Electronic Journal of Statistics
Vol. 4 (2010) 1097–1112
ISSN: 1935-7524
DOI: 10.1214/10-EJS579

A generic algorithm for reducing bias in parametric estimation

Ioannis Kosmidis[*,†]

Department of Statistical Science
University College
London, WC1E 6BT, UK, e-mail: ioannis@stats.ucl.ac.uk

and

David Firth[*]

Department of Statistics
University of Warwick
Coventry, CV4 7AL, UK, e-mail: d.firth@warwick.ac.uk

Abstract: A general iterative algorithm is developed for the computation of reduced-bias parameter estimates in regular statistical models through adjustments to the score function. The algorithm unifies and provides appealing new interpretation for iterative methods that have been published previously for some specific model classes. The new algorithm can usefully be viewed as a series of iterative bias corrections, thus facilitating the adjusted score approach to bias reduction in any model for which the first-order bias of the maximum likelihood estimator has already been derived. The method is tested by application to a logit-linear multiple regression model with beta-distributed responses; the results confirm the effectiveness of the new algorithm, and also reveal some important errors in the existing literature on beta regression.

AMS 2000 subject classifications: Primary 62F10, 62F12; secondary 62F05.
Keywords and phrases: Adjusted score, asymptotic bias correction, beta regression, bias reduction, fisher scoring, prater gasoline data.

Received September 2010.

Contents

[*]This work was carried out under the support of the UK Engineering and Physical Sciences Research Council.
[†]This work was carried out when the first author was a member of the Centre for Research in Statistical Methodology, University of Warwick.

1. Introduction

Suppose that interest is in the estimation of the q-vector of parameters $\beta = (\beta_1, \ldots, \beta_q)$ of a parametric model. If $l(\beta)$ is the log-likelihood for β, the maximum likelihood estimator $\hat{\beta}$ solves the score equations

$$S(\beta) = \nabla_\beta l(\beta) = 0\,,$$

provided that the observed information matrix $I(\beta) = -\nabla_\beta \nabla_\beta^T l(\beta)$ is positive definite when evaluated at $\hat{\beta}$. Under fairly standard regularity conditions (for example, the conditions described in McCullagh [21, §§7.1,7.2], or equivalently the conditions in Cox and Hinkley [8, §9.1]), the maximum likelihood estimator $\hat{\beta}$ has bias of asymptotic order $O(n^{-1})$ where n is the sample size or some other measure of how information accumulates for the parameters of the model. This means that the bias of the maximum likelihood estimator vanishes as $n \to \infty$. Nevertheless, in practice the bias of $\hat{\beta}$ may be considerable for small or moderate values of n.

An approach to the correction of the bias of the maximum likelihood estimator is to define a bias-corrected estimator $\tilde{\beta} = \hat{\beta} - b(\hat{\beta})$, where $b(\hat{\beta})$ is the $O(n^{-1})$ term in the asymptotic expansion of the bias of the maximum likelihood estimator. It may be shown that $\tilde{\beta}$ has bias of asymptotic order $O(n^{-2})$ [see, for example, 11]. An extensive literature has been devoted to obtaining analytical expressions for $b(\beta)$ and studying the properties of the bias-corrected estimator, especially for classes of models where the bias of $\hat{\beta}$ is large enough to affect inferences appreciably. Characteristic examples of such studies are Cox and Snell [9], Schaefer [29], Gart et al. [15], Cook et al. [4], Cordeiro and McCullagh [6], Breslow and Lin [2], Lin and Breslow [20], Cordeiro and Vasconcellos [7] and Cordeiro and Toyama Udo [5].

An alternative family of estimators β^* with $O(n^{-2})$ bias was developed in Firth [14]. These estimators differ from the bias-corrected estimator $\tilde{\beta}$ in that they are not computed directly from the maximum likelihood estimator. The latter fact has motivated the study and use of the bias-reduced estimator β^* instead of $\tilde{\beta}$ [for example, 3, 16, 17, 19, 22, 25, 28, 32], especially in models where there is a positive probability that the maximum likelihood estimate is on the boundary of the parameter space. Leading examples are log-linear, logit-linear and similar models for counts, where the bias-corrected estimator is undefined whenever the maximum likelihood estimate has one or more infinite components [see, for example 1, for multinomial response models]. The estimator β^* results from the solution of a set of adjusted score equations, and hence in all but the simplest cases an iterative scheme needs to be employed to obtain the

bias-reduced estimate. For some specific families of models efficient estimation schemes have been developed by exploiting the specific structure of the adjustments. For example, for generalized linear models Kosmidis and Firth [19] suggest an iterative scheme that operates through appropriate adjustment of the maximum likelihood "working observations" [see also, 13], and for the particular case of binomial regression models Kosmidis [18] develops an appealing iterative scheme based on iterative adjustment of the binomial counts. More generally, however, the special structure needed for the existence of such iterative adjustment schemes is absent.

In the current paper a generic procedure for obtaining the bias-reduced estimate is developed. The procedure directly depends on $b(\beta)$ and hence it can be easily implemented for all the models for which $b(\beta)$ has already been obtained in the literature. Furthermore, as will be shown, for certain prominent members of the family of bias-reduced estimators the algorithm provides a unified computational framework for bias correction and bias reduction.

The new algorithm is then tested through an application to nonlinear regression with beta-distributed responses, a situation in which bias correction of the maximum likelihood estimator has received considerable recent attention in the literature. In addition to demonstrating the effectiveness of the algorithm developed here, a thorough numerical study reveals some errors in the recent literature on such models. The design and analysis of the simulation experiment conducted to detect such errors have special features associated with the large-sample behaviour of bias and variance, and form a template for the numerical study of asymptotic properties more generally.

2. Bias reduction via adjusted score functions

Firth [14] showed that an estimator with $O(n^{-2})$ bias may be obtained through the solution of an adjusted score equation in the general form

$$S^*(\beta) = S(\beta) + A(\beta) = 0\,, \tag{2.1}$$

where $A(\beta)$, suitably chosen, is $O_p(1)$ in magnitude as $n \to \infty$. Firth [14] described two specific instances of the general bias-reducing adjustment A, denoted by $A^{(E)}$ and $A^{(O)}$, based respectively on the expected and observed information matrix. The components of these two alternatives are given by

$$A_t^{(E)}(\beta) = \frac{1}{2}\,\mathrm{tr}\left[\{F(\beta)\}^{-1}\{P_t(\beta) + Q_t(\beta)\}\right] \quad (t = 1, \ldots, q)\,, \tag{2.2}$$

and

$$A^{(O)}(\beta) = I(\beta)\{F(\beta)\}^{-1}A^{(E)}(\beta)\,,$$

where $F(\beta) = E_\beta\{I(\beta)\}$ is the expected information matrix and $P_t(\beta) = E_\beta\{S(\beta)S(\beta)^T S_t(\beta)\}$ and $Q_t(\beta) = E_\beta\{-I(\beta)S_t(\beta)\}$ are higher order joint null moments of log-likelihood derivatives.

