Biometrika (1993), **80**, 1, *pp.* 27–38
Printed in Great Britain

Bias reduction of maximum likelihood estimates

By DAVID FIRTH

Department of Mathematics, University of Southampton, SO9 5NH, U.K.

Summary

It is shown how, in regular parametric problems, the first-order term is removed from the asymptotic bias of maximum likelihood estimates by a suitable modification of the score function. In exponential families with canonical parameterization the effect is to penalize the likelihood by the Jeffreys invariant prior. In binomial logistic models, Poisson log linear models and certain other generalized linear models, the Jeffreys prior penalty function can be imposed in standard regression software using a scheme of iterative adjustments to the data.

Some key words: Asymptotic bias; Biased estimating equations; Exponential family; Generalized linear model; Jeffreys prior; Logistic regression; Modified score; Penalized likelihood; Shrinkage.

1. Introduction

In a regular model with a p-dimensional parameter θ the asymptotic bias of the maximum likelihood estimate $\hat{\theta}$ may be written as

$$b(\theta)=\frac{b_1(\theta)}{n}+\frac{b_2(\theta)}{n^2}+\ldots, \tag{1·1}$$

where n is usually interpreted as the number of observations but may be some other measure of the rate at which information accrues. The focus of this paper is a general method for reducing the bias, a specific aim being removal of the $O(n^{-1})$ term.

Two standard approaches have been extensively studied in the literature. The computationally-intensive jackknife method (Quenouille, 1949, 1956) is very general and does not require calculation of $b_1(\theta)$ for its implementation. The other standard approach simply substitutes $\hat{\theta}$ for the unknown θ in $b_1(\theta)/n$; the bias-corrected estimate is then calculated as

$$\hat{\theta}_{\mathrm{BC}}=\hat{\theta}-\frac{b_1(\hat{\theta})}{n}. \tag{1·2}$$

Both of these methods succeed in removing the term $b_1(\theta)/n$ from the asymptotic bias. The jackknife has the advantage of requiring no theoretical calculation, but this is typically offset by a loss of precision. The estimator $\hat{\theta}_{\mathrm{BC}}$ of (1·2) is, quite generally, second-order efficient. See, for example, Cox & Hinkley (1974, §§ 8.4, 9.2) for a discussion of both methods.

A common feature of the two standard approaches is that they are 'corrective', rather than 'preventive' in character. The maximum likelihood estimate $\hat{\theta}$ is first calculated, then corrected. Quite apart from any philosophical considerations or matters of principle that might pertain here, a practical requirement for the application of either method to a finite sample is the existence of $\hat{\theta}$ for that sample, and in the case of the jackknife for certain sub-samples also. In practice, particularly with small or medium-sized sets of data, it is not uncommon that $\hat{\theta}$ is infinite in some samples; linear logistic models for a binary response, for example, are prone to such behaviour, e.g. Albert & Anderson (1984), Clogg et al. (1991). In such cases the jackknife and $\hat{\theta}_{\mathrm{BC}}$ estimators are bias-reducing only in an asymptotic sense.

Motivated partly by this, we explore an approach to bias reduction which does not depend on the finiteness of $\hat{\theta}$. A systematic correction will be developed for the mechanism that produces the maximum likelihood estimate, namely the score equation, rather than for the estimate itself. The modified score function is derived in § 2. Application to exponential families in canonical parameterization is discussed in § 3, where an interesting connection is found with Jeffreys priors. Section 4 discusses application more generally, both to exponential families in other parameterizations and to models outside the exponential family.

There may be connections between the results given here and work on adjusted profile likelihood, e.g. Barndorff-Nielsen (1983), Cox & Reid (1987), but these are not pursued in this paper. Some very recent work in this direction is mentioned in § 5.

2. Modified score function

In regular problems the maximum likelihood estimate is derived as a solution to the score equation

$$\nabla l(\theta) = U(\theta) = 0,$$

where $l(\theta) = \log L(\theta)$ is the log likelihood function. To motivate the general development, consider initially an exponential family model $l(\theta) = t\theta - K(\theta)$ in which θ is scalar. Then

$$U(\theta) = l'(\theta) = t - K'(\theta),$$

so that the sufficient statistic t affects only the location of $U(\theta)$, not its shape. Now bias in $\hat{\theta}$ arises from the combination of (i) unbiasedness of the score function, $E\{U(\theta)\} = 0$ at the true value of θ, and (ii) curvature of the score function, $U''(\theta) \neq 0$. Clearly if $U(\theta)$ is linear in θ then $E(\hat{\theta}) = \theta$; but positive curvature as shown in Fig. 1, for example, combines with the unbiasedness of the score function to induce a bias in $\hat{\theta}$, in this case in the positive direction.

