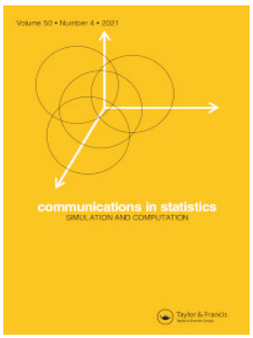

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Small sample bias correction or bias reduction?

Xuemao Zhang, Sudhir Paul & You-Gan Wang

To cite this article: Xuemao Zhang, Sudhir Paul & You-Gan Wang (2021) Small sample bias correction or bias reduction?, Communications in Statistics - Simulation and Computation, 50:4, 1165-1177, DOI: 10.1080/03610918.2019.1577976

To link to this article: https://doi.org/10.1080/03610918.2019.1577976

Published online: 12 Mar 2019.

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
2021, VOL. 50, NO. 4, 1165–1177
https://doi.org/10.1080/03610918.2019.1577976

Small sample bias correction or bias reduction?

Xuemao Zhang[a], Sudhir Paul[b], and You-Gan Wang[c]

[a]Department of Mathematics, East Stroudsburg University, East Stroudsburg, Pennsylvania, USA; [b]Department of Mathematics and Statistics, University of Windsor, Windsor, Ontario, Canada; [c]School of Mathematical Sciences, Queensland University of Technology, Brisbane City, Australia

ABSTRACT

Many problems in biomedical and other sciences are subject to biased estimates (maximum likelihood or of similar types). In two seminal papers Cox and Snell (1968) and Firth (1993) deal with first order bias of maximum likelihood estimates. Cox and Snell obtain a correction term that corrects, approximately, first order bias and Firth uses an adjustment to the score function; the solution of the estimating equation obtained by solving the adjusted score function to zero, removes the first order bias of the maximum likelihood estimates approximately. In many applications authors use one of these two procedures for bias correction without being aware that the other exists or whether these two procedures are equivalent. In this paper we investigate the equivalence issue of the two methods through theoretical analysis, simulation study and data analysis. We show that the two methods yield either exactly the same estimates or that the preventive method has some edge over the other.

ARTICLE HISTORY
Received 12 November 2018
Accepted 18 January 2019

KEYWORDS
Bias correction; Bias reduction; Generalized estimating equations; Longitudinal data; Marginal model

1. Introduction

The existence of bias in maximum likelihood estimates in parametric and semiparametric models, particularly in small samples, is a common phenomenon. One way to correct the bias is using the resampling technique such as Jackknife and Bootstrap (Shao and Tu 1995). Another way is to use Taylor series. Two well-known Taylor series bias correction methods of first order bias correction are by Cox and Snell (1968) and Firth (1993). The method by Cox and Snell (1968) is corrective and that by Firth (1993) is preventive. Many authors use one of these procedures without knowing the difference between them. For example, Saha and Paul use the former for bias correction of the maximum likelihood estimator of the negative binomial dispersion parameter (Saha and Paul 2005a) and of the maximum likelihood estimator of the beta-binomial dispersion parameter (Saha and Paul 2005b), whereas Heinze and Puhr (2010) use the later to reduce bias in conditional logistic regression with small or sparse data sets. Kosmidis and Firth (2009) extend Firth (1993) to bias reduction in exponential family nonlinear models. Paul and Zhang (2014) compare both methods for GEE (generalized estimating equations) estimation of regression parameters for longitudinal data and found them to be equivalent. However, to our knowledge, no systematic study has been conducted to

CONTACT Xuemao Zhang xzhang2@esu.edu Department of Mathematics, East Stroudsburg University, East Stroudsburg, PA 18301-2999, USA.

establish whether, in general, these two methods are equivalent. The purpose of this paper is to show, through theoretical and data analysis, whether these two methods are equivalent in some sense.

Let $l(\boldsymbol{\beta})$ be the log-likelihood for the regression parameter $\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ and $U(\boldsymbol{\beta}) = \partial l/\partial \boldsymbol{\beta}$ be the score function with respect to $\boldsymbol{\beta}$. Then, the first-order bias of the maximum likelihood estimate (MLE) $\hat{\beta}_j$ given by Cox and Snell (1968) is

$$b_s = \frac{1}{2}\sum_i \sum_j \sum_l \kappa^{si}\kappa^{jl}(\kappa_{ijl} + 2\kappa_{ij,l}), \quad s = 1, ..., p,$$

where $\kappa_{ij} = E(\frac{\partial^2 l}{\partial\beta_i \partial\beta_j})$, $\kappa^{ij} = \kappa_{ij}^{-1}$, $\kappa_{ijl} = E(\frac{\partial^3 l}{\partial\beta_i \partial\beta_j \partial\beta_l})$, $\kappa_{ij,l} = E\left[(\frac{\partial^2 l}{\partial\beta_i \partial\beta_j})(\frac{\partial l}{\partial\beta_l})\right]$.

Cordeiro and Klein (1994) show that the above formula for the bias of $\hat{\beta}_s$ can be expressed as

$$b_s = \sum_{i=1}^{p} \kappa^{si} \sum_{j,l=1}^{p} \left[\kappa_{ij}^{(l)} - \frac{1}{2}\kappa_{ijl}\right]\kappa^{jl}, \quad s = 1, ..., p, \tag{1.1}$$

where $\kappa_{ij}^{(l)} = \partial\kappa_{ij}/\partial\beta_l$.

Firth (1993) gives a different form of the first-order bias of the $\hat{\beta}_s$ which is

$$b_s = -\frac{1}{2}\sum_i \sum_j \sum_l \kappa^{si}\kappa^{jl}(\kappa_{i,j,l} + \kappa_{jl,i}), \quad s = 1, ..., p,$$

where $\kappa_{i,j,l} = E(\frac{\partial l}{\partial\beta_i}\frac{\partial l}{\partial\beta_j}\frac{\partial l}{\partial\beta_l})$. These formulas can be derived by symbolic computation with Maple or Mathematica (Stošić and Cordeiro 2009).

