in nature, the transformation is useful only if all the binomial indices are fairly large.

4.3 Models for binary responses

4.3.1 *Link functions*

To investigate the relationship between the response probability π and the covariate vector $(x_1, \ldots, x_p)$, it is convenient, though perhaps not absolutely necessary, to construct a formal model thought capable of describing the effect on π of changes in $(x_1, \ldots, x_p)$. In practice, this formal model usually embodies assumptions such as zero correlation or independence, lack of interaction or additivity, linearity and so on. These assumptions cannot be taken for granted and should, if possible, be checked. Furthermore, the behaviour of the model should, as far as possible, be consistent with known physical, biological or mathematical laws, especially in its limiting behaviour.

Linear models play an important role in both applied and theoretical work — and with good reason. We suppose therefore that the dependence of π on $(x_1, \ldots, x_p)$ occurs through the linear combination

$$\eta = \sum_{j=1}^{p} x_j \beta_j \tag{4.8}$$

for unknown coefficients $\beta_1, \ldots, \beta_p$. Unless restrictions are imposed on $\boldsymbol{\beta}$ we have $-\infty < \eta < \infty$. Thus, to express π as the linear combination (4.8) would be inconsistent with the laws of probability. A simple and effective way of avoiding this difficulty is to use a transformation $g(\pi)$ that maps the unit interval onto the whole real line $(-\infty, \infty)$. This remedy leads to instances of generalized linear models in which the systematic part is

$$g(\pi_i) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j; \qquad i = 1, \ldots, n. \tag{4.9}$$

A wide choice of link functions $g(\pi)$ is available. Three functions commonly used in practice are

1. the logit or logistic function

$$g_1(\pi) = \log\{\pi/(1-\pi)\};$$

2. the probit or inverse Normal function

$$g_2(\pi) = \Phi^{-1}(\pi);$$

3. the complementary log-log function

$$g_3(\pi) = \log\{-\log(1-\pi)\}.$$

A fourth possibility, the log-log function

$$g_4(\pi) = -\log\{-\log(\pi)\},$$

which is the natural counterpart of the complementary log-log function, is seldom used because its behaviour is inappropriate for $\pi < \frac{1}{2}$, the region that is usually of interest. All four functions can be obtained as the inverses of well-known cumulative distribution functions having support on the entire real axis. The corresponding density functions are discussed in Exercises 4.22–4.23. The first two functions are symmetrical in the sense that

$$g_1(\pi) = -g_1(1-\pi).$$

The latter two functions are not symmetrical in this sense, but are related via

$$g_3(\pi) = -g_4(1-\pi).$$

All four functions are continuous and increasing on $(0, 1)$.

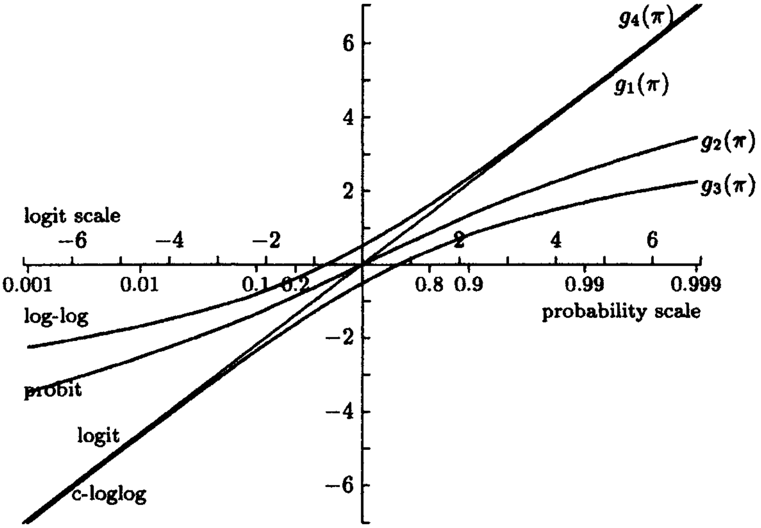

Fig. 4.1. *A graphical comparison of three link functions with the logistic function: the* 45° *line is the logistic function.*

Figure 4.1 compares the four functions. The logistic function is taken as the standard and $g_2(\pi)$, $g_3(\pi)$, $g_4(\pi)$ are plotted against $g_1(\pi)$ for values of π in the range 0.01 to 0.99.

The logistic and the probit function are almost linearly related over the interval $0.1 \le \pi \le 0.9$. For this reason, it is usually difficult to discriminate between these two functions on the grounds of goodness-of-fit; see, for example, Chambers and Cox (1967). For small values of π, the complementary log-log function is close to the logistic, both being close to $\log(\pi)$. As π approaches 1, the complementary log-log function approaches infinity much more slowly than either the logistic or the probit function. Similar comments apply to the log-log function as can be seen from Figure 4.1.

All asymptotic and approximate theory presented in this chapter applies regardless of the choice of link function. However, we shall be concerned mostly with the logistic function, not so much because of its simpler theoretical properties, but because of its simple interpretation as the logarithm of the odds ratio. Apart from this, the logistic function has one important advantage over all alternative transformations in that it is eminently suited for the analysis of data collected retrospectively. See section 4.3.3.

4.3.2 *Parameter interpretation*

In order to summarize the conclusions of an analysis in an easily digested form, it is helpful to state the magnitudes of the estimated effects on an easily understood scale. The scale most suitable for this purpose is often different from the scale or link function used to achieve additivity of effects, namely $g(\pi)$. For instance, if a linear logistic model has been used with two covariates x_1 and x_2, we have the model

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

for the log odds of a positive response. Equivalently, the model may be written in terms of the odds of a positive response, giving

$$\frac{\pi}{1-\pi} = \exp\bigl(\beta_0 + \beta_1 x_1 + \beta_2 x_2\bigr).$$

Finally, the probability of a positive response is

$$\pi = \frac{\exp\bigl(\beta_0 + \beta_1 x_1 + \beta_2 x_2\bigr)}{1 + \exp\bigl(\beta_0 + \beta_1 x_1 + \beta_2 x_2\bigr)}.$$

This is the inverse function of $g_1(\pi)$. Assuming that x_1 and x_2 are functionally unrelated, the conclusions based on such a model may be stated as follows. The effect of a unit change in x_2 is to increase the log odds by an amount β_2. Equivalently, but perhaps preferably, we may say that the effect of a unit change in x_2 is to increase the odds of a positive response multiplicatively by the factor $\exp(\beta_2)$. It is important here that x_1 be held fixed and not be permitted to vary as a consequence of the change in x_2. These statements are fairly easy to comprehend because the direction and magnitude of the stated effect are unaffected by the values of x_1 and x_2.

The corresponding statements given on the probability scale are more complicated because the effect on π of a unit change in x_2 depends on the values of x_1 and x_2. The derivative of π with respect to x_2 is

$$\frac{\partial \pi}{\partial x_2} = \pi(1-\pi)\beta_2.$$

Thus, a small change in x_2 has a larger effect, as measured on the probability scale, if π is near 0.5 than if π is near 0 or 1. Perhaps the simplest device to assist in the presentation of conclusions is to give the graph of

$$\pi(\eta) = \exp(\eta)/\{1+\exp(\eta)\}$$

and to state the effect on η of changes in x_2. The effect on the probability can then be read from the graph. This method works equally well whatever the link function used. The required inverse link functions are

$$\begin{aligned} \pi_2(\eta) &= g_2^{-1}(\eta) = \Phi(\eta), \\ \pi_3(\eta) &= g_3^{-1}(\eta) = 1-\exp(-e^{\eta}), \\ \text{and}\quad \pi_4(\eta) &= g_4^{-1}(\eta) = \exp(-e^{-\eta}). \end{aligned}$$

All of these functions are defined for $-\infty < \eta < \infty$ and increase continuously from zero at $-\infty$ to one at ∞.

4.3.3 *Retrospective sampling*

One important property of the logistic function not shared by the other link functions is that differences on the logistic scale can be estimated regardless of whether the data are sampled *prospectively* or *retrospectively*. To illustrate the difference between these two sampling schemes, suppose that a population is partitioned according to two binary variables, $(D, \bar{D})$ referring to the presence or absence of disease, and $(X, \bar{X})$ referring to exposure or non-exposure to the toxin or carcinogen under investigation. Suppose that the proportions of the population in the four categories thus formed are as shown in Table 4.2.

Table 4.2 *Hypothetical frequencies of disease and exposure status*

		Disease	*status*	
		$\bar{D}$	D	*Total*
Exposure	$\bar{X}$	$\pi_{00} = 0.70$	$\pi_{01} = 0.02$	$\pi_{0.} = 0.72$
status	X	$\pi_{10} = 0.25$	$\pi_{11} = 0.03$	$\pi_{1.} = 0.28$
	Total	$\pi_{.0} = 0.95$	$\pi_{.1} = 0.05$	1.0

In a prospective study, an exposed group of subjects is selected together with a comparable group of non-exposed individuals. The progress of each group is monitored, often over a prolonged period, with a view towards comparing the incidence of disease in the two groups. In this way, the row totals, giving the numbers of subjects in each of the exposure categories, are fixed by design. The column totals are random, reflecting the incidence of disease in the overall population, weighted according to the sizes of exposure groups in the sample.

In a retrospective study, diseased and disease-free individuals are selected — often from hospital records collected over a period of several years. In this design, the column totals are fixed by design and the row totals are random, reflecting the frequency of exposure in the population, weighted according to the sizes of the disease groups in the sample.

Considering the prospective study first, the logits for the two exposure groups are

$$\log(\pi_{01}/\pi_{00}) = -\log(35) = -3.555 \quad \text{and}$$
$$\log(\pi_{11}/\pi_{10}) = -\log(8.3) = -2.120.$$

The difference of logits is thus

$$\Delta = \log(\pi_{11}/\pi_{10}) - \log(\pi_{01}/\pi_{00}) = 1.435.$$

This difference could also be estimated by sampling retrospectively from the two disease groups $\bar{D}$ and D because

$$\Delta = \log(\pi_{11}/\pi_{01}) - \log(\pi_{10}/\pi_{00}).$$

In fact, in the present example, the retrospective design is substantially more efficient than the prospective design. This is because the disease is rare even among those who are exposed to the toxin or carcinogen. Thus, for a prospective study to be effective, a large number of initially healthy subjects must be followed for a prolonged period in order that a sufficiently large number of subjects may eventually fall victim to the disease. In a retrospective study, on the other hand, the investigator has access via hospital records to all cases of the disease recorded over a substantial period of time. In the case of rare diseases, it is common to take a 100% sample of the diseased individuals and to compare these with a similar sized sample of disease-free subjects. Since exposure is fairly common, ranging from 26% among those who are disease-free to 60% among those with the disease, a substantial number of exposed and non-exposed subjects will be observed both among the cases (D) and among the controls $(\bar{D})$.

