Biometrics 64, 106–114
March 2008

DOI: 10.1111/j.1541-0420.2007.00828.x

Analysis of Matched Case–Control Data in Presence of Nonignorable Missing Exposure

Samiran Sinha[1],* and Tapabrata Maiti[2],**

[1]Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.
[2]Department of Statistics, Iowa State University, Ames, Iowa 50011, U.S.A.
**email:* sinha@stat.tamu.edu
***email:* taps@iastate.edu

Summary. The present article deals with informative missing (IM) exposure data in matched case–control studies. When the missingness mechanism depends on the unobserved exposure values, modeling the missing data mechanism is inevitable. Therefore, a full likelihood-based approach for handling IM data has been proposed by positing a model for selection probability, and a parametric model for the partially missing exposure variable among the control population along with a disease risk model. We develop an EM algorithm to estimate the model parameters. Three special cases: (a) binary exposure variable, (b) normally distributed exposure variable, and (c) lognormally distributed exposure variable are discussed in detail. The method is illustrated by analyzing a real matched case–control data with missing exposure variable. The performance of the proposed method is evaluated through simulation studies, and the robustness of the proposed method for violation of different types of model assumptions has been considered.

Key words: Conditional likelihood; EM algorithm; Generalized exponential family; Los Angeles endometrial cancer data; Matched case–control data; Nonignorable missingness.

1. Introduction

The article concerns nonignorable missing covariates in 1:M ($M \geq 1$) matched case–control studies. Let D be the binary disease indicator variable, $\boldsymbol{S}$ be a set of matching variables, $\boldsymbol{X}^* = (\boldsymbol{Z}, X)$ be a set of exposure variables, and R be the indicator variable of whether a subject has complete covariate information or not. We assume $\boldsymbol{Z}$ is always observed and X is partially missing. Generally, matched case–control data are collected according to the following two steps. In the first step cases, $(\boldsymbol{X}^*_{i1}, \boldsymbol{S})$ are independently sampled from $p(\boldsymbol{X}^*, \boldsymbol{S} \mid D = 1)$, for $i = 1, \ldots, n$ and in the second step controls $(\boldsymbol{X}^*_{ij})$ are drawn from $p(\boldsymbol{X}^* \mid D = 0, \boldsymbol{S} = \boldsymbol{s}_i)$, for $j = 2, \ldots, M + 1$. The risk model we are interested in is

$$pr(D = 1 \mid \boldsymbol{S}, \boldsymbol{X}^*) = H\{\beta_0(\boldsymbol{S}) + \boldsymbol{\beta}^t \boldsymbol{X}^*\}, \tag{1}$$

where $H(u) = \{1 + \exp(-u)\}^{-1}$. Here, $\beta_0(\boldsymbol{S})$ is the stratum effect defined by the variable $\boldsymbol{S}$, and $\boldsymbol{\beta}$ is the log–odds ratio parameter associated with $\boldsymbol{X}^*$. The standard method for estimating $\boldsymbol{\beta}$ is to use conditional logistic regression (CLR), conditioning on the number of cases within each matched set. However, in the presence of missing Xs, the CLR method is inefficient. If the missing value occurs for a case subject, the CLR analysis completely ignores the stratum where the case belongs to. On the other hand, for a 1:1 matched case–control study, appearance of a missing value leads to the deletion of the entire stratum where the missing value occurs. Therefore, the CLR analysis not only ignores the subject with missing X, but also sometimes deletes the entire stratum where the missing value occurs.

Following the terminology of Little and Rubin (1987), if the missingness mechanism does not depend on the observed data, i.e., data are missing completely at random (MCAR), or missingness depends on the observed data (MAR), in both situations the CLR analysis produces inefficient estimates of the parameter. Between the two approaches for handling MAR data several authors modeled the distribution of the partially missing exposure variable (Satten and Kupper, 1993; Paik and Sacco, 2000; Satten and Carroll, 2000), and some authors modeled missingness process (Lipsitz, Parzen, and Ewell, 1998; Rathouz, Satten, and Carroll, 2002). However, as noted by Rathouz (2003), if modeling of the distribution of the missing exposure variable is at all possible, then the likelihood approach adopted by Satten and Kupper (1993) yields semiparametric efficient estimator of the parameter of interest. Adopting the similar type of likelihood, Sinha et al. (2005) proposed a semiparametric Bayesian inference for handling unobserved heterogeneity in matched case–control study in the presence of missing exposure variable.

When the missing data are informative (IM), which means missingness mechanism depends on the unobserved data, failure to incorporate the missingness mechanism in the analysis may produce biased and inconsistent results. Ibrahim, Lipsitz, and Chen (1999) proposed a method for handling IM covariate data in generalized linear model. However, until recently there was no work on IM data for matched case–control study. Recently, Paik (2004) proposed a parametric approach for handling nonignorable missing data in matched case–control studies. The beauty of the proposed likelihood is that it reduces to the CLR approach when data are fully observed. The method requires (a) the disease risk model specified in (1); (b) a model for the selection probability; and (c) a model for the exposure distribution conditional on the matching variables, the completely observed covariates, and the disease status. The method comprises the following two steps. First, the parameters of the selection model and the exposure distributions are estimated through a joint likelihood of R and X given $\boldsymbol{Z}, \boldsymbol{S}$, and D. At the second step, a conditional likelihood approach is used to estimate the log–odds ratio parameters by replacing the terms associated with missing covariates by the corresponding expected value. Indeed, the parameter estimates obtained at the first step are used in the second step. Satten and Kupper (1993) and later on Satten and Carroll (2000) noted that the parameters of the case distribution are functions of the parameters of the control distribution and disease–exposure association parameter $\boldsymbol{\beta}$. Paik (2004) indeed recognizes that fact while $pr(D=1 \mid \boldsymbol{Z}, \boldsymbol{S})/pr(D=0 \mid \boldsymbol{Z}, \boldsymbol{S})$ is calculated by plugging in the inputed values X^* for the missing X in Step 2, however that parametric dependence seems to be ignored while the parameters for the exposure distribution and the selection model are estimated in Step 1, and consequently in calculating the inputed value X^*. Hence, the method essentially estimates the same odds-ratio parameter twice—once in Step 1 in terms of the parameter of the exposure distribution among the case population, and again in Step 2, which may entail some loss of efficiency. In addition, the proposed method assumes that the form of the distribution of the partially missing exposure variable is the same for the case and control population. This assumption holds as long as the distribution of the partially missing exposure variable is a member of an exponential family of distributions. As shown in Section 3 the distributions not belonging to exponential family do not satisfy this assumption, hence the method is not applicable in that situation. Therefore, a generalized method is needed to overcome these issues.

The aim of this article is to propose a full likelihood-based approach for handling informative missing (IM) data in matched case–control studies. We start with the risk model (1) and a model for the selection probability. More importantly, instead of assuming the same form of distribution for the exposure variable among the cases and the controls, we pose a parametric form of distribution for the partially missing exposure variable only among the control population. By using the disease–risk model and the exposure distribution among the control population, we are able to obtain the exposure distribution among the cases. This approach avoids the risk of double estimation of the same log–odds-ratio parameter, and importantly it does not restrict the exposure distribution among the cases and controls to be of the same form. Next, we form the joint likelihood of R, X, and D conditional on $\boldsymbol{Z}, \boldsymbol{S}$, and T, the number of cases in each matched set. The maximum likelihood estimates of the parameters are obtained by using an EM algorithm. The full likelihood-based approach allows us to estimate the standard error of the parameters by using observed Fisher's information matrix. The general theory of the proposed method for any type of exposure variable is given in Section 2, while three special cases—Bernoulli, Normal, and Lognormal distribution for the missing exposure variable—are discussed in Section 3. We apply the proposed method to analyze a matched case–control data on endometrial cancer among postmenopausal women living in Los Angeles. Among several measured covariates, obesity was missing for about 16% of the study participants, and we try to analyze the data using different methods considering obesity as a partially missing exposure variable. The details of the data analysis are collected in Section 4. Section 5 contains extensive simulation study exploring different missingness mechanism in matched case–control studies. Moreover, we study how the different methods are affected for different types of model violations. Section 6 contains concluding remarks with some discussion.

