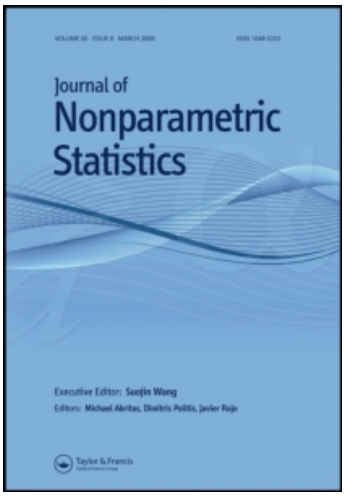

Journal of Nonparametric Statistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713645758

A new semiparametric procedure for matched case-control studies with missing covariates

Samiran Sinha [a]; Suojin Wang [a]
[a] Department of Statistics, Texas A&M University, College Station, TX, USA

Online Publication Date: 01 October 2009

To cite this Article Sinha, Samiran and Wang, Suojin(2009)'A new semiparametric procedure for matched case-control studies with missing covariates',Journal of Nonparametric Statistics,21:7,889 — 905
To link to this Article: DOI: 10.1080/10485250903019523
URL: http://dx.doi.org/10.1080/10485250903019523

Journal of Nonparametric Statistics
Vol. 21, No. 7, October 2009, 889–905

Taylor & Francis
Taylor & Francis Group

A new semiparametric procedure for matched case-control studies with missing covariates

Samiran Sinha* and Suojin Wang

Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA

(*Received 11 July 2008; final version received 2 May 2009*)

In this paper, we propose an easy-to-use semiparametric method for analysing matched case-control data when one of the covariates of interest is partially missing. Missing covariate information in matched case-control studies may create bias and reduce efficiency of the parameter estimates. In order to cope with this situation we consider a robust approach which is comprised of estimating some functionals of the distribution of the partially missing covariate using a kernel regression technique in a conditional likelihood framework. The large sample theory of the proposed estimator is investigated and the asymptotic normality is obtained. A simulation study is conducted to assess the performance of the proposed method in terms of robustness and efficiency. The proposed method is also applied to a real dataset which motivates this work.

Keywords: conditional logistic regression; estimating equation; kernel regression; matched case-control study; missing at random

1. Introduction and background

One of the most common types of epidemiological studies is a case-control design which involves two independent random samples from diseased and non-diseased groups of a target population, and then exposure or covariate information is ascertained from the sampled individuals. In order to control the confounding effect of nuisance factors (factors other than the covariates of interest) matched case-control study is often used where matched strata are defined by the combination of the levels of the nuisance factors [1]. Due to the retrospective nature of the data collection, often covariate(s) contains missing values. One motivating example is the Los Angeles Endometrial Cancer Study which was conducted to investigate the effects of gall-bladder disease, use of estrogen, and hypertension on endometrial cancer which is one of the common cancers among US women who have gone through menopause and are 45 years old or older. Two important covariates, duration of estrogen use and obesity, were missing for approximately 5% and 16% of the subjects. The concern is then how to handle subjects with missing covariates in a matched case-control design. The complete ignorance of the subjects with missing covariates may lead to inefficient and/or biased estimates of the relative risk parameters.

*Corresponding author. Email: sinha@stat.tamu.edu

ISSN 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/10485250903019523
http://www.informaworld.com

There are many articles on missing covariate data in a cohort study, such as Ibrahim [2], Wang *et al.* [3], Lipsitz *et al.* [4], just to name a few. In a Bayesian context Mukherjee *et al.* [5] modelled the exposure variable of an unmatched case-control study without missing data using the Dirichlet process prior. However, missing covariate data in matched case-control studies have received less than deserved attention from researchers. There are two important issues that need to be taken into account while analysing matched case-control data with a missing covariate. According to the study design each matched set is sparse as it typically contains only one case (diseased) subject and maybe a couple of control (non-diseased) subjects, and thus there is a lack of enough degrees of freedom to estimate the matched set's specific nuisance parameter. The second issue is that as the sampling is conditional on the disease status and the matching variable(s), the marginal distribution of the covariate is not identifiable from the data unless the marginal disease prevalence is known for each stratum.

There are two basic approaches for handling missing covariate data in matched case-control studies. Lipsitz *et al.* [6] parametrically modelled the missingness probability in terms of the observed quantities. In this approach the subjects with a missing covariate are used only when modelling the missingness probability. Along the same line of thought Rathouz *et al.* [7] proposed a class of semiparametric estimators which efficiently use data from all subjects, and produce more efficient estimators for the relative-risk parameters corresponding to the covariates which have no missing values. In an alternative school of thought, people have modelled the distribution of the missing covariate. The pioneering work in this approach was done by Satten and Kupper [8], Paik and Sacco [9], and Satten and Carroll [10]. They used different types of conditional likelihood functions which require some type of distributional assumptions regarding the missing covariate. Satten and Carroll [10] modelled the distribution of the missing covariate given the observed covariates among the control population when all the covariates take only finite many values. On the other hand, Paik and Sacco [9] modelled the missing covariate given the completely observed covariate and the disease status with a distribution belonging to an exponential family of distributions.

As pointed out by Satten and Carroll [10], Paik and Sacco's method enjoys consistency and may be more robust toward the specified model for the missing covariate. The reason is that their assumed model affects only the subjects with missing exposure, and the contribution of the other subjects to the conditional likelihood remains the same. A good discussion can be found in [11] regarding the optimality of different procedures for estimating parameters in this context. Sinha *et al.* [12] proposed a nonparametric Bayesian procedure to capture unobserved stratum heterogeneity in the distribution of the missing covariate. Importantly, in the inference procedure they also used a likelihood function similar to that proposed in Satten and Carroll's paper that is quite different from the conditional likelihood function in our current approach. All these methods considered missing at random (MAR) data, and modelled the distribution of the missing covariate parametrically. The above mentioned methods produce consistent and asymptotically unbiased parameter estimates as long as the assumed parametric form of the distribution of the missing covariate is correct; a violation of the model assumption may lead to biased estimates of the parameters. There are two types of model violations. First the assumed distributional form for the missing covariate may not be true, and second the assumed form of relationship between the missing covariate and the observed covariate(s) is incorrect. In each case the estimated parameters may be biased; of course the amount of the bias depends on the amount of missing data, the degree of a model violation, and the type of the likelihood function or the estimating equation used for this purpose.

The aim of this paper is to provide a completely robust method of estimation of the relative risk parameters, which is free from any distributional assumption of the missing covariates, and whose point estimation procedure does not require the estimation of the missingness probability function. In order to achieve this goal we use a set of estimating equations which involve several

functionals of the conditional distribution of the missing covariate given the observed covariates and the disease status, and these functionals are estimated via a kernel method. Though the estimated functionals have slower rate of convergence than $\sqrt{n}$, we show that the estimates for the parameters of interest remain to be $\sqrt{n}$ consistent. We also provide a formula for the calculation of the standard error of the estimates. The proposed method yields conditional logistic regression (CLR) estimates of the parameters when there is no missing data. We conduct a simulation study to assess the performance of the proposed method and to compare it with an alternative parametric approach as well as the complete case analysis (CCA) in several finite sample settings. In addition, we apply the proposed method to a real matched case-control study.

The rest of the paper is outlined as follows. In Section 2 we describe the primary disease risk model and notation while the proposed method is discussed in Section 3. Section 4 contains the related asymptotic theory with technical details given in the Appendix. A simulation study is given in Section 5 to evaluate the performance of the proposed method in terms of robustness and variability of the estimates. Some details are also suggested on how to choose the kernel function and the smoothing parameters. The data analysis is illustrated in Section 6. Some concluding remarks are given in Section 7.

2. Model and notation

Suppose we have an 1:M matched case-control study with n strata. The ith stratum contains one case and $M(\geq 1)$ control subjects, and let j be the index for subjects. Let Y_{ij} be the binary disease variable for the jth subject in the ith stratum which takes on value one for cases and zero for controls. Let $\boldsymbol{Z}$ be a $p \times 1$ vector of covariates which is always observed, and X be a scalar covariate which may not be observed for all subjects. Also, let $\boldsymbol{W}$ be a $q \times 1$ vector of covariates which are used as the matching variables, and $\boldsymbol{W}_i$ be the value of $\boldsymbol{W}$ for the ith case subject. For the ith matched set M controls are drawn from $p(X, \boldsymbol{Z}|Y = 0, \boldsymbol{W} = \boldsymbol{W}_i)$, for $i = 1, \ldots, n$. We assume a logistic disease risk model for prospective data

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

where $H(u) = \{1 + \exp(-u)\}^{-1}$ and β_{0i} is the nonparametric effect of the matching variables $\boldsymbol{W}_i$ on the disease probability. The main parameters of interest $\boldsymbol{\beta}_1$ and β_2 can be interpreted as the global log-odds ratio parameter associated with $\boldsymbol{Z}$ and X, respectively. Note that the log-odds ratio parameters do not vary across the strata, which is a standard assumption in this context. Our main goal here is to estimate $\boldsymbol{\beta}_1$ and β_2 when X is partially missing.

When X is observed for all sampled subjects, a semiparametric inference can be carried out by using the following CLR [13].

$$L_{\text{CLR}} = \prod_{i=1}^{n} \frac{\sum_{j=1}^{M+1} Y_{ij} A_{ij}^*}{\sum_{k=1}^{M+1} A_{ik}^*}, \tag{1}$$

where $A_{ij}^* = P(Y_{ij} = 1|X_{ij},\ \boldsymbol{Z}_{ij},\ \boldsymbol{W}_i)/P(Y_{ij} = 0|X_{ij},\ \boldsymbol{Z}_{ij},\ \boldsymbol{W}_i) = \exp(\beta_{0i} + \boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1 + X_{ij}\beta_2)$. Note that, in L_{CLR}, without loss of generality one can replace A_{ij}^* by $A_{ij} = \exp(\boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1 + X_{ij}\beta_2)$. The conditional maximum likelihood estimate of the parameters are obtained by maximising L_{CLR} with respect to $\boldsymbol{\beta}_1$ and β_2. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \beta_2)^{\mathrm{T}}$. Then the conditional score equations obtained from (1) are $\sum_{i=1}^{n}\sum_{j=1}^{M+1} v_{ij} = 0$, where $v_{ij} = (Y_{ij} - A_{ij}/\sum_{k=1}^{M+1} A_{ik})\partial \log(A_{ij})/\partial\boldsymbol{\theta}$. It should be noted that $E(v_{ij}) = 0$ as the conditional expectation of Y_{ij} given the covariate information, matching variable and $\sum_{k=1}^{M+1} Y_{ik} = 1$ in the ith matched set is $A_{ij}/\sum_{k=1}^{M+1} A_{ik}$.

