APPENDIX A

Elementary likelihood theory

Scalar parameter

This appendix contains a concise summary, without proofs and omitting esoteric details of regularity conditions, of the more important properties of likelihood functions, derivatives and estimates, that are used throughout the book.

Definition: The log likelihood is the logarithm of the joint probability or probability density function of the data, denoted by

$$l(\theta; y) = \log f_Y(y; \theta).$$

If Y is a vector having n independent components, the log likelihood is a sum of n independent terms

$$l(\theta; \mathbf{y}) = \sum \log f_{Y_i}(y_i; \theta).$$

This representation is often used, at least implicitly, in the derivation of asymptotic results for large n.

Derivatives: Under mild regularity conditions the log-likelihood derivatives satisfy the following moment identities:

$$\begin{aligned} E_\theta\left(\frac{\partial l}{\partial \theta}\right) &= 0 \\ E_\theta\left(\frac{\partial^2 l}{\partial \theta^2}\right) + \mathrm{var}_\theta\left(\frac{\partial l}{\partial \theta}\right) &= 0. \end{aligned} \tag{A.1}$$

The notation above is chosen to emphasize the fact that the twin processes of differentiation and averaging take place at the same value of θ. These relations are obtained by differentiating with respect to θ the identity

$$\int f_Y(y; \theta)\, dy \equiv 1.$$

The necessary regularity conditions are those required to justify interchanging the order of differentiation with respect to the parameter and integration over the sample space. In particular, the sample space must be the same for all values of the parameter, or at least for all θ in an open neighbourhood of the true parameter point.

Further differentiation with respect to θ gives higher-order identities, sometimes called the Bartlett identities after Bartlett (1954). The third-order identity is

$$E_\theta\left(\frac{\partial^3 l}{\partial\theta^3}\right) + 3\,\mathrm{cov}_\theta\left(\frac{\partial^2 l}{\partial\theta^2}, \frac{\partial l}{\partial\theta}\right) + E_\theta\left(\frac{\partial l}{\partial\theta}\right)^3 = 0. \qquad (A.2)$$

These results, connecting moments of log-likelihood derivatives, are exact for all sample sizes provided, of course, that all the necessary moments are finite.

Terminology. The log-likelihood derivative $U(\theta; y) = \partial l/\partial\theta$ is sometimes called the *score statistic*. Its variance

$$i(\theta) = \mathrm{var}_\theta\left(\frac{\partial l}{\partial\theta}\right) = -E_\theta\left(\frac{\partial^2 l}{\partial\theta^2}\right)$$

is called the *Fisher information for* θ and plays an important role in much of what follows.

If the components of Y are independent we may write

$$U(\theta; y) = \sum \frac{\partial \log f_{Y_i}(y_i; \theta)}{\partial\theta},$$

$$i(\theta) = \sum E\left(\frac{\partial \log f_{Y_i}(y_i; \theta)}{\partial\theta}\right)^2,$$

showing explicitly that the score statistic is a sum of n independent contributions and that the Fisher information based on the vector Y is the sum of the Fisher informations from the components.

Asymptotic results. The following results hold under further regularity conditions related to the behaviour of the sequence of observations for large n, or to be more precise, as the amount of information, $i(\theta)$, becomes large. In particular, the first derivative suitably normalized converges in distribution to a standard Normal random variable. Thus

$$i(\theta)^{-1/2}\left(\frac{\partial l}{\partial\theta}\right) \sim N(0,1) + O_p(n^{-1/2}) \qquad (A.3)$$

provided that the assumed model is correct and that the derivative is computed at the true parameter point.

The error term is governed more by the magnitude of $i(\theta)$ than by the number of components of Y. Most commonly, however, $i(\theta)$ is roughly proportional to n and it is then immaterial whether we normalize by n or by $i(\theta)$.

Maximum-likelihood estimation: Ordinarily the likelihood function has a single maximum in the interior of the parameter space. The maximum-likelihood estimate, denoted by $\hat{\theta}$, is then obtained as the solution of the equation $U(\hat{\theta}; y) = 0$. For large $i(\theta)$, the distribution of $\hat{\theta}$ is often adequately approximated by

$$\hat{\theta} - \theta \sim N(0, i(\theta)^{-1}), \tag{A.4}$$

assuming, as always, that the model is correct. Higher-order approximations based on Edgeworth series are given by McCullagh (1987, p.210).