The basis of the present work is the idea that the bias in $\hat{\theta}$ can be reduced by introducing a small bias into the score function. The appropriate modification to $U(\theta)$ is given by

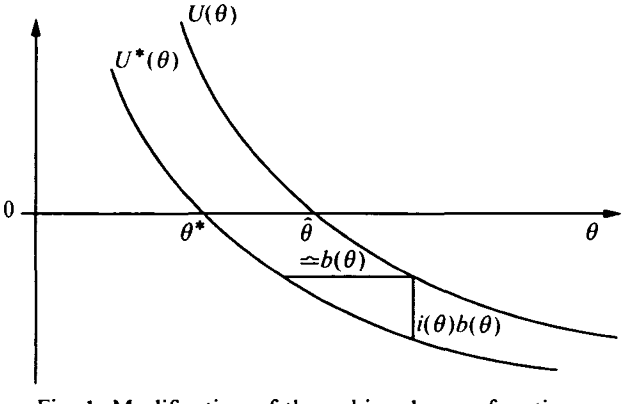

Fig. 1. Modification of the unbiased score function.

simple triangle geometry, illustrated in Fig. 1. If $\hat{\theta}$ is subject to a positive bias $b(\theta)$, the score function is shifted downward at each point θ by an amount $i(\theta)b(\theta)$, where $-i(\theta) = U'(\theta)$ is the local gradient; this defines a modified score function

$$U^*(\theta) = U(\theta) - i(\theta)b(\theta) \tag{2·1}$$

and hence a modified estimate θ^*, given as a solution to $U^*(\theta) = 0$. In the case of a vector parameter, (2·1) should be read as a vector equation, in which $i(\theta)$ is the Fisher information matrix.

To formalize the heuristic argument above and extend it to problems other than canonical exponential families, it will be convenient to employ the notation and methods of McCullagh (1987, § 7·3) for log likelihood derivatives and their null cumulants. The derivatives are denoted by

$$U_r(\theta) = \partial l/\partial\theta^r, \quad U_{rs}(\theta) = \partial^2 l/\partial\theta^r\partial\theta^s,$$

and so on, where $\theta = (\theta^1, \ldots, \theta^p)$ is the parameter vector. The joint null cumulants are

$$\kappa_{r,s} = n^{-1}E\{U_rU_s\}, \quad \kappa_{r,s,t} = n^{-1}E\{U_rU_sU_t\}, \quad \kappa_{r,st} = n^{-1}E\{U_rU_{st}\},$$

and so on. We note here the well-known relationships

$$\kappa_{rs} + \kappa_{r,s} = 0, \quad \kappa_{rst} + \kappa_{r,st} + \kappa_{s,rt} + \kappa_{t,rs} + \kappa_{r,s,t} = 0. \tag{2·2}$$

Consider now a fairly general modification of the score function, of the form

$$U_r^*(\theta) = U_r(\theta) + A_r(\theta),$$

in which A_r is allowed to depend on the data and is $O_p(1)$ as $n \to \infty$. Suppose that $\hat{\theta}$ and θ^* satisfy $U(\hat{\theta}) = 0$ and $U^*(\theta^*) = 0$, and write $\hat{\gamma} = n^{\frac{1}{2}}(\theta^* - \theta)$. Then, by an argument closely following that of McCullagh (1987, p. 209), based on an expansion of $U_r^*(\theta^*)$ about the true value θ, the bias of θ^* is

$$E(n^{-\frac{1}{2}}\hat{\gamma}^r) = n^{-1}\kappa^{r,s}\{-\kappa^{t,u}(\kappa_{s,t,u} + \kappa_{s,tu})/2 + \alpha_s\} + O(n^{-3/2}),$$

where $\kappa^{r,s}$ denotes the inverse of the Fisher information matrix $\kappa_{r,s}$, α_s denotes the null expectation of A_s, and the summation convention applies. The term

$$-n^{-1}\kappa^{r,s}\kappa^{t,u}(\kappa_{s,t,u} + \kappa_{s,tu})/2 = n^{-1}b_1^r(\theta)$$

is the first-order bias of $\hat{\theta}$, for example Cox & Snell (1968). The modification A_r therefore removes the first-order term if it satisfies

$$\kappa^{r,s}\alpha_s = -b_1^r + O(n^{-\frac{1}{2}}),$$

the solution to which is

$$\alpha_r = -\kappa_{r,s}b_1^s + O(n^{-\frac{1}{2}}).$$

In matrix notation, then, the vector A should be such that

$$E(A) = -i(\theta)b_1(\theta)/n + O(n^{-\frac{1}{2}}).$$

Obvious candidates for a bias-reducing choice of A are therefore $A^{(E)} = -i(\theta)b_1(\theta)/n$ and $A^{(O)} = -I(\theta)b_1(\theta)/n$, using expected and observed information respectively. In the case of an exponential family in canonical parameterization the observed information $I(\theta)$ does not involve the data, so $A^{(O)}$ and $A^{(E)}$ coincide. More generally, either of these modifications removes the $O(n^{-1})$ bias term. The difference between $A^{(O)}$ and $A^{(E)}$, in terms of second-order efficiency, is discussed briefly in § 4.