Kosmidis and Firth [19] gave a more general family of bias-reducing adjustments to the score vector. The general adjustment is of the form

$$A(\beta) = -\{G(\beta) + R(\beta)\}\, b(\beta)\,, \tag{2.3}$$

where $G(\beta)$ is either $F(\beta)$ or $I(\beta)$ or some other matrix with expectation $F(\beta)$, and $R(\beta)$ is a $q \times q$ matrix with expectation of order $O(n^{1/2})$. The vector

$$b(\beta) = -\{F(\beta)\}^{-1} A^{(E)}(\beta) \tag{2.4}$$

is the $O(n^{-1})$ asymptotic bias. It is immediately apparent that if $G(\beta) = F(\beta)$ with $R(\beta) = 0$ then the $A^{(E)}$ adjustment results, and if $G(\beta) = I(\beta)$ with $R(\beta) = 0$ the $A^{(O)}$ adjustment results.

3. Bias reduction as iterated bias correction

A full Newton-Raphson iteration for obtaining the bias-reduced estimate would require the evaluation of the matrix $I(\beta) + \nabla_\beta^T A(\beta)$. Even for relatively simple models a closed form expression for $\nabla_\beta^T A(\beta)$ requires cumbersome algebra and, depending on the complexity of the resultant expression, may also be difficult to implement. For this reason, the following *quasi* Newton-Raphson iteration is proposed:

$$\beta^{(j+1)} = \beta^{(j)} + \left\{ I\left(\beta^{(j)}\right) \right\}^{-1} S^*\left(\beta^{(j)}\right)\,, \tag{3.1}$$

where $\beta^{(j)}$ is the candidate value for β^* at the jth iteration. An alternative to the above iteration is the modified Fisher scoring iteration proposed in Kosmidis and Firth [19] in the specific context of generalized nonlinear models. The key difference between (3.1) and the iteration proposed in Kosmidis and Firth [19] is the use of $F(\beta)$ instead of $I(\beta)$ for calculation of the direction. Either iteration may be used but (3.1) seems closer in spirit to a Newton-Raphson iteration; because of the omission of the term $\nabla_\beta^T A(\beta)$, though, the convergence rate is generally linear instead of quadratic.

By substituting (2.1) and (2.3) into (3.1), the iteration may be re-expressed in the form

$$\begin{aligned}\beta^{(j+1)} = \beta^{(j)} &+ \left\{ I\left(\beta^{(j)}\right) \right\}^{-1} S\left(\beta^{(j)}\right) \\ &- \left\{ I\left(\beta^{(j)}\right) \right\}^{-1} \left\{ G\left(\beta^{(j)}\right) + R\left(\beta^{(j)}\right) \right\} b\left(\beta^{(j)}\right).\end{aligned} \tag{3.2}$$

Note that the first two terms on the right hand side of the above expression correspond to a Newton-Raphson iteration for maximizing the log-likelihood and hence (3.2) may be re-expressed as

$$\beta^{(j+1)} = \hat{\beta}^{(j+1)} - \left\{ I\left(\beta^{(j)}\right) \right\}^{-1} \left\{ G\left(\beta^{(j)}\right) + R\left(\beta^{(j)}\right) \right\} b\left(\beta^{(j)}\right)\,, \tag{3.3}$$

where $\hat{\beta}^{(j+1)}$ is the candidate value for the maximum likelihood estimate that would be obtained by taking a single Newton-Raphson step from $\beta^{(j)}$.

Convergence or otherwise of the above iteration depends on the properties of the specific model under consideration. Nevertheless, assuming that it does converge, it is apparent from (3.1) that at convergence iteration (3.3) gives the solution to the equations $S^*(\beta) = 0$.

In regular statistical models the maximum likelihood estimator differs from the bias-reduced estimator by a quantity of order $O(n^{-1})$. Typically, then, the maximum likelihood estimate is a good starting value for the iterative scheme, provided that none of its components is on the boundary of the parameter space.

In the special case of bias reduction based on the $A^{(O)}$ adjustment, iteration (3.3) can be usefully re-expressed as simply

$$\beta^{(j+1)} = \hat{\beta}^{(j+1)} - b\left(\beta^{(j)}\right). \tag{3.4}$$

This has a rather appealing interpretation: at each step, the next candidate value of the maximum likelihood estimate is corrected by subtracting the $O(n^{-1})$ bias evaluated at the current value of the bias-reduced estimate. Hence, if $\beta^{(0)} = \hat{\beta}$, the first step of the proposed scheme delivers the bias-corrected maximum likelihood estimate; and iterating until convergence yields the bias-reduced estimate based on adjustment $A^{(O)}$.

For bias reduction based on the $A^{(E)}$ adjustment, iterated bias correction as in (3.4) can still be used, but with the symbols in (3.4) having a different meaning. In that case $\hat{\beta}^{(j+1)}$ represents the candidate value for the maximum likelihood estimate obtained by taking a single Fisher-scoring step from $\beta^{(j)}$, instead of a Newton-Raphson step. This provides a useful new interpretation of the modified Fisher scoring iteration that was suggested in Kosmidis and Firth [19].

