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

Samiran Sinha¹,∗ and Tapabrata Maiti²,∗∗

¹Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.
²Department 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 ≥ 1) matched case–control studies. Let D be the binary disease indicator variable, S be a set of matching variables, X∗ = (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 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, (Xij, S) are independently sampled from p(X∗, S | D = 1), for i = 1, . . . , n and in the second step controls (Xij∗) are drawn from p(X∗ | D = 0, S = sj), for j = 2, . . . , M + 1. The risk model we are interested in is

\[ pr(D = 1 | S, X^*) = H(\beta_0(S) + \beta^T X^*), \]  

where H(u) = {1 + exp (− u)}−1. Here, \( \beta_0(S) \) is the stratum effect defined by the variable S, and \( \beta \) is the log-odds ratio parameter associated with \( X^* \). The standard method for estimating \( \beta \) is to use conditional logistic regression (CLR), conditioning on the number of cases within each matched set. However, in the presence of missing \( X_s \), 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 $Z, S,$ 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 $\beta$. Paik (2004) indeed recognizes the fact while is not applicable in that situation. Therefore, a generalized exponential family do not satisfy this assumption, the method is a member of an exponential family of distributions. 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.

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 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, $X = (Z, X)$, where $Z$ is a $p \times 1$ vector of completely observed covariates, and $X$ is assumed to be a scaler risk factor. Define $\beta = (\beta_1, \beta_2)$, where $\beta_1$ is a $p \times 1$ vector of the log–odds-ratio parameters associated with $Z$, and $\beta_2$ 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 $\beta$ from the conditional likelihood function

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

$$= \prod_{i=1}^{n} \exp \left( \sum_{j=1}^{M+1} \beta_1^z Z_{ij} + \beta_2^z X_{ij} \right).$$

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

$$L = p(R, X, D \mid Z, S)$$

$$= pr(R \mid X, Z, S, D)pr(X \mid Z, S, D)pr(D \mid Z, S, T).$$
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

\[
pr(R_{ij} = 1 \mid X_{ij}, Z_{ij}, D_{ij}, S_i, \delta) = \frac{H(\delta_i + \delta^2_{ij} Z_{ij} + \delta X_{ij} Z_{ij})}{\delta_i + \delta X_{ij} Z_{ij} + \delta^2_{ij} Z_{ij}}.
\]

Let \( p(X_{ij} \mid Z_{ij}, D_{ij} = 0, S_i, \gamma_i) \) 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 \( p(X_{ij} \mid Z_{ij}, D_{ij} = 1, S_i, \gamma, \delta) \). Next, we rewrite \( L \) as

\[
L = \prod_{i=1}^{n} \prod_{j=1}^{M} \{ pr(R_{ij} = 1 \mid X_{ij}, Z_{ij}, D_{ij}, S_i) \\
\times \frac{pr^{1-R_{ij}}(R_{ij} = 0 \mid X_{ij}, Z_{ij}, D_{ij}, S_i)}{1 - pr(R_{ij} = 1 \mid X_{ij}, Z_{ij}, D_{ij}, S_i)} dF \\
\times (X_{ij} \mid Z_{ij}, D_{ij}, S_i) \times p^{R_{ij}}(X_{ij} \mid Z_{ij}, D_{ij}, S_i) \}
\times \frac{n}{\prod_{i=1}^{n} pr(D_{i1} = 1, D_{i2} = \cdots = D_{iM+1} = 0 \mid \{ Z_{ij} \}_{j=1}^{M+1}, S_i, \sum_{j=1}^{M+1} D_{ij} = 1)}.
\]

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

\[
pr(D_{i1} = 1, D_{i2} = \cdots = D_{iM+1} = 0 \mid \{ Z_{ij} \}_{j=1}^{M+1}, S_i, \sum_{j=1}^{M+1} D_{ij} = 1) \\
= \frac{pr(D_{i1} = 1 \mid Z_{i1}, S_i)/pr(D_{i1} = 0 \mid Z_{i1}, S_i)}/ \\
\sum_{k=1}^{M+1} pr(D_{ik} = 1 \mid Z_{ik}, S_i)/pr(D_{ik} = 0 \mid Z_{ik}, S_i),
\]

which is nothing but a functional odd of the disease \( pr(D = 1 \mid Z, S) / pr(D = 0 \mid Z, S) \).

