Published in final edited form as:

J Nonparametr Stat. 2017 ; 29(3): 577–593. doi:10.1080/10485252.2017.1339306.

Reducing Bias for Maximum Approximate Conditional Likelihood Estimator with General Missing Data Mechanism

Jiwei Zhao*

Department of Biostatistics, State University of New York at Buffalo, Buffalo, NY, USA

Abstract

In missing data analysis, the assumption of the missing data mechanism is crucial. Under different assumptions, different statistical methods have to be developed accordingly; however, in reality this kind of assumption is usually unverifiable. Therefore a less stringent, and hence more flexible, assumption is preferred. In this paper, we consider a generally applicable missing data mechanism, which includes various instances in all three scenarios: missing completely at random, missing at random, and missing not at random. Under this general missing data mechanism, we introduce the conditional likelihood and its approximate version as the base for estimating the unknown parameter of interest. Since this approximate conditional likelihood uses the completely observed samples only, it may result in large estimation bias, which could deteriorate the statistical inference and also jeopardize other statistical procedure. To tackle this problem, we propose to use some resampling techniques to reduce the estimation bias. We consider both the Jackknife and the Bootstrap in our paper. We compare their asymptotic biases through a higher order expansion up to $O(n^{-1})$. We also derive some results for the mean squared error in terms of estimation accuracy. We conduct comprehensive simulation studies under different situations to illustrate our proposed method. We also apply our method to a prostate cancer data analysis.

Keywords

Missing data mechanism; approximate conditional likelihood; resampling; higher order asymptotic expansion; bias

1. Introduction

Substantial instances of missing data are a serious problem that undermines the scientific conclusions from clinical trials, observation data and many other biomedical studies. Various statistical methods compensating for missing data have been developed in the literature. As Little et al. (2012) pointed out, the validity of these methods depends on the mechanism leading to missing data. Rubin (1976) classified the missing data mechanism into three different scenarios. If the propensity of missing data depends only on the observed values, the mechanism is called missing at random, which implies that recorded characteristics can

*Corresponding author. zhaoj@buffalo.edu.

Disclosure Statement

No potential conflict of interest was reported by the author.

account for differences in the distribution of missing variables for observed and missing cases. The missing at random includes the missing completely at random as a special case, which means that the propensity equals a constant. If the propensity depends not only on observed data but also on unobserved ones, the mechanism is called missing not at random. The missing data mechanism is a useful taxonomy that distinguishes missing data methods. Statistical methods developed under each specific scenario can be found in Little and Rubin (2002); Tsiatis (2006); Kim and Shao (2013) and among others.

However, the assumption imposed on the missing data mechanism is usually unfortunately unverifiable. Hence if the mechanism assumption is too restrictive to be correctly specified, it's difficult to tell the validity of the method and the correctness of the conclusion. A plausible direction, and also a promising remedy to this problem, is to impose a generally applicable assumption on the missing data mechanism. Motivated by Kalbfleisch (1978), we propose a conditional likelihood function in a regression model with general missing data mechanism. This general mechanism assumption includes many specific situations in the three scenarios discussed above. This conditional likelihood brings an infeasible computation burden with order $n!$. To resolve this issue, we propose an approximate version of this conditional likelihood, which reduces the computation burden from the order $n!$ to n^2 . Our approximate conditional likelihood includes the pairwise pseudo likelihood in Liang and Qin (2000) as a special case.

The approximate conditional likelihood only gets the completely observed samples involved. Consequently the corresponding maximum approximate conditional likelihood estimator (MACLE), although asymptotically unbiased, may have a large estimation bias in finite samples. This considerable issue was also noted in Liang and Qin (2000). In finite samples, a large bias aggravates the estimation accuracy, distorts the statistical inference (testing a hypothesis and constructing a confidence interval), and also affects other statistical procedures (screening and selecting predictors). For instance, Kosmidis (2014) presented a case study showing some direct consequences in statistical inference that the bias in the estimation of a single nuisance parameter can have, even if all parameters of interest are estimated with negligible bias.

To alleviate this issue, many different methods can be implemented. Some biascorrection methods directly estimate the higher order bias derived from the asymptotic expansion, however, an obvious disadvantage is that, they rely on the closed-form expression of the leading term in the asymptotic bias, which is a tedious or even infeasible task for many complex models. In this paper, we propose to use some resampling methods to reduce the estimation bias in the approximate conditional likelihood with general missing data mechanism. We consider both the (delete-one) Jackknife and the (nonparametric) Bootstrap. Although resampling techniques are also computational intensive, they only rely on a one-step procedure which is always computational feasible and not demanding.

The general theory for the Jackknife and the Bootstrap can be found in Quenouille (1949, 1956); Tukey (1958); Miller (1974); Efron (1982); Efron and Tibshirani (1993); Hall (1992); Shao and Tu (1995); Davidson and Hinkley (1997) and among many others. In this paper, for independent and identically distributed observations with missing data, we use the U-

statistics and higher-order expansion techniques to derive the asymptotic bias up to the order of $O(n^{-1})$ for the original MACLE, the Jackknife estimator, and the Bootstrap estimator, respectively. It can be seen that, the asymptotic bias of the original MACLE estimator is of order $O(n^{-1})$ with an explicit formula of its leading term, while the asymptotic bias of the Jackknife or the Bootstrap estimator is of order $O(n^{-2})$. We also obtain some results for the mean squared error (MSE) in terms of estimation accuracy.

We organize the rest of the article as follows. In Section 2, we introduce the model set-up, the conditional likelihood and its approximation, and the three estimators to be compared: the original MACLE, the Jackknife one and the Bootstrap one. Their corresponding asymptotic behaviors are investigated in Section 3. In Section 4, we conduct comprehensive simulation studies to compare the three estimators in finite samples. We apply our method to a prostate cancer data analysis in Section 5 and conclude the article with a discussion in Section 6. Technical details are collected in the appendix.

2. Methodology

We consider a general regression model $p(Y|X; \theta)$ between a q -dimensional outcome variable Y and a p -dimensional explanatory variable X , where θ is the unknown parameter of main interest. Assume that we have a collection of independent observations $\{y_i, x_i\}$, $i = 1, \dots, n$, where (y_i, x_i) 's are identically distributed realizations of (Y, X) .

When variables from (Y, X) have missing values, we use the indicator R to denote whether the data from each subject are completely observed. Without loss of generality, we assume the first n_1 subjects are fully observed with $r_i = 1$, $i = 1, \dots, n_1$, and the remaining $n - n_1$ subjects contain missing components with $r_i = 0$, $i = n_1 + 1, \dots, n$.

For the ease of presentation, we denote $W = (R, Y^T, X^T)^T$ and its realization as $w_i = (r_i, y_i^T, x_i^T)^T$. In missing data analysis, the model $\Pr(R = 1|Y, X)$ is called the missing data mechanism. In this paper, we consider a general missing data mechanism

$$\Pr(R = 1|Y, X) = s(Y)t(X), \quad (1)$$

where s and t are functions not necessarily to be known or specified.

We emphasize that the assumption (1) is very flexible and generally applicable. Firstly, (1) only assumes that $\Pr(R = 1|Y, X)$ can be written as the multiplier of an X -only function and a Y -only function. We do not impose any specific form on s or t , therefore it is robust against misspecification of s or t function. Also, it includes various specific missing-data situations. Some instances are, if $s(\cdot)$ is a constant and only X is fully observed, the mechanism is missing at random; if $t(\cdot)$ is a constant and only Y is fully observed, the mechanism is missing not at random.

In the literature, to study the complex missing data mechanism, especially for the nonignorable missingness, some other semi/non-parametric assumptions were explored. For

example, Kim and Yu (2011) proposed a semiparametric logistic regression model and Xie, Qian, and Qu (2011) considered a generalized additive missing data model. They were more focusing on either estimating the nonignorable missing data mechanism itself or investigating the impact of the nonignorable missingness. Our assumption (1) here is different from theirs. We propose (1) mainly due to the idea of the conditional likelihood and our focus is more on estimating the unknown parameter θ in the regression model $p(Y|X; \theta)$.

Under assumption (1), one very important fact is

$$p(Y|X, R=1) = \frac{\Pr(R=1|Y, X)}{\Pr(R=1|X)} p(Y|X) = \frac{s(Y)t(X)}{w(X)} p(Y|X), \quad (2)$$

where $w(X) = \Pr(R=1|X)$. (2) reveals that, the *direct* maximum likelihood estimate from a biased sample (the fully observed subjects) will result in biased estimates and incorrect conclusions unless $\Pr(R=1|Y, X) = \Pr(R=1|X)$.