4. Example: Beta regression

4.1. Beta generalized linear model

As an illustrative application of the bias-reduction algorithm, consider the case of a generalized linear model with Beta-distributed responses [for example, 10, 12]. Suppose that $Y_1, \ldots, Y_n$ are independent Beta-distributed random variables, the density of Y_i being

$$f_i(y) = \frac{\Gamma(\delta_i + \epsilon_i)}{\Gamma(\delta_i)\Gamma(\epsilon_i)} y^{\delta_i - 1}(1-y)^{\epsilon_i - 1} \qquad (0 < y_i < 1;\ \delta_i > 0,\ \epsilon_i > 0; i = 1, \ldots, n).$$

Then

$$E(Y_i) = \frac{\delta_i}{\delta_i + \epsilon_i} = \mu_i \quad \text{and} \quad \text{var}(Y_i) = \frac{\mu_i(1-\mu_i)}{1 + \delta_i + \epsilon_i}.$$

For the purposes of the current paper it will be assumed that the precision quantities $\delta_i + \epsilon_i$ $(i = 1, \ldots, n)$ are all equal, and we will write $1/(1 + \delta_i + \epsilon_i) = \sigma^2 < 1$. The response dispersion, relative to the common variance function $V(\mu_i) = \mu_i(1-\mu_i)$, is thus assumed constant:

$$\frac{\text{var}(Y_i)}{V(\mu_i)} = \sigma^2.$$

In some applications this constant-dispersion assumption might need to be relaxed, for example, as in Smithson and Verkuilen [31] or Simas et al. [30].

The dependence of the response mean μ_i upon a p-vector x_i of covariate values is commonly modeled through a link function $g(.)$ to a linear predictor η_i ($i = 1, \ldots, n$). Since $\mu_i \in (0, 1)$, the obvious candidate link functions in this context are inverse cumulative distribution functions (logit, probit and suchlike). The assumed relationship between the expected response and the covariate values is then

$$g(\mu_i) = \eta_i = \sum_{t=1}^{p} \gamma_t x_{it}\,, \qquad (i = 1, \ldots, n)\,,$$

and so the parameters of this model are the vector of regression coefficients $\gamma = (\gamma_1, \ldots, \gamma_p)$ and the dispersion parameter σ^2. Ferrari and Cribari-Neto [12] parameterized the model in terms of the *precision* parameter $\phi = 1/\sigma^2 - 1$ and that representation will also be used here, with $\beta = (\gamma_1, \ldots, \gamma_p, \phi)$ being the full vector of model parameters.

4.2. Bias reduction

For Beta regression models, Ospina et al. [23] express the vector $b(\beta)$ of the first-order biases of $\hat{\beta}$ as the estimator of the regression coefficients of an appropriately weighted linear regression. A similar result is obtained in Simas et al. [30] for nonlinear Beta regression models with dispersion covariates. Despite the analytical elegance of such expressions for $b(\beta)$, patterned after similar results for generalized linear models in Cordeiro and McCullagh [6], they seem to offer no benefit in terms of efficient implementation. In what follows a more direct approach is taken. For general families of models, equations (2.2), (2.4) and (3.3) suggest that bias correction and any bias-reduction method can be directly implemented if $F(\beta)$, $I(\beta)$, $P_t(\beta) + Q_t(\beta)$ ($t = 1, \ldots, q$) and the required form of $G(\beta) + R(\beta)$ are all available in closed form; the matrix inversions and multiplications necessary for implementation can then be done numerically.

For the Beta regression model the log-likelihood can be written in the form

$$l(\beta) = \sum_{i=1}^{n} \left[\phi\mu_i(U_i - Z_i) + \phi Z_i - \log\Gamma(\phi\mu_i) - \log\Gamma\{\phi(1-\mu_i)\} + \log\Gamma(\phi)\right]\,,$$

where $\mu_i = g^{-1}(\eta_i)$, and $U_i = \log Y_i$ and $Z_i = \log(1 - Y_i)$ are the sufficient statistics for the Beta distribution with natural parameters δ_i and ϵ_i, respectively ($i = 1, \ldots, n$). Direct differentiation of $l(\beta)$ with respect to ϕ and γ gives that

$$S(\beta) = \begin{bmatrix} \nabla_\gamma l(\beta) \\ \partial l(\beta)/\partial\phi \end{bmatrix} = \begin{bmatrix} \phi X^T D(\bar{U} - \bar{Z}) \\ 1_n^T M(\bar{U} - \bar{Z}) + 1_n^T \bar{Z} \end{bmatrix}\,, \tag{4.1}$$

where the dependence on β of the quantities appearing in the right hand side has been suppressed for notational convenience. Here, 1_n is an n-vector of ones, X is the $n \times p$ matrix with x_i as its ith row, $D = \mathrm{diag}\{d_1, \ldots, d_n\}$ with $d_i = \mathrm{d}\mu_i/\mathrm{d}\eta_i$

$(i = 1, \ldots, n)$, and $M = \mathrm{diag}\{\mu_1, \ldots, \mu_n\}$. Furthermore, $\bar{U}$ and $\bar{Z}$ are the n-vectors with ith components the centered sufficient statistics $\bar{U}_i = U_i - \lambda_i$ and $\bar{Z}_i = Z_i - \xi_i$, respectively, where $\lambda_i = E(U_i) = \psi^{(0)}(\phi\mu_i) - \psi^{(0)}(\phi)$ and $\xi_i = E(Z_i) = \psi^{(0)}\{\phi(1-\mu_i)\} - \psi^{(0)}(\phi)$ $(i = 1, \ldots, n)$. The function $\psi^{(r)}(k) = \mathrm{d}^{r+1} \log\Gamma(k)/\mathrm{d}k^{r+1}$ is the polygamma function of order r $(r = 0, 1, \ldots)$.

Expressing all likelihood derivatives in terms of the sufficient statistics U_i and Z_i facilitates the calculation of $P_t + Q_t$ because the derivation of higher-order joint cumulants of Z_i and U_i is merely a simple algebraic exercise; any joint cumulant of Z_i and U_i results from appropriate-order partial differentiation of the cumulant transform of the Beta distribution with respect to the natural parameters δ_i and ϵ_i.

