Statistica Sinica **28** (2018), 1677-1701
doi:https://doi.org/10.5705/ss.202015.0437

IMPUTATION-BASED ADJUSTED SCORE EQUATIONS IN GENERALIZED LINEAR MODELS WITH NONIGNORABLE MISSING COVARIATE VALUES

Fang Fang[1], Jiwei Zhao[2] and Jun Shao[1,3]

[1]*East China Normal University,* [2]*State University of New York at Buffalo and* [3]*University of Wisconsin at Madison*

Abstract: We consider the estimation of unknown parameters in a generalized linear model when some covariates have nonignorable missing values. When an instrument, a covariate that helps identifying parameters under nonignorable missingness, is appropriately specified, a pseudo likelihood approach similar to that in Tang, Little and Raghunathan (2003) or Zhao and Shao (2015) can be applied. However, this approach does not work well when the instrument is a weak predictor of the response given other covariates. We show that the asymptotic variances of the pseudo likelihood estimators for the regression coefficients of covariates other than the instrument diverge to infinity as the regression coefficient of the instrument goes to 0. By an imputation-based adjustment for the score equations, we propose a new estimator for the regression coefficients of the covariates other than the instrument. This works well even if the instrument is a weak predictor. It is semiparametric since the propensity of missing covariate data is completely unspecified. To solve the adjusted score equation, we develop an iterative algorithm that can be applied by using standard softwares at each iteration. We establish some theoretical results on the convergence of the proposed iterative algorithm and asymptotic normality of the resulting estimators. A variance estimation formula is also derived. Some simulation results and a data example are presented for illustration.

Key words and phrases: Adjusted likelihood, identifiability, instruments, nonignorable missing covariate data, pseudo-likelihood, semiparametric.

1. Introduction

Missing covariate data commonly exist in such health and biomedical related studies as clinical trials, observational data, environmental studies, and health surveys. We consider the estimation of θ, an unknown parameter vector of interest in a generalized linear model (GLM) on the conditional density $p(Y|X,\theta)$, where Y is a response variable always observed, and X is a covariate vector that may have missing data. Covariate values are missing at random (MAR) when

the probability of whether covariate values are missing, conditioned on (Y, X), does not depend on any unobserved value. Under MAR, statistical approaches for handling missing covariate data are well-developed, e.g., see Little (1992), Little and Rubin (2002), Robins, Rotnitzky and Zhao (1994), Robins, Hsieh and Newey (1995), Lipsitz, Ibrahim and Zhao (1999), Ibrahim et al. (2005), Tsiatis (2006), and Qin, Zhang and Leung (2009). In many biomedical and health related studies, however, missing covariate data are not MAR and are referred to as *nonignorable*, because the occurrence of a missing covariate is related to the covariate value itself even after conditioning on all observed data (Lipsitz et al. (1999)). For example, in a health survey of drug use with income as a covariate, a missing income value may be directly related to the income value after conditioning on drug use and other covariates; in an obesity study for children, adolescents with overweight conditions may hide their weight values, regardless of whether obesity and other covariates are observed.

We focus on nonignorable missing covariate data. This is challenging because, when missing data are nonignorable, there is a model identifiability issue (Robins and Ritov (1997)) and valid parameter estimators can be derived only under some model assumptions that may be hard to verify. To describe our approach, we first introduce two main assumptions. We assume that X can be decomposed into sub-vectors U and Z such that U may have missing values and Z is a fully observed covariate vector that is related to U but unrelated with the propensity of missing data once (Y, U) is conditioned,

$$P(R = 1|Y, U, Z) = P(R = 1|Y, U), \tag{1.1}$$

where $R = 1$ if U is fully observed and $R = 0$ otherwise. For nonignorable missing response Y data, Tang, Little and Raghunathan (2003) and Wang, Shao and Kim (2014) considered an assumption similar to (1.1). It is reasonable since it is typical in practice that not all covariates are related to the propensity of missing data, given other observed and unobserved covariates and the response. Wang, Shao and Kim (2014) showed that the existence of a covariate Z that is unrelated with the propensity of missing data is almost necessary for identifying parameters. Following Wang, Shao and Kim (2014), we call Z a nonresponse instrument or an instrument for short. Some discussion about how to choose an instrument can be found in Section 5.

With nonignorable missing data, Robins and Ritov (1997) showed that, in order to identify all unknown parameters, either the propensity of missing data or the original data distribution must have a parametric component. For covariate

missing data under (1.1), this means that either $P(R = 1|Y, U)$ or the density of $U|Z$ (U given Z) needs to have a parametric component. There exist some methods assuming both $P(R = 1|Y, U)$ and the density of $U|Z$ are parametric, using either maximum likelihood or Bayesian approaches, see, e.g., Lipsitz et al. (1999), Ibrahim, Lipsitz and Chen (1999), Herring and Ibrahim (2002), Stubbendick and Ibrahim (2003, 2006), Huang, Chen and Ibrahim (2005), and Ibrahim and Molenberghs (2009). Our second assumption is a parametric model $p(U|Z, \gamma)$ for $U|Z$ with an unknown parameter vector γ. Our approach is semiparametric since the propensity $P(R = 1|Y, U)$ in (1.1) is unspecified. To the best of our knowledge, semiparametric methods for handling nonignorable missing covariate data are limited.

By (1.1) and Bayes formula,

$$p(Z|Y, U, R = 1) = p(Z|Y, U) = \frac{p(Y|U, Z, \theta)p(U|Z, \gamma)p(Z)}{\int p(Y|U, z, \theta)p(U|z, \gamma)p(z)dz}, \tag{1.2}$$

where $p(Z)$ is the density of Z. Having N sampled subjects with realizations (y_i, u_i, z_i, r_i), $i = 1, ..., N$, independent and identically distributed as (Y, U, Z, R), where u_i is fully observed if and only if $r_i = 1$, we may estimate θ and γ by maximizing the pseudo-likelihood function

$$L(\theta', \gamma') = \prod_{i:r_i=1} \frac{p(y_i|u_i, z_i, \theta')p(u_i|z_i, \gamma')}{\sum_{j=1}^{N} p(y_i|u_i, z_j, \theta')p(u_i|z_j, \gamma')}. \tag{1.3}$$

That is based on (1.2) with $p(Z)$ estimated by the empirical distribution of Z and the true value (θ, γ) replaced by a parameter value (θ', γ') of the likelihood function. This pseudo-likelihood approach is an extension of the approach in Tang, Little and Raghunathan (2003) and Zhao and Shao (2015) to covariate missing data.

If Z is a weak predictor of Y when U is conditioned (e.g., Z is a surrogate of U), the pseudo-likelihood estimator for θ does not work well. To be more specific, denote the conditional density of $Y|(U, Z)$ as

$$p(y_i|u_i, z_i, \theta) = \exp(y_i\eta_i - b(\eta_i) + c(y_i)), \tag{1.4}$$

where b and c are known functions, $\eta_i = \eta(\alpha_c + \alpha_u^\tau u_i + \beta^\tau z_i)$, α_u and β are p- and q-dimensional, the superscript τ denotes transpose, η is a known one-to-one and continuously differentiable function, and $\theta = (\alpha, \beta)$ with $\alpha = (\alpha_c, \alpha_u)$ denotes the true but unknown parameter vector of interest. Without loss of generality, we assume throughout that there is no dispersion parameter or the dispersion parameter is known. When $\beta = 0$, $p(Y|U, Z, \theta)$ and $p(Y|U, z, \theta)$ in (1.2) do not depend on Z or z so that the right hand side of (1.2) does not involve θ.

Consequently, when Z is a weak predictor of Y given U, β is close to 0 and the estimator of α obtained by maximizing (1.3) that is based on (1.2) cannot work well. We show in Section 2.1 that the asymptotic variance of the pseudo likelihood estimator for α diverges to infinity when β goes to 0.

This paper aims to develop an efficient estimation method for α regardless of whether the dependence of $p(Y|U, Z, \theta)$ on Z is weak or not. Our proposed method consists of two parts: in the first we estimate β and γ by maximizing (1.3); in the second we estimate α. Instead of using (1.3) which can lead to an inefficient α estimator, we propose a score equation adjusted for missing covariate values using the estimated β and γ in the first part. When $p(Y|U, Z, \theta)$ depends on Z but the dependence is weak, the new estimator of α is much more efficient than the pseudo-likelihood estimator. This is illustrated in simulation results. In the special case where we know $\beta = 0$, the pseudo-likelihood cannot consistently estimate α while our proposed method can. Our approach utilizes all observed data, whereas (1.3) does not use the partially observed covariate data and y_i for any subject with $r_i = 0$.

Computation is often an issue in the presence of nonignorable missing data. Under a pure parametric framework, some Monte Carlo algorithms are developed in Lipsitz et al. (1999), Ibrahim, Lipsitz and Chen (1999), Herring and Ibrahim (2002), Huang, Chen and Ibrahim (2005), among others. For maximizing a semiparametric likelihood similar to (1.3), some algorithms were developed in Tang, Little and Raghunathan (2003) and Zhao and Shao (2015), but convergence of these algorithms has not been investigated. To solve the adjusted score equation in our method, we propose an iterative algorithm. Any available software packages for GLMs can be directly used in each iteration. We establish the convergence of the algorithm.

Our proposed method is described in Section 2. Section 3 studies the convergence of our algorithm and the asymptotic distributions of the proposed estimators. Results of simulation studies and a data example are presented in Sections 4 and 5, respectively. Proofs are provided in an Appendix.

