7.2.4 *Profile likelihood*

In those instances where they exist, marginal and conditional likelihoods work well, often with little sacrifice of information. However, marginal and conditional likelihoods are available only in very special problems. The profile log likelihood, while less satisfactory from several points of view, does have the important virtue that it can be used in all circumstances.

Let $\hat{\boldsymbol{\lambda}}_\psi$ be the maximum-likelihood estimate of $\boldsymbol{\lambda}$ for fixed $\boldsymbol{\psi}$. This maximum is assumed here to be unique, as it is for most generalized linear models. The partially maximized log-likelihood function,

$$l^\dagger(\boldsymbol{\psi}; y) = l(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_\psi; y) = \sup_{\boldsymbol{\lambda}} l(\boldsymbol{\psi}, \boldsymbol{\lambda}; y)$$

is called the profile log likelihood for $\boldsymbol{\psi}$. Under certain conditions the profile log likelihood may be used just like any other log likelihood. In particular, the maximum of $l^\dagger(\boldsymbol{\psi}; y)$ coincides with the overall maximum-likelihood estimate. Further, approximate confidence sets for $\boldsymbol{\psi}$ may be obtained in the usual way, namely

$$\{\boldsymbol{\psi} : 2l^\dagger(\hat{\boldsymbol{\psi}}; y) - 2l^\dagger(\boldsymbol{\psi}; y) \leq \chi^2_{p,1-\alpha}\}$$

where $p = \dim(\boldsymbol{\psi})$. Alternatively, though usually less accurately, intervals may be based on $\hat{\boldsymbol{\psi}}$ together with the second derivatives of $l^\dagger(\boldsymbol{\psi}; y)$ at the maximum. Such confidence intervals are often satisfactory if $\dim(\boldsymbol{\lambda})$ is small in relation to the total Fisher information, but are liable to be misleading otherwise.

Unfortunately $l^\dagger(\boldsymbol{\psi}; y)$ is not a log likelihood function in the usual sense. Most obviously, its derivative does not have zero mean, a property that is essential for estimating equations. In fact the derivative of $l^\dagger$ may be written in terms of the partial derivatives of l as follows:

$$\begin{aligned}\frac{\partial l^\dagger}{\partial \boldsymbol{\psi}} &= \frac{\partial}{\partial \boldsymbol{\psi}} l(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_\psi; y) \\ &= \frac{\partial l}{\partial \boldsymbol{\psi}} + \frac{\partial^2 l}{\partial \boldsymbol{\psi} \partial \boldsymbol{\lambda}}(\hat{\boldsymbol{\lambda}}_\psi - \boldsymbol{\lambda}) + \tfrac{1}{2}\frac{\partial^3 l}{\partial \boldsymbol{\psi} \partial \boldsymbol{\lambda}^2}(\hat{\boldsymbol{\lambda}}_\psi - \boldsymbol{\lambda})^2 + \ldots \\ &\quad + \left\{\frac{\partial l}{\partial \boldsymbol{\lambda}} + \frac{\partial^2 l}{\partial \boldsymbol{\lambda}^2}(\hat{\boldsymbol{\lambda}}_\psi - \boldsymbol{\lambda}) + \tfrac{1}{2}\frac{\partial^3 l}{\partial \boldsymbol{\lambda}^3}(\hat{\boldsymbol{\lambda}}_\psi - \boldsymbol{\lambda})^2 + \ldots\right\}\frac{\partial \hat{\boldsymbol{\lambda}}_\psi}{\partial \boldsymbol{\psi}}\end{aligned}$$

The expression in parentheses is just $\partial l(\boldsymbol{\psi}, \boldsymbol{\lambda})/\partial \boldsymbol{\lambda}$ evaluated at $\hat{\boldsymbol{\lambda}}_\psi$, and hence is identically zero. Under the usual regularity conditions

for large n, the remaining three terms are $O_p(n^{1/2})$, $O_p(n^{1/2})$ and $O_p(1)$ respectively. The first term has zero mean but the remaining two have mean $O(1)$ if $\hat{\boldsymbol{\lambda}}_\psi$ is a consistent estimate of $\boldsymbol{\lambda}$. Their expectations may be inflated if $\hat{\boldsymbol{\lambda}}_\psi$ is not consistent.

A simple expression for the approximate mean of $\partial l^\dagger/\partial \boldsymbol{\psi}$ in terms of cumulants of the derivatives of l is given by McCullagh and Tibshirani (1988).

In general, if the dimension of $\boldsymbol{\lambda}$ is a substantial fraction of n, the mean of $\partial l^\dagger/\partial \boldsymbol{\psi}$ is not negligible and the profile log likelihood can be misleading if interpreted as an ordinary log likelihood.

