OXFORD JOURNALS
OXFORD UNIVERSITY PRESS

Bias reduction in exponential family nonlinear models

Author(s): IOANNIS KOSMIDIS and DAVID FIRTH

Source: *Biometrika*, DECEMBER 2009, Vol. 96, No. 4 (DECEMBER 2009), pp. 793-804

Published by: Oxford University Press on behalf of Biometrika Trust

Stable URL: https://www.jstor.org/stable/27798867

REFERENCES
Linked references are available on JSTOR for this article:
https://www.jstor.org/stable/27798867?seq=1&cid=pdf-reference#references_tab_contents
You may need to log in to JSTOR to access the linked references.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

and *Oxford University Press* are collaborating with JSTOR to digitize, preserve and extend access to *Biometrika*

JSTOR

This content downloaded from
130.126.162.126 on Sat, 28 Nov 2020 04:47:10 UTC
All use subject to https://about.jstor.org/terms

Biometrika (2009), **96**, 4, *pp.* 793–804
© 2009 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/asp055
Advance Access publication 9 November 2009

Bias reduction in exponential family nonlinear models

BY IOANNIS KOSMIDIS AND DAVID FIRTH

Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.
i.kosmidis@warwick.ac.uk d.firth@warwick.ac.uk

SUMMARY

In Firth (1993, *Biometrika*) it was shown how the leading term in the asymptotic bias of the maximum likelihood estimator is removed by adjusting the score vector, and that in canonical-link generalized linear models the method is equivalent to maximizing a penalized likelihood that is easily implemented via iterative adjustment of the data. Here a more general family of bias-reducing adjustments is developed for a broad class of univariate and multivariate generalized nonlinear models. The resulting formulae for the adjusted score vector are computationally convenient, and in univariate models they directly suggest implementation through an iterative scheme of data adjustment. For generalized linear models a necessary and sufficient condition is given for the existence of a penalized likelihood interpretation of the method. An illustrative application to the Goodman row-column association model shows how the computational simplicity and statistical benefits of bias reduction extend beyond generalized linear models.

Some key words: Asymptotic bias correction; Generalized nonlinear model; Multivariate generalized linear model; Penalized likelihood; Pseudo-data.

1. INTRODUCTION

In regular parametric statistical models the maximum likelihood estimator is consistent, and the leading term in its asymptotic bias expansion is of magnitude $O(n^{-1})$; here n denotes the sample size or other index of information, assumed to be large relative to the number, p say, of parameters. Among methods that have been suggested for removal of the $O(n^{-1})$ bias, the approach taken by Firth (1993) has received considerable recent attention. That method has the advantage of not requiring the value of the maximum likelihood estimate itself. This last fact has, at least in part, motivated detailed empirical studies of the method in some common situations where the maximum likelihood estimate can be infinite, notably logistic or similar regression models for a binary or multinomial response, and models for censored lifetime data; see, for example, Mehrabi & Matthews (1995), Pettitt et al. (1998), Heinze & Schemper (2002), Bull et al. (2002, 2007), Heinze & Schemper (2004), Zorn (2005) and Sartori (2006).

Point estimation and unbiasedness are, of course, not strong statistical principles. The notion of bias, in particular, relates to a specific parameterization of a model: for example, the unbiasedness of the familiar sample variance S^2 as an estimator of σ^2 does not deliver an unbiased estimator of σ itself. Thus, the bias correction of the maximum likelihood estimator, by any method, violates the appealing property of exact equivariance under reparameterization of the model. Moreover, as noted by Firth (1993) and well known previously, the reduction in bias may sometimes be accompanied by inflation of variance, possibly yielding an estimator whose mean squared error is worse than that of the original one. Despite these reservations, it is amply evident from published empirical studies such as those mentioned above that, in some very commonly

This content downloaded from
130.126.162.126 on Sat, 28 Nov 2020 04:47:10 UTC
All use subject to https://about.jstor.org/terms

used models, bias reduction by the method studied here can perform substantially better than unadjusted maximum likelihood, especially when the sample size is not large. Importantly, better here means not only in terms of bias, but also in terms of other properties such as finiteness, mean squared error and the coverage of approximate confidence intervals. The remarks above make plain that such improvements cannot be universal. A primary motivation for the present paper is to provide a unified conceptual and computational framework for further exploration of the properties of bias-reduced maximum likelihood in a large class of models that includes many model types commonly used in practice. The class of multivariate generalized nonlinear models includes univariate Gaussian and generalized nonlinear regressions, and univariate and multivariate generalized linear models, as prominent special cases.

2. BIAS REDUCTION BY ADJUSTMENT OF THE SCORE

2·1. *Preamble and notation*

In regular problems, maximum likelihood estimates are obtained by solving the score equation $U(\beta) = \nabla l(\beta) = 0$, where $l(\beta)$ is the loglikelihood function for the parameter vector β. Under regularity conditions in the spirit of those in Cramér (1946, § 33.2), which guarantee the consistency of the maximum likelihood estimator, Firth (1993) showed how, by solving a suitably adjusted version of the score equation, removal of the $O(n^{-1})$ bias term of the maximum likelihood estimator is achieved. In Firth (1993) the form of the adjusted score vector is given in index notation and using the Einstein summation convention. Despite the general elegance and compactness of such notation, it will be departed from here, since matrix notation will be simpler and more illuminating for the present purposes. First, some notational rules are introduced for the representation of arrays as blocked matrices.

Consider a sequence of four-way arrays $\{E_r;\ r = 1, \ldots, k\}$. Such an array E_r is an arrangement of scalars e_{rstuv} with $s \in \{1, \ldots, l\}$, $t \in \{1, \ldots, m\}$, $u \in \{1, \ldots, n\}$ and $v \in \{1, \ldots, q\}$. The array E_r can be represented as an $ln \times mq$ blocked matrix having the form

$$E_r = \begin{pmatrix} E_{r11} & E_{r12} & \cdots & E_{r1m} \\ E_{r21} & E_{r22} & \cdots & E_{r2m} \\ \vdots & \vdots & \ddots & \vdots \\ E_{rl1} & E_{rl2} & \cdots & E_{rlm} \end{pmatrix},$$

with E_{rst} an $n \times q$ matrix. Denote by E_{rstu} the uth row of the matrix E_{rst} as a row vector, i.e. having dimension $1 \times q$.