3. Exponential families

3·1. *Jeffreys prior as bias-reducing penalty function*

If θ is the canonical parameter of an exponential family model, $\kappa_{r,st}=0$ for all r, s and t. Therefore the rth element of $A^{(E)}(\theta)$, or equivalently of $A^{(O)}(\theta)$, is given by

$$a_r = -n\kappa_{r,s}b_1^s/n = \kappa_{r,s}\kappa^{s,t}\kappa^{u,v}\kappa_{t,u,v}/2 = \kappa^{u,v}\kappa_{r,u,v}/2 = -\kappa^{u,v}\kappa_{ruv}/2,$$

using the identities (2·2). In matrix notation, this may be written as

$$a_r = \frac{1}{2}\operatorname{tr}\left\{i^{-1}\left(\frac{\partial i}{\partial\theta_r}\right)\right\} = \frac{\partial}{\partial\theta_r}\left\{\frac{1}{2}\log|i(\theta)|\right\}.$$

Solution of $U_r^* \equiv U_r + a_r = 0$ therefore locates a stationary point of

$$l^*(\theta) = l(\theta) + \tfrac{1}{2}\log|i(\theta)|$$

or, equivalently, of the penalized likelihood function

$$L^*(\theta) = L(\theta)|i(\theta)|^{\frac{1}{2}}.$$

The penalty function $|i(\theta)|^{\frac{1}{2}}$ here is the Jeffreys (1946) invariant prior for the problem. The arguments of § 2 show that, for the canonical parameter of an exponential family model, the $O(n^{-1})$ bias is removed by calculation of the posterior mode based on this prior.

3·2. *Example: Normal distribution*

Suppose that $y_1, \ldots, y_n$ is a random sample from the normal distribution with mean μ and variance σ^2. The canonical parameterization is $\theta = \{\mu/\sigma^2, -1/(2\sigma^2)\}$, with information matrix

$$i(\theta) = \begin{pmatrix} \sigma^2 & 2\mu\sigma^2 \\ 2\mu\sigma^2 & 4\mu^2\sigma^2+2\sigma^4 \end{pmatrix}.$$

The bias-reducing penalty function is therefore $|i(\theta)|^{\frac{1}{2}} \propto \sigma^3$, which yields estimates

$$\theta^* = \left\{\frac{(n-3)\bar{y}}{s(\bar{y})},\ -\frac{n-3}{2s(\bar{y})}\right\},$$

where $\bar{y} = n^{-1}\Sigma\, y_i$ and $s(\mu) = \Sigma\,(y_i-\mu)^2$. In this example, θ^* is exactly unbiased for $n>3$.

3·3. *Example: Binomial logistic regression*

The calculation and correction of bias in the maximum likelihood estimates of logistic regression parameters have been studied by many authors, including Anderson & Richardson (1979), McLachlan (1980), Schaefer (1983), Copas (1988), McCullagh & Nelder (1989, § 15.2) and Cordeiro & McCullagh (1991). If the success probability for the ith observation is $\pi_i = \exp(\eta_i)/\{1+\exp(\eta_i)\}$, where $\eta_i = \Sigma\, x_{ir}\beta_r$, maximum likelihood estimates of the canonical parameter β are found to be biased away from the point $\beta=0$. Bias correction, therefore, requires some degree of 'shrinkage' of $\hat{\beta}$ towards this point.

In logistic regression the information matrix is $i(\beta) = I(\beta) = X^{\mathrm{T}}WX$, where $X=\{x_{ir}\}$ is the design matrix, $W = \operatorname{diag}\{m_i\pi_i(1-\pi_i)\}$ and m_i is the binomial index for the ith count. The determinant is maximized at $\pi_i = \frac{1}{2}$ $(i=1,\ldots,n)$, that is at $\beta=0$, so the Jeffreys prior shrinks estimates towards this point. The arguments of § 2 show that the amount of shrinkage is that needed to remove the $O(n^{-1})$ bias. It may also be shown that, provided X is of full rank, $\log|i(\beta)|$ is strictly concave and unbounded below as $\beta \to \infty$ in any direction. This, combined with the fact that $l(\beta)$ itself is strictly concave and bounded above, ensures that the maximum penalized likelihood estimate β^* exists and is unique.