2. Methodology

2.1. Estimation of β and γ

Let $\xi = (\theta, \gamma)$ and $\tilde{\xi} = (\tilde{\theta}, \tilde{\gamma})$, with $\tilde{\theta} = (\tilde{\alpha}, \tilde{\beta})$ the estimator of ξ obtained by maximizing $L(\theta', \gamma')$ in (1.3), F the true distribution of Z, F' denote a distribu-

tion, $\xi' = (\theta', \gamma')$, and

$$H_i(\xi', F') = r_i \left\{ \log p(y_i|u_i, z_i, \theta') p(u_i|z_i, \gamma') - \log \int p(y_i|u_i, z, \theta') p(u_i|z, \gamma') dF'(z) \right\}.$$

Then maximizing (1.3) is the same as maximizing $\sum_{i=1}^{N} H_i(\xi', \hat{F})$, where $\hat{F}$ is the empirical distribution of the Z-data. In what follows, we consider $H_i(\xi', F')$ as a random variable and $H(\xi', F')$ as $H_i(\xi', F')$ with (y_i, u_i, z_i, r_i) replaced by (Y, U, Z, R). Throughout, we use notation ∇_φ to denote the first order derivative with respect to φ and $\nabla^2_{\varphi\vartheta}$ to denote the second order derivative with respect to φ and ϑ.

Proposition 1. *Assume that*

(a) $E\{H(\xi', \hat{F}) - H(\xi', F)\} \to 0$ *as* $N \to \infty$, *and there exists* $\epsilon > 0$ *such that*

$$\lim_{N\to\infty} \sup_{\xi', \|F'-F\|<\epsilon} \left| \frac{1}{N} \sum_{i=1}^{N} H_i(\xi', F') - E\{H(\xi', F')\} \right| = 0.$$

(b) $H(\xi', F')$ *is continuously twice differentiable with respect to* ξ', *the matrix* $E\{-\nabla^2_{\xi\xi} H(\xi, F)\}$ *is positive definite, and* $\nabla^2_{\xi'\xi'} H(\xi', F)$ *is bounded by an integrable function in a neighborhood of* ξ.

When ξ *is identifiable from* (1.2),

(1) $\tilde{\xi}$ *is consistent, asymptotically normal and*

$$\sqrt{N}(\tilde{\xi} - \xi) = \frac{-1}{\sqrt{N}} \sum_{i=1}^{N} \left[E\{\nabla^2_{\xi\xi} H(\xi, F)\} \right]^{-1} \begin{pmatrix} \nabla_\theta H_i(\xi, F) + S_1(z_i, \xi, F) \\ \nabla_\gamma H_i(\xi, F) + T_1(z_i, \xi, F) \end{pmatrix} + o_p(1),$$

where S_1 *and* T_1 *are defined in* (A.3) *and* (A.4), *respectively, and* $o_p(1)$ *denotes a quantity converging to* 0 *in probability,*

(2) *if* $\beta \to 0$, *then* $E\{\nabla^2_{\alpha\alpha} H(\xi, F)\} \to 0$.

This result reveals a drawback of the pseudo-likelihood estimator based on (1.3): if β is close to 0, $\tilde{\alpha}$ can be very inefficient, although $\tilde{\beta}$ asymptotically works well whether or not β is close to 0. This result motivates us to develop a method to estimate α more efficiently.

Proposition 1 assumes that ξ is identifiable from (1.2), which holds under some conditions similar to those in Zhao and Shao (2015). The details are omitted here.

2.2. Estimation of α by adjusted score equation

We now consider how to estimate α differently. When there is no missing data, the parameter α under a GLM with a fixed regression coefficient $\tilde{\beta}$ for covariate z_i is typically estimated by solving the score equation

$$\begin{aligned} S(\alpha') &= \sum_{i=1}^{N} \nabla_{\alpha'} \log(p(y_i|u_i, z_i, \theta'))\big|_{\beta'=\tilde{\beta}} \\ &= \sum_{i=1}^{N} g(u_i, z_i, \alpha')\Big\{y_i - h(\alpha'_c + \alpha'^{\tau}_u u_i + \tilde{\beta}^{\tau} z_i)\Big\} = 0, \end{aligned} \tag{2.1}$$

where $h(\alpha'_c+\alpha'^{\tau}_u u_i+\tilde{\beta}^{\tau} z_i)=\nabla_{\eta} b(\eta_i)=E(y_i|u_i, z_i)$ and $g(u_i, z_i, \alpha')=\nabla_{\alpha'} h/\nabla^2_{\eta\eta} b(\eta_i)$ with α in η_i replaced by a parameter value $\alpha' = (\alpha'_c, \alpha'_u)$ and β replaced by $\tilde{\beta}$. The score equation (2.1) is valid in the sense that it satisfies $E\{S(\alpha)\big|_{\tilde{\beta}=\beta}\} = 0$ which, together with some other regularity conditions, ensures that the solution is a consistent estimator of α. The score equation (2.1) can be easily solved by any existing software packages for GLMs with an option of "offset".

When the u_i's have missing data, (2.1) cannot be solved. Since some components of U may be always observed but are not part of the instrument Z, we let $U = (U_1, U_2)$, where U_1 may have missing values and U_2 is always observed, and let $u_i = (u_{i1}, u_{i2})$ be the realization of U for subject i. We consider score equation which is (2.1) adjusted for missing covariate data:

$$\sum_{i=1}^{N} g(u^*_{i1}, u_{i2}, z_i, \alpha')\{y_i - h(\alpha'_c + \alpha'^{\tau}_{u1} u^*_{i1} + \alpha'^{\tau}_{u2} u_{i2} + \tilde{\beta}^{\tau} z_i)\} = 0, \tag{2.2}$$

where u^*_{i1} is a function of observed data. A popular candidate for u^*_{i1} is the conditional expectation $E(u_{i1}|u_{i2}, z_i)$. However, when $u^*_{i1} = E(u_{i1}|u_{i2}, z_i)$, (2.2) is not valid unless h is linear because

$$h(\alpha_c + \alpha^{\tau}_{u1} u^*_{i1} + \alpha^{\tau}_{u2} u_{i2} + \tilde{\beta}^{\tau} z_i) = E\{h(\alpha_c + \alpha^{\tau}_{u1} u_{i1} + \alpha^{\tau}_{u2} u_{i2} + \tilde{\beta}^{\tau} z_i)|u_{i2}, z_i\} \tag{2.3}$$

holds only for linear h when $u^*_{i1} = E(u_{i1}|u_{i2}, z_i)$, where the expectation in (2.3) is taken with respect to u_{i1}.

To have a valid (2.2) we must find an imputed u^*_{i1} satisfying (2.3). For nonlinear h, this can be achieved only when u^*_{i1} is a function of $(u_{i2}, z_i, \tilde{\beta})$ as well as the unknown α. For multivariate u_{i1}, the solution of u^*_{i1} to (2.3) is not unique. Among all the solutions, we propose to choose u^*_{i1} so that (2.3) holds and meanwhile is as close as possible to $E(u_{i1}|u_{i2}, z_i)$, i.e.,

$$u^*_{i1}(\alpha) = u^{(0)}_{i1} + \frac{\alpha_{u1}}{\|\alpha_{u1}\|^2}\left\{h^{-1}(\mu_i(\alpha)) - \alpha_c - \alpha^{\tau}_{u1} u^{(0)}_{i1} - \alpha^{\tau}_{u2} u_{i2} - \tilde{\beta}^{\tau} z_i\right\} \tag{2.4}$$

where $\mu_i(\alpha)$ denotes the quantity on the right hand side of (2.3), $u_{i1}^{(0)} = E(u_{i1}|u_{i2}, z_i)$, and $\|\cdot\|$ is the Euclidean norm. Formula (2.4) is also good for univariate u_{i1}. Using (2.4) in (2.2) leads to the valid score equation

$$\sum_{i=1}^{N} g(u_{i1}^*(\alpha'), u_{i2}, z_i, \alpha') \left\{ y_i - h(\alpha_c' + \alpha_{u1}'^{\tau} u_{i1}^*(\alpha') + \alpha_{u2}'^{\tau} u_{i2} + \tilde{\beta}^{\tau} z_i) \right\} = 0. \quad (2.5)$$

Since (2.5) may be hard to solve directly, we propose to solve it iteratively by introducing another parameter value α'':

$$S(\alpha'|\alpha'') = \sum_{i=1}^{N} g(u_{i1}^*(\alpha''), u_{i2}, z_i, \alpha') \left\{ y_i - h(\alpha_c' + \alpha_{u1}'^{\tau} u_{i1}^*(\alpha'') + \alpha_{u2}'^{\tau} u_{i2} + \tilde{\beta}^{\tau} z_i) \right\} = 0, \quad (2.6)$$

where $u_{i1}^*(\alpha'')$ is defined by (2.4) with α replaced by α''. Score equation (2.5) is the same as $S(\alpha'|\alpha') = 0$, and $E\{S(\alpha|\alpha)\big|_{\tilde{\beta}=\beta}\} = 0$. Having $\hat{\alpha}^{(t)}$ at the tth step in an iteration, we can compute $u_{i1}^*(\hat{\alpha}^{(t)})$ using (2.4) and solve $S(\alpha'|\hat{\alpha}^{(t)}) = 0$ over α' to get $\hat{\alpha}^{(t+1)}$. Here is an algorithm for solving (2.5).

Algorithm:

0. For each subject i, generate a random sample $\{u_{i1}^m, m = 1, \ldots, M\}$ from $p(u_{i1}|u_{i2}, z_i, \tilde{\gamma})$, where the Monte Carlo sample size M is a preset large positive integer, and $\tilde{\gamma}$ is the consistent estimator of γ obtained in Section 2.1.

1. Having $\hat{\alpha}^{(t)}$ at the tth iteration, compute $u_{i1}^*(\hat{\alpha}^{(t)})$ according to (2.4) with $\alpha = \hat{\alpha}^{(t)}$, $u_{i1}^{(0)}$ replaced by $E(u_{i1}|u_{i2}, z_i, \tilde{\gamma})$, and $\mu_i(\hat{\alpha}^{(t)})$ approximated by
$$\mu_i^{(t)}(\hat{\alpha}^{(t)}) = \frac{1}{M} \sum_{m=1}^{M} h\left(\hat{\alpha}_c^{(t)} + \hat{\alpha}_{u1}^{(t)^{\tau}} u_{i1}^m + \hat{\alpha}_{u2}^{(t)^{\tau}} u_{i2} + \tilde{\beta}^{\tau} z_i \right).$$

2. Replace $u_{i1}^*(\alpha'')$ in (2.6) by $u_{i1}^*(\hat{\alpha}^{(t)})$ and compute $\hat{\alpha}^{(t+1)}$ by solving $S(\alpha'|\hat{\alpha}^{(t)}) = 0$, where α' is a parameter value.

