The Canadian Journal of Statistics
Vol. 38, No. 4, 2010, Pages 680–697
La revue canadienne de statistique

An estimated-score approach for dealing with missing covariate data in matched case–control studies

Samiran SINHA*

Department of Statistics, Texas A&M University, College Station, TX 77845, USA

Key words and phrases: Conditional likelihood; efficiency; empirical distribution; influence function; log-odds ratio; missing data.

MSC 2000: Primary 62G05; secondary 62-07.

Abstract: Matched case–control designs are commonly used in epidemiological studies for estimating the effect of exposure variables on the risk of a disease by controlling the effect of confounding variables. Due to retrospective nature of the study, information on a covariate could be missing for some subjects. A straightforward application of the conditional logistic likelihood for analyzing matched case–control data with the partially missing covariate may yield inefficient estimators of the parameters. A robust method has been proposed to handle this problem using an estimated conditional score approach when the missingness mechanism does not depend on the disease status. Within the conditional logistic likelihood framework, an empirical procedure is used to estimate the odds of the disease for the subjects with missing covariate values. The asymptotic distribution and the asymptotic variance of the estimator when the matching variables and the completely observed covariates are categorical. The finite sample performance of the proposed estimator is assessed through a simulation study. Finally, the proposed method has been applied to analyze two matched case–control studies. *The Canadian Journal of Statistics* 38: 680–697; 2010 © 2010 Statistical Society of Canada

Résumé: Des designs cas-contrôle appariés sont souvent utilisés dans des études épidémiologiques afin d'estimer l'effet des variables d'exposition sur le risque d'une maladie en contrôlant l'effet des variables parasites. Par la nature rétrospective de l'étude, l'information sur les covariables peut être manquante chez certains sujets. L'application directe de la vraisemblance logistique conditionnelle pour analyser les données cas-contrôle appariées avec des covariables manquantes partiellement peut conduire à des estimations inefficaces des paramètres. L'auteur propose une méthode robuste pour traiter ce problème en utilisant une approche par score conditionnel estimé lorsque le mécanisme responsable des valeurs manquantes ne dépend pas du statut de la maladie. Dans le cadre de la vraisemblance logistique conditionnelle, il utilise une procédure empirique pour estimer les cotes de la maladie pour les sujets ayant des valeurs de covariables manquantes. L'auteur obtient aussi la distribution et la variance asymptotiques de l'estimateur lorsque les variables d'appariement et les covariables complètement observées sont catégorielles. La performance pour de petits échantillons de l'estimateur proposé est évaluée à l'aide d'une étude de simulations. Finalement, il applique la méthode proposée à l'analyse de deux études cas-contrôle appariés. *La revue canadienne de statistique* 38: 680–697; 2010 © 2010 Société statistique du Canada

1. INTRODUCTION

In this paper we propose a nonparametric method for dealing with a partially missing covariate in matched case–control studies which are commonly used in clinical and epidemiological research. In a matched case–control study each case or diseased subject is matched with a number of controls or nondiseased subjects based on some confounding variables which possibly have influence on

* *Author to whom correspondence may be addressed.*
E-mail: sinha@stat.tamu.edu

the disease risk as well as on the covariate of interest. For analyzing a matched case–control data usually a conditional logistic likelihood is used which allows us to estimate the log-odds ratio parameters without estimating the stratum-specific intercept parameters involved in the disease risk model.

For a missing data scenario a straightforward use of the conditional logistic likelihood yields inefficient and possibly biased estimates of the parameters depending on the missingness mechanism. The reason for loss of efficiency is that in order to apply the conditional logistic likelihood in its standard form to the partially missing data one has to discard completely the subjects with incomplete information. Moreover, if a missing value occurs for a case subject, the entire stratum in which the case belongs has to be removed from the analysis. This method is known as complete-case analysis. For handling binary missing covariate data epidemiologists may use the missing-indicator method proposed by Huberman & Langholz (1999) which can be easily implemented in any statistical software. Although the method can be conveniently applied without much effort, it could produce biased estimates if the covariate status for matched cases and controls are dependent which can arise via the matching variables. As the method is a compromise between the complete-case analysis and an unmatched analysis of the data, it is more efficient than the complete-case analysis (Li, Song & Gary, 2004). However, the method is not applicable if the partially missing covariate is a continuous or a categorical variable with more than two categories.

Therefore, for handling any type of partially missing-at-random covariate data (Little & Rubin, 2002, p. 12) other researchers either suggested to model the distribution of the partially missing covariates parametrically (Paik & Sacco, 2000; Satten & Carroll, 2000; Sinha et al., 2005) or to model the missingness probability parametrically (Lipsitz, Parzen & Ewell, 1998; Rathouz, Satten & Carroll, 2002). Even though the later approaches are usually less efficient, they could be useful if the distribution of the partially missing covariate is difficult to model parametrically. In the context of modelling the distribution of the partially missing covariate, Paik & Sacco (2000) and Satten & Carroll (2000) used different types of conditional likelihood, and the former method is relatively more robust to the distribution of the partially missing covariate and Satten and Carroll's method is more efficient. Sinha et al. (2005) used the Satten and Carroll type of conditional likelihood and modelled the unobserved stratum effect on the parametric distribution of the partially missing covariate via a nonparametric Bayesian approach. Sinha, Mukherjee & Ghosh (2004) extended the idea of Sinha et al. (2005) in the context of multiple disease category data. For handling nonignorable missing data in matched case–control studies, Paik (2004) and Sinha & Maiti (2008) proposed two parametric approaches. The methods proposed in Paik & Sacco (2000) and Paik (2004) are applicable if the partially missing covariate is a member of the canonical exponential family of distributions, whereas the methods of Satten & Carroll (2000), Sinha et al. (2005), Sinha, Mukherjee & Ghosh (2004), and Sinha & Maiti (2008) are applicable to any distribution for the partially missing covariate. In any case, realistic parametric models for partially missing covariates are difficult to construct and the methods based on parametric model assumptions are not robust to model misspecification.

Recently Sinha & Wang (2009) proposed a method for handling missing covariate data where they used a kernel density approach instead of using a parametric model for the distribution of the partially missing covariate. Although the article made an important advancement in this field, there are certain difficulties in the application of the method. First of all, in order to make the asymptotic theory work one needs to work with a higher order kernel (order more than 2) which can be easily constructed from a given kernel. However, a kernel of order more than 2 may produce negative value for the expected disease odds, although logically it must always be positive. Therefore, whenever they encountered this problem in the data analysis and simulation, they had to set the expected value of the disease odds to an arbitrary small positive number without

any real theoretical justification. Second, for handling boundary effects (Karunamuni & Alberts, 2005) one needs to incorporate boundary corrections which involve intensive computation. Third, the choice of an appropriate bandwidth is crucial for timely and accurate parameter estimation. Therefore, a suitable and easy nonparametric method is needed which can circumvent these issues.

In this article we propose an easily implemented nonparametric method which is robust to the distribution of the partially missing covariate. The theoretical justification and practical implementation of this method is much more straightforward than Sinha & Wang (2009). Unlike kernel-based nonparametric approach, we do not need to deal with the issue of bandwidth selection, boundary effects, or occasional negative estimate of $\text{pr}(Y = 1|\boldsymbol{Z}, \boldsymbol{S})/\text{pr}(Y = 0|\boldsymbol{Z}, \boldsymbol{S})$. A brief outline of the proposed method is given below, and the details are given in Section 3. Let Y, $\boldsymbol{S}$, $\boldsymbol{Z}$, and X be the binary disease indicator variable, a set of matching variables, a set of covariates which is always observed, and the partially missing covariate, respectively. Within the framework of a conditional logistic likelihood we replace the odds of the disease $\text{pr}(Y = 1|\boldsymbol{S}, X, \boldsymbol{Z})/\text{pr}(Y = 0|\boldsymbol{S}, X, \boldsymbol{Z})$ by $\text{pr}(Y = 1|\boldsymbol{S}, \boldsymbol{Z})/\text{pr}(Y = 0|\boldsymbol{S}, \boldsymbol{Z})$ if the covariate X is missing, and $\text{pr}(Y = 1|\boldsymbol{S}, \boldsymbol{Z})/\text{pr}(Y = 0|\boldsymbol{S}, \boldsymbol{Z})$ is estimated nonparametrically. One advantage of the conditional likelihood that we are using is that the contribution of the subjects with fully observed covariate remains the same as their contribution to the conditional logistic likelihood without missing data. In our proposed method we assume that both $\boldsymbol{S}$ and $\boldsymbol{Z}$ are categorical, and the missingness mechanism is independent of the disease variable Y and partially missing variable X which will be referred to as case-independent missing-at-random data. We have studied the asymptotic property of the estimator and obtained an analytical formula for the asymptotic standard error of the estimator which are described in Section 4. For deriving the asymptotic distribution we used the empirical process theory which has not been used previously in this context. In essence, our inference is based on an estimated conditional logistic likelihood, and the estimates are obtained by solving an estimated conditional score equation. The original idea of estimated likelihood can be found in Pepe & Fleming (1991) and Carroll & Wand (1991). However, the application of this idea in a conditional likelihood framework is new.