Instead note that the function on the right-hand side of (2) preserves to be the multiplier of an X -only function and a Y -only function. This important property motivates us to consider an *indirect* maximum likelihood estimate. We consider the joint likelihood function for completely observed samples:

$$p(y_1, \dots, y_{n_1} | x_1, \dots, x_{n_1}, r_1 = \dots = r_{n_1} = 1; \theta).$$

In particular, we decompose $(y_1, \dots, y_{n_1})$ into its rank statistic ξ and order statistic

$\eta = (y_{(1)}, \dots, y_{(n_1)})$, and we only focus on the following conditional likelihood

$$\begin{aligned} p(\xi | \eta, x_1, \dots, x_{n_1}, r_1 = \dots = r_{n_1} = 1; \theta) \\ &= \frac{\prod_{i=1}^{n_1} p(y_i | x_i, r_i = 1; \theta)}{\sum_c \prod_{i=1}^{n_1} p(y_{(i)} | x_i, r_i = 1; \theta)} \\ &= \frac{\prod_{i=1}^{n_1} p(y_i | x_i; \theta)}{\sum_c \prod_{i=1}^{n_1} p(y_{(i)} | x_i; \theta)}, \end{aligned}$$

where the summation $\sum_c$ is over all possible permutations of $\{1, \dots, n_1\}$. Note that the second equality above uses the fact (2). The nuisance functions $s(\cdot)$, $t(\cdot)$ and $w(\cdot)$ are all canceled out.

Conceptually we can maximize the above conditional likelihood and use the corresponding estimator as the base for statistical inference, however, in reality we encounter a tremendous computational burden with an order of $n_1!$ (Liang and Qin 2000). To tackle this computation issue, we consider the following m -wise *approximate* conditional likelihood:

$$\prod_{1 \leq i_1 < i_2 < \dots < i_m \leq n_1} \frac{\prod_{j=1}^m p(y_{i_j} | x_{i_j}; \theta)}{\sum_{c_m} \prod_{j=1}^m p(y_{(i_j)} | x_{i_j}; \theta)}, \quad (3)$$

where c_m indicates all possible permutations of $\{i_1, \dots, i_m\}$, $2 \leq m \leq n_1$. Out of all possible values of m , $m = 2$ is of particular interest (Liang and Qin 2000). When $m = 2$, the *approximate* conditional likelihood becomes a pairwise fashion:

$$\prod_{1 \leq i < j \leq n_1} \frac{p(y_i | x_i; \theta) p(y_j | x_j; \theta)}{p(y_i | x_i; \theta) p(y_j | x_j; \theta) + p(y_i | x_j; \theta) p(y_j | x_i; \theta)}. \quad (4)$$

The parameter identification is a notorious issue in missing data analysis, especially with data missing not at random. The previous literature discussed this issue for either missing response (Tang, Little, and Raghunathan 2003; Shao and Zhao 2013; Zhao and Shao 2015) or missing covariate (Fang, Zhao, and Shao 2017). Under our set-up, it can be seen that not all components of the unknown parameter θ retain in (3), i.e., not all components of the unknown parameter θ are estimable. An intuitive explanation is that, we use the conditional likelihood idea and some terms in the likelihood function are canceled out through the conditioning, and therefore the corresponding information contained can not be captured.

To simplify the approximate conditional likelihood and clearly tell what components of the unknown parameter θ are estimable, we consider one commonly used model, the generalized linear model (GLM; McCullagh and Nelder (1989)) for a scalar response variable, i.e., with a canonical link, the conditional distribution of Y given X belongs to a GLM with the following density:

$$p(Y|X; \theta) = \exp[\phi^{-1}\{Y_\eta - b(\eta)\} + c(y; \phi)], \quad (5)$$

where b and c are known functions, $\eta = \alpha + \beta^T X$, $\theta = (\alpha, \beta^T, \phi)^T$, ϕ represents the positive dispersion parameter. Under the GLM assumption, the negative counterpart of the log-version of the approximate conditional likelihood (4), after adding a normalizing constant, can be written as

$$\begin{aligned}\mathcal{L}(\gamma) &= \frac{2}{n_1(n_1-1)} \sum_{1 \leq i < j \leq n_1} \log \{1 + \exp(-y_{i\wedge j} \gamma^T \mathbf{x}_{i\wedge j})\} \\ &\propto \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} r_i r_j \log \{1 + \exp(-y_{i\wedge j} \gamma^T \mathbf{x}_{i\wedge j})\},\end{aligned}\quad (6)$$

where $y_{i\wedge j} = y_i - y_j$, $\mathbf{x}_{i\wedge j} = \mathbf{x}_i - \mathbf{x}_j$ and $\gamma = \phi^{-1} \beta$. It's clear that only γ is estimable, i.e., only dispersion-scaled version of β is estimable due to the presence of missing data and the adoption of the conditional likelihood. We will be only focusing on (6) in the following of the paper. In Section 3, we can show that $\mathcal{L}(\gamma)$ has a unique global minimizer and we denote it as $\hat{\gamma}$. Readers can find some discussion on the higher order approximate conditional likelihood in the Discussion section.

Due to the fact that, the construction of $\hat{\gamma}$ only uses the completely observed samples with $i = 1, \dots, n_1$, in reality the estimation bias and accuracy of $\hat{\gamma}$ becomes a considerable issue. Therefore we propose to reduce the estimation bias using some resampling techniques. We consider both the delete-one Jackknife and the nonparametric Bootstrap methods. For the Jackknife estimator $\hat{\gamma}_J$: firstly, for $k = 1, \dots, n$, we obtain $\hat{\gamma}^{(-k)}$ by minimizing the analog of (6) only using the data index by $1, \dots, n$ except k , i.e.,

$$\hat{\gamma}^{(-k)} = \arg \min_{\gamma} \frac{2}{(n-1)(n-2)} \sum_{\substack{1 \leq i < j \leq n \\ i, j \neq k}} r_i r_j \log \{1 + \exp(-y_{i\wedge j} \gamma^T \mathbf{x}_{i\wedge j})\},$$

then the average of the n estimators is $\bar{\gamma}^{(J)} = \frac{1}{n} \sum_{i=1}^n \hat{\gamma}^{(-i)} = \frac{1}{n} \left(\sum_{k=1}^{n_1} \hat{\gamma}^{(-k)} + (n - n_1) \hat{\gamma} \right)$,

and we propose our Jackknife estimator $\hat{\gamma}_J$ as

$$\hat{\gamma}_J = n \hat{\gamma} - (n-1) \bar{\gamma}^{(J)} = \left(n_1 + 1 - \frac{n_1}{n} \right) \hat{\gamma} - \frac{n-1}{n} \sum_{k=1}^{n_1} \hat{\gamma}^{(-k)}. \quad (7)$$

The construction of the Bootstrap estimator $\hat{\gamma}_B$ is as follows: for $b = 1, \dots, B$, sample data $w_1^{*(b)}, \dots, w_n^{*(b)}$ with replacement from the original data $w_1, \dots, w_n$ then derive the $\hat{\gamma}^{*(b)}$ from the analog of (6) using the Bootstrap data, i.e.,

$$\hat{\gamma}^{*(b)} = \arg \min_{\gamma} \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} r_i^{*(b)} r_j^{*(b)} \log \{1 + \exp(-y_{i\wedge j}^* \gamma^T \mathbf{x}_{i\wedge j}^*)\},$$

and denote the average of these n estimators as $\bar{\gamma}^{(B)} = \frac{1}{B} \sum_{b=1}^B \hat{\gamma}^{*(b)}$. Then we propose our Bootstrap estimator $\hat{\gamma}_B$ as

$$\hat{\gamma}_B = 2\hat{\gamma} - \bar{\gamma}^{(B)} = 2\hat{\gamma} - \frac{1}{B} \sum_{b=1}^B \hat{\gamma}^{*(b)}. \quad (8)$$

To conclude this Section, we discuss the algorithmic aspect of computing $\hat{\gamma}$, $\hat{\gamma}_J$ and $\hat{\gamma}_B$. We claim that the computation of $\hat{\gamma}$ can be implemented by calculating the maximum likelihood estimator of a regular logistic regression without the intercept term, therefore, any available software, for example package `glm` in R, can achieve this and the computation speed is very fast. Afterwards, the computation of $\hat{\gamma}_J$ and $\hat{\gamma}_B$ straightforward and also very fast. To see this, we only have to note

$$\begin{aligned} \mathcal{L}(\gamma) &= \frac{2}{n_1(n_1-1)} \sum_{1 \leq i < j \leq n_1} \log \left\{ 1 + \exp(-y_{\Lambda j} \gamma^T \mathbf{x}_{\Lambda j}) \right\} \\ &= \frac{2}{n_1(n_1-1)} \sum_{1 \leq i < j \leq n_1} \log \left\{ 1 + \exp(-\text{sign}(y_{\Lambda j}) |y_{\Lambda j}| \gamma^T \mathbf{x}_{\Lambda j}) \right\} \\ &= \frac{1}{N} \sum_{k=1}^N \log \left\{ 1 + \exp(z_k \gamma^T \mathbf{v}_k) \right\}, \end{aligned}$$

where we let $\text{sign}(\cdot)$ denote the sign function, $z_k = -\text{sign}(y_{\Lambda j})$ and $\mathbf{v}_k = \mathbf{x}_{\Lambda j} |y_{\Lambda j}|$, $N = n_1(n_1 - 1)/2$. If we define

$$u_k = \begin{cases} 1 & \text{if } y_{\Lambda j} > 0 \\ 0 & \text{if } y_{\Lambda j} < 0, \end{cases}$$

it can be seen that, $\mathcal{L}(\gamma)$ can be regarded as the negative log-likelihood function of a regular logistic regression with response u_k , covariate $\mathbf{v}_k$, without the intercept term.