It is interesting to compare the profile log likelihood with the marginal log likelihood in a model for which both can be calculated explicitly. The covariance-estimation model, considered briefly at the end of section 7.2.1, is such an example. The profile log likelihood for the covariance parameters $\boldsymbol{\theta}$ in that problem is

$$l^\dagger(\boldsymbol{\theta}; y) = -\tfrac{1}{2}\log\det\boldsymbol{\Sigma} - \tfrac{1}{2}Q_2(\mathbf{R}),$$

which differs from the marginal log likelihood given at the end of section 7.2.1 by the term $\frac{1}{2}\log\det(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})$. Both the marginal and profile log likelihoods depend on the data only through the contrasts or residuals, $\mathbf{R}$. The marginal log likelihood is clearly preferable to $l^\dagger$ in this example, because $l^\dagger$ is not a log likelihood. The derivatives of $l^\dagger$, unlike those of the marginal log likelihood, do not have zero mean.

The use of profile likelihoods for the estimation of covariance functions has been studied by Mardia and Marshall (1984).

7.3 Hypergeometric distributions

7.3.1 *Central hypergeometric distribution*

Suppose that a simple random sample of size m_1 is taken from a population of size $m_\cdot$. The population is known to comprise s_1 individuals who have attribute A and $s_2 = m_\cdot - s_1$ who do not. In the sample, Y individuals have attribute A and the remainder, $m_1 - Y$, do not. The following table gives the numbers of sampled and non-sampled subjects who possess the attribute in question.

	Attribute A	$\overline{A}$	*Total*
sampled	$Y \equiv Y_{11}$	$m_1 - Y \equiv Y_{12}$	m_1
non-sampled	$s_1 - Y \equiv Y_{21}$	$m_2 - s_1 + Y \equiv Y_{22}$	m_2
Total	s_1	s_2	$m_\cdot \equiv s_\cdot$

Under the simple random sampling model, the distribution of Y conditionally on the marginal totals $\mathbf{m}, \mathbf{s}$ is

$$\text{pr}(Y = y \mid \mathbf{m}, \mathbf{s}) = \frac{\binom{m_1}{y}\binom{m_2}{s_1 - y}}{\binom{m_\cdot}{s_1}} = \frac{\binom{s_1}{y}\binom{s_2}{m_1 - y}}{\binom{s_\cdot}{m_1}} \tag{7.6}$$

The range of possible values for y is the set of integers satisfying

$$a = \max(0, s_1 - m_2) \le y \le \min(m_1, s_1) = b. \tag{7.7}$$

There are $\min(m_1, m_2, s_1, s_2) + 1$ points in the sample space. If $a = b$, the conditional distribution puts all its mass at the single point a. Degeneracy occurs only if one of the four marginal totals is zero.

The central hypergeometric distribution (7.6) is denoted by $Y \sim H(\mathbf{m}, \mathbf{s})$ or by $Y \sim H(\mathbf{s}, \mathbf{m})$.

An alternative derivation of the hypergeometric distribution is as follows. Suppose that $Y_1 \sim B(m_1, \pi)$ and $Y_2 \sim B(m_2, \pi)$ are independent binomial random variables. Then the conditional distribution of $Y \equiv Y_1$ conditionally on $Y_1 + Y_2 = s_1$ is given by (7.6).

The descending factorial moments of Y are easily obtained from (7.6) as follows:

$$\mu_{[r]} = E\{Y^{(r)}\} = m_1^{(r)} s_1^{(r)} / m_\cdot^{(r)},$$

where $Y^{(r)} = Y(Y-1)\ldots(Y-r+1)$, provided that $r \le \min(m_1, s_1)$. From these factorial moments we may compute the cumulants of Y as follows. First, define the following functions of the marginal frequencies in terms of the sampling fraction $\tau = m_1/m_\cdot$.

$$K_1 = s_1/m_\cdot, \qquad \lambda_1 = m_\cdot \tau_1 = m_1,$$
$$K_2 = s_1 s_s / m_\cdot^{(2)}, \qquad \lambda_2 = m_\cdot \tau_1(1-\tau_1) = m_1 m_2 / m_\cdot,$$

$$K_3 = s_1 s_2 (s_2 - s_1)/m_\cdot^{(3)}, \qquad \lambda_3 = m_\cdot \tau_1(1-\tau_1)(1-2\tau_1) = m_1 m_2 (m_2 - m_1)/m_\cdot^2,$$
$$K_4 = s_1 s_2 \{m_\cdot(m_\cdot + 1) - 6 s_1 s_2\}/m_\cdot^{(4)},$$
$$K_{22} = s_1^{(2)} s_2^{(2)} / m_\cdot^{(4)}, \qquad \lambda_4 = m_\cdot \tau_1(1-\tau_1)\big(1 - 6\tau_1(1-\tau_1)\big).$$