Before we conclude this section, we would like to summarize the main features of this article. We propose a full likelihood–based approach for handling IM data in matched case–control studies, allowing any kind of exposure variable. Asymptotic normality and consistency of the parameters are automatic as long as the exposure distribution and the missingness mechanism are correctly specified. The simulation study indicates superiority of the proposed methods in terms of bias and efficiency compared to the existing alternative procedures in some situations, and robustness for moderate departure from various model assumptions.

2. Method

Suppose we have n matched sets and each set comprises one case and M($\geq$1) controls. As noted earlier, $\boldsymbol{X}^* = (\boldsymbol{Z}, X)$, where $\boldsymbol{Z}$ is a $p \times 1$ vector of completely observed covariates, and X is assumed to be a scaler risk factor. Define $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \beta_2)$, where $\boldsymbol{\beta}_1$ is a $p \times 1$ vector of the log–odds-ratio parameters associated with $\boldsymbol{Z}$, and β_2 is the log–odds-ratio parameter associated with X. The indicator variable $R = 1$ if X is observed and 0 otherwise. Here, we will use i as an index for stratum and j as an index for a subject. Furthermore, we assume that within stratum i, $j = 1$ stands for a case and the rest are controls. If there is no missing value, then we estimate $\boldsymbol{\beta}$ from the conditional likelihood function

$$L_{\text{CLR}} = \prod_{i=1}^{n} P\Bigg[D_{i1}=1, D_{i2}=\cdots=D_{iM+1}=0 \mid \left\{\boldsymbol{Z}_{ij}, X_{ij}\right\}_{j=1}^{M+1}, \times \boldsymbol{S}_i, \sum_{j=1}^{M+1} D_{ij}=1\Bigg]$$

$$= \prod_{i=1}^{n} \frac{\exp(\boldsymbol{\beta}_1^{\tau}\boldsymbol{Z}_{i1}+\beta_2 X_{i1})}{\sum_{j=1}^{M+1}\exp(\boldsymbol{\beta}_1^{\tau}\boldsymbol{Z}_{ij}+\beta_2 X_{ij})}. \qquad (2)$$

Note that due to conditioning on $T = \sum_{j=1}^{M+1} D_{ij}$, L_{CLR} becomes free from the nuisance parameter $\beta_0(\boldsymbol{S}_i)$. However, in the presence of missing values we base our inference on L, the joint conditional likelihood of R, X, D given $\boldsymbol{Z}, \boldsymbol{S}$, and the conditioning event T, and it is

$$L = p(R, X, D \mid \boldsymbol{Z}, \boldsymbol{S})$$
$$= pr(R \mid X, \boldsymbol{Z}, \boldsymbol{S}, D)\, p(X \mid \boldsymbol{Z}, \boldsymbol{S}, D)\, pr(D \mid \boldsymbol{Z}, \boldsymbol{S}, T).$$

The above likelihood function is similar to that of Satten and Carroll (2000) except that they did not consider the selection model. Rathouz (2003) showed that the maximum likelihood estimator obtained from the last two terms of L is semiparametric efficient for the disease–exposure association parameter for the MAR data. We assume that the selection probability is given by

$$\begin{aligned} & pr(R_{ij}=1\mid X_{ij},\boldsymbol{Z}_{ij},D_{ij},\boldsymbol{S}_i,\boldsymbol{\delta}) \\ &\quad = H(\delta_0+\boldsymbol{\delta}^{\tau}_{Z}\boldsymbol{Z}_{ij}+\delta_X X_{ij}+\delta_D D_{ij}+\boldsymbol{\delta}_S\boldsymbol{S}_i). \qquad (3)\end{aligned}$$

Let $p(X_{ij}\mid \boldsymbol{Z}_{ij},D_{ij}=0,\boldsymbol{S}_i,\boldsymbol{\gamma},)$ be the assumed form of the distribution of the exposure variable among the control population, then by using result (A.2) of the Web Appendix, we obtain the distribution exposure among the case population $p(X_{ij}\mid \boldsymbol{Z}_{ij},D_{ij}=1,\boldsymbol{S}_i,\boldsymbol{\gamma},\boldsymbol{\beta})$. Next, we rewrite L as

$$\begin{aligned} L = \prod_{i=1}^{n}\prod_{j=1}^{M+1}\Big\{ & pr^{R_{ij}}(R_{ij}=1\mid X_{ij},\boldsymbol{Z}_{ij},D_{ij},\boldsymbol{S}_i) \\ & \times \int pr^{1-R_{ij}}(R_{ij}=0\mid X_{ij},\boldsymbol{Z}_{ij},D_{ij},\boldsymbol{S}_i)\, dF \\ & (X_{ij}\mid \boldsymbol{Z}_{ij},D_{ij},\boldsymbol{S}_i)\times p^{R_{ij}}(X_{ij}\mid \boldsymbol{Z}_{ij},D_{ij},\boldsymbol{S}_i)\Big\} \\ \times \prod_{i=1}^{n} pr\Bigg(& D_{i1}=1, D_{i2}=\cdots=D_{iM+1}=0\mid \{\boldsymbol{Z}_{ij}\}_{j=1}^{M+1}, \\ & \boldsymbol{S}_i,\sum_{j=1}^{M+1}D_{ij}=1\Bigg). \qquad (4)\end{aligned}$$

Note that, as shown in the Web Appendix (equation A.4), the last term of the likelihood L can be written as

$$\begin{aligned} & pr\Bigg(D_{i1}=1, D_{i2}=\cdots=D_{iM+1}=0\mid \{\boldsymbol{Z}_{ij}\}_{j=1}^{M+1}, \\ &\qquad \boldsymbol{S}_i,\sum_{j=1}^{M+1}D_{ij}=1\Bigg) \\ &\quad = \big\{pr(D_{i1}=1\mid \boldsymbol{Z}_{i1},\boldsymbol{S}_i)/pr(D_{i1}=0\mid \boldsymbol{Z}_{i1},\boldsymbol{S}_i)\big\}/ \\ &\qquad \sum_{k=1}^{M+1} pr(D_{ik}=1\mid \boldsymbol{Z}_{ik},\boldsymbol{S}_i)/pr(D_{ik}=0\mid \boldsymbol{Z}_{ik},\boldsymbol{S}_i),\end{aligned}$$

which is nothing but a function of the marginal odds of the disease $pr(D=1\mid \boldsymbol{Z},\boldsymbol{S})/pr(D=0\mid \boldsymbol{Z},\boldsymbol{S})$.

Remark 1. In forming the likelihood, we assume that conditional on the matching variable $\boldsymbol{S}$, the subjects are independent within a stratum. Also, we assume that strata are independent.