However, it is not generally known whether the bias formula given by Cox and Snell (1968) is the same as that given by Firth (1993). In the Appendix we give a proof that the two formulas are equal. Thus, the first order bias of the maximum likelihood estimates can be calculated using any of the three formulas, although the formula by Cordeiro and Klein (1994) is simplest to use. The R package *mle.tools* (Mazucheli, Menezes, and Nadarajah 2017) can be used to implement the bias correction of maximum likelihood estimates.

Now, the bias corrected estimate $\tilde{\beta}_s$ of β_s is given by $\tilde{\beta}_s = \hat{\beta}_s - b(\hat{\beta}_s)$. Following the "preventive" method of Firth (1993), by introducing a bias term into the score function $U(\boldsymbol{\beta})$, the modified score function is

$$U^*(\boldsymbol{\beta}) = U(\boldsymbol{\beta}) - \mathbf{I}\mathbf{b}(\boldsymbol{\beta}), \tag{1.2}$$

where $b(\boldsymbol{\beta}) = (b_1(\boldsymbol{\beta}), ..., b_p(\boldsymbol{\beta}))'$ and $\mathbf{I}$ in the Fisher information matrix. Then, the bias reduced estimate, denoted by $\boldsymbol{\beta}^*$, of $\boldsymbol{\beta}$ using the method of Firth (1993) is obtained by solving the modified score equation

$$U^*(\boldsymbol{\beta}) = 0. \tag{1.3}$$

Section 2 gives a theoretical discussion of the equivalence of the two methods. Six examples are given in Sec. 3 to show equivalence theoretically or through data analysis. A discussion follows in Sec. 4.

2. Equivalence of bias-correction and bias reduction

Suppose $\hat{\boldsymbol{\beta}}$ is the solution to $U(\boldsymbol{\beta}) = 0$ and $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - b(\hat{\boldsymbol{\beta}})$ is the direct bias corrected estimator. We can show $U^*(\tilde{\boldsymbol{\beta}})$ in general is not 0 implying $\tilde{\boldsymbol{\beta}}$ is not the solution to $U^* = 0$. The bias reduced estimator satisfies $U^*(\boldsymbol{\beta}^*) = 0$. So the two methods differ in general, but we do not know to what extent they differ. For this here we do an approximate analysis.

$$\begin{aligned} U^*(\tilde{\boldsymbol{\beta}}) &= U(\tilde{\boldsymbol{\beta}}) - \mathbf{I}\mathbf{b}(\tilde{\boldsymbol{\beta}}) \\ &= U(\hat{\boldsymbol{\beta}} - b(\hat{\boldsymbol{\beta}})) - \mathbf{I}\mathbf{b}(\hat{\boldsymbol{\beta}} - \mathbf{b}(\hat{\boldsymbol{\beta}})) \\ &\approx U(\hat{\boldsymbol{\beta}}) + \mathbf{I}(\hat{\boldsymbol{\beta}}) * b(\hat{\boldsymbol{\beta}}) + \frac{1}{2} b' K_{ijl} b - \mathbf{I}\mathbf{b}(\hat{\boldsymbol{\beta}}) + \mathbf{I}\frac{\partial \mathbf{b}}{\partial \boldsymbol{\beta}}\mathbf{b} \\ &= \frac{1}{2} b' K_{ijl} b + \mathbf{I}\frac{\partial b}{\partial \beta} = AD(say), \end{aligned}$$

where AD is approximate difference. Here $b'K_{ijl}b$ is a vector obtained by $b'K_{ij,1}b, b'K_{ij,2}b, \ldots, b'K_{ij,n}b$, and $K_{ij,1}$ is the derivative of the matrix $-\mathbf{I}$ with respect to $\boldsymbol{\beta}$. If the bias function $b(\theta)$ is linear in $\theta, \tau_0 + \tau_1\theta$, where both τ_0 and τ_1 are $O(1/n)$ we will have $U^*(\tilde{\boldsymbol{\beta}}) = 0$ indicating $\tilde{\boldsymbol{\beta}}$ coincides with $\boldsymbol{\beta}^*$.

Do they then become the same? So the correction only differs in $o_p(1/n)$ in general or $O_p(1/n^2)$ in some cases. But in general linear approximation is not exact. So the two methods differ as we will show in the numerical examples in Sec. 3.

Even in the case $b(\theta) = (\tau_0 + \tau_1\theta)/n$. The two methods can still differ in the second order $(1/n^2)$. This is because $b(\theta)$ is only the first order approximation in Cox and Snell (1968). Suppose $E(\hat{\theta}) = \theta + B(\theta), b(\theta) = (\tau_0 + \tau_1\theta)/n$ and $B(\theta) = b(\theta) + h$, where h represents higher order bias. Then,

$$\begin{aligned} E\{\hat{\theta} - b(\hat{\theta})\} &= \theta + B(\theta) - Eb(\hat{\theta}) \\ &= \theta + \{\tau_0 + \tau_1\theta\}/n + h(\theta) - \{\tau_0 + \tau_1 E(\hat{\theta})\}/n \\ &= \theta + \tau_1\theta/n + h(\theta) - \tau_1\{\theta + (\tau_0 + \tau_1\theta)/n + h\}/n \\ &= \theta + h(\theta) - \tau_1(\tau_0 + \tau_1\theta)/n^2 - \tau_1 h/n. \end{aligned}$$

Only in the case of $h = \frac{\tau_1}{(n-\tau_1)n} b(\theta)$, we will have the bias corrected estimate $\hat{\theta} - b(\hat{\theta})$ exactly unbiased. This is true for the Example 1 to be shown in Sec. 3.