The first four cumulants of Y are

$$\begin{aligned} E(Y) &= K_1\lambda_1, \quad \text{var}(Y) = K_2\lambda_2, \\ \kappa_3(Y) &= K_3\lambda_3, \quad \kappa_4(Y) = K_4\lambda_4 - 6K_{22}\lambda_2^2/(m_\cdot - 1). \end{aligned} \tag{7.8}$$

Note that λ_r is the rth cumulant of the $B(m_\cdot, \tau_1)$ distribution associated with the sampling fraction, whereas $K_1, \ldots, K_4$, K_{22} are the population k-statistics and polykay up to order four. Details of these symmetric functions are given in McCullagh (1987), Chapter 4, especially section 4.6. For large $m_\cdot$ and for fixed sampling fraction, the λs are $O(m_\cdot)$, whereas the Ks are $O(1)$ for fixed attribute ratio, s_1/s_2.

Note that the third cumulant of Y is zero if either $K_3 = 0$ or $\lambda_3 = 0$. In fact all odd-order cumulants are zero under these conditions and the distribution of Y is symmetric.

7.3.2 *Non-central hypergeometric distribution*

The non-central hypergeometric distribution with odds ratio ψ is an exponentially weighted version of the central hypergeometric distribution (7.6). Thus

$$\text{pr}(Y = y; \psi) = \frac{\binom{m_1}{y}\binom{m_2}{s_1 - y}\psi^y}{P_0(\psi)} \tag{7.9}$$

where $P_0(\psi)$ is the polynomial in ψ,

$$P_0(\psi) = \sum_{j=a}^{b} \binom{m_1}{j}\binom{m_2}{s_1 - j}\psi^j.$$

The range of summation is given by (7.7). This distribution arises in the exponentially weighted sampling scheme in which each of the $\binom{m_\cdot}{m_1}$ possible samples is weighted proportionally to ψ^y, where

y is a particular function of the sample. Here y is the number of individuals in the sample who possess attribute A, but in principle any function of the sample could be chosen.

Alternatively, the non-central hypergeometric distribution may be derived as follows. Suppose that $Y_1 \sim B(m_1, \pi_1)$, $Y_2 \sim B(m_2, \pi_2)$ are independent binomial random variables and that $\psi = \pi_1(1-\pi_2)/\{\pi_2(1-\pi_1)\}$ is the odds ratio. Then the conditional distribution of Y_1 given that $Y_\cdot = s_1$ is non-central hypergeometric with parameter ψ. For conciseness, we write $Y \sim H(\mathbf{m}, \mathbf{s}; \psi)$ to denote the conditional distribution (7.9). Note that $P_0(1) = \binom{m_\cdot}{s_1}$, so that $H(\mathbf{m}, \mathbf{s}; 1)$ is identical to $H(\mathbf{m}, \mathbf{s})$.

An 'observation' from the distribution (7.9) is often presented as a 2×2 table in which the marginal totals are $\mathbf{m}$ and $\mathbf{s}$. The contribution of such an observation to the conditional log likelihood is

$$y \log \psi - \log P_0(\psi),$$

where the dependence on $\mathbf{m}$ and $\mathbf{s}$ has been suppressed in the notation for the polynomial $P_0(\psi)$. This log likelihood has the standard exponential-family form with canonical parameter $\theta = \log \psi$ and cumulant function

$$K(\theta) = \log P_0(e^\theta).$$

The mean and variance of Y are therefore

$$\kappa_1(\theta) = E(Y;\theta) = K'(\theta) = P_1(\psi)/P_0(\psi)$$
$$\kappa_2(\theta) = \mathrm{var}(Y;\theta) = K''(\theta) = P_2(\psi)/P_0(\psi) - \{P_1(\psi)/P_0(\psi)\}^2,$$

where $P_r(\psi)$ is the polynomial

$$P_r(\psi) = \sum_{j=a}^{b} j^r \psi^j \binom{m_1}{j}\binom{m_2}{s_1 - j}. \qquad (7.10)$$