Further differentiation of $l(\beta)$ gives the observed information on β,

$$I(\beta) = F(\beta) - \begin{bmatrix} \phi X^T D' \,\mathrm{diag}\{\bar{U} - \bar{Z}\} X & X^T D(\bar{U} - \bar{Z}) \\ (\bar{U} - \bar{Z})^T D X & 0 \end{bmatrix}, \qquad (4.2)$$

where

$$F(\beta) = \begin{bmatrix} \phi^2 X^T D K_2 D X & \phi X^T D\,(M K_2 - \Psi_1) \\ \phi\,(M K_2 - \Psi_1)\, D X & 1_n^T \{M K_2 M + (1 - 2M)\Psi_1\}\, 1_n - n\psi^{(1)}(\phi) \end{bmatrix} \qquad (4.3)$$

is immediately recognised to be the expected information on β, since the expectation of the second summand in the right hand side of (4.2) is zero. Here, $D' = \mathrm{diag}\{d'_1, \ldots, d'_n\}$ with $d'_i = \mathrm{d}^2\mu_i/\mathrm{d}\eta_i^2$, $K_2 = \mathrm{diag}\{\mathrm{var}(\bar{U}_1 - \bar{Z}_1), \ldots, \mathrm{var}\left(\bar{U}_n - \bar{Z}_n\right)\}$, where $\mathrm{var}\left(\bar{U}_i - \bar{Z}_i\right) = \psi^{(1)}(\phi\mu_i) + \psi^{(1)}\{\phi(1-\mu_i)\}$ and

$$\Psi_r = \mathrm{diag}\left\{\psi^{(r)}\{\phi(1-\mu_1)\}, \ldots, \psi^{(r)}\{\phi(1-\mu_n)\}\right\} \quad (r = 0, 1, \ldots; i = 1, \ldots, n).$$

A careful examination of expressions (4.1) and (4.2) reveals that

$$P_t(\beta) + Q_t(\beta) = E_\beta\{S(\beta)S(\beta)^T S_t(\beta)\} + E_\beta\{-I(\beta)S_t(\beta)\} \quad (t = 1, \ldots, p+1),$$

depends on $\mathrm{var}\left(\bar{U}_i - \bar{Z}_i\right)$, on the cumulants $E\left(\bar{Z}_i^3\right)$, $E\left\{(\bar{U}_i - \bar{Z}_i)^3\right\}$ and on the covariances $E\left\{(\bar{U}_i - \bar{Z}_i)\bar{Z}_i\right\}$, $E\left\{(\bar{U}_i - \bar{Z}_i)^2\bar{Z}_i\right\}$ and $E\left\{(\bar{U}_i - \bar{Z}_i)\bar{Z}_i^2\right\}$ $(i = 1, \ldots, n)$. Re-expressing the above expectations as sums of joint cumulants of U_i and Z_i, direct differentiation of the cumulant transform of the Beta distribution with respect to δ_i and ϵ_i gives

$$\begin{aligned} E\left(\bar{Z}_i^3\right) &= \psi^{(2)}\{\phi(1-\mu_i)\} - \psi^{(2)}(\phi), \\ E\left\{(\bar{U}_i - \bar{Z}_i)^3\right\} &= \psi^{(2)}(\phi\mu_i) - \psi^{(2)}\{\phi(1-\mu_i)\}, \\ E\left\{(\bar{U}_i - \bar{Z}_i)\bar{Z}_i\right\} &= -\psi^{(1)}\{\phi(1-\mu_i)\}, \\ E\left\{(\bar{U}_i - \bar{Z}_i)^2\bar{Z}_i\right\} &= \psi^{(2)}\{\phi(1-\mu_i)\}, \\ E\left\{(\bar{U}_i - \bar{Z}_i)\bar{Z}_i^2\right\} &= -\psi^{(2)}\{\phi(1-\mu_i)\} \quad (i = 1, \ldots, n). \end{aligned}$$

Some algebra then gives

$$P_t + Q_t = \phi \begin{bmatrix} V_{\gamma\gamma} & V_{\gamma\phi} \\ V_{\gamma\phi}^T & V_{\phi\phi} \end{bmatrix} \quad (t = 1, \ldots, p) \qquad (4.4)$$

and

$$P_{p+1} + Q_{p+1} = \phi \begin{bmatrix} W_{\gamma\gamma} & W_{\gamma\phi} \\ W_{\gamma\phi}^T & W_{\phi\phi} \end{bmatrix}, \tag{4.5}$$

where

$$\begin{aligned} V_{\gamma\gamma} &= \phi^2 X^T D \left(\phi D K_3 D + D' K_2\right) T_t X \,, \\ V_{\gamma\phi} &= \phi X^T D \left(\phi M K_3 + \phi \Psi_2 + K_2\right) D T_t 1_n \,, \\ V_{\phi\phi} &= \phi 1_n^T D \left\{M K_3 M + (2M-1)\,\Psi_2\right\} T_t 1_n \,, \end{aligned}$$

and

$$\begin{aligned} W_{\gamma\gamma} &= \phi X^T \left\{\phi D \left(M K_3 + \Psi_2\right) D + D' \left(M K_2 - \Psi_1\right)\right\} X \,, \\ W_{\gamma\phi} &= X^T D \left\{\phi M K_3 M + \phi \left(2M-1\right) \Psi_2 + M K_2 - \Psi_1\right\} 1_n \,, \\ W_{\phi\phi} &= 1_n^T \left\{M M K_3 M + \left(3MM - 3M + 1\right) \Psi_2\right\} 1_n - n\psi^{(2)}(\phi) \,, \end{aligned}$$

with $K_3 = \operatorname{diag}\left\{E\{(\bar{U}_1 - \bar{Z}_1)^3\}, \ldots, E\{(\bar{U}_n - \bar{Z}_n)^3\}\right\}$ and $T_t = \operatorname{diag}\{x_{1t}, \ldots, x_{nt}\}$.

Iteration (3.3) can now be implemented by using expressions (4.1), (4.2), (4.3), (4.4) and (4.5) and the chosen matrices $G(\beta)$ and $R(\beta)$.

4.3. Numerical study

As a partial check on the correctness of the bias-reduction algorithm a model with

$$\log \frac{\mu_i}{1-\mu_i} = \alpha + \sum_{t=1}^{9} \gamma_t s_{it} + \delta t_i \quad (i = 1, \ldots, n) \tag{4.6}$$

is fitted to the $n = 32$ observations of the gasoline yield data of Prater [26]. The response variable is the proportion of crude oil converted to gasoline after distillation and fractionation, and $s_{i1}, \ldots, s_{i9}$ are the values of 9 binary covariates which represent the 10 distinct experimental settings in the data set for the triplet i) temperature in degrees Fahrenheit at which 10% of crude oil has vaporized, ii) crude oil gravity, and iii) vapor pressure of crude oil ($i = 1, \ldots, n$). Lastly, t_i is the temperature in degrees Fahrenheit at which all gasoline has vaporized for the ith observation ($i = 1, \ldots, n$). The same model was also used for illustration in Ospina et al. [23].