It is difficult to show analytically, in general, that the quantity AD is zero. So, in situations in which it is not possible to show that AD is zero, we will illustrate this in the next section through some examples. Through the examples if we can show that AD is zero or close to zero, this will indicate the two methods approximately match up to second order, otherwise they only match the first order.

3. Some examples

Two things will be investigated through examples, (i) the value of AD and as to how close it is to zero and (ii) to what extent the bias corrected and the bias reduced estimates differ (if they do). Six examples are given.

Example 1: Normal Distribution (Firth 1993, Section 3.2)

Suppose $y_1, ..., y_n$ is a random sample from $N(\mu, \sigma^2)$. We are interested in estimating the two dimensional canonical parameter $\theta = (\mu/\sigma^2, -1/2\sigma^2)$. It can be seen that, apart from a constant, the log-likelihood is given by

$$l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \left[\frac{1}{2} \log(-\theta_2) + \frac{\theta_1^2}{4\theta_2} + y_i\theta_1 + y_i^2\theta_2 \right].$$

Using this it is easy to see that the score functions for (θ_1, θ_2) are

$$\begin{pmatrix} \frac{\theta_1 n}{2\theta_2} + n\bar{y} \\ \sum_i \left\{ -\frac{1}{2\theta_2} - \frac{\theta_1^2}{4\theta_2^2} + y_i^2 \right\} \end{pmatrix}.$$

The information matrix is

$$I(\theta) = -n \begin{pmatrix} \frac{1}{2\theta_2} & -\frac{\theta_1}{2\theta_2^2} \\ -\frac{\theta_1}{2\theta_2^2} & \frac{1}{2\theta_2^2} + \frac{\theta_1^2}{2\theta_2^3} \end{pmatrix}.$$

The maximum likelihood estimate of θ is given by

$$\hat{\theta} = \begin{pmatrix} \frac{n\bar{y}}{(n-1)s_y^2} \\ \frac{-n}{2(n-1)s_y^2} \end{pmatrix},$$

where $s_y^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1)$. Clearly, $E(\hat{\theta}) = \theta n/(n-3)$ producing a bias of $B(\theta) = 3\theta/(n-3)$.

Furthermore,

$$K^{(1)} = n \begin{pmatrix} 0 & \frac{1}{2\theta_2^2} \\ \frac{1}{2\theta_2^2} & -\frac{\theta_1}{\theta_2^3} \end{pmatrix} \text{ and } K^{(2)} = n \begin{pmatrix} \frac{1}{2\theta_2^2} & -\frac{\theta_1}{\theta_2^3} \\ -\frac{\theta_1}{\theta_2^3} & \frac{1}{\theta_2^3} + \frac{3\theta_1^2}{2\theta_2^4} \end{pmatrix}.$$

This leads to

$$\begin{pmatrix} b'I^{(1)}b \\ b'I^{(2)}b \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{9}{n\theta_2} \end{pmatrix}.$$

We also have $\frac{\partial b}{\partial \beta'} b = 9(\theta_1, \theta_2)/n^2$ and $\mathbf{I} \frac{\partial b}{\partial \beta'} b = 9(0, 1/(2\theta_2))/n$.

Now, the bias formula given in Eq. (1.1) can be written as

$$b(\theta) = K^{-1} A \text{vec}(K^{-1}),$$

where $K^{-1} = i^{-1}(\theta), A = \{A^{(1)}|A^{(2)}\}$ with $A^{(l)} = \{a_{ij}^{(l)}\} = \{\kappa_{ij}^{(l)} - \frac{1}{2}\kappa_{ijl}\} = \{\frac{1}{2}\kappa_{ij}^{(l)}\}, i, j, l = 1, 2$. After some algebraic calculations, we obtain

$$b(\theta) = \frac{3}{n}\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} = \frac{3}{n}\begin{pmatrix}\mu/\sigma^2\\ -1/2\sigma^2\end{pmatrix} \text{ and } B(\theta) = 3\theta/(n-3).$$

Therefore, the bias-corrected maximum likelihood estimate of θ is

$$\tilde{\theta} = \hat{\theta} - b(\hat{\theta}) = \frac{n-3}{n}\hat{\theta}.$$

We can now check $U^*(\tilde{\theta}) = 0$ implying the two methods become the same in this particular case.

$$\begin{aligned} U(\tilde{\theta}) &= U(\hat{\theta}) - \begin{pmatrix}0\\ \dfrac{3}{2(n-3)\hat{\theta}_2}\end{pmatrix} \\ &= -\begin{pmatrix}0\\ \dfrac{3}{2(n-3)\hat{\theta}_2}\end{pmatrix}. \\ I(\tilde{\theta})b(\tilde{\theta}) &= \begin{pmatrix}0\\ -\dfrac{3}{2n\tilde{\theta}_2}\end{pmatrix} \\ &= \begin{pmatrix}0\\ -\dfrac{3}{2(n-3)\hat{\theta}_2}\end{pmatrix}. \end{aligned}$$

We therefore have $U^*(\tilde{\theta}) = U(\tilde{\theta}) - I(\tilde{\theta})b(\tilde{\theta}) = U(\hat{\theta}) = 0$.

Now, the modified score function given in Eq. (1.2), for this example, can be written as

$$U^*(\theta) = U(\theta) - i(\theta)b(\theta) = \begin{pmatrix}\dfrac{\partial l}{\partial \theta_1}\\ \dfrac{\partial l}{\partial \theta_2} - \dfrac{3}{2\theta_2}\end{pmatrix},$$

which results in the bias-reduced maximum likelihood estimate of θ

$$\theta^* = \begin{pmatrix}\dfrac{(n-3)\bar{y}}{(n-1)s_y^2}\\ \dfrac{-(n-3)}{2(n-1)s_y^2},\end{pmatrix},$$

which is the same as given in Firth (1993, p. 30).