As an illustration of this shrinkage, consider a variant of a one-parameter example used by Copas (1988). Suppose that $\eta_i = x_i\beta$, a regression through the origin on the logistic scale, and let x_i take values in $\{-2, -1, 0, 1, 2\}$. Copas (1988) examines $b_1(\beta)/n$ when 10 binary observations are taken at each of these five points and finds that, for small β, the asymptotic bias away from zero is about 3·4% of the true value. For our illustrative purposes, we take a much more severe case in which only one binary observation is made at each of the five design points, so that complete enumeration is feasible. The sufficient statistic, $t = \Sigma\, y_i x_i$, has only seven possible values; note, incidentally, that the observation at $x_i = 0$ contributes nothing to t, so is redundant in this example. Table 1 gives $\hat\beta$, $\hat\beta_{\text{BC}}$ and β^* corresponding to the seven values of t, and also the sampling distribution at two particular values of β. The mean of β^* is found to be 0·46 when $\beta = 0{\cdot}5$, and 0·82 when the true value is 1·0. This seems satisfactory given the very small sample size and the large probability that $\hat\beta$ is infinite. Note that $\beta = 1$ is a fairly large slope in this context, implying a range of response probabilities from 0·12 to 0·88 at the five design points.

Table 1. *Distribution of estimators in a small logistic regression model*

				Sampling probabilities	
$t(y)$	$\hat\beta$	$\hat\beta_{\text{BC}}$	β^*	$\beta = 0{\cdot}5$	$\beta = 1$
−3	$-\infty$	—	−1·38	0·010	0·001
−2	−1·01	−0·52	−0·68	0·034	0·006
−1	−0·42	−0·27	−0·31	0·084	0·023
0	0	0	0	0·185	0·083
1	0·42	0·27	0·31	0·229	0·168
2	1·01	0·52	0·68	0·251	0·305
3	∞	—	1·38	0·207	0·415

The simplest of all logistic models is that for a single binomial observation, the target parameter being $\beta = \log\{\pi/(1-\pi)\}$. The information is proportional to $\pi(1-\pi)$, so that the penalized likelihood is simply

$$L^* = \pi^{y+\frac{1}{2}}(1-\pi)^{m-y+\frac{1}{2}}.$$

Maximization of L^* yields

$$\beta^* = \log\left(\frac{y+\frac{1}{2}}{m-y+\frac{1}{2}}\right),$$

which is familiar as the bias-reducing form of the empirical logit (Haldane, 1955; Anscombe, 1956; Cox & Snell, 1989, § 2.1.6). For this single-sample model β^* is the maximum likelihood estimate calculated from adjusted data formed by adding $\frac{1}{2}$ to y and 1 to m. The same adjustment does not produce β^* in the case of a general design matrix X, but is a special instance of a more general adjustment procedure that may be used for calculation of β^*.

The $O(n^{-1})$ bias vector in general is given by McCullagh & Nelder (1989, § 15.2) in the form

$$b_1/n = (X^T WX)^{-1} X^T W\xi, \tag{3·1}$$

where $W\xi$ has ith element $h_i(\pi_i - \frac{1}{2})$ and h_i is the ith diagonal element of the 'hat' matrix

$$H = W^{\frac{1}{2}} X (X^T WX)^{-1} X^T W^{\frac{1}{2}}.$$

Hence $U^* = U - X^T W\xi$, with rth component

$$U_r^* = \sum_i \{(y_i + h_i/2) - (m_i + h_i)\pi_i\} x_{ir} \quad (r = 1, \ldots, p). \tag{3·2}$$

Solution of $U^* = 0$ is therefore equivalent to solution of maximum likelihood equations based on adjusted data formed by adding $h_i(\beta^*)/2$ to y_i and $h_i(\beta^*)$ to m_i. This suggests an iterative algorithm in which the adjustments $\{h_i\}$ are updated at each cycle of a standard iteratively weighted least-squares procedure. Calculation of β^* is thus made possible in any regression software that allows both weighting of observations and access to the leverage quantities $\{h_i\}$. Implementation and properties of the algorithm are discussed by Firth (1992a, b).

Rubin & Schenker (1987) and Clogg et al. (1991) suggest adjustments that are in a similar spirit, but different from the ones just described. Their adjustment is noniterative, but does not take account of differences among the leverages $\{h_i\}$, and shrinks estimates towards the mean rather than toward zero; it does not remove the $O(n^{-1})$ bias. A comparative study would be useful.