The parameters $\beta = (\alpha, \gamma_1, \ldots, \gamma_9, \delta, \phi)$ are estimated using maximum likelihood, bias correction and bias reduction with $A^{(O)}$ and $A^{(E)}$ adjustments using the expressions (4.4) and (4.5) to implement iteration (3.3). The results for the actual data are shown in Table 1, while Table 2 presents the results of a simulation study based on this example. Some remarks on these results follow.

Remark 1: Checking correctness of the implementation

The bias-corrected estimates and the bias-reduced estimates with $A^{(E)}$ adjustments in Table 1 differ appreciably from the corresponding values reported in

TABLE 1

Estimates of the parameters of model (4.6) using maximum likelihood, bias correction and bias reduction with $A^{(E)}$ and $A^{(O)}$ adjustments. The parenthesized quantities are the corresponding estimated standard errors based on the expected information matrix

	Maximum likelihood		Bias correction		Bias reduction using $A^{(E)}$		Bias reduction using $A^{(O)}$	
α	−6.15957	(0.18232)	−6.14837	(0.23595)	−6.14171	(0.23588)	−6.14005	(0.23591)
γ_1	1.72773	(0.10123)	1.72484	(0.13107)	1.72325	(0.13106)	1.72273	(0.13108)
γ_2	1.32260	(0.11790)	1.32009	(0.15260)	1.31860	(0.15257)	1.31823	(0.15260)
γ_3	1.57231	(0.11610)	1.56928	(0.15030)	1.56734	(0.15028)	1.56699	(0.15031)
γ_4	1.05971	(0.10236)	1.05788	(0.13251)	1.05677	(0.13249)	1.05651	(0.13251)
γ_5	1.13375	(0.10352)	1.13165	(0.13404)	1.13024	(0.13403)	1.13003	(0.13405)
γ_6	1.04016	(0.10604)	1.03829	(0.13729)	1.03714	(0.13727)	1.03689	(0.13729)
γ_7	0.54369	(0.10913)	0.54309	(0.14119)	0.54242	(0.14116)	0.54253	(0.14118)
γ_8	0.49590	(0.10893)	0.49518	(0.14099)	0.49446	(0.14096)	0.49454	(0.14099)
γ_9	0.38579	(0.11859)	0.38502	(0.15353)	0.38459	(0.15351)	0.38446	(0.15354)
δ	0.01097	(0.00041)	0.01094	(0.00053)	0.01093	(0.00053)	0.01093	(0.00053)
ϕ	440.27839	(110.02562)	261.20610	(65.25866)	261.03777	(65.21640)	260.90168	(65.18234)

TABLE 2

Estimated biases $\times 10^2$ (T1) and mean squared errors $\times 10^2$ (T2) for the parameters of model (4.6) based on a simulation of size 2×10^6 from the maximum likelihood fit. The parenthesised quantities are estimates of the corresponding simulation standard errors

		Maximum likelihood		Bias correction		Bias reduction using $A^{(E)}$		Bias reduction using $A^{(O)}$	
T1	α	−1.129	(0.013)	−0.389	(0.013)	0.114	(0.013)	0.182	(0.013)
	γ_1	0.293	(0.007)	0.101	(0.007)	−0.028	(0.007)	−0.046	(0.007)
	γ_2	0.253	(0.008)	0.087	(0.008)	−0.025	(0.008)	−0.040	(0.008)
	γ_3	0.313	(0.008)	0.112	(0.008)	−0.023	(0.008)	−0.042	(0.008)
	γ_4	0.194	(0.007)	0.073	(0.007)	−0.010	(0.007)	−0.021	(0.007)
	γ_5	0.216	(0.007)	0.076	(0.007)	−0.018	(0.007)	−0.031	(0.007)
	γ_6	0.198	(0.008)	0.074	(0.007)	−0.009	(0.007)	−0.021	(0.007)
	γ_7	0.059	(0.008)	0.019	(0.008)	−0.008	(0.008)	−0.012	(0.008)
	γ_8	0.074	(0.008)	0.026	(0.008)	−0.006	(0.008)	−0.010	(0.008)
	γ_9	0.071	(0.008)	0.020	(0.008)	−0.014	(0.008)	−0.019	(0.008)
	δ	0.002	(< 0.001)	0.001	(< 0.001)	< 0.001	(< 0.001)	< 0.001	(< 0.001)
	ϕ	30162.5	(18.0)	1.8	(10.7)	−18.2	(10.7)	−30.1	(10.7)
T2	α	3.355	(0.003)	3.335	(0.003)	3.329	(0.003)	3.328	(0.003)
	γ_1	1.030	(0.001)	1.027	(0.001)	1.025	(0.001)	1.025	(0.001)
	γ_2	1.397	(0.001)	1.392	(0.001)	1.390	(0.001)	1.389	(0.001)
	γ_3	1.353	(0.001)	1.348	(0.001)	1.346	(0.001)	1.345	(0.001)
	γ_4	1.053	(0.001)	1.049	(0.001)	1.047	(0.001)	1.047	(0.001)
	γ_5	1.076	(0.001)	1.073	(0.001)	1.071	(0.001)	1.071	(0.001)
	γ_6	1.128	(0.001)	1.125	(0.001)	1.123	(0.001)	1.123	(0.001)
	γ_7	1.201	(0.001)	1.197	(0.001)	1.194	(0.001)	1.194	(0.001)
	γ_8	1.191	(0.001)	1.187	(0.001)	1.185	(0.001)	1.184	(0.001)
	γ_9	1.413	(0.001)	1.409	(0.001)	1.406	(0.001)	1.406	(0.001)
	δ	< 0.001	(< 0.001)	< 0.001	(< 0.001)	< 0.001	(< 0.001)	< 0.001	(< 0.001)
	ϕ	15570353.7	(20413.3)	2281835.6	(4257.5)	2281844.7	(4255.0)	2281881.7	(4253.5)

Ospina et al. [23, Table 6] (labeled "CBCE" and "PBCE", respectively, therein) while the maximum likelihood estimates are the same, at least to five significant digits. After some investigation it was found that the differences arise from two distinct sources.

The reason for the fairly substantial difference in the reported bias-reduced estimates is an elementary but serious error: in equation (3.5) of Ospina et al. [23] the sign of the adjustment term is the opposite of that suggested in Firth [14]. As a result of this, and as is also apparent in the simulation studies reported in Ospina et al. [23], the estimator labeled "PBCE" therein approximately doubles the bias of the maximum likelihood estimator instead of eliminating it. The same mistake was made also in equation (12) of Simas et al. [30], with the same unfortunate consequence.

