Journal of Statistical Planning and Inference 140 (2010) 2478–2485

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Bias correction in logistic regression with missing categorical covariates

Ujjwal Das [a], Tapabrata Maiti [b], Vivek Pradhan [c,*]

[a] Division of Statistics, Northern Illinois University, DeKalb, IL 60115, USA
[b] Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA
[c] Boston Scientific Corporation, 100 Boston Scientific Way, Marlborough, MA 01752, USA

ARTICLE INFO

Article history:
Received 24 August 2009
Received in revised form
19 February 2010
Accepted 23 February 2010
Available online 2 March 2010

Keywords:
Bias
EM algorithm
Maximum likelihood estimation
Missing at random

ABSTRACT

Logistic regression plays an important role in many fields. In practice, we often encounter missing covariates in different applied sectors, particularly in biomedical sciences. Ibrahim (1990) proposed a method to handle missing covariates in generalized linear model (*GLM*) setup. It is well known that logistic regression estimates using small or medium sized missing data are biased. Considering the missing data that are missing at random, in this paper we have reduced the bias by two methods; first we have derived a closed form bias expression using Cox and Snell (1968), and second we have used likelihood based modification similar to Firth (1993). Here we have analytically shown that the Firth type likelihood modification in Ibrahim led to the second order bias reduction. The proposed methods are simple to apply on an existing method, need no analytical work, with the exception of a little change in the optimization function. We have carried out extensive simulation studies comparing the methods, and our simulation results are also supported by a real world data.

1. Introduction

Logistic regression is an important statistical tool in applied statistics. In general, in order to estimate the regression coefficients, it is required to have the full covariate information. However, in practice, most of the time data are obtained with missing values due to several reasons; for example, survey non-responses, study subjects failing to report to a clinic, unavailability of some patient's information, etc. In the usual estimation process, the missingness is handled by dropping the subjects whose data contains missing values. In clinical trial and many other applications, often the collected data itself are small, and the deletion of the cases with missing covariates makes it even smaller in the actual estimation process. It is well known that the maximum likelihood estimates (MLE) of a logistic regression from a small sample are biased. The MLEs are estimated using an iterative process. A practical problem of such iterative processes (such as Newton–Raphson method) of finding the estimates using a small sample is separation. For the binary response variable, separation is a phenomenon when a single variable, or the linear combination of the predictor variables perfectly or near perfectly separates the space of success and failure. In such a situation, although the loglikelihood converges to a finite value, at least one or more parameters of interest diverge. An excellent discussion can be found in Heinze and Schemper (2002) and Heinze (2006). In such setup, the occurrences of partially missing covariates add an extra complexity to the whole process.

* Corresponding author.
E-mail address: vivek.pradhan@bsci.com (V. Pradhan).

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.02.018

Following Cox and Snell (1968), Cordeiro and McCullagh (1991) proposed a closed form expression to compute the bias in generalized linear model (GLM). By modifying the likelihood in GLM, Firth (1993) proposed another method of reducing the estimate bias. Heinze and Schemper (2002), and later Heinze (2006) showed that Firth's method also solves the problem of separation in logistic regression.

There is a lot of literature on incomplete data modeling. For early developments we refer to Wilks (1932) and for recent development we refer to Little and Rubin (2002). Most of these works focus on missing outcomes, but in practice covariates in regression models are often missing. Ibrahim (1990) proposed a likelihood based approach for the incomplete covariate problem in the generalized linear model (*GLM*), where the MLEs were estimated using the EM algorithm (Dempster et al., 1977). With missing covariates, Little (1992) pursued a bunch of approaches for estimating the regression parameters. Horton and Laird (2001) modified the maximum likelihood estimation in the presence of missing covariate by incorporating auxiliary information.

In this article, we have proposed two methods to compute the bias of the estimates in logistic regression with missing values, where the covariates are assumed to be categorical and partially missing under the assumption that the probability of observing $\boldsymbol{x}$ given y and other covariates does not depend on any unobserved data. This assumption on missing data mechanism is called missing at random, popularly known as MAR. Considering Ibrahim's (1990) proposed likelihood, first we have derived a closed form bias expression following Cox and Snell (1968); second we have reduced the bias by incorporating the Firth's novel approach in Ibrahim's likelihood. We have further showed that the Firth type modification in Ibrahim's likelihood led to a second order bias reduction. From now onwards we will refer the method as proposed by Ibrahim (1990) as *Ibrahim*, the closed form bias expression as the *CF*, and the Firth type modification in Ibrahim's likelihood as the *Firth*.

In the following, Section 2 presents a motivating example using real data; Section 3 presents the likelihood equation of *Ibrahim*, which is optimized to get the MLE in logistic regression; Section 4 presents the derivation of the proposed closed form bias expression using the log-likelihood of *Ibrahim*, and the Firth type likelihood modification (*Firth*) in *Ibrahim*; Section 5 presents the formula to compute the standard error of the estimates using Louis (1982); Section 6 presents the simulation studies and the data analysis using real data; and finally Section 7 draws the discussion and conclusion. The appendix of this article presents the theoretical justification of the second order bias reduction ($O(n^{-2})$) using *Firth*.

2. Motivating example: liver cancer study using EST data

We consider data from Eastern Cooperative Oncology Group clinical trails EST 2282 (Falkson et al., 1990) and EST 1286 (Falkson et al., 1995). To motivate the importance of bias correction for small sample estimates, we consider a subset of this data containing only the first 45 observations. We consider the covariates *Fetoprtn* (alpha fetoprotein), *Antigen* (antihepatitis B antigen), *Jaundice* (a biochemical marker: 1, if present and 0, if absent), age (in years) and the response variable *Y*, which represents the number of cancer infected liver cells at the beginning of the clinical trial. Since the chance of survival is lower for those patients who have higher cancerous liver cells, we dichotomize the response variable *Y* and create a new response variable *survive* where *survive*=1 if $Y \leq 5$ and 0 otherwise. There are 21% missing values in the covariates, of which the variable *Fetoprtn* has 5.8% missing and the variable *Antigen* has 18% missing. We fit the model *survive=Fetoprtn+Antigen+Jaundice+age* to analyze the binary responses. The results corresponding to the covariates are exhibited in Table 1.