More generally, the moments about the origin are expressible as rational functions in ψ, namely

$$\mu_r(\psi) = P_r(\psi)/P_0(\psi).$$

Unfortunately the functions $\kappa_1(\theta)$ and $\kappa_2(\theta)$ are awkward to compute particularly if the range of summation in (7.10) is extensive. The following approximations are often useful. First, it is easily shown that, conditionally on the marginal totals,

$$E(Y_{11}Y_{22}) = \psi E(Y_{12}Y_{21})$$

and, more generally, that

$$E(Y_{11}^{(r)}Y_{22}^{(r)}) = \psi^r E(Y_{12}^{(r)}Y_{21}^{(r)}).$$

Hence, since $E(Y_{11}Y_{22}) = \mu_{11}\mu_{22} + \kappa_2$, we have

$$\psi = \frac{\mu_{11}\mu_{22} + \kappa_2}{\mu_{12}\mu_{21} + \kappa_2},$$

where $\mu_{11} = E(Y_{11};\theta), \ldots$ are the conditional means for the four cells, and κ_2 is the conditional variance of each cell. Consequently we have the following exact relationship between $\kappa_1 \equiv \mu_{11}$ and κ_2:

$$\kappa_1(m_2 - s_1 + \kappa_1) + \kappa_2 = \psi\{(s_1 - \kappa_1)(m_1 - \kappa_1) + \kappa_2\}. \qquad (7.11)$$

In addition, the following approximate relationship may be derived from asymptotic considerations of the type discussed in section 6.5.6:

$$\kappa_2 \simeq \frac{m_\cdot}{m_\cdot - 1}\left(\frac{1}{\mu_{11}} + \frac{1}{\mu_{12}} + \frac{1}{\mu_{21}} + \frac{1}{\mu_{22}}\right)^{-1}. \qquad (7.12)$$

In addition to being asymptotically correct for large $b - a$, this expression is exact for $m_\cdot = 2$, the smallest non-degenerate value, and also for $\psi = 1$, whatever the marginal configuration.

The simultaneous solution to (7.11) and (7.12) gives a very accurate approximation to the conditional mean and variance provided that either $|\theta| < 2$ or the marginal totals are large: see Breslow and Cologne (1986). An equally accurate but slightly more complicated approximation is given by Barndorff-Nielsen and Cox (1979).

however, gives a simple exact relationship between the conditional mean vector $\boldsymbol{\mu}_1$ of Y and the conditional covariance matrix $\boldsymbol{\Sigma}$.

$$\frac{E(Y_{1j}Y_{2k})}{E(Y_{2j}Y_{1k})} = \psi_j = \frac{\mu_{1j}\mu_{2k} - \sigma_{jk}}{\mu_{2j}\mu_{1k} - \sigma_{jk}}. \tag{7.15}$$

Note that

$$\sigma_{jk} = \operatorname{cov}(Y_{1j}, Y_{1k}) = -\operatorname{cov}(Y_{1j}, Y_{2k})$$

is negative for $j < k$.

The covariance matrix $\boldsymbol{\Sigma}$ of $Y_{11}, \ldots, Y_{1k}$ may be approximated quite accurately as follows. Define the vector $\boldsymbol{\zeta}$ with components ζ_j given by

$$\frac{1}{\zeta_j} = \frac{1}{\mu_{1j}} + \frac{1}{\mu_{2j}}.$$

The approximate covariance matrix $\tilde{\boldsymbol{\Sigma}}$ is then given in terms of $\boldsymbol{\zeta}$ by

$$\tilde{\boldsymbol{\Sigma}} = \frac{m_.}{m_. - 1}\{\operatorname{diag}(\boldsymbol{\zeta}) - \boldsymbol{\zeta}\boldsymbol{\zeta}^T/\zeta_.\}. \tag{7.16}$$

This matrix has rank $k-1$. The simultaneous solution of equations (7.15) and (7.16) gives the approximate mean and covariance matrix of $\mathbf{Y}$ as a function of $\boldsymbol{\psi}$.

7.4 Some applications involving binary data

7.4.1 *Comparison of two binomial probabilities*

Suppose that a clinical trial is undertaken to compare the effect of a new drug or other therapy with the current standard drug or therapy. Ignoring side-effects and other complications, the response for each patient is assumed to be simply 'success' or 'failure'. In order to highlight the differences between the conditional log likelihood and the unconditional log likelihood, it is assumed that the observed data are as shown in Table 7.1. For a single stand-alone experiment, the numbers in this Table are unrealistically small, except perhaps as the information available at an early stage in the experiment when few patients have been recruited. In the context of a large-scale multi-centre clinical trial, however, Table 7.1 might represent the contribution of one of the smaller

Table 7.1 *Hypothetical responses in one segment of a clinical trial*

	Response		
	Success	*Failure*	*Total*
Treatment	$Y_1 = 2$	1	$m_1 = 3$
Control	$Y_2 = 1$	3	$m_2 = 4$
Total	$Y_. = 3$	4	$m_. = 7$

centres to the study. It is in the latter context that the methods described here have greatest impact.