That mistake does not, however, account for the differences seen also between the reported bias-corrected estimates here and in Table 6 of Ospina et al. [23]. Those differences, while relatively small, are still too large to be attributed to presentational rounding error: at least one of the two implementations of the $O(n^{-1})$ bias term $b(\beta)$ must therefore be incorrect. A brief account follows of an extensive simulation exercise designed to determine which of the two reported sets of bias-corrected estimates is incorrect.

The bias of the maximum likelihood estimator can be written in the form $B(\beta)/n + O(n^{-2})$, where $b(\beta) = B(\beta)/n$ is as in (2.4). The following simulation experiment relies on the fact that as n increases the bias of $\hat{\beta}$ is almost completely determined by the value of $B(\beta)$.

Let X be the $n \times p$ model matrix for (4.6) and denote by $Z(j)$ a $n_j \times p$ model matrix with $n_j = nj$, whose rows result from repeating j times each row of X ($j = 1, 2, \ldots; n = 32; p = 11$). Then the bias of the maximum likelihood estimator $\hat{\beta}^{[j]}$ for model matrix $Z(j)$ is $B(\beta)/n_j + O(n_j^{-2})$ ($j = 1, 2, \ldots$). The alternative values of the vector $B(\hat{\beta})$ are calculated as

$$B^{(\mathrm{cur})}(\hat{\beta}) = n\left(\hat{\beta} - \tilde{\beta}^{(\mathrm{cur})}\right),$$
$$B^{(\mathrm{Osp})}(\hat{\beta}) = n\left(\hat{\beta} - \tilde{\beta}^{(\mathrm{Osp})}\right),$$

using the bias-corrected estimates $\tilde{\beta}^{(\mathrm{cur})}$ in Table 1 for the current implementation, and the bias-corrected estimates $\tilde{\beta}^{(\mathrm{Osp})}$ reported in Ospina et al. [23, Table 6], respectively.

For any $j \in \{1, 2, \ldots\}$, consider now simulating some number N_j of samples from the model, using the maximum likelihood estimates in Table 1 as the true value for β. The bias of $\hat{\beta}^{[j]}$ can then be estimated using maximum likelihood fits to those samples and, after multiplication by n_j, can be compared to $B^{(\mathrm{cur})}(\hat{\beta})$ and $B^{(\mathrm{Osp})}(\hat{\beta})$. The standard error of the estimator of n_j times the bias of $\hat{\beta}^{[j]}$ is $O\left(\sqrt{n_j/N_j}\right)$. Hence, the order of that standard error can be stabilized to $O(1/\sqrt{N})$ by choosing $N_j = Nn_j$, for some sufficiently large N.

The plots in Figure 1 give the estimated values of the components of n_j times the bias of $\hat{\beta}^{[j]}$ that correspond to $\alpha, \gamma_1, \ldots, \gamma_9$, and δ, for $N = 5000$ and $j =$

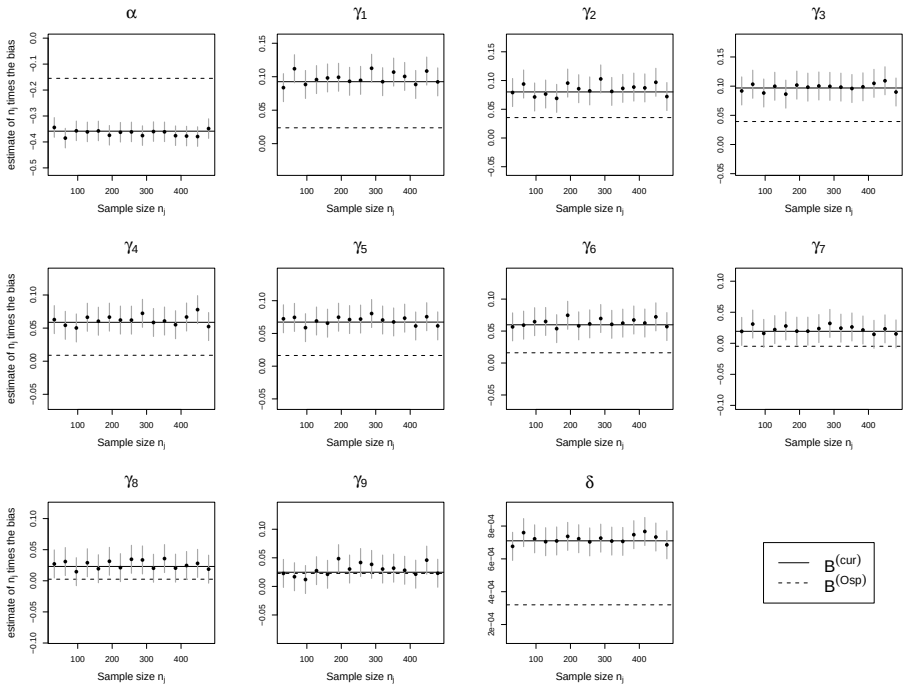

FIG 1. *Simulation-based estimates of the components of n_j times the bias of $\hat{\beta}^{[j]}$ for $j = 1, \ldots, 15$. The solid and dashed horizontal lines are the components of $B^{(\mathrm{cur})}(\hat{\beta})$ and $B^{(\mathrm{Osp})}(\hat{\beta})$, respectively. The vertical lines are approximate 99% confidence intervals.*

$1, \ldots, 15$. To provide an indication of the uncertainty due to simulation error, the estimates are accompanied by 99% confidence intervals based on asymptotic normality. For every parameter the estimate is very close to $B^{(\mathrm{cur})}(\hat{\beta})$. This experiment provides clear evidence of either an algebraic or implementation error in Ospina et al. [23] as far as $b(\beta)$ is concerned, at least for the parameters $\alpha, \gamma_1, \ldots, \gamma_6$ and δ; the differences found here are too large to be accounted for by the presentational rounding to 5 decimal places in Table 6 of Ospina et al. [23].