More generally, if there are several exposure groups and other covariates, we may write the linear logistic model in the form

$$\mathrm{pr}(D \mid \mathbf{x}) = \exp(\alpha + \boldsymbol{\beta}^T\mathbf{x})/[1 + \exp(\alpha + \boldsymbol{\beta}^T\mathbf{x})] \qquad (4.10)$$

for the probability of contracting the disease given that the subject has covariates $\mathbf{x}$. Included in $\mathbf{x}$ is the information on the exposure category to which the individual belongs, together with other factors considered relevant to the incidence of the disease.

Model (4.10) is specified in a form appropriate for data sampled prospectively. Suppose, however that the data are sampled retrospectively. Introduce the dummy variable Z to define whether an individual is sampled or not, and denote the sampling proportions by

$$\pi_0 = \mathrm{pr}(Z = 1 \mid D) \qquad \text{and} \qquad \pi_1 = \mathrm{pr}(Z = 1 \mid \bar{D}).$$

It is essential here that the sampling proportions depend only on D and not on $\mathbf{x}$. We may now use Bayes's theorem to compute the disease frequency among sampled individuals who have a specified covariate vector $\mathbf{x}$.

$$\begin{aligned}\mathrm{pr}(D \mid Z = 1, \mathbf{x}) &= \frac{\mathrm{pr}(Z = 1 \mid D, \mathbf{x})\, \mathrm{pr}(D \mid \mathbf{x})}{\mathrm{pr}(Z = 1 | D, \mathbf{x})\, \mathrm{pr}(D|\mathbf{x}) + \mathrm{pr}(Z = 1 | \bar{D}, \mathbf{x})\, \mathrm{pr}(\bar{D}|\mathbf{x})} \\ &= \frac{\pi_0 \exp(\alpha + \boldsymbol{\beta}^T\mathbf{x})}{\pi_1 + \pi_0 \exp(\alpha + \boldsymbol{\beta}^T\mathbf{x})} \\ &= \frac{\exp(\alpha^* + \boldsymbol{\beta}^T\mathbf{x})}{1 + \exp(\alpha^* + \boldsymbol{\beta}^T\mathbf{x})},\end{aligned}$$

where $\alpha^* = \alpha + \log(\pi_0/\pi_1)$. In other words, although the data have been sampled retrospectively, the logistic model (4.10) continues to apply with the same coefficients $\boldsymbol{\beta}$ but a different intercept. It follows therefore, that the logistic models described here in the context of prospective studies can be applied to retrospective studies provided that the intercept is treated as a nuisance parameter.

This derivation follows the lines of Armitage (1971) and Breslow and Day (1980). No such simple inversion exists for probit or complementary log-log models.

4.4 Likelihood functions for binary data

4.4.1 *Log likelihood for binomial data*

The responses $y_1, \ldots, y_n$ are assumed to be the observed values of independent random variables $Y_1, \ldots, Y_n$ such that Y_i has the binomial distribution with index m_i and parameter π_i. It is convenient initially to consider the log likelihood as a function of the n-vector $\boldsymbol{\pi} = \pi_1, \ldots, \pi_n$. Subsequently, when we wish to study specific linear models such as (4.10), the log likelihood is considered as a function of the coefficients appearing in the model. Using (4.2), the log likelihood may be written in the form

$$l(\boldsymbol{\pi}; \mathbf{y}) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i)\right]. \tag{4.11}$$

The constant function of $\mathbf{y}$ not involving $\boldsymbol{\pi}$, namely

$$\sum \log \binom{m_i}{y_i},$$

has been omitted because it plays no role.

The systematic part of the model specifies the relation between the vector $\boldsymbol{\pi}$ and the experimental or observational conditions as summarized by the model matrix $\mathbf{X}$ of order $n \times p$. For generalized linear models, this relationship takes the form

$$g(\pi_i) = \eta_i = \sum_j x_{ij}\beta_j; \qquad i = 1, \ldots, n, \tag{4.12}$$

so that the log likelihood (4.11) can be expressed as a function of the unknown parameters $\beta_1, \ldots, \beta_p$. It is a good tactical manoeuvre, however, not to make this substitution but to keep the two expressions separate. For instance, we may wish to compare several models by adding or deleting covariates. This operation changes the set of parameters, but leaves expression (4.11) unaltered.

In the case of linear logistic models, we have

$$g(\pi_i) = \eta_i = \log\{\pi_i/(1-\pi_i)\} = \sum_j x_{ij}\beta_j.$$

Substitution into (4.11) gives

$$l(\boldsymbol{\beta}; \mathbf{y}) = \sum_i \sum_j y_i x_{ij}\beta_j - \sum_i m_i \log\Big(1 + \exp\sum_j x_{ij}\beta_j\Big), \quad (4.13)$$

where we have written $l(\boldsymbol{\beta}; \mathbf{y})$ instead of $l(\boldsymbol{\pi}(\boldsymbol{\beta}); \mathbf{y})$. The important point to notice here is that, because the logistic link is also the canonical link, the log likelihood depends on $\mathbf{y}$ only through the linear combinations $\mathbf{X}^T\mathbf{y}$. These p combinations are said to be sufficient for $\boldsymbol{\beta}$. In fact, as will be seen shortly, the likelihood equations in this special case amount to setting the observed linear combinations $\mathbf{X}^T\mathbf{y}$ equal to their expectation, namely $E(\mathbf{X}^T\mathbf{Y}; \hat{\boldsymbol{\beta}})$. This may be viewed a special case of the *method of moments*.

Section 2.4.4 and Table 2.1 give the canonical link functions for other distributions.

4.4.2 *Parameter estimation*

Following the general technique given in section 2.5, we now derive the likelihood equations for the parameters $\boldsymbol{\beta}$ that appear in (4.12). First note that the derivative of the log-likelihood function, in the form given in (4.11), with respect to π_i is

$$\frac{\partial l}{\partial \pi_i} = \frac{y_i - m_i\pi_i}{\pi_i(1-\pi_i)}.$$

Using the chain rule, the derivative with respect to β_r is

$$\frac{\partial l}{\partial \beta_r} = \sum_{i=1}^{n} \frac{y_i - m_i\pi_i}{\pi_i(1-\pi_i)} \frac{\partial \pi_i}{\partial \beta_r}.$$

In the case of generalized linear models, it is convenient to express $\partial\pi_i/\partial\beta_r$ as a product

$$\frac{\partial \pi_i}{\partial \beta_r} = \frac{d\pi_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_r} = \frac{d\pi_i}{d\eta_i} x_{ir}.$$

Thus the derivative with respect to β_r is

$$\frac{\partial l}{\partial \beta_r} = \sum_i \frac{y_i - m_i\pi_i}{\pi_i(1-\pi_i)} \frac{d\pi_i}{d\eta_i} x_{ir}. \tag{4.14}$$

The Fisher information for $\boldsymbol{\beta}$ is

$$\begin{aligned} -E\left(\frac{\partial^2 l}{\partial\beta_r \partial\beta_s}\right) &= \sum_i \frac{m_i}{\pi_i(1-\pi_i)} \frac{\partial\pi_i}{\partial\beta_r} \frac{\partial\pi_i}{\partial\beta_s} \\ &= \sum_i m_i \frac{(d\pi_i/d\eta_i)^2}{\pi_i(1-\pi_i)} x_{ir}x_{is} \\ &= \left\{\mathbf{X}^T\mathbf{W}\mathbf{X}\right\}_{rs}, \end{aligned} \tag{4.15}$$

where $\mathbf{W}$ is a diagonal matrix of weights given by

$$\mathbf{W} = \text{diag}\left\{m_i\left(\frac{d\pi_i}{d\eta_i}\right)^2 \Big/ \pi_i(1-\pi_i)\right\}.$$

In the case of linear logistic models, equation (4.14) reduces to

$$\partial l/\partial\boldsymbol{\beta} = \mathbf{X}^T(\mathbf{Y} - \boldsymbol{\mu})$$

when written in matrix notation. The likelihood equations then amount to equating the sufficient statistic, $\mathbf{X}^T\mathbf{Y}$, to its expectation as a function of $\hat{\boldsymbol{\beta}}$. In addition, the diagonal matrix of weights appearing in the Fisher information reduces to

$$\mathbf{W} = \text{diag}\{m_i\pi_i(1-\pi_i)\}.$$

Following the lines of the general Newton-Raphson procedure described in Chapter 2, parameter estimates may be obtained in the following way. Given initial estimates $\hat{\boldsymbol{\beta}}_0$, we may compute the vectors $\hat{\boldsymbol{\pi}}_0$ and $\hat{\boldsymbol{\eta}}_0$. Using these values, define the adjusted dependent variate, $\mathbf{Z}$, with components

$$z_i = \hat{\eta}_i + \frac{y_i - m_i\hat{\pi}_i}{m_i} \frac{d\eta_i}{d\pi_i},$$

all quantities being computed at the initial estimate $\hat{\boldsymbol{\beta}}_0$. Maximum-likelihood estimates satisfy the equation

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{W}\mathbf{Z}, \tag{4.16}$$

which can be solved iteratively using standard least-squares methods. The revised estimate is

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Z}$$

where all quantities appearing on the right are computed using the initial estimate.

Failure to converge is rarely a problem unless one or more components of $\hat{\boldsymbol{\beta}}$ are infinite, which usually implies that some of the fitted probabilities are either zero or one. Infinite parameter estimates can occur if the data are sparse and $y_i = 0$ or $y_i = m_i$ for certain components of the response vector. Although the iterative procedure does not converge under these circumstances, nevertheless the sequence of fitted probabilities, $\hat{\boldsymbol{\pi}}^{(j)}$ generally tends quite rapidly towards $\hat{\boldsymbol{\pi}}$ and the deviance towards its limiting value. After a few cycles of (4.16) the fitted values $m_i\hat{\pi}_i$ are normally quite accurate but the parameter estimates and their standard errors may not be. Two criteria ought therefore to be tested to detect abnormal convergence of this type. The primary criterion ought to be based on the change in the fitted probabilities, for instance by using the deviance. A supplementary test for parameter divergence can be based on the change in $\hat{\boldsymbol{\beta}}$ or in the linear predictor, $\hat{\boldsymbol{\eta}}$. Abnormal convergence means that the log likelihood is either very flat or, more likely, has an asymptote. Consequently, the computed parameter estimates and their estimated standard errors are not to be trusted.

Some results concerning the existence and uniqueness of parameter estimates have been given by Wedderburn (1976) and by Haberman (1977). These results show that if the link function is log concave, as it is for the four functions discussed in section 4.3.1, and if $0 < y_i < m_i$ for each i, then $\hat{\boldsymbol{\beta}}$ is finite and the log likelihood has a unique maximum at $\hat{\boldsymbol{\beta}}$.

Starting values $\hat{\boldsymbol{\beta}}^{(0)}$ can be obtained using the method described in Chapter 2, beginning with 'fitted values' $\tilde{\mu} = (y + \frac{1}{2})/(m+1)$. A good choice of starting value usually reduces the number of cycles in (4.16) by about one or perhaps two. Consequently, the choice of initial estimate is usually not critical. A bad choice may, however, result in divergence.