3. Execute 1-2 for $t = 1, \ldots, T$ until $\|\hat{\alpha}^{(T)} - \hat{\alpha}^{(T-1)}\|$ is smaller than a preset small threshold value, and take the estimator for α to be $\hat{\alpha} = \hat{\alpha}^{(T)}$.

Although our procedure is iterative, the computational burden is minimum. First, we only need to generate Monte Carlo samples in step 0 once for all iterations. Second, at each iteration, once $\alpha'' = \hat{\alpha}^{(t)}$ is fixed, the estimating equation $S(\alpha'|\hat{\alpha}^{(t)}) = 0$ is just a regular score equation for a GLM similar to (2.1), which can be easily solved by any software packages for GLMs with an option of "offset".

Our final proposed estimator for θ is $\hat{\theta} = (\hat{\alpha}, \tilde{\beta})$. The β estimator is the pseudo-likelihood estimator and our main effort is to improve the estimation efficiency of α. In some applications we know $\beta = 0$. Then the right hand side of (1.2) does not relate to θ so that the pseudo-likelihood (1.3) should not be used to estimate θ. However, γ can still be estimated by maximizing the reduced pseudo-likelihood

$$L(\gamma') = \prod_{i:r_i=1} \frac{p(u_i|z_i, \gamma')}{\sum_{j=1}^{N} p(u_i|z_j, \gamma')}. \tag{2.7}$$

Our proposed method for estimating α can still be applied with $\tilde{\beta} = 0$ and $\tilde{\gamma}$ as the maximizer of (2.7).

3. Theoretical Results

In this section, we study asymptotic properties of our proposed estimators. These include the convergence property of the sequence $\{\hat{\alpha}^{(t)}, t = 1, 2, \ldots\}$ from the proposed algorithm, the consistency, asymptotic normality, and variance estimation for $\hat{\theta} = (\hat{\alpha}, \tilde{\beta})$.

3.1. Convergence of $\hat{\alpha}^{(t)}$

For simplicity, we assume that β and γ are known and $\tilde{\beta} = \beta$ and $\tilde{\gamma} = \gamma$ in this subsection. This does not affect the generality of our result on convergence since $\tilde{\beta}$ and $\tilde{\gamma}$ are consistent as shown in Proposition 1. Let

$$S_i(\alpha'|\alpha'') = g(u_{i1}^*(\alpha''), u_{i2}, z_i, \alpha') \left\{ y_i - h(\alpha_c' + \alpha_{u1}'^{\tau} u_{i1}^*(\alpha'') + \alpha_{u2}'^{\tau} u_{i2} + \beta^{\tau} z_i) \right\}$$

so that $S(\alpha'|\alpha'')$ in (2.6) is the sum of $S_i(\alpha'|\alpha'')$, $i = 1, \ldots, N$. Take $l(\alpha'|\alpha'') = \sum_{i=1}^{N} l_i(\alpha'|\alpha'')$, where $l_i(\alpha'|\alpha'') = \log(p(y_i|u_{i1}^*(\alpha''), u_{i2}, z_i, \alpha', \beta))$. Note that $\nabla_{\alpha'} l_i(\alpha'|\alpha'') = S_i(\alpha'|\alpha'')$.

Theorem 2. *Assume the following:*

(a) *The parameter space of α (also of α' and α'') is an open subset of R^{p+1} with the true value α as an interior point;*

(b) *$l(\alpha'|\alpha)$ as a function of α' is strictly concave in B, a compact and convex neighborhood of α, and $E\{\sup_{\alpha', \alpha'' \in B} \|\nabla_{\alpha''} l_i(\alpha'|\alpha'')\|\} < \infty$;*

(c) *the absolute values of the minimum and maximum eigenvalues of the matrix $I_1(\alpha)^{-1} I_2(\alpha)$ are both less than 1, where $I_1(\alpha) = E\{-\nabla_{\alpha'} S_i(\alpha'|\alpha)\}_{\alpha'=\alpha}$, and $I_2(\alpha) = E\{\nabla_{\alpha''} S_i(\alpha|\alpha'')\}_{\alpha''=\alpha}$.*

If $\hat{\alpha}^{(1)}$ *converges to* α *in probability, then there exists a sequence* $\{\hat{\alpha}^{(t)}, t \geq 1\}$ *such that* $S(\hat{\alpha}^{(t+1)}|\hat{\alpha}^{(t)}) = 0$ *and*

$$\lim_{N\to\infty} P\left(\|\hat{\alpha}^{(t)} - \hat{\alpha}\| \geq \|\hat{\alpha}^{(t+1)} - \hat{\alpha}\| \ \text{ for all } t\right) = 1,$$

where $\hat{\alpha} = \lim_{t\to\infty} \hat{\alpha}^{(t)}$ *is a consistent solution to* $S(\alpha'|\alpha') = 0$.

If we pick a consistent estimator $\hat{\alpha}^{(1)}$ as the starting point in our algorithm, for large enough N there exists a sequence $\{\hat{\alpha}^{(t)}, t \geq 1\}$ that converges monotonically to $\hat{\alpha}$ with probability approaching 1. We can use $\tilde{\alpha}$ obtained in Section 2.1 as our initial $\hat{\alpha}^{(1)}$.

3.2. Asymptotic normality of $\hat{\theta} = (\hat{\alpha}, \tilde{\beta})$ and variance estimation

We have shown the consistency and asymptotic normality of $\tilde{\beta}$ in Proposition 1. Since $\hat{\alpha}$ is a consistent solution to $S(\alpha'|\alpha') = 0$ and $E\{S(\alpha|\alpha)\big|_{\tilde{\beta}=\beta}\} = 0$, we can show that $\hat{\alpha} - \alpha$ is asymptotically normal with mean 0 and some covariance matrix, using a standard asymptotic analysis. For statistical inference, we need to estimate the asymptotic covariance matrix of $\hat{\alpha}$ and the joint asymptotic covariance matrix of $\hat{\theta} = (\hat{\alpha}, \tilde{\beta})$. Here $S(\alpha'|\alpha')$ depends on the estimated $\tilde{\beta}$ as well as the estimated $\tilde{\gamma}$ since it is involved in the conditional expectation $E(u_{i1}|u_{i2}, z_i, \tilde{\gamma})$. The variations of $\tilde{\beta}$ and $\tilde{\gamma}$ have to be considered for variance estimation. We rewrite $S_i(\alpha'|\alpha')$ as $S_i(\alpha', \beta', \gamma')$ in the rest of this subsection.

In the proof of Proposition 1 in the Appendix, we show that

$$\sqrt{N}(\tilde{\beta} - \beta) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} D_{i2} + o_p(1),$$

$$\sqrt{N}(\tilde{\gamma} - \gamma) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} D_{i3} + o_p(1),$$

where D_{i2} and D_{i3} are defined in (A.5). Based on this, we have the asymptotic representation of $\hat{\theta} = (\hat{\alpha}, \tilde{\beta})$ with its explicit influence function.

Theorem 3. *Assume the conditions in Proposition* 1 *and Theorem* 2 *hold. Also assume* $S_i(\alpha', \beta', \gamma')$ *is continuously differentiable with respect to* $\xi' = (\alpha', \beta', \gamma')$, $\nabla_{\xi'} S_i(\alpha', \beta', \gamma')$ *is bounded by an integrable function in a neighborhood of* ξ, *and the matrix* $E\{-\nabla_\alpha S_i(\alpha, \beta, \gamma)\}$ *is positive definite. Then*

$$\sqrt{N}(\hat{\theta} - \theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} E_i + o_p(1) \to_d N(0, \Sigma),$$

where $E_i = \kappa(w_i, \xi, F, A, B_1, B_2, B_3)$ *is defined in* (A.12) *in the Appendix,* $w_i =$

(y_i, u_i, z_i, r_i), $A = E\{\nabla^2_{\xi\xi} H(\xi, F)\}$, $B_1 = E\{\nabla_\alpha S_i(\alpha, \beta, \gamma)\}$, $B_2 = E\{\nabla_\beta S_i(\alpha, \beta, \gamma)\}$, $B_3 = E\{\nabla_\gamma S_i(\alpha, \beta, \gamma)\}$, *and* $\Sigma = Var(E_i)$.

Now we can obtain a consistent variance estimator $\hat{\Sigma}$ of Σ using the substitution technique.

Theorem 4. *Let* $\hat{\Sigma}$ *be the sample covariance matrix based on* $\hat{E}_i = \kappa(w_i, \hat{\xi}, \hat{F}, \hat{A}, \hat{B}_1, \hat{B}_2, \hat{B}_3)$, $i = 1, \ldots, N$, *where* $\hat{F}$ *is the empirical distribution of* Z*-data,* $\hat{A} = \sum_{i=1}^N \nabla^2_{\xi\xi} H_i(\hat{\xi}, \hat{F})/N$, $\hat{B}_1 = \sum_{i=1}^N \nabla_\alpha S_i(\hat{\xi})/N$, $\hat{B}_2 = \sum_{i=1}^N \nabla_\beta S_i(\hat{\xi})/N$, $\hat{B}_3 = \sum_{i=1}^N \nabla_\gamma S_i(\hat{\xi})/N$, *and* $\hat{\xi} = (\hat{\alpha}, \tilde{\beta}, \tilde{\gamma})$. *Assume the conditions in Theorem 3 hold and, for any* $c > 0$,

$$\sup_{\|w\| \le c} \|\kappa(w, \hat{\xi}, \hat{F}, \hat{A}, \hat{B}_1, \hat{B}_2, \hat{B}_3) - \kappa(w, \xi, F, A, B_1, B_2, B_3)\| = o_p(1) \tag{3.1}$$

and there exist a constant $c_0 > 0$ *and an integrable function* $h(w) \ge 0$ *such that*

$$P(\|\kappa(w, \hat{\xi}, \hat{F}, \hat{A}, \hat{B}_1, \hat{B}_2, \hat{B}_3)\|^2 \le h(w) \text{ for all } \|w\| \ge c_0) \to 1. \tag{3.2}$$

Then $\|\hat{\Sigma} - \Sigma\| = o_p(1)$ *as* $N \to \infty$.