Remark 2: Impact of bias in this example

The results in Table 2 are as expected: the bias found in the maximum likelihood estimator $\hat{\beta}$ is reduced by all three of the alternative methods which aim to improve the bias. By far the most important effect of bias reduction here is to reduce substantially the estimated value of ϕ; in regard to estimation of the regression parameters (α, γ, δ), on the other hand, the bias in this example is rather slight and is therefore of no real consequence. The smaller value of ϕ does, however, result in an appreciable increase in the estimated standard errors for all the regression parameter estimates. A further simulation exercise, not reported

in detail here, was conducted to check the accuracy of the asymptotic standard errors reported in Table 1: it was found that those standard errors all agree with simulation-based standard errors to at least 3 decimal places. Because of the large bias in the estimated precision parameter $\hat{\phi}$, the usual standard errors based on the maximum likelihood analysis are systematically too small. The principal effect of bias reduction in this example, then, is to produce more realistic standard errors for the estimates of all the parameters.

It should be noted that the estimated value of ϕ in this example, even after reduction of the bias, is quite large: that is, the residual dispersion in the model is quite small. This accounts for the rather small bias found in the maximum likelihood estimator for the regression parameters. In a different situation with more substantial residual dispersion present, biases in the regression estimates themselves (i.e., not only in their standard errors) would likely become more important.

Remark 3: Parameterization

In general, bias reduction will typically make most sense when applied to estimators whose distribution is approximately symmetric, since it will then most often improve the accuracy of inferences made when using first-order asymptotic normal approximations. In the present application, the distributions for all parameters except ϕ are close to symmetric; $\hat{\phi}$ exhibits substantial positive skewness, as is often found in the estimation of positive-valued parameters.

In this model it seems preferable, then, to consider bias reduction instead for a transformed version of ϕ, the most obvious candidate being $\log\phi$. The distribution of $\log\hat{\phi}$ is much closer to being symmetric than that of $\hat{\phi}$, a fact confirmed by graphical summaries (not presented here) of the simulation experiment underlying Table 2.

Consider a general re-parameterization from β to $\omega = (\alpha, \gamma_1, \ldots, \gamma_9, \delta, \zeta)$, with $\zeta = h(\phi)$ for some appropriate function $h : \Re^+ \to H \subset \Re$. Because the maximum likelihood estimator is equivariant under re-parameterization, the components of the bias vector — and hence of the vector of first-order biases — corresponding to α, $\gamma_1, \ldots, \gamma_9$ and δ will be the same in both the ω and β parameterizations. Assuming that $h(.)$ is at least three times differentiable, and using the consistency of the maximum likelihood estimator $\hat{\phi}$ of ϕ, $\hat{\zeta} - \zeta$ admits the expansion

$$\hat{\zeta}-\zeta = h(\hat{\phi})-h(\phi) = (\hat{\phi}-\phi)\frac{\mathrm{d}h(\phi)}{\mathrm{d}\phi}+\frac{1}{2}(\hat{\phi}-\phi)^2\frac{\mathrm{d}^2h(\phi)}{\mathrm{d}\phi^2}+\frac{1}{6}(\hat{\phi}-\phi)^3\frac{\mathrm{d}^3h(\phi)}{\mathrm{d}\phi^3}+O_p(n^{-2})\,. \tag{4.7}$$

Noting that $E[(\hat{\phi}-\phi)^r]$ is $O(n^{-(r+1)/2})$ if r is odd and $O(n^{-r/2})$ if r is even [see, for example 24, Section 9.4 for the asymptotic expansion of the maximum likelihood estimator], taking expectations in both sides of (4.7) gives that the bias of $\hat{\zeta}$ can be written as

$$E(\hat{\zeta}-\zeta) = b_\phi(\beta)\frac{\mathrm{d}h(\phi)}{\mathrm{d}\phi} + \frac{1}{2}F^{-1}_{\phi\phi}(\beta)\frac{\mathrm{d}^2h(\phi)}{\mathrm{d}\phi^2} + O(n^{-2})\,. \tag{4.8}$$

Here $b_\phi(\beta)$ and $F^{-1}_{\phi\phi}(\beta)$ denote the components of $b(\beta)$ and $\{F(\beta)\}^{-1}$ which correspond to ϕ. Furthermore, the expected information matrix on ω is $F^*(\omega) = J(\omega)F(\beta(\omega))J(\omega)$, where $\beta(\omega) = (\alpha, \gamma_1, \ldots, \gamma_9, \delta, h^{-1}(\zeta))$ and

$$J(\omega) = \begin{bmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & \ldots & 0 & \mathrm{d}h^{-1}(\zeta)/\mathrm{d}\zeta \end{bmatrix},$$

with $h^{-1}(.)$ the inverse of the function $h(.)$.

Expression (4.8) can be used to obtain the first-order bias of $h(\hat{\phi})$ for every $h(.)$, by merely using the first-order bias of $\hat{\phi}$, the inverse of $F(\beta)$ and the first two derivatives of $h(.)$. Also, the bias-reduced estimate of ω based on $A^{(E)}$ adjustments can be obtained by using iteration (3.4) with

$$\hat{\omega}^{(j+1)} = \omega^{(j)} + \left\{F^*\left(\omega^{(j)}\right)\right\}^{-1} S^*\left(\omega^{(j)}\right),$$

where $S^*(\omega) = J(\omega)S(\beta(\omega))$ is the score vector in the ω parameterization. While the maximum likelihood and bias-corrected estimates for $\alpha, \gamma_1, \ldots, \gamma_9$ and δ will be exactly the same in both the β and ω parameterizations, the corresponding bias-reduced estimates will generally differ slightly between parameterizations.

Table 3 gives the maximum likelihood, bias-corrected and bias-reduced estimates of ω when $h(\phi) = \log\phi$, along with estimated standard errors based on $F^*(\omega)$ evaluated at the corresponding estimates. The principal differences between Table 3 and Table 1 are in the implied estimates of ϕ, and consequently in the standard errors for estimates of the regression parameters $\alpha, \gamma_1, \ldots, \gamma_9$ and δ. For example, the bias-reduced estimate of ϕ from Table 3

TABLE 3

Estimates of the parameters of model (4.6) using maximum likelihood, bias correction and bias reduction based on $A^{(E)}$, for the re-parameterization with $\zeta = \log(\phi)$. In parentheses are the corresponding estimated standard errors based on the expected information matrix $F^*(\omega)$