3. Method of estimation

In order to handle the missing covariate we introduce non-missing value indicator δ_{ij}:

$$\delta_{ij} = \begin{cases} 1 & \text{if } X_{ij} \text{is observed} \\ 0 & \text{otherwise.} \end{cases}$$

We assume that data are MAR [14] in the sense that given the observed data the missingness mechanism neither depends on the variable of interest X nor on the parameters of interest $\boldsymbol{\theta}$. Therefore, $\text{pr}(\delta_{ij} = 1|X_{ij}, \boldsymbol{Z}_{ij}, \boldsymbol{W}_{ij}, Y_{ij}) = \text{pr}(\delta_{ij} = 1|\boldsymbol{Z}_{ij}, \boldsymbol{W}_{ij}, Y_{ij})$. Importantly, in a matched case-control study sampling is done from $p(\boldsymbol{Z}_{ij}, X_{ij}|Y_{ij} = y, \boldsymbol{W}_i)$, and following the notation of Hosmer and Lemeshow [15] and Paik and Sacco [9] define $p(\boldsymbol{Z}_{ij}, X_{ij}|Y_{ij} = y, \boldsymbol{W}_i) = h(y; \boldsymbol{Z}_{ij}, X_{ij}, \boldsymbol{W}_i)$. Thus, without loss of generality if $j = 1$ stands for the case and the rest for the controls, the unconditional likelihood is

$$\prod_{i=1}^{n} \left\{ h(1; \boldsymbol{Z}_{i1}, X_{i1}, \boldsymbol{W}_i) \prod_{j=2}^{M+1} h(0; \boldsymbol{Z}_{ij}, X_{ij}, \boldsymbol{W}_i) \right\}.$$

The conditional likelihood function is

$$\prod_{i=1}^{n} \frac{h(1; \boldsymbol{Z}_{i1}, X_{i1}, \boldsymbol{W}_i) \prod_{j=2}^{M+1} h(0; \boldsymbol{Z}_{ij}, X_{ij}, \boldsymbol{W}_i)}{\sum_{k=1}^{M+1} h(1; \boldsymbol{Z}_{ik}, X_{ik}, \boldsymbol{W}_i) \prod_{j=1, j\neq k}^{M+1} h(0; \boldsymbol{Z}_{ij}, X_{ij}, \boldsymbol{W}_i)}. \tag{2}$$

According to Breslow [16], (2) is the conditional likelihood of the event where $(\boldsymbol{Z}_{i1}, X_{i1})$ is for the case and $(\boldsymbol{Z}_{i2}, X_{i2}), \ldots, (\boldsymbol{Z}_{iM+1}, X_{iM+1})$ are for the controls in stratum i, given the set of observed $(M + 1)$ exposures.

For MAR data, the unconditional likelihood is

$$\prod_{i=1}^{n} \left[h^{\delta_{i1}}(1; \boldsymbol{Z}_{i1}, X_{i1}, \boldsymbol{W}_i) g^{1-\delta_{i1}}(1; \boldsymbol{Z}_{i1}, \boldsymbol{W}_i) \prod_{j=2}^{M+1} \left\{ h^{\delta_{ij}}(0; \boldsymbol{Z}_{ij}, X_{ij}, \boldsymbol{W}_i) g^{1-\delta_{ij}}(0; \boldsymbol{Z}_{ij}, \boldsymbol{W}_i) \right\} \right],$$

where $g(y; \boldsymbol{Z}_{ij}, \boldsymbol{W}_i) = p(\boldsymbol{Z}_{ij}|Y_{ij} = y, \boldsymbol{W}_i)$, and the conditional likelihood is

$$L_{\text{c}} = \prod_{i=1}^{n} \frac{\begin{array}{c} h^{\delta_{i1}}(1; \boldsymbol{Z}_{i1}, X_{i1}, \boldsymbol{W}_i) g^{1-\delta_{i1}}(1; \boldsymbol{Z}_{i1}, \boldsymbol{W}_i) \prod_{j=2}^{M+1} \\ \times \{ h^{\delta_{ij}}(0; \boldsymbol{Z}_{ij}, X_{ij}, \boldsymbol{W}_i) g^{1-\delta_{ij}}(0; \boldsymbol{Z}_{ij}, \boldsymbol{W}_i) \} \end{array}}{\begin{array}{c} \sum_{k=1}^{M+1} h^{\delta_{ik}}(1; \boldsymbol{Z}_{ik}, X_{ik}, \boldsymbol{W}_i) g^{1-\delta_{ik}}(1; \boldsymbol{Z}_{ik}, \boldsymbol{W}_i) \prod_{j=1, j\neq k}^{M+1} \\ \times \{ h^{\delta_{ij}}(0; \boldsymbol{Z}_{ik}, X_{ik}, \boldsymbol{W}_i) g^{1-\delta_{ij}}(0; \boldsymbol{Z}_{ij}, \boldsymbol{W}_i) \} \end{array}}.$$

In each term above if we divide the numerator and the denominator by $\prod_{j=1}^{M+1} \{ h^{\delta_{ij}}(0; \boldsymbol{Z}_{ij}, X_{ij}, \boldsymbol{W}_i) g^{1-\delta_{ij}}(0; \boldsymbol{Z}_{ij}, \boldsymbol{W}_i) \}$, we obtain

$$L_{\text{c}} = \prod_{i=1}^{n} \frac{\delta_{i1} A^*_{i1} + (1 - \delta_{i1}) a^*_{i1}}{\sum_{k=1}^{M+1} \{ \delta_{ik} A^*_{ik} + (1 - \delta_{ik}) a^*_{ik} \}}, \tag{3}$$

where a^*_{ij} is the conditional expectation of A^*_{ij} with respect to $f(X|\boldsymbol{Z}, \boldsymbol{W}, Y = 0)$, the density of X given $\boldsymbol{Z}$, $\boldsymbol{W}$, and $Y = 0$, i.e.,

$$\begin{aligned} a^*_{ij} &= \frac{\text{pr}(Y_{ij} = 1|\boldsymbol{Z}_{ij}, \boldsymbol{W}_i)}{\text{pr}(Y_{ij} = 0|\boldsymbol{Z}_{ij}, \boldsymbol{W}_i)} = \int \frac{\text{pr}(Y_{ij} = 1|\boldsymbol{Z}_{ij}, X_{ij}, \boldsymbol{W}_i)}{\text{pr}(Y_{ij} = 0|\boldsymbol{Z}_{ij}, X_{ij}, \boldsymbol{W}_i)} f(X_{ij}|\boldsymbol{Z}_{ij}, \boldsymbol{W}_i, Y_{ij} = 0) \text{d}X_{ij} \\ &= \int A^*_{ij} f(X_{ij}|\boldsymbol{Z}_{ij}, \boldsymbol{W}_i, Y_{ij} = 0) \text{d}X_{ij}, \end{aligned}$$

and A_{ij}^*'s were defined earlier. Note that the second equation above is readily shown following Satten and Kupper [8]. After dividing the numerator and the denominator of (3) by $e^{\beta_{0i}}$, we obtain

$$L_c = \prod_{i=1}^{n} \frac{\sum_{j=1}^{M+1} Y_{ij}\{\delta_{ij}A_{ij} + (1-\delta_{ij})a_{ij}\}}{\sum_{k=1}^{M+1}\{\delta_{ik}A_{ik} + (1-\delta_{ik})a_{ik}\}}, \tag{4}$$

where $a_{ij} = \exp(\boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1)\int \exp(\beta_2 x) f(x|\boldsymbol{Z}_{ij}, \boldsymbol{W}_i, Y=0)\mathrm{d}x$. Note that neither A_{ij} nor a_{ij} is a function of Y_{ij}. Likelihood (4) allows any jth subject to be the case in stratum i. The score function for $\boldsymbol{\theta}$ derived from (4) is then

$$U_n(\boldsymbol{\theta}, a, b) = n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{M+1} S_{ij}(\boldsymbol{\theta}, a, b)\left\{\delta_{ij}\frac{\partial \log(A_{ij})}{\partial \boldsymbol{\theta}} + (1-\delta_{ij})\frac{b_{ij}}{a_{ij}}\right\}, \tag{5}$$

where $S_{ij}(\boldsymbol{\theta}, a, b) = Y_{ij} - \psi_{ij}(\boldsymbol{\theta}, a, b)$, with

$$\psi_{ij}(\boldsymbol{\theta}, a, b) = \frac{\delta_{ij}A_{ij} + (1-\delta_{ij})a_{ij}}{\sum_{k=1}^{M+1}\{\delta_{ik}A_{ik} + (1-\delta_{ik})a_{ik}\}} \tag{6}$$

and $b_{ij} = \partial a_{ij}/\partial\boldsymbol{\theta}$.

Note that both a_{ij} and b_{ij} are the expectation of A_{ij} and $\partial A_{ij}/\partial\boldsymbol{\theta}$, respectively, with respect to the conditional density of X given $\boldsymbol{Z}$, $\boldsymbol{W}$, and $Y=0$. We propose to estimate them by the kernel method. Let $V = (\boldsymbol{Z}^{\mathrm{T}}, \boldsymbol{W}^{\mathrm{T}})^{\mathrm{T}}$, and d be the number of continuous components of V. Let K be an rth-order kernel function of $d(<r)$ variables as in [3]; for example with $d=1$, K will satisfy $\int K(t)\mathrm{d}t = 1$, $\int t^s K(t)\mathrm{d}t = 0$ for $s = 1, \ldots, (r-1)$, and $\int t^r K(t)\mathrm{d}t \neq 0$. Also, assume that $K_h(\cdot) = (1/h)K(\cdot/h)$, where h is the smoothing parameter. For each $\boldsymbol{\theta}$ define Nadaraya–Watson estimator of a_{ij} and b_{ij} as

$$\widehat{a}_{ij} = \widehat{a}_{ij}(\boldsymbol{\theta}) = \exp(\boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1)\frac{\sum_{k=1}^{n}\sum_{l=1}^{M+1}(1-Y_{kl})\delta_{kl}e^{X_{kl}\beta_2}K_h(V_{ij}-V_{kl})}{\sum_{k=1}^{n}\sum_{l=1}^{M+1}(1-Y_{kl})\delta_{kl}K_h(V_{ij}-V_{kl})},$$

$$\widehat{b}_{ij} = \widehat{b}_{ij}(\boldsymbol{\theta}) = \begin{pmatrix}\widehat{E}\left(\dfrac{\partial A_{ij}}{\partial\boldsymbol{\beta}_1}\right)\\ \widehat{E}\left(\dfrac{\partial A_{ij}}{\partial\beta_2}\right)\end{pmatrix} = \begin{pmatrix}\boldsymbol{Z}_{ij}\widehat{a}_{ij}(\boldsymbol{\theta})\\ \exp(\boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1)\dfrac{\sum_{k=1}^{n}\sum_{l=1}^{M+1}(1-Y_{kl})\delta_{kl}X_{kl}e^{X_{kl}\beta_2}K_h(V_{ij}-V_{kl})}{\sum_{k=1}^{n}\sum_{l=1}^{M+1}(1-Y_{kl})\delta_{kl}K_h(V_{ij}-V_{kl})}\end{pmatrix}.$$

We propose to estimate the parameters by solving the following estimated score equations

$$U_n(\boldsymbol{\theta}, \widehat{a}, \widehat{b}) = n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{M+1} S_{ij}(\theta, \widehat{a}, \widehat{b})\left\{\delta_{ij}\frac{\partial \log(A_{ij})}{\partial\boldsymbol{\theta}} + (1-\delta_{ij})\frac{\widehat{b}_{ij}}{\widehat{a}_{ij}}\right\} = 0. \tag{7}$$

For given h, the parameter estimates are obtained as follows:

Step 0 Initialize $\boldsymbol{\theta} = \boldsymbol{\theta}_{(0)}$.