**Remark 1.** In forming the likelihood, we assume that conditional on the matching variable \( S \), 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 \( \{ Z_{ij} \}_{j=1}^{M+1} \) given \( S_i, D_{i1} = 1, D_{ij} = 0 \) 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

\[
p(Z_{i1} \mid D_{i1} = 1, S_i) \prod_{j=2}^{M+1} p(Z_{ij} \mid D_{ij} = 1, S_i) \\
\sum_{m=1}^{M+1} p(Z_{im} \mid D_{i1} = 1, S_i) \prod_{h=m}^{M+1} p(Z_{ih} \mid D_{ih} = 1, S_i).
\]

This result asserts that \( L \) can also be interpreted as a joint likelihood of \( R, \gamma, X, \) and \( Z \) given \( S, D, \) 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 \( P(D = 1 \mid S) \) in the stratum defined by the variable \( S \). Therefore, instead of modeling the marginal distribution of the exposure variable \( p(X \mid S, Z) \), we model the exposure distribution among the control population \( p(X \mid D = 0, S, Z) \).

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 \( p(X \mid Z_{ij}, D_{ij}, S_i, R_{ij} = 0) \) and

\[
p(X \mid Z_{ij}, D_{ij}, S_i, R_{ij} = 0) = \frac{pr(R_{ij} = 1 \mid X, Z_{ij}, D_{ij}, S_i)p(X \mid Z_{ij}, D_{ij}, S_i)}{pr(R_{ij} = 0 \mid Z_{ij}, D_{ij}, S_i)}.
\]

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

\[
\log \{ L_{\text{comp}}(\Theta) \} = \sum_{i=1}^{n} \log \{ pr(D_{i1} = 1, D_{i2} = \cdots = D_{iM+1} = 0 \mid \{ Z_{ij} \}_{j=1}^{M+1}, S_i, \sum_{j=1}^{M+1} D_{ij} = 1) \\
\times \frac{\{ pr(D_{i1} = 1 \mid Z_{i1}, S_i)/pr(D_{i1} = 0 \mid Z_{i1}, S_i)\}}{\sum_{k=1}^{M+1} pr(D_{ik} = 1 \mid Z_{ik}, S_i)/pr(D_{ik} = 0 \mid Z_{ik}, S_i)} + \}
\]

\[
\sum_{(i,j): R_{ij} = 1} \{ \log p(X_{ij} \mid Z_{ij}, D_{ij}, S_i) + \log pr(R_{ij} = 1 \mid X_{ij}, Z_{ij}, D_{ij}, S_i) \} + \}
\sum_{(i,j): R_{ij} = 0} \{ \log p(X_{ij}^{(M)} \mid Z_{ij}, D_{ij}, S_i) + \log pr(R_{ij} = 0 \mid X_{ij}^{(M)} \mid Z_{ij}, D_{ij}, S_i) \},
\]

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

\[
E_{\text{comp}}(\log L_{\text{comp}}(\Theta^{(t+1)})) = L_1(\Theta^{(t+1)}) + L_2(\Theta^{(t+1)}) + L_3(\Theta^{(t+1)} \mid \Theta^{(t)}),
\]
where \( L_1(\Theta) \) and \( L_2(\Theta) \) are the first and the second term on the right-hand side of equation (5), and

\[
L_3(\Theta^{(t+1)} | \Theta^{(t)}) = \sum_{(i,j): R_{ij} = 0} E_{\theta^{(t)}} \left\{ \log p(X_{ij}^{(M)} | Z_j, D_j, S_i) \right. \]
\[
+ \log \pi(\gamma_i X_j - b(\eta_{ij})) \right\}.
\]

