Published in final edited form as:

Scand Stat Theory Appl. 2009 December 1; 36(4): 713–734. doi:10.1111/j.1467-9469.2009.00646.x.

Estimation and inference based on Neumann series approximation to locally efficient score in missing data problems

HUA YUN CHEN

Division of Epidemiology & Biostatistics, School of Public Health, University of Illinois at Chicago, Chicago, IL 60532

HUA YUN CHEN: hychen@uic.edu

Abstract

Theory on semiparametric efficient estimation in missing data problems has been systematically developed by Robins and his coauthors. Except in relatively simple problems, semiparametric efficient scores cannot be expressed in closed forms. Instead, the efficient scores are often expressed as solutions to integral equations. Neumann series was proposed in the form of successive approximation to the efficient scores in those situations. Statistical properties of the estimator based on the Neumann series approximation are difficult to obtain and as a result, have not been clearly studied. In this paper, we reformulate the successive approximation in a simple iterative form and study the statistical properties of the estimator based on the reformulation. We show that a doubly-robust locally-efficient estimator can be obtained following the algorithm in robustifying the likelihood score. The results can be applied to, among others, the parametric regression, the marginal regression, and the Cox regression when data are subject to missing values and the missing data are missing at random. A simulation study is conducted to evaluate the performance of the approach and a real data example is analyzed to demonstrate the use of the approach.

Keywords

auxiliary covariates; information operator; non-monotone missing pattern; weighted estimating equations

1 Introduction

The semiparametric efficient estimation for missing data problems has been extensively studied (Bickel *et al.*, 1993; Robins *et al.*, 1994 and others). One major task in such problems is to project the estimating score onto the orthogonal complement of the nuisance score space. However, the projection often depends on the unknown underlying distribution that generated the data (Robins *et al.*, 1994, 1995; Rotnitzky & Robins, 1995; Rotnitzky *et al.*, 1998; Scharfstein *et al.*, 1999). To overcome the difficulty, working models have been proposed to compute a locally efficient score. It has been shown that when data are missing at random and either the working model for the missing data mechanism or the working model for the nuisance model of the full data is correct, the locally efficient score is asymptotically unbiased (Lipsitz *et al.*, 1999; Robins *et al.*, 1999; Robins & Rotnitzky, 2001; van der Laan & Robins, 2003). Note that, except in simple cases, computing the projection using working models still corresponds to the hard problem of solving an integral equation. Neumann series expansion has been proposed to obtain an approximate solution through successive approximation (Robins *et al.*, 1994, Robins & Wang, 1998, and van der Laan & Robins, 2003). Since the procedure for finding the locally efficient estimator based on the approximate locally efficient score is complicated, the study of the asymptotic properties of the estimator has been left open.

In this article, we reformulate the successive approximation and show that an algorithm based on robustifying the likelihood score yields an estimator having the desired asymptotic properties, *i.e.*, doubly robust and locally efficient.

The remainder of the article is organized as follows. In section 2, we reformulate the successive approximation in a simple form and show the robust property of the algorithm. The asymptotic properties of the estimator are carefully studied in section 3. We show that the algorithm indeed yields an estimator which is doubly-robust and locally-efficient when appropriate care is taken with regard to the number of terms used in the Neumann series approximation. Applications of the theory developed to regression models are briefly discussed in section 4. A simulation study is performed using parametric regression with missing covariates to examine the finite sample performance of the algorithm in section 5 and the algorithm is applied to a real data example. The article is concluded with some discussions in section 6. All proofs are collected in the Appendix.

2 The Neumann series approximation to the efficient score

Let Y be the full data, R be the missing data indicator for Y , and $R(Y)$ and $\bar{R}(Y)$ be respectively the observed and missing parts of Y . Let the density of the distribution for (R, Y) with respect to μ , a product of count measures and Lebesgue measures, be $dP_{(\alpha, \beta, \theta)}/d\mu = \pi(R|Y, \alpha)f(Y, \beta, \theta)$, where $\beta \in \Omega$, $\theta \in \Theta$, $\alpha \in \Xi$. Here β and α are usually Euclidean parameters, and θ is a nuisance parameter, which can be of infinite dimension. Let $\eta = (\alpha, \beta, \theta)$, where β is the parameter of interest and (α, θ) are nuisance parameters. Let $b(Y) \in L_0^2(P_\eta)$, a mean-zero square-integrable function with respect to P_η . Define the nonparametric information operator m_η as

$$m_\eta\{b(Y)\} = E_\eta[E_\eta\{b(Y)|R, R(Y)\}|Y].$$

Neumann series approximation to the efficient score appeared as the successive approximation in Robins *et al.* (1994). The method first finds the efficient score for β under the full data model, denoted by $S_\beta^{F, eff}$. The method then employs the successive approximation,

$$U_N = S_\beta^{F, eff} + \mathcal{P}_\eta(I - m_\eta)U_{N-1},$$

where $\mathcal{P}_\eta$ is the projection to the closure of the nuisance score space of the full data model and $U_0 = S_\beta^{F, eff}$. The semiparametric efficient score is $E_\eta\{U_\infty|R, R(Y)\}$. There are many unanswered questions associated with the use of this approach in practice. First, we need guidelines for choosing a finite N in implementing the algorithm. Second, the successive approximation is only known to converge under L^2 -norm, which is insufficient in studying the properties of the estimator when the underlying distribution used in computing the projection is estimated rather than known. Third, it is not known if the estimator generated from the procedure involving approximations is indeed asymptotically equivalent to the estimator with known underlying distribution.

To answer these questions, we reformulate the successive approximation in another form as $U_N = (I - \mathcal{P}_\eta m_\eta)U_{N-1}$, or equivalently as an explicit expression:

$$U_N = (I - \mathcal{P}_\eta m_\eta)^N S_\beta,$$

where S_β is the likelihood score for β . The equivalence of the simpler form to the successive approximation can be shown by noting that $(I - \mathcal{P}_\eta)\mathcal{P}_\eta = 0$ and $S_\beta^{F,eff} = (I - \mathcal{P}_\eta)S_\beta$. This allows us to express the efficient score for the missing data problem directly as

$$\lim_{N \rightarrow \infty} E\{(I - \mathcal{P}_\eta m_\eta)^N S_\beta | R, R(Y)\}.$$

Note that the approximation based on the new expression does not require us to first find the efficient score under the full data. The approximation is the likelihood score when $N = 0$ and can be regarded as robustification of the likelihood score when $N > 0$. An algorithm for finding the approximate locally efficient score for estimating β based on the expression can be described as follows. First, find an estimate of the nuisance parameters under the working models using methods such as the maximum likelihood approach. Then, compute the approximate efficient score with the nuisance parameters estimated from the working models. Finally, solve the score equation to obtain β estimator. Results in the next section show that it is sufficient that the number N in the approximation be taken in an order higher than the logarithm of the sample size. Let

$$T_N\{R, R(Y), \eta\} = E\{(I - \mathcal{P}_\eta m_\eta)^N S_\beta | R, R(Y)\}$$

and $T\{R, R(Y), \eta\}$ be the limit of $T_N\{R, R(Y), \eta\}$ under $L^2(P_\eta)$. The large sample properties of the approximate locally efficient estimator are studied in the next section. The following proposition describes the robust properties of the estimating scores, which are used in the next section. Similar results can be found in van der Laan & Robins (2003, sections 2.9 & 2.10). Note, however, that part (a) of the proposition is different from their results.

Let $P_0 = P_{(\beta_0, \theta_0, \alpha_0)}$, the true distribution generating the data. If, for any small ε and large M , there exists $\theta_{(\varepsilon, M)}$ such that

$$f(y, \beta_0, \theta) \left(1 + \varepsilon \left[S(y) 1_{\|S(y)\| \leq M} - E_{(\beta_0, \theta)} \{ S(Y) 1_{\|S(Y)\| \leq M} \} \right] \right) = f(y, \beta_0, \theta_{\varepsilon, M}),$$

where $S(Y) = f(Y, \beta_0, \theta_0) / f(Y, \beta_0, \theta) - 1$, then, in this paper, $\{f(y, \beta_0, \theta), \theta \in \Theta\}$ is called a super-convex family for θ at β_0 . Note that a super-convex family of distributions is always a convex family of distributions, which corresponds to $M = \infty$. When the densities are bounded above from infinity and below from zero, a convex family is also a super-convex family.