We would like to point out that our formulation of the likelihood is similar to that of Paik & Sacco (2000), but there is an important distinction. Paik & Sacco (2000) modelled the distribution of the partially missing covariate given the completely observed covariates, the matching variables, and the disease status through a member of the generalized exponential family of distributions, and their method produces consistent parameter estimator under missing-at-random data and if the assumed parametric models are correct. On the contrary, we empirically estimate the distribution of the partially missing covariate only among the controls given the completely observed covariates and the matching variables. Our method produces consistent estimators of the parameters if the data are missing completely-at-random or the missingness mechanism does not depend on the disease status. Through a simulation study we assess the performance of the proposed method. Also, in the simulation study we show how we handle continuous $\boldsymbol{S}$ or $\boldsymbol{Z}$. The details of the simulation study are given in Section 5. The simulation study not only shows the efficiency gain of our method compared to the complete-case analysis but also shows the advantage of our method over an useful parametric method when a realistic parametric model for X is difficult to construct. For the purpose of illustration we apply the proposed method to analyze the Los Angeles Endometrial Cancer data (Breslow & Day, 1980) and the low-birth-weight data for the state of Alabama in the year of 1968, and the details of the data analyses are collected in Section 6. Section 7 contains a discussion.

2. MODEL AND ASSUMPTION

Suppose that we have an 1:M matched case–control data with n strata. For the jth subject in the ith stratum we assume a logistic disease risk model

$$\text{pr}(Y_{ij} = 1|\boldsymbol{S}_i, \boldsymbol{Z}_{ij}, X_{ij}) = H\{\beta_{0i}(\boldsymbol{S}_i) + \boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{Z}_{ij} + \beta_2 X_{ij}\}, \quad j = 1, \ldots, (M+1),\ i = 1, \ldots, n,$$

where $H(u) = \exp(u)/\{1 + \exp(u)\}$. The intercept parameter $\beta_{0i}(\boldsymbol{S}_i)$ is an unknown function of the matching variables, $\boldsymbol{S}$, and $\boldsymbol{\beta}_1$ and β_2 are the log-odds ratio parameters corresponding to $\boldsymbol{Z}$ and X, respectively. For notational convenience and save some space, we will denote $\text{pr}(Y_{ij} = 1|X_{ij}, \boldsymbol{V}_{ij})/\text{pr}(Y_{ij} = 0|X_{ij}, \boldsymbol{V}_{ij})$, the odds of the disease for the jth subject in the ith stratum, by $\exp\{\beta_{0i}(\boldsymbol{S}_i)\}O_{ij}$, where $\boldsymbol{V}_{ij} = (\boldsymbol{S}_i^{\mathrm{T}}, \boldsymbol{Z}_{ij}^{\mathrm{T}})^{\mathrm{T}}$ and $O_{ij} = \exp(\boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{Z}_{ij} + \beta_2 X_{ij})$. Then the conditional logistic likelihood function is

$$L = \prod_{i=1}^{n}\left\{\prod_{j=1}^{M+1} O_{ij}^{Y_{ij}} \Big/ \sum_{k=1}^{M+1} O_{ik}\right\}.$$

The parameter estimates are obtained by solving the estimating equation $S_\beta = 0$, where

$$S_\beta \equiv \sum_{i=1}^{n}\sum_{j=1}^{M+1}\left\{Y_{ij} - \frac{O_{ij}}{\sum_{k=1}^{M+1} O_{ik}}\right\}\frac{\partial \log(O_{ij})}{\partial \boldsymbol{\beta}},$$

and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \beta_2)^{\mathrm{T}}$. For dealing with missing data we introduce nonmissing value indicator R which takes on one if X is observed for a subject and zero otherwise.

2.1. Missingness Mechanism

We assume

$$\text{pr}(R = 1|X, \boldsymbol{V}, Y) = \text{pr}(R = 1|\boldsymbol{V}) \equiv \pi(\boldsymbol{V}),$$

thus the missingness mechanism depends neither on the unobserved variable nor on the disease status. Furthermore, we assume $\pi(\boldsymbol{V})$ does not depend on $\boldsymbol{\beta}$. Observe that this case-independent missing-at-random mechanism is more restrictive than the missing-at-random mechanism (Little & Rubin, 2002) where $\text{pr}(R = 1|X, \boldsymbol{V}, Y)$ may depend on both $\boldsymbol{V}$ and Y. On the other hand, this mechanism is more flexible than missing-completely-at-random where missingness mechanism does not depend on X, $\boldsymbol{V}$, or Y.

In the complete-case analysis $\boldsymbol{\beta}$ is obtained by solving

$$S_\beta^{\text{cc}} \equiv \sum_{\left\{i:\sum_{j=1}^{M+1} R_{ij}Y_{ij}=1, \sum_{j=1}^{M+1} R_{ij}(1-Y_{ij})\geq 1\right\}}\ \sum_{\{j:R_{ij}=1\}}\left\{Y_{ij} - \frac{O_{ij}}{\sum_{k=1}^{M+1} R_{ik}O_{ik}}\right\}\frac{\partial \log(O_{ij})}{\partial \boldsymbol{\beta}} = 0, \tag{1}$$

and it is easy to verify that the complete-case analysis produces consistent estimators for the parameters when the data are missing-completely-at-random.

3. ESTIMATION METHODOLOGY

Let $p(x, \boldsymbol{z}|y, \boldsymbol{s})$ and $p(\boldsymbol{z}|y, \boldsymbol{s})$ be the conditional probability of observing $X = x$ and $\boldsymbol{Z} = \boldsymbol{z}$ given $Y = y$ and $\boldsymbol{S} = \boldsymbol{s}$ and $\boldsymbol{Z} = \boldsymbol{z}$ given $Y = y$ and $\boldsymbol{S} = \boldsymbol{s}$, respectively. In a matched case–control study the sampling is done from $p(X_{ij}, \boldsymbol{Z}_{ij}|Y_{ij} = y, \boldsymbol{S}_i)$ or $p(\boldsymbol{Z}_{ij}|Y_{ij} = y, \boldsymbol{S}_i)$ depending on whether we observe X (i.e., $R_{ij} = 1$) or not (i.e., $R_{ij} = 0$), and following the idea of Hosmer & Lemeshow (2000) and Paik & Sacco (2000) for case-independent missing-at-random data, the unconditional

likelihood is (ignoring the terms which are independent of $\boldsymbol{\beta}$)

$$\prod_{i=1}^{n}\prod_{j=1}^{M+1} p^{R_{ij}}(X_{ij}, \boldsymbol{Z}_{ij}|Y_{ij}, \boldsymbol{S}_i)p^{1-R_{ij}}(\boldsymbol{Z}_{ij}|Y_{ij}, \boldsymbol{S}_i),$$

and the conditional likelihood after conditioning that each stratum has one case subject and M controls is

$$L_{\mathrm{c}} = \prod_{i=1}^{n}\left[\prod_{j=1}^{M+1} p^{R_{ij}}(X_{ij}, \boldsymbol{Z}_{ij}|Y_{ij}, \boldsymbol{S}_i)p^{1-R_{ij}}(\boldsymbol{Z}_{ij}|Y_{ij}, \boldsymbol{S}_i) \div \sum_{k=1}^{M+1}\left\{p^{R_{ik}}(X_{ik}, \boldsymbol{Z}_{ik}|1, \boldsymbol{S}_i)p^{1-R_{ik}}(\boldsymbol{Z}_{ik}|1, \boldsymbol{S}_i)\prod_{j\neq k} p^{R_{ij}}(X_{ij}, \boldsymbol{Z}_{ij}|0, \boldsymbol{S}_i)p^{1-R_{ij}}(\boldsymbol{Z}_{ij}|0, \boldsymbol{S}_i)\right\}\right]. \tag{2}$$