4.4.3 *Deviance function*

The residual deviance is defined to be twice the difference between the maximum achievable log likelihood and that attained under the fitted model. Under any given model, H_0, with fitted probabilities $\hat{\boldsymbol{\pi}}$, the log likelihood is

$$l(\hat{\boldsymbol{\pi}};\mathbf{y}) = \sum_i \{y_i \log \hat{\pi}_i + (m_i - y_i)\log(1-\hat{\pi}_i)\},$$

which is just (4.11) written in a more symmetrical form. The maximum achievable log likelihood is attained at the point $\tilde{\pi}_i = y_i/m_i$, but this point does not usually occur in the model space under H_0. The deviance function is therefore

$$\begin{aligned} D(\mathbf{y};\hat{\boldsymbol{\pi}}) &= 2l(\tilde{\boldsymbol{\pi}};\mathbf{y}) - 2l(\hat{\boldsymbol{\pi}};\mathbf{y}) \\ &= 2\sum_i \Big\{ y_i \log(y_i/\hat{\mu}_i) + (m_i - y_i)\log\Big(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\Big)\Big\}. \end{aligned}$$

This function behaves in much the same way as the residual sum of squares or weighted residual sum of squares in ordinary linear models. The addition of further covariates has the effect of reducing D.

It is often claimed that the random variable $D(\mathbf{Y};\hat{\boldsymbol{\pi}})$ is asymptotically or approximately distributed as χ^2_{n-p}, where p is the number of fitted parameters under H_0. This claim is then used to justify the use of D as a goodness-of-fit statistic for testing the adequacy of the fitted model. Proofs of the limiting χ^2_{n-p} distribution are based on the following assumptions whose relevance in any given application must be open to question.

Assumption 1: The observations are distributed independently according to the binomial distribution. In other words, the possibility of over-dispersion (Section 4.5) is not considered.

Assumption 2: The approximation is based on a limiting operation in which $\dim(\mathbf{Y}) = n$ is fixed, $m_i \to \infty$ for each i, and in fact $m_i\pi_i(1-\pi_i) \to \infty$.

In the limit given by assumption 2, D is approximately independent of the estimated parameters $\hat{\boldsymbol{\beta}}$ and hence approximately independent of the fitted probabilities $\hat{\boldsymbol{\pi}}$. Approximate independence is essential for D to be considered as a goodness-of-fit statistic, but this property alone does not guarantee good power.

If n is large and $m_i\pi_i(1-\pi_i)$ remains bounded the whole theory breaks down in two ways. First, the limiting χ^2 approximation no longer holds. Second, and more importantly, D is not independent of $\hat{\boldsymbol{\pi}}$ even approximately. As a consequence, a large value of D could be obtained with high probability by judicious choice of $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$. In other words, a large value of D cannot necessarily be considered to be evidence of a poor fit. For an extreme instance of this effect, see section 4.4.5.

The deviance function is most directly useful not as an absolute measure of goodness-of-fit but for comparing two nested models. For instance, we may wish to test whether the addition of a further covariate significantly improves the fit. Let H_0 denote the model under test and H_A the extended model containing an additional covariate. The corresponding fitted values are denoted by $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\mu}}_A$ respectively. The reduction in deviance

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}_0) - D(\mathbf{y};\hat{\boldsymbol{\mu}}_A) = 2l(\hat{\boldsymbol{\mu}}_A;\mathbf{y}) - 2l(\hat{\boldsymbol{\mu}}_0;\mathbf{y}) \qquad (4.17)$$

is identical to the likelihood-ratio statistic for testing H_0 against H_A. This statistic is distributed approximately like χ^2_1 independently of $\hat{\boldsymbol{\mu}}$ under assumption 1 above provided that either n is large or that assumption 2 is satisfied. In particular, $D(\mathbf{Y};\hat{\boldsymbol{\mu}}_0)$ need not have an approximate χ^2 distribution nor need it be distributed independently of $\hat{\boldsymbol{\mu}}_0$. The χ^2 approximation is usually quite accurate for differences of deviances even though it is inaccurate for the deviances themselves.

4.4.4 *Bias and precision of estimates*

To a first order of approximation, maximum-likelihood estimates are unbiased with asymptotic variance equal to the inverse Fisher information matrix (4.15). Specifically, for large n,

$$\begin{aligned} E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= O(n^{-1}) \\ \operatorname{cov}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\{1 + O(n^{-1})\}. \end{aligned} \qquad (4.18)$$

These approximate results are also true for the alternative limit in which n is fixed and $\mathbf{m} \to \infty$. The errors are then $O(m_i^{-1})$.

It is possible to give an expression for the bias of $\hat{\boldsymbol{\beta}}$ that covers all link functions. However, in order to keep the expressions as simple as possible, we shall restrict attention to linear logistic models. In that case, the bias of $\hat{\boldsymbol{\beta}}$ involves the 3-way array

$$\kappa_{r,s,t} = \sum_i x_{ir}x_{is}x_{it}\, m_i\pi_i(1-\pi_i)(1-2\pi_i),$$

which is just the skewness array of the log likelihood derivative $\partial l/\partial\boldsymbol{\beta}$. If we denote by $\kappa_{r,s}$ the elements of the Fisher information matrix $\mathbf{X}^T\mathbf{W}\mathbf{X}$, and by $\kappa^{r,s}$ the elements of the inverse matrix, we have

$$\text{bias}(\hat{\beta}_r) \simeq -\sum_{ijk} \kappa^{r,i}\kappa^{j,k}\kappa_{i,j,k}/2.$$

The approximate skewness array of $\hat{\boldsymbol{\beta}}$ is

$$\text{cum}(\hat{\beta}_r, \hat{\beta}_s, \hat{\beta}_t) \simeq -2\sum_{ijk} \kappa^{r,i}\kappa^{s,j}\kappa^{t,k}\kappa_{i,j,k}.$$

Bias and skewness terms represent the major departures of the distribution of $\hat{\boldsymbol{\beta}}$ from the usual Normal approximation. Edgeworth corrections will usually improve the accuracy of the approximation.

The corresponding expressions for other link functions are given by McCullagh (1987, p.209). Computational tactics for generalized linear models are discussed in sections 15.2–15.3.

4.4.5 *Sparseness*

By sparseness we mean that a sizeable proportion of the observed counts are small. An extreme instance of this phenomenon occurs in Table 4.1a, where data are listed by subject number and hence $m_i = 1$ for each i. More generally, we say that the data are sparse if many components of the binomial index vector are small, say 5 or less. Sparseness does not necessarily imply that there is little information in the data about the values of the parameters. On the contrary, if the data recorded are extensive, (n large), the asymptotic approximation (4.18) is usually quite accurate. The effect of sparseness is noticed mainly on the deviance function and Pearson's statistic, which fail to have the properties required for goodness-of-fit statistics.

To illustrate the nature of the effect, suppose that $Y_i \sim B(1, \pi_i)$ and that a linear logistic model such as (4.10) has been fitted by maximum likelihood, yielding fitted values

$$\hat{\pi}_i = \exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})/[1+\exp(\mathbf{x}_i^T\hat{\boldsymbol{\beta}})].$$

The residual deviance function is

$$\begin{aligned}D &= 2\sum\left\{y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\hat{\pi}_i}\right)\right\}\\ &= 2\sum\left\{y_i \log y_i + (1-y_i)\log(1-y_i)\right.\\ &\qquad\left. - y_i\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) - \log(1-\hat{\pi}_i)\right\}.\end{aligned}$$

Since $y = 0$ or 1, we have $y \log y = (1-y)\log(1-y) = 0$. Further, $\log(\hat{\pi}_i/(1-\hat{\pi}_i)) = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$. Thus

$$\begin{aligned}D &= -2\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{Y} - 2\sum\log(1-\hat{\pi}_i)\\ &= -2\hat{\boldsymbol{\eta}}^T\hat{\boldsymbol{\pi}} - 2\sum\log(1-\hat{\pi}_i)\end{aligned}$$

since $\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\hat{\boldsymbol{\mu}}$ is the maximum-likelihood equation. Evidently, therefore, D is a function of $\hat{\boldsymbol{\beta}}$ in this case. In other words, given $\hat{\boldsymbol{\beta}}$, D has a conditionally degenerate distribution and cannot be used to test goodness of fit. Exact degeneracy occurs only for linear logistic models if $m_i = 1$, but near degeneracy occurs for any link function provided that the m_i are small.

The effect of extreme sparseness on Pearson's statistic is less obvious but can be seen from the following example. Suppose that the observations are identically distributed and $Y_i \sim B(1, \pi)$. Then $\hat{\pi} = \bar{y}$ and Pearson's statistic reduces to

$$X^2 = \sum\frac{(y_i - \bar{y})^2}{\bar{y}(1-\bar{y})} = n.$$

The sample size is not very useful as a test for goodness of fit! The deviance function fares no better, for

$$D = -2n\{\bar{y}\log\bar{y} + (1-\bar{y})\log(1-\bar{y})\},$$

is a function of $\hat{\pi}$.

For intermediate cases in which the m_i are small but mostly greater than one, we may use D or X^2 as test statistics. However, in the computation of significance levels it is essential to use the conditional distribution of the statistic given the observed $\hat{\boldsymbol{\beta}}$. Exact conditional moments of X^2 can be computed in some important special cases: see, for example, the Haldane-Dawson formulae for two-way tables in Exercise 6.16. More generally, however, approximate formulae are available for the conditional mean and variance of X^2 for linear logistic models (McCullagh, 1985). If n is large, it is best to use a Normal approximation for X^2 in which the conditional mean and variance are

$$\begin{aligned}E(X^2\,|\,\hat{\beta}) \simeq n - p - \tfrac{1}{2}\sum_i \{1 - 6\hat{\pi}_i(1-\hat{\pi}_i)\}\hat{V}_{ii} \\ + \tfrac{1}{2}\sum_{ij} m_i\hat{\pi}_i(1-\hat{\pi}_i)(1-2\hat{\pi}_i)\hat{V}_{ii}\hat{V}_{ij}(1-2\hat{\pi}_j)\end{aligned}$$

$$\operatorname{var}(X^2\,|\,\hat{\beta}) \simeq (1 - p/n)\Big\{2n + n\hat{p}_4 - \sum_{ij}(1-2\hat{\pi}_i)(1-2\hat{\pi}_j)\hat{V}_{ij}\Big\}$$

where V_{ij} are the elements of $\mathbf{V} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T$, the approximate covariance matrix of $\hat{\eta}$. These expressions are easy to compute following the fitting of a linear logistic model because the matrix $\mathbf{V}$ is readily available. Note that, unlike D, the conditional variance of X^2 is ordinarily not zero for pure binary data.