Proposition 1

Assume that data are missing at random and that the true distribution generating the data is P_0 . Then the following results hold:

- a. For any fixed N , if the nuisance model for the full data is correct, i.e., $\theta = \theta_0$, then $T_N\{R, R(Y), \beta_0, \theta_0, \alpha\}$ is asymptotically unbiased under P_0 . The $L^2(P_{(\beta_0, \theta_0, \alpha)})$ limit of T_N is also asymptotically unbiased if it is in $L^2(P_0)$ and the missing data probabilities are bounded away from zero.
- b. If the model for the missing data mechanism is correct, i.e., $\alpha = \alpha_0$, and $f(y, \beta_0, \theta)$ is a super-convex family with respect to θ , then $E_0[T\{R, R(Y), \beta_0, \theta, \alpha_0\}] = 0$ if $E_0[T^{\otimes 2}\{R, R(Y), \beta_0, \theta, \alpha_0\}] < \infty$, where $T^{\otimes 2} = TT'$.

We suppress the proof of this proposition because it is similar to the proof of such results in the literature such as in Robins *et al.* (2000) and van der Laan & Robins (2003). The proposition suggests that T is doubly robust, *i.e.*, it is unbiased when either the missing data mechanism model or the nuisance model for the full data is correctly specified. For a fixed N , T_N is unbiased only when the nuisance model for the full data is correct. However, note that T_N approximates T . If we allow N to depend on the sample size, we can make T_{N_n} doubly robust as $n \rightarrow \infty$. We explain in the next section how to implement this idea.

3 Estimation and inference based on approximate locally efficient scores

To simplify notation in this section, we use θ to denote the nuisance parameter. That is, we absorb α into θ . Denote the model $\int \pi(R|Y, \alpha) f(Y, \beta, \theta) d\bar{K}(Y)$ by $g(R, R(Y), \beta, \theta)$ after the parameter absorption. Let $\theta(\gamma)$, $\gamma \in \Gamma$ define a working model which is a regular submodel. Let $\theta(\hat{\gamma})$ be a $\sqrt{n}$ consistent estimator of $\theta(\gamma)$ under the working models.

To accommodate β of infinite dimension, we use the functional form to denote T_N and T . Let $T_{N(i)}(\eta)(h_1) = T_N\{R_i, R_i(Y_i), \eta\}(h_1)$ and $T_{(i)}(\eta)(h_1) = T\{R_i, R_i(Y_i), \eta\}(h_1)$ where $h_1 \in H_1$, a Hilbert space. Let $\hat{\theta} = \theta(\hat{\gamma})$. For a given N , let $\hat{\beta}_N$ be the solution to the equation

$$P_n T_N(\hat{\beta}_N, \hat{\theta})(h_1) = \frac{1}{n} \sum_{i=1}^n T_{N(i)}(\hat{\beta}_N, \hat{\theta})(h_1) = 0,$$

for all $h_1 \in H_1$. Let $\tilde{\beta}$ be the solution to the equation

$$P_n T(\tilde{\beta}, \hat{\theta})(h_1) = \frac{1}{n} \sum_{i=1}^n T_{(i)}(\tilde{\beta}, \hat{\theta})(h_1) = 0,$$

for all $h_1 \in H_1$. Define linear operator $\mathcal{Q}_0$ as a map from H_1 to itself satisfying

$$\langle \mathcal{Q}_0 h_1, h_1^* \rangle_{H_1} = E_0 \left\{ \frac{\partial T(\eta_0)}{\partial \beta}(h_1)(h_1^*) \right\}. \quad (1)$$

Assumption 9 in the Appendix guarantees that $\mathcal{Q}_0$ exists, is uniquely defined, and is continuously invertible because of the following. For any given h_1 , the right-hand side of the foregoing equation defines a linear functional on H_1 . By Riesz representation theorem, there exists an $h^{**} \in H_1$ such that, for all $h_1^* \in H_1$,

$$\langle h_1^{**}, h_1^* \rangle_{H_1} = P_0 \left\{ \frac{\partial T(\eta_0)}{\partial \beta}(h_1)(h_1^*) \right\}.$$

We can thus define the map $\mathcal{Q}_0$ from H_1 to H_1 such that $\mathcal{Q}_0 h_1 = h_1^{**}$. By varying $h_1 \in H_1$, we see that $\mathcal{Q}_0$ is well-defined on H_1 . It is straightforward to verify that $\mathcal{Q}_0$ is a linear operator on H_1 . Similarly, we define linear operators, $\mathcal{Q}_N$ and $\mathcal{Q}_{0N}$, that map H_1 to itself, respectively as

$$\langle Q_N h_1, h_1^* \rangle_{H_1} = P_0 \left\{ \frac{\partial T_N(\beta_N, \theta^*)}{\partial \beta}(h_1)(h_1^*) \right\}.$$

and

$$\langle Q_{0N} h_1, h_1^* \rangle_{H_1} = P_0 \left\{ \frac{\partial T_N(\beta_0, \theta^*)}{\partial \beta}(h_1)(h_1^*) \right\},$$

which are respectively continuously invertible from assumption 10 in the Appendix and from the continuity of the right-hand side with respect to β .

We now state theorems on the asymptotic properties of the β estimators when data are missing at random and either the missing data mechanism model or the nuisance full data model is correctly specified. Theorem 1 describes the asymptotic behavior of $\tilde{\beta}$ when $n \rightarrow \infty$. Theorem 2 states the asymptotic property of $\tilde{\beta}_N$ when N is fixed and $n \rightarrow \infty$. Theorem 3 describes the asymptotic behavior of $\tilde{\beta}_N$ when $n \rightarrow \infty$ and N , as a function of n , also tends to ∞ . Assumptions and proofs are given in the Appendix.

Theorem 1

Under assumptions 1–9, as $n \rightarrow \infty$, $\tilde{\beta} \rightarrow \beta_0$ almost surely and

$$\langle \sqrt{n}(\tilde{\beta} - \beta_0), h_1 \rangle_{H_1} \rightarrow N(0, V_0(h_1)),$$

uniformly for $h_1 \in H_{10}$, and

$$V_0(h_1) = E_0 \{ T(\beta_0, \theta^*) (Q_0^{-1} h_1) \}^{\otimes 2},$$

which can be consistently estimated uniformly for $h_1 \in H_{10}$ by

$$\frac{1}{n} \sum_{i=1}^n \left[T_{(i)}(\tilde{\beta}, \hat{\theta}) \{ \hat{Q}_0^{-1}(h_1) \} \right]^{\otimes 2},$$

where $\theta^* = \theta(\gamma^*)$ and γ^* is the limit of $\hat{\gamma}$, and

$$\langle \hat{Q}_0(h_1), h_1^* \rangle_{H_1} = \frac{1}{n} \sum_{i=1}^n \sqrt{n} \left[T_{(i)}(\tilde{\beta} + h_1 / \sqrt{n}, \hat{\theta})(h_1^*) - T_{(i)}(\tilde{\beta}, \hat{\theta})(h_1^*) \right],$$

for $h_1^* \in H_1$. When the nuisance models for θ are correctly specified, $\theta^* = \theta_0$ and V_0 attains the semiparametric efficient variance bound.

Theorem 2

Under assumptions 1–8, and 10–12, for any fixed N , as $n \rightarrow \infty$, $\tilde{\beta}_N$ converges almost surely to β_N satisfying $E_0\{T_N(\beta_N, \theta^*)(h_1)\} = 0$ for all $h_1 \in H_1$. The asymptotic bias of $\tilde{\beta}_N$ can be approximated by

$$\langle \beta_N - \beta_0, h_1 \rangle_{H_1} \approx E_0\{T_N(\beta_0, \theta^*)(-Q_{0N}^{-1}h_1)\}.$$

Further,

$$\langle \sqrt{n}(\tilde{\beta}_N - \beta_N), h_1 \rangle_{H_1} \rightarrow N(0, V_N(h_1))$$

uniformly over $h_1 \in H_{10}$, where

$$V_N(h_1) = E_0 \left[T_N(\beta_N, \theta^*)(Q_N^{-1}h_1) + E_0 \left\{ \frac{\partial T_N(\beta_N, \theta^*)}{\partial \theta}(u)(Q_N^{-1}h_1) \right\}_{u=U_1} \right]^{\otimes 2}$$

can be consistently estimated by

$$\frac{1}{n} \sum_{i=1}^n \left(T_{N(i)}\{\tilde{\beta}_N, \hat{\theta}\}(\hat{Q}_N^{-1}h_1) + \sqrt{n} \left[\bar{T}_N\{\tilde{\beta}_N, \hat{\theta} + \frac{U_i}{\sqrt{n}}\}(\hat{Q}_N^{-1}h_1) - \bar{T}_N\{\tilde{\beta}_N, \hat{\theta}\}(\hat{Q}_N^{-1}h_1) \right] \right)^{\otimes 2}$$

uniformly for $h_1 \in H_{10}$, where $\hat{Q}_N(h_1)$ is defined as

$$\langle \hat{Q}_N(h_1), h_1^* \rangle_{H_1} = \frac{1}{n} \sum_{i=1}^n T_{N(i)}\{\tilde{\beta}_N + h_1/\sqrt{n}, \hat{\theta}\}(h_1^*)$$

for $h_1, h_1^* \in H_1$, and $U_1, \dots, U_n$ are influence functions of $\hat{\theta}$ in estimating θ^* , i.e.,

$$\hat{\theta} - \theta^* = \frac{1}{n} \sum_{i=1}^n U_i + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Theorem 3