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

\[
I_{\theta}(\hat{\Theta}) = \mathbb{E} \left[ \frac{\partial^2}{\partial \Theta^2} \log \{ L_{\text{comp}}(\Theta) \} \right]_{\Theta = \hat{\Theta}}.
\]

(7)

The above equation is with respect to the conditional distribution \( p(X | Z, D, S, R = 0) \). 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 exponential family of distributions. Then,

\[
p(X_{ij} | Z_j, D_j = 0, S_i) = \exp \left\{ \xi [X_{ij} \eta_i - b(\eta_{ij})] + c(X_{ij}, \xi) \right\},
\]

(8)

where \( \eta_i = \gamma_0 + \gamma_j Z_j. \) The mean and variance of the above distribution are \( b'(\eta_i) \) and \( \xi^{-1} b''(\eta_i) \), respectively. \( c(\cdot, \cdot) \) is the normalizing constant. Using the results of Satten and Kupper (1993), we show in the Web Appendix that

\[
\frac{\pi(D_j = 1 | Z_j, S_i)}{\pi(D_j = 0 | Z_j, S_i)} = \exp \left\{ \beta_j(S_i) + \beta_j Z_j + \xi \{ b'\eta_{ij} - b(\eta_{ij}) \} \right\},
\]

(9)

where \( \eta_{ij} = \eta_{ij} + \xi^{-1} \beta_j. \) Equation (9) gives us the odds of the disease conditional only on \( Z \) and \( S \), 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

\[
p(X_{ij} | D_j = 1, Z_j, S_i) = \exp \left\{ \xi \{ X_{ij} \eta_i - b(\eta_{ij}) \} + c(\xi, X_{ij}) \right\}.
\]

(10)

Note that the case distribution is also a member of the exponential family of distributions, and its natural parameter \( \eta_{ij} \) is a function of \( \eta_i, \beta_j \), and \( \xi \).

**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.,

\[
p(X_{ij} | Z_j, D_j = 0, S_i) = \exp \left[ \eta_i X_j - \log \{ 1 + \exp(\eta_i) \} \right].
\]

(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 \( p(X_{ij} | Z_j, D_j = 1, S_i) = \exp \left[ X_j \eta_{ij} - \log \{ 1 + \exp(\eta_i) \} \right] \), where \( \eta_{ij} = \eta_{ij} + \beta_j. \) Furthermore, the conditional probability \( p(X = 1 | Z_j, D_j, S_i, R_j = 0) \) becomes \( H(\psi_i(\Theta)) \), where \( \psi_i(\Theta) = \eta_j + D_j \beta_j + \log \left\{ \pi(\hat{\Theta}(1)) / \pi(\hat{\Theta}(0)) \right\} \). Here, \( \pi(\hat{\Theta}(u)) = H(\delta_j + \delta \xi Z_j + \delta X_j + \delta D_j, \delta \beta_j, \delta \pi(S_i)) \) and \( \pi(\hat{\Theta}(u)) = 1 - \pi(\hat{\Theta}(u)) \). Using all these information, all three functions \( L_1, L_2, \) and \( L_3 \) of the E-step simplify as

\[
L_1(\Theta^{(t+1)}) = \sum_{i,j} \left[ \beta_j^t Z_{ij} + \log \left\{ \frac{1 + e^{\eta_{ij}}}{1 + e^{\eta_{ij}}} \right\} \right. \]
\[
- \log \left\{ \sum_{j=1}^{M+1} \exp \left( \beta_j Z_{ij} + \log \left\{ \frac{1 + e^{\eta_{ij}}}{1 + e^{\eta_{ij}}} \right\} \right) \right\},
\]