	Maximum likelihood		Bias correction		Bias reduction using $A^{(E)}$	
α	−6.15957	(0.18232)	−6.14837	(0.21944)	−6.14259	(0.22998)
γ_1	1.72773	(0.10123)	1.72484	(0.12189)	1.72347	(0.12777)
γ_2	1.32260	(0.11790)	1.32009	(0.14193)	1.31880	(0.14875)
γ_3	1.57231	(0.11610)	1.56928	(0.13978)	1.56758	(0.14651)
γ_4	1.05971	(0.10236)	1.05788	(0.12323)	1.05691	(0.12917)
γ_5	1.13375	(0.10352)	1.13165	(0.12465)	1.13041	(0.13067)
γ_6	1.04016	(0.10604)	1.03829	(0.12767)	1.03729	(0.13383)
γ_7	0.54369	(0.10913)	0.54309	(0.13133)	0.54248	(0.13763)
γ_8	0.49590	(0.10893)	0.49518	(0.13112)	0.49453	(0.13743)
γ_9	0.38579	(0.11859)	0.38502	(0.14278)	0.38465	(0.14966)
δ	0.01097	(0.00041)	0.01094	(0.00050)	0.01093	(0.00052)
ζ	6.08741	(0.24990)	5.71191	(0.24986)	5.61608	(0.24984)

is exp(5.61608) = 274.8; this is slightly larger than the corresponding value 261.0 from Table 1, resulting in slightly smaller estimated standard errors for the regression parameters in Table 3.

5. Concluding remarks

The new algorithm developed here unifies various iterative methods that have been made available previously for specific models, and extends them to cover any new situation for which the $O(1/n)$ bias of the maximum likelihood estimator can be derived. The method was tested and demonstrated here in the context of beta-response nonlinear regression, and was found to perform robustly in all of the very large number of samples that were used in simulation studies.

The particular illustrative example presented here is just one of several beta-regression applications that the authors have worked through carefully, and the results were qualitatively the same in all of them. Bias in estimation of the regression parameters in such models is typically so small as to be of no consequence, at least when the precision parameter ϕ is not unreasonably small; but the standard errors in a maximum-likelihood analysis are systematically under-estimated, with the likely consequence that spuriously strong conclusions would often be drawn. Reducing the bias in the estimated precision parameter increases the estimated standard errors in such a way that they reflect better the true variability of the estimated regression parameters.

The calculations described here were all programmed in *R* [27], and the code is available on request from the first author.

Acknowledgments

The authors gratefully acknowledge the financial support of the UK Engineering and Physical Sciences Research Council for this work.

References

[1] Albert, A. and J. Anderson (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika 71*(1), 1–10. MR0738319

[2] Breslow, N. E. and X. Lin (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika 82*, 81–91. MR1332840

[3] Bull, S. B., C. Mak, and C. Greenwood (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis 39*, 57–74. MR1895558

[4] Cook, R. D., C.-L. Tsai, and B. C. Wei (1986). Bias in nonlinear regression. *Biometrika 73*, 615–623. MR0897853

[5] Cordeiro, G. and M. Toyama Udo (2008). Bias correction in generalized nonlinear models with dispersion covariates. *Communications in Statistics: Theory and Methods 37*(14), 2219–225. MR2526676

[6] Cordeiro, G. M. and P. McCullagh (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society, Series B: Methodological 53*(3), 629–643. MR1125720

[7] Cordeiro, G. M. and K. L. P. Vasconcellos (1997). Bias correction for a class of multivariate nonlinear regression models. *Statistics & Probability Letters 35*, 155–164. MR1483269

[8] Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. London: Chapman & Hall Ltd. MR0370837

[9] Cox, D. R. and E. J. Snell (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society, Series B: Methodological 30*, 248–275. MR0237052

[10] Cribari-Neto, F. and A. Zeileis (2010). Beta regression in R. *Journal of Statistical Software 34*(2), 1–24.

[11] Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion). *The Annals of Statistics 3*, 1189–1217. MR0428531

[12] Ferrari, S. and F. Cribari-Neto (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics 31*(7), 799–815. MR2095753

[13] Firth, D. (1992). Bias reduction, the Jeffreys prior and GLIM. In L. Fahrmeir, B. Francis, R. Gilchrist, and G. Tutz (Eds.), *Advances in GLIM and Statistical Modelling: Proceedings of the GLIM 92 Conference, Munich*, New York, pp. 91–100. Springer.

[14] Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*(1), 27–38. MR1225212

[15] Gart, J. J., H. M. Pettigrew, and D. G. Thomas (1985). The effect of bias, variance estimation, skewness and kurtosis of the empirical logit on weighted least squares analyses. *Biometrika 72*, 179–190.

[16] Heinze, G. and M. Schemper (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine 21*, 2409–2419.

[17] Heinze, G. and M. Schemper (2004). A solution to the problem of monotone likelihood in Cox regression. *Biometrics 57*, 114–119. MR1833296

[18] Kosmidis, I. (2009). On iterative adjustment of responses for the reduction of bias in binary regression models. Technical Report 09-36, CRiSM working paper series.

[19] Kosmidis, I. and D. Firth (2009). Bias reduction in exponential family nonlinear models. *Biometrika 96*(4), 793–804. MR2564491

[20] Lin, X. and N. E. Breslow (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association 91*, 1007–1016. MR1424603

[21] McCullagh, P. (1987). *Tensor Methods in Statistics*. London: Chapman and Hall. MR0907286

[22] Mehrabi, Y. and J. N. S. Matthews (1995). Likelihood-based methods for bias reduction in limiting dilution assays. *Biometrics 51*, 1543–1549.

[23] Ospina, R., F. Cribari-Neto, and K. L. Vasconcellos (2006). Improved point and interval estimation for a beta regression model. *Computational Statistics and Data Analysis 51*(2), 960 – 981. MR2297500

[24] Pace, L. and A. Salvan (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. London: World Scientific. MR1476674

[25] Pettitt, A. N., J. M. Kelly, and J. T. Gao (1998). Bias correction for censored data with exponential lifetimes. *Statistica Sinica 8*, 941–964. MR1651517

[26] Prater, N. H. (1956). Estimate gasoline yields from crudes. *Petroleum Refiner 35*, 236–238.

[27] R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

[28] Sartori, N. (2006). Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions. *Journal of Statistical Planning and Inference 136*, 4259–4275. MR2323415

[29] Schaefer, R. L. (1983). Bias correction in maximum likelihood logistic regression. *Statistics in Medicine 2*, 71–78.

[30] Simas, A. B., W. Barreto-Souza, and A. V. Rocha (2010). Improved estimators for a general class of beta regression models. *Computational Statistics and Data Analysis 54*(2), 348–366.

[31] Smithson, M. and J. Verkuilen (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods 11*(1), 54–71.

[32] Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis 13*, 157–170.