3. Asymptotic Theory

In this section, we concern the theoretical properties of the three estimators introduced in last section: the original estimator $\hat{\gamma}$, the Jackknife estimator $\hat{\gamma}_J$ in (7) and the Bootstrap estimator $\hat{\gamma}_B$ in (8). For ease of presentation we denote $\nabla^s f$ as the s -th order gradient of the function f , where s is a positive integer. From (6), note that our score equation has the following form

$$\nabla \mathcal{L}(\gamma) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \zeta_{ij}(\gamma) = \mathbf{0},$$

where $\zeta_{ij}(\gamma) = \frac{-r_i r_j y_{\Lambda j} \mathbf{x}_{\Lambda j}}{1 + \exp(y_{\Lambda j} \gamma^T \mathbf{x}_{\Lambda j})}$. We also use a subindex $_0$ to denote the true value of its corresponding unknown parameter.

Firstly, due to the complexity of the missing data mechanism and the approximate conditional likelihood, we rigorously justify that $\mathcal{L}(\gamma)$ is a legitimate loss function. In the following, at the population level we show the true value γ_0 is the unique global minimizer of $E(\mathcal{L}(\gamma))$, where $E(\cdot)$ is the expectation under the truth of the unknown parameter.

THEOREM 1 Under assumptions (1), (5), $E(\nabla \mathcal{L}(\gamma))|_{\gamma=\gamma_0} = 0$, and γ_0 is the global minimizer of $E(\mathcal{L}(\gamma))$.

The proof is provided in the Appendix. Furthermore, it can be checked that

$$\nabla^2 \mathcal{L}(\gamma) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \frac{r_i r_j^2 \exp(y_{i,j} \gamma^T x_{i,j})}{\left(1 + \exp(y_{i,j} \gamma^T x_{i,j})\right)^2} x_{i,j} x_{i,j}^T$$

is a positive definite matrix, therefore, even at the finite sample level $\mathcal{L}(\gamma)$ still has a unique minimizer. Recall that we discussed how to efficiently locate this unique minimizer at the end of last section.

Next we investigate the higher order property of estimator $\hat{\gamma}$. We adopt the approach that can synthesize the empirical distribution and the true cumulative distribution of the data. In

detail, we let $\hat{F}$ represent the empirical distribution that puts the weight $\binom{n}{2}^{-1}$ on each of the

pair data $(\mathbf{w}_i, \mathbf{w}_j)$, then our score equation is to say that $\int \zeta_{i,j}(\gamma) d\hat{F} = 0$. Furthermore, let's

denote F as the joint cumulative distribution function of $(\mathbf{w}_1, \mathbf{w}_2)$, and write

$F_\epsilon = F + \epsilon \Delta = F + \epsilon \sqrt{n}(\hat{F} - F)$, where $\epsilon \in [0, 1/\sqrt{n}]$. Denote $\gamma(\epsilon)$ as the solution of

$\int \zeta_{i,j}(\gamma) dF_\epsilon = 0$. It's clear that $\hat{\gamma} = \gamma(1/\sqrt{n})$, and the true value $\gamma_0 = \gamma(0)$. The following result provides the higher order expansion of $\hat{\gamma}$ around the truth γ_0 .

THEOREM 2 Assume (1), (5), also assume the s -th order derivative of $\mathcal{L}(\gamma)$ defined in (6) exists in a neighborhood of γ_0 in the parameter space and $E(\|\nabla^s \mathcal{L}(\gamma)\|^2) < \infty$ for s up to 3, where $\|\mathbf{M}\| = [\text{trace}(\mathbf{M}\mathbf{M}^T)]^{1/2}$, then we have

$$\hat{\gamma} = \gamma_0 + \frac{1}{\sqrt{n}} A_n + \frac{1}{2!} \frac{1}{n} B_n + o(n^{-1}).$$

where we define $S(\gamma) = \sqrt{n} \nabla \mathcal{L}(\gamma)$, $T(\gamma) = \sqrt{n} \{ \nabla^2 \mathcal{L}(\gamma) - E \nabla^2 \mathcal{L}(\gamma) \}$, $\Sigma_1 = E \nabla^2 \mathcal{L}(\gamma)|_{\gamma=\gamma_0}$ and

$A_n = - \sum_1^{-1} S(\gamma_0)$, $B_n = - \sum_1^{-1} \{ E \nabla^3 \mathcal{L}(\gamma_0) A_n \otimes A_n + 2T(\gamma_0) A_n \}$, $E \nabla^3 \mathcal{L}(\gamma_0)$ is a $p \times p^2$ matrix, and its j -th row is the derivative of j -th row of $E \nabla^2 \mathcal{L}(\gamma)$ with respect to γ , evaluated at γ_0 , as a row vector. The notation $\otimes$ represents the Kronecker product.

The proof of this result is given in the Appendix. The existence of Σ_1^{-1} follows from the proof of Theorem 1. Immediately we have the following corollary.

COROLLARY 1 Given the same assumptions as Theorem 2, we have

- (1) The bias of $\hat{\gamma}$, $\text{Bias}(\hat{\gamma}) = \frac{1}{2!} \frac{1}{n} E B_n + o(n^{-1}) = O(n^{-1})$;
- (2) The asymptotic distribution of $\hat{\gamma}$, $\sqrt{n}(\hat{\gamma} - \gamma_0) \xrightarrow{d} N(0, V)$, where $V = \Sigma_1^{-1} \Sigma_2 \Sigma_1^{-1}$, $\Sigma_2 = 4E[\zeta_{12}(\gamma)\zeta_{13}(\gamma)^T]_{\gamma=\gamma_0}$;
- (3) The mean squared error (MSE) of $\hat{\gamma}$, $\text{MSE}(\hat{\gamma}) = \frac{1}{n} A_n \circ A_n + \frac{1}{n^{3/2}} A_n \circ B_n + o(n^{-3/2})$, where $\circ$ represents the Hadamard product.

Our next result summarizes the asymptotic bias of the Jackknife estimator $\hat{\gamma}_J$ and the Bootstrap estimator $\hat{\gamma}_B$.

THEOREM 3 Given the same assumptions as Theorem 2, we have

$$\text{Bias}(\hat{\gamma}_J) = \text{Bias}(\hat{\gamma}_B) = O(n^{-2}).$$

Compared the conclusion of (1) in Corollary 1 with that of Theorem 3, we can see that, through either Jackknife or Bootstrap, the leading term (of order n^{-1}) of the bias of $\hat{\gamma}$ is eliminated and the corresponding bias in $\hat{\gamma}_J$ or $\hat{\gamma}_B$ is reduced substantially.

We see a higher order expansion of $\text{MSE}(\hat{\gamma})$ from the conclusion of (3) in Corollary 1. Actually, following the idea of Theorem 2, we can derive a much higher order expansion of $\hat{\gamma}$ than presented in the paper. For instance, under appropriate assumptions, similar to Theorem 2, we have

$$\hat{\gamma} = \gamma_0 + \frac{1}{\sqrt{n}} A_n + \frac{1}{2!} \frac{1}{n} B_n + \frac{1}{3!} \frac{1}{n^{3/2}} C_n + o(n^{-3/2}),$$

and therefore we can derive

$$\text{MSE}(\hat{\gamma}) = \frac{1}{n} A_n \circ A_n + \frac{1}{n^{3/2}} A_n \circ B_n + \frac{1}{4n^2} B_n \circ B_n + \frac{1}{3n^2} A_n \circ C_n + o(n^{-2}).$$

For $\text{MSE}(\hat{\gamma}_J)$ and $\text{MSE}(\hat{\gamma}_B)$, from the proof of Theorem 3, it is too complicated to spell out its readily used formula. In theory, there is no definite answer of whether the MSE is also reduced through resampling. In the literature, there exist some findings about the reduction of the bias may result in the increase of the MSE. In this paper we mainly examine the performance of our proposed estimators in terms of MSE in the following numerical studies.

4. Simulation Study

In this section, we conduct simulation studies to evaluate the finite sample performance of four estimators: the original MACLE $\hat{\gamma}$, the Jackknife estimator $\hat{\gamma}_J$, the Bootstrap estimator $\hat{\gamma}_{B1}$ with 100 Bootstrap samples and the Bootstrap estimator $\hat{\gamma}_{B2}$ with 200 Bootstrap samples. We examine two situations linear regression and logistic regression. In each of the situation, we replicate the simulation 500 times.