Step 1 Calculate $U_n\{\boldsymbol{\theta}_{(0)}, \widehat{a}(\boldsymbol{\theta}_{(0)}), \widehat{b}(\boldsymbol{\theta}_{(0)})\}$ and $G_n\{\boldsymbol{\theta}_{(0)}, \widehat{a}(\boldsymbol{\theta}_{(0)}), \widehat{b}(\boldsymbol{\theta}_{(0)})\} = -n^{-1/2}[\partial U_n\{\boldsymbol{\theta}, \widehat{a}(\boldsymbol{\theta}), \widehat{b}(\boldsymbol{\theta})\}/\partial\boldsymbol{\theta}^{\mathrm{T}}]_{\boldsymbol{\theta}=\boldsymbol{\theta}_{(0)}}$, where

$$
\begin{aligned}
G_n(\boldsymbol{\theta}, \widehat{a}, \widehat{b}) = -\frac{1}{n}\sum_{i=1}^{n}\Bigg[\sum_{j=1}^{M+1} S_{ij}(\boldsymbol{\theta}, \widehat{a}, \widehat{b})\left\{\delta_{ij}\frac{\partial^2 \log A_{ij}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} + (1-\delta_{ij})\frac{\partial^2 \log \widehat{a}_{ij}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\right\}\\
- \frac{\sum_{j=1}^{M+1}\{\delta_{ij}(\partial A_{ij}/\partial\boldsymbol{\theta})\partial \log(A_{ij})/\partial\boldsymbol{\theta}^{\mathrm{T}} + (1-\delta_{ij})(\partial\widehat{a}_{ij}/\partial\boldsymbol{\theta})\partial \log(\widehat{a}_{ij})/\partial\boldsymbol{\theta}^{\mathrm{T}}}{\sum_{j=1}^{M+1}\{\delta_{ij}A_{ij} + (1-\delta_{ij})\widehat{a}_{ij}\}}\\
+ \frac{\sum_{j=1}^{M+1}\{\delta_{ij}\partial A_{ij}/\partial\boldsymbol{\theta} + (1-\delta_{ij})\partial\widehat{a}_{ij}/\partial\boldsymbol{\theta}\}}{\sum_{j=1}^{M+1}\{\delta_{ij}A_{ij} + (1-\delta_{ij})\widehat{a}_{ij}\}}\\
\times \frac{\sum_{j=1}^{M+1}\{\delta_{ij}\partial A_{ij}/\partial\boldsymbol{\theta}^{\mathrm{T}} + (1-\delta_{ij})\partial\widehat{a}_{ij}/\partial\boldsymbol{\theta}^{\mathrm{T}}\}}{\sum_{j=1}^{M+1}\{\delta_{ij}A_{ij} + (1-\delta_{ij})\widehat{a}_{ij}\}}\Bigg]. \qquad (8)
\end{aligned}
$$

Step 2 Obtain $\boldsymbol{\theta}_{(1)} = \boldsymbol{\theta}_{(0)} + n^{-1/2}G_n^{-1}\{\boldsymbol{\theta}_{(0)}, \widehat{a}(\boldsymbol{\theta}_{(0)}), \widehat{b}(\boldsymbol{\theta}_{(0)})\}U_n\{\boldsymbol{\theta}_{(0)}, \widehat{a}(\boldsymbol{\theta}_{(0)}), \widehat{b}(\boldsymbol{\theta}_{(0)})\}$.

Step 3 Repeat the iterative procedure above until $\boldsymbol{\theta}_{(k)}$ converges to $\widehat{\boldsymbol{\theta}}$.

4. Asymptotic properties of the proposed estimator

We assume that the following regularity conditions hold in an open neighbourhood containing the true parameter value $\boldsymbol{\theta}_0$ of $\boldsymbol{\theta}$. Define the selection probabilities $\pi_0(V) = \mathrm{pr}(\delta = 1|V, Y = 0)$ and $\pi_1(V) = \mathrm{pr}(\delta = 1|V, Y = 1)$, where $V = (\boldsymbol{Z}^{\mathrm{T}}, \boldsymbol{W}^{\mathrm{T}})^{\mathrm{T}}$. In our consideration of asymptotics, M is viewed as fixed while n increases to ∞.

(A1) The selection probabilities have r partial and bounded derivatives with respect to the continuous components of V.
(A2) For every v in the domain of V, $\pi_0(v) \geq k_0$ and $\pi_1(v) \geq k_1$ for some constants $k_0 > 0$ and $k_1 > 0$.
(A3) The probability density functions $f_{[V|\delta=1,Y=0]}(\cdot)$ and $f_{[V|\delta=0]}(\cdot)$ have r continuous and bounded derivatives with respect to the continuous components of V.
(A4) The density functions $f_{[V|\delta=1,Y=0]}(\cdot)$ and $f_{[V|\delta=0]}(\cdot)$ have the same support, and $f_{[V|\delta=0]}(\cdot)/f_{[V|\delta=1,Y=0]}(\cdot)$ is bounded over the support.
(A5) $\sum_{j=1}^{M+1} S_{1j}\{\delta_{1j}\partial \log(A_{1j})/\partial\boldsymbol{\theta} + (1-\delta_{1j})b_{1j}/a_{1j}\}$ has a finite second moment, and

$$
G(\boldsymbol{\theta}) = E\left[-\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\sum_{j=1}^{M+1} S_{1j}(\theta, a, b)\left\{\delta_{1j}\frac{\partial \log(A_{1j})}{\partial\boldsymbol{\theta}} + (1-\delta_{1j})\frac{b_{1j}}{a_{1j}}\right\}\right] \qquad (9)
$$

is positive definite.

Note that Assumption (A2) is crucial for the estimation of the standard errors of the parameter estimates. The reason is that if there are few observed X around the conditioning variables even when the sample size is large, then the kernel method will not work well to produce good estimates of conditional quantities such as a or b. We now present the following main results.

THEOREM 1 *Under Assumptions* (A1)–(A5), *if the bandwidth h satisfies that* $nh^{2d} \to \infty$ *and* $nh^{2r} \to 0$, *as* $n \to \infty$, *then* $\widehat{\boldsymbol{\theta}}$ *obtained by solving* (7) *is a consistent estimator of* $\boldsymbol{\theta}$, *and* $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ *is asymptotically normally distributed with mean* $\mathbf{0}$ *and a variance–covariance matrix* Σ *defined at the end of the Appendix.*

The proof of Theorem 1 is given in the Appendix.

THEOREM 2 *The variance–covariance matrix* Σ *in Theorem* 1 *can be consistently estimated by*

$$G_n^{-1}(\widehat{\boldsymbol{\theta}}, \widehat{a}, \widehat{b}) \left(\frac{1}{n} \sum_{i=1}^{n} \widehat{J}_i \widehat{J}_i^{\mathrm{T}} \right) G_n^{-\mathrm{T}}(\widehat{\boldsymbol{\theta}}, \widehat{a}, \widehat{b}), \tag{10}$$

where

$$\widehat{J}_i = \sum_{j=1}^{M+1} \left(S_{ij}(\widehat{\boldsymbol{\theta}}, \widehat{a}, \widehat{b}) \left\{ \delta_{ij} \frac{\partial \log(A_{ij})}{\partial \boldsymbol{\theta}} + (1 - \delta_{ij}) \frac{\widehat{b}_{ij}}{\widehat{a}_{ij}} \right\} + (1 - Y_{ij}) \delta_{ij} \left[\frac{1}{\widehat{a}_{ij}} \left(\frac{\partial A_{ij}}{\partial \boldsymbol{\theta}} - \frac{\widehat{b}_{ij} A_{ij}}{\widehat{a}_{ij}} \right) \right. \right.$$
$$\left. \left. \times \{ \widehat{S}_{1ij} \widehat{c}_{10}^{*}(V_{ij}) + \widehat{S}_{0ij} \widehat{c}_{00}^{*}(V_{ij}) \} - \left(\frac{\widehat{b}_{ij}}{\widehat{a}_{ij}} \widehat{D}_{ij}^{*} - \widehat{T}_{ij}^{*} \right) (A_{ij} - \widehat{a}_{ij}) \{ \widehat{c}_{10}^{*}(V_{ij}) + \widehat{c}_{00}^{*}(V_{ij}) \} \right] \right),$$

and $\widehat{c}_{\Delta 0}^{*}(V_{ij})$ *is the kernel estimate of* $c_{\Delta 0}^{*}(V) = \mathrm{pr}(\delta = 0 | V, Y = \Delta)/\mathrm{pr}(\delta = 1 | V, Y = 0)$ *for* $\Delta = 0, 1$. *For the sake of simplicity, with a little abuse of notation here* $\widehat{a}$ *and* $\widehat{b}$ *denote the estimates of a and b evaluated at* $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. *Furthermore,*

$$\widehat{D}_{ij}^{*} = \frac{\sum_{k=1}^{n} \sum_{l=1}^{M+1} D_k(\widehat{\boldsymbol{\theta}}, \widehat{a}, \widehat{b}) K_h(V_{kl} - V_{ij})}{\sum_{k=1}^{n} \sum_{l=1}^{M+1} K_h(V_{kl} - V_{ij})}$$

and

$$\widehat{T}_{ij}^{*} = \frac{\sum_{k=1}^{n} \sum_{l=1}^{M+1} T_k(\widehat{\boldsymbol{\theta}}, \widehat{a}, \widehat{b}) K_h(V_{kl} - V_{ij})}{\sum_{k=1}^{n} \sum_{l=1}^{M+1} K_h(V_{kl} - V_{ij})},$$

with $D_k(\boldsymbol{\theta}, a, b) = 1/\sum_{j=1}^{M+1} \{\delta_{kj} A_{kj} + (1 - \delta_{kj}) a_{kj}\}$ *and* $T_k(\boldsymbol{\theta}, a, b) = \sum_{j=1}^{M+1} \{\delta_{kj} \partial A_{kj}/\partial \boldsymbol{\theta} + (1 - \delta_{kj}) \partial a_{kj}/\partial \boldsymbol{\theta}\}/[\sum_{j=1}^{M+1} \{\delta_{kj} A_{kj} + (1 - \delta_{kj}) a_{kj}\}]^2$, *for* $k = 1, \ldots, n$.