Note that in Table 1, the variable *age* is significant in *Ibrahim* and *Firth* (p=0.031 and p=0.047) at the 5% level, however, the same is highly insignificant using CF (p=0.119). Now, the pivotal question is: 'Which method should be preferred for inference?' In the following sections we investigate further to address this question.

Table 1
Estimates from Ibrahim, Firth and CF using the EST dataset.

Method	Parameter	Estimate	Std. err.	Lower CI	Upper CI	*P*-value
Ibrahim	Intercept	−3.856	2.117	−8.005	0.294	0.069
	Fetoprtn	−0.164	0.849	−1.828	1.500	0.847
	Antigen	0.152	0.859	−1.531	1.835	0.859
	Jaundice	−0.792	0.878	−2.514	0.930	0.367
	Age	0.065	0.030	0.006	0.124	0.031
Firth	Intercept	−3.178	1.948	−6.996	0.640	0.103
	Fetoprtn	−0.141	0.806	−1.722	1.439	0.861
	Antigen	0.154	0.821	−1.455	1.762	0.851
	Jaundice	−0.696	0.863	−2.388	0.996	0.420
	Age	0.054	0.027	0.001	0.107	0.047
CF	Intercept	−3.017	2.117	−7.166	1.133	0.154
	Fetoprtn	−0.153	0.849	−1.818	1.511	0.857
	Antigen	0.106	0.859	−1.577	1.789	0.902
	Jaundice	−0.580	0.878	−2.301	1.142	0.509
	Age	0.047	0.030	−0.012	0.106	0.119

3. MLE from the missing data

This section assumes that all responses are fully observed and only covariates are partially missing. For response distribution we use Bernoulli, and all the covariates with missing values are categorical. Suppose $y_1, y_2, \ldots, y_n$ are independent Bernoulli random variables, and let $\boldsymbol{x}$ represent the matrix of regressors' of order $n \times p$, where p is the number of regressors. So the i th row of $\boldsymbol{x}$ that corresponds to y_i consists of the i th component of the predictors. For completely observed data assume $\pi_i \equiv P(y_i = 1|\boldsymbol{x}_i,\boldsymbol{\beta})$, where $\boldsymbol{\beta} \equiv (\beta_0,\beta_1,\beta_2,\ldots,\beta_p)'$ are the unknown parameters. Then, in logistic regression the dependency of π_i and $\boldsymbol{x}_i$ can be written as $\log(\pi_i/(1-\pi_i)) = \boldsymbol{x}_i'\boldsymbol{\beta}$, and the likelihood of the same can be written as $L = \Pi_{i=1}^{n} P(y_i|\boldsymbol{x}_i,\boldsymbol{\beta})$. Let $P(\boldsymbol{x}_i|\boldsymbol{\alpha})$ be the marginal density of $\boldsymbol{x}_i$, where $\boldsymbol{\alpha}$ is the indexing nuisance parameter. Therefore, for $\boldsymbol{\theta} \equiv (\boldsymbol{\beta},\boldsymbol{\alpha})$, the complete log likelihood is $l(\boldsymbol{\theta},\boldsymbol{x},y) = \Sigma_i l(\boldsymbol{\theta},\boldsymbol{x}_i,y_i) = \Sigma_i\{l_{y_i|x_i}(\boldsymbol{\beta}) + l_{x_i}(\boldsymbol{\alpha})\}$,where $l_{y_i|x_i}(\boldsymbol{\beta}) = \log(P(y_i|\boldsymbol{x}_i,\boldsymbol{\beta}))$, and $l_{x_i}(\boldsymbol{\alpha}) = \log(P(\boldsymbol{x}_i|\boldsymbol{\alpha}))$. Suppose $\boldsymbol{x}_i = (\boldsymbol{x}_{obs,i}\ \ \boldsymbol{x}_{mis,i})$; following Ibrahim (1990) the E-step at the $(t+1)$-th iteration of log-likelihood using EM algorithm becomes

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n}\sum_{\mathbf{x}_{mis,i}(j)} w_{ij}^{(t)}\ l(\boldsymbol{\theta};\mathbf{x}_i,y_i) = \sum_{i=1}^{n}\sum_{\mathbf{x}_{mis,i}(j)} w_{ij}^{(t)}\log[P(y_i|\mathbf{x}_i,\boldsymbol{\beta})] + \sum_{i=1}^{n}\sum_{\mathbf{x}_{mis,i}(j)} w_{ij}^{(t)}\log[P(\mathbf{x}_i|\boldsymbol{\alpha})] = Q^{(1)}(\boldsymbol{\beta}|\boldsymbol{\theta}^{(t)}) + Q^{(2)}(\boldsymbol{\alpha}|\boldsymbol{\theta}^{(t)}), \quad (1)$$

where $\boldsymbol{\theta}^{(t)}$ is the estimate obtained at the t th iteration in EM. The inner sum in Eq. (1) is taken on all possible distinct missing pattern indexing j for each subject i. For categorical covariates using the Bayes theorem, $w_{ij}^{(t)}$ can be written as

$$w_{ij}^{(t)} = P(\mathbf{x}_{mis,i}(j)|\mathbf{x}_{obs,i},y_i,\boldsymbol{\theta}^{(t)}) = \frac{P(y_i|\mathbf{x}_{mis,i}(j),\mathbf{x}_{obs,i},\boldsymbol{\theta}^{(t)})\ P(\mathbf{x}_{mis,i}(j),\mathbf{x}_{obs,i}|\boldsymbol{\theta}^{(t)})}{\sum_{\mathbf{x}_{mis,i(j)}} P(y_i|\mathbf{x}_i,\boldsymbol{\theta}^{(t)})\ P(\mathbf{x}_i|\boldsymbol{\theta}^{(t)}).} \quad (2)$$