Similar expressions are available for the conditional cumulants of the deviance statistic but these are too complex for practical use. See, for instance, McCullagh (1986).

It is good statistical practice, however, not to rely on either D or X^2 as an absolute measure of goodness of fit in these circumstances. It is much better to look for specific deviations from the model of a type that is easily understood scientifically. For instance, we may look for interactions among the covariates or non-linear effects by adding suitable terms to the model and observing the reduction in deviance. The reduction in deviance thus induced is usually well approximated by a χ^2 distribution.

4.4.6 *Extrapolation*

Extrapolation beyond the range of the observed x-values in order to predict the probability of failure at extreme x-values is a hazardous exercise because its success depends heavily on the correctness of the assumed model, particularly on the choice of link function. It is common to find that two models that give similar predictions over the range of observed x-values may give very different predictions when extrapolated. The need for extreme extrapolation arises most commonly in reliability experiments, where failure is a rare event under naturally-occurring conditions. Experimentation is carried out under an accelerated testing regime using extreme stresses or high doses to increase the observed failure rate. For instance, in certain toxicology experiments, laboratory animals are subjected to unusually high doses of a suspected toxin or carcinogen. On the basis of the observed responses at high dose levels, it is required either to predict the failure rate at much lower dose levels, or to set confidence limits on the dose x_0 that would produce an acceptably low failure rate, π_0, the so-called maximum safe dose or maximum acceptable dose.

Suppose, by way of example, that the observed dose levels in log units and the responses are as shown in Table 4.3. It is required to predict the failure rate at dose levels equal to 1/50 unit and 1/100 unit, corresponding on the log scale to $x = -3.912$ and -4.605 respectively. On fitting the model

$$g(\pi) = \beta_0 + \beta_1 x$$

for various choices of $g(\pi)$, we find the fitted probabilities as shown in Table 4.3. On treating $(\hat\beta_0, \hat\beta_1)$ as bivariate Normal with covariance matrix (4.18), we find the predicted failure rates and confidence intervals as shown in Table 4.4. Clearly, the predicted failure probabilities are heavily dependent on the choice of link function, although the fitted probabilities in Table 4.3 are almost identical for the four link functions.

Table 4.3 *Hypothetical responses in a toxicology experiment*

Dose	*Response*	*Fitted probability*			
(*log units*)	y/m	logit	probit	c-log log	log-log
0	3/10	0.280	0.281	0.288	0.278
1	5/10	0.540	0.540	0.519	0.558
2	8/10	0.780	0.782	0.793	0.766

The converse problem of setting approximate confidence intervals for the dose x_0 that gives rise to a failure probability π_0 is most easily accomplished using Fieller's method. The linear combination

$$\hat\beta_0 + \hat\beta_1 x_0 - g(\pi_0)$$

Table 4.4 *Failure rates predicted by four models at low doses*

			90% *Confidence interval*	
x_0	link	$\hat{\pi}(x_0)$	g-scale	π-scale
-3.912	logit	0.00513	$(-9.47, -1.07)$	$(7.7\times10^{-5}, 0.26)$
	probit	0.00061	$(-5.72, -0.75)$	$(5.3\times10^{-7}, 0.23)$
	c-log log	0.01684	$(-7.04, -1.11)$	$(8.8\times10^{-4}, 0.28)$
	log log	1.1×10^{-12}	$(-6.13, -0.50)$	$(10^{-200}, 0.19)$
-4.605	logit	0.00239	$(-10.82, -1.25)$	$(2.0\times10^{-5}, 0.22)$
	probit	0.00011	$(-6.54, -0.87)$	$(3.1\times10^{-9}, 0.19)$
	c-log log	0.00994	$(-7.76, -1.26)$	$(3.5\times10^{-4}, 0.25)$
	log log	2.3×10^{-21}	$(-7.09, -0.63)$	$(10^{-522}, 0.15)$

is approximately Normally distributed with mean 0 and variance

$$v^2(x_0) = \mathrm{var}(\hat{\beta}_0) + 2x_0 \,\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) + x_0^2 \,\mathrm{var}(\hat{\beta}_1)$$

The resulting confidence 'interval' is the set of all x_0-values satisfying

$$\left| \frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - g(\pi_0)}{v(x_0)} \right| < k^*_{\alpha/2} \qquad (4.19)$$

where $\Phi(k^*_\alpha) = 1-\alpha$. The set (4.19) may be a finite interval, a semi-infinite interval or the complement of an interval. The numerical values produced by (4.19) are again heavily dependent on the choice of link function.

In practice, it is usually a good idea to compute the set (4.19) for a suitable selection of link functions. Only if these are in reasonable agreement can any real confidence be placed in the predictions.

An alternative method for constructing approximate confidence intervals for x_0 using the likelihood function directly is outlined in Exercises 4.19 and 4.20.

4.5 Over-dispersion

4.5.1 *Genesis*

By the term 'over-dispersion', we mean that the variance of the response Y exceeds the nominal variance — in this case the nominal binomial variance, $m\pi(1-\pi)$. Over-dispersion is not uncommon in practice. In fact, some would maintain that over-dispersion is the norm in practice and nominal dispersion the exception. The incidence and the degree of over-dispersion encountered greatly depend on the field of application. In large-scale epidemiological studies concerning geographical variation in the incidence of disease, the binomial variance is often an almost negligible component of the total variance. Unless there are good external reasons for relying on the binomial assumption, it seems wise to be cautious and to assume that over-dispersion is present to some extent unless and until it is shown to be absent.

Over-dispersion can arise in a number of ways. The simplest, and perhaps the most common mechanism, is clustering in the population, a mechanism previously proposed by Lexis (1879): see Stigler (1986, p. 229–238). Families, households, litters, colonies and neighbourhoods are common instances of naturally-occurring clusters in populations. Clusters usually vary in size, but we shall assume for simplicity that the cluster size, k, is fixed and that the m individuals sampled actually come from m/k clusters. In the ith cluster, the number of positive respondents, Z_i, is assumed to have the binomial distribution with index k and parameter π_i, which varies from cluster to cluster. Thus, the total number of positive respondents is

$$Y = Z_1 + Z_2 + \ldots + Z_{m/k}.$$

If we write $E(\pi_i) = \pi$ and $\mathrm{var}(\pi_i) = \tau^2\pi(1-\pi)$, it may be shown that the unconditional mean and variance of Y are

$$\begin{aligned} E(Y) &= m\pi \\ \mathrm{var}(Y) &= m\pi(1-\pi)\{1+(k-1)\tau^2\} \\ &= \sigma^2 m\pi(1-\pi). \end{aligned} \tag{4.20}$$

Note that the dispersion parameter $\sigma^2 = 1 + (k-1)\tau^2$ depends on the cluster size and on the variability of π from cluster to cluster, but not on the sample size, m. This is important because it enables us to proceed as if the observations were binomially distributed and to estimate the dispersion parameter from the residuals.

Over-dispersion can occur only if $m > 1$. If $m = 1$, the mean necessarily determines the variance and all higher-order cumulants. In general, the preceding derivation via cluster sampling forces the dispersion parameter to lie in the interval

$$1 \le \sigma^2 \le k \le m$$

because $0 \leq \tau^2 \leq 1$. It is often desirable, in order to accommodate under-dispersion, to extend the domain of definition to include values of σ^2 in the interval $0 \leq \sigma^2 \leq 1$.

The beta-binomial distribution (Exercise 4.17), is sometimes used as an alternative model for over-dispersion. This distribution has the property that the variance ratio $\mathrm{var}(Y)/\{m\pi(1-\pi)\}$ is a linear function of m, rather than a constant as in (4.20). By plotting residuals against m it is possible, in principle at least, to discriminate between these two models. The examples that we have examined, however, seem to favour the constant dispersion factor in (4.20) over the beta-binomial model.

4.5.2 *Parameter estimation*

With specific forms of over-dispersion, such as that described in Exercise 4.17 leading to the beta-binomial model, one can use maximum likelihood to estimate the regression parameters and the dispersion parameter jointly. Though this is an attractive option from a theoretical standpoint, in practice it seems unwise to rely on a specific form of over-dispersion, particularly where the assumed form has been chosen for mathematical convenience rather than scientific plausibility. For that reason, in what follows we assume that the effect of over-dispersion is as shown in (4.20). In other words, the mean is unaffected but the variance is inflated by an unknown factor σ^2.

With this form of over-dispersion, the models described in section 4.3 may still be fitted using the methods of section 4.4, as if the binomial distribution continued to apply. The only difference occurs in section 4.3.3 where the χ^2_{n-p} and the χ^2_1 approximations are replaced by $\sigma^2\chi^2_{n-p}$ and $\sigma^2\chi^2_1$ respectively. In section 4.4.4, the covariance matrix of $\hat{\boldsymbol{\beta}}$ is replaced by

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}) \simeq \sigma^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}. \tag{4.21}$$

For further details, see Chapter 9.

The expressions for bias and skewness given in section 4.4.4 are not valid without further assumptions concerning the effect of over-dispersion on the higher-order cumulants: see, for instance, Exercise 4.18.

There remains only the problem of estimating the dispersion factor, which is required for setting confidence limits on $\boldsymbol{\beta}$ and on components of $\boldsymbol{\beta}$. This is exactly analogous to the problem of estimating σ^2 in ordinary Normal-theory linear or non-linear models. Suppose first that there is replication: in other words, for each covariate value $\mathbf{x}$, several observations $(y_1, m_1), \ldots, (y_r, m_r)$ are observed. These observations are independent and essentially identically distributed apart from the fact that the indices $m_1, \ldots, m_r$ may be unequal. The estimate of π based on this covariate class alone is

$$\tilde{\pi} = y_{\cdot}/m_{\cdot}\,.$$

and the expected value of the within-class weighted sum of squares

$$\sum_{j=1}^{r} (y_j - m_j\tilde{\pi})^2/m_j$$

is equal to $(r-1)\sigma^2\pi(1-\pi)$. In other words,

$$s^2 = \frac{1}{r-1}\sum_j \frac{(y_j - m_j\tilde{\pi})^2}{m_j\tilde{\pi}(1-\tilde{\pi})} \qquad (4.22)$$

is an approximately unbiased estimator of σ^2 on $r-1$ degrees of freedom. On pooling together these estimators, one for each covariate class in which replication occurs, we obtain the replication estimate of dispersion on $\sum(r-1)$ degrees of freedom. This estimator has a slight bias of order $O(m_{\cdot}^{-1})$ in the binomial case (for $\sigma^2 = 1$) and has comparable bias otherwise. The value of the replication estimate of σ^2 is independent of the fitted model.