\[
L_2(\Theta^{(t+1)}) = \sum_{(i,j): R_{ij} = 1} \left[ X_{ij} (\eta_j + D_j \beta_j)
\right.
\]
\[
- \log \left( 1 + e^{\eta_{ij} + D_j \beta_j} \right) + \log \pi_j (X_{ij}) \],
\]

\[
L_3(\Theta^{(t+1)} | \Theta^{(t)}) = \sum_{(i,j): R_{ij} = 0} \left[ H(\psi_i(\Theta^{(t)})) (\eta_j + D_j \beta_j)
\right.
\]
\[
- \log \left( 1 + e^{\eta_{ij} + D_j \beta_j} \right) + H(\psi_i(\Theta^{(t)})) \log \pi(\hat{\Theta}(1))
\]
\[
+ \left. [1 - H(\psi_i(\Theta^{(t)}))] \log \pi(\hat{\Theta}(0)) \right].
\]

In the M-step, we maximize \( L_1 + L_2 + L_3 \) with respect to the parameter \( \Theta \) 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

\[
p(X | Z, D, S) = \exp \left[ X (\gamma_0 + \gamma Z_j + \gamma S_j + \gamma D_D) \right] - \log \left( 1 + e^{\gamma Z_j + \gamma S_j + \gamma D_D} \right).
\]

Through our calculations, it turns out that \( \gamma_D = \beta_j. \) 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 imputed values of the missing \( X \), which is denoted by \( X \) in that paper. Therefore, as indicated in the paragraph right before equation (7) of that article that \( \beta_j \) (in the name of \( \gamma_D \)) is estimated once through the likelihood \( p(R, X | D, Z, S), \)
Nonignorable Missingness in Matched Case–Control Study

and then again through 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 \( \mu_j = \gamma_0 + \gamma_2 Z_j + \gamma_3 S \) and variance \( \sigma^2 \), then

\[
p(X_j | Z_j, D_j = 0, S_i) = \exp \left\{ \left( \mu_j - \frac{t^2}{2} \right)/\sigma^2 - X_j^2/2\sigma^2 - \log(2\pi\sigma^2)/2 \right\}.
\]

(12)

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

\[
L_1(\Theta^{(j)}) = \frac{1}{\sigma^2} \exp(\sum_{j=1}^M \left[ \beta_1 Z_{ij} + \beta_2 \mu_{ij} - \log(\sum_{j=1}^M \exp(\beta_1 Z_{ij} + \beta_2 \mu_{ij})) \right]).
\]

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

In the M-step, we maximize \( L_1 + L_2 + L_3 \) with respect to the parameter \( \Theta \) 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 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 residing 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 \( X \) the and 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 \( \beta_1 \) and \( \beta_2 \) are the disease-exposure association parameter for \( Z \) and \( X \), respectively.