If $m = 1$, the representation of a sequence of three-way arrays results, and E_r has sth block the $n \times q$ matrix E_{rs}. In this case, E_{rst} denotes the tth row of E_{rs}, as a row vector, i.e. having dimension $1 \times q$. A sequence $\{E_r;\ r = 1, \ldots, k\}$ of two-way arrays of dimension $n \times q$ results from setting $m = 1$ and $l = 1$ in the four-way case. In this case, E_{rs} denotes the sth row of E_r, as a row vector, i.e. having dimension $1 \times q$. The blocking structure and dimensions of any quantity appearing below will be clearly stated or should be self-evident from the context. For brevity, the dependence of array structures on β will usually be suppressed.

2·2. *The general family of adjustments*

With the above conventions, let F and I denote the expected and observed information on the p-dimensional parameter β. Suppose that the tth component U_t of the score vector is adjusted to

$$U_t^* = U_t + A_t \quad (t = 1. \ldots, p),$$

This content downloaded from
130.126.162.126 on Sat, 28 Nov 2020 04:47:10 UTC
All use subject to https://about.jstor.org/terms

where A_t is $O_p(1)$ as $n \to \infty$. As in Firth (1993) but now in matrix notation, removal of the $O(n^{-1})$ bias term occurs if A_t is one of

$$A_t^{(E)} = \frac{1}{2}\operatorname{tr}\{F^{-1}(P_t + Q_t)\}, \quad A_t^{(O)} = I_t F^{-1} A^{(E)},$$

based on the expected or observed information, respectively. In the above, $P_t = E(UU^{\mathrm{T}}U_t)$ is the tth block of the $p^2 \times p$ third-order cumulant matrix of the score vector, and $Q_t = E(-IU_t)$ is the tth block of the $p^2 \times p$ blocked matrix of covariances of the first and second derivatives of the loglikelihood with respect to the parameters. The $1 \times p$ vector I_t denotes the tth row of I. All expectations are taken with respect to the model and at β. These adjustments, based on observed and expected information, are in fact particular instances of bias-reducing adjustments in the more general family

$$A_t = (G_t + R_t)F^{-1}A^{(E)}, \tag{1}$$

where R_t is $O_p(n^{1/2})$ and G_t is the tth row of either F or I. The adjustments (1) can usefully be re-expressed as a weighted sum of the adjustments based on the expected information, so that the components of the adjusted score vector take the general form

$$U_t^* = U_t + \sum_{u=1}^{p} e_{tu} A_u^{(E)} \quad (t = 1, \ldots, p). \tag{2}$$

Here e_{tu} denotes the (t, u)th component of matrix $(G + R)F^{-1}$; in the special case of adjustment based on the expected information, this is simply the identity matrix.

3. Adjusted score for multivariate exponential family models

3·1. *The multivariate generalized nonlinear model*

Consider a q-dimensional random variable Y from the exponential family having density or probability mass function of the form

$$f(y\,;\theta) = \exp\left\{\frac{y^{\mathrm{T}}\theta - b(\theta)}{\lambda} + c(y, \lambda)\right\},$$

in which the dispersion λ is assumed known. The expectation and the $q \times q$ variance-covariance matrix of Y are

$$E(Y;\theta) = \nabla b(\theta) = \mu(\theta), \quad \operatorname{cov}(Y\,;\theta) = \lambda \mathcal{D}^2\{b(\theta);\theta\} = \Sigma(\theta).$$

The quantity $\mathcal{D}^2\{b(\theta);\theta\}$ is the $q \times q$ Hessian matrix of $b(\theta)$ with respect to θ. Below, if a and b are p and q dimensional vectors, respectively, $\mathcal{D}(a;b)$ is the $p \times q$ matrix of the first derivatives of a with respect to b, and $\mathcal{D}^2\{a;b\}$ is a $pq \times q$ blocked Hessian matrix of the second derivatives of a with respect to b, whose blocks are the $q \times q$ matrices $\mathcal{D}^2\{a_i;b\}$ $(i = 1, \ldots, p)$.

Suppose now that observed response vectors $y_1, \ldots, y_n$ are realizations of independent random vectors $Y_1, \ldots, Y_n$, with θ_r the parameter vector for Y_r $(r = 1, \ldots, n)$ and λ_r the known dispersion. An exponential family nonlinear model, or generalized nonlinear model, connects the expectation of each Y_r with a predictor η_r through a known vector-valued link function $g_r : \Re^q \to \Re^q$,

$$g_r(\mu_r) = \eta_r(\beta) \quad (r = 1, \ldots, n), \tag{3}$$

This content downloaded from
130.126.162.126 on Sat, 28 Nov 2020 04:47:10 UTC
All use subject to https://about.jstor.org/terms

where $\eta_r : \Re^p \to \Re^q$ is a specified vector-valued function of model parameters β that is at least twice continuously differentiable. For technical details on the required regularity conditions and the properties of this class of models, see Wei (1997).

With the notation $\mu_r = \mu\{\theta_r(\beta)\}$ and $\Sigma_r = \Sigma\{\theta_r(\beta)\}$, the score vector for the parameters β of model (3) has the form

$$U = \sum_r X_r^{\mathrm{T}} W_r \mathcal{D}(\eta_r; \mu_r)(y_r - \mu_r), \tag{4}$$

where $W_r = D_r \Sigma_r^{-1} D_r^{\mathrm{T}}$ is the generalization of the working weight, defined exactly as for generalized linear multivariate models (Fahrmeir & Tutz, 2001, Appendix A.1). Also, $X_r = \mathcal{D}(\eta_r; \beta)$, and $D_r^{\mathrm{T}} = \mathcal{D}(\mu_r; \eta_r)$. The observed information on β is given by

$$\begin{aligned} I = \sum_r X_r^{\mathrm{T}} W_r X_r - \sum_r \sum_{s=1}^{q} \lambda_r^{-1} \left(X_r^{\mathrm{T}} V_{rs} X_r\right)(y_{rs} - \mu_{rs}) \\ - \sum_r \sum_{s,u=1}^{q} \mathcal{D}^2\{\eta_{ru}; \beta\} \left(\Sigma_r^{-1} D_r^{\mathrm{T}}\right)_{su}(y_{rs} - \mu_{rs}), \end{aligned} \tag{5}$$

where $V_{rs} = \mathcal{D}^2\{\theta_{rs}; \eta_r\}$ and $(\Sigma_r^{-1} D_r^{\mathrm{T}})_{su}$ is the (s, u)th element of the product $\Sigma_r^{-1} D_r^{\mathrm{T}}$. The expected information is $F = E(I) = \sum_r X_r^{\mathrm{T}} W_r X_r$.