Let N_n be a sequence such that $\log n/N_n \rightarrow 0$ as $n \rightarrow \infty$. Under assumptions 1–9, $\tilde{\beta}_{N_n} - \tilde{\beta} \rightarrow 0$ P_0 -almost surely, and

$$\langle \sqrt{n}(\tilde{\beta}_{N_n} - \tilde{\beta}), h_1 \rangle_{H_1} = o_{P_0}(1)$$

uniformly over $h_1 \in H_{10}$ as $n \rightarrow \infty$. Further, $V_0(h_1)$ can be consistently estimated by

$$\frac{1}{n} \sum_{i=1}^n \left\{ T_{N_n(i)}(\tilde{\beta}_{N_n}, \tilde{\theta}) \right\} (\widehat{Q}_{0N_n}^{-1}(h_1))^{\otimes 2}$$

uniformly over $h_1 \in H_{10}$, where

$$\langle \widehat{Q}_{0N_n}^{-1}(h_1), h_1^* \rangle_{H_1} = \frac{1}{n} \sum_{i=1}^n \sqrt{n} \left[T_{N_n(i)}(\tilde{\beta}_{N_n} + h_1^* / \sqrt{n}, \tilde{\theta})(h_1) - T_{N_n(i)}(\tilde{\beta}_{N_n}, \tilde{\theta})(h_1) \right],$$

for $h_1, h_1^* \in H_1$.

In practice, Theorem 1 is useful only when the locally efficient score has a closed-form expression. This can happen sometimes. In general, Theorem 1 cannot be applied directly because of the unknown form of the locally efficient score. Theorem 2 can almost always be applied to the approximation score with a finite N . It can be seen from Theorem 2 that, although the bias in estimating β cannot be totally avoided, the magnitude of the bias can be controlled by selecting a sufficiently large N . This is because $E_0 T_N(\beta_0, \theta^*)(h_1) \rightarrow E_0 T(\beta_0, \theta^*)(h_1) = 0$ uniformly over $h_1 \in H_{10}$ as $N \rightarrow \infty$. Furthermore, if the nuisance model for the full data is correctly specified in the missing data problem, bias in estimating β is asymptotically zero for any fixed N when $n \rightarrow \infty$ because $E_0 T_N(\beta_0, \theta^*)(h_1) = 0$ from Proposition 1(a). Theorem 3 states that the approximation score with N sufficient large relative to the sample size is asymptotically equivalent to the locally efficient score in estimating β . This confirms that the algorithm indeed finds the locally efficient estimator when carefully implemented.

4 Applications to examples

In this section, we apply the theorems in the previous section to several regression models frequently used in practice.

Example 1: Missing data in parametric regression models

Let $Y = (V, W, X)$ with density $p(v|w, x)f(w|x, \beta)q(x)$ where $f(w|x, \beta)$ is the parametric regression model with a Euclidean parameter $\beta \in R^k$, which is of primary interest, and V is the auxiliary information observed in addition to outcome W and covariates X . The nuisance parameter for the complete data model is (p, q) . We know from Robins *et al.* (1994) that doubly robust estimating scores have the form $E_\eta\{m_\eta^{-1}(\mathcal{D})|R, R(Y)\}$, where $\mathcal{D}$ is in the orthogonal complement of the nuisance score space. Specifically, $\mathcal{D} = S(W, X) - E_\eta\{S(W, X)|X\}$ for any square-integrable function S . The semiparametric efficient score is the special doubly robust estimating score with $\mathcal{D}$ satisfying the integral equation

$$E_\eta\{m_\eta^{-1}(\mathcal{D})|X, W\} - E_\eta\{m_\eta^{-1}(\mathcal{D})|X\} = \frac{\partial}{\partial \beta} \log f(W|X, \beta).$$

When missing data form monotone patterns, m_η^{-1} has a closed-form expression. But the foregoing integral equation does not have a closed-form solution even with the simplest missing data pattern and when no auxiliary covariates are involved. As a result, successive

approximation is needed except when the missing data form monotone patterns and we are satisfied with a doubly robust estimator.

The score operator for the parameters (q, β, p) is

$$A_{\eta}(h_1, h_{21}, h_{22}) = h_1^T \frac{\partial \log f}{\partial \beta} + h_{21}(v, w, x) + h_{22}(x),$$

where $h_1 \in H_1 = R^k$ with k the dimension of β , and

$$h_{21} \in H_{21} = \{h_{21}(v, w, x) \in L^2(P_{\eta}) | E_{\eta}\{h_{21}(V, W, X) | W, X\} = 0\}$$

and

$$h_{22} \in H_{22} = \{h_{22}(x) \in L^2(P_{\eta}) | E_{\eta}\{h_{22}(X)\} = 0\}.$$

Note that H_1 does not vary for different $\beta \in \Omega$. H_{10} can be chosen as $H_{10} = \{h_1 \in R^k | |h_1| \leq 1\}$. Let $H_2 = H_{21} \times H_{22}$. Let $A_{2\eta}(h_{21}, h_{22}) = A_{\eta}(0, h_{21}, h_{22})$. The adjoint operator of $A_{2\eta}$ is

$$A_{2\eta}^* \{s(v, w, x)\} = \{s(v, w, x) - E_{\eta}(s|w, x), E_{\eta}(s|x) - E_{\eta}(s)\}.$$

It follows that $A_{2\eta}^* A_{2\eta}(h_{21}, h_{22}) = (h_{21}(v, w, x), h_{22}(x))$. Hence, $A_{2\eta}^* A_{2\eta}$ is continuously invertible on H_2 and $\mathcal{P}_{\eta} = A_{2\eta}(A_{2\eta}^* A_{2\eta})^{-1} A_{2\eta}^*$ having the form

$$\mathcal{P}_{\eta}(s) = s(v, w, x) - E_{\eta}(s|w, x) + E_{\eta}(s|x),$$

for any mean-zero square integrable function s . Assume that the densities involving the nuisance parameters are uniformly bounded, the convexity requirement for $q(v|w, x)f(w|x, \beta)p(x)$ with respect to qp can be verified as follows.

$$\tau q_1(v|w, x)p_1(x) + (1 - \tau)q_2(v|w, x)p_2(x) = q_{\tau}(v|w, x)p_{\tau}(x),$$

where

$$q_{\tau}(v|w, x) = \frac{\tau q_1(v|w, x)p_1(x) + (1 - \tau)q_2(v|w, x)p_2(x)}{\tau p_1(x) + (1 - \tau)p_2(x)}$$

and

$$p_{\tau}(x) = \tau p_1(x) + (1 - \tau) p_2(x)$$

for $\tau \in [0, 1]$.

Example 2: Marginal mean model

Let $W = (W_1, \dots, W_K)^T$ and $E(W_k|X) = g_k(X_k\beta)$ for $k = 1, \dots, K$. Let $g(\beta) = \{g_1(X_1\beta), \dots, g_K(X_K\beta)\}^T$, and $f(\varepsilon)$ be the joint density of W given X , where $\varepsilon_i = w_i - g_i(X_i\beta)$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_K)$. Let the density for X be q and the density for V given (W, X) be p , where V denotes the auxiliary covariate. The nuisance parameter for the complete data model is (q, f, p) . When data are missing at random, the efficient score for estimating β is $E_{\eta}\{m_{\eta}^{-1}(\mathcal{D})|R, R(Y)\}$ (Robins *et al.*, 1994; 1995), where $\mathcal{D} = \text{Cov}_{\eta}(S, \varepsilon|X)\{\text{Var}_{\eta}(\varepsilon|X)\}^{-1}\varepsilon$ satisfying

$$\text{Cov}_{\eta}\{m_{\eta}^{-1}(\mathcal{D}), \varepsilon|X\} = \frac{dg}{d\beta}.$$

When X is completely observed, the efficient score has the form

$$\frac{dg}{d\beta} [\text{Var}_{\eta}\{m_{\eta}^{-1}(\varepsilon)|X\}]^{-1} m_{\eta}^{-1}(\varepsilon).$$

Successive approximation is needed when either missing data form nonmontone patterns or covariates are subject to missing values.

