CHAPTER 7

Conditional likelihoods*

7.1 Introduction

In many applications the likelihood function involves several parameters, only a few of which are of interest to the investigator. The remaining parameters, often pejoratively called incidental or nuisance parameters, are necessary in order that the model make sense physically, but their values are largely irrelevant to the experiment and to the conclusions that are to be drawn. To take a simple example, consider a comparative experiment with two treatment groups in which the response is ordinal with k categories. The proportional-odds model,

$$\log\{\gamma_{ij}/(1-\gamma_{ij})\} = \theta_j - \beta x_i, \qquad j = 1,\ldots,k-1;\ i = 1,2,$$

in which x_i is an indicator variable for treatment group, involves one parameter of interest, β, and $k-1$ nuisance parameters, $\theta_1,\ldots,\theta_{k-1}$. There are applications in which both components $\boldsymbol{\theta}$ and β are of equal interest to the investigator, but usually, in comparative experiments, the focus is on change or rate of change of the response as the stimulus is increased. Thus, when we use the terms 'nuisance parameter' or 'incidental parameter' it is with certain common applications in mind. The status of a parameter depends on the context.

Two difficulties arise in dealing with likelihood functions that depend on a large number of incidental parameters in addition to the effects of interest. First, from a purely mathematical point of view, there is no guarantee of consistency or optimality in the limit as the number of parameters increases in proportion to the data

*This chapter contains more mathematical material and may be omitted on first reading.

accumulated. Whether this large-sample mathematical difficulty has any relevance in the finite samples actually observed is another matter, and will not be discussed here. The second difficulty is the purely numerical one of maximizing a function of many variables and of obtaining the inverse matrix of second derivatives, but this is a subsidiary consideration in view of the first difficulty. For these reasons, we seek a modified likelihood function that depends on as few of the incidental parameters as possible while, at the same time, sacrificing as little information as possible. Inferences are then based on this modified likelihood function, particularly on its shape in the vicinity of its maximum.

7.2 Marginal and conditional likelihoods

7.2.1 *Marginal likelihood*

One way of eliminating unwanted nuisance parameters is to work with the marginal likelihood for a suitably chosen subset of the complete data vector. This method does not always work satisfactorily, but when it does, it is clearly desirable to choose as large a subset of the original data as possible so that the information loss is minimized.

In the context of the bivariate logistic model (6.25), if $\boldsymbol{\beta}^{(a)}$ is the parameter of interest and $(\boldsymbol{\beta}^{(b)}, \boldsymbol{\beta}^{(ab)})$ are nuisance parameters, we may eliminate the nuisance parameters by working with the log likelihood for the marginal variable A alone, *i.e.* with the marginal totals $(Y_{1\cdot}, Y_{2\cdot})$. Evidently, from the analysis in section 6.6.1, some small loss of information is thereby incurred, but this loss might well be judged acceptable in view of the simplicity achieved. In this example, the loss of efficiency is balanced by a gain in robustness, because the marginal likelihood estimates are consistent whether or not the assumed models for η_a and η_{ab} in (6.19) are correct, whereas the estimates derived from the full likelihood are not protected against such mis-specifications.

To take an unrelated example, suppose $Y_1, \ldots, Y_n$ are observations taken at spatial locations $s_1, \ldots, s_n$, and that the n-vector $\mathbf{Y}$ has the multivariate Normal distribution with cumulants

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}; \qquad \mathrm{cov}(\mathbf{Y}) = \boldsymbol{\Sigma}(\boldsymbol{\theta}),$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is a known covariance function parameterized by $\boldsymbol{\theta}$. For example, we might have

$$\sigma_{ij}(\boldsymbol{\theta}) = \theta_1^2 \exp\{-|s_i - s_j|/\theta_2\},$$

where θ_2 has the physical dimension of length and θ_1 has the same physical dimension as y. If $\boldsymbol{\theta}$ is the parameter of interest and $\boldsymbol{\beta}$ is regarded as a nuisance parameter, we may eliminate $\boldsymbol{\beta}$ from the likelihood by working with the set of contrasts,

$$\mathbf{R} = (\mathbf{I} - \mathbf{P}_X)\mathbf{Y} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Y},$$

whose mean is zero and whose distribution does not depend on $\boldsymbol{\beta}$. Any complete set of $n-p$ linearly independent contrasts with zero mean is a linear transformation of $\mathbf{R}$, so that the choice of projection matrix, $\mathbf{P}_X$, is immaterial. In other words, we could replace $\mathbf{P}_X$ by $\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}$, for any positive definite matrix $\mathbf{W}$, without affecting the likelihood. See Exercises 7.10–7.13. In this example there appears to be no loss of information on $\boldsymbol{\theta}$ by using $\mathbf{R}$ in place of $\mathbf{Y}$, though it is difficult to give a totally satisfactory justification of this claim.

Although $\mathbf{R}$ has a rank-deficient covariance matrix whose determinant is zero, it is still possible to write down explicitly the log likelihood for $\boldsymbol{\theta}$ based on $\mathbf{R}$. The usual method (Kitanidis, 1987), which is to choose an arbitrary full-rank sub-vector, introduces unnecessary and undesirable asymmetry into the formulae. Assuming that $\boldsymbol{\Sigma}$ has rank n and that $\mathbf{X}$ has rank p, the marginal log likelihood for $\boldsymbol{\theta}$ based on $\mathbf{R}$ is

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

where

$$Q_2(\mathbf{R}) = \mathbf{R}^T(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1})\mathbf{R},$$

the weighted residual sum of squares for $\mathbf{R}$, is unaffected by the choice of projection matrix. An equivalent expression in terms of the eigenvalues of the matrix $(\mathbf{I} - \mathbf{P}_X)\boldsymbol{\Sigma}(\mathbf{I} - \mathbf{P}_X)$ is given by Patterson and Thompson (1971). Yet another equivalent expression is given by Harville (1974, 1977).

Note that in the important special case where $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, this marginal log likelihood becomes

$$-\tfrac{1}{2}(n-p)\log\sigma^2 - \tfrac{1}{2}(\text{RSS})/\sigma^2$$

where RSS is the residual sum of squares. This is just the marginal log likelihood derived from RSS $\sim \sigma^2\chi^2_{n-p}$.

For a derivation of the marginal log likelihood, also sometimes called the *restricted log likelihood* (Corbeil and Searle, 1976), see Exercises 7.8–7.13.

7.2.2 *Conditional likelihood*

Suppose that the parameter vector $\boldsymbol{\theta}$ can be partitioned into components $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$, in which $\boldsymbol{\psi}$ is the parameter vector of interest and $\boldsymbol{\lambda}$ is a vector of nuisance parameters. Suppose in addition that for each fixed value $\boldsymbol{\psi}_0$ of $\boldsymbol{\psi}$, the statistic $S_\lambda(\boldsymbol{\psi}_0)$ is sufficient for $\boldsymbol{\lambda}$ and complete. For definitions of completeness and sufficiency, see Cox and Hinkley (1974), section 2.2 or Lehmann (1986), sections 1.9 and 4.3. It is essential to distinguish two cases, (i) where $S_\lambda(\boldsymbol{\psi}_0)$ depends on $\boldsymbol{\psi}_0$, and (ii) where $S_\lambda(\boldsymbol{\psi}_0)$ is independent of $\boldsymbol{\psi}_0$ so that the sufficient statistic for $\boldsymbol{\lambda}$ is the same for all $\boldsymbol{\psi}_0$. In (i) the conditional distribution of $\mathbf{Y}$ given $S_\lambda(\boldsymbol{\psi}_0)$ is independent of $\boldsymbol{\lambda}$ only under $\boldsymbol{\psi} = \boldsymbol{\psi}_0$. Thus we write

$$f_{Y|S_\lambda(\psi_0)}(y \mid S_\lambda(\boldsymbol{\psi}_0); \boldsymbol{\psi}, \boldsymbol{\lambda})$$

for the conditional distribution.