Since the E-step in Eq. (1) is the sum of two functions with two distinct parameters, the maximization M-step involves two separate maximizations. In the M-step, to maximize $Q^{(1)}$ the score function becomes $U(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{\mathbf{x}_{mis,i(j)}} w_{ij}\boldsymbol{x}_i(y_i-\pi_i) = 0$, and the information matrix $I(\boldsymbol{\beta})$ becomes $I(\boldsymbol{\beta}) = \boldsymbol{x}'\boldsymbol{\pi}(1-\boldsymbol{\pi})w\boldsymbol{x}$. The maximization of $Q^{(1)}$ is done iteratively using the relation $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + I(\boldsymbol{\beta}^{(t)})^{-1}U(\boldsymbol{\beta}^{(t)})$. Since all covariates are categorical, the reasonable marginal distribution of $\boldsymbol{x}$ is multinomial, and therefore the maximization of $Q^{(2)}$ can be done in usual way. By repeating the E and M step separately, one can estimate the parameters $\boldsymbol{\theta} \equiv (\boldsymbol{\beta},\boldsymbol{\alpha})$.

4. Bias correction in MLE

This section presents two methods to get bias corrected estimates in logistic regression of data containing missing values. Since the likelihood function satisfy the following regularity conditions as stated in Chapter 9 of Cox and Hinkley (1979)

- The parameter space is of finite dimension as well as a closed set, and the true parameter value is an interior point of that set.
- Any two distinct points from the parameter space yields two distinct probability distributions.
- The first three derivatives of log-likelihood function with respect to the parameter θ exist almost-surely in the neighborhood of the true parametric value. Furthermore, in such a neighborhood, n^{-1} times the absolute value of the third derivative is bounded above by a function of the random variable with finite expectation.
- The identity $E(U(\boldsymbol{\theta})U'(\boldsymbol{\theta})) = -E(\partial U(\boldsymbol{\theta})/\partial\boldsymbol{\theta})$ holds, and is finite positive definite in the neighborhood of the true value of $\boldsymbol{\theta}$ where $U(\boldsymbol{\theta})$ is the score function.

we can derive the bias expressions. In the following, Section 4.1 presents the proposed closed form expression that can be incorporated on existing methods, and then Section 4.2 presents the likelihood based modification as proposed by Firth (1993).

4.1. The proposed corrected form by subtraction: CF

In further discussion, we will use w_{ij} instead of $w_{ij}^{(t)}$ since we will use final estimates of θ obtained by satisfying some convergence criterion. Using the estimates from the previous section, now we reduce its $O(n^{-1})$ order bias by subtraction. For that we calculate the exact $O(n^{-1})$ order bias of the estimates by following equation (20) of Cox and Snell (1968); for s th component of the parameter vector the bias is

$$Bias(\hat{\beta}_s) = \sum_{r,t,u}\frac{1}{2}I^{rs}I^{tu}\{K_{rtu} + 2J_{t,ru}\} = \sum_{r,t,u}\frac{1}{2}\left(\sum_i\sum_{\mathbf{x}_{mis,i}(j)} w_{ij}\pi_i(1-\pi_i)x_{ir}x_{is}\right)^{-1}\left(\sum_i\sum_{\mathbf{x}_{mis,i}(j)} w_{ij}\pi_i(1-\pi_i)x_{it}x_{iu}\right)^{-1}$$
$$\times\left[-\sum_i\sum_{\mathbf{x}_{mis,i}(j)} w_{ij}(1-2\pi_i)\pi_i(1-\pi_i)x_{ir}x_{it}x_{iu} - 2.\left(\sum_i\sum_{\mathbf{x}_{mis,i}(j)} w_{ij}\{E(y_i)-\pi_i\}x_{it}\right)\left(\sum_i\sum_{\mathbf{x}_{mis,i}(j)} w_{ij}\pi_i(1-\pi_i)x_{ir}x_{iu}\right)\right],$$

where

$$K_{rst}=E\left(\sum_j \frac{\partial^3 \log P_j(y_j\boldsymbol{\beta})}{\partial\beta_r\beta_s\beta_t}\right),\quad J_{t,ru}=E\left(\sum_j \frac{\partial \log P_j(y_j\boldsymbol{\beta})}{\partial\beta_t}\frac{\partial^2 \log P_j(y_j,\boldsymbol{\beta})}{\partial\beta_r\beta_u}\right),$$

and $I^{\cdot\cdot}$ are the inverse of the Fisher information matrix. Also, the expectation in the bias is given by

$$E(y_i)=\sum_{y_i} y_i \exp\left(\sum_{\mathbf{x}_{mis,i}(j)} w_{ij} l(\boldsymbol{\theta},\boldsymbol{x}_i,y_i)\right)=\mu_i, \tag{3}$$

where the expectation of y_i is taken with respect to the adjusted likelihood. Therefore, the s th component of bias corrected estimate is

$$\hat{\beta}_{CF_s}=\hat{\beta}_s-Bias(\hat{\beta}_s).$$

The existence of bias is ensured by the regularity conditions, since they justify the Taylor expansion of the log-likelihood to derive the bias expression.