In addition

$$\widehat{S}_{\Delta ij} = \Delta - \frac{\sum_{k=1}^{n} \sum_{l=1}^{M+1} \psi_{kl}(\widehat{\boldsymbol{\theta}}, \widehat{a}, \widehat{b}) K_h(V_{kl} - V_{ij})}{\sum_{k=1}^{n} \sum_{l=1}^{M+1} K_h(V_{kl} - V_{ij})}.$$

The proof of Theorem 2 follows easily from the proof of Theorem 1.

5. A simulation study

In this section we evaluate the performance of the proposed method via a simulation study. First, we generated a cohort of size $N = 16{,}000$ by simulating W, Z, X and then Y. Then from the cohort we constructed 1:1 matched case-control data with $n = 150$ and $n = 300$ strata using W as the matching variable. In addition, the case of $n = 100$ has been added after a comment by a referee. In order to simulate realistic data we closely followed the Los Angeles Endometrial Cancer data, and the parameter values were close to the corresponding estimates which were obtained while the dataset was analysed by a parametric approach. We simulated the data according to the following steps.

- Simulate W from $N(0.53, 0.24^2)$.
- Simulate Z from Gamma(0.23, 1.20), so that $E(Z) = 0.276$.

- Simulate X from
  Case I $N(0.1445 + 0.5430W + 0.1180Z, 0.4^2)$.
  Case II $I(W < 0.5)U[0.9, 3.0] + I(W \geq 0.5)N(W, 0.2^2)$.
  Case III $(1/2)N(1.2Z, 0.4^2) + (1/2)N(-1.2 + Z^2, 0.5^2)$.
- Simulate the disease variable Y from the Bernoulli distribution with success probability $H(-3.5 + 1.1W + Z + 0.5X)$. Therefore, the true values of β_1 and β_2 are 1 and 0.5. For each scenarios MAR data were generated by generating the non-missing value indicator from the Bernoulli distribution with success probability $H(1 + Z + 0.35W)$ and $H(Z + 0.35W)$. They resulted in about 20% and 40% missing data, respectively.

Under each scenario we generated $R = 1000$ datasets, and each dataset was analysed by the following four methods. First we assumed that there is no missing data, and the fully observed data were analysed by the CLR method. This approach is hereafter referred to as full data analysis (FDA). Second we considered only the matched sets which have no missing covariate, and this reduced dataset was analysed via the CLR method. This approach is hereafter referred to as CCA. For the sake of comparison we analysed each missing dataset by a parametric approach. Here we applied Paik and Sacco's method, which is easy to apply and reduces to the CLR approach if there is no missing exposure variable. Since X is a continuous random variable in all three cases, in the parametric approach we model X as $N(\gamma_0 + \gamma_1 Z + \gamma_2 W + \gamma_3 Y, \sigma^2)$. This method will be abbreviated as PARA. Figure 1 shows the distribution of X for cases I, II, and III respectively. The density plots of the last two distributions clearly indicate that the normal model assumption is violated in Cases II and III. Lastly we analysed 1000 datasets using the proposed nonparametric method which is abbreviated as NONP.

In the NONP method $V = (Z, W)$ has two continuous components so that $d = 2$. The bandwidth h of the kernel should satisfy $nh^{2d} \to \infty$ and $nh^{2r} \to 0$, which implies $r > d$. We choose $h = O(n^{-1/p})$. Then r must be the smallest even integer $\geq (p - d)$. Following the idea of Wang *et al.* [3] we take $p = 5$, and consequently we choose a kernel of order $r = 4$, $K(u) = 2\exp(-u^2/2)/\sqrt{2\pi} - \exp(-u^2/4)/\sqrt{4\pi}$ [17]. Note that $K(u) \geq 0$ for $u \in [-2.05, 2.05]$ and $-0.0249 < K(u) \leq 0$ for $u \in [-2.05, 2.05]^C$. It is easily seen that the above symmetric kernel satisfies $\int K(u)\mathrm{d}u = 1$, $\int u^j K(u)\mathrm{d}u = 0$ for $j = 1, 2, 3$, and $\int u^4 K(u)\mathrm{d}u = -6$. One should also note that the above kernel may give biased estimate of the quantities of interest if the point where the kernel is evaluated is near the boundary points of the observed dataset, and that bias may induce bias in the original relative risk parameters for small sample size. Therefore, following Hart and Wehrly [18] we construct a boundary kernel based on $K(u)$. Let $W_{\min}$ and $W_{\max}$ be the observed minimum and maximum values of W. For a given bandwidth h, for a given value of W and any other point W^* we define the boundary kernel as

$$K_b\left(\frac{W - W^*}{h}\right) = \begin{cases} \dfrac{2}{\sqrt{2\pi}h}\exp\left\{-\dfrac{(W - W^*)^2}{2h^2}\right\} & \\ \quad -\dfrac{1}{\sqrt{4\pi}h}\exp\left\{-\dfrac{(W - W^*)^2}{4h^2}\right\}, & \text{if } W \in (W_{\min} + h, W_{\max} - h) \\ (\alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3)\left[\dfrac{2}{\sqrt{2\pi}h}\exp\{-(5t)^2/2\}\right. & \\ \qquad \left. -\dfrac{1}{\sqrt{4\pi}h}\exp\{-(5t)^2/4\}\right], & \text{otherwise,} \end{cases}$$

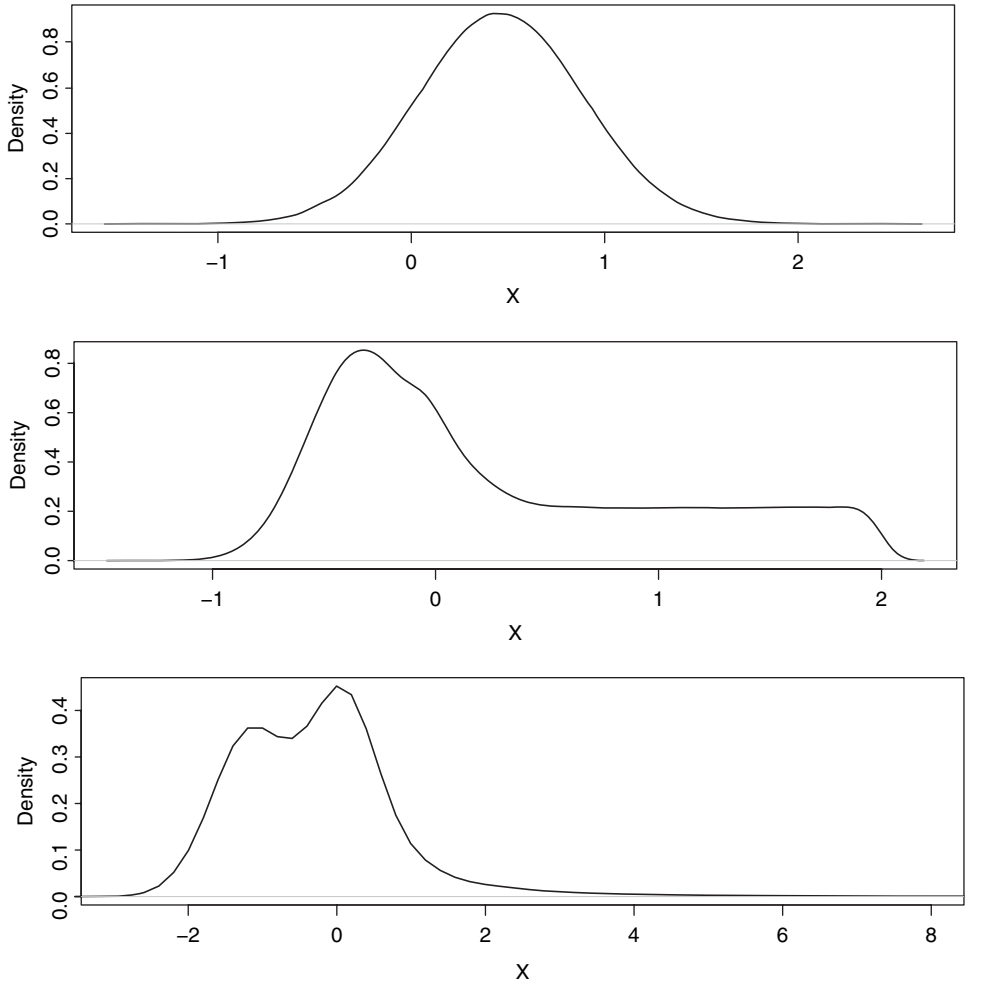

Figure 1. Marginal density plot of X in Cases I, II, and III, respectively.

where $t = (W - W^*)/(5h)$ and $\alpha_0, \ldots, \alpha_3$ are constants such that

$$\int_{-q_2}^{q_1} (\alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3) 5 \left[\frac{2}{\sqrt{2\pi}} \exp\{-(5t)^2/2\} - \frac{1}{\sqrt{4\pi}} \exp\{-(5t)^2/4\} \right] \mathrm{d}t = 1$$

$$\int_{-q_2}^{q_1} t^j (\alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3) 5 \left[\frac{2}{\sqrt{2\pi}} \exp\{-(5t)^2/2\} - \frac{1}{\sqrt{4\pi}} \exp\{-(5t)^2/4\} \right] \mathrm{d}t = 0$$

$$\text{for } j = 1, 2, 3,$$

and q_1 and q_2 are defined as $q_1 = \min\{1, (W - W_{\min})/(5h)\}$ and $q_2 = \min\{1, (W_{\max} - W)/(5h)\}$. The constants are obtained by solving

$$\boldsymbol{A}\boldsymbol{\alpha} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \text{ where } \boldsymbol{A} = \begin{pmatrix} e_0 & e_1 & e_2 & e_3 \\ e_1 & e_2 & e_3 & e_4 \\ e_2 & e_3 & e_4 & e_5 \\ e_3 & e_4 & e_5 & e_6 \end{pmatrix} \text{ and } \boldsymbol{\alpha}^{\mathrm{T}} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3).$$

Here $e_j = (2/5^j)I_j(-5q_2, 5q_1) - (\sqrt{2}/5)^j I_j(-5q_2/\sqrt{2}, 5q_1/\sqrt{2})$ for $j = 0, 1 \ldots, 6$, and $I_j(x_1, x_2) = \int_{x_1}^{x_2} u^j \exp(-u^2/2)\mathrm{d}u/\sqrt{2\pi}$ for $x_1 \leq x_2$. We adopt the following iterative way to compute the integrals $I_0(x_1, x_2) = \Phi(x_2) - \Phi(x_1)$, $I_1(x_1, x_2) = \{\exp(-x_1^2/2) - \exp(-x_2^2/2)\}/\sqrt{2\pi}$, and $I_j(x_1, x_2) = \{x_1^{j-1}\exp(-x_1^2/2) - x_2^{j-1}\exp(-x_2^2/2)\}/\sqrt{2\pi} + (j-1)I_{j-2}(x_1, x_2)$. The constants vary from observation to observation, so for each boundary point we need to solve these constants.