4. Simulation Studies

Some simulation studies were conducted to examine the finite sample performance of the proposed estimator $\hat{\theta}$, especially $\hat{\alpha}$, and to compare it with some other estimators. We considered univariate Y, U, and Z in thirteen cases, a combination of continusous/discrete Y and U, and different values of β:

(A1) $Y|U, Z \sim N(-2 + U + 0.1Z, 1)$, $U|Z \sim N(2 - 4Z^2, 1)$, $Z \sim U(0, 1)$, and $P(R = 1|Y, U, Z) = \Phi(1 - U + |Y|)$, where Φ is the cumulative distribution function of the standard normal distribution.

(A2) the same as (A1) except that $Y|U, Z \sim N(-2 + U + Z, 1)$.

(A3) the same as (A1) except that $Y|U, Z \sim N(-2 + U + 2Z, 1)$.

(A4) the same as (A2) except that $P(R = 1|Y, U, Z) = \Phi(1 + |Y|)$, i.e., missing is ignorable.

(B1) Y is binary with $P(Y = 1|U, Z) = \{1 + \exp(-(-1 - U + 0.1Z))\}^{-1}$, $U|Z \sim N(-2 + 2Z^2, 1)$, $Z \sim N(1, 1)$, and $P(R = 1|Y, U, Z) = \Phi(2 + U - Y)$.

(B2) the same as (B1) except that $P(Y = 1|U, Z) = \{1 + \exp(-(-1 - U + Z))\}^{-1}$.

(B3) the same as (B1) except that $P(Y = 1|U, Z) = \{1 + \exp(-(-1 - U + 2Z))\}^{-1}$.

(C1) $Y|U,Z \sim N(-1+U+0.1Z,1)$, U is binary with $P(U=1|Z)=\{1+\exp(-(-2+4Z^2))\}^{-1}$, $Z\sim N(1,1)$, and $P(R=1|Y,U,Z)=\Phi(1.5-U+Y)$.

(C2) the same as (C1) except that $Y|U,Z \sim N(-1+U+Z,1)$ and $P(R=1|Y,U,Z)=\Phi(0.5-U+Y)$.

(C3) the same as (C1) except that $Y|U,Z \sim N(-1+U+2Z,1)$ and $P(R=1|Y,U,Z)=\Phi(-U+Y)$.

(D1) Y is binary with $P(Y=1|U,Z)=\{1+\exp(-(-1+U+0.1Z))\}^{-1}$, U is binary with $P(U=1|Z)=\{1+\exp(-(1-2Z^2))\}^{-1}$, $Z\sim N(1,1)$, and $P(R=1|Y,U,Z)=\Phi(0.1+Y+YU)$.

(D2) the same as (D1) except that $P(Y=1|U,Z)=\{1+\exp(-(-1+U+Z))\}^{-1}$ and $P(R=1|Y,U,Z)=\Phi(-0.3+Y+YU)$.

(D3) the same as (D1) except that $P(Y=1|U,Z)=\{1+\exp(-(-1+U+2Z))\}^{-1}$ and $P(R=1|Y,U,Z)=\Phi(-0.5+Y+YU)$.

For each continuous/discrete combination of Y and U, we considered $\beta=0.1$, 1, 2 and adjusted the propensity function to get similar percentages of complete data, around 74% and 63% for the first six cases and the last six cases, respectively. The sample size was $N=500$. Four methods were compared: the full data method (assume no missing data, which is not applicable in practice and was just used as a standard), complete case analysis, the pseudo-likelihood method maximizing (1.3), and the proposed method. For the proposed method, the Monte Carlo sample size was $M=10{,}000$, and the threshold value for the algorithm convergence was 0.001. In cases A1, A2, A3, and A4, we also considered the maximum likelihood estimation assuming MAR.

Based on 1,000 simulation replications, Tables 1-4 report, for each method under consideration, the simulation average of the relative bias in %, standard deviation, standard error (the estimate of standard deviation), and empirical coverage probability in % of approximate 95% confidence interval for θ based on the normal approximation. The standard errors based on the full data method and complete case analysis were obtained by the standard formulas in GLMs. The standard error for the pseudo-likelihood estimator was estimated by the asymptotic representation in (A.5). The standard error for the proposed method was based on $\hat{\Sigma}$ defined in Theorem 4.

The simulation results in Tables 1-4 can be summarized as follows. (1) The complete case analysis has large bias and low coverage probability as expected.

Table 1. Simulation results for normal Y and normal U.

		Case (A1)			Case (A2)		
	method	$\alpha_c = -2$	$\alpha_u = 1$	$\beta = 0.1$	$\alpha_c = -2$	$\alpha_u = 1$	$\beta = 1$
relative	full	−0.2	−0.1	3.8	−0.3	0.1	1.0
bias %	CC	−3.8	4.2	97.4	−5.4	8.3	19.6
	PL	−1.7	−2.8	98.4	−3.3	−2.3	2.4
	proposed	−0.2	1.6	98.4	4.0	1.0	2.4
	MLE-MAR	10.8	3.2	−198.1	10.1	4.2	−15.4
standard	full	0.142	0.043	0.227	0.143	0.043	0.229
deviation	CC	0.171	0.058	0.275	0.173	0.055	0.277
	PL	3.663	1.675	0.351	2.281	1.003	0.276
	proposed	0.361	0.103	0.351	0.487	0.120	0.276
	MLE-MAR	0.165	0.052	0.261	0.163	0.050	0.255
standard	full	0.146	0.043	0.232	0.146	0.043	0.232
error	CC	0.176	0.056	0.281	0.176	0.055	0.282
	PL	5.533	2.625	0.205	2.337	1.025	0.273
	proposed	0.375	0.103	0.205	0.483	0.121	0.273
	MLE-MAR	0.164	0.051	0.260	0.163	0.050	0.255
coverage	full	95.4	95.2	95.0	95.4	95.0	95.4
probability	CC	93.4	87.4	94.0	90.7	67.3	89.9
%	PL	96.1	97.1	64.0	97.8	97.7	95.2
	proposed	93.6	96.1	64.0	95.5	96.6	95.2
	MLE-MAR	73.4	88.2	86.9	76.0	86.8	90.6
		Case (A3)			Case (A4), MAR		
		$\alpha_c = -2$	$\alpha_u = 1$	$\beta = 2$	$\alpha_c = -2$	$\alpha_u = 1$	$\beta = 1$
relative	full	0.1	−0.1	−0.2	−0.5	0.3	1.5
bias %	CC	−5.4	11.0	12.4	−12.6	10.2	17.7
	PL	−1.7	−0.3	1.5	−3.2	−2.4	3.8
	proposed	3.4	0.9	1.5	5.8	2.3	3.8
	MLE-MAR	9.5	4.4	−7.1	−0.7	0.4	2.2
standard	full	0.148	0.044	0.235	0.140	0.042	0.224
deviation	CC	0.177	0.052	0.279	0.194	0.054	0.312
	PL	1.093	0.473	0.315	2.659	0.980	0.313
	proposed	0.511	0.128	0.315	0.635	0.151	0.313
	MLE-MAR	0.165	0.050	0.262	0.163	0.047	0.258
standard	full	0.146	0.043	0.232	0.145	0.043	0.232
error	CC	0.174	0.053	0.277	0.200	0.056	0.322
	PL	1.088	0.472	0.305	2.786	1.019	0.312
	proposed	0.501	0.129	0.305	0.613	0.151	0.312
	MLE-MAR	0.165	0.050	0.262	0.162	0.048	0.260
coverage	full	94.9	94.7	94.6	95.9	95.6	95.8
probability %	CC	89.9	44.9	84.8	75.8	55.5	92.1
	PL	95.9	95.9	94.3	98.1	97.8	95.7
	proposed	96.3	96.3	94.3	95.5	96.5	95.7
	MLE-MAR	76.1	83.6	89.8	95.1	94.9	95.2

full: the estimator assuming no missing data
CC: the estimator using samples with no missing data
PL: the pseudo-likelihood estimator by maximizing (1.3)
proposed: our proposed estimator
MLE-MAR: maximum likelihood estimation assuming MAR

Table 2. Simulation results for binary Y and normal U.

		Case (B1)			Case (B2)			Case (B3)		
	method	$\alpha_c=-1$	$\alpha_u=-1$	$\beta=0.1$	$\alpha_c=-1$	$\alpha_u=-1$	$\beta=1$	$\alpha_c=-1$	$\alpha_u=-1$	$\beta=2$
relative	full	−2.5	−2.2	12.5	−3.9	−2.4	4.4	−2.3	−1.6	1.8
bias %	CC	−35.7	22.4	20.6	−44.0	13.4	12.7	−50.3	5.3	11.5
	PL	17.0	−9.3	−6.3	−16.6	−35.6	7.3	−10.7	−12.3	4.0
	proposed	−3.2	−3.6	−6.3	−5.9	−4.4	7.3	−4.3	−3.3	4.0
standard	full	0.226	0.121	0.242	0.225	0.113	0.262	0.244	0.105	0.308
deviation	CC	0.266	0.141	0.303	0.295	0.128	0.337	0.349	0.124	0.412
	PL	2.251	1.437	0.262	1.125	1.093	0.332	0.491	0.452	0.410
	proposed	0.266	0.176	0.262	0.281	0.148	0.332	0.310	0.138	0.410
standard	full	0.220	0.119	0.235	0.227	0.111	0.261	0.239	0.103	0.302
error	CC	0.257	0.147	0.289	0.292	0.134	0.335	0.326	0.121	0.388
	PL	2.751	1.642	0.157	0.930	1.002	0.329	0.471	0.377	0.394
	proposed	0.234	0.165	0.157	0.279	0.147	0.329	0.300	0.133	0.394
coverage	full	94.3	95.2	94.7	95.6	95.2	95.5	94.9	95.0	95.1
probability	CC	76.9	62.0	94.6	72.7	78.6	95.1	70.4	90.0	92.6
%	PL	91.0	79.3	61.8	96.3	91.2	94.0	95.6	93.9	94.4
	proposed	94.0	95.1	61.8	94.1	94.1	94.0	94.5	94.6	94.4

full: the estimator assuming no missing data
CC: the estimator using samples with no missing data
PL: the pseudo-likelihood estimator by maximizing (1.3)
proposed: our proposed estimator

Table 3. Simulation results for normal Y and binary U.