In (ii), where $S_\lambda(\boldsymbol{\psi}_0) \equiv S_\lambda$, we write the conditional density in the form

$$f_{Y|S_\lambda}(y \mid S_\lambda; \boldsymbol{\psi})$$

emphasizing that the conditional distribution is independent of $\boldsymbol{\lambda}$. Here there is no conceptual difficulty in using

$$l_c(\boldsymbol{\psi}) = \log f_{Y|S_\lambda}(y \mid S_\lambda; \boldsymbol{\psi}) = \log f_Y(y; \boldsymbol{\psi}, \boldsymbol{\lambda}) - \log f_{S_\lambda}(s_\lambda; \boldsymbol{\psi}, \boldsymbol{\lambda}) \tag{7.1}$$

as the conditional log likelihood for $\boldsymbol{\psi}$. The maximizing value $\hat{\boldsymbol{\psi}}_c$, and the Fisher information $\mathbf{i}_{\psi|S_\lambda}$ based on the conditional log likelihood (7.1) are, in general, different from those derived from the full likelihood. For a simple but important example of this conditioning argument, see sections 7.4.1 and 7.4.2.

When $S_\lambda(\psi_0)$ depends on ψ_0, however, it is most important in writing the conditional density to distinguish between the two values of ψ. Thus

$$l_c(\psi, \lambda; \psi_0) = \log f_{Y|S_\lambda(\psi_0)}(y \mid S_\lambda(\psi_0); \psi, \lambda),$$

considered as a function of ψ and λ for fixed ψ_0, is a log-likelihood function. For instance, under the usual regularity conditions, the conditional score statistic

$$\frac{\partial l_c(\psi, \lambda; \psi_0)}{\partial \psi}$$

has zero expectation given $S_\lambda(\psi_0)$, and conditional covariance matrix equal to

$$-E\left(\frac{\partial^2 l_c(\psi, \lambda; \psi_0)}{\partial \psi^2}\right).$$

It follows that the score statistic is uncorrelated with $S_\lambda(\psi_0)$ for all parameter values.

By contrast, the reduced function

$$l^*(\psi) = l_c(\psi, \lambda; \psi)$$

is not the logarithm of a density with respect to any fixed measure, and hence does not ordinarily have the properties of a log-likelihood function. The reason for this is that the transformation from the original variables $\mathbf{Y}$ to the pair $(S_\lambda(\psi_0), \tilde{S})$, where $\tilde{S}$ is a complementary statistic, involves a Jacobian that depends on ψ_0. So long as ψ_0 is regarded as fixed, the Jacobian has no effect on the log likelihood and can be ignored. But if the Jacobian depends on ψ an extra term must be included in the log likelihood. Ordinarily, therefore, when computing likelihood functions it does not make sense to condition on parameter-dependent statistics. See Exercises 7.1–7.6 for several examples in which these differences are important.

Ideally we would like to choose a value ψ_0 for the conditioning statistic that is as near as possible to the conditional maximum-likelihood estimate, $\hat{\psi}_c$. For purposes of estimation, this effect is achieved by solving the conditional likelihood equation $U_\psi = \mathbf{0}$, where

$$U_\psi = \frac{\partial l_c(\boldsymbol{\psi}, \boldsymbol{\lambda}; \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}\bigg|_{\psi_0=\psi} \qquad (7.2)$$

evaluated at $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi})$. This is not the same as finding the roots of $\partial l^*(\boldsymbol{\psi})/\partial\boldsymbol{\psi} = 0$ because the log likelihood is differentiated only with respect to the first argument. Another way of deriving (7.2) is to bias-correct the unconditional log-likelihood derivatives as follows:

$$\frac{\partial l(\boldsymbol{\psi}, \boldsymbol{\lambda})}{\partial \boldsymbol{\psi}} - E\left(\frac{\partial l(\boldsymbol{\psi}, \boldsymbol{\lambda})}{\partial \boldsymbol{\psi}} \mid S_\lambda(\psi)\right).$$

This bias-corrected derivative is identical to (7.2) with $\boldsymbol{\lambda}$ replaced by $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi})$. This interpretation via estimating functions has been emphasized by Godambe (1976) and by Lindsay (1982).

The asymptotic variance of $\boldsymbol{\psi}_c$ is the inverse of

$$-E\left(\frac{\partial^2 l_c(\boldsymbol{\psi}, \boldsymbol{\lambda}; \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}^2}\right)\bigg|_{\psi_o=\psi}$$

evaluated at $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi})$.

The conditional score function as defined above is unaffected by the parameterization chosen for the nuisance parameters because

$$\frac{\partial l_c(\boldsymbol{\psi}, \boldsymbol{\lambda}; \boldsymbol{\psi})}{\partial \boldsymbol{\lambda}} \equiv \mathbf{0}$$

for all parameter values. Consequently there is no ambiguity regarding the meaning of (7.2).

This line of argument produces a usable score statistic having zero mean at the true parameter point, a 'conditional likelihood' estimator and an approximate standard error, but it does not produce a likelihood function for $\boldsymbol{\psi}$ directly. For this purpose the modified profile likelihood of Barndorff-Nielsen (1985, 1986) may be used. For a related derivation via orthogonal parameters, see Cox and Reid (1987).

The following example illustrates several aspects of the conditioning argument. Suppose that

$$Y_1 \sim N(\mu_1, 1) \quad \text{and} \quad Y_2 \sim N(\mu_2, 1)$$

are independent and that the ratio $\psi = \mu_2/\mu_1$ is the parameter of interest. To complete the parameterization, we may take $\lambda_1 = \mu_1$, $\lambda_2 = \mu_2$ or any other suitable complementary parameter such as $\lambda_3 = \mu_1 + \mu_2$ or the orthogonal parameter $(\mu_1^2 + \mu_2^2)^{1/2}$. A sufficient statistic for the nuisance parameter given $\psi = \psi_0$ is

$$S_\lambda(\psi_0) = Y_1 + \psi_0 Y_2.$$

Other equivalent forms of the sufficient statistic are

$$\begin{aligned}\hat{\mu}_1(\psi_0) &= \hat{\lambda}_1 = (Y_1 + \psi_0 Y_2)/(1+\psi_0^2),\\ \hat{\mu}_2(\psi_0) &= \hat{\lambda}_2 = \psi_0(Y_1 + \psi_0 Y_2)/(1+\psi_0^2),\\ \text{and}\quad \hat{\lambda}_3 &= (1+\psi_0)(Y_1 + \psi_0 Y_2)/(1+\psi_0^2).\end{aligned}$$

The conditional log likelihood given $S_\lambda(\psi_0)$ is

$$l_c(\psi, \mu_1; \psi_0) = -\tfrac{1}{2}\frac{(y_2 - \psi_0 y_1 - (\psi - \psi_0)\mu_1)^2}{1+\psi_0^2} - \tfrac{1}{2}\log(1+\psi_0^2).$$

Differentiation with respect to ψ followed by setting $\psi_0 = \psi$ gives

$$U_\psi = \left.\frac{\partial l_c(\psi, \mu_1; \psi_0)}{\partial \psi}\right|_{\psi_0=\psi} = \frac{\mu_1(y_2 - \psi y_1)}{1+\psi^2}. \qquad (7.3)$$

Note that the Jacobian term vanishes on differentiation. Further,

$$\begin{aligned}E(U_\psi) &= 0,\\ \mathrm{var}(U_\psi) &= \mu_1^2/(1+\psi^2),\\ -E\left(\frac{\partial^2 l_c(\psi, \mu_1; \psi_0)}{\partial \psi^2}\right)\Big|_{\psi_0=\psi} &= \frac{\mu_1^2}{1+\psi^2},\end{aligned}$$

so that the usual likelihood properties are satisfied. Also $\hat{\psi}_c = y_2/y_1$ with 'asymptotic' variance $(1+\psi^2)/\mu_1^2$, the usual approximation for the variance of a ratio estimator. In all of these expressions μ_1 is to be replaced by $\hat{\mu}_1(\psi)$.