Remark 2. As noted in Breslow (1996), the third term of the above likelihood can be written as a joint probability of $\{\boldsymbol{Z}_{ij}\}_{j=1}^{M+1}$ given $\boldsymbol{S}_i, D_{i1}=1, D_{ij}=0$ for $j=2,\ldots,M+1$, and the conditioning event T^{*}, the set of unordered exposure variable. Thus, the third term of likelihood (4) is

$$\frac{p(\boldsymbol{Z}_{i1}\mid D_{i1}=1,\boldsymbol{S}_i)\displaystyle\prod_{j=2}^{M+1}p(\boldsymbol{Z}_{ij}\mid D_{ij}=1,\boldsymbol{S}_i)}{\displaystyle\sum_{m=1}^{M+1}p(\boldsymbol{Z}_{im}\mid D_{i1}=1,\boldsymbol{S}_i)\prod_{h\neq m}p(\boldsymbol{Z}_{ih}\mid D_{ih}=1,\boldsymbol{S}_i)}.$$

This result asserts that L can also be interpreted as a joint likelihood of R, X, and $\boldsymbol{Z}$ given $\boldsymbol{S}, D$, and the conditioning event T^{*}. This representation of the likelihood conforms with the retrospective nature of the design by which matched case–control data are collected.

Remark 3. There is a clear difference between the standard approach used in prospective or cross-sectional study design and the proposed method outlined here in terms of modeling of the exposure distribution. In a case–control study, more specifically in a retrospective study with outcome-dependent sampling design, the marginal distribution of the exposure variable and the intercept parameter of equation (1) are not identifiable, unless we know the disease prevalence $P(D=1\mid \boldsymbol{S})$ in the stratum defined by the variable $\boldsymbol{S}$. Therefore, instead of modeling the marginal distribution of the exposure variable $p(X\mid \boldsymbol{S},\boldsymbol{Z})$, we model the exposure distribution among the control population $p(X\mid D=0,\boldsymbol{S},\boldsymbol{Z})$.

In order to estimate the parameters we develop an EM algorithm. The EM algorithm consists of two steps at each iteration: (i) the E-step and (ii) the M-step. In the E-step, we take expectation of the complete data likelihood with respect to the conditional distribution of the unobserved X, i.e., with respect to $p(X\mid \boldsymbol{Z}_{ij},D_{ij},\boldsymbol{S}_i,R_{ij}=0)$ and

$$\begin{aligned} & p(X\mid \boldsymbol{Z}_{ij},D_{ij},\boldsymbol{S}_i,R_{ij}=0) \\ &\quad = \frac{pr(R_{ij}=0\mid X,\boldsymbol{Z}_{ij},D_{ij},\boldsymbol{S}_i)p(X\mid \boldsymbol{Z}_{ij},D_{ij},\boldsymbol{S}_i)}{pr(R_{ij}=0\mid Z_{ij},D_{ij},\boldsymbol{S}_i)}.\end{aligned}$$

Define $\Theta=(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta})$, then log of the complete data likelihood is

$$\begin{aligned} \log\{L_{\text{comp}}(\Theta)\} = \sum_{i=1}^{n}\log pr\Bigg(& D_{i1}=1, D_{i2}=\cdots=D_{iM+1} \\ & =0\mid \{\boldsymbol{Z}_{ij}\}_{j=1}^{M+1},\boldsymbol{S}_i,\sum_{j=1}^{M+1}D_{ij}=1\Bigg) \\ + \sum_{(i,j):R_{ij}=1} & \big\{\log p(X_{ij}\mid \boldsymbol{Z}_{ij},D_{ij},\boldsymbol{S}_i) \\ & +\log pr(R_{ij}=1\mid X_{ij},\boldsymbol{Z}_{ij},D_{ij},\boldsymbol{S}_i)\big\} \\ + \sum_{(i,j):R_{ij}=0} & \big\{\log p(X_{ij}^{(M)}\mid \boldsymbol{Z}_{ij},D_{ij},\boldsymbol{S}_i) \\ & +\log pr(R_{ij}=0\mid X_{ij}^{(M)},\boldsymbol{Z}_{ij},D_{ij},\boldsymbol{S}_i)\big\}, \\ & \qquad (5)\end{aligned}$$

where $X_{ij}^{(M)}$ denotes the unobserved value of X. Therefore, at the $(t+1)$st iteration of the EM algorithm, the E-step is

$$\begin{aligned} E_{\Theta^{(t)}}\{\log L_{\text{comp}}(\Theta^{(t+1)})\} &= L_1(\Theta^{(t+1)})+L_2(\Theta^{(t+1)}) \\ &\quad + L_3(\Theta^{(t+1)}\mid \Theta^{(t)}), \qquad (6)\end{aligned}$$

where $L_1(\Theta)$ and $L_2(\Theta)$ are the first and the second term on the right-hand side of equation (5), and

$$L_3(\Theta^{(t+1)} \mid \Theta^{(t)}) = \sum_{(i,j):R_{ij}=0} E_{\Theta^{(t)}} \Big\{ \log p(X_{ij}^{(M)} \mid \boldsymbol{Z}_{ij}, D_{ij}, \boldsymbol{S}_i) + \log pr(R_{ij} = 0 \mid X_{ij}^{(M)}, \boldsymbol{Z}_{ij}, D_{ij}, \boldsymbol{S}_i) \Big\}.$$

Here, $E_{\Theta^{(t)}}$ denotes the expectation with respect to the conditional distribution of the unobserved X conditioning on the observed data, and evaluated at the previous iterative value of $\Theta = \Theta^{(t)}$. In the M-step of $(t+1)$st iteration, we maximize (6) with respect to $\Theta^{(t+1)}$ treating $\Theta^{(t)}$ as a constant. We iterate between the two steps until convergence is achieved.

Estimating the Observed Information

Following Louis (1982), we calculate the information matrix as

$$I_\Theta(\hat{\Theta}) = -E\left[\frac{\partial^2}{\partial\Theta\partial\Theta^\tau} \log\{L_{\text{comp}}(\Theta)\}\right]_{\Theta=\hat{\Theta}} - E\left[\frac{\partial}{\partial\Theta} \log\{L_{\text{comp}}(\Theta)\} \frac{\partial}{\partial\Theta^\tau} \log\{L_{\text{comp}}(\Theta)\}\right]_{\Theta=\hat{\Theta}}. \tag{7}$$

The above expectation is with respect to the conditional distribution $p(X \mid \boldsymbol{Z}, D, \boldsymbol{S}, R = 0)$. For continuous X, the expectations can be evaluated by the Monte Carlo method. Estimates of the variance–covariance matrix of the parameter can be obtained by inverting the observed information matrix.