In the linear regression, we first generate the covariate X from the standard normal distribution $N(0, 1)$ and also the random error s independently, then we generate the response Y following $Y = \alpha_0 + \beta_0 X + \epsilon$, where $\alpha_0 = 0$ and $\beta_0 \in \{-1, 0.8, \dots, 0.8, 1\}$ (11 scenarios). Note that the marginal distribution of Y is $N(0, 1 + \beta_0^2)$. For the distribution of the missing indicator R , we generate $R = 1(Y < z_{0.7}\sqrt{1 + \beta_0^2})$ and $R = 1(Y < z_{0.7}\sqrt{1 + \beta_0^2})1(X < z_{0.9})$ (two situations) so that we will have around 60–70% samples completely observed. We let the total sample size $n = 100$. Note that this is a GLM with dispersion parameter $\sigma = 1$. Under our consideration, we can only estimate $\gamma = \beta/\sigma = \beta$. We collect three measures for the estimators of γ : the empirical bias (Bias), the average of the biases of the estimators across 500 simulation runs; the standard error (SE), the standard deviation of the estimators across 500 simulation runs; and the mean squared error (MSE), the average of the squared biases across 500 simulation runs.

Similarly, in the logistic regression, we generate the covariate X from the standard normal distribution $N(0, 1)$ and the binary outcome Y following

$$\Pr(Y = 1|X) = \frac{\exp(\alpha_0 + \beta_0 X)}{1 + \exp(\alpha_0 + \beta_0 X)},$$

where $\alpha_0 = -0.5$ and $\beta_0 \in \{-1, -0.8, \dots, 0.8, 1\}$ (11 scenarios). We generate the missing indicator R following $\Pr(R = 1|Y, X) = (1 + Y)/2$, i.e., we keep all samples with $Y = 1$ and keep around half samples with $Y = 0$, and $\Pr(R = 1|Y, X) = \mathbb{1}(X < z_{0.9})(1 + Y)/2$ (two situations). Under our model set-up, we have around 70% samples completely observed in the first situation and around 60% in the second situation. Note that in the logistic regression we can estimate $\gamma = \beta$. We consider the total sample size $n = 200$. Similarly to the linear regression situation, we collect three measures for the estimators of γ .

The simulation results for the linear regression are contained in Tables 1–2, and those for the logistic regression are in Tables 3–4. From these results we can reach the following conclusions. First, for the bias, the initial MACLE $\hat{\gamma}$ has a larger bias when $|\gamma_0|$ is getting away from zero. The three proposed estimators $\hat{\gamma}_J$, $\hat{\gamma}_{B1}$ and $\hat{\gamma}_{B2}$ all obtain a substantial reduction for the bias, especially when $|\gamma_0|$ is far from zero. This finding is consistent with our theoretical results in Corollary 1 (1) and Theorem 3. Second, for the MSE, although we don't have a theoretical justification, both Jackknife and Bootstrap estimators get a smaller MSE than the initial MACLE in our numerical study. For the linear regression the reduction is about 10–15% and for the logistic regression the reduction is about 5–10%. Third, the

difference between the Jackknife and the Bootstrap estimators is always negligible. Their performances match each other very well. This confirms the same theoretical result of the Jackknife and the Bootstrap estimators. In particular, the Bootstrap estimators with both 100 samples and 200 samples perform almost exactly the same. This finding reveals that 200 is large enough for the number of Bootstrap samples in our numerical study.

5. The Prostate Cancer Study

We analyze a data set from a study (GEO GDS3289) investigating the association between prostate cancer tumors and genomic biomarkers (Tomlins et al. 2007). The whole data set can be accessed from the website of the National Center for Biotechnology Information of the National Institutes of Health. Briefly, this data set contains $n = 104$ samples, out of which 34 are benign epithelium samples and 70 non-benign samples. In the literature, scientists have identified genes that are possibly related to the prostate cancer in a significant way in similar studies, for example, AMACR (Luo et al. 2002), SLC25A6 (Dunn et al. 2006), MME (Kalin et al. 2011) and EZH2 (Varambally et al. 2002). In our analysis we only focus on these four meaningful biomarkers. However, there are about 50% samples that don't have the values of EZH2 (53 out of 104). Too many samples are eliminated if we include EZH2 in our study but proceed to a direct logistic regression model.

To apply our proposed method, we focus on the $n_1 = 51$ samples with fully available information of all four genes. Specifically, we use $Y = 1$ if a non-benign sample and $Y = 0$ if a benign sample. There are 36 non-benign samples in our complete data set. Recall that our original logistic regression model follows

$$\text{logit Pr}(Y = 1 | \text{prostate cancer genes}) = \alpha + \gamma_1 \text{AMACR} + \gamma_2 \text{SLC25A6} + \gamma_3 \text{MME} + \gamma_4 \text{EZH2},$$

where $\text{logit}(p) = \log(p/(1-p))$. We implement the approximate conditional likelihood (6) and compute its global minimizer $\hat{\gamma}_1, \dots, \hat{\gamma}_4$, the original MACLE. Then we obtain their corresponding Jackknife estimator in (7) and Bootstrap estimator in (8). We calculate the Bootstrap estimators using both 100 and 200 Bootstrap samples. The results are contained in Table 5.

From the results, compared to the original MACLE, both Jackknife and Bootstrap estimators get a relatively smaller coefficient for positive biomarkers AMACR and SLC25A6. For negative biomarker MME, the resampling estimators obtain a smaller effect; while for negative biomarker EZH2 which has missing values in the original data set, the resampling estimators obtain a larger effect. In all, our proposed resampling estimators reveal some potential estimation bias of the original MACLE. Also, the Bootstrap estimators using both 100 and 200 samples perform fairly the same. This shows that 200 is large enough for the number of Bootstrap samples in this real data analysis. Although there is some estimation difference between the Jackknife and the Bootstrap estimators, they move to the same direction and their performances are reasonably close to each other.

6. Discussion

This paper considers an estimation problem with a general missing data mechanism. Without specifying a parametric distribution of the mechanism, we propose the conditional likelihood and its approximate version. To reduce the estimation bias and potentially increase the estimation accuracy, we propose some resampling based estimators. Another advantage of our proposed framework is that, different from Tang et al. (2003), we don't need a condition that the true value of γ can not equal to zero. Note that this kind of condition is stringent and unverifiable in reality. From the numerical studies, our method performs very well when γ_0 is close to zero.

Following the theoretical development in this paper, although it's cumbersome, we can also derive the asymptotic expansion of the MSE of $\hat{\gamma}_J$ and $\hat{\gamma}_B$. However, as we can see, even in the most simple model, due to the use of the approximate conditional likelihood, it's almost infeasible to exactly compare the MSE of $\hat{\gamma}_J$ and $\hat{\gamma}_B$ to that of the MACLE $\hat{\gamma}$. Hence, in terms of MSE, although our numerical studies show positive direction for our proposed resampling based estimators, we could not reach a theoretical conclusion in general.

The general missing data mechanism we considered includes nonignorable missingness. When the mechanism becomes nonignorable, the parameter identifiability issue is crucial. In this paper, we show that a dispersion-scaled version of the original parameter is estimable for a GLM with the canonical link. For the GLM with a non-canonical link, Zhao and Shao (2017) provided some conditions under which all components of the unknown parameter are identifiable. It's an interesting topic for the future research to consider whether the instrumental variable approach studied in Zhao and Shao (2015) can be used in the framework considered in this paper.

It's also interesting to consider higher order (than 2) approximate conditional likelihood in our problem. For example, similar to (6), the approximate conditional likelihood of order 3 is the following

$$\mathcal{L}_3(\gamma) = \frac{6}{n(n-1)(n-2)} \sum_{1 \leq i < j < k \leq n} r_i r_j r_k \log \left\{ 1 + A_{ijk}(\gamma) + B_{ijk}(\gamma) + C_{ijk}(\gamma) + D_{ijk}(\gamma) + E_{ijk}(\gamma) \right\},$$

where $A_{ijk}(\gamma) = \exp\{-y_{i\bar{j}}\gamma^T \mathbf{x}_{i\bar{j}}\}$, $B_{ijk}(\gamma) = \exp\{-y_{j\bar{k}}\gamma^T \mathbf{x}_{j\bar{k}}\}$, $C_{ijk}(\gamma) = \exp\{-y_{k\bar{i}}\gamma^T \mathbf{x}_{k\bar{i}}\}$, $D_{ijk}(\gamma) = \exp\{-y_i\gamma^T \mathbf{x}_{i\bar{j}} - y_j\gamma^T \mathbf{x}_{j\bar{k}} - y_k\gamma^T \mathbf{x}_{k\bar{i}}\}$, $E_{ijk}(\gamma) = \exp\{y_j\gamma^T \mathbf{x}_{i\bar{j}} + y_k\gamma^T \mathbf{x}_{j\bar{k}} + y_i\gamma^T \mathbf{x}_{k\bar{i}}\}$.

However, how to find a tractable and efficient algorithm to optimize this complicated function is challenging. Certainly this is already beyond the scope of this current paper and it warrants further investigation in the future.