		Case (C1)			Case (C2)			Case (C3)		
	method	$\alpha_c=-1$	$\alpha_u=1$	$\beta=0.1$	$\alpha_c=-1$	$\alpha_u=1$	β=1	$\alpha_c=-1$	$\alpha_u=1$	$\beta=2$
relative	full	0.1	0.2	−1.2	0.1	−0.2	−0.1	0.1	−0.5	0.1
bias %	CC	41.2	0.3	−26.5	69.8	−5.3	−21.9	72.0	−20.1	−11.3
	PL	−21.0	4.6	22.6	−3.2	1.4	0.8	−4.7	2.9	0.4
	proposed	−0.4	−2.6	22.6	0.2	−1.8	0.8	−0.9	−0.8	0.4
standard	full	0.082	0.118	0.053	0.083	0.116	0.056	0.083	0.115	0.055
deviation	CC	0.084	0.120	0.055	0.108	0.147	0.070	0.120	0.158	0.080
	PL	1.718	1.718	0.069	0.353	0.394	0.092	0.259	0.281	0.095
	proposed	0.093	0.158	0.069	0.102	0.183	0.092	0.123	0.180	0.095
standard	full	0.081	0.115	0.053	0.081	0.115	0.053	0.081	0.115	0.053
error	CC	0.087	0.123	0.056	0.111	0.153	0.070	0.129	0.166	0.078
	PL	2.243	2.342	0.072	0.334	0.373	0.091	0.253	0.283	0.096
	proposed	0.094	0.167	0.072	0.100	0.182	0.091	0.120	0.183	0.096
coverage	full	95.0	94.3	95.2	94.1	94.1	94.4	94.1	94.4	94.9
probability	CC	0.3	94.9	92.9	0.0	94.9	12.0	0.0	77.9	20.4
%	PL	96.6	98.6	88.5	94.6	94.2	94.1	94.4	94.5	94.2
	proposed	95.6	94.9	88.5	95.1	94.8	94.1	95.6	95.5	94.2

full: the estimator assuming no missing data
CC: the estimator using samples with no missing data
PL: the pseudo-likelihood estimator by maximizing (1.3)
proposed: our proposed estimator

(2) When β is small, the standard error for the pseudo-likelihood estimator $\tilde{\alpha}$ is large. (3) The proposed estimator $\hat{\alpha}$ works quite well in terms of bias, standard deviation, and coverage probability; compared with the pseudo-likelihood method, the estimation efficiency for α improves a lot in terms of smaller stan-

Table 4. Simulation results for binary Y and binary U.

		Case (D1)			Case (D2)			Case (D3)		
	method	$\alpha_c=-1$	$\alpha_u=1$	$\beta=0.1$	$\alpha_c=-1$	$\alpha_u=1$	β=1	$\alpha_c=-1$	$\alpha_u=1$	$\beta=2$
relative	full	−0.8	−0.7	1.3	−1.6	1.9	1.3	−2.2	2.0	1.5
bias %	CC	46.5	12.2	1.5	66.3	26.9	2.4	77.8	34.3	3.1
	PL	−548.8	135.1	−1.1	−3.0	−1.8	2.5	−5.4	1.4	3.2
	proposed	0.4	1.3	−1.1	−3.2	7.0	2.5	−5.4	12.8	3.2
standard	full	0.196	0.233	0.107	0.211	0.251	0.138	0.257	0.295	0.214
deviation	CC	0.230	0.278	0.130	0.284	0.354	0.191	0.378	0.460	0.332
	PL	8.190	9.063	0.111	0.423	1.003	0.190	0.518	0.771	0.329
	proposed	0.208	0.350	0.111	0.312	0.481	0.190	0.457	0.708	0.329
standard	full	0.193	0.235	0.113	0.212	0.250	0.136	0.256	0.292	0.212
error	CC	0.226	0.278	0.132	0.283	0.347	0.190	0.373	0.451	0.324
	PL	3.409	5.132	0.097	0.426	1.131	0.188	0.513	0.770	0.319
	proposed	0.200	0.347	0.097	0.312	0.480	0.188	0.446	0.727	0.319
coverage	full	95.2	94.7	96.4	95.5	95.0	95.0	95.3	95.1	95.3
probability	CC	43.9	94.0	96.1	34.4	89.5	95.2	43.1	89.7	95.3
%	PL	54.7	72.3	83.3	93.4	98.8	94.9	93.4	95.4	95.0
	proposed	94.3	95.7	83.3	95.3	96.2	94.9	94.6	95.8	95.0

full: the estimator assuming no missing data
CC: the estimator using samples with no missing data
PL: the pseudo-likelihood estimator by maximizing (1.3)
proposed: our proposed estimator

dard error especially when β is small. (4) The pseudo likelihood estimator $\tilde{\beta}$ works well when β is not close to 0. When $\beta = 0.1$, $\tilde{\beta}$ could have large relative bias and low coverage probability; when β is close to 0, however, the effect of α dominates so that a not-so-accurate $\tilde{\beta}$ may not be a serious problem if α can be estimated accurately. (5) The proposed estimator works well under MAR in case A4, although it is less efficient than the maximum likelihood estimator (MLE-MAR). The MLE-MAR, however, may have large relative biases and low coverage probabilities when missing is nonignorable, in cases A1, A2, and A3.

To check the dependence of the pseudo-likelihood and our proposed methods on assumption (1.1), we conducted some further simulations under cases A5 and B4, which are A2 and B2 but with propensity functions $P(R = 1|Y, U, Z) = \Phi(1 - U + |Y| + Z)$ and $P(R = 1|Y, U, Z) = \Phi(2 + U - Y + Z)$, respectively. The sample size was 500 and the simulation size was 1,000. Results with the format as the previous simulations are given in Table 5.

The pseudo likelihood and proposed methods are not robust against the violation of assumption (1.1) although, under A5, the performance of our proposed method is still good. When the propensity depends on Z, the pseudo likelihood and proposed methods could have large relative bias and low coverage probability, as case B4 shows. The selection of an instrument with nonignorable missing

Table 5. Simulation results with violation of assumption (1.1).

		Case (A5)			Case (B4)		
	method	$\alpha_c = -2$	$\alpha_u = 1$	$\beta = 1$	$\alpha_c = -1$	$\alpha_u = -1$	$\beta = 1$
relative	full	−0.3	0.1	1.0	−4.1	−2.2	4.7
bias %	CC	−4.3	7.4	16.8	−54.2	3.9	37.7
	PL	−1.2	−38.3	3.1	−11.6	−51.5	16.9
	proposed	2.3	5.1	3.1	−13.0	−8.4	16.9
standard	full	0.143	0.043	0.227	0.231	0.111	0.264
deviation	CC	0.167	0.051	0.262	0.344	0.133	0.381
	PL	2.196	0.961	0.268	2.712	1.933	0.601
	proposed	0.483	0.113	0.268	0.440	0.237	0.601
standard	full	0.146	0.043	0.232	0.227	0.111	0.262
error	CC	0.169	0.051	0.268	0.328	0.133	0.364
	PL	2.207	0.972	0.265	1.197	1.001	0.300
	proposed	0.456	0.111	0.265	0.263	0.143	0.300
coverage	full	95.4	95.6	95.5	94.8	95.6	95.2
probability %	CC	92.1	69.8	90.9	66.3	92.7	86.1
	PL	97.4	96.6	95.0	41.1	67.7	58.5
	proposed	95.1	94.4	95.0	66.7	70.4	58.5

full: the estimator assuming no missing data
CC: the estimator using samples with no missing data
PL: the pseudo-likelihood estimator by maximizing (1.3)
proposed: our proposed estimator

covariates is an interesting and difficult research topic. It will be a part of the authors' future research.

5. An Example

For illustration, especially for a discussion about how to select the instrument Z, we analyzed a data set from the National Health and Nutrition Examination Survey (NHANES 2005), which was designed to assess the health and nutritional status of adults and children in the United States. The data are available at http://www.cdc.gov/nchs/nhanes.htm. We focused on how middle-aged and old people's hypertension is related to body fat, age and gender. Dual-energy x-ray absorptiometry (*dxa*) has been accepted as the gold standard direct measurement of body fat. However, some of the *dxa* data are missing. Typically, NHANES data sets with missing data/variables are released without statistical adjustment and, as officially pointed out by the United States Centers for Disease Control and Prevention, examination of missing items in the *dxa* data files indicates that there seems to be systematic, non-random patterns to the missing data in *dxa*. Use of only the measured variables could lead to biased results.