Normal approximation for the distribution of the ratio is unsatisfactory unless μ_1 is large compared to the standard deviation of Y_1. Fieller confidence intervals generated via the score statistic U_ψ by

$$\{\psi : U_\psi^2/\mathrm{var}(U_\psi) \le k_{\alpha/2}^2\}$$

are exact and are preferred over Normal approximations.

By contrast, differentiation of the reduced function $l^*(\cdot)$ gives

$$\frac{\partial l_c(\psi,\mu_1;\psi)}{\partial\psi} = \frac{y_1(y_2-\psi y_1)}{1+\psi^2} + \frac{\psi(y_2-\psi y_1)^2}{(1+\psi^2)^2} - \frac{\psi}{1+\psi^2}.$$

The latter derivative has mean $-\psi/(1+\psi^2)$. For further discussion of this point see Exercise 7.19.

7.2.3 *Exponential-family models*

Suppose that the log likelihood for $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$ can be written in the exponential-family form

$$l(\boldsymbol{\theta}; y) = \boldsymbol{\theta}^T \mathbf{s} - b(\boldsymbol{\theta})$$

and admits a decomposition of the form

$$l(\boldsymbol{\theta}; y) = \boldsymbol{\psi}^T \mathbf{s}_1 + \boldsymbol{\lambda}^T \mathbf{s}_2 - b(\boldsymbol{\psi}, \boldsymbol{\lambda}), \qquad (7.4)$$

where $\mathbf{s} = (\mathbf{s}_1(y), \mathbf{s}_2(y))$ are functions of the data. Likelihood functions of this type occur most commonly when the observations are independent and the model considered is linear in the canonical parameter. Then the sufficient statistic is a linear function of the data, namely $\mathbf{s} = \mathbf{X}^T\mathbf{y}$. A decomposition of the form (7.4) can be achieved only if the parameter of interest, $\boldsymbol{\psi}$, is a linear function of $\boldsymbol{\theta}$. The choice of nuisance parameter, $\boldsymbol{\lambda}$, is to a large extent arbitrary and the inferences regarding $\boldsymbol{\psi}$ should be unaffected by the parameterization chosen for $\boldsymbol{\lambda}$.

It may be helpful at this stage to consider a simple example. Suppose that Y_1, Y_2 are independent Poisson random variables with means μ_1, μ_2 and that we are interested in the ratio $\psi' = \mu_1/\mu_2$. Here $\theta_i = \log\mu_i$ are the canonical parameters, and the parameter of interest $\psi = \log\psi' = \theta_1 - \theta_2$ is a linear contrast of the canonical parameters. For the nuisance parameter, we may choose any one of a variety of complementary parameters such as

$$\lambda_1' = \mu_1, \quad \lambda_2' = \mu_2, \quad \lambda_3' = \mu_1 + \mu_2 \quad \text{or} \quad \lambda_4' = \mu_1\mu_2.$$

The log likelihood expressed initially in terms of $\boldsymbol{\theta}$ is

$$\begin{aligned} l(\theta; \mathbf{y}) &= y_1\theta_1 + y_2\theta_2 - \exp(\theta_1) - \exp(\theta_2) \\ &= (y_1+y_2)\lambda_1 - y_2\psi - \exp(\lambda_1)(1+e^{-\psi}) \\ &= (y_1+y_2)\lambda_2 + y_1\psi - \exp(\lambda_2)(1+e^{\psi}) \\ &= \tfrac{1}{2}(y_1+y_2)\lambda_4 + \tfrac{1}{2}(y_1-y_2)\psi - 2\exp(\lambda_4/2)\cosh(\tfrac{1}{2}\psi), \end{aligned}$$

where $\psi = \log \psi'$ and $\lambda_j = \log \lambda_j'$.

It follows from (7.4) that for any given value of $\boldsymbol{\psi}$, $\mathbf{s}_2$ is sufficient for the nuisance parameter and that $\mathbf{s}_2$ is the same whatever parameterization is chosen for the nuisance parameter. In the above example, $\mathbf{s}_2 = Y_{\cdot}$ for each of the parameterizations considered. The conditional distribution of the data $\mathbf{Y}$ given $\mathbf{s}_2$ does not depend on $\boldsymbol{\lambda}$ and hence the conditional log likelihood

$$l(\boldsymbol{\psi} \,|\, \mathbf{s}_2) = \boldsymbol{\psi}^T \mathbf{s}_1 - b^*(\boldsymbol{\psi}; s_2) \tag{7.5}$$

may be used for inferences regarding $\boldsymbol{\psi}$. Given the sufficient statistic $\mathbf{s}_2$, the value of the nuisance parameter is irrelevant in subsequent calculations.

Note that the conditional log likelihood retains the exponential-family form in which $\mathbf{s}_1$ is conditionally sufficient for $\boldsymbol{\psi}$ given the value of $\mathbf{s}_2$. In the Poisson example, the required conditional distribution is that of the pair (Y_1, Y_2) given that $Y_{\cdot} = m$, and this is well known to yield the binomial distribution. It is immaterial here whether we work with the conditional distribution of $Y_1 \,|\, Y_{\cdot}$, $Y_2 \,|\, Y_{\cdot}$ or $Y_1 - Y_2 \,|\, Y_{\cdot}$ because these conditional distributions differ by a fixed linear transformation.

For a second example, suppose that $Y_1 \sim B(1, \pi_1)$ and $Y_2 \sim B(1, \pi_2)$ are independent and that the odds ratio

$$\psi' = \frac{\pi_1}{1-\pi_1} \bigg/ \frac{\pi_2}{1-\pi_2}$$

is the parameter of interest. The log-likelihood function is

$$\begin{aligned} &y_1 \log\left(\frac{\pi_1}{1-\pi_1}\right) + y_2 \log\left(\frac{\pi_2}{1-\pi_2}\right) + \log(1-\pi_1) + \log(1-\pi_2) \\ &= y_1\psi + (y_1 + y_2)\log\left(\frac{\pi_2}{1-\pi_2}\right) + \log(1-\pi_1) + \log(1-\pi_2) \end{aligned}$$

where $\psi = \log \psi'$. By the argument just given, we are led to consider the conditional distribution of (Y_1, Y_2) given that $Y_{\cdot} = y_{\cdot}$. If $y_{\cdot} = 0$ or $y_{\cdot} = 2$ the conditional distribution is degenerate. Otherwise if $y_{\cdot} = 1$, we have

$$\begin{aligned} \operatorname{pr}(Y_1 = 0 \,|\, Y_{\cdot} = 1) &= 1/(1+\psi') \\ \operatorname{pr}(Y_1 = 1 \,|\, Y_{\cdot} = 1) &= \psi'/(1+\psi') \end{aligned}$$

This is a particular instance of the hypergeometric distribution studied in section 7.3.

7.2.4 *Profile likelihood*

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

Let $\hat{\boldsymbol{\lambda}}_\psi$ 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

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

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:

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

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 $m_{\bullet}$. The population is known to comprise s_1 individuals who have attribute A and $s_2 = m_{\bullet} - s_1$ 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	$\bar{A}$	*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	$m_. \equiv s_.$

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

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

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

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

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:

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

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

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

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

The first four cumulants of Y are

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

Note that λ_r is the rth cumulant of the $B(m_{\bullet}, \tau_1)$ 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 $m_{\bullet}$ and for fixed sampling fraction, the λs are $O(m_{\bullet})$, 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

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

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 $\binom{m_{\bullet}}{m_1}$ 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

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

where $P_r(\psi)$ is the polynomial

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

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:

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

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

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

7.3.3 *Multivariate hypergeometric distribution*

Suppose that $\mathbf{Y}_1 \sim M(m_1, \boldsymbol{\pi})$ and $\mathbf{Y}_2 \sim M(m_2, \boldsymbol{\pi})$ are independent multinomial vectors, each on k categories. Then the conditional distribution of the vector $\mathbf{Y} \equiv \mathbf{Y}_1$ given that $\mathbf{Y}_1 + \mathbf{Y}_2 = \mathbf{s}$ is as follows.

$$\mathrm{pr}(\mathbf{Y} = \mathbf{y} \mid \mathbf{s}) = \frac{\binom{m_1}{\mathbf{y}}\binom{m_2}{\mathbf{s} - \mathbf{y}}}{\binom{m_.}{\mathbf{s}}} = \frac{\binom{s_1}{y_1} \cdots \binom{s_k}{y_k}}{\binom{s_.}{y_.}} \tag{7.13}$$

where $s_. \equiv m_.$ and $y_. \equiv m_1$. From a statistical perspective, one important aspect of this conditional distribution is that it does not depend on the multinomial probability vector $\boldsymbol{\pi}$.