Thus, for this example, the bias-corrected and bias-reduced maximum likelihood estimates of θ are of the same.

In this case, we also have

$$AD = -\frac{1}{2}\left(b'K^{(1)}b, b'K^{(2)}b, \ldots, b'K^{(p)}b\right) + \mathbf{I}\frac{\partial b}{\partial \beta'}b = 0.$$

Example 2: Logistic regression

Suppose $\mathbf{y}=(y_1,...,y_n)'$ is a vector of independent binary variables and the success probability π_i, for the ith response is explained by a linear regression $\eta_i=\mathbf{x}_i'\boldsymbol{\beta}$ in the logit scale, where $\boldsymbol{\beta}=(\beta_1,...,\beta_p)'$, $\mathbf{x}_i=(x_{i1},...,x_{ip})'$, $i=1,...,n$. Then, the linear logistics regression model can be written as $P(y_i=1|\mathbf{x}_i,\boldsymbol{\beta})=\pi_i(\mathbf{x}_i'\boldsymbol{\beta})=\frac{\exp(\eta_i)}{1+\exp(\eta_i)}=\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}$, $i=1,...,n$.

Therefore, the likelihood function is

$$L=\prod_{i=1}^{n}\pi_i^{y_i}(1-\pi_i)^{1-y_i}=\prod_{i=1}^{n}\left(\frac{\pi_i}{1-\pi_i}\right)^{y_i}(1-\pi_i)=\prod_{i=1}^{n}\left(\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)^{y_i}\frac{1}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)},y_i=0,1$$

and the log-likelihood function is

$$l=\sum_{i=1}^{n}\left[y_i\mathbf{x}_i'\boldsymbol{\beta}-\log\left(1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)\right],y_i=0,1.$$

The maximum likelihood estimator of $\boldsymbol{\beta}$ is obtained by solving $U(\boldsymbol{\beta})=\frac{\partial l}{\partial\boldsymbol{\beta}}=0$ or solving the following equations

$$\frac{\partial l}{\partial\beta_j}=\sum_{i=1}^{n}x_{ij}\left[y_i-\frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}\right]=0,j=1,...,p.$$

Furthermore,

$$\frac{\partial^2 l}{\partial\beta_j^2}=-\sum_{i=1}^{n}x_{ij}^2\left[\frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}-\left(\frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}\right)^2\right]=-\sum_{i=1}^{n}x_{ij}^2\pi_i(1-\pi_i),j=1,...,p$$

and

$$\begin{aligned}\frac{\partial^2 l}{\partial\beta_j\partial\beta_k}&=-\sum_{i=1}^{n}x_{ij}x_{ik}\left[\frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}-\left(\frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}\right)^2\right]\\&=-\sum_{i=1}^{n}x_{ij}x_{ik}\pi_i(1-\pi_i),j,k=1,...,p.\end{aligned}$$

Thus, the information matrix can be written as

$$i(\boldsymbol{\beta})=\mathbf{X}'\mathbf{W}\mathbf{X},$$

where $\mathbf{W}=\mathrm{diag}\{\pi_i(1-\pi_i)\}$ and $\mathbf{X}=\{\mathbf{x}_1,...,\mathbf{x}_n\}'$ is a $n\times p$ matrix.

Now,

$$\begin{aligned}\kappa_{jk}^{(l)}=\frac{\partial^3 l}{\partial\beta_j\partial\beta_k\partial\beta_l}&=-\sum_{i=1}^{n}x_{ij}x_{ik}x_{il}\left[\frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{\left[1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right]^2}-2\frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}\frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{\left[1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right]^2}\right]\\&=-\sum_{i=1}^{n}x_{ij}x_{ik}x_{il}\pi_i(1-\pi_i)(1-2\pi_i),j,k,l=1,...,p.\end{aligned}$$

And again $\kappa_{jkl}=\kappa_{jk}^{(l)},j,k,l=1,...,p$.

In matrix notation, the bias of the maximum likelihood estimate of $\boldsymbol{\beta}$ is given by

$$b(\boldsymbol{\beta})=K^{-1}A\mathrm{vec}(K^{-1}),$$

Table 1. Maximum likelihood estimation in the one-parameter logistic regression model.

n	$\hat{\beta}$	$\tilde{\beta}$	β^*	AD
5	0.0808	0.0768	0.0673	5.06×10^{-3}
10	0.110	0.103	0.0988	6.40×10^{-3}
15	0.167	0.158	0.154	4.20×10^{-3}
20	0.246	0.237	0.232	3.44×10^{-3}
30	0.312	0.303	0.299	2.61×10^{-3}
40	0.428	0.420	0.416	2.49×10^{-3}
50	0.521	0.514	0.509	2.30×10^{-3}

where $K^{-1} = i^{-1}(\boldsymbol{\beta})$, $A = \{A^{(1)}|A^{(2)}|\cdots|A^{(p)}\}$ with $A^{(l)} = \{a_{jk}^{(l)}\} = \{\kappa_{jk}^{(l)} - \frac{1}{2}\kappa_{jkl}\} = \{\frac{1}{2}\kappa_{jk}^{(l)}\}$, $j, k, l = 1, ..., p$.

Note that the modified score function is

$$U^*(\boldsymbol{\beta}) = U(\boldsymbol{\beta}) - i(\boldsymbol{\beta})b(\boldsymbol{\beta}) = U(\boldsymbol{\beta}) - \mathbf{I}_p A \text{vec}\left(i^{-1}(\boldsymbol{\beta})\right),$$

where $\mathbf{I}_p$ is a p-dimensional Identity matrix.