3·4. *Other generalized linear models*

The argument given above for logistic regression may be extended to any generalized linear model, with canonical link or otherwise, using formula (15.4) of McCullagh & Nelder (1989) for the vector ξ in (3·1). In the canonical link case the general form of the modified score function, of which (3·2) above is a special case, is

$$U_r^* = U_r + \frac{1}{2\phi} \sum_i \left(\frac{\kappa_{3i}}{\kappa_{2i}}\right) h_i x_{ir} \quad (r = 1, \ldots, p),$$

where κ_{ti} is the tth cumulant of y_i and ϕ is the usual dispersion parameter. Thus, for example, in a Poisson log linear model where $\kappa_{ti} = \mu_i$ $(t = 1, 2, \ldots)$,

$$U_r^* = \sum_i \{(y_i + h_i/2) - \mu_i\} x_{ir} \quad (r = 1, \ldots, p),$$

and β^* can be computed by iteratively making adjustments of $\{h_i/2\}$ to the counts $\{y_i\}$. In the case of the normal distribution, $\kappa_{3i} = 0$: maximum likelihood estimates in normal linear regression are unbiased, so no adjustment is made.

4. OTHER MODELS

4·1. *Modified score function*

We now discuss application of the results of § 2 in a more general setting that includes exponential family models in noncanonical parameterization, as well as nonexponential models.

The modified score function in this general setting has the form $U_r^* = U_r + A_r$, where $A_r(\theta)$ is based either on the expected information,

$$A_r = A_r^{(E)} = n\kappa_{r,s}\kappa^{s,t}\kappa^{u,v}(\kappa_{t,u,v} + \kappa_{t,uv})/(2n)$$

$$= \kappa^{u,v}(\kappa_{r,u,v} + \kappa_{r,uv})/2, \qquad (4\cdot1)$$

or on the observed information,

$$A_r = A_r^{(O)} = -U_{rs}\kappa^{s,t}\kappa^{u,v}(\kappa_{t,u,v} + \kappa_{t,uv})/(2n).$$

Intuition suggests that estimates derived using $A_r^{(O)}$ may be preferable in terms of efficiency. To explore this further, consider an expansion of $U^*(\theta^*)$ about $\hat{\theta}$. By definition,

$$0 = U_r^*(\theta^*) = U_r(\theta^*) + A_r(\theta^*).$$

If $A_r(\theta) = A_r^{(O)}(\theta) = U_{rs}(\theta)b_1^s(\theta)/n$, we have that

$$(\theta^* - \hat{\theta})^r = -b_1^r(\hat{\theta})/n + O_p(n^{-2}), \qquad (4\cdot2)$$

while, if $A_r(\theta) = A_r^{(E)}(\theta) = -i_{rs}(\theta)b_1^s(\theta)/n$,

$$(\theta^* - \hat{\theta})^r = -b_1^r(\hat{\theta})/n - i^{rs}(\hat{\theta})\{U_{st}(\hat{\theta}) + i_{st}(\hat{\theta})\}b_1^t(\hat{\theta})/n + O_p(n^{-2}). \qquad (4\cdot3)$$

The difference $U_{st}(\hat{\theta}) + i_{st}(\hat{\theta})$ between expected and observed information at the maximum likelihood estimate is $O_p(n^{-\frac{1}{2}})$ in general, e.g. Pierce (1975), so that the extra term in (4·3) is $O_p(n^{-3/2})$. In the special case of a full exponential family model, with any parameterization, this term vanishes.

From (4·2) it may be concluded that if U^* is calculated using the observed information function, θ^* agrees with $\hat{\theta}_{\text{BC}}$ to second order. This is not the case if expected information is used, unless the model is a full exponential family. Thus both forms of U^* yield estimators that are first-order efficient, and the results of Efron (1975) show that both forms are second-order efficient in full exponential family models. In curved exponential families and more generally, use of the modification $A^{(E)}$ involves a second-order loss of precision relative to use of $A^{(O)}$.

There follow some simple examples to illustrate the approach.

4·2. *Example: Normal distribution*

Here we re-consider the example of § 3·2 in the more familiar (μ, σ^2) parameterization. For convenience, denote σ^2 by ϕ. The score vector has components

$$U_\mu = (y_. - n\mu)/\phi, \quad U_\phi = s(\mu)/(2\phi^2) - n/(2\phi),$$

so the observed and expected information matrices are

$$I = \begin{pmatrix} n/\phi & (y_. - n\mu)/\phi^2 \\ (y_. - n\mu)/\phi^2 & s(\mu)/\phi^3 - n/(2\phi^2) \end{pmatrix}, \quad i = \begin{pmatrix} n/\phi & 0 \\ 0 & n/(2\phi^2) \end{pmatrix}.$$