In the absence of replication, or if the number of degrees of freedom for replication is small, an estimate of σ^2 may be based on the residual sum of squares appropriately weighted. If the fitted model is correct,

$$\tilde{\sigma}^2 = \frac{1}{n-p}\sum_i \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)} = X^2/(n-p) \qquad (4.23)$$

is approximately unbiased for σ^2 provided that p is small compared with n. The estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ is then

$$\text{estimated var}(\hat{\boldsymbol{\beta}}) = \tilde{\sigma}^2(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}.$$

Note that if $m_i = 1$ for each i we must have $\sigma^2 = 1$. The replication estimator (4.22) has this property to a close approximation but (4.23) based on Pearson's statistic does not.

The alternative estimator of σ^2 based on the normalized residual deviance is approximately equal to $\tilde{\sigma}^2$ in the non-sparse case for which all m_i are large. In the sparse case, however, $\tilde{\sigma}^2$ is consistent for σ^2 whereas $D(\mathbf{y}; \hat{\boldsymbol{\mu}})/(n-p)$ is not. The latter claim is evident from the discussion in section 4.4.5. For instance, if $Y_i \sim B(1, \pi)$ for each i, (4.23) and (4.22) give

$$\tilde{\sigma}^2 = s^2 = n/(n-1),$$

which tends to unity as n becomes large. By contrast,

$$\frac{D}{n-1} = -\frac{2n}{n-1}\{\hat{\pi}\log\hat{\pi} + (1-\hat{\pi})\log(1-\hat{\pi})\},$$

whose value ranges from 0 to $1.386 = 2\log 2$ as $\hat{\pi}$ ranges from 0 to 0.5.

The approximate bias and variance of $\tilde{\sigma}^2$ in the absence of over-dispersion are given in section 4.4.5. In the presence of over-dispersion satisfying (4.20), the bias of $\tilde{\sigma}^2$ is of order $O(n^{-1})$. Both the bias and the variance depend on the effect of over-dispersion on the third and fourth cumulants of Y. If the effect of over-dispersion on these cumulants is as described in Exercise 4.18, explicit expressions can be obtained for the approximate bias and variance of $\tilde{\sigma}^2$. These formulae are moderately complicated and are of limited usefulness in practice because the higher-order dispersion factors must be estimated from the available data. Even for linear models, the estimation of higher-order cumulants is seldom worthwhile unless the data are very extensive.

4.6 Example

4.6.1 *Habitat preferences of lizards*

The following data are in many ways typical of social-science investigations, although the example concerns the behaviour of lizards rather than humans. The data, taken from Schoener (1970), have subsequently been analysed by Fienberg (1970b) and by

Table 4.5 *A comparison of site preferences of two species of lizard,* grahami *and* opalinus

			T								
	Perch		*Early*			*Mid-day*			*Late*		
	D	H									
S	(in)	(ft)	G	O	*Total*	G	O	*Total*	G	O	*Total*
Sun	≤ 2	< 5	20	2	22	8	1	9	4	4	8
		≥ 5	13	0	13	8	0	8	12	0	12
	> 2	< 5	8	3	11	4	1	5	5	3	8
		≥ 5	6	0	6	0	0	0	1	1	2
Shade	≤ 2	< 5	34	11	45	69	20	89	18	10	28
		≥ 5	31	5	36	55	4	59	13	3	16
	> 2	< 5	17	15	32	60	32	92	8	8	16
		≥ 5	12	1	13	21	5	26	4	4	8

H, perch height; D, perch diameter; S, sunny/shady; T, time of day; G, *grahami*; O, *opalinus*.

Bishop *et al.* (1975). Data concerning the daytime habits of two species of lizard, *grahami* and *opalinus*, were collected by observing occupied sites or perches and recording the appropriate description, namely species involved, time of day, height and diameter of perch and whether the site was sunny or shaded. Time of day is recorded here as early, mid-day or late.

As often with such problems, several analyses are possible depending on the purpose of the investigation. We might, for example, wish to compare how preferences for the various perches vary with the time of day regardless of the species involved. We find by inspection of the data (Table 4.5) that shady sites are preferred to sunny sites at all times of day but particularly so at mid-day. Furthermore, again by inspection, low perches are preferred to high ones and small-diameter perches to large ones. There is, of course, the possibility that these conclusions are produced by an artefact of the data-collection process and that, for instance, occupied sites at eye level or below are easier to spot than occupied perches higher up. In fact, selection bias of this type seems inevitable unless some considerable effort is devoted to observing all lizards in a given area.

A similar analysis, but with the same deficiencies, can be made for each species separately.

Suppose instead that an occupied site, regardless of its position, diameter and so on, is equally difficult to spot whether occupied by a *grahami* or an *opalinus* lizard. This assumption would be plausible if the two species were similar in size and colour. Suppose in addition that the purpose of the investigation is to compare the two species with regard to their preferred perches. Thus we see that, of the 22 occupied perches of small diameter low in the tree observed in a sunny location early in the day, only two, or 9%, were occupied by *opalinus* lizards. For similar perches observed later in the day, the proportion is four out of eight, i.e. 50%. On this comparison, therefore, it appears that, relative to *opalinus*, *grahami* lizards prefer to sun themselves early in the day.

To pursue this analysis more formally, we take as fixed the total number m_{ijkl} of occupied sites observed for each combination of i = perch height, j = perch diameter, k = sunny/shady and l = time of day. In the language of statistical theory, these totals or covariate-class sizes are *ancillary* provided that the purpose of the investigation is to compare preferences or to examine the differences between the site preferences of the two species. The response variable y_{ijkl} gives the observed number or, equivalently, the observed proportion of the m_{ijkl} occupied sites that were occupied by *grahami* lizards. By symmetry, we could equally well work with $m_{ijkl} - y_{ijkl}$, the number of sites occupied by *opalinus* lizards. We take the random variable Y_{ijkl} to be binomially distributed with index m_{ijkl} and parameter π_{ijkl}. Thus π is the probability that an observed occupied site is in fact occupied by a *grahami* lizard. Of course, the possibility of over-dispersion relative to the binomial distribution must be borne in mind.

At the exploratory stage, probably the simplest analysis of these data is obtained by transforming to the logistic scale. Using the empirical logistic transformation (4.7), we have the transformed value for $y_1/m_1 = 20/22$, namely

$$z_1 = \log(20.5/2.5) = 2.1041$$

with approximate variance $1/20.5 + 1/2.5 = 0.4488$. A straightforward linear analysis of the transformed values is usually a satisfactory method of analysis if all the observed counts are moderately large. In this example not all the counts are large and for that reason, we must confirm our findings using a different technique. To

Table 4.6 *Computation of logistic factorial standardized contrasts for lizard data*

Transformed value	*Estimated variance*	*Raw contrast*	*Estimated variance*	*Parameter*	*Absolute standardized contrast*
2.1041	0.4488	30.9092	20.16	I	—
3.2958	2.0741	11.0986	20.16	H	2.47
0.8873	0.4034	−12.4508	20.16	D	2.77
2.5649	2.1538	−5.3629	20.16	HD	1.19
1.0986	0.1159	−5.4698	20.16	S	1.22
1.7451	0.2136	−0.1739	20.16	HS	0.04
0.1214	0.1217	3.9168	20.16	DS	0.87
2.1203	0.7467	5.4014	20.16	HDS	1.20
1.7346	0.7843	−8.3505	11.79	T_L	2.43
2.8332	2.1176	−1.9645	11.79	HT_L	0.57
1.0986	0.8889	−2.1333	11.79	DT_L	0.62
0.0000	4.0000	−6.2926	11.79	HDT_L	1.83
1.2209	0.0632	2.0122	11.79	ST_L	0.59
2.5123	0.2402	−1.7595	11.79	HST_L	0.51
0.6214	0.0473	−0.4950	11.79	DST_L	0.14
1.3633	0.2283	2.0210	11.79	$HDST_L$	0.59
0.0000	0.4444	−3.2438	45.27	T_Q	0.48
3.2189	2.0900	4.9987	45.27	HT_Q	0.74
0.4520	0.4675	3.2022	45.27	DT_Q	0.48
0.0000	1.3333	2.8773	45.27	HDT_Q	0.43
0.5664	0.1493	−5.6243	45.27	ST_Q	0.84
1.3499	0.3598	−6.2738	45.27	HST_Q	0.93
0.0000	0.2353	−1.2453	45.27	DST_Q	0.19
0.0000	0.4444	0.4582	45.27	$HDST_Q$	0.07

maintain balance, the observation $(0, 0)$ is transformed to $z_{12} = 0.0$ with 'variance' 4.0.

The first two columns of Table 4.6 give the transformed values and their estimated variances listed in the usual standard order corresponding to the factors H, D, S and T. Four steps of Yates's algorithm (not given) produce the raw contrasts, again associated with the four factors in the same standard order. In the case of the factor T, which has three ordered levels, linear and quadratic contrasts were used to complete the decomposition. Variances are computed in a similar way, the coefficients being squared. Thus, all main effects and interactions involving H, D and S only have the

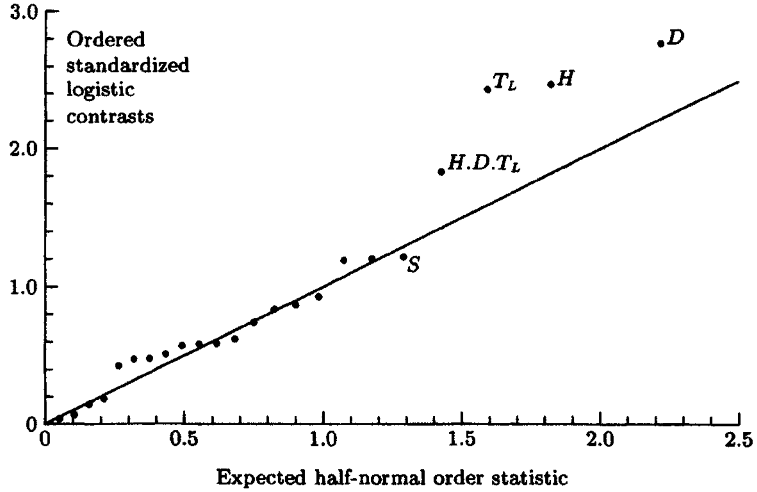

Fig. 4.2. *Half-normal plot of ordered absolute standardized logistic contrasts for the lizard data. The solid line is at* 45°.

same variance, 20.16, which is the total of column 2. Similarly for terms involving T_L and for terms involving T_Q. Finally we compute the standardized contrasts: of these, only the main effects of H and D and the linear effect of time, with standardized contrasts in excess of 2.4, appear to be significant.

A half-Normal plot (Daniel, 1959) of the ordered absolute standardized logistic contrasts against their Normal-theory expected values (Fig. 4.2), suggests that the main effects of height and diameter and the linear effect of time are significant though not overwhelmingly so. The three-factor interaction $H.D.T_L$ also deviates from the theoretical line, but this appears to be an aberration well within the sampling limits especially when due allowance is made for the effect of selection. As a matter of policy, no allowance for selection would normally be made when judging the significance of main effects in a full factorial design. Such effects that are not expected in advance to be null should be excluded from the half-normal plot, though this has not been done in Fig. 4.2.