An alternative derivation of (7.13) based on simple random sampling from a finite population of size $m_.$ is as follows. Suppose that attribute G has k levels and that the k levels of G are mutually exclusive and exhaustive. For instance G might denote a particular genetic marker such as blood group with levels O, A, B, AB. In a different context, G might denote the species of salmon in Lake Michigan, with levels whose labels are *coho*, *chinook*,.... Under simple random sampling, the distribution of species in a sample of size m_1 is given by (7.13), where $s_1, s_2, \ldots$ are the numbers of *coho*, *chinook*,... in the lake.

If the sampled and non-sampled individuals are arranged in a two-way table, the entries appear as follows.

	Attribute level				
	G_1	G_2	...	G_k	*Total*
Sampled	Y_{11}	Y_{12}	...	Y_{1k}	m_1
Not sampled	Y_{21}	Y_{22}	...	Y_{2k}	m_2
Total	s_1	s_2	...	s_k	$m_. \equiv s_.$

Evidently, in this table we may reverse the roles played by the rows and the columns. Thus, suppose that $Y_1 \sim B(s_1, \tau), \ldots, Y_k \sim B(s_k, \tau)$ are independent binomial random variables. Then the joint conditional distribution of $\mathbf{Y} = (Y_1, \ldots, Y_k)$, conditional on the event $Y_. = m_1$, is again given by (7.13). Note that in the first derivation, conditioning reduces the dimension of the sample space from $2(k-1)$ to $k-1$, in the process eliminating $k-1$ parameters $\boldsymbol{\pi}$.

In the latter derivation, conditioning reduces the dimension from k to $k-1$, in the process eliminating the single parameter τ.

The joint conditional mean vector and covariance matrix of $\mathbf{Y}$ are given by

$$\begin{aligned} E(Y_j) &= \tilde{\pi}_j m_1 \\ \operatorname{var}(Y_j) &= \tilde{\pi}_j(1-\tilde{\pi}_j)m_1 m_2/(m_\cdot - 1) \\ \operatorname{cov}(Y_i, Y_j) &= -\tilde{\pi}_i\tilde{\pi}_j m_1 m_2/(m_\cdot - 1), \end{aligned}$$

where $\tilde{\pi}_j = s_j/s_\cdot$ is the proportion in the population who have attribute j.

7.3.4 *Multivariate non-central hypergeometric distribution*

Suppose that $\mathbf{Y}_1 \sim M(m_1, \boldsymbol{\pi}_1)$ and $\mathbf{Y}_2 \sim M(m_2, \boldsymbol{\pi}_2)$ are independent multinomial random variables on k categories each. Then the conditional distribution of $\mathbf{Y} \equiv \mathbf{Y}_1$ given that $\mathbf{Y}_1 + \mathbf{Y}_2 = \mathbf{s}$ is as follows:

$$\operatorname{pr}(\mathbf{Y} = \mathbf{y} \mid \mathbf{s}, \boldsymbol{\psi}) = \frac{\binom{m_1}{\mathbf{y}}\binom{m_2}{\mathbf{s}-\mathbf{y}}\psi_1^{y_1}\cdots\psi_k^{y_k}}{\sum_j \binom{m_1}{\mathbf{j}}\binom{m_2}{\mathbf{s}-\mathbf{j}}\psi_1^{j_1}\cdots\psi_k^{j_k}}. \qquad (7.14)$$

The sum in the denominator runs over the entire conditional sample space, which comprises all non-negative integer-valued vectors $\mathbf{y}$ satisfying the positivity conditions

$$0 \le y_j \le s_j, \qquad \sum y_j = m_1.$$

The odds-ratio parameters ψ_j are defined as contrasts relative to category k by

$$\psi_j = \frac{\pi_{1j}\pi_{2k}}{\pi_{2j}\pi_{1k}}$$

so that $\psi_k \equiv 1$.

The exact non-central moments and cumulants are complicated functions of $\boldsymbol{\psi}$ and the marginal frequencies $\mathbf{s}$. Direct computation is awkward because of the nature of the sample space and because of the large number of points it contains. The following equation, however, gives a simple exact relationship between the conditional mean vector $\boldsymbol{\mu}_1$ 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

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

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

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

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

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

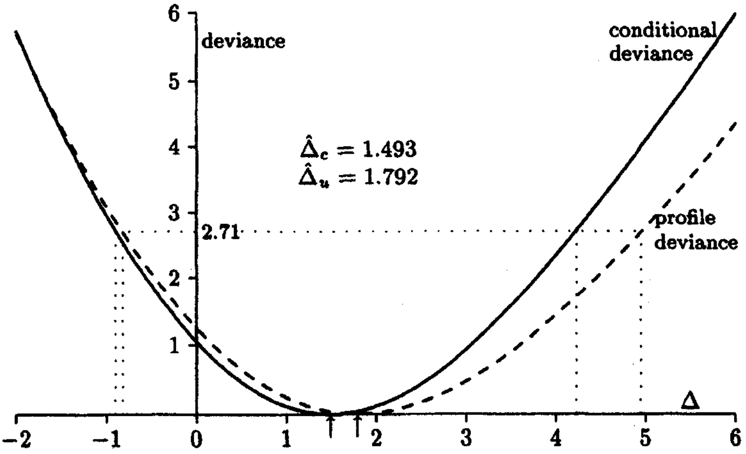

Fig. 7.1 *Graphical comparison of hypergeometric and binomial deviance functions for the data in Table 7.1. Nominal 90% intervals for the log odds ratio, Δ, are indicated.*

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

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

The conditional deviance function

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

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

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

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

Consequently, if we write

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

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

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

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

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

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

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

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

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

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

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

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*	
	80+	80−	80+	80−		*under model* (ii)	
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

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

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.

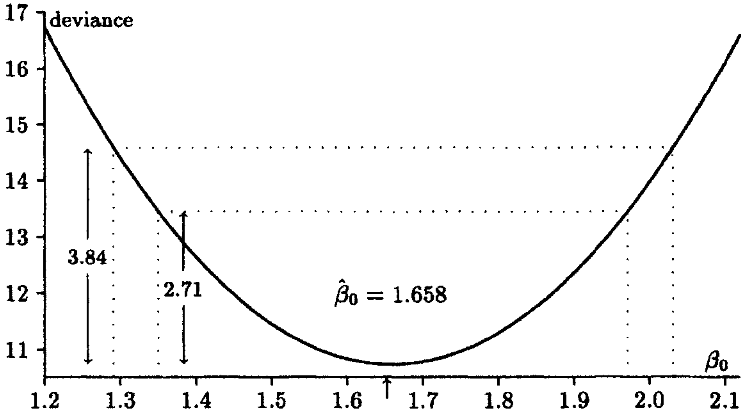

Fig. 7.2 *Hypergeometric deviance for model* (7.20). *Nominal* 90% *and* 95% *intervals for the log odds ratio,* β_0, *are indicated.*

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, $Z_{\bullet} = Z_1 + Z_2$, of the observed responses. If $Z_{\bullet} = (0, \ldots, 2, \ldots, 0)$, both R_1 and R_2 are determined by $Z_{\bullet}$ and the conditional distribution given $Z_{\bullet}$ is degenerate. However, if

$$Z_{\bullet} = (0, \ldots, 1, \ldots, 1, \ldots, 0),$$

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

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

which is independent of $\boldsymbol{\lambda}$ as required. Every ordered pair of responses (i, j) with $i \neq j$ contributes a factor (7.21) to the conditional likelihood. Identical pairs, for which the control response is the same as the treatment response, contribute a factor of unity and can be ignored in the conditional likelihood. Thus, if Y_{ij} is the number of ordered pairs responding (i, j), the symmetric total $m_{ij} = Y_{ij} + Y_{ji}$ is just the number of vector sums $Z_{\bullet}$ that have values in positions i and j. Hence, conditionally

$$\begin{aligned} Y_{ij} &\sim B(m_{ij}, \pi_{ij}) \qquad i < j \\ \operatorname{logit}(\pi_{ij}) &= \Delta_j - \Delta_i. \end{aligned} \tag{7.22}$$

The conditional log likelihood is therefore the product of $k(k-1)/2$ independent binomial factors satisfying the model shown above. As usual, the levels of the treatment factor $\boldsymbol{\Delta}$ can be chosen to satisfy $\Delta_1 = 0$ or $\Delta_k = 0$ or $\sum \Delta_j = 0$. Evidently, from (7.22) only the differences are relevant.