We consider the product of the two boundary kernels $K_b\{(z - Z)/h_1\}K_b\{(w - W)/h_2\}$. In order to properly define the kernel based estimates of a_{ij} and b_{ij}'s which are outside the observed boundary of Z among the controls we define $Z_{\min}$ and $Z_{\max}$ as the minimum and maximum values of Z in the observed data irrespective of the case and control status. By taking product of two boundary kernels we avoid the selection of bandwidth matrix which is a fairly difficult task. Let h_1 and h_2 be the smoothing parameters for Z and W respectively. We choose $h_1 = c_1\widehat{\sigma}_Z n^{-1/5}$ and $h_2 = c_2\widehat{\sigma}_W n^{-1/5}$, where $\widehat{\sigma}_Z$ and $\widehat{\sigma}_W$ are the sample standard deviation of Z and W in the overall dataset. These bandwidths satisfy the bandwidth criteria. The constants c_1 and c_2 play a crucial role in determining the proper bandwidth for the distribution of X among the controls. Their data-driven choice is as follows. These constants are varied over a grid of values from 0.05 to 2.5 with increment 0.05. The joint grid point which maximises the log of the estimated likelihood function L_{ec} is taken as our (c_1, c_2), where

$$L_{ec} = \prod_{i=1}^{n} \frac{\sum_{j=1}^{M+1} Y_{ij}\{\delta_{ij}A_{ij} + (1 - \delta_{ij})\widehat{a}_{ij}\}}{\sum_{k=1}^{M+1}\{\delta_{ik}A_{ik} + (1 - \delta_{ik})\widehat{a}_{ik}\}}.$$

Alternatively one may use a cross-validation method to choose a reasonable bandwidth.

Table 1. Results of the simulation study based on 1000 replications of 1:1 matched case-control data with the number of strata $n = 100$.

		$\beta_1 = 1.0$				$\beta_2 = 0.50$			
Case	Method	Mean	True SE	SE	CP	Mean	True SE	SE	CP
I		1.048	0.333	0.324	0.957	0.539	0.443	0.417	0.937
I(a)	CCA	1.064	0.372	0.385	0.955	0.565	0.557	0.527	0.955
	PARA	1.055	0.336	0.357	0.956	0.538	0.493	0.488	0.958
	NONP	1.057	0.340	0.317	0.932	0.539	0.539	0.497	0.932
I(b)	CCA	1.126	0.509	0.504	0.956	0.583	0.773	0.724	0.970
	PARA	1.057	0.341	0.360	0.953	0.550	0.550	0.570	0.967
	NONP	1.056	0.347	0.315	0.931	0.572	0.636	0.616	0.927
II	FDA	1.068	0.331	0.329	0.962	0.534	0.387	0.371	0.953
II(a)	CCA	1.091	0.380	0.394	0.966	0.534	0.485	0.475	0.970
	PARA	1.067	0.323	0.361	0.956	0.450	0.386	0.403	0.963
	NONP	1.066	0.328	0.319	0.950	0.532	0.453	0.443	0.944
II(b)	CCA	1.134	0.486	0.509	0.958	0.553	0.697	0.673	0.969
	PARA	1.071	0.327	0.363	0.966	0.391	0.420	0.436	0.952
	NONP	1.085	0.333	0.326	0.958	0.549	0.507	0.481	0.929
III	FDA	1.083	0.552	0.555	0.967	0.535	0.260	0.257	0.963
III(a)	CCA	1.115	0.677	0.685	0.973	0.553	0.336	0.325	0.960
	PARA	1.099	0.595	0.644	0.969	0.486	0.285	0.296	0.949
	NONP	1.108	0.578	0.554	0.949	0.535	0.313	0.299	0.949
III(b)	CCA	1.162	0.880	0.921	0.980	0.572	0.455	0.446	0.958
	PARA	1.109	0.651	0.703	0.967	0.436	0.321	0.332	0.929
	NONP	1.111	0.593	0.589	0.953	0.545	0.360	0.394	0.942

Note: 'Mean' and 'True SE' represent the average value of the estimates and the square root of the variance of the estimates across the simulated datasets. 'SE' represents the average value of the standard error.

The results for $n = 100$, 150, and 300 strata are presented in Tables 1–3, respectively, where missingness mechanism (a) and (b) indicate selection probability $H(1 + Z_{ij} + 0.35W_{ij})$ and $H(Z_{ij} + 0.35W_{ij})$, respectively, so that 'Case I(a)' means Case I with missingness mechanism (a), etc. The tables show the empirical means of the parameter estimates, the empirical standard errors of the estimates ('True SE') across the simulations, and the average values of the standard errors. For the NONP method the standard error was calculated based on formula (10). First of all the CCA produces estimates with large variances. The reason for this is that it ignores many subjects with missing covariate data. Consequently, the variance increases with the proportion of missing data. Both the parametric and nonparametric approaches work well in Case I where X is generated from a normal distribution. For Cases II and III and when $n = 150$ and $n = 300$ the parametric method produces a significantly biased estimate of β_2 compared to the proposed nonparametric method. The advantage of the proposed method over the parametric method is not clear for the smaller sample size of $n = 100$. Overall the variances of the estimates decrease with the sample size, and by increasing the sample size one can remove the small sample bias in the parameter estimates. For $n = 150$ the parameter estimates due to CCA method did not converge for approximately 7% of the datasets. The results presented in the tables are based on the converged estimates. This problem is reduced when the sample size n increases. Neither the parametric approach nor the nonparametric approach faced any convergence problem. For the PARA method the standard error of the estimates was calculated via the Jackknife method. Note that in the calculation of the standard error using formula (10) we estimated $\text{pr}(\delta = 1|V, Y = y)$ for $y = 0, 1$ by a kernel method.

Table 2. Results of the simulation study based on 1000 replications of 1:1 matched case-control data with the number of strata $n = 150$.

		$\beta_1 = 1.0$				$\beta_2 = 0.50$			
Case	Method	Mean	True SE	SE	CP	Mean	True SE	SE	CP
I		1.038	0.263	0.261	0.967	0.520	0.342	0.335	0.953
I(a)	CCA	1.063	0.313	0.311	0.965	0.529	0.430	0.422	0.950
	PARA	1.040	0.263	0.276	0.958	0.519	0.380	0.385	0.955
	NONP	1.039	0.267	0.254	0.946	0.520	0.418	0.385	0.948
I(b)	CCA	1.102	0.401	0.397	0.975	0.549	0.594	0.573	0.962
	PARA	1.040	0.263	0.278	0.962	0.528	0.448	0.449	0.962
	NONP	1.044	0.270	0.256	0.949	0.532	0.514	0.458	0.948
II	FDA	1.052	0.267	0.264	0.949	0.517	0.297	0.298	0.957
II(a)	CCA	1.036	0.271	0.306	0.973	0.551	0.389	0.381	0.964
	PARA	1.051	0.264	0.281	0.958	0.455	0.302	0.318	0.959
	NONP	1.060	0.271	0.259	0.946	0.533	0.345	0.340	0.932
II(b)	CCA	1.090	0.399	0.395	0.959	0.553	0.526	0.526	0.966
	PARA	1.051	0.261	0.281	0.961	0.408	0.327	0.344	0.956
	NONP	1.066	0.269	0.263	0.949	0.545	0.405	0.393	0.928
III	FDA	1.047	0.449	0.44	0.963	0.521	0.204	0.206	0.959
III(a)	CCA	1.081	0.549	0.540	0.959	0.525	0.255	0.258	0.958
	PARA	1.057	0.476	0.489	0.963	0.471	0.217	0.231	0.963
	NONP	1.065	0.461	0.437	0.950	0.529	0.230	0.232	0.960
III(b)	CCA	1.111	0.716	0.714	0.971	0.539	0.349	0.348	0.973
	PARA	1.078	0.521	0.537	0.962	0.413	0.242	0.259	0.939
	NONP	1.077	0.490	0.457	0.938	0.535	0.270	0.287	0.952

Note: 'Mean' and 'True SE' represent the average value of the estimates and the square root of the variance of the estimates across the simulated datasets. 'SE' represents the average value of the standard error.

Table 3. Results of the simulation study based on 1000 replications of 1:1 matched case-control data with the number of strata $n = 300$.

		$\beta_1 = 1.0$				$\beta_2 = 0.50$			
Case	Method	Mean	True SE	SE	CP	Mean	True SE	SE	CP
I		1.022	0.187	0.181	0.945	0.509	0.233	0.235	0.953
I(a)	CCA	1.031	0.215	0.213	0.953	0.515	0.286	0.293	0.958
	PARA	1.022	0.187	0.187	0.948	0.508	0.256	0.265	0.959
	NONP	1.022	0.188	0.179	0.939	0.519	0.261	0.268	0.949
I(b)	CCA	1.053	0.273	0.265	0.968	0.525	0.406	0.392	0.951
	PARA	1.022	0.187	0.187	0.951	0.521	0.307	0.307	0.945
	NONP	1.023	0.189	0.180	0.942	0.521	0.315	0.329	0.947
II	FDA	1.031	0.189	0.183	0.946	0.506	0.197	0.207	0.965
II(a)	CCA	1.038	0.220	0.215	0.953	0.516	0.262	0.263	0.958
	PARA	1.031	0.189	0.189	0.944	0.440	0.205	0.217	0.951
	NONP	1.036	0.192	0.183	0.938	0.525	0.226	0.236	0.967
II(b)	CCA	1.056	0.276	0.268	0.946	0.537	0.363	0.356	0.961
	PARA	1.031	0.189	0.188	0.947	0.384	0.252	0.244	0.932
	NONP	1.038	0.193	0.184	0.941	0.527	0.265	0.281	0.946
III	FDA	1.015	0.313	0.303	0.944	0.512	0.145	0.143	0.949
III(a)	CCA	1.029	0.384	0.369	0.957	0.515	0.184	0.179	0.955
	PARA	1.007	0.329	0.328	0.950	0.467	0.160	0.159	0.935
	NONP	1.035	0.316	0.308	0.942	0.520	0.162	0.165	0.954
III(b)	CCA	1.084	0.501	0.484	0.961	0.524	0.248	0.240	0.947
	PARA	1.032	0.360	0.361	0.952	0.407	0.182	0.179	0.903
	NONP	1.033	0.327	0.330	0.944	0.520	0.184	0.209	0.958

Note: 'Mean' and 'True SE' represent the average value of the estimates and the square root of the variance of the estimates across the simulated datasets. 'SE' represents the average value of the standard error.