3·2. *Derivation of the adjusted score*

From (4) and (5) and after some algebra, the sum of the cumulant matrices P_t and Q_t is found to be

$$P_t + Q_t = \sum_r \sum_{s=1}^{q} \left[X_r^{\mathrm{T}}\{(D_r \Sigma_r^{-1})_s \otimes 1_q\}\mathcal{D}^2(\mu_r; \eta_r) X_r + (W_{rs} \otimes 1_q)\mathcal{D}^2(\eta_r; \beta)\right] x_{rst}, \tag{6}$$

where W_{rs} is the sth row of the $q \times q$ matrix W_r as a $1 \times q$ vector, and $(D_r \Sigma_r^{-1})_s$ is the sth row of $D_r \Sigma_r^{-1}$ as a $1 \times q$ vector.

After substituting (6) into (2), and some tedious but straightforward algebraic manipulation, the adjusted score components for exponential family nonlinear models with known dispersion are found to take the form

$$\begin{aligned} U_t^* = U_t + \frac{1}{2} \sum_r \sum_{s=1}^{q} \operatorname{tr}\left[H_r W_r^{-1}\{(D_r \Sigma_r^{-1})_s \otimes 1_q\}\mathcal{D}^2(\mu_r; \eta_r)\right. \\ \left. + F^{-1}(W_{rs} \otimes 1_q)\mathcal{D}^2(\eta_r; \beta)\right] x_{rst}^* \quad (t = 1, \ldots, p), \end{aligned} \tag{7}$$

where $x_{rst}^* = \sum_{u=1}^{p} e_{tu} x_{rsu}$, and $H_r = X_r F^{-1} X_r^{\mathrm{T}} W_r$ $(r = 1, \ldots, n)$ are the $q \times q$ diagonal blocks of the projection or hat matrix that have the well-known leverage interpretation in linear models (see, for example, Fahrmeir & Tutz, 2001, § 4.2.2).

Expression (7) is convenient for computational purposes, especially, as the quantities involved are all routinely available after fitting a model by maximum likelihood in a typical software package. A further appealing feature of this representation is that its structure is unchanged for different choices from the generic family of adjustments (1): only the coefficients e_{tu} change.

Consider now the special case of the canonical link function g such that $\theta_r = g(\mu_r) = \eta_r$ $(r = 1, \ldots, n)$. This definition of canonical link extends the corresponding definition for generalized linear models. In the canonical-link case, $D_r = \lambda_r^{-1} \Sigma_r$, so $W_r^{-1} = \lambda_r^2 \Sigma_r^{-1}$ and

$$\{(D_r \Sigma_r^{-1})_s \otimes 1_q\}\mathcal{D}^2(\mu_r; \eta_r) = \lambda_r^{-1} \mathcal{D}^2(\mu_{rs}; \eta_r) = \lambda_r^{-3} K_{rs},$$

This content downloaded from
130.126.162.126 on Sat, 28 Nov 2020 04:47:10 UTC
All use subject to https://about.jstor.org/terms

where K_{rs} denotes the sth block of rows of K_r, $s \in \{1, \ldots, q\}$, with K_r the blocked $q^2 \times q$ matrix of third-order cumulants of the random vector Y_r. Hence, expression (7) is considerably simplified, taking the form

$$U_t^* = U_t + \frac{1}{2} \sum_r \sum_{s=1}^{q} \lambda_r^{-1} \operatorname{tr} \{ H_r \Sigma_r^{-1} K_{rs} + \lambda_r^{-1} F^{-1} (\Sigma_{rs} \otimes 1_p) \mathcal{D}^2(\eta_r; \beta) \} x_{rst}^*. \tag{8}$$

3·3. *A special case: The multivariate generalized linear model*

Generalized linear models have the form (3), with $\eta_r(\beta) = X_r \beta$, where the $q \times p$ design matrix X_r is some appropriate function of a covariate vector and does not depend on β. Thus, in equation (7), $\mathcal{D}^2 \{\eta_r; \beta\} = 0$ and the adjusted score reduces to

$$U_t^* = U_t + \frac{1}{2} \sum_r \sum_{s=1}^{q} \operatorname{tr} \left[H_r W_r^{-1} \{ (D_r \Sigma_r^{-1})_s \otimes 1_q \} \mathcal{D}^2(\mu_r; \eta_r) \right] x_{rst}^*. \tag{9}$$

By the same argument, for canonical link functions, expression (8) simplifies to

$$U_t^* = U_t + \frac{1}{2} \sum_r \sum_{s=1}^{q} \lambda_r^{-1} \operatorname{tr} (H_r \Sigma_r^{-1} K_{rs}) x_{rst}^*.$$

If the link is canonical, the expected and observed information coincide, so that e_{ts} simplifies to the (t, s)th element of $1_p + RF^{-1}$, with 1_p the $p \times p$ identity matrix.

3·4. *Notes on implementation*

An obvious way of solving the adjusted score equation is a modified Fisher scoring iteration, in which the likelihood score is replaced by the adjusted score; the maximum likelihood estimates, if available and finite, often provide a good choice of starting values. However, more convenient and possibly more efficient schemes can be constructed by exploiting the special structure of the adjusted score for any given model. For example, Firth (1992a,1992b), in the special case of univariate generalized linear models with canonical link, derived a modified iterative reweighted least squares procedure. In later sections this problem is revisited in order to generalize such procedures to univariate generalized linear and nonlinear models with any link function.

A generalized linear model with noncanonical link can always be expressed as a generalized nonlinear model with canonical link, and consequently (9) is equivalent to (8) for the corresponding generalized nonlinear model with canonical link. This fact can sometimes be exploited to allow the use of existing software, with relatively minor modifications, for implementation of the bias-reducing adjustments.

4. Univariate generalized linear models

4·1. *The adjusted score*

Suppose now that the response variable is scalar, i.e., $q = 1$. For notational simplicity, in the univariate case write $\kappa_{2,r}$ and $\kappa_{3,r}$ for the variance and the third cumulant of Y_r, respectively, and $w_r = d_r^2/\kappa_{2,r}$ for the working weights, where $d_r = \mathrm{d}\mu_r/\mathrm{d}\eta_r$ $(r = 1, \ldots, n)$. For a univariate generalized linear model with general link function, the adjusted score components (9) reduce to

$$U_t^* = U_t + \frac{1}{2} \sum_r h_r \frac{d_r'}{d_r} \sum_{s=1}^{p} e_{ts} x_{rs} \quad (t = 1, \ldots, p), \tag{10}$$

This content downloaded from
130.126.162.126 on Sat, 28 Nov 2020 04:47:10 UTC
All use subject to https://about.jstor.org/terms

where $d_r' = \mathrm{d}^2\mu_r/\mathrm{d}\eta_r^2$ and x_{rt} is the (r, t)th element of the $n \times p$ design matrix X. The quantity h_r is the rth diagonal element of the projection matrix $H = XF^{-1}X^{\mathrm{T}}W$, where $W = \mathrm{diag}(w_1, \ldots, w_n)$, and $F = X^{\mathrm{T}}WX$. Note that d_r'/d_r depends solely on the link function.