In our analysis, the binary response variable Y equals 1 if the subject has hypertension, i.e., the systolic blood pressure (the average of BPXSY1-4) is greater than 140 or the diastolic blood pressure (the average of BPXDI1-4) is greater than 90, and equals 0 otherwise. The covariate U_1 that may have missing values is the body fat percentage measured by *dxa* (variable DXDTOPF). There are $N = 1{,}591$ subjects and 393 (24.7%) of them have missing *dxa*. There are other covariates fully observed: *age* (variable RIDAGEYR), *gender* (variable RIAGENDR, 1 for male and 0 for female), and *bmi* (body mass index, logarithm of variable BMXBMI). The *bmi* can be considered as a surrogate variable for *dxa*, although it is less accurate than *dxa*, and hence the conditional independence assumption is commonly made (Reilly and Pepe (1995); Bashir and Duffy (1997); Horton and Laird (2001)): Y and *bmi* are conditionally independent given *dxa*, *age* and *gender*. Thus, the GLM is

$$P(Y = 1|dxa, age, gender, bmi) = \frac{\exp(\theta_1 + \theta_2 dxa + \theta_3 age + \theta_4 gender)}{1 + \exp(\theta_1 + \theta_2 dxa + \theta_3 age + \theta_4 gender)}. \tag{5.1}$$

For estimating the parameter $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$, we considered the complete case analysis, the maximum likelihood estimator assuming MAR, the pseudo likelihood method, and the proposed method. For the latter two, an instrument Z needs to be selected from *age*, *gender*, and *bmi*. Since the fewer components Z has, the more likely condition (1.1) holds, we only considered univariate instruments. As indicated in Zhao and Shao (2015), as an instrument, a binary variable (*gender*) alone is not enough to identify θ. Thus we considered two options for instrument: $Z = age$ and $Z = bmi$. Once Z was selected, U_2 contained components of $(age, gender, bmi)$ that are not in Z. Except for the complete case analysis, a parametric model for $U|Z$ was also needed. For this data set, *age*, *gender* and *bmi* are almost independent so that only a parametric model for $dxa|age, gender, bmi$ was needed. We assumed that $dxa|age, gender, bmi \sim N(\gamma_1 + \gamma_2 age + \gamma_3 gender + \gamma_4 bmi, \gamma_5)$.

When $Z = bmi$, (5.1) means $\beta = 0$. We could apply the proposed method by setting $\tilde{\beta} = 0$ and estimating γ with (2.7).

Table 6 reports the estimate, standard error and p-value for θ based on the methods considered. When *age* is used as the instrument, as pointed out by Zhao and Shao (2015), the hypothesis that the effect of *age* equals zero cannot be tested, so the p-value for *age* is not available. The proposed method using different covariates as the instrument Z produces similar results. When $Z = age$, in which both pseudo-likelihood and the proposed method are applicable, the

Table 6. Analysis results of the NHANES data.

method	effect	estimate	standard error	p-value
complete case	intercept	−5.3175	0.7156	0.000
	dxa	0.0076	0.0126	0.549
	gender	0.0593	0.2094	0.777
	age	0.0674	0.0100	0.000
MLE-MAR	intercept	−5.0807	0.6044	0.000
	dxa	0.0223	0.0102	0.028
	gender	0.1821	0.1780	0.306
	age	0.0538	0.0080	0.000
proposed	intercept	−5.4514	0.5993	0.000
Z = bmi	dxa	0.0313	0.0103	0.002
	gender	0.2958	0.1712	0.084
	age	0.0534	0.0080	0.000
proposed	intercept	−6.3600	0.7585	0.000
Z = age	dxa	0.0293	0.0135	0.030
	gender	0.2724	0.2042	0.182
	age	0.0702	0.0099	n.a.
pseudo likeliood	intercept	−5.9139	5.1676	0.253
Z = age	dxa	0.0003	0.0279	0.992
	gender	0.3122	0.9007	0.729
	age	0.0702	0.0099	n.a.

n.a.: not available

proposed estimators for *intercept*, *dxa* and *gender* have much smaller standard errors than the pseudo likelihood estimators. The pseudo-likelihood method indicates the effect of *dxa* is not significant, which is inconsistent to our common knowledge. The reason might be that the variance of this estimator is too large. The complete case analysis has comparable standard errors with the proposed method but it also fails to detect the significant effect of *dxa* probably due to the estimation bias for *dxa*. This supports the official statement made by the United States Centers for Disease Control and Prevention. Although the maximum likelihood method assuming MAR produces smaller estimator of *gender* effect than the proposed method, they are comparable in this example. This may be because *dxa* and *bmi* are highly correlated so that a nonignorable propensity is close to MAR, only one of *dxa* and *bmi* is needed in the propensity of missing data and it seems more reasonable to include *dxa* in the propensity rather than its surrogate *bmi*. It is important to try different methods under different assumptions and to obtain robust results, especially because we cannot check the assumptions on the propensity.

In this example we tried different choices of the instrument Z to check the

robustness of the proposed estimator. Overall, our proposed method is more flexible in choosing an instrument than the pseudo-likelihood method. Based on the analysis results, $Z = bmi$ may be preferred for several reasons: if we use $Z = age$, the significance of the effect of age cannot be tested; since bmi can be considered as a surrogate of dxa, after dxa is conditioned, the indicator R is independent of bmi, which means assumption (1.1) holds; it seems common to believe that the missingness of dxa is related to age, although we are not sure if it is true after dxa is conditioned.

Acknowledgment

We thank an associate editor and two referees for their helpful comments. Fang Fang's research was partially supported by Shanghai Nature Science Foundation 15ZR1410300, Program of Shanghai Subject Chief Scientist (14XD 1401600), Shanghai Rising Star Program (16QA1401700), National Scientific Foundation of China (11601156), and the 111 Project (B14019). Jiwei Zhao's research was partially supported by the US National Institutes of Health award UL1TR001412. Jun Shao's research was partially supported by the 111 Project (B14019) and the US National Science Foundation grant DMS-1305474.

Appendix: Proofs of Theorems

Proof of Proposition 1

(1) Under (a), we have

$$\begin{aligned}\tilde{\xi} &= \arg\max_{\xi'} \frac{1}{N}\sum_{i=1}^{N} H_i(\xi', \hat{F}) \\ &= \arg\max_{\xi'} \left[\left\{ \frac{1}{N}\sum_{i=1}^{N} H_i(\xi', \hat{F}) - EH(\xi', F')|_{F'=\hat{F}} \right\} + EH(\xi', F')|_{F'=\hat{F}} \right] \\ &= o_p(1) + \arg\max_{\xi'} [\{EH(\xi', F')|_{F'=\hat{F}} - EH(\xi', F)\} + EH(\xi', F)] \\ &= o_p(1) + \arg\max_{\xi'} E\{R \log p(Z|Y, U, \xi', F) - R \log p(Z|F)\} \\ &= o_p(1) + \xi.\end{aligned}$$

Let $l(\xi', \hat{F}) = (1/N)\sum_{i=1}^{N} H_i(\xi', \hat{F})$. Under (b) and by Taylor's expansion, we have

$$0 = \nabla_{\xi} l(\tilde{\xi}, \hat{F}) = \nabla_{\xi} l(\xi, \hat{F}) + E\{\nabla^2_{\xi\xi} H(\xi, F)\}(\tilde{\xi} - \xi) + o_p(N^{-1/2}). \tag{A.1}$$

Using the theory of V-statistics and similar arguments to Zhao and Shao (2015), we have

$$\nabla_\xi l(\xi, \hat{F}) - \nabla_\xi l(\xi, F) = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} S_1(z_i, \xi, F) \\ T_1(z_i, \xi, F)\end{pmatrix} + o_p(N^{-1/2}), \tag{A.2}$$

where

$$S_1(z_i, \xi, F) = E\left[\frac{r\int \nabla_\theta p(y|u,z,\theta)p(u|z,\gamma)dF(z)p(y|u,z_i,\theta)p(u|z_i,\gamma)}{\{\int p(y|u,z,\theta)p(u|z,\gamma)dF(z)\}^2} - \frac{r\nabla_\theta p(y|u,z_i,\theta)p(u|z_i,\gamma)}{\int p(y|u,z,\theta)p(u|z,\gamma)dF(z)}\right], \tag{A.3}$$

$$T_1(z_i, \xi, F) = E\left[\frac{r\int p(y|u,z,\theta)\nabla_\gamma p(u|z,\gamma)dF(z)p(y|u,z_i,\theta)p(u|z_i,\gamma)}{\{\int p(y|u,z,\theta)p(u|z,\gamma)dF(z)\}^2} - \frac{rp(y|u,z_i,\theta)\nabla_\gamma p(u|z_i,\gamma)}{\int p(y|u,z,\theta)p(u|z,\gamma)dF(z)}\right], \tag{A.4}$$

and the expectation is taken with respect to (r, y, u). Then by (A.1) and (A.2), we have

$$\begin{aligned}\sqrt{N}(\tilde{\xi} - \xi) &= \big[-E\{\nabla^2_{\xi\xi}H(\xi, F)\}\big]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\begin{pmatrix}\nabla_\theta H_i(\xi, F) + S_1(z_i, \xi, F) \\ \nabla_\gamma H_i(\xi, F) + T_1(z_i, \xi, F)\end{pmatrix} + o_p(1) \\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N} D_i + o_p(1) \to_d N(0, \Lambda),\end{aligned}$$

where

$$D_i = \begin{pmatrix} D_{i1} \\ D_{i2} \\ D_{i3}\end{pmatrix} = \big[-E\{\nabla^2_{\xi\xi}H(\xi, F)\}\big]^{-1}\begin{pmatrix}\nabla_\theta H_i(\xi, F) + S_1(z_i, \xi, F) \\ \nabla_\gamma H_i(\xi, F) + T_1(z_i, \xi, F)\end{pmatrix}, \tag{A.5}$$

D_{i1}, D_{i2}, and D_{i3} are the rows of D_i corresponding to $\tilde{\alpha}$, $\tilde{\beta}$ and $\tilde{\gamma}$ respectively, and $\Lambda = \mathrm{Var}(D_i)$.