After dividing the numerator and denominator of (2) by $\prod_{i=1}^{n}\prod_{j=1}^{M+1} p^{R_{ij}}(X_{ij}, \boldsymbol{Z}_{ij}|0, \boldsymbol{S}_i)p^{1-R_{ij}}(\boldsymbol{Z}_{ij}|0, \boldsymbol{S}_i)$, we obtain the conditional likelihood under missing data

$$L_{\mathrm{c}} = \prod_{i=1}^{n}\left[\prod_{j=1}^{M+1}\{R_{ij}O_{ij} + (1-R_{ij})Q_{ij}\}^{Y_{ij}} \Big/ \sum_{k=1}^{M+1}\{R_{ik}O_{ik} + (1-R_{ik})Q_{ik}\}\right],$$

where

$$Q_{ij} \equiv \mathrm{pr}(Y_{ij}=1|\boldsymbol{Z}_{ij}, \boldsymbol{S}_i)/\mathrm{pr}(Y_{ij}=0|\boldsymbol{Z}_{ij}, \boldsymbol{S}_i).$$

Let $P(x|\boldsymbol{V}, Y=0)$ be the conditional cumulative distribution function of X among the controls given $\boldsymbol{V}$, that is, $P(x|\boldsymbol{V}, Y=0) \equiv \mathrm{pr}(X \leq x|\boldsymbol{V}, Y=0)$. Then after a few steps of algebra one can show that

$$Q_{ij} = \int \{\mathrm{pr}(Y_{ij}=1|x, \boldsymbol{Z}_{ij}, \boldsymbol{S}_i)/\mathrm{pr}(Y_{ij}=0|x, \boldsymbol{Z}_{ij}, \boldsymbol{S}_i)\}\,\mathrm{d}P(x|\boldsymbol{V}_{ij}, Y=0). \tag{3}$$

When $P(x|\boldsymbol{V}, Y=0)$ is known, $\boldsymbol{\beta}$ is estimated by solving the score equation $S_\beta^{\mathrm{m}} \equiv \partial L_{\mathrm{c}}/\partial\boldsymbol{\beta} = 0$, where

$$S_\beta^{\mathrm{m}} = \sum_{i=1}^{n}(Y_{ij} - D_{ij})\left\{R_{ij}\frac{\partial \log(O_{ij})}{\partial\boldsymbol{\beta}} + (1-R_{ij})\frac{\partial \log(Q_{ij})}{\partial\boldsymbol{\beta}}\right\}.$$

with

$$D_{ij} = \frac{R_{ij}O_{ij} + (1-R_{ij})Q_{ij}}{\sum_{k=1}^{M+1}\{R_{ik}O_{ik} + (1-R_{ik})Q_{ik}\}}.$$

For case-independent missing-at-random assumption, $E(S_\beta^{\mathrm{m}}) = 0$ as $E(Y_{ij}|\{R_{ik}, X_{ik}, \boldsymbol{V}_{ik}\}_{k=1}^{M+1}) = D_{ij}$.

Since, $P(x|\boldsymbol{V}, Y = 0)$ is not known, we estimate Q_{ij} using an empirical estimate of $P(x|\boldsymbol{V}, Y = 0)$ (Pepe & Fleming, 1991). The empirical estimates of $P(x|\boldsymbol{V}, Y = 0)$ are

$$\hat{P}(x|\boldsymbol{V}, Y = 0) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{M+1} I(X_{ij} \le x, R_{ij} = 1, \boldsymbol{V}_{ij} = \boldsymbol{V}, Y_{ij} = 0)}{\sum_{i=1}^{n}\sum_{j=1}^{M+1} I(R_{ij} = 1, \boldsymbol{V}_{ij} = \boldsymbol{V}, Y_{ij} = 0)},$$

where I is the indicator function, and this yields an unbiased estimate of Q_{ij}

$$\begin{aligned}\hat{Q}_{ij} &= \frac{\sum_{k=1}^{n}\sum_{l=1}^{M+1} O_{kl} I(R_{kl} = 1, \boldsymbol{V}_{kl} = \boldsymbol{V}_{ij}, Y_{kl} = 0)}{\sum_{k=1}^{n}\sum_{l=1}^{M+1} I(R_{kl} = 1, \boldsymbol{V}_{kl} = \boldsymbol{V}_{ij}, Y_{kl} = 0)} \\ &= \int \frac{\mathrm{pr}(Y_{ij} = 1|x, \boldsymbol{Z}_{ij}, \boldsymbol{S}_i)}{\mathrm{pr}(Y_{ij} = 0|x, \boldsymbol{Z}_{ij}, \boldsymbol{S}_i)}\, \mathrm{d}\hat{P}(x|\boldsymbol{V}_{ij}, Y = 0).\end{aligned}$$

When X is a categorical variable, $\mathrm{d}P(x|\boldsymbol{V}_{ij}, Y = 0) = \mathrm{pr}(X = x|\boldsymbol{V}_{ij}, Y = 0)$ and it is estimated by

$$\mathrm{d}\hat{P}(x|\boldsymbol{V}_{ij}, Y = 0) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{M+1} I(X_{ij} = x, R_{ij} = 1, \boldsymbol{V}_{ij} = \boldsymbol{V}, Y_{ij} = 0)}{\sum_{i=1}^{n}\sum_{j=1}^{M+1} I(R_{ij} = 1, \boldsymbol{V}_{ij} = \boldsymbol{V}, Y_{ij} = 0)}.$$

Now, we estimate $\boldsymbol{\beta}$ by solving the estimated-score equation

$$S_{\beta}^{\mathrm{em}} \equiv \sum_{i=1}^{n}(Y_{ij} - \hat{D}_{ij})\left\{R_{ij}\frac{\partial \log(O_{ij})}{\partial \boldsymbol{\beta}} + (1 - R_{ij})\frac{\partial \log(\hat{Q}_{ij})}{\partial \boldsymbol{\beta}}\right\} = 0,$$

where

$$\hat{D}_{ij} \equiv \frac{R_{ij}O_{ij} + (1 - R_{ij})\hat{Q}_{ij}}{\sum_{k=1}^{M+1}\{R_{ik}O_{ik} + (1 - R_{ik})\hat{Q}_{ik}\}}.$$

We would like to point out that we estimate Q_{ij} given in (3) empirically based on the observed data on X among the control population. In matched case–control study usually the covariate distribution varies across the strata. The dependence of the covariate distribution on the strata could be of two types. First, the covariate X belongs to a family of distributions whose mean, variance, or other parameters can depend on the stratification variable functionally. Second, the covariate X has different distributions for each stratum. For instance, X has a normal distribution for stratum $i = 1$ and an exponential distribution for stratum $i = 2$. In our estimation procedure we cannot accommodate the second possibility as it would require theoretically infinite number of controls in each stratum which is practically impossible. None of Satten & Carroll (2000), Paik & Sacco (2000), Rathouz, Satten & Carroll (2002), and Sinha et al. (2005) tries to address the second possibility in the missing data context. Furthermore, even under the first type of dependence between the covariate X and the stratification variables $\boldsymbol{S}$, the disease probability does vary across the strata, and consequently, it requires a conditional likelihood method for efficient estimation of the parameters.

4. ASYMPTOTIC RESULTS

In this section we study the asymptotic behaviour of the proposed estimator. We assume that the number of controls in each stratum, M is fixed. Let n_{yr} be the number of observations with $Y = y$ and $R = r$, for $r, y = 0, 1$, and define $n_{+r} \equiv n_{0r} + n_{1r}$. We assume that $n_{+0}/n_{01} \to \rho_1$ and