We begin with the usual assumption that responses are independent and homogeneous within each of the two groups. Allowance can be made for the differential effect of covariates measured on individuals, but to introduce such effects at this stage would only complicate the argument. Strict adherence to protocol, together with randomization and concealment, are essential to ensure comparability, internal homogeneity and independence. With these assumptions, the numbers of successes in each treatment group may be regarded as independent binomial variables $Y_i \sim B(m_i, \pi_i)$, where

$$\begin{aligned} \operatorname{logit}\pi_1 &= \lambda + \Delta \\ \operatorname{logit}\pi_2 &= \lambda. \end{aligned} \tag{7.17}$$

For a single experiment or 2×2 table, (7.17) is simply a re-parameterization from the original probability scale to the more convenient logistic scale. Implicit in the re-parameterization, however, is the assumption that the logistic difference, Δ is a good and useful measure of the treatment effect. In particular, when it is required to pool information gathered at several participating sites or hospitals, it is often assumed that λ may vary from site to site but that Δ remains constant over all sites regardless of the success rate for the controls.

In order to set approximate confidence limits for Δ, there are two principal ways in which we may proceed. The simplest way is to fit the linear logistic model (7.17) using the methods described in Chapter 4. Approximate confidence limits may be based on $\hat{\Delta}$ and its large-sample standard error. For the present example this gives

$$\hat{\Delta} = \log\left(\frac{2 \times 3}{1 \times 1}\right) = 1.792, \qquad \text{s.e.}(\hat{\Delta}) \simeq 1.683.$$

Note that the large-sample variance of $\hat{\Delta}$ is

$$\operatorname{var}\hat{\Delta} = 1/2 + 1/1 + 1/1 + 1/3 = 17/6.$$

More accurate intervals are obtained by working with the profile deviance,

$$D(y;\Delta) = 2l(\hat{\Delta},\hat{\lambda}) - 2l(\Delta,\hat{\lambda}_\Delta)$$

where $\hat{\lambda}_\Delta$ is the maximum-likelihood estimate of λ for given Δ. This statistic is easy to compute using standard computer packages. For the data in Table 7.1, the profile deviance is plotted in Fig. 7.1. The nominal 90% large-sample confidence interval, determined graphically, is

$$\{\Delta : D(y;\Delta) - D(y;\hat{\Delta}) < 2.71\} = (-0.80, 4.95).$$

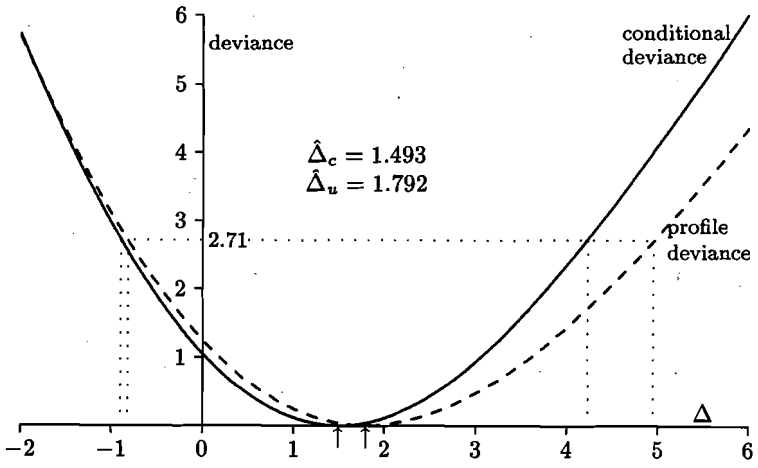

Fig. 7.1 *Graphical comparison of hypergeometric and binomial deviance functions for the data in Table 7.1. Nominal 90% intervals for the log odds ratio, Δ, are indicated.*

The alternative approach advocated here is to eliminate λ by using the conditional likelihood given $Y_{\cdot}$. The hypergeometric log likelihood is

$$l_c(\Delta) = y_1\Delta - \log P_0(e^\Delta),$$

where, for Table 7.1, $P_0(\psi)$ is equal to the cubic polynomial

$$P_0(\psi) = 4 + 18\psi + 12\psi^2 + \psi^3.$$

The hypergeometric likelihood has its maximum at a point $\hat{\Delta}_c$ different from the unconditional maximum $\hat{\Delta}$. In general $|\hat{\Delta}_c| \leq |\hat{\Delta}|$, with equality only at the origin. More precisely $\hat{\Delta}_c$ satisfies the standard exponential-family condition

$$y_1 = e^{\hat{\Delta}_c}P_0'(e^{\hat{\Delta}_c})/P_0(e^{\hat{\Delta}_c}) = E(Y_1 \,|\, Y_{\cdot};\hat{\Delta}_c).$$