The exposure distribution among the controls is modeled as \( p(X = 1 | Z, S, D = 0) = H(\gamma_0 + \gamma_2 Z + \gamma_3 S) \), and so the exposure distribution among the cases becomes \( p(X = 1 | Z, S, D = 1) = H(\gamma_0 + \gamma_2 Z + \gamma_3 S + \beta_1) \). We analyze the data by using the CLR technique by the method proposed in Paik (2004), we call it as Paik, and by using the newly proposed method, we call it as 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
Table 1
Results of the analysis of the Los Angeles Endometrial Cancer Data. Est. and SD represent the estimate and the standard error, respectively.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>CLR</th>
<th>Paik</th>
<th>SM</th>
</tr>
</thead>
<tbody>
<tr>
<td>β₁</td>
<td>1.2844 ± 0.3921</td>
<td>1.3098 ± 0.3760</td>
<td>1.2857 ± 0.3741</td>
</tr>
<tr>
<td>β₂</td>
<td>0.4405 ± 0.3800</td>
<td>0.6879 ± 0.3633</td>
<td>0.5346 ± 0.3991</td>
</tr>
<tr>
<td>γ₀</td>
<td>0.0288 ± 0.2875</td>
<td>−0.0139 ± 0.3716</td>
<td></td>
</tr>
<tr>
<td>γ₁</td>
<td>0.1582 ± 0.3623</td>
<td>0.1692 ± 0.3625</td>
<td></td>
</tr>
<tr>
<td>γ₂</td>
<td>0.5390 ± 0.4868</td>
<td>0.5337 ± 0.4850</td>
<td></td>
</tr>
<tr>
<td>γ₃</td>
<td>0.5224 ± 0.3104</td>
<td></td>
<td></td>
</tr>
<tr>
<td>δ₀</td>
<td>1.1606 ± 0.4031</td>
<td>1.0827 ± 0.4013</td>
<td></td>
</tr>
<tr>
<td>δ₁</td>
<td>0.4933 ± 0.5616</td>
<td>0.4875 ± 0.5639</td>
<td></td>
</tr>
<tr>
<td>δ₂</td>
<td>0.6560 ± 0.3112</td>
<td>0.8717 ± 0.3146</td>
<td></td>
</tr>
<tr>
<td>δ₃</td>
<td>0.5750 ± 0.4694</td>
<td>0.5493 ± 0.4711</td>
<td></td>
</tr>
<tr>
<td>δ₄</td>
<td>−0.0403 ± 0.6456</td>
<td>−0.0635 ± 0.6526</td>
<td></td>
</tr>
</tbody>
</table>

from 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 β₁ and β₂ for a set of values of δₓ. Note that when δₓ = 0, the missingness mechanism becomes MAR. The sensitivity of the Paik and SM method is studied when δₓ ∈ [−4, 4]. We set δₓ 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 β₂ changes with δₓ for both the methods along with the 95% confidence interval. Estimate of β₂ obtained by using the SM method ranges from 0.41 to 0.53 whereas β₂ of the Paik varies from 0.44 to 0.82. For a moderate range of values of δₓ, the regression parameter β₂ remains insignificant for both the methods. For the SM method, the other disease–risk parameter β₁ varies from 1.2876 to 1.3013 for δₓ ∈ [−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,
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, 0.24²). Therefore, first, generate the matching variable \( S \) from the Normal(0.53, 0.24²) 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(\( a = 0.23, b = 0.8333 \)), 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 \( \beta_1 \), the log-odds-ratio parameter for \( Z \), and \( \beta_2 \), 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 true and the estimated value. Therefore, for fully observed data, if 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 \( \beta \) 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 \( \beta_1 \) and \( \beta_2 \), 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

\[
L = \prod_{i=1}^{n} \prod_{j=1}^{M+1} p(X_{ij}Z_{ij}, D_{ij}, S_i) \times \prod_{i=1}^{n} p(D_{i1} = 1, D_{i2} = \cdots = D_{iM+1} = 0 \mid Z_{ij})^{M+1} \left( S_i \sum_{j=1}^{M+1} D_{ij} = 1 \right). \tag{15}
\]

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 \( \beta \) 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.
mechanism MM5, as it considers one kind of violation of the model assumption of the selection probability. It turns out that the estimate of \( \beta_2 \) 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 the fact. After a thorough numerical investigation, we found that in the domain of \( S \), 0.5470 \( S^2 \) 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 0.5470 \( S^2 \), 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.