$n_{10}/n_{01} \to \rho_2$ as the sample size $n \to \infty$. Furthermore, define $\mathcal{I} = \lim_{n\to\infty} -(1/n)(\partial S_\beta^{\text{em}}/\partial \boldsymbol{\beta})$. In the Appendix we show that in a neighbourhood of the true $\boldsymbol{\beta}$, the solution $\hat{\boldsymbol{\beta}}$ obtained by solving $S_\beta^{\text{em}} = 0$ is consistent for $\boldsymbol{\beta}$. We also show that the $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed with the following influence function representation, that is,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathcal{I}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathcal{Q}_{ni} + o_p(1),$$

where $\mathcal{Q}_{ni}$ are independently and identically distributed (i.i.d.) mean zero random variables and

$$\begin{aligned}
\mathcal{Q}_{ni} = \sum_{j=1}^{M+1} \Bigg[& (Y_{ij} - D_{ij}) \left\{ R_{ij} \frac{\partial \log(O_{ij})}{\partial \beta} + (1 - R_{ij}) \frac{\partial \log(Q_{ij})}{\partial \beta} \right\} \\
& - \rho_1 (1 - Y_{ij}) R_{ij} \left\{ \frac{\partial O_{ij}}{\partial \boldsymbol{\beta}} E\left(\frac{D}{Q} \middle| \boldsymbol{V} = \boldsymbol{V}_{ij}, R = 0 \right) - O_{ij} E\left(\frac{D}{Q^2} \frac{\partial Q}{\partial \boldsymbol{\beta}} \middle| \boldsymbol{V} = \boldsymbol{V}_{ij}, R = 0 \right) \right\} \\
& + \rho_2 (1 - Y_{ij}) R_{ij} \left\{ \frac{\partial O_{ij}}{\partial \boldsymbol{\beta}} E\left(\frac{1}{Q} \middle| \boldsymbol{V} = \boldsymbol{V}_{ij}, Y = 1, R = 0 \right) \right. \\
& \left. - O_{ij} E\left(\frac{1}{Q^2} \frac{\partial Q}{\partial \boldsymbol{\beta}} \middle| \boldsymbol{V} = \boldsymbol{V}_{ij}, Y = 1, R = 0 \right) \right\} \Bigg].
\end{aligned}$$

Therefore, the sandwich variance estimator of $\hat{\boldsymbol{\beta}}$ is $(1/n^2)\hat{\mathcal{I}}^{-1}(\sum_{i=1}^{n} \hat{\mathcal{Q}}_{ni}\hat{\mathcal{Q}}_{ni}^{\mathrm{T}})\hat{\mathcal{I}}^{-1}$, where $\hat{\mathcal{I}} = -(1/n)\partial S_\beta^{\text{em}}/\partial \boldsymbol{\beta}$ and its expression is given in the Appendix. In $\hat{\mathcal{Q}}_{ni}$ we replace D_{ij}, Q_{ij}, and $\partial Q/\partial \boldsymbol{\beta}$ by $\hat{D}_{ij}$, $\hat{Q}_{ij}$, $\partial \hat{Q}/\partial \boldsymbol{\beta}$ and the expectations are replaced by the corresponding empirical averages.

5. SIMULATION STUDY

We assessed the performance of our proposed method through two sets of simulation studies. In the first set of simulations we considered 1:1 matched case–control data sets with a binary X and in the second set of simulations we considered 1:2 matched case–control data sets with a continuous X. In order to create matched case–control data we initially simulated (S, Z, X, Y) for every subject of a large cohort, and then based on the values of S we created 1:1 and 1:2 matched case–control data with $n = 500$ strata. We simulated S from a Normal(0, 1) distribution, and Z from a Bernoulli($p = 0.35$) distribution. For the first and second scenarios, we simulated X given S and Z from a Bernoulli$\{p = H(-2 + Z + S)\}$ distribution and a two-component mixture of normal distributions (1/2)Normal$(1.2 + S + Z, 0.6^2)$ + (1/2)Normal$(-1.2 - S - Z, 0.6^2)$, respectively. For given values of S, X, and Z, we simulated Y from a Bernoulli$\{p = H(\beta_0 + 1S + \beta_1 Z + \beta_2 X)\}$ distribution, and we set $\beta_0 = -3.5$, $\beta_1 = 0.5$, and $\beta_2 = 1$ which resulted in approximately 8% diseased subjects in every cohort.

In order to create missing values in X for every subject of a matched case–control data we simulated a nonmissing value indicator R from a Bernoulli distribution. If R takes on one we consider X is observed, otherwise X is missing. We considered four different missingness mechanisms with logit$\{\text{pr}(R = 1|S, X, Y, Z)\} = 0.6 + 0.5Z + 0.35S$, logit$\{\text{pr}(R = 1|S, X, Y, Z)\} = -0.1 + 0.5Z + 0.35S$, logit$\{\text{pr}(R = 1|S, X, Y, Z)\} = 0.45 + Y + 0.35Z$, and logit$\{\text{pr}(R = 1|S, X, Y, Z)\} = 0.25 + Y + Z$, respectively. These four missingness mechanisms resulted in approximately 30%, 50%, 30% and 30% missing data, respectively. The first two missing mechanisms are case-independent missing-at-random. The third and fourth missingness mechanisms were considered to assess the amount of bias in our proposed method

when the missingness mechanism actually depends on the disease status and other observed variables.

Under each scenario and each missingness mechanism we generated 500 data sets, and each simulated data set was analyzed by the following methods. First, we assumed that there was no missing value, and the fully observed data were analyzed by the conditional logistic likelihood method. Next, we considered the data with missing values in X and estimated the parameters by the complete-case method, by our proposed method which is referred to as estimated-score approach, and by the missing-indicator method of Huberman & Langholz (1999). Note that the missing-indicator method was applied only for the case of binary X. Furthermore, each simulated data set was analyzed by a parametric method. Between the two alternative parametric approaches, Satten & Carroll (2000) and Paik & Sacco (2004), we chose to compare with the method of Paik and Sacco because their conditional likelihood is similar to our conditional likelihood and easy to implement. However, we point out that the method of Paik and Sacco is not exactly the parametric counterpart of our method. Their method required us to model the distribution of X parametrically in terms of S, Y, and Z. Note that for the analysis of the data sets by our proposed method we had to transform the continuous matching variable to a categorical variable, and we redefined S as follows:

$$S = \begin{cases} 0 & \text{if } S < a_0 \\ k & \text{if } a_{k-1} < S < a_k \text{ for } k = 1, 2, 3 \\ 4 & \text{if } S > a_3, \end{cases}$$

where a_0, a_1, a_2, and a_3 are the 25th, 40th, 60th, and 75th percentiles of the observed values of continuous S. While we categorize continuous variables we should keep in mind that there should be at least a few observations for each possible combinations of the resultant set of categorical variables $\boldsymbol{V}$, the nonmissing indicator R, and the disease status Y. Indeed with categorized S our model is slightly misspecified which, in turn, allows us to judge the performance of our method under slight model misspecification. For both parameters β_1 and β_2, we present the estimates (EST), the mean squared error (MSE), the true standard error (TSE) which is obtained from the estimates across the simulated data sets, the mean estimated standard error (ESE), and the coverage probability (CP) based on the 95% Wald-type confidence intervals. The ESE for the parametric approach was calculated based on the jackknife method, and the ESE for the estimated-score approach was obtained by the sandwich formula given in Section 4.

The results for scenario 1 are presented in Table 1. The complete-case analysis shows maximum bias in the estimate of β_2 for case-independent missing-at-random. The missing-indicator method shows maximum bias in the estimate of β_1 for all types of missing data except the third missingness mechanism. Intuitive explanations of high bias in β_1 for the missing-indicator method are (1) the association between X and Z and (2) incorporation of the missing indicator (R^c) in the analysis where $\text{logit}\{\text{pr}(R^c = 1|S, X, Y, Z)\} = \text{logit}\{\text{pr}(R = 0|S, X, Y, Z)\}$. Note that R^c depends on Z which may cause some multicollinearity problem. This fact is somewhat evident in the simulation results as the bias in β_1 increases with the percentage of missing data. Also, the magnitude of this bias depends on how the missing indicator is associated with Z and other variables. For case-independent missing-at-random data, the parametric and the estimated-score method perform equally well in terms of bias and variance. Here we point out that for the parametric analysis, we correctly modelled the distribution of X given S, Z, and Y using a linear-logistic model. Importantly, the gain in efficiency in the estimated-score approach compared to the complete-case method is very significant. For case-dependent missing data, the estimated-score method shows maximum bias in the estimate of β_2 compared to any other methods. The simulation results indicate that the standard error of the estimated-score approach is somewhat smaller than the parametric approach. Intuitively, we can say that this parametric approach is not the exact para-

TABLE 1: Simulation results for 1:1 matched data.