When data are fully observed, the score for estimating β is

$$A_{1\eta}h_1 = - \sum_{k=1}^K h_1^T X_k^T g'_k(X_k\beta) \frac{\partial \log f}{\partial \varepsilon_k}(\varepsilon_1, \dots, \varepsilon_K),$$

where $h_1 \in H_1 = R^d$ and d is the dimension of parameter β . H_{10} can be taken as the unit ball in R^d . The nuisance score is

$$A_{2\eta}\{h_{21}(V, X, W), h_{22}(X, W), h_{23}(X)\} = h_{22}(X, W) + h_{21}(V, X, W) + h_{23}(X),$$

where $H_2 = H_{21} \times H_{22} \times H_{23}$ with

$$H_{21} = \{h_{21}(v, w, x) \in L^2(P_{\eta}) | E_{\eta}\{h_{21}(V, W, X) | W, X\} = 0\},$$

$$H_{22} = \{h_{22}(w, x) \in L^2(P_{\eta}) | E_{\eta}\{h_{22}(X, W) | X\} = 0, E_{\eta}\{\varepsilon h_{22}(X, W) | X\} = 0\},$$

and

$$H_{23} = \{h_{23}(x) \in L^2(P_\eta) | E_\eta\{h_{23}(X)\} = 0\}.$$

The adjoint operator of $A_{2\eta}$ is

$$A_{2\eta}^* S(V, X, W) = \{S(V, X, W) - E_\eta(S|X, W), E_\eta(S|X, W) - E_\eta\{S|X\} \\ - \text{Cov}_\eta(S, \varepsilon|X)\{\text{Var}_\eta(\varepsilon|X)\}^{-1}\varepsilon, E_\eta(S|X) - E_\eta(S)\}.$$

It follows that $A_{2\eta}^* A_{2\eta}\{h_{21}, h_{22}, h_{23}\} = (h_{21}(v, w, x), h_{22}(w, x), h_{23}(x))$. Hence, $A_{2\eta}^* A_{2\eta}$ has continuous inverse on $H_{21} \times H_{22} \times H_{23}$ and $\mathcal{P}_\eta = A_{2\eta}(A_{2\eta}^* A_{2\eta})^{-1} A_{2\eta}^*$ appears as

$$\mathcal{P}_\eta s = s(V, X, W) - E_\eta(s\varepsilon|X)\{\text{Var}_\eta(\varepsilon|X)\}^{-1}\varepsilon.$$

for mean-zero square-integrable function s . The efficient score for β under the full data is

$$S_\beta^{\text{eff}, F} = \{X_1^T g'_1(X_1\beta), \dots, X_K^T g'_K(X_K\beta)\} \{\text{Var}_\eta(W|X)\}^{-1}\varepsilon$$

When the densities involving the nuisance parameters are uniformly bounded, the convexity requirement can be verified as follows.

$$\tau q_1(v|w, x) f_1\{w - g(\beta)\} p_1(x) + (1 - \tau) q_2(v|w, x) f_2\{w - g(\beta)\} p_2(x) = q_\tau(v|w, x) f_\tau\{w - g(\beta)\} p_\tau(x),$$

where

$$q_\tau(v|w, x) = \frac{\tau q_1(v|w, x) f_1\{w - g(\beta)\} p_1(x) + (1 - \tau) q_2(v|w, x) f_2\{w - g(\beta)\} p_2(x)}{\tau f_1\{w - g(\beta)\} p_1(x) + (1 - \tau) f_2\{w - g(\beta)\} p_2(x)}, \\ f_\tau\{w - g(\beta)\} = \frac{\tau f_1\{w - g(\beta)\} p_1(x) + (1 - \tau) f_2\{w - g(\beta)\} p_2(x)}{\tau p_1(x) + (1 - \tau) p_2(x)},$$

and $p_\tau(x) = \tau p_1(x) + (1 - \tau) p_2(x)$ for $\tau \in [0, 1]$.

Example 3: The missing covariate problem in Cox regression model

Suppose that T is the survival time of a subject, which is subject to censoring by censoring time C . Given covariate Z (time-independent), T and C are independent. $X = T \wedge C = \min(T, C)$ and $\delta = 1_{\{T \leq C\}}$ rather than (T, C) are observed. Z is subject to missing values. Assume that, given (T, C, Z) , the missing data mechanism depends on the observed data $R(Y) = \{X, \delta, R, R(Z)\}$ only. Suppose that the Cox proportional hazards model holds, that is

$$\lim_{\Delta \rightarrow 0} \frac{1}{\Delta} P(t < T \leq t + \Delta | T \geq t, Z) = \lambda(t) \varphi(\beta Z),$$

where φ is a known function and $\lambda(t)$ is an unknown baseline hazard function. The nuisance parameter includes the censoring distribution, baseline hazard, and covariate distribution. The efficient score for estimating β when data are subject to MAR missing values is

$E_\eta[m_\eta^{-1}\{\mathcal{D}(X, \delta, Z)\}|R, R(X, \delta, Z)]$ (Robins *et al.*, 1994; Nan *et al.*, 2004) where $\mathcal{D}$ is the unique solution to

$$\frac{\partial \log \varphi}{\partial \beta} - \frac{E_\eta\{\xi(u) \frac{\partial \varphi}{\partial \beta}\}}{E_\eta\{\xi(u) \varphi\}} = m_\eta^{-1}(\mathcal{D}(u, 1, Z) - \frac{E_\eta\{m_\eta^{-1}(\mathcal{D})\xi(u)|Z\}}{E_\eta\{\xi(u)|Z\}}) - E_\eta\left\{\xi(u) \varphi \left[m_\eta^{-1}(\mathcal{D})(u, 1, Z) - \frac{E_\eta\{m_\eta^{-1}(\mathcal{D})\xi(u)|Z\}}{E_\eta\{\xi(u)|Z\}} \right] \right\} / E_\eta\{\xi(u) \varphi\}$$

and

$$\mathcal{D} = \int \left\{ b_1(u, Z) - \frac{E_\eta\{\xi(u) \varphi b_1(u, Z)\}}{E_\eta\{\xi(u) \varphi\}} \right\} \{dN(u) - \xi(u) \varphi(\beta Z) d\Lambda(u)\}$$

for some b_1 , where $\xi(u) = 1_{\{X \geq u\}}$ and $N(u) = 1_{\{X \leq u, \delta=1\}}$. The successive approximation is needed in obtaining a locally efficient estimator of β .

The density for (X, δ, Z) is

$$f(x, \delta|z, \beta, \lambda) p(z) = \lambda^\delta(x) \varphi^\delta(\beta z) \exp\{-\Lambda(x) \varphi(\beta z)\} g_c^{1-\delta}(x|z) \bar{G}_c^\delta(x|z) p(z),$$

where $\Lambda(x) = \int_0^x \lambda(t) dt$ and g_c is the density function of the censoring time distribution G_c and $\bar{G}_c = 1 - G_c$. Let $\Lambda_c(x, z) = \int_0^x \lambda_c(t, z) dt$, $\lambda_c = g_c / \bar{G}_c$, $dM_T(t, z) = dN_T(t) - Y(t) \lambda_T(t, z) dt$ with $N_T(t) = 1_{\{X \leq t, \delta=1\}}$, and $dM_C(t, z) = dN_C(t) - Y(t) \lambda_C(t, z) dt$ with $N_C(t) = 1_{\{X \leq t, \delta=0\}}$. The score operator for the parameters (β, λ, g_c, p) is

$$A_\eta\{h_{11}, h_{12}(x), h_{21}(x, z), h_{22}(z)\} = \int h_{11}^T \frac{\partial}{\partial \beta} \log \varphi(\beta Z) dM_T(t, z) + \int h_{12}(t) dM_T(t, z) + \int h_{21}(t, Z) dM_C(t, Z) + h_{22}(Z),$$

where $H_1 = H_{11} \times H_{12}$, $H_2 = H_{21} \times H_{22}$, and $H_{11} = R^k$ with k being the dimension of β ,

$$H_{12} = \{h_{12}(t) | h_{12}(t) \in L^2\{d\Lambda(t)\}\}, \\ H_{21} = \{h_{21}(t, z) | h_{21}(t, z) \in L^2\{d\Lambda_c(t|z) dP(z)\}\},$$

and

$$H_{22} = \{h_{22}(z) \in L^2(P_\eta) | E_\eta\{h_{22}(Z)\} = 0\}.$$

For Λ s that are bounded at T_0 , the study stopping time, H_{12} does not change with Λ . Hence, H_1 is fixed. Define the inner product on H_2 as

$$\langle (h_{21}, h_{22}), (h_{21}^*, h_{22}^*) \rangle_{H_2} = E_\eta \left\{ \int h_{21}(t, Z) h_{21}^*(t, Z) d\Lambda_c(t|Z) \right\} + E_\eta \{ h_{22}(Z) h_{22}^*(Z) \}.$$

It is not difficult to see that H_2 is a Hilbert space under the inner product. Similarly, we can define an inner product on H_1 as

$$\langle (h_{11}, h_{12}), (h_{11}^*, h_{12}^*) \rangle_{H_1} = h_{11}^T h_{11}^* + \int h_{12}(t) h_{12}^*(t) d\Lambda(t)$$

to make it a Hilbert space. Let H_{10} be the subset of H_1 such that for any $(h_{11}, h_{12}) \in H_{10}$, h_{11} is bounded by 1 and $h_{12} \in BV[0, T_0]$, i.e., h_{12} has bounded variation on $[0, T_0]$.