6. A real data example

We now return to the example briefly discussed in the introduction. Endometrial cancer is one common type of malignancy that occurs in the inner membrane of the uterus. It is found that this cancer is associated with a high level of estrogen use. Use of estrogen or other hormone therapy is widely prevalent among post-menopausal women. In order to study the effect of several risk factors on this cancer a study was conducted among post-menopausal women in an affluent retirement community of Los Angeles. The data comprises of $n = 63$ strata and each stratum consists of 1 case and $M = 4$ controls [1]. The 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. Obesity is treated as the partially missing exposure variable (X) and presence of gall-bladder disease is considered as a binary completely observed covariate (Z). For the purpose of illustration, in this article we assume that the data are MAR as is done by other researchers, although there is no clear way to validate this assumption.

The reason of considering gall-bladder disease as one of the risk factors is that estrogen, which is one of the well known regulating factors of endometrial cancer, raises the level of cholesterol in bile. Bile, a substance produced by the liver and stored in the gall-bladder, promotes the growth of gallstones and other gall-bladder diseases. Also, since fat tissues can increase a women's estrogen levels, overweight or presence of obesity is considered as a potential risk factor. In the original study obesity was determined, as is customarily defined, according to whether the body mass index (BMI) value exceeds the normal value of 30 or not [19]. In the analysis age is transformed into [0, 1] scale and then used as a matching variable W. The disease risk model of our interest

Table 4. The results of the analyses of the Los Angeles Endometrial Cancer Data by three different methods.

Method		Presence of gall bladder disease	Obesity
CCA	Estimate	1.279	0.440
	Std err	0.394	0.376
PARA	Estimate	1.270	0.597
	Std err	0.365	0.294
NONP	Estimate	1.304	0.636
	Std err	0.366	0.321

Note: Std err represents the standard error of the parameter estimates.

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 dataset is analysed by three methods. First we analysed it with the CLR method ignoring the subjects with missing X, ignoring the stratum with missing X for the case subject, and ignoring the stratum where all controls have missing X. With a little abuse of terminology we will call this method 'CCA'. Second we analysed the data using the parametric approach. For the parametric method the distribution of the missing covariate was modeled as $\text{logit}\{\text{pr}(X = 1|Z, W, Y)\} = \gamma_0 + \gamma_1 Z + \gamma_2 W + \gamma_3 Y$. The estimate (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). These estimates are used to obtain the estimates of β_1 and β_2. Lastly, we analysed the data using the proposed nonparametric approach.

Note that in this data example $V = (Z, W)$ and the number of continuous components in V is $d = 1$, and that is the matching variable age (W). We used the same kernel function as in the simulation study. Two bandwidths h_0^* and h_1^* were needed for W corresponding to $Z = 0$ and $Z = 1$. We chose $h_0^* = c_0^* n_{\text{CO}}^{-1/5} \widehat{\sigma}_{W0}$ and $h_1^* = c_1^* n_{\text{CO}}^{-1/5} \widehat{\sigma}_{W1}$, where $\widehat{\sigma}_{W0}$ and $\widehat{\sigma}_{W1}$ are the observed standard deviation of W among $Z = 0$ and $Z = 1$, respectively, and $n_{\text{CO}} = 252$ is the total number of controls. The point (c_0^*, c_1^*) that maximises the estimated likelihood function L_{ec} over a grid of values (0.25, 0.5, 1.0, 1.5, 2.0, 2.25, 2.5, 2.75, 3) for each of c_0^* and c_1^* was used in our analysis. For this data example we found $(c_0^*, c_1^*) = (2, 2.5)$.

The estimate and standard error of β_1 and β_2 are presented in Table 4. It is seen that the presence of gall-bladder disease appears to increase the risk of having endometrial cancer. From the CCA and NONP analyses we do not find a significant association between the cancer and obesity while the PARA method shows a statistically significant association. As expected the CCA method produces largest standard error for $\hat{\beta}_1$ and $\hat{\beta}_2$. The standard errors in the parametric approach are somewhat smaller than that of the nonparametric approach. Since the model assumption used in the PARA method is difficult to verify, we would be more comfortable with the conclusion drawn from the NONP analysis.

7. Discussion

This paper provides a flexible method for analysing matched case-control data with missing covariate data. The proposed method relies on a simple kernel technique, and the estimates are easy to compute. A formula for standard error calculations has been derived, and some guidelines on how to choose the smoothing parameters and the appropriate kernel function have also been given. The use of a boundary kernel has been considered for a given kernel. The proposed method is not only robust against model misspecification for X, but also reduces to the standard conditional

likelihood analysis when there is no missing data. One limitation of the proposed method is that this paper only deals with a single missing covariate. Indeed theoretically the method can be extended to handle multiple missing covariates. On the other hand, in practice due to a slower rate of convergence the kernel method might not be a very viable option to handle multiple missing covariates. Like many nonparametric approaches the proposed method works well for large sample sizes as is suggested in our limited simulation study. The computer code is available from the authors upon request.

The proposed method has a broader applicability beyond this particular problem. It is possible to generalise the new method to $K:M$ matched case-control studies where $K \geq 1$ and $M \geq 1$ can vary over different strata. Furthermore, the method can be extended in principle to derive robust estimates of the parameters in presence of covariate measurement error in matched case-control studies. It may also be extended to handle missing covariate data in the Cox's proportional hazard (CPH) model as the partial likelihood derived from the CPH model and the conditional likelihood of matched studies have some similarities.

As commented by a reviewer, one may address the missing covariate problem in matched case-control studies by using partially linear models to model the association between the disease and the matching variables which can alleviate the issue of high-dimensional parameters. With the above modelling one can analyse the matched data as if the data are collected prospectively, and without using the conditional likelihood. However, the details of this method will be considered in a future work.

Acknowledgements

We thank the editor and two referees for their constructive comments and suggestions that helped us greatly improve this paper.

References

[1] N. Breslow and N.E. Day, *Statistical Methods in Cancer Research*, International Agency for Research in Cancer, Lyon, 1980.
[2] J.G. Ibrahim, *Incomplete data in generalized linear models*, J. Am. Stat. Assoc. 85 (1990), pp. 765–769.
[3] C.Y. Wang, S. Wang, L.-P. Zhao, and S.-T. Ou, *Weighted semiparametric estimation in regression analysis with missing covariate data*, J. Am. Stat. Assoc. 92 (1997), pp. 512–525.
[4] S.R. Lipsitz, J.G. Ibrahim, and L.P. Zhao, *A weighted estimating equation for missing covariate data with properties similar to maximum likelihood*, J. Am. Stat. Assoc. 94 (1999), pp. 1147–1160.
[5] B. Mukherjee, L. Zhang, M. Ghosh, and S. Sinha, *Semiparametric Bayesian analysis of case-control data under conditional gene-environment independence*, Biometrics 63 (2007), pp. 834–844.
[6] S.R. Lipsitz, M. Parzen, and M. Ewell, *Inference using conditional logistic regression with missing covariates*, Biometrics 54 (1998), pp. 295–303.
[7] P. Rathouz, G. Satten, and R.J. Carroll, *Semiparametric inference in matched case-control studies with missing covariate data*, Biometrika 89 (2002), pp. 905–916.
[8] G.A. Satten and L.L. Kupper, *Inferences about exposure-disease associations using probability-of-exposure information*, J. Am. Stat. Assoc. 88 (1993), pp. 200–208.
[9] M.C. Paik and R.L. Sacco, *Matched case-control data analyses with missing covariates*, J. R. Stat. Soc. C 49 (2000), pp. 145–156.
[10] G. Satten and R.J. Carroll, *Conditional and unconditional categorical regression models with missing covariates*, Biometrics 56 (2000), pp. 384–388.
[11] P.J. Rathouz, *Likelihood methods for missing covariate data in highly stratified studies*, J. R. Stat. Soc. B, 65 (2003), pp. 711–723.
[12] S. Sinha, B. Mukherjee, M. Ghosh, B.K. Mallick, and R.J. Carroll, *Semiparametric Bayesian analysis of matched case-control studies with missing exposure*, J. Am. Stat. Assoc. 100 (2005), pp. 591–601.
[13] N. Breslow, *Covariance analysis of censored survival data*, Biometrics 30 (1974), pp. 89–99.
[14] R.J.A. Little and D.B. Rubin, *Statistical Analysis with Missing Data*, John Wiley & Sons, New York, 1987.
[15] D.W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, John Wiley & Sons, 2000, New York.
[16] N. Breslow, *Statistics in epidemiology: the case-control study*, J. Am. Stat. Assoc. 91 (1996), pp. 14–28.
[17] B. Abdous, *Computationally efficient classes of higher-order kernels*, Can. J. Stat. 23 (1995), pp. 21–27.

[18] J. Hart and T. Wehrly, *Kernel regression when boundary region is large, with an application to testing the adequacy of polynomial models*, J. Am.. Stat. Assoc. 87 (1992), pp. 1018–1024.
[19] T.M. Mack, M.C. Pike, B.E. Henderson, R.I. Pfeffer, V.R. Gerkins, B.S. Arthur, and S.E. Brown, *Estrogen and endometrial cancer in a retirement community*, New Engl. J. Med. 294 (1976), pp. 1262–1267.
[20] S. Wang and C.Y. Wang, *A note on kernel assisted estimators in missing covariate regression*, Stat. Probab. Lett. 55 (2001), pp. 439–449.

Appendix

Proof of Theorem 1

In this proof the main and lengthy task is to approximate $U_n(\boldsymbol{\theta}, \widehat{a}, \widehat{b})$ by a sum of independent and identically distributed (iid) random variables with mean zero, as in Equation (A5).