	β_1					β_2				
Method	EST	MSE	TSE	ESE	CP	EST	MSE	TSE	ESE	CP
Fully observed data	0.50	0.02	0.14	0.14	0.95	1.01	0.03	0.16	0.16	0.94
Logit{pr($R = 1\|S, X, Y, Z$)} $= 0.6 + 0.5Z + 0.35S$, approximately 30% missing data										
Complete case	0.50	0.04	0.20	0.20	0.96	1.03	0.05	0.23	0.22	0.95
Missing indicator	0.56	0.03	0.14	0.14	0.93	0.99	0.03	0.18	0.18	0.95
Parametric	0.50	0.02	0.15	0.14	0.94	1.00	0.04	0.19	0.19	0.95
Estimated score	0.51	0.02	0.15	0.14	0.93	1.01	0.03	0.19	0.19	0.96
Logit{pr($R = 1\|S, X, Y, Z$)} $= -0.1 + 0.5Z + 0.35S$, approximately 50% missing data										
Complete case	0.51	0.07	0.27	0.26	0.94	1.05	0.09	0.29	0.29	0.96
Missing indicator	0.59	0.03	0.14	0.14	0.90	0.98	0.04	0.20	0.20	0.94
Parametric	0.50	0.02	0.15	0.15	0.95	1.01	0.05	0.22	0.21	0.95
Estimated score	0.51	0.02	0.15	0.15	0.94	1.00	0.04	0.20	0.21	0.96
Logit{pr($R = 1\|S, X, Y, Z$)} $= 0.45 + Y + 0.35Z$, approximately 30% missing data										
Complete case	0.44	0.04	0.20	0.19	0.94	1.02	0.05	0.22	0.22	0.95
Missing indicator	0.49	0.02	0.15	0.15	0.95	1.01	0.04	0.16	0.19	0.93
Parametric	0.50	0.02	0.15	0.14	0.94	0.98	0.03	0.18	0.18	0.95
Estimated score	0.52	0.02	0.14	0.14	0.96	0.89	0.04	0.16	0.16	0.90
Logit{pr($R = 1\|S, X, Y, Z$)} $= 0.25 + Y + Z$, approximately 30% missing data										
Complete case	0.33	0.07	0.20	0.19	0.82	1.02	0.05	0.22	0.22	0.95
Missing indicator	0.38	0.04	0.15	0.15	0.86	1.00	0.04	0.19	0.19	0.94
Parametric	0.50	0.02	0.15	0.14	0.94	0.99	0.03	0.18	0.18	0.94
Estimated score	0.54	0.02	0.14	0.14	0.97	0.91	0.03	0.16	0.16	0.91

The partially missing covariate X was simulated from the Bernoulli distribution with success probability $p(X = 1|Z, S, Y) = H(-2 + Z + S)$. The true values of β_1 and β_2 are 0.5 and 1, respectively. EST, MSE, TSE, ESE, and CP stand for estimate, mean squared error, true standard error, estimated standard error, and the coverage probability based on the 95% Wald-type confidence intervals.

metric counterpart of our nonparametric approach. Secondly, the parametric approach essentially estimates the log-odds ratio parameter twice ignoring the relationship between the distributions of the covariate among the cases and controls which has been discussed at length in Sinha & Maiti (2008).

The results for scenario 2 are presented in Table 2. For case-independent missing-at-random data, the estimated-score approach shows least bias in the estimates of β_1 and β_2, and the maximum bias is observed in the parametric method. Note that in the parametric method we modelled the distribution of X by a normal distribution whose mean was a linear function of S, Z, and Y. Thus, the bias of the parametric method is due to a model misspecification. Also, the estimated-score approach has least MSE values for case-independent missing-at-random data. When the missingness depends on the disease status strongly, the estimated score and the parametric approaches show significant bias in the estimate of β_2.

Our approach requires that $\boldsymbol{V}$ must be categorical with a relatively small number of categories. This may actually be a potential problem if precise matching variables are unknown and a surrogate

TABLE 2: Simulation results for 1:2 matched data.

	β_1					β_2				
Method	EST	MSE	TSE	ESE	CP	EST	MSE	TSE	ESE	CP
Fully observed data	0.49	0.04	0.19	0.20	0.96	1.02	0.01	0.09	0.09	0.97
Logit{pr($R = 1\|S, X, Y, Z$)} $= 0.6 + 0.5Z + 0.35S$), approximately 30% missing data										
Complete case	0.49	0.07	0.27	0.27	0.96	1.03	0.02	0.14	0.13	0.98
Parametric	0.81	0.13	0.18	0.18	0.62	0.93	0.01	0.07	0.07	0.77
Estimated score	0.51	0.04	0.19	0.18	0.93	1.01	0.01	0.10	0.10	0.95
Estimated score*	0.55	0.03	0.18	0.16	0.92	0.97	0.01	0.08	0.08	0.92
Logit{pr($R = 1\|S, X, Y, Z$)} $= -0.1 + 0.5Z + 0.35S$, approximately 50% missing data										
Complete case	0.48	0.14	0.37	0.34	0.95	1.05	0.05	0.21	0.17	0.96
Parametric	0.93	0.22	0.18	0.18	0.36	0.90	0.02	0.07	0.07	0.64
Estimated score	0.51	0.05	0.21	0.18	0.92	1.02	0.01	0.12	0.11	0.95
Estimated score*	0.57	0.04	0.19	0.16	0.86	0.97	0.01	0.09	0.09	0.90
Logit{pr($R = 1\|S, X, Y, Z$)} $= 0.45 + Y + 0.35Z$, approximately 30% missing data										
Complete case	0.42	0.07	0.24	0.24	0.93	1.02	0.01	0.12	0.11	0.95
Parametric	0.59	0.04	0.19	0.19	0.94	1.10	0.02	0.10	0.10	0.87
Estimated score	0.48	0.03	0.17	0.15	0.93	0.92	0.01	0.08	0.07	0.74
Estimated score*	0.52	0.03	0.16	0.15	0.92	0.88	0.02	0.06	0.06	0.50
Logit{pr($R = 1\|S, X, Y, Z$)} $= 0.25 + Y + Z$, approximately 30% missing data										
Complete case	0.31	0.10	0.25	0.24	0.87	1.02	0.01	0.12	0.11	0.95
Parametric	0.57	0.04	0.19	0.20	0.96	1.09	0.02	0.10	0.10	0.88
Estimated score	0.49	0.03	0.17	0.15	0.94	0.92	0.01	0.08	0.07	0.77
Estimated score*	0.54	0.03	0.16	0.15	0.92	0.89	0.02	0.07	0.07	0.55

The partially missing covariate X was simulated from $(1/2)\text{normal}(1.2 + S + Z, 0.6^2) + (1/2)\text{normal}(-1.2 - S - Z, 0.6^2)$. The true values of β_1 and β_2 are 0.5 and 1, respectively. EST, MSE, TSE, ESE, and CP stand for estimate, mean squared error, true standard error, estimated standard error, and the coverage probability based on the 95% Wald-type confidence intervals.

*The results of the estimated-score approach when the matching variable S was ignored from the analysis.

is used (e.g., same household or same community) which is assumed to ensure matching on a range of unmeasured covariates. Thus, in order to judge the performance of our method when matching variables are not properly taken into account in the analysis, we analyzed the simulated data sets under scenario 2 without considering S in the estimation of Q_{ij}. The corresponding results are presented in Table 2 and indicated by * notation. The results indicate that the magnitude of these biases depend heavily on the missingness mechanism. Also, these biases depend on how strongly S and X are associated (results not shown here).

In summary, we can conclude that the proposed estimated-score approach produces least biased estimate of the parameters when the missingness mechanism does not depend on the disease status. The gain in efficiency for estimating β_2 could be as high as 57% compared to the complete-case analysis. If the missingness mechanism depends on disease status, the estimated-score method produces biased estimates and the magnitude of bias depends on the percentage of missing data and the degree of association between the missing indicator and disease status.

6. DATA EXAMPLES

6.1. Los Angeles Endometrial Cancer Data

The Los Angeles Endometrial data (Breslow & Day, 1980) have been analyzed in many articles, such as, Satten & Carroll (2000), Sinha & Maiti (2008), and Sinha & Wang (2009), among others. In order to study the effect of several risk factors on endometrial cancer, a study was conducted among post-menopausal women in an affluent retirement community of Los Angeles. The data consist of $n = 63$ cases and each case was matched with $M = 4$ controls based on age and residence in the same community as the case subject. Among several measured risk factors, the binary exposure variable obesity was missing for about 16% of the study participants. Obesity, a binary indicator variable, is treated as the partially missing exposure variable (X) and presence of gall-bladder disease is considered as a binary completely observed covariate (Z). We will treat age as the matching variable (S). Furthermore, Y indicates the disease status and R stands for the nonmissing value indicator.

We assume that data are missing-at-random, although there is no way to validate this assumption. Next, we fit a simple logistic regression model to find the effect of S, Y, and Z on the missing probability, and the results do not indicate that missingness mechanism is associated with the disease status Y.

In the analysis age was used as a matching variable. The disease risk model of our interest is $H(\beta_{0i} + \beta_1 Z_{ij} + \beta_2 X_{ij})$, where β_1 and β_2 are the disease-exposure association parameter for Z and X, respectively. The data were analyzed by four different methods.