Let $A_{2\eta}(h_{21}, h_{22}) = A_\eta(0, 0, h_{21}, h_{22})$. The adjoint of $A_{2\eta}$ satisfies

$$\langle A_{2\eta}(h_{21}, h_{22}), s(x, \delta, z) \rangle_{L^2(P_\eta)} = \langle (h_{21}, h_{22}), A_{2\eta}^* \{s(x, \delta, z)\} \rangle_{H_2}.$$

Note that, for any $s(X, \delta, Z) \in L^2(P_\eta)$, s can be represented as (Sasieni, 1992; Nan *et al.*, 2004)

$$\begin{aligned} s(X, \delta, Z) &= \int \left[s(t, 1, Z) - \frac{E_\eta \{s(X, \delta, Z)Y(t)|Z\}}{E_\eta \{Y(t)|Z\}} \right] dM_T(t, z) \\ &= \int \left[s(t, 0, Z) - \frac{E_\eta \{s(X, \delta, Z)Y(t)|Z\}}{E_\eta \{Y(t)|Z\}} \right] dM_c(t, z) \\ &\quad + E_\eta \{s(X, \delta, Z)|Z\}, \end{aligned}$$

and the three components are orthogonal to each other. It follows that

$$\begin{aligned} \langle A_{2\eta}(h_{21}, h_{22}), s(X, \delta, Z) \rangle_{L^2(P_\eta)} &= E_\eta \{h_{22}(Z) E_\eta(s|Z)\} \\ &+ E_\eta \left[\int h_{21}(t, Z) \left\{ s(t, 0, Z) - \frac{E_\eta \{s(X, \delta, Z)Y(t)|Z\}}{E_\eta \{Y(t)|Z\}} \right\} d\langle M_c \rangle(t, z) \right], \end{aligned}$$

where $d\langle M_c \rangle(t, z) = Y(t) d\Lambda_c(t, z)$. It can be seen that the adjoint operator $A_{2\eta}^*$ can be obtained as

$$\begin{aligned} A_{2\eta}^*(s) &= \left(s(t, 0, z) E_\eta \{Y(t)|Z\} - E_\eta \{s(X, \delta, Z)Y(t)|Z\}, \right. \\ &\quad \left. E_\eta \{s(X, \delta, Z)|Z\} - E_\eta \{s(X, \delta, Z)\} \right). \end{aligned}$$

By direct calculation, it follows that

$$A_{2\eta}^* A_{2\eta}(h_{21}, h_{22}) = (h_{21}(t, Z) E\{Y(t)|Z\}, h_{22}(Z)),$$

which implies that $A_{2\eta}^* A_{2\eta}$ is continuously invertible on H_2 when $E_\eta\{Y(t)|Z\} > \sigma > 0$ for all $t \leq T_0$. The projection operator appears as

$$\mathcal{P}_\eta(s) = \int \left[s(t, 0, Z) - \frac{E_\eta\{s(X, \delta, Z)Y(t)|Z\}}{E_\eta(Y(t)|Z)} \right] dM_c(t, Z) + E_\eta\{s(X, \delta, Z)|Z\}.$$

for a mean-zero square-integrable function s . The efficient score for estimating (β, Λ) can be expressed as

$$\lim_{N \rightarrow \infty} E_\eta \left\{ (I - \mathcal{P}_\eta m_\eta)^N a(X, \delta, Z) | X, \delta, R, R(Z) \right\},$$

where

$$a(X, \delta, Z) = \int h_{11}^T \frac{\partial}{\partial \beta} \log \varphi(\beta Z) dM_\tau(t, Z) + \int h_{12}(t) dM_\tau(t, Z).$$

The convexity requirement can be verified as follows.

$$\tau g_{c1}^{1-\delta}(x|z) \bar{G}_{c1}^\delta(x|z) p_1(z) + (1 - \tau) g_{c2}^{1-\delta}(x|z) \bar{G}_{c2}^\delta(x|z) p_2(z) = g_{c\tau}^{1-\delta}(x|z) \bar{G}_{c\tau}^\delta(x|z) p_\tau(z),$$

where

$$g_{c\tau}(t|z) = \frac{\tau g_{c1}(t|z) p_1(z) + (1 - \tau) g_{c2}(t|z) p_2(z)}{\tau p_1(z) + (1 - \tau) p_2(z)}$$

and

$$p_\tau(z) = \tau p_1(z) + (1 - \tau) p_2(z),$$

for $\tau \in [0, 1]$. Note that we considered the regression parameter and the baseline hazard as the parameters of interest rather than the regression parameter alone because of the convexity requirement. This treatment is different from those treated in the literature.

5 Numeric study

5.1 Simulation studies

We perform a simulation study on missing data in parametric regression with/without auxiliary covariates. Two parametric regression models were simulated. The first was the logistic regression. The second was the linear regression with a normal error.

In the logistic regression model, two independent covariates were simulated. One was binary and the other was normally distributed. One normally distributed auxiliary covariate was also simulated in this case. It was assumed that, given the covariates, the outcome and the auxiliary

covariates were independent. But the auxiliary covariate depended on the other covariates. In the simulation, both covariates were subject to missing values and the missingness depended on the outcome and the auxiliary covariate only. More specifically, we assumed that $E(Y|X_1, X_2) = g(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$, where $g(t) = (1 + e^{-t})^{-1}$. The parameter $(\beta_0, \beta_1, \beta_2) = (1, -0.5, 0.5)$. The model for the auxiliary covariate V given X_1 and X_2 was set to $E(V|X_1, X_2) = \psi_0 + \psi_1 X_1 + \psi_2 X_2 + \psi_3 X_1 X_2$ with a standard normal error. In the simulation, we set $(\psi_0, \psi_1, \psi_2) = (0.5, 0.3, 1)$ and $\psi_3 = 0$ which corresponds to a correct model for V given X_1 and X_2 , and $\psi_3 = -2$, which corresponds to a severely misspecified model in the data analysis. Three missing covariate patterns were simulated. They were completely observed, observed X_1 only, and observed X_2 only. Let R_1 and R_2 denote the missingness indicators for X_1 and X_2 respectively. The missing data were generated by the model

$$\log \frac{P(R_1=i, R_2=j|Y, V)}{P(R_1=1, R_2=1|Y, V)} = \alpha_0 + \alpha_1 Y + \alpha_2 V + \alpha_3 YV,$$

where $(i, j) = (1, 0)$ or $(0, 1)$ and $(\alpha_0, \alpha_1, \alpha_2) = (-0.5, -0.5, 0.5)$, and $\alpha_3 = 0$ corresponding to a correct missing data mechanism model and $\alpha_3 = -1.5$ corresponding to an incorrect missing data mechanism model in the data analysis.