Obviously, we cannot show theoretically that the bias corrected and the bias reduced estimates will be the same. So, in what follows we use two data sets; one from Copas (1988) and the other from Miles and Shevlin (2001) to see whether they are different and if so to what extent.

Example 2.1: Copas considers a one-parameter logistic regression example used by Copas (1988). This is also example 3.3 in Firth (1993). The regression go through the origin and the explanatory variable x takes values in $\{-2, -1, 0, 1, 2\}$. To conduct a simulation study, for each x value we generate $1, 2, 3, 4, 6, 8, 10$ random binary responses (resulting in samples of size 5, 10, 15, 20, 30, 40 and 50) using the regression parameter $\beta = 0.5$. And we calculate the maximum likelihood estimate $\hat{\beta}$, bias-corrected estimate $\tilde{\beta}$ and bias-reduced estimate β^*. Repeat this procedure 1000 times and the mean value of the estimates of β are summarized in Table 1. The values of AD are also given in the this example.

Table 1 shows that the bias-corrected and the bias-reduced estimates are not the same but very close.

Example 2.2: Here we consider the employee selection procedures data from Miles and Shevlin (2001). An organization wanted to assess the efficiency of its employee selection procedures. The first variable, score, is an applicant's score on a test of aptitude towards the job. The second variable, experience, is the number of months of relevant prior experience that the applicant has had before the job. The response variable, pass, is whether the applicant actually passed the test after their training period (1 indicates Yes, 0 indicates No). The data for the 26 subjects are given in Table 2.

For these data we consider the logistic regression model

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \text{score} + \beta_2 \text{experience}, i = 1, ..., 26.$$

The maximum likelihood estimates, the bias corrected estimates and the bias reduced estimates are given in Table 3.

It appears that for this example also there is a difference between $\tilde{\beta}$ and β^*. However, the difference is very small.

Table 2. Employee selection procedures data from Miles and Shevlin (2001).

Score	5	1	1	4	1	1	4	1	3	4	5	1	3
Experience	6	15	12	6	15	6	16	10	12	26	2	12	18
Pass	0	0	0	0	1	0	1	1	0	1	1	0	0
Score	3	1	2	1	4	4	5	4	4	2	2	1	5
Experience	3	24	8	9	18	22	3	12	24	18	6	8	12
Pass	0	1	0	0	0	1	1	0	1	1	0	0	0

Table 3. Maximum likelihood estimation in a logistic regression model for the data in Miles and Shevlin (2001).

	β_0	β_1	β_2
$\hat{\beta}$	−3.17	0.29	0.15
β	−2.68	0.25	0.12
β^*	−2.74	0.26	0.12

Table 4. Data for O. cernua seed in bean root extract.

Dilution level	Proportions
1/1	2/43, 9/51, 5/44, 16/71, 2/24, 0/7
1/25	17/19, 43/56, 79/87, 50/55, 9/10
1/625	11/13, 47/62, 90/104, 46/51, 9/11

Table 5. The Format of Data for O. cernua seed in bean root extract.

Dilution level	Proportions
1	$y_{11}/n_{11}, y_{12}/n_{12}, \ldots, y_{1j}/n_{1j}, \ldots, y_{1m_1}/n_{1m_1}$
2	$y_{21}/n_{21}, y_{22}/n_{22}, \ldots, y_{2j}/n_{2j}, \ldots, y_{2m_2}/n_{2m_2}$
3	$y_{31}/n_{31}, y_{32}/n_{32}, \ldots, y_{3j}/n_{3j}, \ldots, y_{3m_3}/n_{3m_3}$

Example 3. Beta-binomial regression with a logit link

We consider data for *O. cernua* seed in bean root extract given in Table 1 of Crowder (1978). These data are reproduced here as Table 4. The data refer to seed Orobanche cernua cultivated in three dilutions 1/1, 1/25 and 1/625 of a bean root extract. The data are in 3 dilution levels, $i\,th$ dilution level having m_i litters, $i = 1, 2, 3$. Then the data are of the form of Table 5.

We assume that $y_{ij}|p_i \sim binomial(n_{ij}, p_i)$ and p_i is a beta random variable with mean π_i and variance $\pi_i(1-\pi_i)\phi_i$. Then the unconditional distribution of y_{ij} is

$$Pr(y_{ij}|\pi_i, \phi_i) = \binom{n_{ij}}{y_{ij}} \frac{\prod_{r=0}^{y_{ij}-1}[\pi_i(1-\phi_i)+r\phi_i]\prod_{r=0}^{n_{ij}-y_{ij}-1}[(1-\pi_i)(1-\phi_i)+r\phi_i]}{\prod_{r=0}^{n_{ij}-1}[(1-\phi_i)+r\phi_i]} \tag{3.1}$$

with mean $n_{ij}\pi_i$ and variance $n_{ij}\pi_i(1-\pi_i)(1+(n_{ij}-1)\phi_i)$, where $0 \le \pi_i \le 1$, and $\phi_i \ge \max[-\pi_i/(n_{ij}-1), -(1-\pi_i)/(n_{ij}-1)]$. This is the extended beta-binomial distribution of Prentice (1986). The parameter ϕ_i is the extra-dispersion parameter. We assume further that $log\{\pi/(1-\pi)\} = \alpha + \beta x_i$. The purpose here is to obtain the bias corrected and bias reduced estimates of the parameters α, β, and ϕ_i, $i = 1, 2, 3$. Due to convergence problem, we consider a common extra-dispersion parameter ϕ and the covariates Dilution level $x_1 = 100/1, x_2 = 100/25$ and $x_3 = 100/625$ respectively. Table 6 summarizes the estimation of α, β, and ϕ.