In the example under discussion we find

$$\hat{\Delta}_c = 1.493, \qquad \text{s.e.}(\hat{\Delta}_c) \simeq 1.492,$$

where the standard error is computed in the usual way, namely

$$\operatorname{var}(\hat{\Delta}_c) \simeq 1/\operatorname{var}(Y;\hat{\Delta}_c) = 1/0.4495.$$

The conditional deviance function

$$2l_c(\hat{\Delta}_c) - 2l_c(\Delta)$$

is plotted as the solid line in Fig 7.1 and departs markedly from the profile deviance for large values of Δ.

7.4.2 *Combination of information from several 2×2 tables*

Suppose that data in the form of Table 7.1 are available from several sources, centres or strata, all cooperating in the same investigation. In the context of a multi-centre clinical trial, the strata are the medical centres participating in the trial. In some trials there may be many such centres, each contributing only a small proportion of the total patients enrolled. At each centre, one would expect that the pool of patients suitable for inclusion in the trial would differ in important respects that are difficult to measure. For instance, pollution levels, water hardness, rainfall, noise levels and other less tangible variables might have an effect on the response. In addition, nursing care and staff morale could have an appreciable effect on patients who are required to remain in hospital. Consequently, one

would expect the success rate for any medical treatment to vary appreciably from centre to centre.

Consequently, if we write

$$\pi_{1i} = \text{pr}(\text{success} \mid \text{treatment})$$
$$\pi_{2i} = \text{pr}(\text{success} \mid \text{control})$$

for the success probabilities at centre i, we may consider the linear logistic model

$$\begin{aligned} \text{logit}\, \pi_{1i} &= \lambda_i + \Delta \\ \text{logit}\, \pi_{2i} &= \lambda_i, \qquad i = 1, \ldots, n. \end{aligned} \tag{7.18}$$

The idea behind this parameterization is that $\Delta > 0$ implies that treatment is uniformly beneficial at all centres regardless of the control success rate: $\Delta < 0$ implies that the new treatment is uniformly poorer than the standard procedure. There is, of course, the possibility that Δ varies from centre to centre, even to the extent that $\Delta > 0$ for some centres and $\Delta < 0$ for others. Such interactions require careful investigation and detailed plausible explanation.

One obvious difficulty with the linear logistic model (7.18) is that it contains $n+1$ parameters to be estimated on the basis of $2n$ observed binomial proportions. In such circumstances, maximum likelihood need not be consistent or efficient for large n. However, following the general argument outlined in section 7.2.2, if we condition on the observed success totals, $Y_{\cdot i}$, at each of the centres, we have

$$Y_{1i} \mid Y_{\cdot i} \sim H(\mathbf{m}_i, y_{\cdot i}; \psi). \tag{7.19}$$

The hypergeometric log likelihood is thus the sum of n conditionally independent terms and depends on only one parameter, namely $\psi = e^{\Delta}$. Provided that the total conditional Fisher information is sufficiently large, standard large-sample likelihood theory applies to the conditional likelihood.

The conditional log likelihood for Δ is

$$l_c(\Delta) = \sum_i \{y_{1i}\Delta - \log P_0(e^{\Delta}; m_{1i}, m_{2i}, y_{\cdot i})\},$$

where additional arguments have been appended to the polynomial $P_0(\cdot)$ to emphasize its dependence on the marginal totals for stratum i.

The score statistic for no treatment effect is

$$U = \partial l_c/\partial\Delta\big|_{\Delta=0} = \sum_i \{Y_{1i} - E(Y_{1i})\} = \sum_i \{Y_{1i} - m_{1i}y_{\cdot i}/m_{\cdot i}\}.$$

The exact null variance of U is the sum of hypergeometric variances, namely

$$\text{var}(U) = \sum_i m_{1i}m_{2i}y_{\cdot i}(m_{\cdot i} - y_{\cdot i})/\{m_{\cdot i}^2(m_{\cdot i} - 1)\}.$$

The approximate one-sided significance level for the hypothesis of no treatment effect is $1 - \Phi(z^-)$, where

$$Z^- = (U - \tfrac{1}{2})/\sigma_U$$

is the continuity-corrected value. This test, first proposed by Mantel and Haenszel (1959), is known as the Mantel-Haenszel test. The Mantel-Haenszel estimator, which is different from the conditional likelihood estimator, is derived in Exercise 9.10.