The correct model for Y given X_1 and X_2 was always used in the analysis of the simulated data. The distributions of the covariates and the auxiliary covariate were assumed unknown in the analysis of the simulated data. The semiparametric odds ratio models with bilinear log-odds ratio functions were used for modeling the distributions of the covariates (X_1, X_2) and of the auxiliary covariate given the outcome and the covariates (X_1, X_2) in the analysis. The polytomous logit regression model with different sets of parameters for different missing patterns and without the interaction term was always assumed in the data analysis. This implies that the missing data mechanism model was misspecified in the analytical model if the model generating the missing data included the interaction term. To compare the performance of different methods, we computed the following estimators for the regression parameter. The first one was the estimator from the complete-case analysis (CC), which is the solution to the estimating equation,

$$\sum_{i=1}^n \mathbf{1}_{\{R_i=1\}} \frac{\partial}{\partial \beta} \log f(Y_i|X_i, \beta) = 0,$$

where $\mathbf{1}$ is a vector with 1 in all of its components. The second one was the simple missing-data-probability weighted estimator (SW), which is the solution to the estimating equation,

$$\sum_{i=1}^n \frac{\mathbf{1}_{\{R_i=1\}}}{\pi_i(\mathbf{1})} \frac{\partial}{\partial \beta} \log f(Y_i|X_i, \beta) = 0,$$

where $\pi_i(r) = \pi\{r, r(V_i, X_i, Y_i), \hat{\alpha}\}$ for all missing-data pattern r and $\hat{\alpha}$ is the maximum likelihood estimate under the missing-data mechanism model, that is, the polytomous linear logit model without interaction. The third was the augmented missing data probability weighted estimator (AW), which is the solution to the estimating equation,

$$\sum_{i=1}^n \left[\frac{1_{\{R_i=1\}}}{\pi_i(\mathbf{1})} \frac{\partial}{\partial \beta} \log f(Y_i|X_i, \beta) + \sum_r \left\{ 1_{\{R_i=r\}} - \frac{1_{\{R_i=1\}}}{\pi_i(\mathbf{1})} \pi_i(r) \right\} \right. \\ \left. \times \widehat{E} \left\{ \frac{\partial}{\partial \beta} \log f(Y_i|X_i, \beta) \middle| R=r, r(V_i, X_i, Y_i) \right\} \right] = 0,$$

where $\widehat{E}$ was computed using the distribution estimated from the following maximum likelihood estimator. The fourth was the maximum likelihood estimator (ML) with the bilinear odds ratio model and without interaction for the covariate distribution. The last two were the likelihood robustification estimators as proposed in this paper using approximation with $N = 10$ (LR-10) and $N = 20$ (LR-20) respectively.

The simulation results were based on 500 replicates of a sample size of 400. The missing proportions for X_1 and for X_2 were approximately 25%. The average number of complete cases thus obtained was close to 200. Table 1 lists the simulation results for the binary outcome data. As expected, when all models are correct, except the CC estimator, the biases of all the other five estimators are relatively small. But the efficiency is different: with the ML estimator the most efficient and the SW estimator the least efficient. When the covariate model is correct and the missing data mechanism is incorrect, the CC and SW estimators can have substantial biases. Biases of all the other estimators are small. When the missing data mechanism model is correct and the covariate model is incorrect, the SW estimator is unbiased. The ML estimator is subject to a sizable bias. Both LR-10 and LR-20 estimators largely correct the bias in the ML estimator. When neither model is correct, the LR-10 and LR-20 estimators along with the AW estimator appear to have much smaller biases when compared with the CC, SW, and ML estimators although all the estimators are biased. The variance estimates for the likelihood robustification estimators appear to work well. The AW estimator has good performance in all the above cases both in terms of bias and variation. This is partly because it has the doubly robust property in the narrow sense that the estimator is consistent when either the covariate models or the missing data mechanism model is correct as long as both the missing data mechanism and its model depend only on the fully observed covariates and the outcome.

In the linear regression model, the variance of the residual error was set to 1. Variables were generated in the same way as in the logistic regression model with the exception that the normal error was used in generating Y and $g(t) = t$. To simplify the computation involved in the analysis of the simulated data, we included V as the third covariate in the linear regression model. However, V had no effect on Y conditional on X_1 and X_2 . The integral with respect to y in computing the expectations in the robustification procedure was approximated by 10 points Gauss-Hermite quadrature. LR-5 ($N = 5$) and LR-10 ($N = 10$) estimators were computed. Five hundred replicates of a sample of 200 were used in the simulation. The results are shown in Table 2. The behavior of the estimators is almost the same as that observed in the previous scenario for the logistic regression model. The difference between LR-5 and LR-10 is still relatively small, which indicates that the convergence rate of the likelihood robustification approximation is reasonably fast in the simulated cases.

In summary, the SW estimator is sensitive to misspecification of the missing data mechanism. The ML estimator can have sizable bias when the covariate models are severely misspecified. We have also simulated other scenarios (not shown) which suggest that the ML estimator with the semiparametric odds ratio model for the covariates is relatively robust against covariate model misspecification. The AW estimator is very robust although it does not have the doubly robust property in general. The likelihood robustification estimators perform better than the AW estimator in all the cases. The estimators from LR-5, LR-10 and LR-20 are nearly indistinguishable, which suggests that approximation using $N = 10$ or even $N = 5$ is good enough in the simulated cases. Other simulations not shown indicate that the number N that gives good

approximation depends on the amount of missing data. In general, the higher the percentage of missing data is, the larger the number N is required. In practice, N can be empirically determined by comparing estimators using different numbers of approximation. In the computation, the covariates that were subject to missing were rounded to the nearest 0.05 in the logistic regression, and to 0.1 in the linear regression. The effect of the rounding on the parameter estimates was nearly negligible as indicated in the results (not shown) when finer roundings were used.

5.2 Application to hip fracture data

The hip fracture data were collected by Dr. Barengolts at the College of Medicine of the University of Illinois at Chicago in studying the hip fracture in veterans. The study matched a case and a control by age and race. Risk factors on bone fracture were assessed. As in Chen (2004), we concentrated on 9 of the risk factors in the analysis. One of the challenging problems in analyzing this dataset is that most of the risk factors are subject to missing values and there are a large number (38 altogether) of missing patterns. This dataset was analyzed in Chen (2004) by the likelihood method using the semiparametric odds ratio models proposed there for the covariates. Since the covariate models applied there are not guaranteed to be correctly specified, it is of interest to verify whether any substantial bias is introduced into the parameter estimator due to the potential covariate model misspecification. This is assessed here by computing the doubly robust estimators of the parameter and comparing them with the maximum likelihood estimator.

There were a few obstacles in actually implementing the proposed method to this dataset. The primary problem was to estimate the missing data probabilities. Since many missing patterns (26 out of 38) have less than 5 observations, it is virtually impossible to estimate the missing data probabilities that depend on one or more variables. As a compromise, we assumed that the missing data did not depend on the observed or unobserved data, *i.e.*, MCAR. Under this assumption, the simple missing data probability weighted estimator is the same as the estimator from the complete-case analysis. We computed the estimator from the complete-case analysis, the maximum likelihood estimator, the augmented weighted estimator, and the likelihood robustification estimators with $N = 10$ and $N = 20$ respectively. In computing these estimators, we rounded data for the three continuous variables: BMI, log(HGB), and Albumin to allow each of them to have about 10 categories. This reduces the computation time and the storage space required. However, the effect of rounding on the parameter estimators is small as discussed in Chen (2004). All the parameter estimates except LR-20, which is the same as LR-10, are shown in Table 3.

The regression coefficients for LevoT4 and dementia estimated from the complete-case analysis are substantially different from those estimated by the other methods. The estimates from the maximum likelihood, the augmented weighted estimating equation, and the likelihood robustification are very close. Estimates from the latter two are even closer. This suggests that the covariate models used in the likelihood approach appear to be reasonable in the sense that it may be close to correctly specified or even if it is incorrectly specified, the influence of the misspecification on the parameter estimates is very small.

6 Discussion

We have shown that the Neumann series approximation can be used to find a locally efficient estimator in missing data problems under the assumption that all configurations of the full data can be observed with a probability bounded away from zero. This helps to close a gap between the semiparametric efficient theory for the missing data problem and the implementation of the procedure in finding such an estimator. The results can be easily modified to be applied to the study of the asymptotic behavior of the doubly robust estimators when missing data are

nonmonotone. Note that the results do not cover the case where a continuous inverse of m does not exist. Similar results in the latter case is expected to be much harder to obtain.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The author thanks the editor for the detailed comments which have greatly improved the presentation of the paper. The author would also like to thank Professor James Robins for his insightful comments on the earlier versions of the paper. Comments from Drs. Y. Q. Chen and C. Y. Wang at FHCRC on the earlier versions of the paper are also very much appreciated. The research was supported by a grant from NCI/NIH on statistical methods for missing data.