Acknowledgements

Research reported in this publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award Number UL1TR001412. The content is solely the responsibility of the author and does not necessarily represent the official views of the NIH. The author would like to thank the Editor, the Associate Editor and two anonymous referees for their constructive comments and insightful suggestions, which have led to a significantly improved paper.

Appendix

Proof of Theorem 1. We denote ξ_{ij} and η_{ij} as the rank and order statistics of pair (y_i, y_j) . Note that $\zeta_{ij}(\gamma)$ is a continuous function with respect to γ and $y_{i\wedge j}$, then we have

$$\begin{aligned} & \mathbb{E}\left\{\zeta_{ij}(\gamma) \mid \eta_{ij}, x_i, x_j, r_i = r_j = 1\right\} \Big|_{\gamma=\gamma_0} \\ & \propto \mathbb{E}\left\{\nabla_{\gamma} \log \Pr(\xi_{ij} \mid \eta_{ij}, x_i, x_j, r_i = r_j = 1, \gamma = \gamma_0) \mid \eta_{ij}, x_i, x_j, r_i = r_j = 1\right\} \\ & = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \nabla_{\gamma} \Pr(\xi_{ij} \mid \eta_{ij}, x_i, x_j, r_i = r_j = 1, \gamma = \gamma_0) d\xi_{ij} \\ & = \nabla_{\gamma} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Pr(\xi_{ij} \mid \eta_{ij}, x_i, x_j, r_i = r_j = 1, \gamma = \gamma_0) d\xi_{ij} = 0, \end{aligned}$$

where in the last step we interchange the integration with the differentiation. Note that although the whole unknown parameter θ should be in $\Pr(\xi_{ij} \mid \eta_{ij}, x_i, x_j, r_i = r_j = 1)$, only γ takes effect in the expressions above. By the double expectation formula,

$$\mathbb{E}\left\{\zeta_{ij}(\gamma) \mid \eta_{ij}, x_i, x_j, r_i = r_j = 1\right\} \Big|_{\gamma=\gamma_0} = \mathbb{E}\left\{\mathbb{E}\left\{\zeta_{ij}(\gamma) \mid \eta_{ij}, x_i, x_j, r_i = r_j = 1\right\} \mid \eta_{ij}, x_i, x_j, r_i = r_j = 1\right\} \Big|_{\gamma=\gamma_0} = 0. \text{ This completes the proof of } \mathbb{E}(\nabla \mathcal{L}(\gamma)) \Big|_{\gamma=\gamma_0} = 0.$$

Again, by interchanging the integration with the differentiation, we obtain

$$\nabla \mathbb{E}(\mathcal{L}(\gamma)) \Big|_{\gamma=\gamma_0} = \mathbb{E}(\nabla \mathcal{L}(\gamma)) \Big|_{\gamma=\gamma_0} = 0. \text{ Hence, } \gamma_0 \text{ is a stationary point of } \mathbb{E}(\mathcal{L}(\gamma)).$$

In addition,

$$\nabla^2 \mathbb{E}(\mathcal{L}(\gamma)) = \mathbb{E}\left\{\frac{r_i r_j y_{i\wedge j}^2 \exp(y_{i\wedge j} \gamma^T x_{i\wedge j})}{\left(1 + \exp(y_{i\wedge j} \gamma^T x_{i\wedge j})\right)^2} x_{i\wedge j} x_{i\wedge j}^T\right\}$$

is positive definite. These combined together imply that γ_0 is the unique global minimizer of $\mathbb{E}(\mathcal{L}(\gamma))$, which completes the proof. ■

Proof of Theorem 2. The basic idea of this result follows Taylor's asymptotic expansion. Here we only need to derive the formulas of A_n and B_n . Throughout the proof, we denote γ^ϵ as the first order derivative of γ as the function of ϵ , and $\gamma^\epsilon(0)$ this function evaluated at 0, and similar notation for the second order derivative $\gamma^{\epsilon\epsilon}$ and realization $\gamma^{\epsilon\epsilon}(0)$.

First we have

$$0 = \int \zeta_{12}(\gamma(\epsilon)) dF_\epsilon.$$

If we take the first order derivative with respect to ϵ on both sides, we have

$$\begin{aligned}
0 &= \int \frac{d\zeta_{12}(\gamma(\epsilon))}{d\epsilon} dF_\epsilon + \int \zeta_{12}(\gamma(\epsilon)) d\Delta \\
&= \int \zeta_{12}^\gamma \gamma^\epsilon dF_\epsilon + \int \zeta_{12}(\gamma(\epsilon)) d\Delta \\
&= E_\epsilon(\zeta_{12}^\gamma) \gamma^\epsilon + \int \zeta_{12}(\gamma(\epsilon)) d\Delta,
\end{aligned}$$

and hence

$$\gamma^\epsilon = -\left(E_\epsilon(\zeta_{12}^\gamma)\right)^{-1} \int \zeta_{12}(\gamma(\epsilon)) d\Delta.$$

Therefore A_n has the following expression

$$\gamma^{\epsilon(0)} = -\left(E \nabla^2 \mathcal{L}(\gamma_0)\right)^{-1} \sqrt{n} \nabla \mathcal{L}(\gamma_0) = -\sum_1^{-1} s(\gamma_0).$$

Next if we take the second order derivative with respect to s on both sides, we similarly have

$$\begin{aligned}
0 &= \int \frac{d^2 \zeta_{12}(\gamma(\epsilon))}{d\epsilon^2} dF_\epsilon + 2 \int \frac{d\zeta_{12}(\gamma(\epsilon))}{d\epsilon} d\Delta \\
&= \int \left((\gamma^\epsilon)^T (\zeta_{12}^{\gamma\gamma})_{j=1}^p \gamma^\epsilon + \zeta_{12}^\gamma \gamma^{\epsilon\epsilon} \right) dF_\epsilon + 2 \int \zeta_{12}^\gamma \gamma^\epsilon d\Delta \\
&= (\gamma^\epsilon)^T E_\epsilon(\zeta_{12}^{\gamma\gamma})_{j=1}^p \gamma^\epsilon + E_\epsilon(\zeta_{12}^\gamma) \gamma^{\epsilon\epsilon} + 2 \int \zeta_{12}^\gamma d\Delta \gamma^\epsilon,
\end{aligned}$$

where $(\gamma^\epsilon)^T (\zeta_{12}^{\gamma\gamma})_{j=1}^p \gamma^\epsilon$ represents a p -dimensional vector, with the j -th coordinate equals to $(\gamma^\epsilon)^T (\zeta_{12}^{\gamma\gamma})_j \gamma^\epsilon$ and $(\zeta_{12}^{\gamma\gamma})_j$ represents the derivative matrix of the j -th row of ζ_{12}^γ with respect to the vector $\boldsymbol{\gamma}$ and hence

$$\gamma^{\epsilon\epsilon} = -\left(E_\epsilon(\zeta_{12}^\gamma)\right)^{-1} \left\{ (\gamma^\epsilon)^T E_\epsilon(\zeta_{12}^{\gamma\gamma})_{j=1}^p \gamma^\epsilon + 2 \int \zeta_{12}^\gamma d\Delta \gamma^\epsilon \right\}.$$

Therefore B_n has the following expression

$$\begin{aligned}
\gamma^{\epsilon\epsilon(0)} &= -\sum_1^{-1} \left\{ s(\gamma_0)^T \sum_1^{-1} (\Sigma_1^\gamma)_{j=1}^p \sum_1^{-1} s(\gamma_0) - 2T(\gamma_0) \sum_1^{-1} s(\gamma_0) \right\} \\
&= -\sum_1^{-1} \left\{ E \nabla^3 \mathcal{L}(\gamma_0) A_n \otimes A_n + 2T(\gamma_0) A_n \right\},
\end{aligned}$$

where we denote $(\Sigma_1^\gamma)_{j=1}^p$ as a sequence of matrices where the j -th matrix equals the derivative matrix of the j -th row of Σ_1 (as a vector) with respect to the vector $\boldsymbol{\gamma}$, $E \nabla^3 \mathcal{L}(\gamma_0)$ is

a $p \times p^2$ matrix, and its j -th row is the derivative of j -th row of $E \nabla^2 \mathcal{L}(\gamma_0)$ with respect to γ , evaluated at γ_0 , as a row vector. ■

Proof of Corollary 1. Notice that $EA_n = \mathbf{0}$, then the conclusions of (1) and (3) are straightforward. To prove (2), firstly Taylor expansion gives us

$$0 = \nabla \mathcal{L}(\hat{\gamma}) = \nabla \mathcal{L}(\gamma_0) + (\hat{\gamma} - \gamma_0) \nabla^2 \mathcal{L}(\gamma_0) + o_p(n^{-1/2}).$$

Therefore,

$$\sqrt{n}(\hat{\gamma} - \gamma_0) = - \left[\nabla^2 \mathcal{L}(\gamma_0) \right]^{-1} \sqrt{n} \nabla \mathcal{L}(\gamma_0) + o_p(1).$$