Model (7.22), known as the model of quasi-symmetry, was first suggested by Caussinus (1965), though no derivation was given. The same model occurs in studies of population migrations where the term *gravity model* is used. In that context the k categories are k geographical locations and Y_{ij} is the number of families or individuals who migrate from area i to area j in the time period under study. For further details see Scholten and van Wissen (1985) or Upton (1985).

Model (7.22) is formally identical to the Bradley-Terry (1952) model used for ranking individuals in a paired competition. If the probability π_{ij} that subject i beats subject j satisfies (7.22), then the Δs give a linear ranking of subjects.

A curious and unusual feature of the linear logistic model (7.22) is that the model matrix $\mathbf{X}$ corresponding to the formula $\Delta_j - \Delta_i$ does not include the intercept nor does the constant vector lie in the column space of $\mathbf{X}$. For instance if $k = 3$ the model formula may be stated explicitly using matrix notation as

$$\begin{pmatrix} \operatorname{logit} \pi_{12} \\ \operatorname{logit} \pi_{13} \\ \operatorname{logit} \pi_{23} \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \Delta_1 \\ \Delta_2 \\ \Delta_3 \end{pmatrix}$$

In this example $\mathbf{X}$ is 3×3 with rank 2 and the sum of the three columns is $\mathbf{0}$.

The adequacy of the linear logistic model (7.22) can be tested by using the residual deviance or Pearson's statistic, each of which has $(k-1)(k-2)/2$ degrees of freedom.

7.5.2 *Ordinal responses*

Consider the application of the proportional-odds model (5.1) for the comparison of two multinomial responses in which the categories are ordered. The observations comprise two independent multinomial vectors

$$\mathbf{Y}_1 \sim M(m_1, \boldsymbol{\pi}_1), \qquad \mathbf{Y}_2 \sim M(m_2, \boldsymbol{\pi}_2),$$

in which the cumulative probabilities γ_{1j}, γ_{2j} satisfy

$$\begin{aligned} \log\{\gamma_{1j}/(1-\gamma_{1j})\} &= \theta_j; \\ \log\{\gamma_{2j}/(1-\gamma_{2j})\} &= \theta_j - \Delta; \end{aligned} \qquad j = 1, \ldots, k-1. \tag{7.23}$$

In this model there is a single parameter of interest, Δ measuring the effect of treatment, and $k-1$ base-line parameters that determine the response probabilities for the control. Since the categories are ordered it is helpful to form cumulative totals not just for the probabilities but for the responses themselves. Thus we write $s_j = y_{1j} + y_{2j}$ for the response category totals and

$$\begin{aligned} Z_{1j} &= Y_{11} + \cdots + Y_{1j} \\ Z_{2j} &= Y_{21} + \cdots + Y_{2j} \\ S_j &= s_1 + \cdots + s_j = Z_{\cdot j} \end{aligned}$$

for the cumulative responses in each group and the cumulative totals respectively. With this notation we have $Z_{ij} \sim B(m_i, \gamma_{ij})$ with γ_{ij} satisfying the linear logistic model (7.23). The nuisance parameters may be eliminated by conditioning on S_j. The resulting hypergeometric distribution is

$$\operatorname{pr}(Z_{1j} = z_{1j} \mid S_j) = \binom{m_1}{z_{1j}}\binom{m_2}{S_j - z_{1j}}\psi^{z_{1j}} \Big/ \sum_r \binom{m_1}{r}\binom{m_2}{S_j - r}\psi^r, \tag{7.24}$$

where $\psi = e^{\Delta}$.

Thus, for each $j = 1, \ldots, k-1$, we have the conditional distribution

$$Z_{1j} \mid S_j \sim H(\mathbf{m}, S_j; \psi)$$

independently of the nuisance parameters. The aim in the discussion that follows is to construct an efficient estimate of ψ on the basis of these $k-1$ conditional distributions. Unfortunately there is no joint conditional distribution of $\{Z_{1j}\}$ that depends only on ψ because the conditional distribution of Z_{1j}, given the vector $\boldsymbol{S} = (S_1, \ldots, S_k)$, depends on both $\boldsymbol{\theta}$ and ψ.

The following method of estimation seems to work well and has much in common with quasi-likelihood as discussed in Chapter 9. The argument runs as follows. Let $\chi_{1j}(\psi)$ be the conditional mean of Z_{1j} given S_j as derived from the hypergeometric distribution (7.24). For each $j = 1, \ldots, k-1$, the difference

$$Z_{1j} - \chi_{1j}(\psi)$$

has zero mean conditionally on S_j, and hence unconditionally. Note that $\chi_{1j}(\psi)$ depends also on S_j. For all generalized linear models the likelihood equations take the form $U(\hat{\psi}; y) = 0$ where $U(\psi; y)$ is a linear function of the data. By analogy, therefore, we seek to construct a function

$$U(\psi; Z) = \sum w_j^* \{Z_{1j} - \chi_{1j}(\psi)\}, \tag{7.25}$$

where $w_j^* = w_j^*(\psi, S_j)$, such that $U(\psi; Z)$ behaves 'like' a log-likelihood derivative. By construction, $U(\psi; Z)$ has zero mean whatever the choice of weights. The weights are chosen so that the mean of $-\partial U/\partial\Delta$ is equal to the variance of U.

The choice of weights is normally not critically important. At worst, a poor choice of weights may lead to a small loss of efficiency. With this in mind, we may choose $w_j^* = 1$ or $w_j^* = S_j(m_{\bullet} - S_j)$. These are safe choices that lead to consistent estimates of ψ, but the variance of the resulting estimate is not simply related to $\partial U/\partial\Delta$.

More formally, however, standard theory for linear estimating equations shows that the optimal weights are given in vector form by

$$\mathbf{w}^* = \mathbf{V}^{-1}\mathbf{d},$$

where $\mathbf{V}$ is the covariance matrix of $Z_{11},\ldots,Z_{1\,k-1}$ and $d_j = \partial\chi_{1j}(\psi)/\partial\Delta$. By a common property of exponential-family models,

$$d_j = \operatorname{var}(Z_{1j} \mid S_j; \psi),$$

which is easily computed either exactly using (7.24) or approximately using (7.12). However, it is unclear what is meant by the covariance matrix $\mathbf{V}$ because the random variables $Z_{1j},\ldots,Z_{1\,j-1}$, whose probability distributions are given by (7.24), are defined on different sample spaces. To use the unconditional covariance matrix would be to violate the spirit of the exercise. Yet it does not make sense to talk of the covariance of two random variables unless they can be defined on a common sample space.