4.2. The likelihood modification like firth: firth

To reduce bias of order $O(n^{-1})$ in GLM, Firth (1993) proposed a method where the likelihood is penalized by a non-informative Jeffreys' prior. Therefore, in a logistic regression, the estimates with reduced bias can be obtained by maximizing the likelihood $L^*(\boldsymbol{\beta})=L(\boldsymbol{\beta})|I(\boldsymbol{\beta})|^{1/2}$ instead of $L(\boldsymbol{\beta})$, where $|I(\boldsymbol{\beta})|^{1/2}$ is Jeffreys (1946) invariant prior. Heinze and Schemper (2002), and Heinze (2006) found that Firth's modified likelihood also solves the problem of separation in logistic regression. Following Firth (1993), in missing value setup if we modify the log-likelihood, the E-step at $(t+1)$-th iteration can be written as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\equiv l^*(\boldsymbol{\theta},\boldsymbol{x},y)=\sum_{i=1}^{n}\sum_{\mathbf{x}_{mis,i}(j)} w_{ij}^{(t)}\{\log[P(y_i|\mathbf{x}_i,\boldsymbol{\beta})]+\frac{1}{2}\log|I(\boldsymbol{\beta})|\}+\sum_{i=1}^{n}\sum_{\mathbf{x}_{mis,i}(j)} w_{ij}^{(t)}\log[P(\mathbf{x}_i|\boldsymbol{\alpha})].$$

For M-step, the score equation $U(\boldsymbol{\beta})$ is replaced by the following:

$$U(\boldsymbol{\beta})^*=\sum_{i=1}^{n}\sum_{\mathbf{x}_{mis,i}(j)} w_{ij}[y_i-\pi_i+h_i(1/2-\pi_i)]\boldsymbol{x}_i=0, \tag{4}$$

where h_i's are the diagonal element of the 'hat' matrix $H=\varsigma^{1/2}\boldsymbol{x}(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}\boldsymbol{x}'\varsigma^{1/2}$ with $\varsigma=diag(\boldsymbol{\pi}(1-\boldsymbol{\pi}))$.

Now, as before following the Cox and Snell (1968), we derive the exact bias expression of the s th component $Bias(\hat{\beta}_s)$ of the estimates:

$$\begin{aligned}
&=\sum_{r,t,u}\frac{1}{2}\left(\sum_i\sum_{\mathbf{x}_{mis,i}(j)} w_{ij}\left[1+h_i+\pi_i\frac{\partial h_i}{\partial\pi_i}-\frac{1}{2}\frac{\partial h_i}{\partial\pi_i}\right]\frac{\partial\pi_i}{\partial\beta_s}x_{i_r}\right)^{-1}\left(\sum_i\sum_{\mathbf{x}_{mis,i}(j)} w_{ij}\left[1+h_i+\pi_i\frac{\partial\, h_i}{\partial\pi_i}-\frac{1}{2}\frac{\partial h_i}{\partial\pi_i}\right]\frac{\partial\pi_i}{\partial\beta_u}x_{i_t}\right)^{-1}\\
&\times\left[\sum_i\sum_{\mathbf{x}_{mis,i}(j)} w_{ij}\left(\left\{2\frac{\partial h_i}{\partial\pi_i}+\frac{\partial^2 h_i}{\partial\pi_i}\left(\pi_i-\frac{1}{2}\right)\right\}\pi_i(1-\pi_i).+\left\{\left(1+h_i+\left(\pi_i-\frac{1}{2}\right)\frac{\partial h_i}{\partial\pi_i}\right)(1-2\pi_i)\right\}\right)\pi_i(1-\pi_i)x_{i_r}x_{i_t}x_{i_u}\right.\\
&\left.+2\left(E\left\{\sum_i\sum_{\mathbf{x}_{mis,i}(j)} w_{ij}\left(y_i-\pi_i+\frac{h_i}{2}-h_i\pi_i\right)\right\}x_{i_t}\right)\left(\sum_i\sum_{\mathbf{x}_{mis,i(j)}} w_{ij}\left\{-1-h_i-\pi_i\frac{\partial h_i}{\partial\pi_i}+\frac{1}{2}\frac{\partial h_i}{\partial\pi_i}\right\}\pi_i(1-\pi_i)x_{i_r}x_{i_u}\right)\right],
\end{aligned}$$

where $\partial\pi_i/\partial\beta_a=\pi_i(1-\pi_i)x_{ia}$ with β_a as the a th component of $\boldsymbol{\beta}$ and $E(y_i)=\sum_{y_i} y_i\exp(\sum_{\mathbf{x}_{mis,i}(j)} w_{ij}l(\boldsymbol{\theta},\boldsymbol{x}_i,y_i))$. The above bias expression involves first order and second order derivative of h_i. In the appendix, we have shown the exact form of the derivatives in terms of matrix notation. Under the regularity conditions and using all those expressions, we get the bias corrected estimate of Firth to be of order $O(n^{-2})$.