The remaining quantities required for calculation of U^* are

$$\kappa_{\mu,\mu,\phi} = 1/(\phi^2), \quad \kappa_{\phi,\phi,\phi} = 1/\phi^3, \quad \kappa_{\mu,\mu\phi} = -1/(\phi^2), \quad \kappa_{\phi,\phi\phi} = -1/\phi^3,$$

$$\kappa_{\mu,\mu,\mu} = \kappa_{\mu,\phi,\phi} = \kappa_{\mu,\mu\mu} = \kappa_{\phi,\mu\mu} = \kappa_{\mu,\phi\phi} = \kappa_{\phi,\phi\mu} = 0.$$

The two alternative modifications to $U(\mu, \phi)$ are calculated as

$$A_\mu^{(E)} = 0, \quad A_\phi^{(E)} = 1/(2\phi),$$

$$A_\mu^{(O)} = 0, \quad A_\phi^{(O)} = s(\mu)/(n\phi^2) - 1/(2\phi).$$

Solution of $U+A^{(E)}=0$ gives $\phi^*=s(\bar{y})/(n-1)$, while the alternative equation $U+A^{(O)}=0$ yields $\phi^*=(n+2)s(\bar{y})/\{n(n+1)\}$. The first of these estimators is, of course, exactly unbiased. In accord with the arguments of § 4·1, both estimators are second-order efficient, with variance $2\phi^2(n+1)/n^2+O(n^{-3})$.

4·3. *Example: Reciprocal mean of a Poisson distribution*

Suppose that $y_1, \ldots, y_n$ are drawn independently from the Poisson distribution with mean μ, and interest is in $\phi=1/\mu$. This could arise, for example, in connection with analysis of count data from a Poisson process, where ϕ is the mean inter-event time.

In this problem we have

$$U=n/\phi^2-y_./\phi, \quad I=2n/\phi^3-y_./\phi^2, \quad i=n/\phi^3,$$

and the maximum likelihood estimate is $\hat{\phi}=1/\bar{y}$ with asymptotic bias $\phi^2/n+O(n^{-2})$. The two alternative modifications to the score function are calculated as

$$A^{(E)}(\phi)=-1/\phi, \quad A^{(O)}(\phi)=\bar{y}-2/\phi.$$

The first of these yields

$$\phi^*=\frac{1}{\bar{y}+1/n},$$

while the second gives

$$\phi^*=\begin{cases} n\{\bar{y}+2/n-\surd(\bar{y}^2+4/n^2)\}/(2\bar{y}) & (\bar{y}>0), \\ n/2 & (\bar{y}=0). \end{cases}$$

Both of these estimators are finite for all samples, have bias that is $O(n^{-2})$, and are second-order efficient with variance $1/(n\mu^3)+2/(n^2\mu^4)+O(n^{-3})$.

4·4. *Example: Normal distribution with known coefficient of variation*

The $N(\mu, c\mu^2)$ distribution, in which c is known, is perhaps the most tractable instance of a curved exponential family, and has been studied by Efron (1975) and Hinkley (1977) among others. For simplicity, consider the case $c=1$. The score function for μ from a random sample $y_1, \ldots, y_n$ is

$$U(\mu)=\frac{\sum y^2}{\mu^3}-\frac{\sum y}{\mu^2}-\frac{n}{\mu},$$

so that the observed information is

$$I(\mu)=\frac{3\sum y^2}{\mu^4}-\frac{2\sum y}{\mu^3}-\frac{n}{\mu^2},$$

with expectation $i(\mu)=3n/\mu^2$. The asymptotic bias calculation in this case yields $b_1(\mu)=-2\mu/9$. Bias-reducing modifications to $U(\mu)$ are therefore

$$A^{(E)}(\mu)=2/(3\mu),$$

$$A^{(O)}(\mu)=(9n^{-1})\left(\frac{6\sum y^2}{\mu^3}-\frac{4\sum y}{\mu^2}-\frac{2n}{\mu}\right).$$

Use of $A^{(E)}$ corresponds to using $(n-\frac{2}{3})$ in place of n; while from (4·2), use of $A^{(O)}$ is approximately equivalent to multiplication of $\hat{\mu}$ by $1+2/(9n)$.

Since this is not a full exponential family model, the discussion of § 4·1 indicates that use of $A^{(E)}$ will not be second-order efficient. Straightforward but tedious calculation yields that

$$\operatorname{var}(\hat{\mu})=\mu^2\left\{\frac{1}{3n}-\frac{2}{81n^2}+O(n^{-3})\right\},$$

and that use of $A^{(O)}$ to reduce bias adds $12\mu^2/(81n^2)$ to the variance, while use of $A^{(E)}$ adds $36\mu^2/(81n^2)$. In this example, bias reduction is variance-inflating, and $A^{(E)}$ inflates the variance by approximately three times as much as does $A^{(O)}$.