Occasionally, however, it may be found that the maximum occurs at a boundary point of the parameter space, which may be finite or infinite. The above approximation is then inappropriate. Approximate confidence limits can be obtained directly from the likelihood function or the likelihood-ratio statistic.

Likelihood-ratio statistics: For large n, the log likelihood at $\hat{\theta}$ differs from the log likelihood at the true parameter point by a random amount whose approximate distribution is given by

$$2l(\hat{\theta}; Y) - 2l(\theta; Y) \sim \chi^2_1 + O(n^{-1}). \tag{A.5}$$

This approximation is often quite accurate for small values of n even when the Normal approximation (A.4) is unsatisfactory. The set of all θ-values satisfying

$$2l(\hat{\theta}; y) - 2l(\theta, y) \leq \chi^2_{1,\alpha}$$

is an approximate $100(1 - \alpha)\%$ confidence set for the parameter and is usually more accurate in terms of coverage probability than intervals based on (A.4).

The above approximations can be improved further using methods given in Appendix C.

Vector parameter

For vector-valued parameters, the same results apply with suitable minor changes of notation. The score statistic is the gradient vector of the log likelihood at θ and the Fisher information $i(\theta)$ is now to be interpreted as a matrix. With obvious modifications, identities $(A.1)$ apply to the vector case. The vector version of $(A.2)$ is given by McCullagh (1987, p.202).

The asymptotic results $(A.3)$ and $(A.4)$ extend readily to vector-valued parameters provided that the limit $i(\theta) \to \infty$ is understood to apply to the eigenvalues of the information matrix and not to the components. An important regularity condition is that $i(\theta)$ have constant rank for all θ in the region of interest.

Nuisance parameters: Suppose that θ is partitioned into two components $\theta = (\psi, \lambda)$, both of which may be vector-valued. The first component is to be regarded as the parameter of interest. The joint Fisher information matrix for θ may then be partitioned as follows:

$$i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix}.$$

Its inverse is denoted by

$$i(\theta)^{-1} = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix}$$

so that, from the formulae for the inverse of a partitioned matrix,

$$\{i^{\psi\psi}\}^{-1} = i_{\psi\psi} - i_{\psi\lambda} i_{\lambda\lambda}^{-1} i_{\lambda\psi} \qquad (A.6)$$

is the approximate inverse covariance matrix of $\hat{\psi}$. Moreover, if $\hat{\lambda}_\psi$ is the maximum likelihood estimate of λ for fixed ψ, the gradient vector of the log likelihood, calculated at $(\psi, \hat{\lambda}_\psi)$, has approximate covariance matrix $\{i^{\psi\psi}\}^{-1}$ rather than $i_{\psi\psi}$. By contrast, the derivative with respect to ψ at (ψ, λ) has exact covariance matrix $i_{\psi\psi}$. Thus it is reasonable to regard $(A.6)$ rather than $i_{\psi\psi}$ as the Fisher information for ψ when λ is unknown. Note that two matrix inversions are required in order to produce the expression in $(A.6)$.

Likelihood-ratio statistics: For large n, the log likelihood at $\hat{\theta}$ differs from the log likelihood at the true parameter point by a random amount whose approximate distribution is given by

$$2l(\hat{\theta}; Y) - 2l(\theta; Y) \sim \chi_p^2 + O(n^{-1}), \qquad (A.7)$$

where p is the dimension of θ or the rank of $i(\theta)$. If there are nuisance parameters in the model, the maximized likelihood-ratio statistic has approximate distribution

$$2l(\hat{\psi}, \hat{\lambda}; Y) - 2l(\psi, \hat{\lambda}_\psi; Y) \sim \chi^2_{p-q} + O(n^{-1}), \qquad (A.8)$$

where $p - q$ is the dimension of ψ or the rank of $i^{\psi\psi}$ in (A.6). These approximations are often quite accurate for small values of n even when Normal approximations for parameter estimates are unsatisfactory. The set of all ψ-values satisfying

$$2l(\hat{\psi}, \hat{\lambda}; y) - 2l(\psi, \hat{\lambda}_\psi, y) \leq \chi^2_{p-q,\alpha}$$

is an approximate $100(1 - \alpha)\%$ confidence set for ψ.