The unit slope observed in Fig. 4.2 is evidence that $\sigma = 1$ and hence there is no suggestion of over-dispersion.

Because of the numerous small observed counts in this particular example, some caution is required in the interpretation of contrasts in Table 4.6. It is possible, for example, that the addition of 1/2 to each count before transforming could swamp the data. Indeed such an effect seems to have occurred for the sunny/shady contrast. Here, few *opalinus* lizards were observed in sunny locations, so that the addition of 1/2 to each of these counts has a marked effect on the S contrast, reducing it towards zero and so diluting its apparent significance.

We now consider an alternative analysis using a generalized linear model fitted by maximum likelihood, which avoids transformation problems. The preceding analysis in Table 4.6 and Fig. 4.2 suggests that the structure of these data is fairly simple; there appear to be no strong interactions on the logistic scale. We are therefore led initially to consider the linear logistic model including all four main effects. Such a model can be written symbolically as $H + D + S + T$ or, in subscript notation,

$$\operatorname{logit}(\pi_{ijkl}) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l, \qquad (4.24)$$

where α, β, γ and δ refer to the four factors H, D, S and T. In fact this model fits the data quite well with no evidence of over-dispersion. All main effects including S are significant at the 5% level, illustrating the drawbacks of the previous analysis where S appeared to be insignificant. None of the two-factor interactions appears significant; the relevant statistics are given in Table 4.7. Parameter estimates associated with the model (4.24) are given in Table 4.8, where we use the convention of setting the first level of each factor to zero. It is possible here to replace T by a single contrast corresponding to late afternoon versus earlier in the day without unduly affecting the fit. This replacement reduces the number of parameters by one but does not greatly simplify the model or statements of conclusions.

Finally, an informal examination of the standardized residuals reveals no unexpected features or patterns.

The principal conclusions to be drawn are as follows. An occupied high perch is more likely to be occupied by a *grahami* lizard than is an occupied low perch. The ratio of the odds for high versus low perches is an estimated 3.10 = exp(1.13), and this ratio applies under all conditions of shade, perch diameter and time of day. It would be false to conclude from this analysis that *grahami* lizards prefer high perches to low perches. We may,

Table 4.7 *Examination of two-factor interactions for lizard data*

*Model description**	*Degrees of freedom*	*Deviance*	*First difference*
Main effects only	17†	14.20	
Main + $T.S$	15	12.93	1.27
Main + $T.H$	15	13.68	0.52
Main + $T.D$	15	14.16	0.04
Main + $S.H$	16	11.98	2.22
Main + $S.D$	16	14.13	0.07
Main + $H.D$	16	13.92	0.28

* The factors here are time of day (T), sunny/shady (S), height (H) and diameter (D).

† Degrees of freedom are reduced by one because no occupied sites were observed for $(i, j, k, l) = (2, 2, 2, 2)$.

Table 4.8 *Parameter estimates for the linear logistic model* (4.24)

Parameter	*Estimate*	*S.E.*
μ	1.945	0.34
H, height > 5ft	1.13	0.26
D, diameter > 2in	−0.76	0.21
S, shady	−0.85	0.32
$T(2)$, mid-day	0.23	0.25
$T(3)$, late	−0.74	0.30

however, conclude that *grahami* lizards have less aversion to high perches than do *opalinus* lizards, so that an occupied high perch is more likely to contain a *grahami* lizard than an occupied low perch.

Similar conclusions may be drawn regarding the effect of shade, perch diameter and time of day on the probability that an occupied site contains a *grahami* lizard. The odds are largest for small-diameter lofty perches observed in a sunny location at mid-day (or in the morning). In fact, only *grahami* and no *opalinus* lizards were observed under these conditions. Because there is no interaction among the effects, the odds are smallest for the converse factor combinations.

These conclusions differ from those of Fienberg (1970b) and Bishop *et al.* (1975), who found an interaction between H and D and between S and T regarding their effect on species' preferences. The principal reason for this difference appears to be the fact that these authors attempted to consider several unrelated issues simultaneously using only a single model, and did not condition on the totals m_{ijkl}, which are regarded as ancillary in the analysis given here.

4.7 Bibliographic notes

The statistical literature on the analysis of discrete data is very extensive and there is a wealth of excellent text-books treating the subject from a number of different angles. Cox (1970) offers a good introduction to the subject and combines a pleasant blend of theory and application in a slim volume. Plackett (1981) is a good introductory text covering much of the material in this chapter and in the following three chapters, but with a slightly different emphasis. Breslow and Day (1980) concentrate on applications in cancer research. Fleiss (1981) discusses applications in the health sciences generally. Haberman (1978, 1979) concentrates mainly on social-science applications. Engel (1987) gives an extensive discussion of over-dispersion.

There is some overlap with the survival-theory literature, where success is sometimes defined rather arbitrarily as two-year or five-year survival: see, for example, Kalbfleisch and Prentice (1980) or Cox and Oakes (1984).

Other books dealing partially or wholly with binary data include Adena and Wilson (1982), Aickin (1983), Armitage (1971), Ashton (1972), Bishop, Fienberg and Holland (1975), Bock (1975), Everitt (1977), Fienberg (1980), Finney (1971), Gokhale and Kullback (1978), Maxwell (1961), Plackett (1981) and Upton (1978).

4.8 Further results and exercises 4

4.1 Suppose that $Y_1, \ldots, Y_m$ are independent Bernoulli random variables for which

$$\mathrm{pr}(Y_i = 0) = 1 - \pi \quad \text{and} \quad \mathrm{pr}(Y_i = 1) = \pi.$$

Show that any fixed sequence comprising y ones and $m - y$ zeros has probability $\pi^y(1 - \pi)^{m-y}$. Hence deduce that the total $Y_{.} =$

$Y_1 + \ldots + Y_m$ has the binomial distribution (4.2) with index m and parameter π.

4.2 Suppose that $Y_1 \sim B(m_1, \pi)$ and $Y_2 \sim B(m_2, \pi)$ are independent. Deduce from Exercise 4.1 that $Y_{\bf .} \sim B(m_{\bf .}, \pi)$.

4.3 Suppose that $Y_1 \sim B(m_1, \pi_1)$ and $Y_2 \sim B(m_2, \pi_2)$ are independent. Show that

$$\mathrm{pr}(Y_{\bf .} = y_{\bf .}) = (1-\pi_1)^{m_1} \pi_2^{y_{\bf .}} (1-\pi_2)^{m_2 - y_{\bf .}} P_0(\psi;\ m_1, m_2, y_{\bf .}),$$

where $\psi = \pi_1(1-\pi_2)/\{\pi_2(1-\pi_1)\}$ is the odds ratio and $P_0(\psi;\ \cdot)$ is the polynomial in ψ

$$P_0(\psi;\ m_1, m_2, y_{\bf .}) = \sum_{j=a}^{b} \binom{m_1}{j} \binom{m_2}{y_{\bf .} - j} \psi^j.$$

The range of summation extends from $a = \max(0, y_{\bf .} - m_2)$ to $b = \min(m_1, y_{\bf .})$. Show also that

$$P_0(1;\ m_1, m_2, y_{\bf .}) = \binom{m_{\bf .}}{y_{\bf .}},$$

which is consistent with the previous exercise.

4.4 Suppose that Y_1, Y_2 are independent Poisson random variables with means μ and $\rho\mu$ respectively. Show that

$$\begin{aligned} Y_{\bf .} = Y_1 + Y_2 &\sim P(\mu + \rho\mu) \\ Y_1 \,|\, Y_{\bf .} = m &\sim B(m,\ 1/(1+\rho)). \end{aligned}$$

Show how you might use this result to test the composite null hypothesis $H_0\colon \rho = 1$ against the one-sided alternative $H_A\colon \rho > 1$.

4.5 Let $Y_1, \ldots, Y_n$ be independent random variables such that $Y_i \sim B(m, \pi_i)$ and let $Y = \sum Y_i$ be the sum. Show that, given $\pi_1, \ldots, \pi_n$,

$$\begin{aligned} E(Y) &= m_{\bf .}\bar{\pi} \\ \mathrm{var}(Y) &= m_{\bf .}\bar{\pi}(1-\bar{\pi}) - m(n-1)k_2(\pi) \end{aligned}$$

where $m_{\bf .} = nm$. Give the expression for $k_2(\pi)$ in terms of $\pi_1, \ldots, \pi_n$.

4.6 In the notation of the previous exercise, assume that $\pi_1, \ldots, \pi_n$ are independent random variables with common mean π and common variance $\tau^2\pi(1-\pi)$. Show that, unconditionally,

$$E(Y) = m_{\bullet}\pi$$
$$\operatorname{var}(Y) = m_{\bullet}\pi(1-\pi)\{1+(m-1)\tau^2\}$$

Deduce that $0 \le \tau^2 \le 1$, so that $\operatorname{var}(Y) \ge m_{\bullet}\pi(1-\pi)$.

4.7 Define

$$B(y) = \binom{m}{y}\pi^y(1-\pi)^{m-y},$$
$$P(y) = e^{-\mu}\mu^y/y!.$$

Let $\pi = \mu/m$. Show that, for fixed μ, as $m - y \to \infty$,

$$\frac{B(y)}{P(y)} \simeq \left(\frac{m}{m-y}\right)^{1/2}.$$

4.8 Suppose that $Y \sim B(m, \pi)$ and that m is large. By expanding in a Taylor series, show that the random variable

$$Z = \arcsin\{(Y/m)^{1/2}\}$$

has approximate first two moments

$$E(Z) \simeq \arcsin(\pi^{1/2}) - \frac{1-2\pi}{8\sqrt{m(1-\pi)}}$$
$$\operatorname{var}(Z) \simeq (4m)^{-1}.$$

4.9 Let $K(\theta)$ be a *cumulant function* such that the rth cumulant of X is the rth derivative of $mK(\theta)$. Let $\mu = mK'(\theta)$ be the mean of X and let $\kappa_2(\mu), \kappa_3(\mu)$ be the variance and third cumulant respectively of X, expressed in terms of μ rather than in terms of θ. Show that

$$\kappa_3(\mu) = \kappa_2(\mu)\kappa_2'(\mu) \quad \text{and} \quad \frac{\kappa_3}{\kappa_2^2} = \frac{d}{d\mu}\log\kappa_2(\mu).$$

Verify that the binomial cumulants have this form with

$$K(\theta) = \log(1+e^\theta).$$

4.10 Show that if the cumulants of X are all $O(m)$ for large m, then $Y = g(\bar{X})$ is approximately symmetrically distributed if $g(\cdot)$ satisfies the second-order differential equation

$$3\kappa_2^2(\mu)g''(\mu) + g'(\mu)\kappa_3(\mu) = 0.$$

Show that if $\kappa_2(\mu)$ and $\kappa_3(\mu)$ are related as in the previous exercise, then

$$g(x) = \int^x \kappa_2^{-1/3}(\mu)\,d\mu.$$

[N.B. $\kappa_2(\mu)$ is the variance function denoted by $V(\mu)$ in section 2.2: $\kappa_3(\mu)$ is an obvious extension.]