4·5. *Example: Precision of duplicate measurements*

This is a severe case of an example discussed by Neyman & Scott (1948), in which the number of parameters increases with n and the maximum likelihood estimate is inconsistent. Suppose that $\{y_{jk}: k=1,\ldots,K; j=1,2\}$ are drawn independently from normal distributions with $E(y_{jk})=\mu_k$ and $\operatorname{var}(y_{jk})=\sigma^2$. Here $\phi=\sigma^2$ is of interest and $\mu_1,\ldots,\mu_K$ are incidental parameters. The maximum likelihood estimate of ϕ is

$$\hat{\phi}=\frac{1}{2K}\sum_k \sum_j (y_{jk}-\bar{y}_k)^2,$$

which has expectation $\phi/2$ for all values of K and probability limit $\phi/2$ as $K\rightarrow\infty$.

We now calculate modification (4·1) to the score function for this problem. Using index k to stand for μ_k, we have that

$$U_\phi=-\frac{K}{\phi}+\frac{1}{2\phi^2}\sum\sum(y_{jk}-\mu_k)^2,\quad U_k=(y_{.k}-2\mu_k)/\phi,$$

$$\kappa^{\phi,\phi}=2\phi^2,\quad \kappa^{k,k}=K\phi\quad(k=1,\ldots,K),$$

and all other $\kappa^{r,s}$ are zero. It may be verified that formula (4·1) now yields

$$A_\phi^{(E)}=K/(2\phi),\quad A_k^{(E)}=0\quad(k=1,\ldots,K),$$

so that $U_k^*=U_k$ for all k, and

$$U_\phi^*=-\frac{K}{2\phi}+\frac{1}{2\phi^2}\sum\sum(y_{jk}-\mu_k)^2.$$

The resultant estimate ϕ^* is $2\hat{\phi}$, which is exactly unbiased and consistent. Note that the 'profile' modified score function,

$$U_\phi^*(\phi,\mu_\phi^*)=-\frac{K}{2\phi}+\frac{1}{2\phi^2}\sum\sum(y_{jk}-\bar{y}_k)^2,$$

in this case is the same as the marginal score function based on the statistic $\Sigma\,\Sigma\,(y_{jk}-\bar{y}_k)^2$.

The use of $A^{(O)}$ in place of $A^{(E)}$ here also yields a consistent, though not unbiased, estimate of ϕ.

This problem may alternatively be considered in the canonical parameterization, (θ, λ) say, with

$$\theta = -1/(2\sigma^2), \quad \lambda_k = \mu_k/\sigma^2 \quad (k = 1, \ldots, K).$$

The Jeffreys prior is found to be proportional to σ^{K+2}, so that

$$\theta^* = -(K-2)/\{2\sum\sum(y_{jk} - \bar{y}_k)^2\},$$

which again is consistent for θ, and exactly unbiased if $K > 2$.

5. DISCUSSION

It has been shown how, in regular problems, the $O(n^{-1})$ bias may be removed from the maximum likelihood estimator by introduction of an appropriate bias term into the score function. If the target parameter is the canonical parameter of an exponential family, the method simply penalizes the likelihood by the Jeffreys invariant prior. For other parameterizations of exponential family models, and for nonexponential families, a choice is available between corrections using observed and expected information. Outside exponential family models, use of the expected information results in a loss of second-order efficiency.

It is not an assumption of this work that bias reduction is always desirable. The merits of bias reduction in any particular problem will depend on a number of factors, including the skewness of the maximum likelihood estimator and any sacrifice in precision that might result; in the $N(\mu, \mu^2)$ problem of § 4·4, for example, it was found that bias reduction inflates the asymptotic variance by, at best, $12\mu^2/(81n^2) + O(n^{-3})$, which is approximately three times the reduction in squared bias. The choice of parameter is crucial in this respect. In binomial logistic regression, for example, reparameterization to the mean-value parameters $\{\tau_r = \Sigma\, m_i\pi_i x_{ir}\colon r = 1, \ldots, p\}$ yields a maximum likelihood estimator that is unbiased without any correction. However, the distribution of $\hat{\tau}$ is typically far from normal, so unbiasedness on the τ scale is not necessarily of great value. The distribution of the regression coefficients $\hat{\beta}$ is usually closer to normality, and moreover the 'shrinkage' effect of bias reduction brings with it a reduction in variance (Copas, 1988). In logistic regression, then, bias reduction on the scale of the canonical parameters seems desirable.

Nothing has yet been said in this paper about standard errors and confidence regions based on bias-reduced estimates. The first order asymptotic covariance matrix of θ^* is the same as that of $\hat{\theta}$, namely $i^{-1}(\theta)$, and this can be used in the usual way for providing standard errors. A study of the second-order term and its implications for inference would be useful.