If the link is canonical, $d_r = \lambda_r^{-1}\kappa_{2,r}$, $d_r' = \lambda_r^{-2}\kappa_{3,r}$ and e_{ts} simplifies to the (t, s)th element of $1_p + RF^{-1}$ in the latter expression. Furthermore, if R is taken to be a matrix of zeros then, as given by Firth (1992a,1992b), $U_t^* = U_t + \sum_r h_r\{\kappa_{3,r}/(2\lambda_r\kappa_{2,r})\}x_{rt}$.

4·2. *Existence of penalized likelihoods for univariate generalized linear models*

For a generalized linear model with canonical link, the bias-reducing score adjustment corresponds to penalization of the likelihood by the Jeffreys (1946) invariant prior (Firth, 1993). More generally, in models with noncanonical link and $p \geqslant 2$, there need not exist a penalized loglikelihood l^* such that $\nabla l^*(\beta) \equiv U^*(\beta)$. The following theorem identifies those noncanonical link functions that always, regardless of the dimension or structure of X, do admit such a penalized likelihood interpretation.

THEOREM 1. *In the class of univariate generalized linear models, there exists a penalized loglikelihood l^* such that $\nabla l^*(\beta) \equiv U(\beta) + A^{(E)}(\beta)$, for all possible specifications of design matrix X, if and only if the inverse link derivatives $d_r = 1/g_r'(\mu_r)$ satisfy*

$$d_r \equiv \alpha_r\kappa_{2,r}^{\omega} \quad (r = 1, \ldots, n), \tag{11}$$

where α_r $(r = 1, \ldots, n)$ and ω do not depend on the model parameters. When condition (11) *holds, the penalized loglikelihood is*

$$l^*(\beta) = \begin{cases} l(\beta) + \dfrac{1}{4}\displaystyle\sum_r \log \kappa_{2,r}(\beta)^{h_r} & (\omega = 1/2), \\ l(\beta) + \dfrac{\omega}{4\omega - 2}\log|F(\beta)| & (\omega \neq 1/2). \end{cases} \tag{12}$$

The proof is in the Appendix.

In the above theorem the canonical link is the special case $\omega = 1$. With $\omega = 0$, condition (11) refers to identity links for which the loglikelihood penalty is identically zero. The case $\omega = 1/2$ is special as the working weights, and hence F and H, do not in that case depend on β.

Example 1. Consider a Poisson generalized linear model with link function from the power family $\eta = (\mu^{\nu} - 1)/\nu$ (McCullagh & Nelder, 1989, § 2.2.3). Then $d_r = \mu_r^{1-\nu}$ and $\kappa_{2,r} = \mu_r$ $(r = 1, \ldots, n)$. Bias reduction based on expected information is equivalent to maximization of penalized likelihood (12) with $\omega = 1 - \nu$.

If $\omega \notin [0, 1/2]$, it is immediate from (12) that bias reduction of the maximum likelihood estimator also increases the value of $|F(\beta)|$ at the maximum; hence approximate confidence ellipsoids, based on asymptotic normality of the estimator, are reduced in volume. When $|F(\beta)|$ is log-concave, as in binomial logistic regressions for example, this shrinkage towards the point of minimum generalized variance also ensures finiteness of the estimator, see chapter 4 of an unpublished University of Warwick thesis by Kosmidis.

Theorem 1 has direct practical consequences, notably for the construction of confidence sets. The use of profile penalized likelihood as suggested, for example, by Heinze & Schemper (2002) and Bull et al. (2007) is always available for a generalized linear model whose link function satisfies condition (11), but typically is not possible otherwise. Models that fail to meet condition (11) include, for example, probit and complementary log-log models for binary responses.

This content downloaded from
130.126.162.126 on Sat, 28 Nov 2020 04:47:10 UTC
All use subject to https://about.jstor.org/terms

4·3. *Pseudo-responses and implementation via modified working observations*

The likelihood score components for univariate generalized linear models with general link have the form $U_t = \sum_r \{w_r(y_r - \mu_r)/d_r\} x_{rt}$ $(t = 1, \ldots, p)$. A simple substitution in expression (10) reveals an important feature of the adjusted score, which was recognized previously (Firth, 1993) in the more specific context of canonical-link models. Consider the case of adjustments based on the expected information, i.e. e_{ts} is unity if $t = s$ and zero otherwise. If $h_r d'_r/(2w_r)$ $(r = 1, \ldots, n)$ were known constants (for example, when $\nu = 1/2$ in Example 1) then the bias reduction method would be formally equivalent to maximum likelihood with the pseudo-responses $y^*_r = y_r + h_r d_r/(2w_r)\,(r = 1, \ldots, n)$ used in place of y_r. Table 1 gives the form of the pseudo-responses for some commonly used generalized linear models. The pseudo-responses suggest a simple approach to implementation, using a familiar likelihood-maximization algorithm such as iterative reweighted least squares but with y_r replaced by y^*_r. In general, $h_r d'_r/w_r$ depends on the parameters, and so the value of y^*_r will be updated according to the current estimates at each step of the algorithm. This modified iterative reweighted least squares step is equivalent to the modified Fisher scoring algorithm of § 3·4 and can be more conveniently described in terms of the modified working observations

$$\zeta^*_r = \eta_r + \frac{y^*_r - \mu_r}{d_r} = \zeta_r - \xi_r \quad (r = 1, \ldots, n).$$

Here, $\zeta_r = \eta_r + (y_r - \mu_r)/d_r$ is the working observation for maximum likelihood, and $\xi_r = -d'_r h_r/(2w_r d_r)$, as defined by Cordeiro & McCullagh (1991). Thus, if the working observation ζ_r is modified by adding $-\xi_r$, the resulting iterative reweighted least squares scheme returns the bias-reduced estimates.

This result stems from the initial definition of the modifications in terms of the $O(n^{-1})$ bias of the maximum likelihood estimator and the fact that, as shown by Cordeiro & McCullagh (1991), the vector of $O(n^{-1})$ biases can be written as $n^{-1}b_1 = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W\xi$.