5. Calculation of information by louis method

We calculate the standard error of the estimate by using Louis (1982) method which decomposes the complete data information into two parts: information from observed data and information associated with missing data. Louis method keeps up the notion of the EM algorithm. So the observed information for all the estimates in this incomplete set up is given by

$$\begin{aligned}
I(\boldsymbol{\theta}|obs)&=E(I(\boldsymbol{\theta}|com)|obs)-E(U(\boldsymbol{\theta}|com)U'(\boldsymbol{\theta}|com)|obs)+E(U(\theta|com)|obs)E(U'(\boldsymbol{\theta}|com)|obs)\\
&=-\ddot{Q}(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})-E(U(\boldsymbol{\theta}|com)U'(\boldsymbol{\theta}|com)|obs)+\sum_i\dot{Q}_i(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})\dot{Q}_i(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})',
\end{aligned}$$

where $\boldsymbol{\theta}=(\boldsymbol{\beta},\boldsymbol{\alpha})$ as mentioned before; *com* and *obs* represent complete and observed data, respectively. $\mathbf{Q}$ is also mentioned before for the appropriate method and the number of dots on $\mathbf{Q}$ represents the order of the derivative. $U(\boldsymbol{\theta}|com)$ is the score

function based on the complete data. Following Ibrahim et al. (2005) the form of the components of observed information matrix can be given as $\ddot{Q}(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) = \sum_i \sum_{\boldsymbol{x}_{mis,i(j)}} w_{ij} \partial^2 l(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$. The score function of complete data is $U(\boldsymbol{\theta}|com) = \sum_i \partial l(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)/\partial\boldsymbol{\theta}$ and lastly the observed score function is $\dot{Q}_i = \sum_{x_{mis,i(j)}} w_{ij} \partial l(\boldsymbol{\theta}; \boldsymbol{x}_i, y_i)/\partial\boldsymbol{\theta}$.

Next, we provide the component-wise expression of the observed information for the Firth type modified likelihood. The kl th term of the first component matrix is $\ddot{Q}(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})_{kl} = \sum_i \sum_{\boldsymbol{x}_{mis,i(j)}} w_{ij}[-1-h_i+\partial h_i/\partial\pi_i(\frac{1}{2}-\pi_i)]\pi_i(1-\pi_i)x_{ik}x_{il}$. The score function for complete data is as before, and $\dot{Q}_i(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})$ is the observed score function provided in Section 4.2. Here h_i is the hat matrix as described during bias calculation and its derivative is shown in the appendix. So, by substituting them properly we get the observed information matrix for the entire parameter vector $\boldsymbol{\theta}$. Now, from the likelihood it is clear that $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are orthogonal and hence $I(\hat{\boldsymbol{\theta}})$ will be a block diagonal matrix. After inverting this matrix we have considered the upper $p \times p$ block as the dispersion matrix of $\boldsymbol{\beta}$ for both methods. $\hat{\boldsymbol{\theta}}$ and w_{ij} are replaced by the closed form estimates and the firth type estimates in their respective formula.

6. Numerical study

6.1. Simulation study

This section presents a simulation study to determine the impact of the bias correction by the discussed methods for sample sizes of n=45, 60, 80, 120. For each sample size n, the simulation is carried out based on 5000 replications. First, we have generated three covariates, x_1, x_2 and x_3 from Normal (0,0.5), Bernoulli (0.5) and Bernoulli ($logit(-1+x_1)$) distributions, respectively. For each replication we have generated the responses from Bernoulli distribution with the probability model $\log(\pi_i/(1-\pi_i)) = (-1-x_1+x_2-x_3)$, where as mentioned earlier $\pi_i = P(y_i = 1)$. Throughout the simulation study, x_1 is kept fully observed and the Bernoulli random variables x_2 and x_3 are allowed to be missing at random in some cases. By using the mechanism, $logit(P(x_2 = missing)) = -1+2x_1-3y$ and $logit(P(x_3 = missing)) = -3+x_1-y+I(m)$, we have generated some missing values in x_2 and x_3, respectively, where $I(m)$=1 if x_2 is missing. So the missingness of x_2 and x_3 depends on both x_1 and the response variable y. The missing probability of x_3 also depends on the missingness of x_2. The percentage of missing values varies from sample to sample. In the following table we provide the output of our simulation studies. The first column of the table contains **n** representing the sample size. The next three columns represent the average missing percentage (*Miss. Pct.*), the parameter of interest (*Param*), and the used method (*Method*). The first four columns represent the average parameter estimates (*Estimate*), the average standard errors (*Std. Error*), the expected coverage (*Expt. Coverage*), and the expected length (*Expt. Length*). The expected coverage and the expected lengths of the confidence intervals are computed using the formulae $\Sigma I(\hat{\beta}_j)/R$ and $\Sigma I(\lambda(\hat{\beta}_j))/R$, where R represents the number of valid replications. I(.) is an indicator function indicating whether or not the true $\boldsymbol{\beta}$ is in the estimated confidence interval, and $\lambda(\hat{\beta}_j)$ is the length of the confidence interval.

Bias correction is meaningful when the sample size is small or medium sized. The simulation results in Table 2 using small or medium sized data exhibit that all the parameters are over estimated using *Ibrahim*. This over estimation is severe when the sample size is small, and it reduces with an increase in the sample size. For example, throughout the simulation corresponding to x_2 variable, the true parameter value is $\beta_2 = 1$. When the sample sizes are 45, 60, 80, 120, the estimated values corresponding to β_2 using Ibrahim are 1.209, 1.195, 1.123, 1.030. Except for a few cases, this behavior is consistent through the simulation. The few cases occur because of the randomness of the data generation and the employed missing data generation pattern. When the sample size is small, the method *Firth* is expected to do well because of its ability to handle separation situations. For n=45, we find that corresponding to x_3 the *Firth* estimate is -1.001 (when the true $\beta_3 = -1$), and same from closed form *CF* is -0.871. There is a partial improvement for n=60 where the estimate corresponding to x_2 is better from *CF* than *Firth* but for x_3 *Firth* is still better. This is because of having separation in the simulation where now the number of separation is fewer than the same in n=45. We further find that for n=80 the estimates using closed form *CF* are the winner, as the estimated values are closer to the true values. For n=120 the estimates from *CF* and *Firth* are close to each other. The simulations with medium to large data are expected to have fewer separations. As we find the fewer number of separation cases for n=80, 120 than the same for lower samples, and the fact that *CF* is the winner, we can safely assume that *CF* works better on medium sized data when there is no separation.