Table 6. Maximum likelihood estimation in a logistic regression model.

	α	β	ϕ
$\hat{\beta}$	0.000175	0.668	−0.0726
$\bar{\beta}$	0.000175	0.668	−0.0725
β^*	0.000133	0.602	−0.0589

For this example it shows that the bias corrected and bias reduced estimates of α, β, and ϕ are not very similar, although the difference does not seem to be significant.

Example 4. Estimation of the Dispersion Parameter in Over-dispersed Poisson Model.

A popular extra(over/under)-dispersed count data model is the Negative binomial distribution, denoted by $NB(m, c)$, where m is the mean and c is the extra-dispersion parameter, having pmf

$$Pr(Y = y|m, c) = \frac{\Gamma(y + c^{-1})}{y!\Gamma(c^{-1})}\left(\frac{cm}{1 + cm}\right)^{y}\left(\frac{1}{1 + cm}\right)^{c^{-1}}, 0 < m, c < \infty, y = 0, 1, 2, \ldots$$

Let $y_1, \ldots, y_n$ be a random sample from $NB(m, c)$. We are interested in estimating the parameters m and c. Following Cox and Snell (1968), Saha and Paul (2005a) derive bias of the maximum likelihood estimates of m and c. They show that the maximum likelihood estimate of m is $\bar{y}$ and hence it is unbiased. Further, the maximum likelihood estimate of c is obtained by solving the maximum likelihood estimating equation

$$U(c) = \sum_{i=1}^{n}\left[\frac{1}{c^2}ln(1 + c\bar{y}) - \frac{y_i - \bar{y}}{c(1 + c\bar{y})} - \sum_{j=0}^{y_i - 1}\frac{1}{c(1 + cj)}\right] = 0.$$

Denote the maximum likelihood estimate of c by $\hat{c}$. Then, bias of the maximum likelihood estimate of c, after replacing m by $\bar{y}$ and c by $\hat{c}$, obtained by Saha and Paul (2005a), is given by

$$b(m, c) = -\frac{m\psi c^4}{2n(1 + cm)} + \frac{c^5\psi^2}{n}\Delta - \frac{c^8\psi^2}{n}(\Phi_1 + \Phi_2 - \Delta_{34}),$$

where expressions for ψ, Δ, Φ_1, Φ_1 and Δ_{34} are given in Appendix of Saha and Paul (2005a, p 185). So, the bias corrected estimate of c is $\hat{c}_{bc} = \hat{c} - b(\hat{m}, \hat{c})$ and the bias reduced estimate of c, denoted by $\hat{c}_{br}$, is obtained by solving

$$U(c) - Ib(\hat{m}, c) = 0, \tag{3.2}$$

where $I = \frac{n}{c^4}\sum_{i=1}^{\infty}\frac{i!(cb)^{i+1}}{(i+1)d_i}$, with $b = \frac{cm}{1+cm}$ and $d_i = \prod_{j=0}^{i}(1 + jc)$ (see Saha and Paul 2005a, p. 185).

Since no closed form solution of the Eq. (3.2) is possible, equality of $\hat{c}_{bc}$ and $\hat{c}_{br}$ cannot be established theoretically. So, these estimates are evaluated for an example. Consider the data on the number of European red mites on apple leaves reported by Bliss and Fisher (1953). Saha and Paul (2005a) report that for the red mites data $\hat{c} = 0.976$ and $\hat{c}_{bc} = 0.9805$. By solving Eq. (3.2) for these data we obtain $\hat{c}_{br} = 0.9806$. For this problem, $\hat{c}_{bc}$ and $\hat{c}_{br}$ are almost identical and the value of AD is 0.04153.

Example 5. Equivalence in Semi-parametric Models: Generalized Estimating Function.

In absence of the assumptions of full distributional model a very popular method of analyzing longitudinal data is the generalized estimating equations (GEEs) approach of Liang and Zeger (1986) and Zeger and Liang (1986). However, as in the parametric models estimates obtained by solving the generalized estimating equations (GEEs) may not be unbiased. The bias correction method of Cox and Snell (1968) and the Bias reduction method of Firth (1993) do not in principle apply to the generalized estimating functions as these are not strictly likelihood score functions. Paul and Zhang (2014) obtain bias-corrected estimates (GEEBc) and bias reduced estimates (GEEBr) of the regression parameters for longitudinal data by treating the generalized estimating functions as if they were likelihood score functions. By an extensive simulation study they show that the performance of both the bias-corrected and the bias reduced methods are similar in terms of bias, efficiency, coverage probability, average coverage length, impact of misspecification of correlation structure, and impact of cluster size on bias correction. They also show that both these methods show superior properties over the GEE estimates for small samples. For details of theory and the simulation study we refer the reader to Paul and Zhang (2014). A summary of results is given in what follows.

Let $R(\boldsymbol{\alpha})$ be a working correlation matrix completely specified by the parameter vector $\boldsymbol{\alpha}$. Then, $\phi \boldsymbol{W}_n = \phi \mathbf{A}_n^{1/2} R(\boldsymbol{\alpha}) \mathbf{A}_n^{1/2}$ is the corresponding working covariance matrix, where $\mathbf{A}_n(\boldsymbol{\beta}) = \text{diag}\{v(\mu_{nj})\}$, $j = 1, ..., d, n = 1, ..., N$. For given consistent estimates of ϕ and $\boldsymbol{\alpha}$, the GEE estimate of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}$, is obtained by solving the generalized estimating equations

$$\sum_{n=1}^{N} \mathbf{D}_n' \mathbf{W}_n^{-1} (\mathbf{y}_n - \boldsymbol{\mu}_n) = 0, \tag{3.3}$$

where $\mathbf{D}_n = \frac{\partial \boldsymbol{\mu}_n}{\partial \boldsymbol{\beta}} = \boldsymbol{\Delta}_n \mathbf{X}_n, \boldsymbol{\Delta}_n = \text{diag}(f(\mathbf{x}_{n1}'\boldsymbol{\beta}), ..., f(\mathbf{x}_{nd}'\boldsymbol{\beta}))$ with $f = F', n = 1, ..., N$.