Since $\nabla^2 \mathcal{L}(\gamma_0)$ and $\nabla \mathcal{L}(\gamma_0)$ are both U-statistics, from the theory of U-statistic, we have the following

$$\begin{aligned} \nabla^2 \mathcal{L}(\gamma_0) &\xrightarrow{p} \Sigma_1, \\ \sqrt{n} \nabla \mathcal{L}(\gamma_0) &\xrightarrow{d} N(0, \Sigma_2), \end{aligned}$$

which completes the proof. ■

Proof of Theorem 3. For Jackknife, we have for each $i = 1, \dots, n$,

$$\hat{\gamma}^{(-i)} = \gamma_0 + \frac{1}{\sqrt{n-1}} A_{n-1}^{(-i)} + \frac{1}{2!} \frac{1}{n-1} B_{n-1}^{(-i)} + o_p((n-1)^{-1}),$$

and hence

$$\bar{\gamma}^{(J)} = \gamma_0 + \frac{1}{\sqrt{n-1}} \bar{A}_{n-1} + \frac{1}{2!} \frac{1}{n-1} \bar{B}_{n-1} + o_p((n-1)^{-1}),$$

where $\bar{A}_{n-1} = \frac{1}{n} \sum_{i=1}^n A_{n-1}^{(-i)}$, $\bar{B}_{n-1} = \frac{1}{n} \sum_{i=1}^n B_{n-1}^{(-i)}$. Due to the fact that $EA_n = 0$, we have

$EA_{n-1}^{(-i)} = E\bar{A}_{n-1} = 0$. We also have $EB_n = E\bar{B}_{n-1}$. Therefore, we have

$\text{Bias}(\hat{\gamma}) = \frac{1}{2n} EB_n + o(n^{-1})$ and $\text{Bias}(\bar{\gamma}^{(J)}) = \frac{1}{2(n-1)} E\bar{B}_{n-1} + o((n-1)^{-1})$. Based on the

Jackknife theory in Quenouille (1949, 1956); Tukey (1958), following the idea of the proof of Theorem 2, we can reinforce these two results to $\text{Bias}(\hat{\gamma}) = \frac{1}{2n} EB_n + O(n^{-2})$ and

$\text{Bias}(\bar{\gamma}^{(J)}) = \frac{1}{2(n-1)} E\bar{B}_{n-1} + o((n-1)^{-2})$. Hence, we have

$\text{Bias}(\hat{\gamma}_J) = n \text{Bias}(\hat{\gamma}) - (n-1) \text{Bias}(\bar{\gamma}^{(J)}) = O(n^{-2})$.

For Bootstrap we have for each $b = 1, \dots, B$,

$$\hat{\gamma}^{*(b)} = \hat{\gamma} + \frac{1}{\sqrt{n}}A_n^{*(b)} + \frac{1}{2n}B_n^{*(b)} + o(n^{-1}),$$

then

$$\bar{\gamma}^{(B)} = \hat{\gamma} + \frac{1}{\sqrt{n}}A_n^* + \frac{1}{2n}B_n^* + o(n^{-1}),$$

where $A_n^* = \frac{1}{B} \sum_{b=1}^B A_n^{*(b)}$ and $B_n^* = \frac{1}{B} \sum_{b=1}^B B_n^{*(b)}$. We can also derive that

$$\hat{\gamma}_B = \hat{\gamma} + \frac{1}{\sqrt{n}}(A_n - A_n^*) + \frac{1}{2n}(B_n - B_n^*) + o(n^{-1}).$$

For the b -th Bootstrap sample, if we take the expectation with respect to the Bootstrap sample, based on the Bootstrap theory, we have

$$E^{*(b)}(\hat{\gamma}^{*(b)} - \hat{\gamma}) = \frac{1}{2n}E^{*(b)}B_n^{*(b)} + o(n^{-1}) = \frac{1}{2n}E^{*(b)}B_n^{*(b)} + O(n^{-2}),$$

and if we then take the expectation with respect to the original sample on both sides, we have

$$E(\hat{\gamma}^{*(b)} - \hat{\gamma}) = \frac{1}{2n}EB_n + O(n^{-2}).$$

Therefore, we have

$$E(\bar{\gamma}^{(B)} - \hat{\gamma}) = \frac{1}{2n}EB_n + O(n^{-2}),$$

and

$$\text{Bias}(\hat{\gamma}_B) = \text{Bias}(\hat{\gamma}) - E(\bar{\gamma}^{(B)} - \hat{\gamma}) = O(n^{-2}).$$

References

- Davidson A and Hinkley D (1997), *Bootstrap Methods and Their Application*, Cambridge University Press.
- Dunn TA et al. (2006), "A novel role of myosin VI in human prostate cancer," *The American Journal of Pathology*, 169, 1843–1854. [PubMed: 17071605]
- Efron B (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, vol. 38, SIAM.
- Efron B and Tibshirani R (1993), *An Introduction to the Bootstrap*, Chapman & Hall Ltd.
- Fang F, Zhao J, and Shao J (2017), "Imputation-based adjusted score equations in generalized linear models with nonignorable missing covariate values," *Statistica Sinica*, To Appear
- Hall P (1992), *The Bootstrap and Edgeworth Expansion*, Springer.

- Kalbfleisch JD (1978), "Likelihood methods and nonparametric tests," *Journal of the American Statistical Association*, 73, 167–170.
- Kalin M et al. (2011), "Novel prognostic markers in the serum of patients with castration-resistant prostate cancer derived from quantitative analysis of the pten conditional knockout mouse proteome," *European Urology*, 60, 1235–1243. [PubMed: 21741162]
- Kim JK and Shao J (2013), *Statistical Methods for Handling Incomplete Data*, CRC Press.
- Kim JK and Yu CL (2011), "A semiparametric estimation of mean functionals with nonignorable missing data," *Journal of the American Statistical Association*, 106, 157–165.
- Kosmidis I (2014), "Bias in parametric estimation: reduction and useful side-effects," *Wiley Interdisciplinary Reviews: Computational Statistics*, 6, 185–196.
- Liang K-Y and Qin J (2000), "Regression analysis under non-standard situations: a pairwise pseudolikelihood approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 773–786.
- Little RJ and Rubin DB (2002), *Statistical Analysis with Missing Data*, Wiley, 2nd ed.
- Little RJ et al. (2012), "The prevention and treatment of missing data in clinical trials," *New England Journal of Medicine*, 367, 1355–1360. [PubMed: 23034025]
- Luo J et al. (2002), " α -Methylacyl-CoA Racemase A New Molecular Marker for Prostate Cancer," *Cancer Research*, 62, 2220–2226. [PubMed: 11956072]
- McCullagh P and Nelder JA (1989), *Generalized Linear Models*, Chapman & Hall/CRC, 2nd ed.
- Miller RG (1974), "The jackknife-a review," *Biometrika*, 61, 1–15.
- Quenouille M (1949), "Approximate Tests of Correlation in Time-Series," *Journal of the Royal Statistical Society. Series B (Methodological)*, 11, 68–84.
- Quenouille MH (1956), "Notes on bias in estimation," *Biometrika*, 43, 353–360.
- Rubin DB (1976), "Inference and missing data," *Biometrika*, 63, 581–592.
- Shao J and Tu D (1995), *The Jackknife and Bootstrap*, Springer.
- Shao J and Zhao J (2013), "Estimation in longitudinal studies with nonignorable dropout," *Statistics and Its Interface*, 6, 303–313.
- Tang G, Little RJ, and Raghunathan TE (2003), "Analysis of multivariate missing data with nonignorable nonresponse," *Biometrika*, 90, 747–764.
- Tomlins SA et al. (2007), "Integrative molecular concept modeling of prostate cancer progression," *Nature Genetics*, 39, 41–51. [PubMed: 17173048]
- Tsiatis AA (2006), *Semiparametric Theory and Missing Data*, Springer.
- Tukey JW (1958), "Bias and confidence in not-quite large samples," in *Annals of Mathematical Statistics*, vol. 29, pp. 614–614.
- Varambally S et al. (2002), "The polycomb group protein EZH2 is involved in progression of prostate cancer," *Nature*, 419, 624–629. [PubMed: 12374981]
- Xie H, Qian Y, and Qu L (2011), "A semiparametric approach for analyzing nonignorable missing data," *Statistica Sinica*, 21, 1881–1899.
- Zhao J and Shao J (2015), "Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data," *Journal of the American Statistical Association*, 110, 1577–1590.
- Zhao J and Shao J (2017), "Approximate conditional likelihood for generalized linear models with general missing data mechanism," *Journal of System Science and Complexity*, 30, 139–153.

Table 1.

Empirical Bias, Standard Error (SE), and Mean Squared Error (MSE) of four estimators of γ in linear regression with $R = 1(Y < z_{0.7}\sqrt{1 + \beta_0^2})$. $\hat{\gamma}$ =the original MACLE; $\hat{\gamma}_J$ =the Jackknife estimator; $\hat{\gamma}_{B1}$ =the Bootstrap estimator with 100 Bootstrap samples; $\hat{\gamma}_{B2}$ =the Bootstrap estimator with 200 Bootstrap samples.