A pragmatic solution that has the merit of simplicity is to use the approximate hypergeometric covariance matrix $\tilde{\boldsymbol{\Sigma}}$ as defined in (7.16). Since $\mathbf{Z}$ is the vector of cumulative totals, we have

$$\tilde{\mathbf{V}} = \mathbf{L}\tilde{\boldsymbol{\Sigma}}\mathbf{L}^T,$$

where $\mathbf{L}$ is the lower triangular matrix forming cumulative totals. The matrix $\tilde{\mathbf{V}}$ thus constructed is a Green's matrix similar to the cumulative multinomial covariance matrix (5.13). With this choice for $\mathbf{V}$, we find

$$\begin{aligned} U(\psi; Z) &= \mathbf{d}^T\mathbf{D}^T\tilde{\boldsymbol{\Sigma}}^{-}\mathbf{D}(\mathbf{Z} - \boldsymbol{\chi}) \\ &= \mathbf{d}^T\mathbf{D}^T\tilde{\boldsymbol{\Sigma}}^{-}(\mathbf{Y}_1 - \boldsymbol{\mu}_1) \end{aligned}$$

where $\mathbf{D} = \mathbf{L}^{-1}$, $\mu_{1j} = \chi_{1j} - \chi_{1\,j-1}$ and $\mathbf{Y}_1^T = (Y_{11},\ldots,Y_{1k})$. We are free to choose the simplest generalized inverse of $\tilde{\boldsymbol{\Sigma}}$, namely

$$\tilde{\boldsymbol{\Sigma}}^{-} = \frac{m_{\cdot} - 1}{m_{\cdot}}\operatorname{diag}\{\zeta^{-1}\} = \frac{m_{\cdot} - 1}{m_{\cdot}}\operatorname{diag}\{\mu_{1j}^{-1} + \mu_{2j}^{-1}\}.$$

Thus

$$\begin{aligned} U(\psi; Z) &= \frac{m_{\cdot} - 1}{m_{\cdot}}\sum_{j=1}^{k}(d_j - d_{j-1})\Big(\frac{1}{\mu_{1j}} + \frac{1}{\mu_{2j}}\Big)(y_{1j} - \mu_{1j}) \\ &\simeq \sum_{j=1}^{k-1}\Big(\frac{\zeta_j + \zeta_{j+1}}{\zeta_{\cdot}}\Big)(Z_{1j} - \chi_{1j}). \qquad (7.26) \end{aligned}$$

The latter approximation comes from replacing $\mathbf{d}$ by the diagonal elements of $\tilde{\mathbf{V}}$. Note that the weights in the first expression above are not all positive.

The 'conditional likelihood' estimate, $\hat{\Delta}_c$, defined as the solution to the equation $U(\hat{\psi}_c; Z) = 0$, has asymptotic variance

$$\begin{aligned} \operatorname{var}(\hat{\Delta}_c) &\simeq \zeta_{.}\Big\{\sum_{j=1}^{k-1}(\zeta_j + \zeta_{j+1})d_j\Big\}^{-1} \\ &\simeq \frac{3(m_{.} - 1)}{m_{.}\zeta_{.}}\Big\{1 - \sum\{\zeta_j/\zeta_{.}\}^3\Big\}^{-1}, \qquad (7.27) \end{aligned}$$

which is essentially the same as the variance of the unconditional maximum-likelihood estimate. See Exercise 5.3.

7.5.3 *Example*

It is only for very sparse tables that there is any appreciable difference between the 'conditional' likelihood estimator described in the previous section, and the unconditional maximum likelihood estimate as described in Chapter 5. By way of example, we consider here the first two rows of Table 5.1, involving the comparison of two cheese additives, A and B, ignoring C and D. The unconditional maximum-likelihood estimate of Δ in the proportional odds model (7.23) is -3.028 with asymptotic standard error 0.455. Table 7.3 shows the steps involved in one cycle of the iteration, beginning from $\Delta_0 = -3.028$.

Table 7.3 *Steps required in one cycle of the iteration to compute* $\hat{\Delta}_c$

S	$\mathbf{Z}_1$	$\mathbf{d}$	$\boldsymbol{\chi}_1$	$\boldsymbol{\mu}_1$	$\boldsymbol{\zeta}$	$\mathbf{w}^*$
6	0	0.2842	0.3021	0.3021	0.2869	0.0615
15	0	0.8220	0.8997	0.5976	0.5579	0.1307
28	1	1.8891	2.2860	1.3863	1.2385	0.3288
46	8	3.6281	6.5994	4.3134	3.2798	0.5105
61	16	3.4019	14.5594	7.9600	3.7359	0.4485
75	24	1.9867	25.4325	10.8731	2.4285	0.3050
95	43	0.4461	43.4791	18.0466	1.7626	0.1580
103	51	0.0440	51.0462	7.5671	0.4095	0.0330
104	52	—	52.0	0.9538	0.0441	—
				52.0	13.7437	

The first column gives the cumulative category totals, S_j, and the second column gives the cumulative observations Z_{1j} for the first group. The third and fourth columns give the variances and means

$$d_j = \mathrm{var}(Z_j \mid S_j) \quad \text{and} \quad X_{1j} = E(Z_{1j} \mid S_j),$$

both computed from the hypergeometric distribution with odds ratio $\psi_0 = \exp(-3.028)$. The fifth column gives the cell means for the first group, $\mu_{1j} = X_{1j} - X_{1\,j-1}$. Finally, the last two columns give

$$\zeta_j = \left(\frac{1}{\mu_{1j}} + \frac{1}{\mu_{2j}}\right)^{-1}$$

where $\mu_{2j} = s_j - \mu_{1j}$ are the cell means for the second group, and $w_j^* = (\zeta_j + \zeta_{j+1})/\zeta_.$ are the required weights.

The score statistic $U = \sum w_j^*(Z_{1j} - X_{1j})$ is equal to 0.2881, while $\sum w_j^* d_j = 4.8016$. Thus the updated estimate of Δ is

$$\hat{\Delta}_c = -3.028 + 0.2881/4.8016 = -2.9680.$$

One further cycle gives $\hat{\Delta}_c = -2.9743$ with asymptotic variance

$$\left(\sum w_j^* d_j\right)^{-1} = (4.9047)^{-1} = 0.2039$$

and standard error 0.4515.

The difference between the conditional and unconditional estimates is only 12% of a standard error and is unlikely to have much effect on the conclusions reached. As usual, the conditional estimate is smaller in magnitude than the unconditional estimate.

7.6 Bibliographic notes

Conditional and marginal likelihoods for the elimination of nuisance parameters have been in use since the early part of this century. It can be argued that Student's usage of degrees of freedom rather than sample size as divisor in the estimation of σ^2 is an application of marginal likelihood, as shown in section 7.2.1. Bartlett (1936, 1937) made further important contributions, particularly to the problem of estimating a common mean when the sample variances are unequal and unknown.

Neyman and Scott (1948), in an important and influential paper, pointed out that when the number of nuisance parameters grows in proportion to n, maximum-likelihood estimates need not be consistent. Even if they are consistent, they need not be efficient.

The use of marginal likelihoods based on error contrasts, for the estimation of variance components, has been recommended by Patterson and Thompson (1971) and further studied by Harville (1974, 1977), Fraser (1968, 1979) and Corbeil and Searle (1976). The method is known as restricted maximum likelihood (REML). The application of the same technique to spatial covariance estimation is due to Kitanidis (1983, 1987), who points out that the marginal likelihood is superior in this context to full, or profile, likelihood.

The matched pairs design is an extreme case where the number of parameters grows in proportion to n, and consequently this design is used as a testing ground for procedures that purport to handle large numbers of nuisance parameters. For binary responses, Cox (1958b) showed that the conditional likelihood ignores all pairs for which the responses are equal. The test for no treatment effect is a simple comparison of the number of pairs responding $(0, 1)$ with those responding $(1, 0)$. This test had been proposed earlier by McNemar (1947). For further details, see Andersen (1973).

Kalbfleisch and Sprott (1970) give an excellent review of various modifications to the likelihood when there is a large number of nuisance parameters. In particular, they discuss the thorny problem of conditioning, when the conditioning statistic for the removal of $\boldsymbol{\lambda}$, $S_{\lambda}(\boldsymbol{\psi})$, depends on $\boldsymbol{\psi}$. See also Godambe (1976) and Lindsay (1982) for a discussion of the same topic from the vantage of optimal estimating equations.