(2) Direct calculation shows

$$E\{\nabla^2_{\alpha\alpha}H(\xi, F)\} = E[E(r|y,u)E\{\nabla^2_{\alpha\alpha}\log p(y|u,z,\theta) - J|y,u\}], \tag{A.6}$$

where

$$J = \frac{\int \nabla^2_{\alpha\alpha}p(y|u,z,\theta)p(u|z,\gamma)dF(z)}{\int p(y|u,z,\theta)p(u|z,\gamma)dF(z)} - \frac{\int \nabla_\alpha p(y|u,z,\theta)p(u|z,\gamma)dF(z)\int \nabla_{\alpha^\tau} p(y|u,z,\theta)p(u|z,\gamma)dF(z)}{\{\int p(y|u,z,\theta)p(u|z,\gamma)dF(z)\}^2}.$$

As $\beta \to 0$, we have $p(y|u,z,\theta) \to \exp(y\eta(\alpha_c + \alpha_u^\tau u) - b\circ\eta(\alpha_c + \alpha_u^\tau u) + c(y))$,

which does not depend on z. Also,

$$\nabla_\alpha \log p(y|u,z,\theta) = \frac{\nabla h}{\nabla^2_{\eta\eta} b \circ \eta}(\alpha_c + \alpha_u^\tau u + \beta^\tau z)\{y - h(\alpha_c + \alpha_u^\tau u + \beta^\tau z)\}\begin{pmatrix}1\\ u\end{pmatrix}$$
$$\to \frac{\nabla h}{\nabla^2_{\eta\eta} b \circ \eta}(\alpha_c + \alpha_u^\tau u)\{y - h(\alpha_c + \alpha_u^\tau u)\}\begin{pmatrix}1\\ u\end{pmatrix},$$

$$\begin{aligned}
&\nabla^2_{\alpha\alpha} \log p(y|u,z,\theta)\\
&= \nabla\left(\frac{\nabla h}{\nabla^2_{\eta\eta} b \circ \eta}\right)(\alpha_c + \alpha_u^\tau u + \beta^\tau z)\{y - h(\alpha_c + \alpha_u^\tau u + \beta^\tau z)\}\begin{pmatrix}1 & u^\tau\\ u & uu^\tau\end{pmatrix}\\
&\quad - \frac{\nabla h}{\nabla^2_{\eta\eta} b \circ \eta} \circ \nabla h(\alpha_c + \alpha_u^\tau u + \beta^\tau z)\begin{pmatrix}1 & u^\tau\\ u & uu^\tau\end{pmatrix}\\
&\to \nabla\left(\frac{\nabla h}{\nabla^2_{\eta\eta} b \circ \eta}\right)(\alpha_c + \alpha_u^\tau u)\{y - h(\alpha_c + \alpha_u^\tau u)\}\begin{pmatrix}1 & u^\tau\\ u & uu^\tau\end{pmatrix}\\
&\quad - \frac{\nabla h}{\nabla^2_{\eta\eta} b \circ \eta} \circ \nabla h(\alpha_c + \alpha_u^\tau u)\begin{pmatrix}1 & u^\tau\\ u & uu^\tau\end{pmatrix},
\end{aligned}$$

which do not depend on z. The definition of h is given in Section 2.2. So as $\beta \to 0$, $\nabla_\alpha p(y|u,z,\theta)$ and $\nabla^2_{\alpha\alpha} p(y|u,z,\theta)$ do not depend on z, and

$$J \to \frac{\nabla^2_{\alpha\alpha} p(y|u,z,\theta)}{p(y|u,z,\theta)} - \frac{\nabla_\alpha p(y|u,z,\theta)\nabla_{\alpha^\tau} p(y|u,z,\theta)}{p^2(y|u,z,\theta)} = \nabla^2_{\alpha\alpha} \log p(y|u,z,\theta)$$

which, together with (A.6), finishes the proof.

Proof of Theorem 2

For any $\epsilon > 0$, take

$$A_N = \left\{\omega \in \Omega : N^{-1}\left|l(\alpha'|\alpha_1'') - l(\alpha'|\alpha_2'')\right| \le K\|\alpha_1'' - \alpha_2''\|, \text{ for any } \alpha', \alpha_1'', \alpha_2'' \in B\right\},$$

where $K = 2E\{\sup_{\alpha',\alpha'' \in B} \|\nabla_{\alpha''} l_i(\alpha'|\alpha'')\|\}$ and Ω is the sample space for the random variables. There exist $\alpha^* \in B$ between α_1'' and α_2'' such that

$$\begin{aligned}
&P\left(N^{-1}\left|l(\alpha'|\alpha_1'') - l(\alpha'|\alpha_2'')\right| \le K\|\alpha_1'' - \alpha_2''\|\right)\\
&= P\left(N^{-1}\left|\nabla_{\alpha''} l(\alpha'|\alpha^*)(\alpha_1'' - \alpha_2'')\right| \le K\|\alpha_1'' - \alpha_2''\|\right)\\
&\ge P\left(N^{-1}\sum_{i=1}^N \sup_{\alpha',\alpha'' \in B} \|\nabla_{\alpha''} l_i(\alpha'|\alpha'')\| \le K\right)\\
&\ge P\left(\left|N^{-1}\sum_{i=1}^N \sup_{\alpha',\alpha'' \in B} \|\nabla_{\alpha''} l_i(\alpha'|\alpha'')\| - \mathrm{E}\sup_{\alpha',\alpha'' \in B} \|\nabla_{\alpha''} l_1(\alpha'|\alpha'')\|\right| \le \frac{K}{2}\right)\\
&\to 1.
\end{aligned}$$

So we have $P(A_N) \to 1$, as $N \to \infty$.

Since $E\{S_i(\alpha'|\alpha)\big|_{\alpha'=\alpha}\} = 0$, there exists $\bar{\alpha}$ such that $S(\bar{\alpha}|\alpha) = 0$ and $\bar{\alpha}$ converges to α in probability. Then for large enough N, $\bar{\alpha} \in B$ with arbitrary large probability. Since $l(\alpha'|\alpha)$ is strictly concave in B and $\nabla_{\alpha'} l(\alpha'|\alpha) = S(\alpha'|\alpha)$, $\bar{\alpha}$ is the unique maximizer of $l(\alpha'|\alpha)$ in B. Let

$$\tau_N = N^{-1}\left\{ l(\bar{\alpha}|\alpha) - \sup_{\alpha' \in B, \|\alpha' - \bar{\alpha}\| \geq \epsilon/2} l(\alpha'|\alpha) \right\},$$

$$B_N = \{\omega \in \Omega : \tau_N(\omega) \geq \tau/2\},$$

$$\tau = \mathrm{E}\{l_i(\alpha|\alpha)\} - \sup_{\alpha' \in B, \|\alpha' - \alpha\| \geq \epsilon/2} \mathrm{E}\{l_i(\alpha'|\alpha)\}.$$

Since $\bar{\alpha}$ converges to α in probability, τ_N converges to τ in probability as B is compact. Additionally, τ is a positive constant, guaranteed by (b), and $E\{l_i(\alpha'|\alpha)\}$ is strictly concave in a neighborhood of α with $\nabla_{\alpha'} E\{l_i(\alpha'|\alpha)\}|_{\alpha'=\alpha} = \mathrm{E}\{S_i(\alpha|\alpha)\} = 0$, so $P(B_N) \to 1$ as $N \to \infty$.

For any fixed t, if $\hat{\alpha}^{(t)}$ converges to α in probability, then for large enough N, $\hat{\alpha}^{(t)} \in B$ with arbitrary large probability. For any $\omega \in A_N \bigcap B_N$, once $\|\hat{\alpha}^{(t)} - \alpha\| \leq (\tau/8)K$, we have

$$N^{-1}\left| l(\alpha'|\hat{\alpha}^{(t)}) - l(\alpha'|\alpha) \right| \leq K\|\hat{\alpha}^{(t)} - \alpha\| \leq \frac{\tau}{8} \text{ for any } \alpha' \in B. \tag{A.7}$$

Hence, when $\alpha' \in B$ and $\|\alpha' - \bar{\alpha}\| \geq \epsilon/2$, we have

$$\begin{aligned} N^{-1} l(\alpha'|\hat{\alpha}^{(t)}) &\leq N^{-1} l(\alpha'|\alpha) + \frac{\tau}{8} && \text{(A.8)} \\ &\leq N^{-1} \sup_{\alpha' \in B, \|\alpha' - \bar{\alpha}\| \geq \epsilon/2} l(\alpha'|\alpha) + \frac{\tau}{8} \\ &= N^{-1} l(\bar{\alpha}|\alpha) - \tau_N(\omega) + \frac{\tau}{8} \\ &\leq N^{-1} l(\bar{\alpha}|\alpha) - \frac{\tau}{2} + \frac{\tau}{8} && \text{(A.9)} \\ &\leq N^{-1} l(\bar{\alpha}|\hat{\alpha}^{(t)}) + \frac{\tau}{8} - \frac{\tau}{2} + \frac{\tau}{8} && \text{(A.10)} \\ &= N^{-1} l(\bar{\alpha}|\hat{\alpha}^{(t)}) - \frac{\tau}{4}, \end{aligned}$$

where (A.8) and (A.10) follow from (A.7), and (A.9) holds since $\omega \in B_N$. Therefore, there exists a local maximizer of $l(\alpha'|\hat{\alpha}^{(t)})$, denoted as $\hat{\alpha}^{(t+1)}$, such that $S(\hat{\alpha}^{(t+1)}|\hat{\alpha}^{(t)}) = 0$ and $\|\hat{\alpha}^{(t+1)} - \bar{\alpha}\| < \epsilon/2$ when $\omega \in A_N \bigcap B_N$ and $\|\hat{\alpha}^{(t)} - \alpha\| \leq \tau/8K$. Hence,

$$\begin{aligned} P\left(\|\hat{\alpha}^{(t+1)} - \alpha\| < \epsilon\right) &\geq P\left(\|\hat{\alpha}^{(t+1)} - \bar{\alpha}\| < \frac{\epsilon}{2}, \|\bar{\alpha} - \alpha\| < \frac{\epsilon}{2}\right) \\ &\geq P\left(\omega \in A_N \bigcap B_N, \|\hat{\alpha}^{(t)} - \alpha\| \leq \frac{\tau}{8K}, \|\bar{\alpha} - \alpha\| < \frac{\epsilon}{2}\right) \end{aligned}$$

$$\to 1, \tag{A.11}$$

where (A.11) holds since $P(A_N) \to 1$, $P(B_N) \to 1$, and both $\hat{\alpha}^{(t)}$ and $\bar{\alpha}$ converge to α in probability. So when $\hat{\alpha}^{(1)}$ converges to α in probability, there exists a sequence $\{\hat{\alpha}^{(t)}, t \geq 1\}$ such that $S(\hat{\alpha}^{(t+1)}|\hat{\alpha}^{(t)}) = 0$ and $\hat{\alpha}^{(t)}$ converges to α in probability as N goes to infinity.