First, we analyzed the data by the complete-case method. Second, we analyzed the data using the parametric approach. For the parametric method the distribution of the missing covariate was modelled as $\text{logit}\{\text{pr}(X = 1|S, Z, Y)\} = \gamma_0 + \gamma_1 S + \gamma_2 Z + \gamma_3 Y$. For this parametric analysis the matching variable (S) was first transformed into [0, 1] scale by subtracting the minimum of the observed ages and then dividing it by the range of the observed ages. The estimates (standard error) of γ_0, γ_1, γ_2, and γ_3 were 0.115(0.319), 0.118(0.387), 0.543(0.539), and 0.479(0.335), respectively. These estimates are used to obtain the estimates of β_1 and β_2. Third, we analyzed the data using the missing-indicator method. In the missing-indicator method, missing values of X are replaced by 0, and we refer to the new variable as X^*. Then, we use simple conditional logistic regression analysis using Z, X^*, and $(1 - R)$ as the covariates.

Finally, we apply the estimated-score approach to this data set. We categorize the matching variable age into five groups as described in Section 5, and make sure that none of the cell frequencies of the contingency table based on S, Z, and Y when $R = 1$, is zero. When we categorize a continuous variable we need to keep in mind that a larger number of categories may be more flexible in terms of modelling the association between the variable and X at the cost of increased variance of the estimators. On the other hand, a smaller number of categories may produce a smaller variance at the expense of losing some important features of that association. The results of these analyses are presented in Table 3, and we found that, based on all four methods, presence of gall-bladder disease has a statistically significant effect of the risk of endometrial cancer at 5% level of significance. None of the methods, except the parametric approach, shows any statistical evidence of association between obesity and the risk of endometrial cancer. However, the standard errors for the parametric and the estimated-score approach are somewhat lower than that of the complete-case and the missing-indicator methods.

6.2. Low-Birth-Weight Data for the State of Alabama

For the purpose of illustration of the proposed method we considered the birth data for the state of Alabama during the year of 1968. The micro data file is available on the Center for Disease Control Website, `http://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm`. The data contain a total of 31,801 live births in Alabama which consist of 66.2% White, 0.20%

TABLE 3: The results of the analyses of the Los Angeles Endometrial Cancer Data.

Method	Presence of gall bladder disease	Obesity
Complete case		
EST	1.28	0.44
SE	0.39	0.38
P-value	<0.01	0.24
Parametric		
EST	1.27	0.60
SE	0.36	0.29
P-value	<0.01	0.04
Missing indicator		
EST	1.28	0.65
SE	0.38	0.36
P-value	<0.01	0.07
Estimated score		
EST	1.35	0.63
SE	0.37	0.34
P-value	<0.01	0.06

EST, estimate; SE, standard error.

Negro, and 33.6% other non-White. A total of 2,831 newborns had birth weight <2,500 g which is generally considered as low-birth-weight. Low-birth-weight is one of the possible causes of infant mortality, and the children with low-birth-weight are at increased risk of lifelong disabilities such as blindness, deafness, and cerebral palsy, making this an extremely important indicator of child well-being. The data contain information on several variables, such as mother's age, race, gestation period, number of live born children, etc. One important feature of the data is that gestation period was missing for many subjects. The subjects with birth weight <2,500 g will be considered as cases ($Y = 1$) and otherwise controls ($Y = 0$). From this cohort of live births who were White we randomly selected $n = 500$ cases and each case was matched with two controls based on mother's age, and the controls were also chosen from the White population. After forming the 1:2 matched case–control data we found gestation period was missing for about 10% of the subjects. The goal was to estimate the effect of number of live born children and gestation period on the risk of having low-birth-weight baby in the White population taking into account the confounding effect of mother's age. Here we treated the number of live born children as a categorical variable, Z, which was defined as

$$Z = \begin{cases} 1 & \text{if the number of children born alive is one} \\ 2 & \text{if the number of children born alive is two} \\ 3 & \text{if the number of children born alive is three} \\ 4 & \text{if the number of children born alive is more than three.} \end{cases}$$

We treated $Z = 1$ as the reference category. Gestation period was measured in weeks, and in our sampled data the minimum and maximum values of gestation period were 20 and 43 weeks, respectively. We defined $X = (\text{gestation period} - 20)/4.28$, and analyzed the data treating X

TABLE 4: Results of the analysis of the low-birth-weight data.

Method	Number of live born children			Gestation period
	2	3	>3	
Complete case				
EST	−0.29	0.42	0.67	−3.61
SE	0.24	0.29	0.32	0.34
P-value	0.23	0.14	0.04	<0.01
Parametric				
EST	0.16	0.12	0.19	−3.18
SE	0.21	0.25	0.28	0.26
P-value	0.46	0.62	0.50	<0.01
Estimated score				
EST	0.16	0.15	0.25	−3.16
SE	0.22	0.24	0.27	0.23
P-value	0.47	0.54	0.36	<0.01

EST, estimate; SE, standard error.

and Z as the two risk factors. Importantly, we found that among the cases, X was missing for approximately 8.6% of the subjects whereas among the controls, it was missing for about 10% of the subjects. Therefore, ignoring the effect of S and Z, a two-sample test for proportions indicates no statistical evidence of the claim that the missingness mechanism depends on the disease status of the subjects.

First, we analyzed the data using the complete-case method. Second, we analyzed the data using the parametric approach described in Section 5. Since, gestation period was a continuous variable, we modelled it by a normal distribution in the parametric method. Third, we analyzed the data using the estimated-score approach. For our analysis, we categorized the matching variable age into five groups as described in Section 5.

The results are presented in Table 4. In all three methods we find that gestation period has a statistically significant effect on the risk of having low-birth-weight baby. Overall, longer gestation period reduces the risk of low-birth-weight baby. More specifically, based on the proposed method the risk of a low-birth-weight baby is $\exp(3.159) = 23.54$-fold higher for a gestation period of <4.28 weeks. The strong association between gestation period and low-birth-weight is not surprising as longer gestation period allows the baby to grow larger in mother's womb, and thus reduces the risk of low-birth-weight. Although the distribution of gestation period is clearly non-normal (Figure 1), due to small percentage of missing values there is no significant difference in the results due to the estimated-score and the parametric approach. This phenomenon is somewhat supported by the results from the simulation scenario 2 and when missingness mechanism is independent of disease status. In that scenario, we observe that the difference in the estimates of β_2 for the parametric and the estimated-score approach increases with overall percentage of missing data.

Only the complete-case analysis shows a statistically significant effect of more than three live born children on the risk. The estimate of the effect of more than three live born children is different for the parametric and the estimated-score approach. An intuitive explanation of this discrepancy

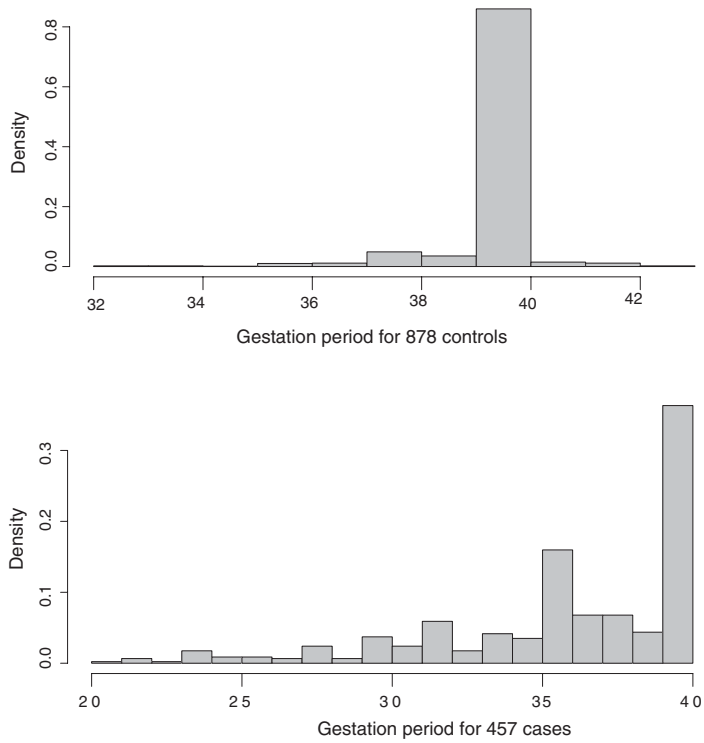

FIGURE 1: Histogram of the observed gestation period measured in weeks in the matched case–control data on low-birth-weight study.

could be the following. First, a simple logistic regression of nonmissing value indicator (R) on S, Y, and Z reveals that $Z = 4$ is strongly associated with R. Second, the simulation scenario 2 revealed that when the missingness mechanism is strongly associated with Z and the model for X is incorrect, the parametric method could produce highly biased estimate for β_1, yielding a big difference in the estimates of β_1 for the parametric and the estimated-score approach. Thus, in this example, possibly a wrong model for the partially missing covariate and strong dependence of R on Z cause a difference in the effect of more than three live born children.