3. Special Cases

3.1 *Exponential Family of Distributions*

Suppose the distribution of the partially missing exposure variable among the control population is a member of the generalized exponential family of distributions. Then,

$$p(X_{ij} \mid \boldsymbol{Z}_{ij}, D_{ij} = 0, \boldsymbol{S}_i) = \exp\big[\xi\{X_{ij}\eta_{ij} - b(\eta_{ij})\} + c(X_{ij}, \xi)\big], \tag{8}$$

where $\eta_{ij} = \gamma_0 + \boldsymbol{\gamma}_Z^\tau \boldsymbol{Z}_{ij} + \boldsymbol{\gamma}_S^\tau \boldsymbol{S}_i$. The mean and variance of the above distribution are $b'(\eta_{ij})$ and $\xi^{-1}b''(\eta_{ij})$, respectively. $c(\cdot, \cdot)$ is the normalizing constant. Using the results of Satten and Kupper (1993), we show in the Web Appendix that

$$\frac{pr(D_{ij} = 1 \mid \boldsymbol{Z}_{ij}, \boldsymbol{S}_i)}{pr(D_{ij} = 0 \mid \boldsymbol{Z}_{ij}, \boldsymbol{S}_i)} = \exp\big[\beta_0(\boldsymbol{S}_i) + \boldsymbol{\beta}_1^\tau \boldsymbol{Z}_{ij} + \xi\{b(\eta_{ij}^*) - b(\eta_{ij})\}\big], \tag{9}$$

where $\eta_{ij}^* = \eta_{ij} + \xi^{-1}\beta_2$. Equation (9) gives us the odds of the disease conditional only on $\boldsymbol{Z}$ and $\boldsymbol{S}$, which is used in the likelihood equation (4). We also show in the Web Appendix that the distribution of the exposure variable among the cases is

$$p(X_{ij} \mid D_{ij} = 1, \boldsymbol{Z}_{ij}, \boldsymbol{S}_i) = \exp\big[\xi\{X_{ij}\eta_{ij}^* - b(\eta_{ij}^*)\} + c(\xi, X_{ij})\big]. \tag{10}$$

Note that the case distribution is also a member of the exponential family of distributions, and its natural parameter η_{ij}^* is a function of η_{ij}, β_2, and ξ.

Binary Exposure Variable

Between the two special cases, first we assume that the distribution of the exposure variable among the control population follows a Bernoulli distribution, i.e.,

$$p(X_{ij} \mid \boldsymbol{Z}_{ij}, D_{ij} = 0, \boldsymbol{S}_i) = \exp\big[\eta_{ij}X_{ij} - \log\{1 + \exp(\eta_{ij})\}\big]. \tag{11}$$

Note the mean and variance are $H(\eta_{ij})$ and $H(\eta_{ij})\{1 - H(\eta_{ij})\}$, respectively. By applying equation (10), we obtain the exposure distribution among the case population $p(X_{ij} \mid \boldsymbol{Z}_{ij}, D_{ij} = 1, \boldsymbol{S}_i) = \exp[X_{ij}\eta_{ij}^* - \log\{1 + \exp(\eta_{ij}^*)\}]$, where $\eta_{ij}^* = \eta_{ij} + \beta_2$. Furthermore, the conditional probability $p(X = 1 \mid \boldsymbol{Z}_{ij}, D_{ij}, \boldsymbol{S}_i, R_{ij} = 0)$ becomes $H\{\psi_{ij}(\boldsymbol{\theta})\}$, where $\psi_{ij}(\boldsymbol{\theta}) = \eta_{ij} + D_{ij}\beta_2 + \log\{\bar{\pi}_{ij}(1)/\bar{\pi}_{ij}(0)\}$. Here, $\pi_{ij}(u) = H(\delta_0 + \boldsymbol{\delta}_Z \boldsymbol{Z}_{ij} + \delta_X u + \delta_D D_{ij} + \boldsymbol{\delta}_S \boldsymbol{S}_i)\}$ and $\bar{\pi}_{ij}(u) = 1 - \pi_{ij}(u)$. Using all this information, all three functions L_1, L_2, and L_3 of the E-step simplify as

$$L_1(\Theta^{(t+1)}) = \sum_{i=1}^{n}\left[\boldsymbol{\beta}_1^\tau \boldsymbol{Z}_{i1} + \log\frac{\big(1 + e^{\eta_{i1}^*}\big)}{\big(1 + e^{\eta_{i1}}\big)} - \log\left[\sum_{j=1}^{M+1} \exp\left\{\boldsymbol{\beta}_1^\tau \boldsymbol{Z}_{ij} + \log\frac{\big(1 + e^{\eta_{ij}^*}\big)}{\big(1 + e^{\eta_{ij}}\big)}\right\}\right]\right],$$

$$L_2(\Theta^{(t+1)}) = \sum_{(i,j):R_{ij}=1}\big[X_{ij}(\eta_{ij} + D_{ij}\beta_2) - \log(1 + e^{\eta_{ij}+D_{ij}\beta_2}) + \log\pi_{ij}(X_{ij})\big],$$

$$L_3(\Theta^{(t+1)} \mid \Theta^{(t)}) = \sum_{(i,j):R_{ij}=0}\big[H\{\psi_{ij}(\Theta^{(t)})\}(\eta_{ij} + D_{ij}\beta_2) - \log(1 + e^{\eta_{ij}+D_{ij}\beta_2}) + H\{\psi_{ij}(\Theta^{(t)})\}\log\bar{\pi}_{ij}(1) + [1 - H\{\psi_{ij}(\Theta^{(t)})\}]\log\bar{\pi}_{ij}(0)\big].$$

In the M-step, we maximize $L_1 + L_2 + L_3$ with respect to the parameter Θ by the Newton–Raphson method.

Remark 4. The difference between the present approach and that of Paik (2004) should be noted. In that paper, the author modeled a binary exposure variable as

$$p(X \mid \boldsymbol{Z}, D, \boldsymbol{S}) = \exp\big[X(\gamma_0 + \boldsymbol{\gamma}_Z \boldsymbol{Z} + \boldsymbol{\gamma}_S \boldsymbol{S} + \gamma_D D) - \log(1 + e^{\gamma_0 + \boldsymbol{\gamma}_Z \boldsymbol{Z} + \boldsymbol{\gamma}_S \boldsymbol{S} + \gamma_D D})\big].$$

Through our calculations, it turns out that $\gamma_D = \beta_2$. Though this relation has been recognized in equation (5) of that paper and subsequently used in forming the conditional likelihood, the relation has not been taken into account to calculate the inputed values of the missing Xs, which is denoted by X^* in that paper. Therefore, as indicated in the paragraph right before equation (7) of that article that β_2 (in the name of γ_D) is estimated once through the likelihood $p(R, X \mid D, \boldsymbol{Z}, \boldsymbol{S})$, and then again it is estimated using the conditional likelihood right before equation (6) of that paper.

Normally Distributed Exposure Variable

Here, we assume that the exposure distribution among the controls follows the normal distribution with mean $\mu_{ij} = \gamma_0 + \boldsymbol{\gamma}_Z^\tau \boldsymbol{Z}_{ij} + \boldsymbol{\gamma}_S^\tau \boldsymbol{S}$ and variance σ^2, then

$$p(X_{ij} \mid \boldsymbol{Z}_{ij}, D_{ij} = 0, \boldsymbol{S}_i) = \exp\left\{(\mu_{ij}X_{ij} - \tfrac{\mu_{ij}^2}{2})/\sigma^2 - X_{ij}^2/2\sigma^2 - \log(2\pi\sigma^2)/2\right\}. \quad (12)$$

Applying equation (10), we obtain the exposure distribution among the case population as the normal distribution with mean $\mu_{ij}^* = \eta_{ij} + \sigma^2\beta_2$ and variance σ^2. Using all this information, L_1 and L_2 of the E-step simplify as

$$L_1(\Theta^{(t+1)}) = \sum_{i=1}^{n}\left[\boldsymbol{\beta}_1^\tau \boldsymbol{Z}_{i1} + \beta_2\mu_{i1} - \log[\sum_{j=1}^{M+1}\exp\{\boldsymbol{\beta}_1^\tau \boldsymbol{Z}_{ij} + \beta_2\mu_{ij}\}]\right],$$

$$L_2(\Theta^{(t+1)}) = \sum_{(i,j):R_{ij}=1}\left[-\frac{1}{2\sigma^2}(X_{ij} - \mu_{ij} + D_{ij}\beta_2\sigma^2)^2 + \log\pi_{ij}(X_{ij})\right].$$