In the examples of §§ 3·3, 4·3, θ^* was found to exist in any finite sample, whereas $\hat{\theta}$ has positive probability of being infinite. It is not known in precisely what range of problems θ^* can be guaranteed finite, and a systematic study of this aspect would be valuable.

A special role for the Jeffreys prior has been indicated previously by Welch & Peers (1963) and Hartigan (1965), though from rather different perspectives. Use of the Jeffreys prior as a bias-reducing penalty function in exponential family problems, as in the present paper, is discussed also in a 1991 Habilitationsschrift, 'Statistical problems with many parameters: critical quantities for approximate normality and posterior density based inference', by Dr W. Ehm at the University of Heidelberg.

In the example of § 4·5, with an increasing number of parameters, the standard assumptions underlying the asymptotic development in § 2 fail to hold, so the success of the bias-reducing modification to U is somewhat surprising. The example is, however, very special. In other problems of this type, such as the binary matched pairs problem (Breslow, 1981), it is found that the $O(1)$ bias is greatly reduced, but not completely eliminated, by adjustment of the score function as described in § 2; details are in an unpublished technical report available from the author. An extensive discussion of connections between bias reduction and approximate conditional inference in problems with many parameters is given in the aforementioned report by W. Ehm.

Acknowledgements

The author is grateful to Susan Lewis and Mary Tuck for providing the experimental data that motivated this work, to Trevor Sweeting for perceptive comments on an early version, and to the Associate Editor and two referees for suggestions that have improved the presentation.

References

Albert, A. & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**, 1-10.

Anderson, J. A. & Richardson, S. C. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics* **21**, 71-8.

Anscombe, F. J. (1956). On estimating binomial response relations. *Biometrika* **43**, 461-4.

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343-65.

Breslow, N. E. (1981). Odds ratio estimators when the data are sparse. *Biometrika* **68**, 73-84.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B. & Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *J. Am. Statist. Assoc.* **86**, 68-78.

Copas, J. B. (1988). Binary regression models for contaminated data. *J. R. Statist. Soc.* B **50**, 225-65.

Cordeiro, G. M. & McCullagh, P. (1991). Bias correction in generalized linear models. *J. R. Statist. Soc.* B **53**, 629-43.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

Cox, D. R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. R. Statist. Soc.* B **49**, 1-39.

Cox, D. R. & Snell, E. J. (1968). A general definition of residuals. *J. R. Statist. Soc.* B **30**, 248-75.

Cox, D. R. & Snell, E. J. (1989). *Analysis of Binary Data*, 2nd ed. London: Chapman and Hall.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* **3**, 1189-242.

Firth, D. (1992a). Bias reduction, the Jeffreys prior and GLIM. In *Advances in GLIM and Statistical Modelling*, Ed. L. Fahrmeir, B. Francis, R. Gilchrist and G. Tutz, pp. 91-100. New York: Springer-Verlag.

Firth, D. (1992b). Generalized linear models and Jeffreys priors: an iterative weighted least-squares approach. In *Computational Statistics*, **1**, Ed. Y. Dodge and J. Whittaker, pp. 553-7. Heidelberg: Physica-Verlag.

Haldane, J. B. S. (1955). The estimation and significance of the logarithm of a ratio of frequencies. *Ann. Hum. Gen.* **20**, 309-11.

Hartigan, J. A. (1965). The asymptotically unbiased prior distribution. *Ann. Math. Statist.* **36**, 1137-52.

Hinkley, D. V. (1977). Conditional inference about a normal mean with known coefficient of variation. *Biometrika* **64**, 105-8.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc.* A **186**, 453-61.

McCullagh, P. (1987). *Tensor Methods in Statistics.* London: Chapman and Hall.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

McLachlan, G. J. (1980). A note on bias correction in maximum likelihood estimation with logistic discrimination. *Technometrics* **22**, 621-7.

Neyman, J. & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1-32.

PIERCE, D. A. (1975). Discussion of paper by B. Efron. *Ann. Statist.* **3**, 1219-21.

QUENOUILLE, M. H. (1949). Approximate tests of correlation in time-series. *J. R. Statist. Soc.* B **11**, 68-84.

QUENOUILLE, M. H. (1956). Notes on bias in estimation. *Biometrika* **43**, 353-60.

RUBIN, D. B. & SCHENKER, N. (1987). Logit-based interval estimation for binomial data using the Jeffreys prior. In *Sociological Methodology 1987*, Ed. C. C. Clogg, pp. 131-44. Washington, DC: American Sociological Association.

SCHAEFER, R. L. (1983). Bias correction in maximum likelihood logistic regression. *Statist. Med.* **2**, 71-8.

WELCH, B. L. & PEERS, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. R. Statist. Soc.* B **25**, 318-29.

[*Received September* 1991. *Revised September* 1992]