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