In the M-step, we maximize $L_1 + L_2 + L_3$ with respect to the parameter Θ by the Newton–Raphson method. Note that neither the conditional distribution of the unobserved X given the observed data nor L_3 has closed analytic form, and

$$p(X_{ij} \mid \boldsymbol{Z}_{ij}, D_{ij}, \boldsymbol{S}_{ij}, R_{ij} = 0) \propto \{1 - H(\delta_0 + \boldsymbol{\delta}_Z^\tau \boldsymbol{Z}_{ij} + \delta_X X_{ij} + \delta_D D_{ij} + \boldsymbol{\delta}_S \boldsymbol{S}_i)\} \exp\{-(X_{ij} - \mu_{ij})^2/2\sigma^2\}. \quad (13)$$

Therefore, the expectation with respect to the conditional distribution (13) needs to be calculated by some numerical method. Let $X_{ij}^{(b)}$, $b = 1, \ldots, B$ be B random numbers drawn from (13), then the Monte Carlo approximation of L_3 is

$$L_3(\Theta^{(t+1)} \mid \Theta^{(t)}) = B^{-1}\sum_{(i,j):R_{ij}=0}\sum_{b=1}^{B}\Big\{\log p(X_{ij}^{(b)} \mid \boldsymbol{Z}_{ij}, D_{ij}, \boldsymbol{S}_i) + \log pr(R_{ij} = 0 \mid X_{ij}^{(b)}, \boldsymbol{Z}_{ij}, D_{ij} = 0, \boldsymbol{S}_i)\Big\},$$

and the B random numbers can be drawn by using the Metropolis–Hastings algorithm.

3.2 *Nonexponential Distribution—Lognormal Distribution*

So far, we have discussed exponential family of distributions for the partially missing exposure variable; however the proposed method is applicable to any distribution. Therefore, now we discuss the case where the distribution of the exposure variable among the controls is a member of nonexponential distribution, such as lognormal distribution. Let $p(X_{ij} \mid \boldsymbol{Z}_{ij}, D_{ij} = 0, \boldsymbol{S}_i) = \exp[-\{\log(X_{ij}) - \mu_{ij}\}^2/2\sigma^2]/\{X_{ij}\sqrt{2\pi\sigma^2}\}$, for $X_{ij} > 0$. Here $\mu_{ij} = \gamma_0 + \boldsymbol{\gamma}_Z^\tau \boldsymbol{Z}_{ij} + \boldsymbol{\gamma}_S^\tau \boldsymbol{S}_i$. Using result (A.2) of the Web Appendix, we obtain the exposure distribution among the cases, and it is $p(X_{ij}\ \boldsymbol{Z}_{ij}, D_{ij} = 1, \boldsymbol{S}_i) \propto \exp[\beta_2 X_{ij} - \frac{\{\log(X_{ij})-\mu_{ij}\}^2}{2\sigma^2}]/\{X_{ij}\sqrt{2\pi\sigma^2}\}$, with the normalizing constant $Q_{ij}(\beta_2,\ \sigma^2) = E\{\exp\ (\beta_2 e^t)\}$, where $t \sim$ Normal $(\mu_{ij},\ \sigma^2)$. Note that the case and control distributions are not of the same form. Plugging in $pr(D_{ij} = 1\ \boldsymbol{Z}_{ij}, S_i)/pr(D_{ij} = 0\ \boldsymbol{Z}_{ij}, \boldsymbol{S}_i) = \exp\{\beta_0(S_i) + \boldsymbol{\beta}_1^\tau \boldsymbol{Z}_{ij}\}Q_{ij}(\beta_2, \sigma^2)$ in L_1, we obtain

$$L_1(\Theta) = \frac{\exp\{\boldsymbol{\beta}_1^\tau \boldsymbol{Z}_{i1}\}Q_{i1}(\beta_2, \sigma^2)}{\displaystyle\sum_{j=1}^{M+1}\exp\{\boldsymbol{\beta}_1^\tau \boldsymbol{Z}_{ij}\}Q_{ij}(\beta_2, \sigma^2)}.$$

Note that none of L_2 or L_3 has a closed form expression. Also, the conditional distribution of the unobserved X given the observed data does not match with any standard distribution. Hence, like the normal distribution scenario of the previous section, one may adopt Metropolis–Hastings algorithm to generate random numbers from the conditional distribution, and then use the Monte Carlo method to approximate the integrals in the E-step. Maximization can be done by the Newton–Raphson method.

4. Example

The Los Angeles endometrial cancer data comprise 63 strata and each stratum consists of one case and four controls (Breslow and Day, 1980). The goal of the study was to assess important risk factors for endometrial cancer among the postmenopausal women of an affluent retirement community in Los Angeles. Controls were chosen from a roster of all women in the same community, and then matched with a case based on their age. Among several measured risk factors, the binary exposure variable obesity was missing for about 16% of the study participants. We treat obesity as the partially missing exposure variable (X) and the presence of gall bladder disease is considered as a binary completely observed covariate (Z). Obesity was determined accordingly as the BMI value exceeds the normal value of 30 or not. In the analysis, age is transformed into [0, 1] scale and then used as a matching variable S. In the data set, obesity was missing in six out of 63 cases and in 45 out of 252 controls. Though this finding is not statistically significant (p-value=0.13), we include D in the selection model. The disease risk model of our interest is $H(\beta_0(S) + \beta_1 Z + \beta_2 X)$, where β_1 and β_2 are the disease-exposure association parameters for Z and X, respectively.

The exposure distribution among the controls is modeled as $p(X = 1 \mid Z, S, D = 0) = H(\gamma_0 + \gamma_Z Z + \gamma_S\ S)$, and so the exposure distribution among the cases becomes $p(X = 1 \mid Z, S, D = 1) = H(\gamma_0 + \gamma_Z Z + \gamma_S S + \beta_2)$. We analyze the data by using the CLR technique by the method proposed in Paik (2004), we call it *Paik*, and by using the newly proposed method, we call it *SM*. The results are presented in Table 1. The standard error (SD) of the Paik method was calculated by the jackknife method, whereas the SD of the SM method was calculated by equation (7). The results indicate that the presence of gall bladder disease seems to increase the risk of having endometrial cancer. From the analyses we do not find any significant association between the cancer and obesity. As expected, the CLR method produces largest standard error for $\hat{\beta}_1$ and $\hat{\beta}_2$. The SM produces least standard error for the parameters among the competitive methods. The estimates due to the Paik and SM are different from

Table 1

Results of the analysis of the Los Angeles Endometrial Cancer Data. Est. and SD represent the estimate and the standard error, respectively.

	Method					
	CLR		Paik		SM	
Parameter	Est.	SD	Est.	SD	Est.	SD
β_1	1.2844	0.3921	1.3098	0.3760	1.2857	0.3741
β_2	0.4405	0.3800	0.6879	0.3633	0.5346	0.3091
γ_0			0.0288	0.2875	−0.0139	0.3611
γ_Z			0.1582	0.3623	0.1692	0.3625
γ_S			0.5390	0.4868	0.5337	0.4850
γ_D			0.5224	0.3104		
δ_0			1.1606	0.4031	1.0827	0.4013
δ_Z			0.4933	0.5616	0.4875	0.5639
δ_X			0.6560	0.3112	0.8717	0.3146
δ_D			0.5750	0.4694	0.5493	0.4711
δ_S			−0.0403	0.6456	−0.0635	0.6526

the CLR approach which does not take into account all the observed data. According to both the methods, the Paik and SM, we find that missingness mechanism is significantly associated with the missing observations with p-values 0.0175 and 0.00279, respectively.