References

- Begun JM, Hall WJ, Huang WM, Wellner J. Information and asymptotic efficiency in parametric-nonparametric models. *Ann Statist* 1983;11:432–452.
- Bickel, P.; Klassen, C.; Ritov, Y.; Wellner, J. *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: John Hopkins University Press; 1993.
- Chen HY. Nonparametric and semiparametric models for missing covariates in parametric regression. *J Amer Statist Assoc* 2004;99:1176–1189.
- Huang Y. Calibration regression of censored lifetime medical cost. *J Amer Statist Assoc* 2002;97:318–327.
- Lipsitz SR, Ibrahim JG, Zhao LP. A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *J Amer Statist Assoc* 1999;94:1147–1160.
- Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. 2. New York: John Wiley; 2002.
- Nan B, Emond MJ, Wellner JA. Information bounds for Cox regression models with missing data. *Ann Statist* 2004;32:723–735.
- Robins JM, Hsieh FS, Newey W. Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *J Roy Statist Soc, Ser B* 1995;57:409–424.
- Robins JM, Rotnitzky A. Comments on “Inference for semiparametric models: Some questions and an answer” by Bickel, P. J. and Kwon, J. in the millennium series of *Statist. Sinica* 2001;11:920–936.
- Robins JM, Rotnitzky A, van der Laan MJ. Discussion of “On profile likelihood” by Murphy, S.A. and van der Vaart, A. W. *J Amer Statist Assoc* 1999;94:477–482.
- Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Amer Statist Assoc* 1994;89:846–866.
- Robins JM, Rotnitzky A, Van der Laan M. Comment on “On profile likelihood” by Murphy and Van der Vaart. *J Amer Statist Assoc* 2000;95:477–485.
- Robins JM, Wang N. Discussion on the papers by Forster and Smith and Clayton et al. *J Roy Statist Soc, Ser B* 1998;60:91–93.
- Rotnitzky A, Robins JM. Analysis of semiparametric models with nonignorable nonresponses. *Stat Med* 1997;16:81–102. [PubMed: 9004385]
- Rubin DB. Inference and missing data. *Biometrika* 1976;63:581–92.
- Sasieni P. Information bounds for the conditional hazard ratio in a nested family of regression models. *J Roy Statist Soc, Ser B* 1992;54:617–635.
- Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *J Amer Statist Assoc* 1999;94:1096–1120.
- Van der Laan, MJ.; Robins, JM. *Unified methods for censored longitudinal data and causality*. New York: Springer; 2003.
- Van der Vaart, AW.; Wellner, JA. *Weak Convergence and Empirical Processes With Application to Statistics*. New York: Springer; 1996.
- Yu M, Nan B. Semiparametric regression models with missing data: the mathematical review and a new application. *Statist Sinica* 2006;16:1193–1212.

Appendix: Proof of Theorems

Assume that the semiparametric model under consideration is $\pi(r/y, \eta)p(y, \eta)$, where $p(y, \eta)$ is the density of the full data y with respect to μ , a product of Lebesgue measures and counting measures. Assume that $\eta = (\beta, \theta) \in \Omega \times \Theta$ and β is the parameter of interest and θ is the nuisance parameter. The marginal distribution for observed data $(R, R(Y))$ is $g(r, r(y), \eta) = \int \pi(r/y, \eta)p(y, \eta)d\bar{r}(y)$, where $\bar{r}(y)$ denote those components of y that are missing in the missing pattern r . Assume that $\theta(\gamma)$, $\gamma \in \Gamma$ is a restricted parameterization of θ such that $\theta(\Gamma) \subset \Theta$. Let $\eta_0 = (\beta_0, \theta^*)$ and P_0 be the true probability measure that generated the data. The following regularity conditions are used in the theorems.

1. For any $\beta \in \Omega$ and $\gamma \in \Gamma$, aside from a μ -zero set, $\{(r, y)/g(r, r(y), \beta, \theta(\gamma)) > 0\}$ is the same for any fixed r . $\pi(r/y, \eta)$ and $p(y, \eta)$ are bounded away from zero and ∞ uniformly for all y and $\eta \in \mathcal{E} = \Omega \times \theta(\Gamma) \subset \Omega \times \Theta$, if $\pi(r/y, \eta) > 0$ for a μ -nonzero set of y . The full data model $p(y, \eta)$ is a convex family with respect to θ .
2. Ω is a compact subset of a Hilbert space. The true parameter value β_0 is an inner point of Ω . Θ is a subset of another Hilbert space. As $n \rightarrow \infty$, $\hat{\gamma} \rightarrow \gamma^*$ in the norm defined on Γ . $\theta(\gamma)$ is continuous.
3. As a $L^2(\mu)$ function of $(\beta, \theta) \in \Omega \times \Theta$, $\{\pi(r/y, \eta)p(y, \eta)\}^{1/2}$ is Fréchet differentiable with respect to $\eta \in \Omega \times \Theta$. The score operator defined as $2\{\pi(r/y, \eta)p(y, \eta)\}^{-1/2}$ times the derivative is denoted by $(A_{1\eta}(h_1), A_{2\eta}(h_2))$ with $h_1 \in H_1$ and $h_2 \in H_2$. Both H_1 and H_2 are Hilbert spaces.
4. $A_{2\eta}^*A_{2\eta}$ is continuously invertible at η_0 , where $A_{2\eta}^*$, mapping $L^2\{P_\eta(y)\}$ to H_2 , is the adjoint operator of $A_{2\eta}$.
5. For $\eta \in \mathcal{E}$, $(\pi_\eta p_\eta)^{1/2}A_{1\eta}(h_1)$ and $(\pi_\eta p_\eta)^{1/2}A_{2\eta}(h_2)$ are continuous with respect to η in $L^2(\mu)$ and are Fréchet differentiable with respect to β in a neighborhood of η_0 in $L^2(\mu)$ and $A_{2\eta}^*((\pi_\eta p_\eta)^{-1/2}s)$ is continuous with respect to η in H_2 norm for $s \in L^2(\mu)$ and is Fréchet differentiable with respect to β in a neighborhood of η_0 in H_2 norm.
6. Suppose that $p(y, \eta) = dP_\eta/d(\mu_1 \times \dots \times \mu_J)$ where μ_j is Lebesgue measure on R^1 or a counting measure, $j = 1, \dots, J$. Suppose that, for any missing pattern r , $\pi(r/y, \eta)p(y, \eta)$, $A_{1\eta}(h_1)$, $A_{2\eta}(h_2)$, and $A_{2\eta}^*(s)$, and their derivatives with respect to β for $\eta \in \mathcal{E}$, are all continuous with respect to the j th argument of y if μ_j is Lebesgue measure, $j = 1, \dots, J$ and h_1 , h_2 , and s are continuous with respect to y_j .

There exists a norm on $\mathcal{E}$, denoted by $\|\cdot\|$, such that

$$\begin{aligned} \|\pi(r/y, \eta_1) - \pi(r/y, \eta_2)\|_{L^\infty(P_{\eta_0})} &\leq C_1 \|\eta_1 - \eta_2\|, \\ \|p(y, \eta_1) - p(y, \eta_2)\|_{L^\infty(P_{\eta_0})} &\leq C_2 \|\eta_1 - \eta_2\|, \\ \|A_{1\eta_1}(h_1) - A_{1\eta_2}(h_1)\|_{L^\infty(P_{\eta_0})} &\leq C_3 \|\eta_1 - \eta_2\| \|h_1\|_{H_1}, \\ \|A_{2\eta_1}(h_2) - A_{2\eta_2}(h_2)\|_{L^\infty(P_{\eta_0})} &\leq C_4 \|\eta_1 - \eta_2\| \|h_2\|_{H_2}, \end{aligned}$$

and

$$\|A_{2\eta_1}^*(s) - A_{2\eta_2}^*(s)\|_{H_2} \leq C_5 \|\eta_1 - \eta_2\| \|s\|_{L^\infty(P_{\eta_0})}$$

for any $\eta_1, \eta_2 \in \mathcal{E}$ and some constants $C_i, i = 1, \dots, 5$. The ε -covering number for $\mathcal{E}$ under $\|\cdot\|, N(\mathcal{E}, \varepsilon, \|\cdot\|)$ satisfies

$$\int_0^\infty \sqrt{\log N(\mathcal{E}, \varepsilon / \|\log \varepsilon\|, \|\cdot\|)} d\varepsilon < \infty.$$

7. There exists an $H_{10} \subset H_1$, for any fixed continuously invertible map $\mathcal{A}$ from H_1 to itself, if $\langle \mathcal{A}h_1, \beta \rangle_{H_1} = 0$ for all $h_1 \in H_{10}$, then $\beta = 0$. The covering number of H_{10} under supremum norm, $N(H_{10}, \varepsilon, \|\cdot\|_\infty)$ satisfies

$$\int_0^\infty \sqrt{\log N(H_{10}, \varepsilon, \|\cdot\|_\infty)} d\varepsilon < \infty.$$

8.

$$\inf_{\|h_1\|_{H_1}=1, \|h_1^*\|_{H_1}=1} \left| E_0 \left\{ \frac{\partial T(\eta_0)}{\partial \beta}(h_1)(h_1^*) \right\} \right| > 0.$$

9.

$$\inf_{\|h_1\|_{H_1}=1, \|h_1^*\|_{H_1}=1} \left| E_0 \left\{ \frac{\partial T_N(\beta_N, \theta^*)}{\partial \beta}(h_1)(h_1^*) \right\} \right| > 0.$$

10. There exists $U_i, i = 1, \dots, n$, iid with $E_0 U^{\otimes 2}$ finite such that

$$\theta(\widehat{\gamma}) - \theta^* = \frac{1}{n} \sum_{i=1}^n U_i + o_p\left(\frac{1}{\sqrt{n}}\right),$$

where $\theta^* = \theta(\gamma^*)$.