7.4.3 *Example: Ille-et-Vilaine study of oesophageal cancer*

The data shown in Table 7.2 is a summary of the Ille-et-Vilaine retrospective study of the effect of alcohol consumption on the incidence of oesophageal cancer. A more complete list of the data, including information on tobacco consumption, is given in Appendix 1 of Breslow and Day (1980). In a retrospective study the numbers of *cases* (subjects with cancer) and the number of *controls* is to be regarded as fixed by the study design. The alcohol consumption rate (high/low) is the effective response. However, for the reasons given in section 4.4.3, the roles of these two variables can be reversed. We may, therefore, regard alcohol consumption rate as the explanatory covariate and outcome (cancer/no cancer) as the response even though such a view is not in accord with the sampling scheme. Since the analysis that follows is conditional on both sets of marginal totals, this role-reversal presents no conceptual difficulty.

It is common to find that the incidence of cancer increases with age. The cases in this study are older on average than the controls. If age were ignored in the analysis, the apparent effect of alcohol

Table 7.2 *Ille-et-Vilaine retrospective study of the relationship between alcohol consumption and the incidence of oesophageal cancer*

	Cancer		*No cancer*				
	Alcohol consumption					*Fitted values under model* (ii)	
	80+	80−	80+	80−			
Age	y_{11}	y_{12}	y_{21}	y_{22}	$\tilde{\psi}_c$	$\hat{\mu}_{11}$	*Residual*
25–34	1	0	9	106	∞	0.33	1.42
35–44	4	5	26	164	4.98	4.11	−0.07
45–54	25	21	29	138	5.61	24.49	0.18
55–64	42	34	27	139	6.30	40.09	0.59
65–74	19	36	18	88	2.56	23.74	−1.89
75+	5	8	0	31	∞	3.24	1.75
Total	96	104	109	666		96.01	$X^2 = 9.04$

consumption would be inflated. For that reason it is advisable to stratify the data by age. In other words, cases are matched with controls of a similar age. The treatment effect is therefore a comparison of cancer incidence rates between subjects of similar age.

Three models are considered.

1. a model in which the log odds-ratio is zero, meaning that alcohol consumption has no effect on the incidence of oesophageal cancer.
2. a model in which the log odds-ratio is constant, meaning that increased alcohol consumption increases the odds for oesophageal cancer by the factor e^{ψ} uniformly over all age groups.
3. a model in which the log odds-ratio increases or decreases linearly with increasing age.

Algebraically, these models may be written in the form

$$\begin{aligned} &\text{(i)} && \log \psi_i = 0, \\ &\text{(ii)} && \log \psi_i = \beta_0, \\ &\text{(iii)} && \log \psi_i = \beta_0 + \beta_1(i - 3.5), \end{aligned} \tag{7.20}$$

where $i = 1, \ldots, 6$ indexes the age strata. The residual deviances for these three models are 89.83, 10.73 and 10.29 on 6, 5 and 4 degrees of freedom respectively.

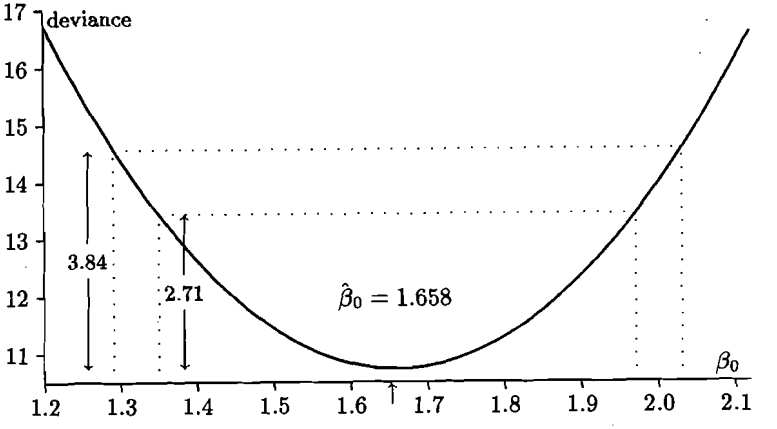

Fig. 7.2 *Hypergeometric deviance for model* (7.20). *Nominal* 90% *and* 95% *intervals for the log odds ratio,* β_0, *are indicated.*

The model formula for (i) is unusual in that it is entirely empty, excluding even the intercept.

The estimate of β_0 for the model of constant odds-ratio is 1.658 with standard error 0.189. Fitted values and residuals under this model are shown in the final two columns of Table 7.2. The residuals, calculated by the formula

$$(y_{11} - \hat{\mu}_{11})/\sqrt{V(\hat{\mu}_{11})},$$

exhibit no patterns that would suggest systematic deviation from constancy of the odds-ratio. The fact that we have chosen the (1,1) cell is immaterial because the residuals are equal in magnitude for the four cells of the response.