7. DISCUSSION

We propose a nonparametric method for handling missing covariate data in a matched case–control study when the missingness mechanism does not depend on the disease status. Although, due to the matching and the constraint on the number of cases on each strata, the actual test for testing this assumption regarding the missingness mechanism is difficult to construct, a logistic regression for R using S, Y, and Z as explanatory variables, may provide some indications of possible association between R and Y.

For case-independent missing-at-random data, the proposed method produces consistent solution for any type of distribution of the partially missing covariate, and the method is very efficient compared to the complete-case analysis. For instance, the results of the simulation scenario 1 indicates that the gain in efficiency in using the proposed method could be as much as 43.9% and 52% for the estimation of β_1 and β_2, respectively, compared to the complete-case analysis. Indeed the gain in efficiency depends primarily on the percentage of missing data.

The estimated-score method produces biased estimate of β_2 when the missingness mechanism depends on disease status. For instance, the simulation scenario 2 indicates that when the odds of missing data among controls is 2.72 times the odds of missing data among cases, the bias in the estimate of β_2 is about 8–9% in the estimated-score method. Another limitation of our method is that $\boldsymbol{V} = (\boldsymbol{S}, \boldsymbol{Z})$ must be categorical. If the number of categories based on $\boldsymbol{V}$ is large for a data set with limited sample size, the method may not work properly. More clearly, in the estimated-score method, we partition the control subjects without missing data based on $\boldsymbol{V}$, and large number of categories of $\boldsymbol{V}$ may result in some groups without any observation, or very few observations. In that situations, we either ignore some of the components of $\boldsymbol{V}$ or merge some of the categories, and in either cases we may lose some important information which may result in biased estimators. Finally, in this paper we have not dealt with more than one missing covariate which is a part of our future research plan.

We believe that the proposed method could be useful for handling missing covariate data in case-cohort or nested case–control studies. The computer code for the simulation and data analysis is available from the author upon request.

ACKNOWLEDGMENTS

This research was partially supported by NSF grant SES-0961618. The author thanks two reviewers, an associate editor, and the editor for their constructive comments.

BIBLIOGRAPHY

N. E. Breslow & N. E. Day (1980), "*Statistical Methods in Cancer Research*, Volume 1, *The Analysis of Case–Control Studies*," International Agency for Research on Cancer, Lyon, France.

R. J. Carroll & M. P. Wand (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society, Series B*, 53, 573–585.

M. Huberman & B. Langholz (1999). Application of the missing-indicator method in matched case–control studies with incomplete data. *American Journal of Epidemiology*, 150, 1340–1345.

D. W. Hosmer & S. Lemeshow (2000), "*Applied Logistic Regression*," John Wiley & Sons, New York.

R. J. Karunamuni & T. Alberts (2005). A generalized reflection method of boundary correction in kernel density estimation. *The Canadian Journal of Statistics*, 33, 497–509.

X. Li, X. Song & R. H. Gary (2004). Comparison of the missing-indicator method and conditional logistic regression in 1:m matched case–control studies with missing exposure values. *American Journal of Epidemiology*, 159, 603–610.

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

R. J. A. Little & D. B. Rubin (2002), "*Statistical Analysis With Missing Data*," John Wiley & Sons, New York.

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

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

M. S. Pepe & T. R. Fleming (1991). A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association*, 86, 108–113.

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

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

S. Sinha & S. Wang (2009). A new semiparametric procedure for matched case–control studies with missing covariates. *Journal of Nonparametric Statistics*, 21, 889–905.

S. Sinha & T. Maiti (2008). Analysis of matched case–control data in presence of nonignorable missing exposure. *Biometrics*, 64, 106–114.

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

S. Sinha, B. Mukherjee & M. Ghosh (2004). Bayesian semiparametric modeling for matched case–control studies with multiple disease states. *Biometrics*, 60, 41–49.

A. W. van der Vaart (2007), "*Asymptotic Statistics*," Cambridge University Press, New York.

APPENDIX

Expressions for $\partial S_\beta^{\mathrm{em}}/\partial \boldsymbol{\beta}$

Partition the estimated-score function as $S_\beta^{\mathrm{em}} = ({S^{\mathrm{em}}}^{\mathrm{T}}_{\beta_1}, S_{\beta_2}^{\mathrm{em}})^{\mathrm{T}}$.

$$\frac{\partial S_{\beta_1}^{\mathrm{em}}}{\partial \boldsymbol{\beta}_1} = -\sum_{i=1}^{n}\sum_{j=1}^{M+1}\left(\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^{\mathrm{T}}\hat{D}_{ij} - \boldsymbol{Z}_{ij}\hat{D}_{ij}\sum_{k=1}^{M+1}\boldsymbol{Z}_{ik}^{\mathrm{T}}\hat{D}_{ik}\right)$$

$$\frac{\partial S_{\beta_1}^{\mathrm{em}}}{\partial \beta_2} = -\sum_{i=1}^{n}\sum_{j=1}^{M+1}\left[\frac{R_{ij}O_{ij}X_{ij} + (1-R_{ij})(\partial\hat{Q}_{ij}/\partial\beta_2)}{\sum_{k=1}^{M+1}\{R_{ik}O_{ik} + (1-R_{ik})\hat{Q}_{ik}\}} - \hat{D}_{ij}\frac{\sum_{k=1}^{M+1}\{R_{ik}O_{ik}X_{ik} + (1-R_{ik})(\partial\hat{Q}_{ik}/\partial\beta_2)\}}{\sum_{k=1}^{M+1}\{R_{ik}O_{ik} + (1-R_{ik})\hat{Q}_{ik}\}}\right]\boldsymbol{Z}_{ij}$$

$$\frac{\partial S_{\beta_2}^{\mathrm{em}}}{\partial \beta_2} = \sum_{i=1}^{n}\sum_{j=1}^{M+1}(Y_{ij}-\hat{D}_{ij})(1-R_{ij})\frac{\partial^2\log(\hat{Q}_{ij})}{\partial\beta_2^2} - \sum_{i=1}^{n}\sum_{j=1}^{M+1}\left\{R_{ij}X_{ij} + (1-R_{ij})\frac{\partial\log(\hat{Q}_{ij})}{\partial\beta_2}\right\} \times\left[\frac{R_{ij}O_{ij}X_{ij} + (1-R_{ij})(\partial\hat{Q}_{ij}/\partial\beta_2)}{\sum_{k=1}^{M+1}\{R_{ik}O_{ik} + (1-R_{ik})\hat{Q}_{ik}\}} - \hat{D}_{ij}\frac{\sum_{k=1}^{M+1}\{R_{ik}O_{ik}X_{ik} + (1-R_{ik})(\partial\hat{Q}_{ik}/\partial\beta_2)\}}{\sum_{k=1}^{M+1}\{R_{ik}O_{ik} + (1-R_{ik})\hat{Q}_{ik}\}}\right]$$

Regularity Conditions

(A1) $\pi(\boldsymbol{V}) > 0$ almost surely in $\boldsymbol{V}$.

(A2) $Q(\boldsymbol{V}) = E\{\mathrm{pr}(Y=1|\boldsymbol{V}, X)/\mathrm{pr}(Y=0|\boldsymbol{V}, X)\}$ is bounded away from zero uniformly in $\boldsymbol{V}$ on any compact set and in $\boldsymbol{\beta}$ in an open neighbourhood of $\boldsymbol{\beta}_0$, the true value of $\boldsymbol{\beta}$.

(A3) The information matrix $\mathcal{I}$ is positive definite.

(A4) $\partial^2(S_\beta^{\mathrm{em}})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}$ can be bounded by an integrable function of $(Y, R, XR, \boldsymbol{Z})$ in an open neighbourhood of $\boldsymbol{\beta}_0$.

Proof of Consistency

The asymptotic unbiasedness of the estimated-score function follows from

$$n^{-1}S_\beta^{\mathrm{em}} \approx n^{-1}S_\beta^{\mathrm{m}},$$

and the unbiasedness of S_β^{m} is followed from the fact that $E(Y_{ij}|\{R_{ik}, X_{ik}, \boldsymbol{V}_{ik}\}_{k=1}^{M+1}) = D_{ij}$. The rest of the consistency proof easily follows from Theorem 1 of Pepe & Fleming (1991).