Since $E\{S_i(\alpha|\alpha)\} = 0$, as N goes to infinity, there exists a sequence $\{\bar{\alpha}_N\}$ such that $P(S(\bar{\alpha}_N|\bar{\alpha}_N) = 0) \to 1$ and $\bar{\alpha}_N$ converges to α in probability. We can use Taylor's expansion technique on $S(\hat{\alpha}^{(t+1)}|\hat{\alpha}^{(t)})$ around the point $(\bar{\alpha}_N, \bar{\alpha}_N)$:

$$\begin{aligned} 0 &= S(\hat{\alpha}^{(t+1)}|\hat{\alpha}^{(t)}) \\ &= S(\bar{\alpha}_N|\bar{\alpha}_N) + \nabla_{\alpha'} S(\alpha'^*|\alpha''^*)(\hat{\alpha}^{(t+1)} - \bar{\alpha}_N) + \nabla_{\alpha''} S(\alpha'^*|\alpha''^*)(\hat{\alpha}^{(t)} - \bar{\alpha}_N), \end{aligned}$$

where α'^* is between $\hat{\alpha}^{(t+1)}$ and $\bar{\alpha}_N$, and α''^* is between $\hat{\alpha}^{(t)}$ and $\bar{\alpha}_N$. Since both α'^* and α''^* converge to α in probability, we have

$$\begin{aligned} \hat{\alpha}^{(t+1)} - \bar{\alpha}_N &= \Big\{ - N^{-1} \nabla_{\alpha'} S(\alpha'^*|\alpha''^*) \Big\}^{-1} \Big\{ N^{-1} \nabla_{\alpha''} S(\alpha'^*|\alpha''^*) \Big\} \Big(\hat{\alpha}^{(t)} - \bar{\alpha}_N \Big) \\ &\cong \big\{ -N^{-1} \nabla_{\alpha'} S(\alpha'|\alpha) \big|_{\alpha'=\alpha} \big\}^{-1} \big\{ N^{-1} \nabla_{\alpha''} S(\alpha|\alpha'') \big|_{\alpha''=\alpha} \big\} \Big(\hat{\alpha}^{(t)} - \bar{\alpha}_N \Big), \end{aligned}$$

where the notation $C_N \cong D_N$ means $D_N^{-1} C_N = 1 + o_p(1)$. Therefore,

$$\hat{\alpha}^{(t)} - \bar{\alpha}_N \cong \big\{ I_1(\alpha)^{-1} I_2(\alpha) \big\}^{t-1} (\hat{\alpha}^{(1)} - \bar{\alpha}_N),$$

which converges to 0 as $t \to \infty$ with condition (c). This finishes the proof with $\hat{\alpha} = \bar{\alpha}_N$.

Proof of Theorem 3

Let $S(\alpha', \beta', \gamma') = \sum_{i=1}^N S_i(\alpha', \beta', \gamma')$. Under the regularity conditions and by Taylor's expansion, we have

$$0 = S\frac{(\hat{\alpha}, \tilde{\beta}, \tilde{\gamma})}{N} = S\frac{(\alpha, \beta, \gamma)}{N} + B_1(\hat{\alpha} - \alpha) + B_2(\tilde{\beta} - \beta) + B_3(\tilde{\gamma} - \gamma) + o_p(N^{-1/2}).$$

Therefore

$$\begin{aligned} \sqrt{N}(\hat{\alpha} - \alpha) &= (-B_1)^{-1} \left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^N S_i(\alpha, \beta, \gamma) + B_2 \frac{1}{\sqrt{N}} \sum_{i=1}^N D_{i2} \right. \\ &\quad \left. + B_3 \frac{1}{\sqrt{N}} \sum_{i=1}^N D_{i3} \right\} + o_p(1), \sqrt{N}(\hat{\theta} - \theta) \\ &= \sqrt{N} \begin{pmatrix} \hat{\alpha} - \alpha \\ \tilde{\beta} - \beta \end{pmatrix} = \frac{1}{\sqrt{N}} \sum_{i=1}^N E_i + o_p(1) \to_d N(0, \Sigma), \end{aligned}$$

where

$$E_i = \begin{bmatrix} (-B_1)^{-1}\{S_i(\alpha,\beta,\gamma) + B_2 D_{i2} + B_3 D_{i3}\} \\ D_{i2} \end{bmatrix}, \tag{A.12}$$

and $\Sigma = \mathrm{Var}(E_i)$.

Proof of Theorem 4

Let $P(w)$ denote the true distribution of $W = (Y, U, Z, R)$ and $\hat{P}(w)$ its empirical distribution based on data $w_i, i = 1, \ldots, N$. Then

$$\Sigma = \mathrm{Var}(E_i) = \int \kappa(w,\xi,F,A,B_1,B_2,B_3)\kappa(w,\xi,F,A,B_1,B_2,B_3)^\tau dP(w),$$

$$\hat{\Sigma} = \int \kappa(w,\hat{\xi},\hat{F},\hat{A},\hat{B}_1,\hat{B}_2,\hat{B}_3)\kappa(w,\hat{\xi},\hat{F},\hat{A},\hat{B}_1,\hat{B}_2,\hat{B}_3)^\tau d\hat{P}(w).$$

Let

$$\begin{aligned} Q_i = {} & \kappa(w,\hat{\xi},\hat{F},\hat{A},\hat{B}_1,\hat{B}_2,\hat{B}_3)\kappa(w,\hat{\xi},\hat{F},\hat{A},\hat{B}_1,\hat{B}_2,\hat{B}_3)^\tau \\ & -\kappa(w,\xi,F,A,B_1,B_2,B_3)\kappa(w,\xi,F,A,B_1,B_2,B_3)^\tau. \end{aligned}$$

From the triangular inequality and law of large numbers,

$$\|\hat{\Sigma} - \Sigma\| \le \frac{1}{N}\|Q_i\| + o_p(1).$$

By (3.1), for any $\epsilon > 0$,

$$\frac{1}{N}\sum_{i=1}^{N} \|Q_i\| I_{[0,c]}(\|w_i\|) < \frac{\epsilon}{2}$$

when N is sufficiently large.

For any $\tilde{\epsilon} > 0$, we can choose c such that $E\{h(w)I_{(c,\infty)}(\|w\|)\} < \epsilon\tilde{\epsilon}/4$. By Chebyshev's inequality and (3.2),

$$P\left(\frac{1}{N}\sum_{i=1}^{N} \|Q_i\| I_{(c,\infty)}(\|w_i\|) > \frac{\epsilon}{2}\right) < \tilde{\epsilon}.$$

Therefore

$$P\left(\frac{1}{N}\sum_{i=1}^{N} \|Q_i\| > \epsilon\right) \to 0,$$

which completes the proof.

References

Bashir, S. and Duffy, S. (1997). The correction of risk estimates for measurement error. *Ann. Epidemol.* **7**, 154–164.

Herring, A. H. and Ibrahim, J. G. (2002). Maximum likelihood estimation in random effects

cure rate models with nonignorable missing covariates. *Biostatistics* **3**, 387–405.

Horton, N. J. and Laird, N. M. (2001). Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information. *Biometrics* **57**, 34–42.

Huang, L., Chen, M.-H. and Ibrahim, J. G. (2005). Bayesian analysis for generalized linear models with nonignorably missing covariates. *Biometrics* **61**, 767–780.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R. and Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *J. Amer. Statist. Assoc.* **100**, 332–346.

Ibrahim, J. G., Lipsitz, S. R. and Chen, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61**, 173–190.

Ibrahim, J. G. and Molenberghs (2009). Missing data methods in longitudinal studies: a review. *Test* **18**, 1–43.

Lipsitz, S. R., Ibrahim, J. G., Chen, M. H. and Peterson, H. (1999). Non-ignorable missing covariates in generalized linear models. *Stat. Med.* **18**, 2435–2448.

Lipsitz, S. R., Ibrahim, J. G. and Zhao, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *J. Amer. Statist. Assoc.* **94**, 1147–1160.

Little, R. J. (1992). Regression with missing X's: a review. *J. Amer. Statist. Assoc.* **87**, 1227–1237.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data* 2nd Edition Wiley, New York.

Qin, J., Zhang, B. and Leung, D. H. (2009). Empirical likelihood in missing data problems. *J. Amer. Statist. Assoc.* **104**, 1492–1503.

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* **82**, 299–314.

Robins, J. M., Hsieh, F. and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **57**, 409–424.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Stat. Med.* **16**, 285–319.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846–866.

Stubbendick, A. L. and Ibrahim, J. G. (2003). Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics* **59**, 1140–1150.

Stubbendick, A. L. and Ibrahim, J. G. (2006). Likelihood-based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. *Statist. Sinica* **16**, 1143–1167.

Tang, G., Little, R. J. and Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika* **90**, 747–764.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer.

Wang, S., Shao, J. and Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statist. Sinica* **24**, 1097–1116.

Zhao, J. and Shao, J. (2015). Semiparametric pseudo likelihoods in generalized linear models with nonignorable missing data. *J. Amer. Statist. Assoc.* **110**, 1577–1590.

School of Statistics, East China Normal University, 500 Dongchuan Road, Shanghai, 200241, China.

E-mail: ffang@sfs.ecnu.edu.cn

Department of Biostatistics, State University of New York at Buffalo, Buffalo, NY 14214, USA.

E-mail: E-mail: zhaoj@buffalo.edu

Department of Statistics, University of Wisconsin - Madison, 1300 University Ave., Madison, WI, 53706, USA.

E-mail: shao@stat.wisc.edu

(Received December 2015; accepted August 2016)