These appealingly simple results do not extend in any obvious way to the case of multivariate responses, because W is no longer diagonal but rather is block diagonal and because the vector of $O(n^{-1})$ biases can be expressed at best as a function of traces of products of matrices with no other apparent simplification.

5. Univariate generalized nonlinear models

With $q = 1$ in expression (7), and under the notational conventions of § 4, the adjusted score components for a univariate exponential family nonlinear model with known dispersion are

$$U^*_t = U_t + \frac{1}{2}\sum_r \left[h_r \frac{d'_r}{d_r} + w_r \operatorname{tr}\{F^{-1}\mathcal{D}^2(\eta_r; \beta)\}\right] x^*_{rt} \quad (t = 1, \ldots, p). \tag{13}$$

In the above expression, $x^*_{rt} = \sum_{s=1}^{p} e_{ts}x_{rs}$, $x_{rs} = \partial\eta_r/\partial\beta_s$ and $\mathcal{D}^2(\eta_r; \beta)$ is the $p \times p$ Hessian matrix of η_r with respect to β. The remaining quantities in (13) are the same as for generalized linear models.

For canonical links, in expression (13), $d'_r/d_r = \lambda^{-1}\kappa_{3,r}/\kappa_{2,r}$ and $w_r = \lambda_r^{-2}\kappa_{2,r}$.

In the case of adjustments based on the expected information, the fitting procedures for univariate generalized linear models can be used in the nonlinear case with only slight changes to the definition of the pseudo-responses and the modified working observations. By (13), the pseudo-responses of Table 1 are straightforwardly adapted to the nonlinear case by adding the extra term $d_r \operatorname{tr}\{F^{-1}\mathcal{D}^2(\eta_r; \beta)\}/2$. The $O(n^{-1})$ bias of the maximum-likelihood

This content downloaded from
130.126.162.126 on Sat, 28 Nov 2020 04:47:10 UTC
All use subject to https://about.jstor.org/terms

Table 1. *Pseudo-responses for several commonly used generalized linear models*

	Link function	Pseudo-responses
Distribution of Y	$\eta = g(\mu)$	$y^* = y + hd'/(2w)$
Binomial (m, π)	$\eta = \log\{\pi/(1-\pi)\}^{\dagger}$	$y^* = y + h(1/2 - \pi)$
	$\eta = \Phi^{-1}(\pi)$	$y^* = y - h\pi(1-\pi)\eta/\{2\phi(\eta)\}$
	$\eta = \log\{-\log(1-\pi)\}$	$y^* = y + h\pi(1-e^{\eta})/(2e^{\eta})$
	$\eta = -\log\{-\log(\pi)\}$	$y^* = y + h(1-\pi)(e^{-\eta}-1)/(2e^{-\eta})$
Poisson (μ)	$\eta = \log\mu^{\dagger}$	$y^* = y + h/2$
	$\eta = \mu$	$y^* = y$
Gamma (μ, ν)	$\eta = -1/\mu^{\dagger}$	$y^* = y + h\mu/\nu$
$\operatorname{var}(Y) = \mu^2/\nu$	$\eta = \log\mu$	$y^* = y + h\mu/(2\nu)$
	$\eta = \mu$	$y^* = y$
Inverse Gaussian (λ, μ)	$\eta = -1/(2\mu^2)^{\dagger}$	$y^* = y + 3h\lambda\mu^2/2$

The normal distribution function and density are denoted by Φ and ϕ, respectively.

†Canonical link. The pseudo-responses for Binomial, Poisson and Gamma models with canonical link are also given by Firth (1992a,1992b).

estimator of the parameter vector in the case of a nonlinear predictor can be written in the form $n^{-1}b_1 = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W\xi^{(N)}$, which generalizes the corresponding expression for normal nonlinear regression models given by Cook et al. (1986). The vector $\xi^{(N)}$ has components

$$\xi_r^{(N)} = -\frac{1}{2}\left[\frac{h_r d_r'}{w_r d_r} + \operatorname{tr}\{F^{-1}\mathcal{D}^2(\eta_r;\beta)\}\right] \quad (r = 1, \ldots, n),$$

and by similar arguments to those used for generalized linear models, the rth modified working variate takes the form $\zeta_r^* = \zeta_r - \xi_r^{(N)}$, with $\zeta_r = \sum_{t=1}^{p}\beta_t x_{rt} + (y_r - \mu_r)/d_r$.

6. Illustration: bias reduction for the rc(1) association model

6·1. *Modified working variates*

The utility of the preceding developments and the properties of the bias-reduced estimator are illustrated here by application to the row-column association model of order 1, often denoted by rc(1), for the analysis of two-way contingency tables (Goodman, 1979, 1985).

Consider a two-way cross-classification by factors X and Y with R and S levels, respectively. The entries of the table are assumed to be realizations of independent Poisson random variables Y_{rs} with means μ_{rs} $(r = 1, \ldots, R; s = 1 \ldots, S)$. For the rc(1) model, μ_{rs} is linked to a nonlinear function η_{rs} of model parameters according to

$$\log\mu_{rs} = \eta_{rs} = \lambda + \lambda_r^X + \lambda_s^Y + \rho\gamma_r\delta_s, \tag{14}$$

where λ_r^X and λ_s^Y are the row and column main effects, γ_r and δ_s are the row and column score parameters and ρ is an association parameter. The rc(1) model is viewed here as a univariate generalized nonlinear model with canonical link.

Write the parameter vector as

$$\beta = \left(\lambda, \lambda_1^X, \ldots, \lambda_R^X, \lambda_1^Y, \ldots, \lambda_S^Y, \rho, \gamma_1, \ldots, \gamma_R, \delta_1, \ldots, \delta_S\right)^{\mathrm{T}},$$

with $p = 2(R + S + 1)$. The full parameter vector β is unidentified, a set of six constraints being required in order to fix the location of $\{\lambda_r^X\}$ and $\{\lambda_s^Y\}$ and the location and scale of $\{\gamma_r\}$ and $\{\delta_s\}$. In what follows, it will be assumed for simplicity that six suitably chosen elements of β are constrained at fixed values.