The standard errors of all methods are computed using Louis (1982). Since the covariance matrices corresponding to *Ibrahim* and *CF* are same, the estimated average standard errors from the simulation are same. However, since the covariance matrix in *Firth* is obtained by penalizing the likelihood, the standard errors are different from that of the *Ibrahim* or *CF*. Throughout the simulation, we find that all the methods achieve the required 95% coverage, however, in terms of expected length *Firth* is always shorter. In fact, the smaller standard error obtained in *Firth* translates into this shorter expected length.

6.2. Example: effect of consolidation on Illinois school district study

This section presents an example of 'The effects of consolidation on Illinois School Districts' study from Gilliland (2008). In the original study the impact of the consolidation of school districts on different characteristics of the schools were examined. In this section, we focus on a particular question of the data that deals with the activities of schools to reduce

Table 2
Table showing simulation result using the model $\log(\pi_i/1-\pi_i)=(-1-x_1+x_2-x_3)$.

n	Miss. Pct.	Param.	Method	Estimate	Std. error	Expt. coverage	Expt. length
45	24.594	Intercept	Ibrahim	−1.167	0.610	0.968	2.367
			CF	−1.095	0.610	0.990	2.394
			Firth	−1.042	0.584	0.977	2.283
		x_1	Ibrahim	−1.346	0.874	0.962	3.369
			CF	−1.024	0.874	0.995	3.426
			Firth	−1.148	0.825	0.980	3.215
		x_2	Ibrahim	1.209	0.861	0.959	3.341
			CF	0.909	0.861	0.987	3.344
			Firth	1.061	0.809	0.975	3.157
		x_3	Ibrahim	−1.135	0.940	0.972	3.689
			CF	−0.871	0.940	0.987	3.659
			Firth	−1.001	0.884	0.975	3.467
60	22.019	Intercept	Ibrahim	−1.140	0.527	0.963	2.045
			CF	−1.098	0.527	0.990	2.069
			Firth	−1.040	0.508	0.971	1.985
		x_1	Ibrahim	−1.181	0.823	0.955	3.198
			CF	−0.994	0.823	0.989	3.229
			Firth	−1.055	0.792	0.968	3.090
		x_2	Ibrahim	1.195	0.697	0.950	2.703
			CF	1.041	0.697	0.985	2.732
			Firth	1.077	0.674	0.966	2.631
		x_3	Ibrahim	−1.093	0.815	0.967	3.206
			CF	−0.940	0.815	0.982	3.205
			Firth	−0.978	0.783	0.969	3.079
80	23.496	Intercept	Ibrahim	−1.102	0.451	0.953	1.751
			CF	−1.068	0.451	0.977	1.764
			Firth	−1.027	0.439	0.961	1.718
		x_1	Ibrahim	−1.191	0.552	0.959	2.141
			CF	−1.012	0.552	0.983	2.163
			Firth	−1.087	0.533	0.970	2.078
		x_2	Ibrahim	1.123	0.625	0.949	2.432
			CF	1.013	0.625	0.972	2.446
			Firth	1.034	0.609	0.961	2.379
		x_3	Ibrahim	−1.075	0.730	0.963	2.858
			CF	−0.984	0.730	0.980	2.867
			Firth	−0.968	0.706	0.967	2.769
120	22.094	Intercept	Ibrahim	−1.045	0.384	0.952	1.495
			CF	−1.028	0.384	0.970	1.500
			Firth	−0.992	0.376	0.957	1.472
		x_1	Ibrahim	−1.060	0.537	0.955	2.094
			CF	−0.941	0.537	0.971	2.107
			Firth	−0.978	0.523	0.960	2.045
		x_2	Ibrahim	1.030	0.520	0.950	2.029
			CF	0.955	0.520	0.964	2.036
			Firth	0.967	0.509	0.957	1.994
		x_3	Ibrahim	−1.086	0.685	0.971	2.696
			CF	−0.980	0.685	0.976	2.696
			Firth	−0.988	0.660	0.969	2.599

violence among their students. The binary response variable *Q3a* is coded 1 if the participants believe that there is a change in the activities and 0 otherwise. The data consists of the covariates *Before_A*, for 'before after', with two levels corresponding to before and after consolidation; *Position* consisting of five categories coded 1 to 5 representing the position of the person who has filled the survey on behalf of the school – assistant principal, principal, superintendent, dean and board member; *Years_In* represents administrative or board member position in years classified into four categories 0–2 years, 3–5 years, 5–10 years and over 10 years; *District Enrolment* abbreviated as *District* is with four categories coded 1 to 4 –under 300, 301–600, 601–1000 and over 1000, respectively; *School* is the school enrolment with four categories. There are 85 observations in this study among which there are 49 (58%) observations where at least one of the variables *Position*, *Years_In*, *District* and *School* contains a missing value. In the original study Gilliland (2008) fit the following model:

$$Q3a \sim Before\_A + Position + Years\_In + Distric + School$$

and reported the variable *Before_A* as a significant factor of the above model. However, considering the small size of the data (85 observations) and the huge presence of missing values (more than 50% missing values) raise a question of the validity of the above inference. We re-run the same model using *Ibrahim*, *Firth* and *CF* and the results are presented in Table 3.

Table 3
Results of the school children behavior study.