4.11 Find the corresponding equations that give the variance-stabilizing transformation of X.

4.12 *Logistic discrimination:* Suppose that a population of individuals is partitioned into two sub-populations or groups, G_1 and G_2, say. It may be helpful to think of G_1 in a epidemiological context as the carriers of a particular virus, comprising $100\pi_1$% of the population, and G_2 as the non-carriers. Measurements Z made on individuals have the following distributions in the two groups:

$$\begin{aligned} G_1&: \quad Z \sim N_p(\boldsymbol{\mu}_1, \Sigma) \\ G_2&: \quad Z \sim N_p(\boldsymbol{\mu}_2, \Sigma). \end{aligned}$$

Let $\mathbf{z}^*$ be an observation made on an individual drawn at random from the combined population. The prior odds that the individual belongs to G_1 are $\pi_1/(1-\pi_1)$. Show that the posterior odds given $\mathbf{z}^*$ are

$$\text{odds}(Y = 1 \mid \mathbf{Z}^*) = \frac{\pi_1}{1-\pi_1} \times \exp(\alpha + \boldsymbol{\beta}^T\mathbf{z}^*)$$

where the logistic regression coefficients are given by

$$\begin{aligned} \alpha &= \tfrac{1}{2}\boldsymbol{\mu}_2^T\Sigma^{-1}\boldsymbol{\mu}_2 - \tfrac{1}{2}\boldsymbol{\mu}_1^T\Sigma^{-1}\boldsymbol{\mu}_1 \\ \boldsymbol{\beta} &= \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \end{aligned}$$

Comment briefly on the differences between maximum likelihood estimation of α and $\boldsymbol{\beta}$ via the Normal-theory likelihood and estimation via logistic regression. [Efron, 1975].

4.13 Go through the calculations of the previous exercise, replacing the Normal distributions by exponential distributions having the same means.

4.14 Suppose that $Y \sim B\big(m, e^{\lambda}/(1+e^{\lambda})\big)$. Show that $m - Y$ also has the binomial distribution and that the induced parameter is $\lambda' = -\lambda$. Consider

$$\tilde{\lambda} = \log\left(\frac{Y + c_1}{m - Y + c_2}\right)$$

as an estimator of λ. Show that, in order to achieve consistency under the transformation $Y \rightarrow m - Y$, we must have $c_1 = c_2$.

4.15 Using the notation of the previous exercise, write

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

where $Z = O_p(1)$ for large m. Show that

$$E\big\{\log(Y + c)\big\} = \log(m\pi) + \frac{c}{m\pi} - \frac{1-\pi}{2m\pi} + O(m^{-3/2}).$$

Find the corresponding expansion for $E\big\{\log(m - Y + c)\big\}$ Hence, if $c_1 = c_2 = c$, deduce that

$$E(\tilde{\lambda}) = \lambda + \frac{(1-2\pi)(c-\frac{1}{2})}{m\pi(1-\pi)} + O(m^{-3/2}).$$

[Cox, 1970, section 3.2].

4.16 Suppose that $Y_1, \ldots, Y_r$ are independent and that $Y_i \sim B(m_i, \pi)$. Show that the maximum-likelihood estimate is $\hat{\pi} = Y_{\cdot}/m_{\cdot}$, and

$$s^2 = \frac{1}{r-1}\sum_i (Y_i - m_i\hat{\pi})^2/\{m_i\hat{\pi}(1-\hat{\pi})\}$$

has expectation

$$E(s^2) = \frac{m_{\cdot}}{m_{\cdot} - 1}.$$

Hence show how the estimator (4.22) may be modified to eliminate bias in the null case of no dispersion. [Haldane, 1937].

4.17 Show that if $Y\,|\,P \sim B(m,p)$, where P has the beta distribution

$$f_P(p) = p^{\alpha-1}(1-p)^{\beta-1}/B(\alpha,\beta), \qquad (0 \le p \le 1),$$

then Y has the beta-binomial distribution

$$\mathrm{pr}(Y=y) \;=\; \binom{m}{y}\frac{B(\alpha+y,\, m+\beta-y)}{B(\alpha,\beta)}$$

for $y = 0,\ldots,m$ and $\alpha,\beta > 0$. Show that

$$\begin{aligned} E(Y) &= m\pi \quad \text{and} \\ \mathrm{var}(Y) &= m\pi(1-\pi)\{1+(m-1)\tau^2\} \end{aligned}$$

and express π and τ^2 in terms of α and β. [Crowder, 1978; Plackett, 1981 p.58; Williams, 1982; Engel, 1987].

4.18 Suppose, following the cluster-sampling mechanism described in section 4.5.1, that

$$Y = Z_1 + Z_2 + \ldots + Z_{m/k}$$

where $Z_i \sim B(k,\pi_i)$ are independent. Assume in addition that the cluster probabilities are independent random variables satisfying

$$E(\pi_i) = \pi, \quad \mathrm{var}(\pi_i) = \tau_2\pi(1-\pi), \quad \kappa_3(\pi_i) = \tau_3\pi(1-\pi)(1-2\pi).$$

Show that the marginal cumulants of Y are

$$\begin{aligned} E(Y) &= m\pi \\ \mathrm{var}(Y) &= m\pi(1-\pi)\{1+(k-1)\tau_2\} \\ \kappa_3(Y) &= m\pi(1-\pi)(1-2\pi)\{1+3(k-1)\tau_2+(k-1)(k-2)\tau_3\}. \end{aligned}$$

[With obvious extensions, similar calculations for the fourth cumulant give

$$\begin{aligned} \kappa_4(Y) = m\pi(1-\pi)\big\{&1+7(k-1)\tau_2+6(k-1)(k-2)\tau_3 \\ &+(k-1)(k-2)(k-3)\tau_4\big\} \\ -6m\pi^2(1-\pi)^2\big\{&1+6(k-1)\tau_2+4(k-1)(k-2)\tau_3 \\ &+(k-1)(k-2)(k-3)\tau_4+(k-1)(2k-3)\tau_2^2\big\}, \end{aligned}$$

breaking the early pattern.]

4.19 Consider the dose-response model

$$g(\pi) = \beta_0 + \beta_1 x.$$

Under the hypothesis that the response probability at x_0 is equal to π_0, show that the model reduces to

$$g(\pi) = \beta_0(1 - x/x_0) + g(\pi_0)x/x_0\,.$$

How would you fit such a model using your favourite computer program?

4.20 Let $D(x_0)$ be the residual deviance under the reduced model in the previous exercise. How would you use a graph of $D(x_0)$ against x_0 to construct an approximate confidence set for the parameter x_0? For the logistic model, compute this interval using the data in Table 4.3 for $\pi_0 = 0.01$. Compare the answer with that given by (4.19).

4.21 Suppose that in a given population, the proportions of individuals in the various categories are as shown in Table 4.2. In a prospective study, 100 subjects in each of the exposure groups are observed over the requisite period. Find the expected numbers in each of the four cells. Show that the estimate of the log odds-ratio has approximate variance

$$\text{var}(\hat{\Delta}_1) \simeq 0.472.$$

In a retrospective study, 100 cases with the disease and 100 disease-free controls are obtained. Their exposure status is subsequently ascertained. Find the expected numbers in the four cells. Show that the estimate of the log odds-ratio has approximate variance

$$\text{var}(\hat{\Delta}_2) \simeq 0.093.$$

Hence compute the relative efficiency of the retrospective design.

4.22 Show that the logistic density

$$f_X(x) = \exp(x)/[1 + \exp(x)]^2$$

is symmetrical about zero. Find the cumulative distribution function and show that the 100p percentile occurs at

$$x_p = \log(p/(1 - p)).$$

Show that the moment generating function of X is

$$M_X(t) = \pi t/\sin(\pi t) = \Gamma(1+t)\Gamma(1-t).$$

for $-1 < t < 1$. Hence deduce that the even cumulants of X are

$$\kappa_2 = \pi^2/3, \quad \kappa_4 = 2\pi^4/15, \quad \kappa_6 = 16\pi^6/63, \quad \kappa_8 = 16\pi^8/15, \ldots .$$

Deduce that $\kappa_{2r} \sim 2(2r-1)!\,\{1 + 2^{-2r}\}$ for large r. Check this approximation numerically for the cumulants listed above.

The exact cumulants are given by the series expansion

$$\kappa_{2r} = 2(2r-1)!\,\zeta(2r) = 2(2r-1)!\,\{1 + 2^{-2r} + 3^{-2r} + 4^{-2r} + \ldots\},$$

where $\zeta(x)$ is the Riemann zeta function.

4.23 Let X be a unit exponential random variable. Show that the density function of $Y = \log X$ is

$$f_Y(y) = \exp(y - e^y) \quad \text{for} \quad -\infty < y < \infty.$$

Plot the density. Find the cumulative distribution function and obtain an expression for the $100p$ percentile.

Show that the moment generating function of Y is

$$M_Y(t) = \Gamma(1+t).$$

Hence deduce that the cumulants of Y are

$$\kappa_{r+1}(Y) = \psi^{(r)}(1) = (-1)^{r+1} r!\,\zeta(r+1).$$

Show in particular that the first four cumulants are

$$\kappa_1 = -\gamma \simeq -0.57721, \quad \kappa_2 = \pi^2/6, \quad \kappa_3 = -2.40411, \quad \kappa_4 = \pi^4/15.$$

Comment briefly on the relation between the even cumulants of Y and those of the logistic density.

Table 4.9 *Number of eggs recovered after* 2 *days out of* 50 *of each type*[†]

Adult species											
A				*B*				*C*			
	egg species				*egg species*				*egg species*		
	a	*b*	*c*		*a*	*b*	*c*		*a*	*b*	*c*
	25	24	15		25	15	22		35	21	28
day 1	26	14	26	*day* 1	31	22	33	*day* 1	36	19	34
	26	24	32		24	12	30		33	16	31
	29	14	32		14	8	13		24	24	23
day 2	28	13	19	*day* 2	18	12	20	*day* 2	38	24	27
	27	19	16		–	–	–		34	36	27
	26	10	13		13	6	14				
day 3	20	7	15	*day* 3	18	11	19				
	14	14	23		8	5	12				

[†]Data courtesy of Mr S. Teleky, University of Chicago.

4.24 Show how you would use your friendly computer program to compute the approximate conditional mean and variance of Pearson's statistic using the formulae given at the end of section 4.4.5. [Hint: express $\sum_j V_{ij}(1 - 2\pi_j)$ as the vector of fitted values in a supplementary weighted linear regression problem. This step is unnecessary if your friendly program permits matrix multiplication.]