		Bias	SE	MSE			Bias	SE	MSE
$\gamma_0 = -1$	$\hat{\gamma}$	-0.0563	0.2700	0.0759	$\gamma_0 = 1$	$\hat{\gamma}$	0.0619	0.2836	0.0841
	$\hat{\gamma}_J$	0.0069	0.2627	0.0689		$\hat{\gamma}_J$	-0.0019	0.2741	0.0750
	$\hat{\gamma}_{B1}$	0.0037	0.2618	0.0684		$\hat{\gamma}_{B1}$	0.0030	0.2732	0.0745
	$\hat{\gamma}_{B2}$	0.0031	0.2614	0.0682		$\hat{\gamma}_{B2}$	0.0035	0.2734	0.0746
$\gamma_0 = -0.8$	$\hat{\gamma}$	-0.0480	0.2401	0.0598	$\gamma_0 = 0.8$	$\hat{\gamma}$	0.0466	0.2451	0.0621
	$\hat{\gamma}_J$	0.0051	0.2328	0.0541		$\hat{\gamma}_J$	-0.0057	0.2359	0.0556
	$\hat{\gamma}_{B1}$	0.0015	0.2323	0.0539		$\hat{\gamma}_{B1}$	-0.0025	0.2349	0.0551
	$\hat{\gamma}_{B2}$	0.0011	0.2315	0.0535		$\hat{\gamma}_{B2}$	0.0018	0.2352	0.0552
$\gamma_0 = -0.6$	$\hat{\gamma}$	-0.0360	0.2267	0.0526	$\gamma_0 = 0.6$	$\hat{\gamma}$	0.0281	0.2148	0.0468
	$\hat{\gamma}_J$	0.0057	0.2187	0.0478		$\hat{\gamma}_J$	-0.0129	0.2047	0.0420
	$\hat{\gamma}_{B1}$	0.0026	0.2187	0.0477		$\hat{\gamma}_{B1}$	-0.0094	0.2049	0.0420
	$\hat{\gamma}_{B2}$	0.0018	0.2183	0.0476		$\hat{\gamma}_{B2}$	-0.0094	0.2050	0.0420
$\gamma_0 = -0.4$	$\hat{\gamma}$	-0.0298	0.1933	0.0382	$\gamma_0 = 0.4$	$\hat{\gamma}$	0.0233	0.2034	0.0419
	$\hat{\gamma}_J$	0.0006	0.1830	0.0334		$\hat{\gamma}_J$	-0.0051	0.1907	0.0363
	$\hat{\gamma}_{B1}$	-0.0014	0.1842	0.0339		$\hat{\gamma}_{B1}$	-0.0017	0.1915	0.0366
	$\hat{\gamma}_{B2}$	-0.0023	0.1838	0.0337		$\hat{\gamma}_{B2}$	-0.0021	0.1917	0.0367
$\gamma_0 = -0.2$	$\hat{\gamma}$	-0.0191	0.1962	0.0388	$\gamma_0 = 0.2$	$\hat{\gamma}$	0.0072	0.2005	0.0402
	$\hat{\gamma}_J$	-0.0039	0.1835	0.0336		$\hat{\gamma}_J$	-0.0065	0.1866	0.0348
	$\hat{\gamma}_{B1}$	-0.0038	0.1855	0.0344		$\hat{\gamma}_{B1}$	-0.0038	0.1883	0.0354
	$\hat{\gamma}_{B2}$	-0.0048	0.1857	0.0344		$\hat{\gamma}_{B2}$	-0.0048	0.1882	0.0354
$\gamma_0 = 0$	$\hat{\gamma}$	-0.0094	0.1919	0.0369		$\hat{\gamma}$			
	$\hat{\gamma}_J$	-0.0079	0.1787	0.0319		$\hat{\gamma}_J$			
	$\hat{\gamma}_{B1}$	-0.0070	0.1815	0.0329		$\hat{\gamma}_{B1}$			
	$\hat{\gamma}_{B2}$	-0.0083	0.1814	0.0329		$\hat{\gamma}_{B2}$			

Table 2.

Empirical Bias, Standard Error (SE), and Mean Squared Error (MSE) of four estimators of γ in linear regression with $R = 1(Y < z_{0.7}\sqrt{1 + \beta_0^2})1(X < z_{0.9})$. $\hat{\gamma}$ =the original MACLE; $\hat{\gamma}_J$ =the Jackknife estimator; $\hat{\gamma}_{B1}$ =the Bootstrap estimator with 100 Bootstrap samples; $\hat{\gamma}_{B2}$ =the Bootstrap estimator with 200 Bootstrap samples.

		Bias	SE	MSE			Bias	SE	MSE
$\gamma_0 = -1$	$\hat{\gamma}$	-0.0705	0.3558	0.1313	$\gamma_0 = 1$	$\hat{\gamma}$	0.0630	0.2970	0.0920
	$\hat{\gamma}_J$	0.0103	0.3433	0.1177		$\hat{\gamma}_J$	-0.0021	0.2870	0.0822
	$\hat{\gamma}_{B1}$	0.0050	0.3408	0.1159		$\hat{\gamma}_{B1}$	0.0017	0.2862	0.0818
	$\hat{\gamma}_{B2}$	0.0032	0.3417	0.1165		$\hat{\gamma}_{B2}$	0.0027	0.2865	0.0819
$\gamma_0 = -0.8$	$\hat{\gamma}$	-0.0597	0.3212	0.1065	$\gamma_0 = 0.8$	$\hat{\gamma}$	0.0462	0.2593	0.0693
	$\hat{\gamma}_J$	0.0084	0.3094	0.0956		$\hat{\gamma}_J$	-0.0086	0.2479	0.0614
	$\hat{\gamma}_{B1}$	0.0023	0.3079	0.0946		$\hat{\gamma}_{B1}$	-0.0051	0.2478	0.0613
	$\hat{\gamma}_{B2}$	0.0014	0.3076	0.0945		$\hat{\gamma}_{B2}$	-0.0040	0.2482	0.0615
$\gamma_0 = -0.6$	$\hat{\gamma}$	-0.0485	0.3145	0.1011	$\gamma_0 = 0.6$	$\hat{\gamma}$	0.0244	0.2300	0.0534
	$\hat{\gamma}_J$	0.0051	0.3012	0.0906		$\hat{\gamma}_J$	-0.0188	0.2176	0.0476
	$\hat{\gamma}_{B1}$	-0.0028	0.3020	0.0910		$\hat{\gamma}_{B1}$	-0.0157	0.2174	0.0474
	$\hat{\gamma}_{B2}$	-0.0025	0.3006	0.0902		$\hat{\gamma}_{B2}$	-0.0149	0.2182	0.0477
$\gamma_0 = -0.4$	$\hat{\gamma}$	-0.0465	0.2762	0.0783	$\gamma_0 = 0.4$	$\hat{\gamma}$	0.0186	0.2246	0.0507
	$\hat{\gamma}_J$	-0.0066	0.2609	0.0680		$\hat{\gamma}_J$	-0.0107	0.2081	0.0433
	$\hat{\gamma}_{B1}$	-0.0133	0.2624	0.0689		$\hat{\gamma}_{B1}$	-0.0085	0.2089	0.0436
	$\hat{\gamma}_{B2}$	-0.0129	0.2717	0.0685		$\hat{\gamma}_{B2}$	-0.0079	0.2096	0.0439
$\gamma_0 = -0.2$	$\hat{\gamma}$	-0.0307	0.2691	0.0732	$\gamma_0 = 0.2$	$\hat{\gamma}$	0.0023	0.2427	0.0588
	$\hat{\gamma}_J$	-0.0078	0.2529	0.0639		$\hat{\gamma}_J$	-0.0107	0.2224	0.0495
	$\hat{\gamma}_{B1}$	-0.0117	0.2538	0.0644		$\hat{\gamma}_{B1}$	-0.0093	0.2250	0.0506
	$\hat{\gamma}_{B2}$	-0.0117	0.2540	0.0645		$\hat{\gamma}_{B2}$	-0.0095	0.2255	0.0508
$\gamma_0 = 0$	$\hat{\gamma}$	-0.0176	0.2501	0.0627					
	$\hat{\gamma}_J$	-0.0120	0.2326	0.0541					
	$\hat{\gamma}_{B1}$	-0.0135	0.2348	0.0552					
	$\hat{\gamma}_{B2}$	-0.0139	0.2353	0.0555					

Table 3.

Empirical Bias, Standard Error (SE), and Mean Squared Error (MSE) of four estimators of γ in logistic regression with $\Pr(R = 1 | Y, X) = (1 + Y)/2$. $\hat{\gamma}$ =the original MACLE; $\hat{\gamma}_J$ =the Jackknife estimator; $\hat{\gamma}_{B1}$ =the Bootstrap estimator with 100 Bootstrap samples; $\hat{\gamma}_{B2}$ =the Bootstrap estimator with 200 Bootstrap samples.