APPENDIX B

Edgeworth series

Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed random variables having finite cumulants, $\kappa_1 \equiv \mu$, $\kappa_2 \equiv \sigma^2$, κ_3, κ_4, up to order four. Define the standardized sum

$$X_n = \frac{Y_1 + \ldots + Y_n - n\mu}{\sigma\sqrt{n}}.$$

Denote by $F_n(x)$ the probability $\mathrm{pr}(X_n \leq x)$. By the central limit theorem, X_n is asymptotically standard normal and $F_n(x) \to \Phi(x)$ as $n \to \infty$. If the distribution of Y has a continuous component, then $F_n(x)$ may be approximated more accurately for large n by an Edgeworth series as follows:

$$E_n(x) = \Phi(x) - \phi(x)\Big\{\rho_3(x^2-1)/(6n^{1/2}) + \rho_4(x^3-3x)/(24n) + \rho_3^2(x^5-10x^3+15x)/(72n)\Big\} \qquad (B.1)$$

where $\rho_3 = \kappa_3/\kappa_2^{3/2}$ and $\rho_4 = \kappa_4/\kappa_2^2$ are the standardized cumulants of Y. The difference $F_n(x) - E_n(x)$ is $o(n^{-1})$ uniformly in x on bounded intervals.

In the case of lattice distributions, this expansion is incorrect because $F_n(x)$ is discontinuous with jumps of order $O(n^{-1/2})$ at the possible values of X_n. The Edgeworth approximation is continuous and hence must involve an error of order $O(n^{-1/2})$ near the discontinuities of $F_n(x)$. However, the Edgeworth series can be adjusted to accommodate these discontinuities in $F_n(x)$ as follows. Suppose that the possible values of Y_i are the integers $0, 1, 2, \ldots$. Define the continuity-corrected abscissa and a 'precision

adjustment' or Sheppard correction as follows:

$$z = \frac{y_1 + \ldots + y_n - n\mu + \frac{1}{2}}{\sigma\sqrt{n}}$$
$$\tau = 1 + \frac{1}{24n\sigma^2}. \qquad (B.2)$$

Then the Edgeworth series with the usual two correction terms may be written

$$F_n(z) = E_n(\tau z) + o(n^{-1}). \qquad (B.3)$$

This approximation is valid only when computed at the 'continuity-corrected' points as defined by $(B.2)$. The distribution function is constant over intervals of the form $\left[z - \frac{1}{2\sigma\sqrt{n}}, z + \frac{1}{2\sigma\sqrt{n}}\right)$.

Note that the correction terms in $(B.3)$ are identical to the correction terms in $(B.1)$. The only difference is the correction for continuity and the adjustment of the argument. The continuity correction has an effect of order $O(n^{-1/2})$ and the Sheppard correction has an effect of order $O(n^{-1})$.

The discrete Edgeworth approximation may be used for the binomial distribution where $Y \sim B(m, \pi)$, provided that m is sufficiently large. The relevant coefficients are

$$z = (y - m\pi + \tfrac{1}{2})/\sqrt{m\pi(1-\pi)}$$
$$\tau = 1 + 1/\{24m\pi(1-\pi)\}$$
$$\rho_3 = (1 - 2\pi)/\sqrt{m\pi(1-\pi)}$$
$$\rho_4 = \{1 - 6\pi(1-\pi)\}/\{m\pi(1-\pi)\}.$$

The sample size, n, is built into these coefficients through the binomial index, m. Thus we may take $n = 1$ in $(B.1)$ and $(B.3)$. In this case, the approximation seems to be quite accurate if $m\pi(1-\pi)$ exceeds 2.0.

Approximation $(B.3)$ is a simplified version of a series expansion given by Esseen (1945), who gives the expansion to higher order than that considered here.

APPENDIX C

Likelihood-ratio statistics

The deviance or deviance difference is just a log-likelihood ratio statistic. In this appendix, we derive the approximate distribution of the log likelihood-ratio statistic for testing a simple null hypothesis concerning a scalar parameter. The corresponding derivations when there are several parameters of interest or when there are nuisance parameters follow similar lines but are considerably more complicated than the proofs presented here.

Suppose that the log likelihood for θ based on data y can be written in the exponential family form

$$l(\theta; y) = n\{t\theta - K(\theta)\},$$

where $t \equiv t(y)$ is the sufficient statistic and θ is called the canonical parameter. The cumulants of the random variable $T = t(Y)$ are

$$\kappa_r(T) = K^{(r)}(\theta)/n^{r-1}.$$

The maximum-likelihood estimate of θ satisfies

$$K'(\hat{\theta}) = t \quad \text{or} \quad \hat{\theta} = g(t), \text{ say.}$$

Hence the log likelihood-ratio statistic for testing $H_0 : \theta = \theta_0$ is

$$W^2 = 2l(\hat{\theta}) - 2l(\theta_0) = 2n\{tg(t) - tg(\mu_0) - h(t) - h(\mu_0)\} \quad (C.1)$$

where $h(\cdot) = K(g(\cdot))$ and $\mu_0 = K'(\theta_0)$ is the null mean of T. Thus $h(t) = K(\hat{\theta})$ and $h(\mu) = K(\theta)$.