Sensitivity Analysis

Usually, little information is available to estimate the parameters of the missingness mechanism. Therefore, several authors have suggested to perform sensitivity analysis by varying certain parameters of the missingness mechanism (see Rotnitzky, Robins, and Scharfstein, 1998; Roy and Lin, 2005). We study the sensitivity of the estimate of β_1 and β_2 for a set of values of δ_X. Note that when $\delta_X = 0$, the missingness mechanism becomes MAR. The sensitivity of the Paik and SM method is studied when $\delta_X \in [-4, 4]$. We set δ_X to a desired value, and then determine other parameters of the model by the Paik and SM method. For a fair comparison, for both the methods we calculate the standard error of the estimate by the jackknife method. Figure 1 shows how $\hat{\beta}_2$ changes with δ_X for both the methods along with the 95% confidence interval. Estimate of β_2 obtained by using the SM method ranges from 0.41 to 0.53 whereas $\widehat{\beta_2}$ of the Paik varies from 0.44 to 0.82. For a moderate range of values of δ_X, the regression parameter β_2 remains insignificant for both the methods. For the SM method, the other disease–risk parameter β_1 varies from 1.2876 to 1.3013 for $\delta_X \in [-4, 4]$, and for the Paik method it varies from 1.3070 to 1.3378.

5. Simulation

We evaluate the performance of the proposed method using simulation study. In order to elicit realistic parameter value,

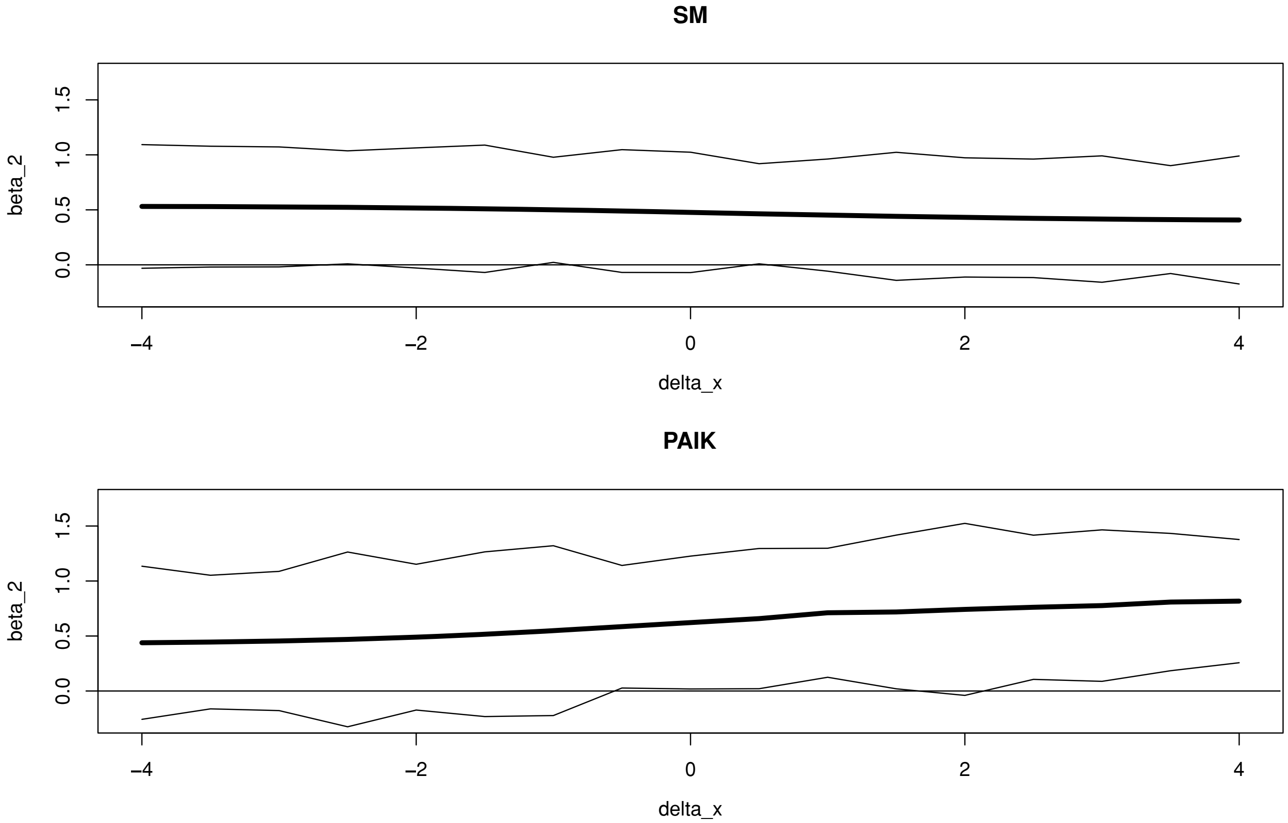

Figure 1. The log odds ratio parameter for obesity and its 95% confidence interval for varying δ_X for the SM and Paik method.

we are going to use the real data as a prototype. We found that after rescaling the matching variable age of the original data set into [0, 1] interval, it approximately follows Normal$(0.53, 0.24^2)$. Therefore, first generate the matching variable S from the Normal$(0.53, 0.24^2)$ for a cohort of $N =$ 20,000 subjects. The presence of gall bladder disease was found to be 19% in the data set. However, in the simulation study, instead of generating Z from a Bernoulli distribution, it is generated from the Gamma$(\mathrm{a} = 0.23, \mathrm{b} = 0.8333)$, where the parameters a and b are obtained by equating the first two moments of the gamma distribution with the observed moments of Z of the original data set. Note that mean of Gamma(a, b) distribution is ab. Then, we generate the binary covariate X from the Bernoulli distribution with the success probability $H(0.2239 + 0.5470S + 0.2332Z)$. The true parameter values are set to the estimated value of the parameters obtained by regressing X on Z and S in the original data set. Now, we generate the disease variable D from the Bernoulli distribution with success probability

$$H(-3.1 + 1.1S + 1.3Z + 0.55X). \tag{14}$$

Here, we set β_1, the log–odds-ratio parameter for Z, and β_2, the log-odds-ratio parameter for X, as 1.3 and 0.55 which are close to the estimates obtained by using the SM method in the data analysis. The coefficient of S in (14) reflects twofold increase in the disease risk for changing S from its 10th quantile to 90th quantile. The intercept parameter is so chosen that the marginal disease probability is 0.10. Next, we randomly choose 100 cases from the cohort and then corresponding to each case we choose one control so that the absolute difference between the values of the matching variable for the case and the control subject is not more than 0.009.

We simulate 300 data sets and then for each data set we create missing data by generating a binary random variable R according to the following missingness mechanism (MM).

MM1. Missing Completely at Random (MCAR): logit $P(R_{ij} = 1) = 0.66$,

MM2. Missing at Random (MAR): logit$P(R_{ij} = 1) = D_{ij}$,

MM3. Informative Missing (IM): logit$P(R_{ij} = 1) = X_{ij}$,

MM4. IM: logit$P(R_{ij} = 1) = 0.5X_{ij} + 0.5D_{ij}$,

MM5. IM: logit$P(R_{ij} = 1) = X_{ij}Z_{ij} + D_{ij}X_{ij}$.