		Bias	SE	MSE			Bias	SE	MSE
$\gamma_0 = -1$	$\hat{\gamma}$	-0.0182	0.2335	0.0548	$\gamma_0 = 1$	$\hat{\gamma}$	0.0274	0.2326	0.0547
	$\hat{\gamma}_J$	0.0121	0.2277	0.0519		$\hat{\gamma}_J$	-0.0034	0.2265	0.0512
	$\hat{\gamma}_{B1}$	0.0128	0.2258	0.0511		$\hat{\gamma}_{B1}$	-0.0034	0.2263	0.0511
	$\hat{\gamma}_{B2}$	0.0130	0.2251	0.0507		$\hat{\gamma}_{B2}$	-0.0041	0.2255	0.0507
$\gamma_0 = -0.8$	$\hat{\gamma}$	-0.0257	0.2286	0.0528	$\gamma_0 = 0.8$	$\hat{\gamma}$	0.0307	0.2279	0.0528
	$\hat{\gamma}_J$	-0.0005	0.2215	0.0490		$\hat{\gamma}_J$	0.0063	0.2212	0.0489
	$\hat{\gamma}_{B1}$	-0.0034	0.2245	0.0503		$\hat{\gamma}_{B1}$	0.0069	0.2216	0.0490
	$\hat{\gamma}_{B2}$	-0.0025	0.2239	0.0500		$\hat{\gamma}_{B2}$	0.0061	0.2216	0.0490
$\gamma_0 = -0.6$	$\hat{\gamma}$	-0.0100	0.1975	0.0390	$\gamma_0 = 0.6$	$\hat{\gamma}$	0.0086	0.1997	0.0399
	$\hat{\gamma}_J$	0.0084	0.1921	0.0369		$\hat{\gamma}_J$	-0.0088	0.1941	0.0377
	$\hat{\gamma}_{B1}$	0.0087	0.1914	0.0366		$\hat{\gamma}_{B1}$	-0.0100	0.1934	0.0374
	$\hat{\gamma}_{B2}$	0.0076	0.1917	0.0367		$\hat{\gamma}_{B2}$	-0.0107	0.1932	0.0373
$\gamma_0 = -0.4$	$\hat{\gamma}$	-0.0239	0.1880	0.0358	$\gamma_0 = 0.4$	$\hat{\gamma}$	0.0191	0.1902	0.0365
	$\hat{\gamma}_J$	-0.0121	0.1832	0.0336		$\hat{\gamma}_J$	0.0064	0.1848	0.0341
	$\hat{\gamma}_{B1}$	-0.0116	0.1837	0.0338		$\hat{\gamma}_{B1}$	0.0063	0.1859	0.0345
	$\hat{\gamma}_{B2}$	-0.0111	0.1831	0.0336		$\hat{\gamma}_{B2}$	0.0068	0.1848	0.0341
$\gamma_0 = -0.2$	$\hat{\gamma}$	-0.0001	0.1796	0.0322	$\gamma_0 = 0.2$	$\hat{\gamma}$	0.0070	0.1737	0.0302
	$\hat{\gamma}_J$	0.0056	0.1744	0.0304		$\hat{\gamma}_J$	0.0006	0.1691	0.0285
	$\hat{\gamma}_{B1}$	0.0054	0.1747	0.0305		$\hat{\gamma}_{B1}$	0.0005	0.1702	0.0289
	$\hat{\gamma}_{B2}$	0.0054	0.1750	0.0306		$\hat{\gamma}_{B2}$	0.0007	0.1695	0.0287
$\gamma_0 = 0$	$\hat{\gamma}$	-0.0019	0.1788	0.0319					
	$\hat{\gamma}_J$	-0.0019	0.1735	0.0300					
	$\hat{\gamma}_{B1}$	-0.0028	0.1755	0.0307					
	$\hat{\gamma}_{B2}$	-0.0019	0.1746	0.0304					

Table 4.

Empirical Bias, Standard Error (SE), and Mean Squared Error (MSE) of four estimators of γ in logistic regression with $\Pr(R=1|Y, X) = \mathbb{1}(X < z_{0.9})(1 + Y)/2$. $\hat{\gamma}$ =the original MACLE; $\hat{\gamma}_J$ =the Jackknife estimator; $\hat{\gamma}_{B1}$ =the Bootstrap estimator with 100 Bootstrap samples; $\hat{\gamma}_{B2}$ =the Bootstrap estimator with 200 Bootstrap samples.

		Bias	SE	MSE			Bias	SE	MSE
$\gamma_0 = -1$	$\hat{\gamma}$	-0.0282	0.2835	0.0848	$\gamma_0 = 1$	$\hat{\gamma}$	0.0276	0.2870	0.0829
	$\hat{\gamma}_J$	0.0112	0.2777	0.0819		$\hat{\gamma}_J$	-0.0051	0.2800	0.0783
	$\hat{\gamma}_{B1}$	0.0118	0.2758	0.0811		$\hat{\gamma}_{B1}$	-0.0055	0.2784	0.0774
	$\hat{\gamma}_{B2}$	0.0120	0.2751	0.0807		$\hat{\gamma}_{B2}$	-0.0054	0.2774	0.0768
$\gamma_0 = -0.8$	$\hat{\gamma}$	-0.0257	0.2786	0.0728	$\gamma_0 = 0.8$	$\hat{\gamma}$	0.0207	0.2779	0.0728
	$\hat{\gamma}_J$	-0.0015	0.2715	0.0690		$\hat{\gamma}_J$	0.0073	0.2712	0.0689
	$\hat{\gamma}_{B1}$	-0.0034	0.2745	0.0603		$\hat{\gamma}_{B1}$	0.0096	0.2716	0.0690
	$\hat{\gamma}_{B2}$	-0.0055	0.2739	0.0600		$\hat{\gamma}_{B2}$	0.0065	0.2716	0.0690
$\gamma_0 = -0.6$	$\hat{\gamma}$	-0.0200	0.2575	0.0690	$\gamma_0 = 0.6$	$\hat{\gamma}$	0.0186	0.2497	0.0599
	$\hat{\gamma}_J$	0.0094	0.2521	0.0669		$\hat{\gamma}_J$	-0.0079	0.2441	0.0577
	$\hat{\gamma}_{B1}$	0.0078	0.2514	0.0666		$\hat{\gamma}_{B1}$	-0.0105	0.2434	0.0574
	$\hat{\gamma}_{B2}$	0.0086	0.2517	0.0667		$\hat{\gamma}_{B2}$	-0.0117	0.2432	0.0573
$\gamma_0 = -0.4$	$\hat{\gamma}$	-0.0380	0.2324	0.0553	$\gamma_0 = 0.4$	$\hat{\gamma}$	0.0191	0.2202	0.0465
	$\hat{\gamma}_J$	-0.0227	0.2352	0.0511		$\hat{\gamma}_J$	0.0075	0.2148	0.0441
	$\hat{\gamma}_{B1}$	-0.0238	0.2261	0.0516		$\hat{\gamma}_{B1}$	0.0086	0.2159	0.0445
	$\hat{\gamma}_{B2}$	-0.0234	0.2249	0.0510		$\hat{\gamma}_{B2}$	0.0071	0.2148	0.0441
$\gamma_0 = -0.2$	$\hat{\gamma}$	0.0010	0.2283	0.0520	$\gamma_0 = 0.2$	$\hat{\gamma}$	0.0085	0.2137	0.0402
	$\hat{\gamma}_J$	0.0083	0.2209	0.0488		$\hat{\gamma}_J$	0.0017	0.2091	0.0485
	$\hat{\gamma}_{B1}$	0.0066	0.2219	0.0492		$\hat{\gamma}_{B1}$	0.0015	0.2102	0.0489
	$\hat{\gamma}_{B2}$	0.0065	0.2220	0.0492		$\hat{\gamma}_{B2}$	0.0017	0.2095	0.0487
$\gamma_0 = 0$	$\hat{\gamma}$	-0.0035	0.1988	0.0419					
	$\hat{\gamma}_J$	-0.0029	0.1935	0.0400					
	$\hat{\gamma}_{B1}$	-0.0038	0.1955	0.0407					
	$\hat{\gamma}_{B2}$	-0.0025	0.1946	0.0404					

Table 5.

Real data analysis on the association between prostate cancer tumors and four genomic biomarkers. A positive coefficient shows a direction for the non-benign sample.

Method		Prostate Cancer Genes			
		AMACR	SLC25A6	MME	EZH2
MACLE	Estimate	2.2185	8.6912	-1.1137	-0.8004
	SE	1.2042	2.2580	0.6995	1.5292
Jackknife	Estimate	1.4752	6.9139	-0.5271	-2.6656
	SE	4.0435	7.2659	1.1131	2.9620
Bootstrap1	Estimate	0.7547	5.9993	-0.6658	-1.2669
	SE	3.1252	5.5266	1.3209	3.7074
Bootstrap2	Estimate	0.7410	5.5220	-0.5813	-1.2621
	SE	3.0910	5.4649	1.4182	3.1822