Regression models for the log odds-ratio, based on the non-central hypergeometric distribution, are now widely used in retrospective studies of disease. This line of work can be traced back to Mantel and Haenszel (1959). For more recent work, see Breslow (1976, 1981) and Breslow and Day (1980). There are close connections here with Cox's (1972a, 1975) partial likelihood, which is also designed for the removal of nuisance parameters.

The formulae in section 7.3.2 for non-central hypergeometric moments were given by Harkness (1965). They were subsequently discussed by Mantel and Hankey (1975) and used by Mantel (1977) for approximating the non-central hypergeometric mean.

Saddlepoint methods for approximating conditional likelihoods for generalized linear models with canonical links are discussed by Davison (1988).

Section 7.5 is based on McCullagh (1982, 1984c).

7.7 Further results and exercises 7

7.1 Suppose that $Y_1, \ldots, Y_n$ are *i.i.d.* $N(\mu, \sigma^2)$. Show that if $\mu = \mu_0$ is given, then

$$S_0 \equiv S(\mu_0) = \sum (Y_j - \mu_0)^2$$

is a complete sufficient statistic for σ^2.

7.2 Show that the log likelihood for (μ, σ^2) in the previous exercise is

$$l(\mu, \sigma^2) = -\frac{1}{2\sigma^2} S(\mu) - \frac{n}{2} \log \sigma^2.$$

Show also that the statistic $S(\mu_0)$ has the non-central χ^2 distribution on n degrees of freedom given by

$$\frac{\exp\{-(S_0 + \lambda)/(2\sigma^2)\}\left(S_0/(2\sigma^2)\right)^{n/2-1}}{2\sigma^2 \pi^{1/2} \Gamma\left(\frac{n-1}{2}\right)} \sum_{r=0}^{\infty} \left(\frac{\lambda}{2\sigma^2}\right)^r \frac{r^r}{r!} B\left(\frac{n-1}{2}, r + \tfrac{1}{2}\right)$$

where $\lambda = n(\mu - \mu_0)^2$. Hence show that the conditional log likelihood given S_0 is

$$\begin{aligned} l_c(\mu, \sigma^2; \mu_0) = & -\frac{1}{2\sigma^2}\{S(\mu) - S(\mu_0)\} - \left(\frac{n}{2} - 1\right) \log S(\mu_0) \\ & + \frac{\lambda}{2\sigma^2} - \log \sum_{r=0}^{\infty} \left(\frac{\lambda}{2\sigma^2}\right)^r \frac{r^r}{r!} B\left(\frac{n-1}{2}, r + \tfrac{1}{2}\right), \end{aligned}$$

whereas the reduced function $l^*(\mu, \sigma^2)$ is

$$l^*(\mu, \sigma^2) = l_c(\mu, \sigma^2; \mu) = -\tfrac{1}{2}(n-2) \log S(\mu).$$

7.3 For the function l^* defined in the previous exercise, show that

$$U^* = \frac{\partial l^*}{\partial \mu} = \frac{(n - \sum(y}{S(\mu)} \sum (y - \mu) = \frac{n(n-2)(\bar{y} - \mu)}{(n-1)s^2 + n(\bar{y} - \mu)^2}$$

whereas the derivative of $l_c(\mu, \sigma^2; \mu_0)$ with respect to μ, evaluated at $\mu_0 = \mu$ is

$$\left.\frac{\partial l_c}{\partial \mu}\right|_{\mu_0 = \mu} = \frac{1}{\sigma^2} \sum (y - \mu),$$

which is monotonely increasing in $\bar{y}$. Comment briefly on the differences between the two derivatives for $n = 1, 2$.

7.4 For the problem discussed in the previous three exercises, show that Y_i and Y_j are conditionally uncorrelated with variance $S(\mu)/n$ and that

$$\operatorname{var}(\bar{Y} \mid S(\mu)) = S(\mu)/n^2.$$

Hence deduce that the conditional moments of the derivatives of l^* are

$$\begin{aligned} E(U^* \mid S(\mu)) &= 0 \\ \operatorname{var}(U^* \mid S(\mu)) &= (n-2)^2/S(\mu) \\ -E\{\partial^2 l^*/\partial \mu^2 \mid S(\mu)\} &= (n-2)^2/S(\mu). \end{aligned}$$

[Bartlett, 1936].

7.5 Suppose that $Y_1, \ldots, Y_n$ are observations taken at spatial locations $s_1, \ldots, s_n$ and that the vector $\mathbf{Y}$ may be taken as multivariate Normal with mean and variance given by

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}, \qquad \operatorname{cov}(\mathbf{Y}) = \boldsymbol{\Sigma}(\boldsymbol{\theta}; \mathbf{s}),$$

where $\mathbf{X}$ is given and $\boldsymbol{\beta}$ is a nuisance parameter. The parameters $\boldsymbol{\theta}$ appearing in the covariance function, $\boldsymbol{\Sigma}(\boldsymbol{\theta}; \mathbf{s})$, are the focus of investigation. Show that for any given value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, the statistic

$$S_0 = \mathbf{X}^T \mathbf{W}_0 \mathbf{Y}, \quad \text{where} \quad \mathbf{W}_0^{-1} = \boldsymbol{\Sigma}(\boldsymbol{\theta}_0; \mathbf{s}),$$

is sufficient for $\boldsymbol{\beta}$. Show that the conditional log likelihood, $l_{Y|S_0}(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{\theta}_0)$, for $(\boldsymbol{\beta},\boldsymbol{\theta})$ given S_0 is

$$-\tfrac{1}{2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\{\boldsymbol{\Sigma}^{-1}-\mathbf{W}_0\mathbf{X}(\mathbf{X}^T\mathbf{W}_0\boldsymbol{\Sigma}\mathbf{W}_0\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_0\}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$$
$$-\tfrac{1}{2}\log\det\boldsymbol{\Sigma}+\tfrac{1}{2}\log\det(\mathbf{X}^T\mathbf{W}_0\boldsymbol{\Sigma}\mathbf{W}_0\mathbf{X}),$$

and hence that the reduced function $l^*(\boldsymbol{\theta}) = l_{Y|S}(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{\theta})$ satisfies

$$\begin{aligned} l^*(\boldsymbol{\beta},\boldsymbol{\theta}) = &-\tfrac{1}{2}\mathbf{Y}^T\boldsymbol{\Sigma}^{-1}(\mathbf{I}-\mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1})\mathbf{Y} \\ &-\tfrac{1}{2}\log\det\boldsymbol{\Sigma}+\tfrac{1}{2}\log\det(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}), \end{aligned}$$

which is independent of $\boldsymbol{\beta}$. By considering the special case $\mathbf{X} = \mathbf{I}$ or otherwise, show that $l^*(\boldsymbol{\beta},\boldsymbol{\theta})$ is not a log-likelihood function.

7.6 Suppose that Y_1, Y_2 are independent, Normally distributed with means μ_1, μ_2 and unit variances. Let $\psi = \mu_2/\mu_1$. Find the conditional distributions of Y_1, Y_2 and $\psi Y_1 - Y_2$ given $Y_1 + \psi Y_2 = C$. Show that these conditional distributions lead to different 'likelihoods'. Explain this phenomenon.

7.7 Find the conditional maximum-likelihood equations for $\boldsymbol{\theta}$ in exercise 7.5 using equation (7.2). Compare this with the marginal maximum-likelihood estimate based on the residuals as described in section 7.2.1.

7.8 Let $\mathbf{H} = [\mathbf{H}_1 : \mathbf{H}_2]$ be an orthogonal matrix partitioned into $\mathbf{H}_1$ of order $n \times n-p$ and $\mathbf{H}_2$ of order $n \times p$. Let $\boldsymbol{\Sigma}_0$ be an arbitrary symmetric matrix of rank n. Show that

$$\det\boldsymbol{\Sigma}_0 = \det(\mathbf{H}_1^T\boldsymbol{\Sigma}_0\mathbf{H}_1)/\det(\mathbf{H}_2^T\boldsymbol{\Sigma}_0^{-1}\mathbf{H}_2).$$

provided that $\det(\mathbf{H}_2^T\boldsymbol{\Sigma}_0^{-1}\mathbf{H}_2) \neq 0$.