4.25 Beetles of the genus *Tribolium* are cannibalistic in the sense that adults eat the eggs of their own species as well as those of closely related species. Any species whose adults can recognize and avoid eggs of their own species while foraging has a distinct evolutionary advantage. Table 4.9 presents the results of one experiment conducted at the University of Chicago by Mr S. Teleky of the Department of Evolutionary Biology. The aim of this study was to determine whether any of the three *Tribolium* species, *castaneum* (A), *confusum* (B), or *madens* (C) has evolved such an advantage.

The experimental procedure used was to isolate a number of adult beetles of the same species and to present them with a vial of 150 eggs – 50 of each type – the eggs being thoroughly mixed to ensure a uniform distribution on the vial. The number of eggs of each type remaining after two days was counted and recorded and is displayed in Table 4.9. Eggs are coded here using the lowercase letters of the adult species. Typically, several such experiments with the same adult species were run in parallel, the adults for each experiment being chosen from a large population of that species. Thus, for adult species A, three experiments were run in parallel beginning on each of three days. Days 1, 2 and 3 are not necessarily consecutive, nor is day 1 for species A the same as day 1 for species B or C.

Analyse the data bearing in mind the objective of the experiment and the design of the experiment. Make due allowance for over-dispersion in the computation of standard errors and confidence intervals. Is there any evidence that any of the three adult species has evolved a preference for eggs of the other species?

4.26 The data shown in Table 4.10 were collected by Sir Francis Galton as part of his study of natural inheritance—in this case the study of the inheritance of eye colour in human populations.

1. Set up a six-level factor, P, one level corresponding to each of the distinguishable eye-colour combinations of the two parents.
2. Set up the corresponding factor, G, for the distinguishable eye-colour combinations of the grandparents. How many levels does this factor have?
3. Fit the linear logistic model P, treating the number of light-eyed children as the binomial response. Examine the standardized residuals and set aside any obviously discrepant points.
4. Re-fit the previous model to the remaining data. Compute the fitted probabilities for all eye-colour combinations of the two parents. Arrange these fitted probabilities in a 3×3 table. Comment on any marked trends or other patterns.
5. Add the factor G to the previous model. Look for trends or other patterns among the levels of G. What evidence is there for a grandparent effect above and beyond the parental effect?
6. Outliers are often caused by transposing digits or otherwise misrecording the data. What alternative explanation can you offer for the most discrepant points in this example?
7. Is there any evidence of over-dispersion? Estimate the dispersion parameter.
8. What additional information could be extracted from these data if the eye-colours of the father and mother were separately recorded? Comment on the relevance of this information

Table 4.10 *Number of light-eyed children in each of 78 families of not less than six brothers or sisters each, classified by eye-colour of parents and grandparents.*†

Number of parents			*Number of grandparents*			*Total*	*Light-eyed*
Light	*Hazel*	*Dark*	*Light*	*Hazel*	*Dark*	*children*	*children*
2	0	0	4	0	0	6	6
2	0	0	4	0	0	6	6
2	0	0	4	0	0	6	6
2	0	0	4	0	0	6	5
2	0	0	4	0	0	7	7
2	0	0	4	0	0	7	7
2	0	0	4	0	0	7	7
2	0	0	4	0	0	7	7
2	0	0	4	0	0	7	7
2	0	0	4	0	0	8	8
2	0	0	4	0	0	8	8
2	0	0	4	0	0	8	8
2	0	0	4	0	0	8	8
2	0	0	4	0	0	8	7
2	0	0	4	0	0	8	7
2	0	0	4	0	0	12	12
2	0	0	3	1	0	7	7
2	0	0	3	1	0	10	4
2	0	0	3	1	0	12	12
2	0	0	3	0	1	7	6
2	0	0	3	0	1	8	8
2	0	0	3	0	1	9	9
2	0	0	3	0	1	9	9
2	0	0	3	0	1	9	7
2	0	0	3	0	1	10	10
2	0	0	2	2	0	7	7
2	0	0	2	2	0	10	9
2	0	0	2	1	1	6	6
2	0	0	2	1	1	10	10
0	2	0	2	1	1	7	4
0	2	0	2	0	2	8	5
0	0	2	3	0	1	6	2
0	0	2	2	0	2	9	1
0	0	2	1	0	3	6	1
0	0	2	1	0	3	11	3
0	0	2	1	1	2	6	0
0	0	2	1	1	2	7	4

Continued

Table 4.10 *Continued.*

Number of parents			*Number of grandparents*			*Total*	*Light-eyed*
Light	*Hazel*	*Dark*	*Light*	*Hazel*	*Dark*	*children*	*children*
1	1	0	3	1	0	6	6
1	1	0	3	1	0	7	6
1	1	0	3	1	0	8	6
1	1	0	3	1	0	9	7
1	1	0	3	1	0	11	10
1	1	0	3	0	1	9	6
1	1	0	3	0	1	11	7
1	1	0	2	2	0	7	6
1	1	0	2	2	0	9	9
1	1	0	2	2	0	11	1
1	1	0	2	0	2	6	6
1	1	0	2	0	2	6	4
1	1	0	2	0	2	8	5
1	1	0	2	0	2	9	7
1	1	0	2	1	1	6	6
1	1	0	2	1	1	10	9
1	1	0	1	3	0	9	4
1	1	0	1	1	2	8	5
1	0	1	4	0	0	7	3
1	0	1	3	0	1	6	4
1	0	1	3	0	1	7	3
1	0	1	3	0	1	8	6
1	0	1	3	0	1	8	5
1	0	1	3	0	1	8	4
1	0	1	3	0	1	9	6
1	0	1	3	0	1	9	5
1	0	1	2	0	2	6	5
1	0	1	2	0	2	6	3
1	0	1	2	0	2	8	4
1	0	1	2	0	2	10	2
1	0	1	2	0	2	14	9
1	0	1	2	1	1	7	5
1	0	1	1	2	1	7	3
1	0	1	1	1	2	7	4
1	0	1	1	0	3	8	4
1	0	1	1	0	3	8	3
1	0	1	0	1	3	6	3
0	1	1	2	0	2	6	3
0	1	1	2	1	1	9	4
0	1	1	1	0	3	13	8
0	1	1	0	4	0	7	2

†Source: Galton (1889, p.216–217).

(a) from a biological viewpoint and (b) from a sociological viewpoint.

9. Fit the linear logistic model G. Compute the fitted probability for each level of G. Label the levels of G appropriately and comment on any trends or patterns in the fitted probabilities.
10. Examine the residuals from the previous model. Comment briefly on any unusual patterns, particularly in families 32 and 56.

4.27 Using the notation of section 4.4.3 in which $Y_i \sim B(m_i, \pi_i)$, let $H_0 \subset H_1$ denote two nested models for the probability vector $\boldsymbol{\pi}$, with deviances $D(\mathbf{y}, \hat{\boldsymbol{\pi}}_0)$ and $D(\mathbf{y}, \hat{\boldsymbol{\pi}}_1)$ respectively. Show that, in the case of linear logistic models, the deviances satisfy the Pythagorean relationship

$$D(\mathbf{y}, \hat{\boldsymbol{\pi}}_0) = D(\mathbf{y}, \hat{\boldsymbol{\pi}}_1) + D(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\pi}}_0).$$

Hence deduce that for logistic models, but not otherwise, $D(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\pi}}_0)$ is the likelihood-ratio statistic for testing H_0 against H_1 as alternative.

4.28 In the previous exercise, suppose that H_0 and H_1 denote a constant and a single-factor model respectively. Show that the fitted values and the deviance functions are then independent of the link used for model specification. Show also that weighted squared Euclidean distance with weights m_i satisfies the Pythagorean relationship. What other discrepancy functions satisfy the Pythagorean relationship in this special case? [Efron, 1978].

4.29 The asymptotic bias of the components of $\hat{\boldsymbol{\beta}}$ in linear logistic models is given by

$$E(\hat{\beta}^r - \beta^r) \simeq -\tfrac{1}{2}\kappa^{r,s}\kappa^{t,u}\kappa_{s,t,u},$$

using the index notation of McCullagh (1987, p. 209), in which $\kappa^{r,s}$ is the inverse Fisher information matrix. Express $\kappa_{s,t,u}$ in terms of the components of the model matrix.

For small $\boldsymbol{\beta}$, justify the approximation

$$E(\hat{\boldsymbol{\beta}}) \simeq \boldsymbol{\beta} \times (1 + p/m_.),$$

showing that the bias vector is approximately collinear with the parameter vector.

4.30 Let R_i be the unobserved true response for unit i with $\pi_i^* = \mathrm{pr}(R_i = 1)$ satisfying the linear logistic model

$$\mathrm{logit}(\pi_i^*) = \boldsymbol{\beta}^T \mathbf{x}_i.$$

Suppose that the observed response is subject to mis-classification as follows.

$$\mathrm{pr}(Y_i = 1 \mid R_i = 0) = \delta_i$$
$$\mathrm{pr}(Y_i = 0 \mid R_i = 1) = \epsilon_i.$$

Show that if the mis-classification errors satisfy

$$\frac{\delta_i}{\epsilon_i} = \frac{\pi_i^*}{1 - \pi_i^*},$$

then the observed response probability $\pi_i = \mathrm{pr}(Y_i = 1)$ satisfies

$$\mathrm{logit}(\pi_i) = \boldsymbol{\beta}^T \mathbf{x}_i.$$

Discuss briefly the plausibility of the assumption concerning the mis-classification probabilities. [Bross, 1954; Ekholm and Palmgren, 1982; Palmgren, 1987; Copas, 1988].

4.31 Consider the likely sampling scheme by which the data in Table 4.5 were obtained. For each of the n occupied sites let G_i be the event that site i is occupied by a *Grahami* lizard. Conversely, O_i denotes occupation by an *Opalinus* lizard. Suppose that

$$\mathrm{pr}(G_i \mid \mathbf{x}_i) = \exp(\alpha + \boldsymbol{\beta}^T \mathbf{x}_i)/\{1 + \exp(\alpha + \boldsymbol{\beta}^T \mathbf{x}_i)\}$$

where $\mathbf{x}$ denotes the factors H, D, S and T. Let Z_i, an indicator variable identifying the sites that were recorded, satisfy

$$\mathrm{pr}(Z_i = 1 \mid G_i, \mathbf{x}_i) = \pi(\mathbf{x}_i)\phi_g,$$
$$\mathrm{pr}(Z_i = 1 \mid O_i, \mathbf{x}_i) = \pi(\mathbf{x}_i)\phi_o,$$

where ϕ_g/ϕ_o is the sampling bias. Show that, for the sampled sites,

$$\mathrm{pr}(G_i \mid Z_i = 1, \mathbf{x}_i) = \exp(\alpha^* + \boldsymbol{\beta}^T \mathbf{x}_i)/\{1 + \exp(\alpha^* + \boldsymbol{\beta}^T \mathbf{x}_i)\}$$

and give the expression for α^*. Explain why the selection probability $\mathrm{pr}(Z = 1)$ almost certainly depends on $\mathbf{x}$.