In order to derive the asymptotic distribution of the estimator we need the following Lemma. Let $n_{yr}(\boldsymbol{v})$ be the number of observations with $Y = y$, $R = r$, and $\boldsymbol{V} = \boldsymbol{v}$, for $y = 0, 1$ and $r = 0, 1$, and $n_{+r}(\boldsymbol{v}) \equiv n_{0r}(\boldsymbol{v}) + n_{1r}(\boldsymbol{v})$.

Lemma 1. *Uniformly for all $\boldsymbol{v}$ in any given compact set,*

$$\frac{\partial \log\{\hat{Q}(\boldsymbol{v})\}}{\partial \boldsymbol{\beta}} - \frac{\partial \log\{Q(\boldsymbol{v})\}}{\partial \boldsymbol{\beta}}$$
$$= \frac{1}{n_{01}(\boldsymbol{v})Q(\boldsymbol{v})} \sum_{k=1}^{n} \sum_{l=1}^{M+1} (1 - Y_{kl}) R_{kl} I(\boldsymbol{V}_{kl} = \boldsymbol{v}) \left\{ \frac{\partial O_{kl}}{\partial \boldsymbol{\beta}} - \frac{O_{kl}}{Q(\boldsymbol{v})} \frac{\partial Q(\boldsymbol{v})}{\partial \boldsymbol{\beta}} \right\} + o_p(1/\sqrt{n})$$

Proof. Let $P(\cdot|\boldsymbol{v}, Y = 0)$ be the distribution of X given $Y = 0$ and $\boldsymbol{V} = \boldsymbol{v}$. Note that $Q = \int O \, \mathrm{d}P$ and $(\partial Q/\partial \boldsymbol{\beta}) = \int (\partial O/\partial \boldsymbol{\beta}) \, \mathrm{d}P$ are both linear functionals of P and both are Hadamard differentiable. Now, under assumption (A2), the result follows from to the application of the functional delta Theorem 20.8 and its chain rule of van der Vaart (2007, p. 298). ■

Influence Function Representation of $\hat{\boldsymbol{\beta}}$

We write

$$n^{-1/2} S_{\beta}^{\mathrm{em}} = n^{-1/2} S_{\beta}^{\mathrm{m}} + T_2 + T_3 + T_4,$$

where

$$T_2 \equiv n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{M+1} (D_{ij} - \hat{D}_{ij}) \left\{ R_{ij} \frac{\partial}{\partial \boldsymbol{\beta}} \log(O_{ij}) + (1 - R_{ij}) \frac{\partial}{\partial \boldsymbol{\beta}} \log(Q_{ij}) \right\} = o_p(1),$$

$$T_3 \equiv -n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{M+1} (1 - R_{ij}) \left\{ \frac{\partial \log(\hat{Q}_{ij})}{\partial \boldsymbol{\beta}} - \frac{\partial \log(Q_{ij})}{\partial \boldsymbol{\beta}} \right\} \hat{D}_{ij},$$

and

$$T_4 \equiv n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{M+1} Y_{ij} (1 - R_{ij}) \left\{ \frac{\partial}{\partial \boldsymbol{\beta}} \log(\hat{Q}_{ij}) - \frac{\partial}{\partial \boldsymbol{\beta}} \log(Q_{ij}) \right\}.$$

Now, using Lemma 1 and then rearrangement of the summations, and then using the strong law of large number we can write

$$T_3 = -n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{M+1} (1 - R_{ij}) \frac{1}{n_{01}(\boldsymbol{V}_{ij}) Q_{ij}}$$
$$\times \sum_{k=1}^{n} \sum_{l=1}^{M+1} (1 - Y_{kl}) R_{kl} I(\boldsymbol{V}_{kl} = \boldsymbol{V}_{ij}) \left\{ \frac{\partial O_{kl}}{\partial \boldsymbol{\beta}} - \frac{O_{kl}}{Q_{ij}} \frac{\partial Q_{ij}}{\partial \boldsymbol{\beta}} \right\} D_{ij} + o_p(1)$$
$$= -n^{-1/2} \sum_{k=1}^{n} \sum_{l=1}^{M+1} (1 - Y_{kl}) R_{kl} \frac{n_{+0}(\boldsymbol{V}_{kl})}{n_{01}(\boldsymbol{V}_{kl})} \left\{ \frac{(\partial O_{kl}/\partial \boldsymbol{\beta})}{n_{+0}(\boldsymbol{V}_{kl})} \sum_{i=1}^{n} \sum_{j=1}^{M+1} \frac{(1 - R_{ij}) D_{ij}}{Q_{ij}} I(\boldsymbol{V}_{ij} = \boldsymbol{V}_{kl}) \right.$$
$$\left. - O_{kl} \sum_{i=1}^{n} \sum_{j=1}^{M+1} \frac{(1 - R_{ij}) D_{ij}}{Q_{ij}^2} \frac{\partial Q_{kl}}{\partial \boldsymbol{\beta}} I(\boldsymbol{V}_{ij} = \boldsymbol{V}_{kl}) \right\} + o_p(1)$$

$$= -\rho_1 n^{-1/2} \sum_{k=1}^{n} \sum_{l=1}^{M+1} (1 - Y_{kl}) R_{kl} \left\{ \frac{\partial O_{kl}}{\partial \boldsymbol{\beta}} E\left(\frac{D}{Q} \middle| \boldsymbol{V} = \boldsymbol{V}_{kl}, R = 0 \right) \right.$$
$$\left. - O_{kl} E\left(\frac{D}{Q^2} \frac{\partial Q}{\partial \boldsymbol{\beta}} \middle| \boldsymbol{V} = \boldsymbol{V}_{kl}, R = 0 \right) \right\} + o_p(1).$$

Following the similar approach we can write

$$T_4 = n^{-1/2} \sum_{k=1}^{n} \sum_{l=1}^{M+1} (1 - Y_{kl}) R_{kl} \frac{n_{10}(\boldsymbol{V}_{kl})}{n_{01}(\boldsymbol{V}_{kl})} \left\{ \frac{(\partial O_{kl}/\partial \boldsymbol{\beta})}{n_{+0}(\boldsymbol{V}_{kl})} \sum_{i=1}^{n} \sum_{j=1}^{M+1} \frac{(1 - R_{ij}) Y_{ij}}{Q_{ij}} I(\boldsymbol{V}_{ij} = \boldsymbol{V}_{kl}) \right.$$
$$\left. - O_{kl} \sum_{i=1}^{n} \sum_{j=1}^{M+1} \frac{(1 - R_{ij}) Y_{ij}}{Q_{ij}^2} \frac{\partial Q_{kl}}{\partial \boldsymbol{\beta}} I(\boldsymbol{V}_{ij} = \boldsymbol{V}_{kl}) \right\} + o_p(1)$$
$$= \rho_2 n^{-1/2} \sum_{k=1}^{n} \sum_{l=1}^{M+1} (1 - Y_{kl}) R_{kl} \left\{ \frac{\partial O_{kl}}{\partial \boldsymbol{\beta}} E\left(\frac{1}{Q} \middle| \boldsymbol{V} = \boldsymbol{V}_{kl}, Y = 1, R = 0 \right) \right.$$
$$\left. - O_{kl} E\left(\frac{1}{Q^2} \frac{\partial Q}{\partial \boldsymbol{\beta}} \middle| \boldsymbol{V} = \boldsymbol{V}_{kl}, Y = 1, R = 0 \right) \right\} + o_p(1).$$

After adding the four terms and ignoring the terms of order $o_p(1)$ we obtain the expression of $\mathcal{Q}_{ni}$. Now a Taylor's series expansion of S_β^{em} yields $n^{-1/2} S_\beta^{\text{em}} = \{-n^{-1} \partial S_{\beta^*}^{\text{em}} / \partial \boldsymbol{\beta}^*\} n^{1/2} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, where $\boldsymbol{\beta}^*$ is between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$. Since $\{-n^{-1} \partial S_{\beta^*}^{\text{em}} / \partial \boldsymbol{\beta}^*\}$ converges to $\mathcal{I}$, an invertible matrix, in probability, and $n^{-1/2} S_\beta^{\text{em}}$ is approximately a sum of i.i.d. random variables, the asymptotic normality of $\hat{\boldsymbol{\beta}}$ now follows by the application of the central limit theorem.

Received 19 December 2009
Accepted 24 June 2010