The left-hand side of Eq. (3.3) denoted by $U(\boldsymbol{\beta}; \boldsymbol{\alpha}, \phi)$ is the generalized estimating function for $\boldsymbol{\beta}$ given $\boldsymbol{\alpha}$ and ϕ. Let $U(\boldsymbol{\beta}; \boldsymbol{\alpha}, \phi) = (U_1, U_2, ..., U_p)'$. For obtaining bias-corrected (Cox and Snell 1968) and bias-reduced (Firth 1993) GEE estimates, we treat U_i as if it were a likelihood score function for β_i, $i = 1, ..., p$.

Now, define $\kappa_{ij} = \text{E}(\partial U_i / \partial \beta_j | \boldsymbol{X}_1, ..., \boldsymbol{X}_n)$ for $i, j = 1, ..., p$. Further, define $\kappa_{ijl} = \text{E}(\partial^2 U_i / \partial \beta_j \partial \beta_l | \boldsymbol{X}_1, ..., \boldsymbol{X}_n)$ and $\kappa_{ij}^{(l)} = \partial \kappa_{ij} / \partial \beta_l$ for $i, j, l = 1, ..., p$. Derivation of the quantities κ_{ij}, $\kappa_{ij}^{(l)}$, and κ_{ijl} are given in the Web Appendix A of Paul and Zhang (2014). Then, the Fisher information matrix analog of order p for $\boldsymbol{\beta}$ is $\mathbf{I} = \{-\kappa_{ij}\}$. Now, let $\mathbf{I}^{-1} = \{\kappa^{ij}\}$ be the inverse of $\mathbf{I}$. Then following the bias formula (1.1) of Cordeiro and Klein (1994), the bias corrected estimate $\tilde{\beta}_s$ of β_s is given by $\tilde{\beta}_s = \hat{\beta}_s - b(\hat{\beta}_s)$. The estimates $\tilde{\beta}_s$ will also be referred to as GEEBc estimates.

Following the "preventive" method of Firth (1993), by introducing a bias term into the score function $U(\boldsymbol{\beta}; \boldsymbol{\alpha}, \phi)$, the modified score function is

$$U^*(\boldsymbol{\beta}; \boldsymbol{\alpha}, \phi) = U(\boldsymbol{\beta}; \boldsymbol{\alpha}, \phi) - \mathbf{I} b(\boldsymbol{\beta}),$$

where $b(\boldsymbol{\beta}) = (b_1(\boldsymbol{\beta}), ..., b_p(\boldsymbol{\beta}))'$.

The bias reduced GEE estimate, denoted by $\boldsymbol{\beta}^*$, of $\boldsymbol{\beta}$ using the method of Firth (1993) is obtained by solving the modified score equation

$$U^*(\boldsymbol{\beta}; \boldsymbol{\alpha}, \phi) = 0. \tag{3.4}$$

This equation needs to be solved iteratively (see Paul and Zhang 2014). The estimates obtained are referred to as GEEBr. The striking similarity of the GEEBc and GEEBr was observed in the analysis of data of a clinical trial on cerebrovascular deficiency with a crossover design from Diggle, Liang, and Zeger (1994). The purpose of this crossover trial was to compare an active drug (A) and a placebo (B). A total of 67 patients were enrolled into the clinical trial of which 34 patients received the active drug (A) followed by the placebo (B), and another 33 patients were treated in the reverse order. The response variable is defined as 0 for an abnormal and 1 for a normal electrocardiogram reading. As indicated in Paul and Zhang (2014) the 2×2 crossover trial can be viewed as a longitudinal study with 2 observations for each patient. The two major covariates, period (x_{n1}) and treatment (x_{n2}), are both time-dependent. They are coded as

$$x_{n1} = \begin{cases} 1, & \text{period 2} \\ 0, & \text{period 1} \end{cases} \quad \text{and} \quad x_{n2} = \begin{cases} 1, & \text{active drug (A)} \\ 0, & \text{placebo (B)} \end{cases}, \quad \text{respectively.}$$

For the data see Table II in Paul and Zhang (2014).

By analyzing a full regression model Diggle, Liang, and Zeger (1994, p. 153) show insignificant treatment-by-period interaction effect. So, Paul and Zhang (2014) consider the logit regression model

$$\text{logit } \Pr(Y_{nj} = 1) = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2}$$

and show that the GEE, GEEBc, and GEEBr estimates of the regression parameter β_0, with their standard errors in parenthesis, are $0.6659(0.2939), 0.6527(0.2924)$ and $0.6527(0.2924)$ respectively. The corresponding estimates of β_1 are $-0.2950(0.2382), -0.2883(0.2367)$ and $-0.2876(0.2367)$ and those for β_2 are $0.5689(0.2398), 0.5557(0.2380)$ and $0.5556(0.2380)$ respectively.

As this data set has 67 patients this represents a large sample. Paul and Zhang then analyze many simple random samples, taken from these data, of size 15 (7, 8) of which one sample was given in their Table III. Analysis of this data set shows that the GEE, GEEBc, and GEEBr estimates of the regression parameters β_0 with their standard errors in parenthesis are $-0.4006(0.6866), -0.3623(0.6776)$ and $-0.3656(0.6784)$ respectively. The corresponding estimates of β_1 are $-0.6655(0.7311)$, $-0.5956(0.7158)$ and $-0.6022(0.7171)$ and those for β_2 are $1.1718(0.7316), 1.0554(0.7161)$ and $1.0660(0.7173)$ respectively.