Define $\eta_n = \{nh^{2r} + (nh^{2d})^{-1}\}^{1/2}$. Let n_{oc} and n_{oco} denote the number of cases and controls with observed X, respectively, whereas n_{mc} and n_{mco} are the corresponding values with unobserved X in a dataset of n strata. Therefore, $n = n_{\text{oc}} + n_{\text{mc}}$ and $n \times M = n_{\text{oco}} + n_{\text{mco}}$. Let

$$U_n(\boldsymbol{\theta}, \widehat{a}, \widehat{b}) - U_n(\boldsymbol{\theta}, a, b) = U_{1n} - U_{2n},$$

where

$$U_{1n} = n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{M+1} S_{ij}(\boldsymbol{\theta}, a, b)(1 - \delta_{ij}) \left(\frac{\widehat{b}_{ij}}{\widehat{a}_{ij}} - \frac{b_{ij}}{a_{ij}} \right),$$

and

$$U_{2n} = n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{M+1} \left[\frac{\delta_{ij} A_{ij} + (1 - \delta_{ij})\widehat{a}_{ij}}{\sum_{k=1}^{M+1} \{\delta_{ik} A_{ik} + (1 - \delta_{ik})\widehat{a}_{ik}\}} - \frac{\delta_{ij} A_{ij} + (1 - \delta_{ij}) a_{ij}}{\sum_{k=1}^{M+1} \{\delta_{ik} A_{ik} + (1 - \delta_{ik}) a_{ik}\}} \right] \times \left\{ \delta_{ij} \frac{\partial A_{ij} / \partial \boldsymbol{\theta}}{A_{ij}} + (1 - \delta_{ij}) \frac{\widehat{b}_{ij}}{\widehat{a}_{ij}} \right\}.$$

Define

$$P_{nij} = \frac{1}{n_{\text{oco}} h^d} \sum_{k=1}^{n} \sum_{l=1}^{M+1} (1 - Y_{kl}) \delta_{kl} \exp(\boldsymbol{Z}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_1 + X_{kl} \beta_2) K_h(V_{ij} - V_{kl}),$$

$$Q_{nij} = \frac{1}{n_{\text{oco}} h^d} \sum_{k=1}^{n} \sum_{l=1}^{M+1} (1 - Y_{kl}) \delta_{kl} K_h(V_{ij} - V_{kl}),$$

$$R_{nij} = \frac{1}{n_{\text{oco}} h^d} \begin{bmatrix} \sum_{k=1}^{n} \sum_{l=1}^{M+1} (1 - Y_{kl}) \delta_{kl} \boldsymbol{Z}_{ij} \exp(\boldsymbol{Z}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_1 + X_{kl} \beta_2) K_h(V_{ij} - V_{kl}) \\ \sum_{k=1}^{n} \sum_{l=1}^{M+1} (1 - Y_{kl}) \delta_{kl} X_{kl} \exp(\boldsymbol{Z}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_1 + X_{kl} \beta_2) K_h(V_{ij} - V_{kl}) \end{bmatrix}.$$

Note that $\widehat{a}_{ij} = P_{nij}/Q_{nij}$ and $\widehat{b}_{ij} = R_{nij}/Q_{nij}$, and $Q_{nij} = \widehat{f}_{[V|\delta=1,Y=0]}(V_{ij})$. Using Lemma 1 of Wang and Wang [20, p. 446] we have

$$\widehat{a}_{ij} - a_{ij} = \frac{P_{nij} - a_{ij} Q_{nij}}{f_{[V|\delta=1,Y=0]}(V_{ij})} + O_p(\xi_n) \quad \text{and} \quad \widehat{b}_{ij} - b_{ij} = \frac{R_{nij} - b_{ij} Q_{nij}}{f_{[V|\delta=1,Y=0]}(V_{ij})} + O_p(\xi_n), \tag{A1}$$

where $\xi_n = h^{2r} + (nh^d)^{-1}$. Furthermore, it follows from Lemma 1 of Wang and Wang [20] that $E(\widehat{a}_{ij} - a_{ij}) = O(h^r)$ and $\text{var}(\widehat{a}_{ij} - a_{ij}) = O\{(nh^d)^{-1}\}$, which together imply $(\widehat{a}_{ij} - a_{ij})^2 = O_p(h^{2r}) + O_p\{(nh^d)^{-1}\} = O_p(\xi_n)$. Similarly, we can show that $(\widehat{b}_{ij} - b_{ij})^2 = O_p(\xi_n)$ and thus $(\widehat{a}_{ij} - a_{ij})(\widehat{b}_{ij} - b_{ij}) = O_p(\xi_n)$. Therefore, using the Taylor series expansion we can write

$$\frac{\widehat{b}_{ij}}{\widehat{a}_{ij}} = \frac{b_{ij}\{1 + (\widehat{b}_{ij} - b_{ij})/b_{ij}\}}{a_{ij}\{1 + (\widehat{a}_{ij} - a_{ij})/a_{ij}\}} = \frac{b_{ij}}{a_{ij}} \left\{ 1 + \frac{\widehat{b}_{ij} - b_{ij}}{b_{ij}} - \frac{\widehat{a}_{ij} - a_{ij}}{a_{ij}} \right\} + O_p(\xi_n). \tag{A2}$$

Replacing $\widehat{b}_{ij}/\widehat{a}_{ij}$ by the dominating term on the right-hand side of (A2) in U_{1n} and since $n^{1/2}\xi_n$ has a smaller order than η_n, we obtain

$$U_{1n} = n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{M+1} S_{ij}(\boldsymbol{\theta}, a, b)(1 - \delta_{ij}) \left\{ \frac{\widehat{b}_{ij} - b_{ij}}{a_{ij}} - \left(\frac{b_{ij}}{a_{ij}} \right) \frac{\widehat{a}_{ij} - a_{ij}}{a_{ij}} \right\} + O_p(\eta_n). \tag{A3}$$

Now consider the first term in the summand of U_{2n}:

$$\begin{aligned}
&\frac{\delta_{ij}A_{ij}+(1-\delta_{ij})\widehat{a}_{ij}}{\sum_{k=1}^{M+1}\{\delta_{ik}A_{ik}+(1-\delta_{ik})\widehat{a}_{ik}\}}-\frac{\delta_{ij}A_{ij}+(1-\delta_{ij})a_{ij}}{\sum_{k=1}^{M+1}\{\delta_{ik}A_{ik}+(1-\delta_{ik})a_{ik}\}}\\
&\quad=\frac{\delta_{ij}A_{ij}+(1-\delta_{ij})a_{ij}}{\sum_{k=1}^{M+1}\{\delta_{ik}A_{ik}+(1-\delta_{ik})a_{ik}\}}\left\{1+\frac{(1-\delta_{ij})(\widehat{a}_{ij}-a_{ij})}{\delta_{ij}A_{ij}+(1-\delta_{ij})a_{ij}}\right\}\\
&\qquad\times\left\{1+\frac{\sum_{k=1}^{M+1}(1-\delta_{ik})(\widehat{a}_{ik}-a_{ik})}{\sum_{k=1}^{M+1}\{\delta_{ik}A_{ik}+(1-\delta_{ik})a_{ik}\}}\right\}^{-1}-\frac{\delta_{ij}A_{ij}+(1-\delta_{ij})a_{ij}}{\sum_{k=1}^{M+1}\{\delta_{ik}A_{ik}+(1-\delta_{ik})a_{ik}\}}\\
&\quad=\frac{\delta_{ij}A_{ij}+(1-\delta_{ij})a_{ij}}{\sum_{k=1}^{M+1}\{\delta_{ik}A_{ik}+(1-\delta_{ik})a_{ik}\}}\left[\frac{(1-\delta_{ij})(\widehat{a}_{ij}-a_{ij})}{\delta_{ij}A_{ij}+(1-\delta_{ij})a_{ij}}-\frac{\sum_{k=1}^{M+1}(1-\delta_{ik})(\widehat{a}_{ik}-a_{ik})}{\sum_{k=1}^{M+1}\{\delta_{ik}A_{ik}+(1-\delta_{ik})a_{ik}\}}\right]+O_p(\xi_n).
\end{aligned}$$

Using the derivation above and (A2) we obtain

$$\begin{aligned}
U_{2n}&=n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{M+1}\frac{\delta_{ij}A_{ij}+(1-\delta_{ij})a_{ij}}{\sum_{k=1}^{M+1}\{\delta_{ik}A_{ik}+(1-\delta_{ik})a_{ik}\}}\times\left\{\frac{(1-\delta_{ij})(\widehat{a}_{ij}-a_{ij})}{\delta_{ij}A_{ij}+(1-\delta_{ij})a_{ij}}-\frac{\sum_{k=1}^{M+1}(1-\delta_{ik})(\widehat{a}_{ik}-a_{ik})}{\sum_{k=1}^{M+1}\{\delta_{ik}A_{ik}+(1-\delta_{ik})a_{ik}\}}\right\}\\
&\quad\times\left\{\delta_{ij}\frac{\partial A_{ij}/\partial\boldsymbol{\theta}}{A_{ij}}+(1-\delta_{ij})\frac{b_{ij}}{a_{ij}}\right\}+O_p(\eta_n)\\
&=n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{M+1}\left\{\frac{(1-\delta_{ij})(\widehat{a}_{ij}-a_{ij})}{\sum_{k=1}^{M+1}\{\delta_{ik}A_{ik}+(1-\delta_{ik})a_{ik}\}}-\frac{\{\delta_{ij}A_{ij}+(1-\delta_{ij})a_{ij}\}\sum_{k=1}^{M+1}(1-\delta_{ik})(\widehat{a}_{ik}-a_{ik})}{[\sum_{k=1}^{M+1}\{\delta_{ik}A_{ik}+(1-\delta_{ik})a_{ik}\}]^2}\right\}\\
&\quad\times\left\{\delta_{ij}\frac{\partial A_{ij}/\partial\boldsymbol{\theta}}{A_{ij}}+(1-\delta_{ij})\frac{b_{ij}}{a_{ij}}\right\}+O_p(\eta_n)\\
&=n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{M+1}\frac{(1-\delta_{ij})b_{ij}(\widehat{a}_{ij}-a_{ij})/a_{ij}}{\sum_{k=1}^{M+1}\{\delta_{ik}A_{ik}+(1-\delta_{ik})a_{ik}\}}-n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{M+1}\frac{\{\delta_{ij}\partial A_{ij}/\partial\boldsymbol{\theta}+(1-\delta_{ij})b_{ij}\}\sum_{k=1}^{M+1}(1-\delta_{ik})(\widehat{a}_{ik}-a_{ik})}{\left[\sum_{k=1}^{M+1}\{\delta_{ik}A_{ik}+(1-\delta_{ik})a_{ik}\}\right]^2}+O_p(\eta_n).
\end{aligned}\tag{A4}$$