Finally, the data sets are analyzed by the three methods—the CLR, the Paik method, and the SM method. The results are given in Table 2. We report the estimated value which is the average of the estimates across the simulated data sets, and the mean square error which is the average of the squared differences between the true and the estimated value.

In all the analyses, we assume that missingness mechanism is unknown; therefore we model the logit of the selection probability as a linear function of Z, S, X, and D. Also, no matter what distribution the exposure variable is generated from, in the analyses we model the logit of the success probability of X among the controls as a linear function of S and Z.

Table 2

Results of the simulation study when $P(X = 1 \mid S, Z) = H(0.2239 + 0.5470S + 0.2332Z)$. The true values of β_1 and β_2 are 1.3 and 0.55, respectively. Est. and MSE stand for estimate and mean square error.

	β_1		β_2	
Method	Est.	MSE	Est.	MSE
		Fully observed data		
CLR&Paik	1.4034	0.1714	0.6075	0.1416
SM	1.3857	0.1568	0.5793	0.1091
		MM1: logit$P(R = 1) = 0.66$		
CLR	1.6196	0.8383	0.5645	0.4786
Paik	1.5213	0.3019	0.5294	0.2919
SM	1.3899	0.2067	0.5426	0.1861
		MM2: logit$P(R = 1) = D$		
CLR	1.7006	0.9199	0.7887	0.7764
Paik	1.6784	0.4927	0.7328	0.3149
SM	1.4263	0.1964	0.6110	0.1750
		MM3: logit$P(R = 1) = X$		
CLR	1.5114	0.5158	0.6042	0.5201
Paik	1.4680	0.1714	0.5911	0.3239
SM	1.3344	0.1131	0.5846	0.2181
		MM4: logit$P(R = 1) = 0.5D + 0.5X$		
CLR	1.5711	0.6595	0.5679	0.3925
Paik	1.5840	0.3344	0.5059	0.3066
SM	1.3798	0.1631	0.4708	0.2063
		MM5: logit$P(R = 1) = DX + XZ$		
CLR	1.4701	0.5669	0.9890	0.5826
Paik	1.3716	0.2517	1.0096	0.4968
SM	1.3236	0.1568	0.9029	0.2908

Note that when the data are completely observed, the Paik method reduces to the CLR method, whereas the parameter estimates of the SM method is then obtained by maximizing

$$\begin{aligned} L = {} & \prod_{i=1}^{n} \prod_{j=1}^{M+1} p(X_{ij} Z_{ij}, D_{ij}, S_i) \\ & \times \prod_{i=1}^{n} pr \Bigg(D_{i1} = 1, D_{i2} = \cdots = D_{iM+1} = 0 \{\boldsymbol{Z}_{ij}\}_{j=1}^{M+1}, \\ & \qquad \boldsymbol{S}_i, \sum_{j=1}^{M+1} D_{ij} = 1 \Bigg). \end{aligned} \tag{15}$$

For fully observed data, the difference between these two approaches is that the SM method models the exposure distribution explicitly whereas the CLR (or Paik) does not model it. Therefore, for fully observed data, if the SM method models the distribution of X correctly, it should perform better than the CLR method by extracting information on $\boldsymbol{\beta}$ contained in $p(X \mid D, S, Z)$. In fact, Table 2 actually shows little bit efficiency gain in the SM compared to the CLR approach for fully observed data.

As expected, in the presence of the missing observations the CLR method produces biased and inefficient estimate for both the parameters β_1 and β_2, among the three methods. The Paik method outperforms the CLR in terms of bias and efficiency. However, among the three methods, the SM shows better performance in terms of bias and efficiency of the parameters. Special attention should be given to the missingness mechanism MM5, as it considers one kind of violation of the model assumption of the selection probability. It turns out that the estimate of β_2 is most affected due to this model violation; however, among the three methods the SM method is least affected.

Robustness of the different methods is studied under two types of model violations of the exposure distribution. First, we consider the situation where the logit of the success probability of X is not linear in S. Hence, in the simulation, X is generated from the Bernoulli distribution with the success probability $H(0.2239 + 0.5470S^2 + 0.2332Z)$, and everything else remains the same as before. While we analyze the data, we model the logit of the success probability of X as a linear function of S and Z. The results are presented in Table 3. There are two situations—when moderate percentage of data is missing and when the data have no missing observation. For missing data situation, the SM outperforms the other methods. It turns out that the estimates are severely affected by the violation of the model assumption of the selection probability. As mentioned by the associate editor, for fully observed data scenario and misspecified model, we expect to see more biased estimate due to full likelihood–based method (i.e., SM) compared to the CLR approach. However, the simulation result does not reflect that fact. After a thorough numerical investigation, we found that in the domain of S, $0.5470S^2$ is almost a linear function of S, and it has slight curvature at the boundary of the domain of S. Therefore, for sample size n = 100, the SM method still performs better than the CLR approach in terms of bias and MSE. However, if the sample size (n) is large ($n \geq 500$), the data are more likely to include some of the boundary values of S which show nonlinearity in the function $0.5470S^2$, and then one may see more biased estimate in the SM compared to the CLR (the results are not presented here). In general, for fully observed data the likelihood–based approach is expected to produce more biased estimator than the CLR method when the underlying assumptions are violated. However, that difference in bias depends on the degree of model violation, numerical effect of the violation on the variables under consideration, and the sample size of the data.

Table 3

Results of the simulation study when $P(X = 1 \mid S, Z) = H(0.2239 + 0.5470S^2 + 0.2332Z)$. *The true values of* β_1 *and* β_2 *are* 1.3 *and* 0.55, *respectively. Est. and MSE stand for estimate and mean square error.*

	β_1		β_2	
Method	Est.	MSE	Est.	MSE
Fully observed data				
CLR&Paik	1.3318	0.1013	0.5086	0.1447
SM	1.3228	0.0978	0.5111	0.1296
MM1: logit$P(R = 1) = 0.66$				
CLR	1.5420	0.8412	0.4924	0.2724
Paik	1.4751	0.2047	0.5041	0.2033
SM	1.3616	0.1578	0.5222	0.1545
MM2: logit$P(R = 1) = D$				
CLR	1.6589	0.9999	0.5476	0.5312
Paik	1.5529	0.2565	0.5858	0.3319
SM	1.3408	0.1219	0.5323	0.1913
MM3: logit$P(R = 1) = X$				
CLR	1.5617	0.4823	0.4957	0.4865
Paik	1.5473	0.3012	0.5249	0.3616
SM	1.4145	0.2115	0.5001	0.2621
MM4: logit$P(R = 1) = 0.5D + 0.5X$				
CLR	1.5671	0.7896	0.4533	0.3468
Paik	1.5048	0.1814	0.4601	0.2865
SM	1.3555	0.1075	0.4729	0.1908
MM5: logit$P(R = 1) = DX + XZ$				
CLR	1.4723	0.7850	0.9677	0.8653
Paik	1.4098	0.2460	0.9953	0.5730
SM	1.3620	0.1665	0.8867	0.3678

Lastly, we consider the situation where the success probability of the binary exposure variable is not in linear-logistic form, but a discontinuous function of Z, which is given by

$$X \sim \begin{cases} \text{Bernoulli}(p), \text{where } p = \\ \quad \Phi(0.2239 + 0.2332Z), \quad \text{for } S > 0.7 \\ \text{Bernoulli}(p), \text{where } p = \\ \quad \Phi(-0.7Z), \quad \text{for } S \leq 0.7. \end{cases} \tag{16}$$