This content downloaded from
130.126.162.126 on Sat, 28 Nov 2020 04:47:10 UTC
All use subject to https://about.jstor.org/terms

Let $\eta_r = (\eta_{r1}, \ldots, \eta_{rS})^{\mathrm{T}}$ and denote the reduced vector of parameters, after removal of the constrained elements, by β^-. For the RC(1) model, $d'_{rs}/d_{rs} = 1$ and $w_{rs} = \mu_{rs}$. Thus, by the results in § 5, the vector of modified working variates $\zeta^* = (\zeta^*_{11}, \ldots, \zeta^*_{1S}, \ldots, \zeta^*_{R1}, \ldots, \zeta^*_{RS})$ has components

$$\zeta^*_{rs} = \zeta_{rs} + \frac{h_{rs}}{2\mu_{rs}} + M_{rs}. \tag{15}$$

Here $M_{rs} = \mathrm{tr}\{F^{-1}\mathcal{D}^2(\eta_{rs}; \beta^-)\}/2$, and h_{rs} is the sth diagonal element of the $S \times S$ matrix $X_r F^{-1} X_r^{\mathrm{T}} W_r$, with W_r a diagonal matrix with sth diagonal element μ_{rs}, $s = 1, \ldots, S$ and with X_r the $S \times (p-6)$ matrix of derivatives of η_{rs} with respect to β^-.

The sth row of X_r results from the deletion of all components of $\mathcal{D}(\eta_{rs}; \beta)$ that correspond to constrained parameters. The derivatives are

$$\mathcal{D}(\eta_{rs}; \beta) = \left(1, i^R_r, i^S_s, \gamma_r\delta_s, \rho\delta_s i^R_r, \rho\gamma_r i^R_r\right),$$

with, for example, i^R_r denoting a $1 \times R$ row vector of zeros with 1 at the rth position. Thus, by noting that $\mathcal{D}^2\{\eta_r; \beta^-\} = \mathcal{D}(X_r; \beta^-)$, and after some straightforward algebra, M_{rs} can be expressed conveniently in the form

$$M_{rs} = \gamma_r C(\rho, \delta_s) + \delta_s C(\rho, \gamma_r) + \rho C(\gamma_r, \delta_s). \tag{16}$$

Here $C(\kappa, \nu)$ for any given pair of unconstrained parameters κ and ν denotes the corresponding element of F^{-1}; if either κ or ν is constrained, $C(\kappa, \nu) = 0$.

Expressions (15) and (16) combine to provide a rather simple implementation of the bias-reduction method in this case via the modified iterative weighted least squares procedure.

6·2. *Simulation study*

For illustration, consider the periodontal condition and calcium intake data given by Goodman (1981, Table 1a). The 135 women under study are cross-classified according to two factors: periodontal condition (X) and calcium intake (Y), each with four levels ($R = 4$, $S = 4$). The constraints used are $\lambda^X_1 = \lambda^Y_1 = 0$, $\gamma_1 = \delta_1 = -2$ and $\gamma_4 = \delta_4 = 2$. With these constraints, the maximum likelihood estimates are obtained and 250 000 tables are simulated from the maximum likelihood fit. For each simulated table, the following are computed: the maximum likelihood estimates, the bias-reduced estimates, corresponding estimated standard errors computed as square roots of the diagonal of the inverse expected information and a nominally 95% Wald-type confidence interval for each parameter separately.

In the RC(1) model, the maximum likelihood estimator has a nonzero probability of at least one of its components being infinite-valued; the mean and other moments of that estimator are undefined. In the present simulation study, almost 3.5% of the tables generated had one or more components of the maximum likelihood vector at infinity. This creates a difficulty for direct comparisons with the bias-reduced estimator, which is always finite. In Table 2, the summaries presented for the maximum likelihood estimator are computed from only that subset of tables in which the estimates were all finite; those summaries are thus estimates of conditional bias, conditional mean squared error and conditional coverage probability of confidence intervals. A direct comparison with corresponding unconditional summaries for the bias-reduced estimator is likely to be misleading, in the direction of favouring the method of maximum likelihood.

Despite the reservation just mentioned, the results in Table 2 indicate that the bias-reduced estimator has bias and mean squared error properties that are better, even, than the corresponding conditional summaries for maximum likelihood. For some components of the parameter vector, the difference is substantial. In terms of coverage frequency of nominally 95% confidence intervals, the results for the two methods are fairly similar; but again it should be noted that

This content downloaded from
130.126.162.126 on Sat, 28 Nov 2020 04:47:10 UTC
All use subject to https://about.jstor.org/terms

Table 2. *Results for the dental health data. Simulation standard errors are in parentheses; for the coverage probabilities they are approximately* 0·04%. *For the method of maximum likelihood, simulation results are all conditional upon finiteness of the estimates*

			Simulation results					
	Estimates		Bias ($\times 10^2$)		MSE ($\times 10$)		Coverage (%)	
	ML	BR	ML	BR	ML	BR	ML	BR
λ	2·31	2·35	−4·19 (0·10)	−0·25 (0·08)	2·28 (0·01)	1·49 (0·01)	96·9	96·6
λ_2^X	−0·13	−0·13	0·48 (0·08)	−0·01 (0·07)	1·45 (0·01)	1·16 (< 0·01)	95·8	96·2
λ_3^X	0·55	0·52	2·97 (0·08)	−0·22 (0·07)	1·50 (0·01)	1·18 (< 0·01)	95·7	96·0
λ_4^X	0·07	0·10	−5·00 (0·12)	0·02 (0·09)	3·34 (0·02)	1·87 (0·01)	97·1	97·3
λ_2^Y	−0·53	−0·53	−0·59 (0·06)	0·06 (0·06)	1·00 (< 0·01)	0·80 (< 0·01)	96·0	96·4
λ_3^Y	−1·17	−1·05	−16·81 (0·16)	1·19 (0·11)	6·55 (0·05)	2·80 (0·01)	97·1	96·1
λ_4^Y	−0·80	−0·75	−7·21 (0·11)	0·22 (0·08)	3·19 (0·02)	1·69 (< 0·01)	97·3	97·3
ρ	−0·20	−0·18	−1·76 (0·01)	−0·03 (0·01)	0·05 (< 0·01)	0·03 (< 0·01)	95·5	95·0
γ_2	−1·55	−1·48	−6·08 (0·16)	0·68 (0·15)	6·30 (0·03)	5·37 (0·02)	95·6	96·7
γ_3	0·90	0·91	1·88 (0·17)	1·43 (0·15)	6·94 (0·03)	5·34 (0·02)	93·8	95·2
δ_2	−1·16	−1·11	−7·00 (0·19)	−0·27 (0·17)	9·00 (0·05)	7·20 (0·03)	94·7	96·4
δ_3	3·11	2·84	37·42 (0·37)	−4·92 (0·27)	35·55 (0·23)	18·13 (0·06)	92·8	92·4

ML, maximum likelihood; BR, bias-reduced maximum likelihood; MSE, mean squared error.

the coverage probability for maximum likelihood excludes those tables with infinite-valued estimates, for which no Wald-type interval can be computed. From this experiment it appears that the bias-reduced estimator has similar advantages in the RC(1) model to those that have been found by other authors in linear logistic regressions and various other contexts.