Furthermore, let $\delta_i=\prod_{j=1}^{M+1}\delta_{ij}$, $V_{ij}^*=\delta_i V_{ij}$, and $X_{ij}^*=\delta_i X_{ij}$. We will use Assumptions (A1), (A3), and (A4), and Lemma 1 of Wang and Wang [20] in the following derivation. Note that we can write $\sum_{i=1}^{n}\sum_{j=1}^{M+1}(1-\delta_{ij})(S_{ij}/a_{ij})(\widehat{b}_{ij}-b_{ij})=\sum_{i=1}^{n}\sum_{j=1}^{M+1}(1-\delta_{ij})\{Y_{ij}+(1-Y_{ij})\}(S_{ij}/a_{ij})(\widehat{b}_{ij}-b_{ij})$. Consider

$$\begin{aligned}
&E\left\{\frac{S_{ij}}{a_{ij}}(\widehat{b}_{ij}-b_{ij})|\delta_{ij}=0,Y_{ij}=\Delta,\ \text{all}\ (\delta,Y,V^*,X^*)\right\}\\
&\quad=\frac{1}{n_{\text{oco}}}\sum_{k=1}^{n}\sum_{l=1}^{M+1}(1-Y_{kl})\delta_{kl}\int\frac{S_{ij}}{a_{ij}}\left(\frac{\partial A_{kl}}{\partial\boldsymbol{\theta}}-b_{ij}\right)K_h(V_{ij}-V_{kl})\frac{f_{[V|\delta=0,Y=\Delta]}(V_{ij})}{h^d f_{[V|\delta=1,Y=0]}(V_{ij})}\mathrm{d}V_{ij}+O_p(\xi_n)\\
&\quad=\frac{1}{n_{\text{oco}}}\sum_{k=1}^{n}\sum_{l=1}^{M+1}(1-Y_{kl})\delta_{kl}\frac{S_{\Delta kl}}{a_{kl}}\left(\frac{\partial A_{kl}}{\partial\boldsymbol{\theta}}-b_{kl}\right)c_{\Delta 0}(V_{kl})+O_p(h^r+\xi_n),
\end{aligned}$$

where $c_{\Delta 0}(V_{kl})=f_{[V|\delta=0,Y=\Delta]}(V_{kl})/f_{[V|\delta=1,Y=0]}(V_{kl})$ and $S_{\Delta kl}=\Delta-E\{\psi_{ij}(\boldsymbol{\theta},a,b)|V_{kl}\}$ for $\Delta=0,1$ and ψ_{ij} is defined in (6). Now let

$$\begin{aligned}
E_n=n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{M+1}(1-\delta_{ij})\Bigg\{&\frac{S_{ij}}{a_{ij}}(\widehat{b}_{ij}-b_{ij})-\frac{Y_{ij}}{n_{\text{oco}}}\sum_{k=1}^{n}\sum_{l=1}^{M+1}(1-Y_{kl})\delta_{kl}\frac{S_{1kl}}{a_{kl}}\left(\frac{\partial A_{kl}}{\partial\boldsymbol{\theta}}-b_{kl}\right)c_{10}(V_{kl})\\
&-\frac{(1-Y_{ij})}{n_{\text{oco}}}\sum_{k=1}^{n}\sum_{l=1}^{M+1}(1-Y_{kl})\delta_{kl}\frac{S_{0kl}}{a_{kl}}\left(\frac{\partial A_{kl}}{\partial\boldsymbol{\theta}}-b_{kl}\right)c_{00}(V_{kl})\Bigg\}
\end{aligned}$$

which represents the first dominating term of U_{1n} given in (A3) minus its conditional expected value. Using Lemma 1 of Wang and Wang [20] it can be shown that $\text{var}\{E_n|\text{all } (\delta, Y, V^*, X^*)\} = O_p(\xi_n)$, resulting in $E_n = O_p(\eta_n)$. Therefore,

$$
\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{M+1}(1-\delta_{ij})\frac{S_{ij}}{a_{ij}}(\widehat{b}_{ij}-b_{ij}) \\
&= n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{M+1}(1-\delta_{ij})\frac{Y_{ij}}{n_{\text{oco}}}\sum_{k=1}^{n}\sum_{l=1}^{M+1}(1-Y_{kl})\delta_{kl}\frac{S_{1kl}}{a_{kl}}\left(\frac{\partial A_{kl}}{\partial\boldsymbol{\theta}}-b_{kl}\right)c_{10}(V_{kl}) \\
&\quad + n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{M+1}(1-\delta_{ij})\frac{(1-Y_{ij})}{n_{\text{oco}}}\sum_{k=1}^{n}\sum_{l=1}^{M+1}(1-Y_{kl})\delta_{kl}\frac{S_{0kl}}{a_{kl}}\left(\frac{\partial A_{kl}}{\partial\boldsymbol{\theta}}-b_{kl}\right)c_{00}(V_{kl})+O_p(\eta_n) \\
&= n^{-1/2}\sum_{k=1}^{n}\sum_{l=1}^{M+1}(1-Y_{kl})\delta_{kl}\frac{S_{1kl}}{a_{kl}}\left(\frac{\partial A_{kl}}{\partial\boldsymbol{\theta}}-b_{kl}\right)c_{10}^{*}(V_{kl}) \\
&\quad + n^{-1/2}\sum_{k=1}^{n}\sum_{l=1}^{M+1}(1-Y_{kl})\delta_{kl}\frac{S_{0kl}}{a_{kl}}\left(\frac{\partial A_{kl}}{\partial\boldsymbol{\theta}}-b_{kl}\right)c_{00}^{*}(V_{kl})+O_p(\eta_n) \\
&= n^{-1/2}\sum_{k=1}^{n}\sum_{l=1}^{M+1}(1-Y_{kl})\delta_{kl}\frac{1}{a_{kl}}\left(\frac{\partial A_{kl}}{\partial\boldsymbol{\theta}}-b_{kl}\right)\{S_{1kl}c_{10}^{*}(V_{kl})+S_{0kl}c_{00}^{*}(V_{kl})\}+O_p(\eta_n),
\end{aligned}
$$

where $c_{10}^{*}(V_{kl}) = \text{pr}(\delta=0|V,Y=1)\text{pr}(Y=1|V)/\text{pr}(\delta=1|V,Y=0)\text{pr}(Y=0|V)$ and $c_{00}^{*}(V_{kl}) = \text{pr}(\delta=0|V,Y=0)/\text{pr}(\delta=1|V,Y=0)$. Similarly, the second term of U_{1n} given in (A3) is

$$
-n^{-1/2}\sum_{k=1}^{n}\sum_{l=1}^{M+1}(1-Y_{kl})\delta_{kl}\frac{b_{kl}}{a_{kl}^{2}}(A_{kl}-a_{kl})\{S_{1kl}c_{10}^{*}(V_{kl})+S_{0kl}c_{00}^{*}(V_{kl})\}+O_p(\eta_n).
$$

Hence,

$$
U_{1n} = n^{-1/2}\sum_{k=1}^{n}\sum_{l=1}^{M+1}(1-Y_{kl})\delta_{kl}\frac{1}{a_{kl}}\left(\frac{\partial A_{kl}}{\partial\boldsymbol{\theta}}-\frac{b_{kl}A_{kl}}{a_{kl}}\right)\{S_{1kl}c_{10}^{*}(V_{kl})+S_{0kl}c_{00}^{*}(V_{kl})\}+O_p(\eta_n).
$$

Using similar arguments as above we can write U_{2n} as

$$
\begin{aligned}
U_{2n} &= n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{M+1}(1-\delta_{ij})\left(\frac{b_{ij}D_i}{a_{ij}}-T_i\right)(\widehat{a}_{ij}-a_{ij})+O_p(\eta_n) \\
&= n^{-1/2}\sum_{k=1}^{n}\sum_{l=1}^{M+1}(1-Y_{kl})\delta_{kl}\left(\frac{b_{kl}}{a_{kl}}D_{kl}^{*}-T_{kl}^{*}\right)(A_{kl}-a_{kl})\{c_{10}^{*}(V_{kl})+c_{00}^{*}(V_{kl})\}+O_p(\eta_n),
\end{aligned}
$$

where $D_i = D_i(\boldsymbol{\theta}, a, b)$ and $T_i = T_i(\boldsymbol{\theta}, a, b)$ are defined in Theorem 2. Also, $D_{kl}^{*} = E(D_k|V_{kl})$ and $T_{kl}^{*} = E(T_k|V_{kl})$. Therefore,

$$
U_n(\boldsymbol{\theta}, \widehat{a}, \widehat{b}) = U_n(\boldsymbol{\theta}, a, b) + U_{1n} - U_{2n} + O_p(\eta_n) = n^{-1/2}\sum_{i=1}^{n}J_i + O_p(\eta_n), \tag{A5}
$$

where

$$
\begin{aligned}
J_i = \sum_{j=1}^{M+1}\Bigg(& S_{ij}\left\{\delta_{ij}\frac{\partial\log(A_{ij})}{\partial\boldsymbol{\theta}}+(1-\delta_{ij})\frac{b_{ij}}{a_{ij}}\right\}+(1-Y_{ij})\delta_{ij}\Bigg[\frac{1}{a_{ij}}\left(\frac{\partial A_{ij}}{\partial\boldsymbol{\theta}}-\frac{b_{ij}A_{ij}}{a_{ij}}\right) \\
&\times\ \{S_{1ij}c_{10}^{*}(V_{ij})+S_{0ij}c_{00}^{*}(V_{ij})\}-\left(\frac{b_{ij}}{a_{ij}}D_{ij}^{*}-T_{ij}^{*}\right)(A_{ij}-a_{ij})\{c_{10}^{*}(V_{ij})+c_{00}^{*}(V_{ij})\}\Bigg]\Bigg)
\end{aligned}
$$

are identically and independently distributed. Hence it is seen that $U_n(\boldsymbol{\theta}, \widehat{a}, \widehat{b})$ is asymptotically a sum of iid mean zero random variables. If $\widehat{\boldsymbol{\theta}}$ is the estimate of $\boldsymbol{\theta}$, then by Taylor's series expansion, the influence function type representation is given by

$$
n^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}) = G_n^{-1}(\boldsymbol{\theta}, \widehat{a}, \widehat{b})n^{-1/2}\sum_{i=1}^{n}J_i + O_p(\eta_n),
$$

where $G_n(\boldsymbol{\theta}, a, b) = -n^{-1/2}\partial U_n(\boldsymbol{\theta}, a, b)/\partial\boldsymbol{\theta}^{\text{T}}$ as in (8). By calculating the mean and variance, and since $\sqrt{\xi_n}$ has a smaller order than η_n, it can also be shown that $G_n(\boldsymbol{\theta}, \widehat{a}, \widehat{b}) - G_n(\boldsymbol{\theta}, a, b) = O_p(\eta_n)$, and that $G_n(\boldsymbol{\theta}, \widehat{a}, \widehat{b})$ converges in probability to $G(\boldsymbol{\theta})$ which is given in (9). Now under the assumption that $G(\boldsymbol{\theta})$ is positive definite, $n^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ follows an asymptotic normal distribution with mean zero. Moreover, using Slutsky's theorem, we obtain its asymptotic variance-covariance matrix as $\Sigma = G^{-1}(\boldsymbol{\theta})\text{cov}(J_1)G^{-\text{T}}(\boldsymbol{\theta})$.