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

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

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 \( \beta_2 \) than the CLR approach. However, \( \text{MSE}_{\text{CLR}} \), the MSE due to the CLR, is still higher than that of the SM, \( \text{MSE}_{\text{SM}} \). The intuitive reason for this behavior is in general \( \sigma_{\text{SM}}^2 < \sigma_{\text{CLR}}^2 \), where \( \sigma_{\text{SM}}^2 \) and \( \sigma_{\text{CLR}}^2 \) are the variance of the estimator under the SM and CLR approach, respectively. Therefore, \( \text{MSE}_{\text{SM}} = \text{Bias}_{\text{SM}}^2 + \sigma_{\text{SM}}^2 < \text{Bias}_{\text{CLR}}^2 + \sigma_{\text{CLR}}^2 = \text{MSE}_{\text{CLR}} \) even though \( \text{Bias}_{\text{SM}} > \text{Bias}_{\text{CLR}} \). When sample size (\( n \)) is large, \( \sigma_{\text{SM}}^2, \sigma_{\text{CLR}}^2 \rightarrow 0 \), 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 \( \beta_1 \) and \( \beta_2 \) are 1.3 and 0.55, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>( \beta_1 )</th>
<th>MSE</th>
<th>( \beta_2 )</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully observed data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLR&amp;Paik</td>
<td>1.3368</td>
<td>0.1689</td>
<td>0.5411</td>
<td>0.1424</td>
</tr>
<tr>
<td>SM</td>
<td>1.2950</td>
<td>0.1394</td>
<td>0.4633</td>
<td>0.0956</td>
</tr>
<tr>
<td>MM1: ( \text{logit}(P(R = 1) = 0.66 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLR</td>
<td>1.6332</td>
<td>0.7014</td>
<td>0.5965</td>
<td>0.4100</td>
</tr>
<tr>
<td>Paik</td>
<td>1.5278</td>
<td>0.2460</td>
<td>0.5277</td>
<td>0.2028</td>
</tr>
<tr>
<td>SM</td>
<td>1.3854</td>
<td>0.1624</td>
<td>0.5072</td>
<td>0.1825</td>
</tr>
<tr>
<td>MM2: ( \text{logit}(P(R = 1) = D )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLR</td>
<td>1.5400</td>
<td>0.6694</td>
<td>0.5877</td>
<td>0.4387</td>
</tr>
<tr>
<td>Paik</td>
<td>1.5960</td>
<td>0.3193</td>
<td>0.6508</td>
<td>0.3022</td>
</tr>
<tr>
<td>SM</td>
<td>1.3816</td>
<td>0.1480</td>
<td>0.5554</td>
<td>0.1543</td>
</tr>
<tr>
<td>MM3: ( \text{logit}(P(R = 1) = X )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLR</td>
<td>1.6461</td>
<td>0.8460</td>
<td>0.6267</td>
<td>0.4033</td>
</tr>
<tr>
<td>Paik</td>
<td>1.5589</td>
<td>0.2297</td>
<td>0.5632</td>
<td>0.2242</td>
</tr>
<tr>
<td>SM</td>
<td>1.3923</td>
<td>0.1503</td>
<td>0.4978</td>
<td>0.1692</td>
</tr>
<tr>
<td>MM4: ( \text{logit}(P(R = 1) = 0.5D + 0.5X )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLR</td>
<td>1.6021</td>
<td>0.7482</td>
<td>0.4175</td>
<td>0.3903</td>
</tr>
<tr>
<td>Paik</td>
<td>1.6091</td>
<td>0.3267</td>
<td>0.5335</td>
<td>0.3463</td>
</tr>
<tr>
<td>SM</td>
<td>1.3895</td>
<td>0.1736</td>
<td>0.4516</td>
<td>0.1978</td>
</tr>
<tr>
<td>MM5: ( \text{logit}(P(R = 1) = DX + XZ )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLR</td>
<td>1.6916</td>
<td>1.1595</td>
<td>0.9521</td>
<td>0.7818</td>
</tr>
<tr>
<td>Paik</td>
<td>1.7161</td>
<td>0.5496</td>
<td>1.0567</td>
<td>0.7405</td>
</tr>
<tr>
<td>SM</td>
<td>1.4971</td>
<td>0.2744</td>
<td>0.8638</td>
<td>0.3822</td>
</tr>
</tbody>
</table>

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 article 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

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

Q1 Author: Please check the edit in the sentence “As shown in” retains the intended meaning.

Q2 Author: Please check the insertion of emdash for corrections in the sentence “The general”.

Q3 Author: Please check the sentence for correctness.