7. CONCLUDING REMARKS

Explicit, general formulae have been derived for the adjusted score equations that produce second-order unbiased estimators, starting from the wide class of multivariate-response exponential family nonlinear models and narrowing down to the simplest case of canonically-linked generalized linear models. As shown in the case of the RC(1) model, simplification of the formulae is also possible in some other special cases, such as generalized bilinear models, by exploiting the specific structure.

The apparent focus here on models with known dispersion does not affect the wide applicability of the results. In generalized linear and nonlinear models where the dispersion λ is unknown, it is usually estimated separately from the parameters β which determine the mean; given an estimate of λ, the methods developed here can simply be applied with λ fixed at its current estimate. The well-known orthogonality of mean and dispersion parameters plays an important role in this.

Generally, as for the maximum likelihood estimator, approximate confidence intervals for the bias-reduced estimator can be constructed by the usual Wald method. However, as noted by Heinze & Schemper (2002) and Bull et al. (2007), for logistic regressions the Wald-type intervals can have poor coverage properties. Heinze & Schemper (2002) and Bull et al. (2007) propose the use, instead, of intervals based on the profile penalized likelihood, which are found to have better properties. Theorem 1 shows that such use of profile penalized likelihood can be extended beyond logistic regressions but is not universally available.

ACKNOWLEDGEMENT

The authors thank the editor, the associate editor and a referee for helpful comments that substantially improved the presentation. Financial support from the Engineering and Physical Sciences Research Council and the Economic and Social Research Council is gratefully acknowledged.

This content downloaded from
130.126.162.126 on Sat, 28 Nov 2020 04:47:10 UTC
All use subject to https://about.jstor.org/terms

Appendix

Proof of Theorem 1. Note that $d_r'/d_r = \mathrm{d}\log d_r/\mathrm{d}\eta_r$ and so, for adjustments based on the expected information, (10) can be written as

$$U_t^* = U_t + \frac{1}{2}\,\mathrm{tr}\,(HET_t) \quad (t = 1, \ldots, p), \tag{A1}$$

with $E = \mathrm{diag}(\mathrm{d}\log d_1/\mathrm{d}\eta_1, \ldots, \mathrm{d}\log d_n/\mathrm{d}\eta_n)$ and $T_t = \mathrm{diag}(x_{1t}, \ldots, x_{nt})$. As, for example, in the case of the existence of quasilikelihoods in McCullagh & Nelder (1989, § 9.3.2), there exists a penalized loglikelihood l^* such that $\nabla l^*(\beta) \equiv U(\beta) + A^{(E)}(\beta)$, if and only if $\partial U_s^*(\beta)/\partial\beta_t = \partial U_t^*(\beta)/\partial\beta_s$ for every $s, t \in \{1, 2, \ldots, p\}$, $s \neq t$. By (A1), this holds if and only if

$$\frac{\partial\,\mathrm{tr}\,(HET_t)}{\partial\beta_s} = \mathrm{tr}\left(H\frac{\partial E}{\partial\beta_s}T_t\right) + \mathrm{tr}\left(\frac{\partial H}{\partial\beta_s}ET_t\right)$$

is invariant under interchange of the subscripts s and t. The first term in the right-hand side of the above expression is invariant because $\partial E/\partial\beta_s = ET_s$ and T_s and T_t are by definition diagonal so that matrix multiplication is commutative for them. For the second term,

$$\frac{\partial H}{\partial\beta_s} = -X(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W_sX(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W + X(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W_s,$$

where $W_s = \partial W/\partial\beta_s = W(2E - \Lambda)T_s$ with $\Lambda = \mathrm{diag}(\mathrm{d}\log\kappa_{2,1}/\mathrm{d}\eta_1, \ldots, \mathrm{d}\log\kappa_{2,n}/\mathrm{d}\eta_n)$. Thus,

$$\mathrm{tr}\left(\frac{\partial H}{\partial\beta_s}ET_t\right) = 2\,\mathrm{tr}\,(HET_sET_t) - 2\,\mathrm{tr}\,(HET_sHET_t) - \mathrm{tr}\,(H\Lambda T_sET_t) + \mathrm{tr}\,(H\Lambda T_sHET_t).$$

By the properties of the trace function, the first three terms in the right-hand side of the above expression are invariant under interchange of s and t. Thus the condition is reduced to the invariance of $\mathrm{tr}(H\Lambda T_sHET_t)$. The projection matrix H can be written as $H = SW$, with $S = X(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}$. Let $\tilde{H} = W^{1/2}SW^{1/2}$. Taking into account the symmetry of $\tilde{H}$, some trivial algebra and an application of Theorem 3 of Magnus & Neudecker (1999, chapter 2) gives

$$\mathrm{tr}(H\Lambda T_sHET_t) = \mathrm{tr}(\tilde{H}T_s\Lambda\tilde{H}ET_t) = (\mathrm{vec}\,T_t)^{\mathrm{T}}\{(E\tilde{H}\Lambda) \otimes \tilde{H}\}\,\mathrm{vec}\,T_s. \tag{A2}$$

The columns of X are by assumption linearly independent and thus (A2) is invariant under interchanges of s and t if and only if $a^{\mathrm{T}}\{(E\tilde{H}\Lambda) \otimes \tilde{H}\}b$ is a symmetric bilinear form for distinct vectors a and b of appropriate dimension, or equivalently if and only if $E\tilde{H}\Lambda$ is symmetric, i.e.

$$\frac{\mathrm{d}\log d_r}{\mathrm{d}\eta_r}\frac{\mathrm{d}\log\kappa_{2,i}}{\mathrm{d}\eta_i}\tilde{h}_{ri} = \frac{\mathrm{d}\log\kappa_{2,r}}{\mathrm{d}\eta_r}\frac{\mathrm{d}\log d_i}{\mathrm{d}\eta_i}\tilde{h}_{ri} \quad (r, i = 1, \ldots, n; r > i), \tag{A3}$$

with $\tilde{h}_{ri}$ the (r, i)th element of $\tilde{H}$.