As before, we analyze the data sets using all the three methods, and in the analyses we model the success probability of X among the controls using a logistic regression in S and Z. The results are presented in Table 4. The SM produces least value of MSE among all the three methods. Overall, though the Paik offers a significant improvement over the CLR in terms of bias and efficiency, using SM method one may have more gain. One should note that for fully observed data situation, due to misspecified model the SM shows more biased estimate of β_2 than the CLR approach. However, MSE_{CLR}, the MSE due to the CLR, is still higher than that of the SM, MSE_{SM}. The intuitive reason for this behavior is in general $\sigma^2_{\text{SM}} \ll \sigma^2_{\text{CLR}}$, where σ^2_{SM} and σ^2_{CLR} are the variance of the estimator under the SM and CLR approach, respectively. Therefore, $MSE_{\text{SM}} = \text{Bias}^2_{\text{SM}} + \sigma^2_{\text{SM}} < \text{Bias}^2_{\text{CLR}} + \sigma^2_{\text{CLR}} = MSE_{\text{CLR}}$ even though Bias $_{\text{SM}} >$ Bias $_{\text{CLR}}$. When sample size (n) is large, $\sigma^2_{\text{SM}}, \sigma^2_{\text{CLR}} \longrightarrow 0$, then the difference in bias is truly reflected through MSE.

Summarizing the results, we conclude that both the Paik and the SM are sensitive towards the selection model. However, the SM performs better than the Paik in terms of MSE. The intuitive explanation of this behavior may be that in the SM method all parameters are estimated simultaneously through a joint conditional likelihood. Both the methods are fairly robust under the misspecification of the distributional form of the binary exposure variable, and under the misspecified form of logit$p(X = 1|Z, S)$. Overall, in the presence of moderate percentage of missing covariate data, both bias and variance are lower in the SM than the CLR and the Paik method. The EM algorithm and the CLR method did not converge for approximately 2% of the data sets. However, that convergence problem can be avoided by choosing very large sample size. The results are presented based only on 300 data sets for which all the methods converged. All the computations were done by using `R` statistical software and the necessary subroutines were written in `Fortran` 77. For the maximization of the likelihoods, we used nlm()

Table 4

Results of the simulation study when $P(X = 1 \mid S, Z) = \Phi(0.2239 + 0.2332Z)I(S > 0.70) + \Phi(-0.7Z)I(S \leq 0.70)$. *The true values of* β_1 *and* β_2 *are* 1.3 *and* 0.55, *respectively. Est. and MSE stand for estimate and mean square error.*

	β_1		β_2	
Method	Est.	MSE	Est.	MSE
	Fully observed data			
CLR&Paik	1.3368	0.1689	0.5411	0.1424
SM	1.2950	0.1394	0.4633	0.0956
	MM1: $\text{logit}P(R = 1) = 0.66$			
CLR	1.6332	0.7014	0.5965	0.4100
Paik	1.5278	0.2460	0.5277	0.2028
SM	1.3854	0.1624	0.5072	0.1825
	MM2: $\text{logit}P(R = 1) = D$			
CLR	1.5400	0.6694	0.5877	0.4387
Paik	1.5960	0.3193	0.6508	0.3022
SM	1.3816	0.1480	0.5554	0.1543
	MM3: $\text{logit}P(R = 1) = X$			
CLR	1.6461	0.8460	0.6267	0.4033
Paik	1.5589	0.2297	0.5632	0.2242
SM	1.3923	0.1503	0.4978	0.1692
	MM4: $\text{logit}P(R = 1) = 0.5D + 0.5X$			
CLR	1.6021	0.7482	0.4175	0.3903
Paik	1.6091	0.3267	0.5335	0.3463
SM	1.3895	0.1736	0.4516	0.1978
	MM5: $\text{logit}P(R = 1) = DX + XZ$			
CLR	1.6916	1.1595	0.9521	0.7818
Paik	1.7161	0.5496	1.0567	0.7405
SM	1.4971	0.2744	0.8638	0.3822

function, and for the CLR analysis we used clogit function of library(survival).

6. Discussion

This article proposes a full likelihood-based method for handling nonignorable missing exposure variable in matched case–control study. The proposed method can handle any kind of distribution for the partially missing exposure variable, and it can also handle varying number of controls in the matched sets. The simulation study shows that the proposed method outperforms the existing methods in terms of bias and efficiency under different scenarios of missing data mechanism. Though this is a model-based approach, the proposed method is not severely affected by moderate type of model violations. Depending on circumstances, such as if all the model assumptions are true, we may even see that for fully observed data the full likelihood-based method performs better than the CLR approach. However, unless there is moderate amount of missing data one should prefer to use the CLR approach which does not require to model its covariate distribution.

7. Supplementary Materials

Web Appendices referenced in Sections 2 and 3, and the data set along with the computer code, are available under the Paper Information link at the *Biometrics* website `http://www.tibs.org/biometrics`.

Acknowledgements

The authors are thankful to Distinguished Professor Raymond J. Carroll, Texas A&M University, for very useful discussions regarding this article. The second author's research is partially supported by a NSF grant. The authors would like to thank the associate editor and the referee for very insightful comments that led to a much improved manuscript.

References

Breslow, N. E. (1996). Statistics in epidemiology: The case–control study. *Journal of the American Statistical Association* **91,** 14–28.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research*, Volume 1. Lyon, France: International Agency for Research on Cancer.

Ibrahim, J. G., Lipsitz, S. R., and Chen, M H. (1999). Missing covariates in generalized linear mixed models when the missing data mechanism is nonignorable. *Journal of Royal Statistical Society, Series B* **61,** 173–190.

Lipsitz, S. R., Parzen, M., and Ewell, M. (1998). Inference using conditional logistic regression with missing covariates. *Biometrics* **54,** 148–160.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis With Missing Data.* New York: Wiley.

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.

Paik, M. C. (2004). Nonignorable missingness in matched case–control data analyses. *Biometrics* **60,** 306–314.

Paik, M. C. and Sacco, R. L. (2000). Matched case–control data analyses with missing covariates. *Journal of the Royal Statistical Society, Series C* **49,** 145–156.

Rathouz, P. J. (2003). Likelihood methods for missing covariate data in highly stratified studies. *Journal of the Royal Statistical Society, Series B* **65,** 711–723.

Rathouz, P. J., Satten, G. A., and Carroll, R. J. (2002). Semiparametric inference in matched case–control studies with missing covariate data. *Biometrika* **89,** 905–916.

Rotnitzky, A., Robins, J., and Scharfstein, D. (1998). Semiparametric regression for repeated outcomes with nonignorable non-response. *Journal of the American Statistical Association* **93,** 1321–1339.

Roy, J. and Lin, X. (2005). Missing covariates in longitudinal data with informative dropouts: Bias analysis and inference. *Biometrics* **61,** 837–846.

Satten, G. and Carroll, R. J. (2000). Conditional and unconditional categorical regression models with missing covariates. *Biometrics* **56,** 384–388.

Satten, G. A. and Kupper, L. (1993). Inferences about exposure—disease associations using probability-of-exposure information. *Journal of the American Statistical Association* **88,** 200–208.

Sinha, S., Mukherjee, B., Ghosh, M., Mallick, B. K., and Carroll, R. J. (2005). Semiparametric Bayesian analysis of matched case–control studies with missing exposure. *Journal of the American Statistical Association* **100,** 591–601.

Received March 2006. *Revised February* 2007.
Accepted February 2007.