4. Discussion

We show, through theoretical discussions that the bias corrected estimates (Cox and Snell 1968) and bias reduced estimates (Firth 1993), in general, are not the same. However, through 6 examples, some of them similar to those analyzed by Firth (1993), using derivations or data analysis we show that, in general, these are either exactly the same or are very close to each other. The models that we dealt with are similar to those by Firth (1993). However, Kosmidis and Firth (2009) extend Firth (1993) to bias reduction in exponential family nonlinear models. As a future study it would be interesting to study equivalence of the two methods in exponential family nonlinear models and non-linear models in general.

ORCID

Xuemao Zhang http://orcid.org/0000-0002-6113-4710

Appendix

To show that the two formulas are equivalent, we need to show

$$\frac{1}{2}\sum_i\sum_j\sum_l \kappa^{si}\kappa^{jl}(\kappa_{ijl}+2\kappa_{ij,l}) = -\frac{1}{2}\sum_i\sum_j\sum_l \kappa^{si}\kappa^{jl}(\kappa_{i,j,l}+\kappa_{jl,i}),$$

or

$$\sum_i \kappa^{si}\sum_j\sum_l \kappa^{jl}(\kappa_{ijl}+2\kappa_{ij,l}+\kappa_{i,j,l}+\kappa_{jl,i}) = 0.$$

By (2.2) in Firth (1993),

$$\kappa_{ijl}+\kappa_{ij,l}+\kappa_{il,j}+\kappa_{jl,i}+\kappa_{i,j,l} = 0,$$

we only need to show that

$$\sum_i \kappa^{si}\sum_j\sum_l \kappa^{jl}(2\kappa_{ij,l}-\kappa_{j,il}-\kappa_{l,ij}) = 0.$$

It is apparent that $\kappa_{ij,l} = \kappa_{l,ij}$. Therefore, we only need to show

$$\sum_i \kappa^{si}\sum_j\sum_l \kappa^{jl}(\kappa_{ij,l}-\kappa_{j,il}) = 0.$$

Note that since $\kappa^{ij} = \kappa^{ji}$, we have that

$$\sum_i\sum_j\sum_l \kappa^{si}\kappa^{jl}\kappa_{ij,l} = \sum_i\sum_j\sum_l \kappa^{si}\kappa^{jl}\kappa_{il,j} = \sum_i\sum_j\sum_l \kappa^{si}\kappa^{jl}\kappa_{j,il}$$

and thus the two formula are equivalent.

References

Bliss, C. I., and R. A. Fisher. 1953. Fitting the negative binomial distribution to biological data - note on the efficient fitting of the negative binomial. *Biometrics* 9 (2):176–200. doi:10.2307/3001850.

Copas, J. B. 1988. Binary regression models for contaminated data. *Journal of the Royal Statistical Society. Series B* 50:225–65. doi:10.1111/j.2517-6161.1988.tb01723.x.

Cordeiro, G. M., and R. Klein. 1994. Bias correction in ARMA models. *Statistics & Probability Letters* 19 (3):169–76. doi:10.1016/0167-7152(94)90100-7.

Cox, D. R., and E. J. Snell. 1968. A general definition of residuals (with discussion). *Journal of the Royal Statistical Society. Series B* 30:248–75. doi:10.1111/j.2517-6161.1968.tb00724.x.

Crowder, M. J. 1978. Beta-binomial anova for proportions. Journal *of Applied Statistics*. 27 (1): 34–7. doi:10.2307/2346223.

Diggle, P. J., K. Y. Liang, and S. L. Zeger. 1994. *Analysis of longitudinal data*. Oxford: Oxford University Press.

Firth, D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika* 80 (1):27–38. doi: 10.1093/biomet/80.1.27.

Heinze, G., and R. Puhr. 2010. Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. Statistics *in Medicine* 29 (7-8):770–7.

Kosmidis, I., and D. Firth. 2009. Bias reduction in exponential family nonlinear models. *Biometrika* 96 (4):793–804. doi:10.1093/biomet/asp055.

Liang, K.-Y., and S. L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73:45–51.

Mazucheli, J., A. Menezes, and S. Nadarajah. 2017. mle.tools: an R package for maximum likelihood bias correction. *The R Journal* 9 (2):268–90.

Miles, J., and M. Shevlin. 2001. *Applying regression and correlation: A guide for students and researchers*. London: SAGE.

Paul, S., and X. Zhang. 2014. Small sample GEE estimation of regression parameters for longitudinal data. Statistics *in Medicine* 33 (22):3869–81.

Prentice, R. L. 1986. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association* 81 (394):321–7. doi:10.1080/01621459.1986.10478275.

Saha, K., and S. Paul. 2005a. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics* 61 (1):179–85. doi:10.1111/j.0006-341X.2005.030833.x.

Saha, K., and S. Paul. 2005b. Bias-corrected maximum likelihood estimator of the intraclass correlation parameter for binary data. Statistics *in Medicine* 24:3497–512. doi:10.1002/sim.2197.

Shao, J., and D. Tu. 1995. *The jackknife and bootstrap*. New York: Springer.

Stošić, B., and G. Cordeiro. 2009. Using maple and mathematica to derive bias corrections for two parameter distributions. Journal *of Statistical Computation and Simulation*. 79 (6):751–67. doi:10.1080/00949650801911047.

Zeger, S. L., and K.-Y. Liang. 1986. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42 (1):121–30.