7.9 Let $\mathbf{P}_X = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ be a projection matrix of order n and rank p. By writing $\mathbf{I} - \mathbf{P}_X = \mathbf{H}\mathbf{I}_{[n-p]}\mathbf{H}^T$, where $\mathbf{H}$ is orthogonal, show that the matrix

$$\boldsymbol{\Sigma}^* = (\mathbf{I}-\mathbf{P}_X)\boldsymbol{\Sigma}_0(\mathbf{I}-\mathbf{P}_X)$$

satisfies

$$\log\mathrm{DET}(\boldsymbol{\Sigma}^*) = \log\det\boldsymbol{\Sigma}_0 + \log\det(\mathbf{X}^T\boldsymbol{\Sigma}_0^{-1}\mathbf{X}) - 2\log\mathrm{DET}(\mathbf{X}),$$

where $\mathrm{DET}(A)$ is defined as the product of the singular values of A.

$$\mathrm{DET}(\mathbf{A}) = \prod_{\lambda \neq 0} \lambda_j(\mathbf{A}).$$

It is assumed here that $\mathbf{X}$ is $n \times p$ with rank p and that $\mathbf{\Sigma}_0$ is positive definite and symmetric. Thus $\det(\mathbf{\Sigma}_0) = \mathrm{DET}(\mathbf{\Sigma}_0)$.

7.10 Suppose that $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed with density $\frac{1}{\sigma} f\left(\frac{\epsilon}{\sigma}\right)$, depending on the unknown parameter σ. Suppose also that the observed values $y_1, \ldots, y_n$ satisfy

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

for fixed known matrices $\mathbf{X}, \mathbf{Z}$ and unknown parameters $\boldsymbol{\beta}, \boldsymbol{\gamma}$. Show that the distribution of $\mathbf{R} = (\mathbf{I} - \mathbf{P}_X)\mathbf{Y}$ does not depend on $\boldsymbol{\beta}$. $[\mathbf{P}_X = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T.]$

Show also that if the ϵs are Normally distributed, the marginal likelihood based on $\mathbf{R}$ is identical to the conditional likelihood given $\mathbf{P}_X\mathbf{Y}$.

7.11 Define the projection matrices $\mathbf{P}$ and $\mathbf{P}_W$ by

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \quad \text{and} \quad \mathbf{P}_W = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}$$

where $\mathbf{W} = \mathbf{\Sigma}^{-1}$. The corresponding residual vectors are

$$\mathbf{R} = (\mathbf{I} - \mathbf{P})\mathbf{Y} \quad \text{and} \quad \mathbf{R}_W = (\mathbf{I} - \mathbf{P}_W)\mathbf{Y}.$$

Assuming that $\operatorname{cov}(\mathbf{Y}) = \mathbf{\Sigma}$, show that

$$\operatorname{cov}(\mathbf{R}_W) = (\mathbf{I} - \mathbf{P}_W)\mathbf{\Sigma} \quad (= \mathbf{\Sigma}_W),$$

and that

$$\mathbf{\Sigma}_W^- = \mathbf{\Sigma}^{-1}(\mathbf{I} - \mathbf{P}_W)$$

is the Moore-Penrose inverse of $\mathbf{\Sigma}_W$. Hence deduce that

$$\mathbf{\Sigma}_I^- = (\mathbf{I} - \mathbf{P})\mathbf{\Sigma}^{-1}(\mathbf{I} - \mathbf{P}_W) \quad \text{and} \quad \mathbf{\Sigma}_W^-$$

are both generalized inverses of $(\mathbf{I} - \mathbf{P})\mathbf{\Sigma}(\mathbf{I} - \mathbf{P})$.

7.12 Show, using the results of the previous exercise, that

$$\mathbf{R}^T\Sigma_I^-\mathbf{R} = \mathbf{R}_W^T\Sigma_W^-\mathbf{R}_W = \mathbf{Y}^T\mathbf{W}\mathbf{Y} - \mathbf{Y}^T\Sigma^{-1}\mathbf{P}_W\mathbf{Y}.$$

7.13 Show that $\mathbf{R}_W = (\mathbf{I} - \mathbf{P}_W)\mathbf{R}$. Hence deduce that the log likelihood based on $\mathbf{R}$ is identical to that based on $\mathbf{R}_W$ for fixed $\mathbf{W}$.

7.14 Show that the simultaneous solution to equations (7.11) and (7.12) can be obtained by iteration using the following steps, beginning with initial estimates $\mu_{ij}^{(1)}$ for the conditional means.

(i) $$\kappa_2^{(r)} = \frac{m_\cdot}{m_\cdot - 1}\left(\frac{1}{\mu_{11}^{(r)}} + \frac{1}{\mu_{12}^{(r)}} + \frac{1}{\mu_{21}^{(r)}} + \frac{1}{\mu_{22}^{(r)}}\right)^{-1}$$

(ii) $$\psi^{(r)} = \frac{\mu_{11}^{(r)}\mu_{22}^{(r)} + \kappa_2^{(r)}}{\mu_{12}^{(r)}\mu_{21}^{(r)} + \kappa_2^{(r)}},$$

(iii) $$\mu_{11}^{(r+1)} = \mu_{11}^{(r)} + \kappa_2^{(r)} \times \{\log\psi - \log\psi^{(r)}\}.$$

Hint: for part (iii) use the property of exponential families that $\partial\mu/\partial\theta = \kappa_2$. If the initial estimates are poorly chosen, care must be taken in step (iii) to ensure that μ_{11} does not get out of range. Otherwise the algorithm seems to converge rapidly. [Liao, 1988].

7.15 Show that the two asymptotic variance formulae (7.27) are not identical but are numerically very similar. Compute both expressions for the data in Table 7.3 and show that the difference is approximately one half of 1%.

7.16 Show that if a particular response category is not used, that category may be deleted without affecting (7.26).

7.17 The estimating equation (7.26) does not have the form specified in (7.25) because the weight $(\zeta_j + \zeta_{j+1})/\zeta_\cdot$ depends on the whole vector S and not just on S_j. Discuss the possible implications of this.

7.18 Fit a model to the data in Table 7.2 in which the odds ratio is constant up to age 75, but different in the 75+ age-group, Show that this model fits better than the linear regression model in (7.20). Give an approximate significance level for the observed difference, making due allowance for selection effects. Use the deviance reduction rather than the parameter estimate as your test statistic.

7.19 For the Fieller-Creasy problem discussed at the end of section 7.2.2, in which the parameter of interest is the ratio of Normal means, show that the bias-corrected derivative of $l^*(\psi)$ is

$$\frac{\partial l^*(\psi)}{\partial\psi} - E\left(\frac{\partial l^*(\psi)}{\partial\psi}\,\Big|\, S_\lambda(\psi), \psi\right) = \frac{y_2 - \psi y_1}{1+\psi^2}\,\frac{s_\lambda(\psi)}{1+\psi^2}.$$

Discuss briefly the use of this unbiased estimating equation as an alternative to (7.3). Under what circumstances do the two methods produce identical estimating equations?

7.20 Suppose that (Y_1, Y_2) are bivariate Normal with mean vector $(\cos\theta, \sin\theta)$ and covariance matrix $n^{-1}\mathbf{I}_2$. The parameter space is taken to be $0 \le \theta < 2\pi$. Prove that $R = (Y_1^2 + Y_2^2)^{1/2}$ is ancillary for θ. Show that the conditional density of $\hat\theta$ given $R = r$ has the von Mises-Fisher form

$$\exp(nr\cos(\hat\theta - \theta))/I_0(nr), \qquad 0 \le \hat\theta < 2\pi$$

where $I_0(\cdot)$ is the modified Bessel function of order zero. Compute the conditional Fisher information for θ as a function of r, and construct a plot of the conditional information for θ given r in the range $0.5 \le r \le 2$. Comment on the difference between the conditional Fisher information and the observed Fisher information for θ.