In general the above equations are not satisfied simultaneously, except possibly for special structures of the design matrix X, which cause $\tilde{h}_{ri} = 0$ for a set of pairs (r, i). Hence, assuming that $\tilde{h}_{ri} \neq 0$ $(r, i = 1, \ldots, n; r > i)$, the general equation in (A3) reduces to $\mathrm{d}\log d_r/\mathrm{d}\eta_r\mathrm{d}\log\kappa_{2,i}/\mathrm{d}\eta_i = \mathrm{d}\log\kappa_{2,r}/\mathrm{d}\eta_r\mathrm{d}\log d_i/\mathrm{d}\eta_i$. Now, if $\mathrm{d}\log\kappa_{2,r}/\mathrm{d}\eta_r = \mathrm{d}\log\kappa_{2,i}/\mathrm{d}\eta_i = 0$ for some pair (r, i) then the equation for this (r, i) is satisfied. Thus, without loss of generality assume that $\mathrm{d}\log\kappa_{2,r}/\mathrm{d}\eta_r \neq 0$ for every $r \in \{1, \ldots, n\}$. Under these assumptions condition (A3) can be written as $\mathrm{d}\log d_r/\mathrm{d}\eta_r = \omega\mathrm{d}\log\kappa_{2,r}/\mathrm{d}\eta_r$ $(r = 1, \ldots, n)$, where ω does not depend on the model parameters. By integration of both sides of $\mathrm{d}\log d_r/\mathrm{d}\eta_r = \omega\mathrm{d}\log\kappa_{2,r}/\mathrm{d}\eta_r$ with respect to η_r, a necessary condition for the adjusted score to be the gradient of a penalized likelihood is thus

$$d_r \equiv \alpha_r\kappa_{2,r}^{\omega} \quad (r = 1, \ldots, n), \tag{A4}$$

where $\{\alpha_r : r = 1, \ldots, n\}$ are real constants not depending on the model parameters. Sufficiency of (A4) follows from the symmetry of $E\tilde{H}\Lambda$.

In addition, if condition (A4) is satisfied for some ω and α_r $(r = 1, \ldots, n)$ then the rth diagonal element of E is $\mathrm{d}\log d_r/\mathrm{d}\eta_r = \omega\kappa_{2,r}'/\kappa_{2,r}$ for every $r \in \{1, \ldots, n\}$, with $\kappa_{2,r}' = \mathrm{d}\kappa_{2,r}/\mathrm{d}\eta_r$. On the other

This content downloaded from
130.126.162.126 on Sat, 28 Nov 2020 04:47:10 UTC
All use subject to https://about.jstor.org/terms

hand, $\mathrm{d}\,w_r/\mathrm{d}\eta_r = (2\omega - 1)w_r\kappa'_{2,r}/\kappa_{2,r}$. Hence, by (A1) and for $\omega \neq 1/2$ the tth component of the adjusted score vector is

$$U_t(\beta) + \frac{\omega}{4\omega - 2}\,\mathrm{tr}[X\{X^{\mathrm{T}}W(\beta)X\}^{-1}X^{\mathrm{T}}W_t(\beta)] = \frac{\partial}{\partial\beta_t}\left\{l(\beta) + \frac{\omega}{4\omega - 2}\log|X^{\mathrm{T}}W(\beta)X|\right\},$$

where and $W_t(\beta) = \partial W(\beta)/\partial\beta_t$.

If $\omega = 1/2$ then $w_r = \alpha_r^2$ $(r = 1, \ldots, n)$. Hence, the hat matrix H does not depend on the model parameters. Thus, by (10), the tth component of the adjusted score vector is

$$U_t(\beta) + \frac{1}{4}\sum_r h_r \frac{\kappa'_{2,r}(\beta)}{\kappa_{2,r}(\beta)} x_{rt} = \frac{\partial}{\partial\beta_t}\left\{l(\beta) + \frac{1}{4}\sum_r \log\kappa_{2,r}(\beta)^{h_r}\right\}. \qquad \square$$

References

BULL, S. B., LEWINGER, J. B. & LEE, S. S. F. (2007). Confidence intervals for multinomial logistic regression in sparse data. *Statist. Med.* **26**, 903–18.

BULL, S. B., MAK, C. & GREENWOOD, C. (2002). A modified score function estimator for multinomial logistic regression in small samples. *Comp. Statist. Data Anal.* **39**, 57–74.

COOK, R. D., TSAI, C.-L. & WEI, B. C. (1986). Bias in nonlinear regression. *Biometrika* **73**, 615–23.

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

CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

FAHRMEIR, L. & TUTZ, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer.

FIRTH, D. (1992a). Bias reduction, the Jeffreys prior and GLIM. In *Advances in GLIM and Statistical Modelling: Proc. GLIM 92 Conf.*, Ed. L. Fahrmeir, B. Francis, R. Gilchrist & G. Tutz, pp. 91–100. New York: Springer.

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

FIRTH, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

GOODMAN, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *J. Am. Statist. Assoc.* **74**, 537–52.

GOODMAN, L. A. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *J. Am. Statist. Assoc.* **76**, 320–34.

GOODMAN, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Ann. Statist.* **13**, 10–69.

HEINZE, G. & SCHEMPER, M. (2002). A solution to the problem of separation in logistic regression. *Statist. Med.* **21**, 2409–19.

HEINZE, G. & SCHEMPER, M. (2004). A solution to the problem of monotone likelihood in Cox regression. *Biometrics* **57**, 114–9.

JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond.* **186**, 453–61.

MAGNUS, J. R. & NEUDECKER, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Chichester: Wiley.

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

MEHRABI, Y. & MATTHEWS, J. N. S. (1995). Likelihood-based methods for bias reduction in limiting dilution assays. *Biometrics* **51**, 1543–9.

PETTITT, A. N., KELLY, J. M. & GAO, J. T. (1998). Bias correction for censored data with exponential lifetimes. *Statist. Sinica* **8**, 941–64.

SARTORI, N. (2006). Bias prevention of maximum likelihood estimates for scalar skew normal and skew *t* distributions. *J. Statist. Plan. Infer.* **136**, 4259–75.

WEI, B. (1997). *Exponential Family Nonlinear Models*. New York: Springer.

ZORN, C. (2005). A solution to separation in binary response models. *Polit. Anal.* **13**, 157–70.

[*Received February* 2008. *Revised April* 2009]

This content downloaded from
130.126.162.126 on Sat, 28 Nov 2020 04:47:10 UTC
All use subject to https://about.jstor.org/terms