For the third model, the estimates are

$$\begin{aligned} \hat{\beta}_0 &= 1.7026 \qquad \text{s.e.}(\hat{\beta}_0) \simeq 0.2000 \\ \hat{\beta}_1 &= -0.1255 \qquad \text{s.e.}(\hat{\beta}_1) \simeq 0.1879 \end{aligned}$$

confirming that there is no evidence of a linear trend in the log odds-ratios.

Both Pearson's statistic and the residual deviance statistic are a little on the large side, though of borderline statistical significance when compared to the nominal χ^2_5 distribution. This inflation may be due to factors that have been ignored in the present analysis.

The unconditional analysis for these data, in which each row of Table 7.2 is treated as a pair of independent binomial variables, gives very similar, though not identical, answers in this example. The unconditional residual deviances for the three models (7.20) are 90.56, 11.04 and 10.61. The unconditional maximum-likelihood estimate of β_0 in the second model is 1.670 with asymptotic standard error 0.190. As usual, the unconditional estimate is larger in magnitude than the conditional estimate. The unconditional estimate is biased away from the origin, though in this example the bias is small because the counts are, for the most part, moderately large. There are similar slight differences between the unconditional and conditional estimates for the third model. None of these differences is of sufficient magnitude to affect the conclusions reached.

Thus it appears that the habitual tippler will find no comfort in these data. The odds for oesophageal cancer are higher by an estimated factor of $5.251 = \exp(1.6584)$ in the high alcohol-consumption group than in the low alcohol group. This odds factor applies to all age groups even though the incidence of cancer increases with age. Approximate 95% confidence limits for the odds-ratio are

$$\exp(1.658 \pm 1.96{\times}0.189) = \exp(1.288,\ 2.028) = (3.624,\ 7.602),$$

which is almost identical to the interval $(3.636, 7.622)$ obtained from the deviance plot in Fig. 7.2. Normal approximations tend to be more accurate when used on the $\log\hat{\psi}$-scale rather than the $\hat{\psi}$-scale.

7.5 Some applications involving polytomous data

7.5.1 *Matched pairs: nominal response*

Suppose that subjects in a study are matched in pairs and that a single polytomous response is observed for each subject. Following the usual procedure for matched pairs, we shall suppose that the logarithmic response probabilities for the control member of the ith pair are

$$\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{ik}),$$

which are free to vary in any haphazard or other way from pair to pair. We shall suppose in addition that the treatment effect as measured on the logarithmic scale is the same for all pairs. The logarithmic response probabilities for the treated member of the ith pair are therefore

$$\boldsymbol{\lambda}_i + \boldsymbol{\Delta} = (\lambda_{i1} + \Delta_1, \ldots, \lambda_{ik} + \Delta_k).$$

The probability of observing response category j for control subject i is

$$\exp(\lambda_{ij}) \Big/ \sum_r \exp(\lambda_{ir}),$$

while the probabilities for the treated subject are

$$\exp(\lambda_{ij} + \Delta_j) \Big/ \sum_r \exp(\lambda_{ir} + \Delta_j).$$

Each response can be represented either as an integer R in the range $(1, k)$, or as an indicator vector Z having k components. The components of Z are

$$Z_j = \begin{cases} 1 & \text{if } R = j \\ 0 & \text{otherwise.} \end{cases}$$

Consider now a given pair having logarithmic response probabilities $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda} + \boldsymbol{\Delta}$, for which the observed categories are r_1 and r_2 respectively. For any given value of $\boldsymbol{\Delta}$, the sufficient statistic for $\boldsymbol{\lambda}$ is the vector sum, $Z_\cdot = Z_1 + Z_2$, of the observed responses. If $Z_\cdot = (0, \ldots, 2, \ldots, 0)$, both R_1 and R_2 are determined by $Z_\cdot$ and the conditional distribution given $Z_\cdot$ is degenerate. However, if

$$Z_\cdot = (0, \ldots, 1, \ldots, 1, \ldots, 0),$$

with non-zero values in positions i and j, we must have

$$(R_1, R_2) = (i, j) \quad \text{or} \quad (j, i).$$

For $i \neq j$, the required conditional distribution is

$$\begin{aligned} \operatorname{pr}(R_1 = i \mid Z_\cdot) &= \frac{e^{\lambda_i} e^{\lambda_j + \Delta_j}}{e^{\lambda_i} e^{\lambda_j + \Delta_j} + e^{\lambda_j} e^{\lambda_i + \Delta_i}} \\ &= e^{\Delta_j} \Big/ (e^{\Delta_i} + e^{\Delta_j}), \end{aligned} \tag{7.21}$$