		Estimate	Std. err.	Lower	Upper	*P*-value
Ibrahim	Intercept	−9.493	3.264	−15.891	−3.096	0.004
	Before_A	1.887	0.811	0.298	3.477	0.020
	Position	0.421	0.453	−0.468	1.309	0.353
	Years_In	1.921	0.640	0.666	3.176	0.003
	District	−0.436	0.484	−1.385	0.513	0.368
	School	1.043	0.790	−0.507	2.592	0.187
Firth	Intercept	−7.560	2.798	−13.044	−2.076	0.007
	Before_A	1.535	0.712	0.140	2.929	0.031
	Position	0.297	0.421	−0.528	1.122	0.481
	Years_In	1.517	0.556	0.428	2.606	0.006
	District	−0.330	0.446	−1.204	0.544	0.460
	School	0.825	0.704	−0.554	2.204	0.241
CF	Intercept	−6.975	3.264	−13.373	−0.578	0.033
	Before_A	1.444	0.811	−0.146	3.033	0.075
	Position	0.233	0.453	−0.656	1.121	0.608
	Years_In	1.408	0.640	0.153	2.663	0.028
	District	−0.296	0.484	−1.245	0.653	0.541
	School	0.559	0.790	−0.990	2.109	0.479

Notice that in Table 3 the results corresponding to *Before_A* is not significant at 5% level in *CF*. In fact, according to the survey, almost all of the data were collected in rural school districts, where the need for a formal program to reduce violence may be unnecessary.

7. Conclusion and discussion

In this article, through simulation and real data analysis we have demonstrated the importance of bias correction of estimates in missing value setup. In the estimation process, using small sample, *Ibrahim* often encounters separation as a result of violation of some of the regularity conditions. In such situations the method *CF* does not work and then *Firth* comes as a rescuer. We have analytically shown that both methods correct bias of order two, although most of the times *CF* is slightly better. In a complete case analysis, the proposed method *CF* reduces to Cordeiro and McCullagh (1991). In complete case analysis Maiti and Pradhan (2008) studied the bias of the estimates; they found that the methods proposed by Cordeiro and McCullagh (1991) and Firth (1993) work equally well, however, for samples with less likelihood of separation Cordeiro and McCullagh (1991) performs slightly better; in missing value setup, we also concur with the same conclusion.

In computing the standard error, and hence all the confidence intervals, we applied Louis (1982) in all methods. In logistic regression, generally the main point of interest is the estimate of the regression coefficients, often people relies more on confidence intervals than on the point estimates of the regression coefficients. Here, we find that *Firth* produces the lower standard errors. Our simulation study indicates that irrespective of the sample size, the expected lengths of the confidence interval from *Firth* are always smaller, and yet preserve the nominal coverage level.

The confidence intervals based on the Wald method is known to perform poor on small-sample. Here, in this article, all of the intervals are computed using Wald method. In our simulation study, we find that even though the estimates from *CF* have the smallest bias, the CIs from the same have the highest coverage among the three. Since the higher coverage probabilities translate into a lower significance level, the confidence interval from *CF* is the most conservative one among the three. In a complete case analysis, Heinze (2006) showed that confidence intervals inverting likelihood ratio statistic based on profile penalized likelihood work better than Wald. A similar kind of approach may be appropriate in the current set up; however, we are leaving this problem for future research.

The proposed *CF* is derived using the likelihood as proposed by *Ibrahim*. There are many likelihood based approaches to handle missing data (for example see Ibrahim et al., 2005). Our proposed methods can be used in any likelihood based approach that handles missing data; one may find it interesting, and we leave it for future research.

Acknowledgments

The authors would like to thank the anonymous referees for their helpful comments. The second author's research was partially supported by the U.S. National Science Foundation Grant SES 0904055.

Appendix

Here we have shown the derivatives involved in the bias expression in matrix notation. The expressions are

$$\frac{\partial h}{\partial \pi} = \varsigma^* \boldsymbol{x}(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}\boldsymbol{x}'\sqrt{\varsigma} + \sqrt{\varsigma}\boldsymbol{x}(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}\boldsymbol{x}'\varsigma^* - \sqrt{\varsigma}\boldsymbol{x}(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}(w\boldsymbol{x}'\varsigma_1^*\boldsymbol{x})(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}\boldsymbol{x}'\sqrt{\varsigma},$$

$$\begin{aligned}\frac{\partial^2 h}{\partial \pi^2} &= \varsigma^{**}\boldsymbol{x}(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}\boldsymbol{x}'\sqrt{\varsigma} + 2\varsigma^*\boldsymbol{x}(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}\boldsymbol{x}'\varsigma^* - \varsigma^*\boldsymbol{x}(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}(w\boldsymbol{x}'\varsigma_1^*\boldsymbol{x})(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}\boldsymbol{x}'\sqrt{\varsigma} + \sqrt{\varsigma}\boldsymbol{x}(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}\boldsymbol{x}'\varsigma^{**} \\ &\quad - \sqrt{\varsigma}\boldsymbol{x}(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}(w\boldsymbol{x}'\varsigma_1^*\boldsymbol{x})(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}\boldsymbol{x}'\varsigma^* - \frac{\partial}{\partial\pi}(\sqrt{\varsigma}\boldsymbol{x}(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}(w\boldsymbol{x}'\varsigma_1^*\boldsymbol{x})(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}\boldsymbol{x}'\sqrt{\varsigma}),\end{aligned}$$