Under H_0, the statistic W^2 has an approximate χ^2_1 distribution and hence it is reasonable to expect that the signed version

$$W = \pm\{2l(\hat{\theta}) - 2l(\theta_0)\}^{1/2}$$

might have an approximate Normal distribution. The sign of W is taken to be the same as that of $t - \mu_0$ and, in fact, W is a monotone increasing function of $t - \mu_0$. To the crudest first-order of approximation, W is the standardized version of T, namely

$$X = (T - \mu_0)/\kappa_2^{1/2}.$$

If we expand W as a power series in X and keep together terms that are of the same asymptotic order in n, we find after a little effort that

$$W = X - \tfrac{1}{6}\rho_3 X^2 + \tfrac{1}{72}(8\rho_3^2 - 3\rho_4)X^3 + O_p(n^{-3/2}), \qquad (C.2)$$

where ρ_3 and ρ_4 are the standardized cumulants of T or the unstandardized cumulants of X. Note that $\rho_3 = O(n^{-1/2})$ and $\rho_4 = O(n^{-1})$.

It is readily verified that the first two moments of W are

$$\begin{aligned} E(W) &= -\rho_3/6 + O(n^{-3/2}), \\ \text{var}(W) &= 1 + (14\rho_3^2 - 9\rho_4)/36 + O(n^{-2}). \end{aligned} \qquad (C.3)$$

If we define the adjusted statistic or re-standardized statistic

$$W' = \{W + \tfrac{1}{6}\rho_3\}\{1 + (9\rho_4 - 14\rho_3^2)/72\},$$

it is easily checked that, with error $O_p(n^{-3/2})$,

$$W' = X - \tfrac{1}{6}\rho_3(X^2 - 1) - \tfrac{1}{24}\rho_4(X^3 - 3X) + \tfrac{1}{36}\rho_3^2(4X^3 - 7X).$$

This series can be recognized as the inverse Cornish-Fisher expansion or the polynomial normalizing transformation. See, for example, Kendall and Stuart (1977, (6.54)) or McCullagh (1987, section 5.7 and Exercise 5.15). Provided that X has an Edgeworth expansion, it follows that

$$W' \sim N(0,1) + O(n^{-3/2})$$

in the Edgeworth sense. This conclusion also follows from the observation that W, as given in $(C.2)$, has third and fourth cumulants of orders $O(n^{-3/2})$ and $O(n^{-2})$ respectively.

In the discrete case, the support points of the distribution of T are usually equally spaced and the discrete Edgeworth approximation given in Appendix A may be used to approximate the distribution of T. The support points of W are only approximately equally spaced. It appears therefore, that the normal approximation with continuity correction for the distribution of W' has an error of order $O(n^{-1})$. The Sheppard correction does not appear to eliminate the $O(n^{-1})$ error term entirely, though it may reduce it substantially.

As a corollary in the continuous case, it follows directly that

$$W^2 \sim (1+b)\chi^2_1 + O(n^{-3/2})$$

where $1+b$ is the sum of the variance and the squared bias of W. The correction term

$$b = (5\rho_3^2 - 3\rho_4)/12, \qquad (C.4)$$

which is of order $O(n^{-1})$, is known as the Bartlett adjustment factor. Its use greatly improves the accuracy of the chi-squared approximation for the likelihood-ratio statistic. In both the discrete and the continuous case, the cumulants of $W^2/(1+b)$ differ from those of χ^2_1 by terms of order $O(n^{-2})$.

Although the derivations in general are considerably more complicated than that presented here, these results can be extended in the following ways:

1. to models not in the exponential family provided that the usual regularity conditions are satisfied.
2. to multi-parameter problems.
3. to problems involving nuisance parameters.

Computational details for generalized linear models are discussed in section 15.2. For proofs and additional information, see Lawley (1956), Barndorff-Nielsen and Cox (1984), McCullagh (1984a, 1987 Chapter 7) and McCullagh and Cox (1986).*

* Annals of Statistics vol 14, 1419–30.