where the last term above is given by

$$\begin{aligned}\frac{\partial}{\partial\pi}(\sqrt{\varsigma}\boldsymbol{x}(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}(w\boldsymbol{x}'\varsigma_1^*\boldsymbol{x})(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}\boldsymbol{x}'\sqrt{\varsigma}) = [&\varsigma^*\boldsymbol{x}(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}(w\boldsymbol{x}'\varsigma_1^*\boldsymbol{x})(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}\boldsymbol{x}'\sqrt{\varsigma} \\ &- \sqrt{\varsigma}\boldsymbol{x}(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}(w\boldsymbol{x}'\varsigma_1^*\boldsymbol{x})(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}(w\boldsymbol{x}'\varsigma_1^*\boldsymbol{x})(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}\boldsymbol{x}'\sqrt{\varsigma} \\ &+ \sqrt{\varsigma}\boldsymbol{x}(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}(w\boldsymbol{x}'\varsigma_2^*\boldsymbol{x})(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}\boldsymbol{x}'\sqrt{\varsigma} \\ &- \sqrt{\varsigma}\boldsymbol{x}(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}(w\boldsymbol{x}'\varsigma_1^*\boldsymbol{x})(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}(w\boldsymbol{x}'\varsigma_1^*\boldsymbol{x})(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}\boldsymbol{x}'\sqrt{\varsigma} \\ &+ \sqrt{\varsigma}\boldsymbol{x}(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}(w\boldsymbol{x}'\varsigma_1^*\boldsymbol{x})(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}\boldsymbol{x}'\varsigma^*],\end{aligned}$$

and these ς^*,ς^{**},ς_1^* and ς_2^* are the derivatives of the matrices with specific forms. They are $\varsigma^* = \partial\sqrt{\varsigma}/\partial\pi = diag\{\sqrt{(1-2\pi_i)/2\sqrt{\pi_i(1-\pi_i)}}\}$, $\varsigma^{**} = \partial\varsigma^*/\partial\pi$, $\varsigma_1^* = \partial\varsigma/\partial\pi = diag\{(1-2\pi_i)\}$, $\varsigma_2^* = \partial\varsigma_1^*/\partial\pi = diag(-2)$ and $\varsigma_2^* = \partial\varsigma_1^*/\partial\pi = diag(-2)$.

The derivative involves the i th diagonal element of the above matrices. Now the terms involved in the Firth-corrected bias expression are as follows with their respective order. First, the 'hat' matrix $H = \varsigma^{1/2}\boldsymbol{x}(w\boldsymbol{x}'\varsigma\boldsymbol{x})^{-1}\boldsymbol{x}'\varsigma^{1/2}$ whose i th element is used with $\varsigma = diag(\boldsymbol{\pi}(1-\boldsymbol{\pi}))$. Assuming the elements of $\boldsymbol{x}$ are bounded, we have h_i=O(n^{-1}). Using this for $\partial h_i/\partial\pi_i$ from the above expression in matrix we get it is O(n^{-1}). Similarly we get $\partial^2 h/\partial\pi^2 = O(n^{-1})$. Also $\sum_{i=1}^{n}\boldsymbol{x}_{ir}\boldsymbol{x}_{is} = O(n)$. So combining all these information order of the first two expressions are come out as O(n^{-1}) each and hence the product has the order O(n^{-2}). Along with these we also have $\sum_i h_i\boldsymbol{x}_{it} = O(1)$, $\sum_i \partial h_i/\partial\pi_i\boldsymbol{x}_{ir}\boldsymbol{x}_{iu} = O(1)$ and $\sum_i \partial^2 h_i/\partial\pi_i^2\boldsymbol{x}_{ir}\boldsymbol{x}_{it}\boldsymbol{x}_{iu} = O(1)$. Hence the order of the bias term as a whole is O(n^{-2}).

References

Cordeiro, G.M., McCullagh, P., 1991. Bias reduction in generalized linear models. Journal of the Royal Statistical Society, Series B 53, 629–643.

Cox, D.R., Hinkley, D.V., 1979. Theoretical Statistics. Chapman & Hall, London.

Cox, D.R., Snell, E.J., 1968. A general distribution of residuals (with discussion). Journal of the Royal Statistical Society, Series B 30, 248–275.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39, 1–22.

Falkson, G., Cnaan, A., Simson, I.W., 1990. A randomized phase II study of acivicin and 4 deoxydoxorubicin in patients with hepatocellular carcinoma in an Eastern Cooperative Oncology Group Study. American Journal of Clinical Oncology 13, 510–515.

Falkson, G., Lipsitz, S., Borden, E., Simson, W., Haller, D., 1995. A ECOG randomized phase II study of beta interferon and Menogoril. American Journal of Clinical Oncology 18, 287–292.

Firth, D., 1993. Bias reduction of maximum likelihood estimates. Biometrika 80, 27–38.

Gilliland, D., 2008. The effects of consolidation on Illinois school districts. Ph.D. Thesis, Western Illinois University.

Heinze, G., 2006. A comparative investigation of methods for logistic regression with separated or nearly separated data. Statistics in Medicine 25, 4216–4226.

Heinze, G., Schemper, M., 2002. A solution to the problem of separation in logistic regression. Statistics in Medicine 21, 2409–2419.

Horton, N.J., Laird, N.M., 2001. Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information. Biometrics 57, 34–42.

Ibrahim, J.G., 1990. Incomplete data in generalized linear model. Journal of American Statistical Association 85, 765–769.

Ibrahim, J.G., Chen, M.H., Lipsitz, S.R., Herring, A.H., 2005. Missing-data methods in generalized linear model: a comparative review. Journal of American Statistical Association 100, 332–346.

Jeffreys, H., 1946. An invariant form for the prior probability in estimation problem. Proceedings of Royal Society, Series A 186, 453–461.

Little, R.J.A., 1992. Regression with missing X's: a review. Journal of American Statistical Association 87, 1227–1237.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data. John Wiley and Sons, New York.

Louis, T.A., 1982. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society, Series B 44, 226–233.

Maiti, T., Pradhan, V., 2008. A comparative study of the bias corrected estimates in logistic regression. Statistical Methods in Medical Research 27, 621–634.

Wilks, S.S., 1932. Moments and distributions of estimates of population parameters from fragmentary samples. Annals of Mathematical Statistics 3, 163–195.