11. $A_{1\eta}(h_1)$ and $A_{2\eta}(h_2)$ are Fréchet differentiable with respect to θ along the path $\theta(\gamma)$, $\gamma \in \Gamma$ in a neighborhood of η_0 in $L^2(P_{\eta_0})$ and $A_{2\eta}^*(s)$ is Fréchet differentiable with respect to θ along the path $\theta(\gamma)$, $\gamma \in \Gamma$ in a neighborhood of η_0 in H_2 .

Before we prove Theorems 1–3, we first establish a set of lemmas for the proofs of the theorems. These lemmas are mostly for showing that T and T_N are differentiable and that

$\mathcal{F} = \bigcup_{N=0}^\infty \{T_N(\eta)(h_1) | \eta \in \mathcal{E}, h_1 \in H_{10}\} \cup \{T(\eta)(h_1) | \eta \in \mathcal{E}, h_1 \in H_{10}\}$ is a P_0 -Donsker class. Let

$D_\eta = (I - \mathcal{P}_\eta m_\eta)$ where $\mathcal{P}_\eta = A_{2\eta}(A_{2\eta}^* A_{2\eta})^{-1} A_{2\eta}^*$. It follows that $T_N(\eta, h_1) = E_\eta \{D_\eta^N A_{1\eta} h_1 | R, R(Y)\}$.

We start from lemmas on the differentiable of T_N with respect to β . Proofs of the lemmas are suppressed and can be found in the supplement materials.

Lemma 1

- a. Under assumptions 1–5, $g_\eta^{1/2} T_N(\eta)(h_1)$, for any N , is Fréchet differentiable with respect to β in $L^2(\mu)$ in a neighborhood of η_0 in $\mathcal{E}$, and both $g_\eta^{1/2} T_N(\eta)(h_1)$, for any N , and the derivatives are continuous at η_0 in $L^2(\mu)$. If we define the derivative of T_N with respect to β as

$$\frac{\partial T_N(\eta)}{\partial \beta}(h_1)(h_1^*) = g_\eta^{-1/2} \left\{ \frac{\partial \{g_\eta^{1/2} T_N(\eta)(h_1)\}}{\partial \beta}(h_1^*) - \frac{1}{2} g_\eta^{1/2} B_{1\eta}(h_1^*) T_N(\eta)(h_1) \right\},$$

where the first term on the right-hand side of the equation refers to the derivative of $g_\eta^{1/2} T_N(\eta)(h_1)$ with respect to β , then

$$\|\varepsilon^{-1} \{T_N(\eta + \varepsilon h_1^*)(h_1) - T_N(\eta)(h_1)\} - \frac{\partial T_N(\eta)}{\partial \beta}(h_1)(h_1^*)\|_{L^2(P_\eta)} \rightarrow 0,$$

as $n \rightarrow \infty$.

- b.** (b) Under assumptions 1–5 and 12, $g_\eta^{1/2} T_N(\eta)(h_1)$, for any N , is Fréchet differentiable with respect to θ along the path $\theta(\gamma)$, $\gamma \in \Gamma$, in $L^2(\mu)$ in a neighborhood of η_0 and the derivatives are continuous at η_0 in $L^2(\mu)$. The derivative of T_N with respect to θ is defined similarly.

Lemma 2

Under the assumptions 1–5, $T(\eta)(h_1)$ is Fréchet differentiable with respect to β in $L^2(P_{\eta_0})$ in a neighborhood of η_0 and both $T(\eta)(h_1)$ and the derivative are continuous at η_0 in $L^2(P_{\eta_0})$. If we define the derivative of T with respect to β as

$$\frac{\partial T(\eta)}{\partial \beta}(h_1)(h_1^*) = g_\eta^{-1/2} \left\{ \frac{\partial \{g_\eta^{1/2} T(\eta)(h_1)\}}{\partial \beta}(h_1^*) - \frac{1}{2} g_\eta^{1/2} B_{1\eta}(h_1^*) T(\eta)(h_1) \right\},$$

where the first term on the right-hand side of the equation refers to the derivative of $g_\eta^{1/2} T(\eta)(h_1)$ with respect to β , then

$$\|\varepsilon \{T(\eta + \varepsilon h_1^*)(h_1) - T(\eta)(h_1)\} - \frac{\partial T(\eta)}{\partial \beta}(h_1)(h_1^*)\|_{L^2(P_\eta)} \rightarrow 0,$$

as $n \rightarrow \infty$.

Lemma 3

Under assumptions 1–3, for any $p > 0$, $N > 0$, and s , a function of Y in $L^\infty(P_\eta)$,

$$\left\| D_\eta^{N+p} s - D_\eta^N s \right\|_{L^\infty(P_\eta)} \leq K N^{c(Y)} (1 - \sigma)^N \|s\|_{L^\infty(P_\eta)},$$

where $c(Y)$ denotes the cardinality of Y and K is a constant independent of N .

Lemma 4

Suppose that there exists a measurable set $\mathcal{N}$ with $P_0(\mathcal{N}) = 0$ such that for all $t(Y) \in \cup_{N=0}^{\infty} \{D_{\eta}^N s | s \in S, \eta \in \mathcal{E}\} \cup \{\lim_{N \rightarrow \infty} D_{\eta}^N s \text{ in } L^{\infty}(P_{\eta}) | s \in S, \eta \in \mathcal{E}\}$,

$$\|t(Y)\|_{L^{\infty}(P_{\eta})} \geq b \sup_{y \in \mathcal{N}^c} |t(y)|, \quad (2)$$

where S is a set of functions of y and $0 < b \leq 1$ is a constant. Then,

$$\cup_{N=0}^{\infty} \{D_{\eta}^N s | s \in S, \eta \in \mathcal{E}\} \cup \{\lim_{N \rightarrow \infty} D_{\eta}^N s \text{ in } L^{\infty}(P_{\eta}) | s \in S, \eta \in \mathcal{E}\}$$

is a P_0 -Donsker class if

- a. S is P_0 -measurable and $L^{\infty}(P_0)$ bounded satisfying

$$\int_0^{\infty} \sqrt{\log N(S, \varepsilon, L^{\infty}(P_{\eta}))} d\varepsilon < \infty, \quad (3)$$

for a fixed $\eta \in \mathcal{E}$.

- b. For any $\eta_1, \eta_2 \in \mathcal{E}$, $s \in S$, and a fixed $\eta \in \mathcal{E}$, there exists a $C(\eta) < \infty$ such that

$$|D_{\eta_1} s - D_{\eta_2} s| \leq C(\eta) \|\eta_1 - \eta_2\| \|s\|_{L^{\infty}(P_{\eta})},$$

and $\mathcal{E}$ has covering numbers under $\|\cdot\|$ satisfying

$$\int_0^{\infty} \sqrt{\log N(\mathcal{E}, \varepsilon / \|\log \varepsilon\|, \|\cdot\|)} d\varepsilon < \infty. \quad (4)$$

Lemma 5

Under assumption 6, if $s(y, \eta)$ is continuous with respect to argument j when μ_j is Lebesgue measure, then there exists a measurable set $\mathcal{N}$ with $P_{\eta}(\mathcal{N}) = 0$ such that

$$\|t(Y, \eta)\|_{L^{\infty}(P_{\eta})} = \sup_{Y \in \mathcal{N}^c} |t(y, \eta)|,$$

for any

$$t \in \cup_{N=0}^{\infty} \{D_{\eta}^N s | s \in S, \eta \in \mathcal{E}\} \cup \{\lim_{N \rightarrow \infty} D_{\eta}^N s | s \in S, \eta \in \mathcal{E}\},$$

where the limit is in the sense of the $L^{\infty}(P_{\eta})$ norm.

Lemma 6

Under assumptions 1–7, we have

$$\|D_{\eta_1} s - D_{\eta_2} s\|_{L^\infty(P_\eta)} \leq C_1(\eta) \|\eta_1 - \eta_2\| \|s\|_{L^\infty(P_\eta)},$$

and

$$\|E_{\eta_1}\{D_{\eta_1} s | R, R(y)\} - E_{\eta_2}\{D_{\eta_2} s | R, R(y)\}\|_{L^\infty(P_\eta)} \leq C_2(\eta) \|\eta_1 - \eta_2\| \|s\|_{L^\infty(P_\eta)},$$

for some constants $C_1(\eta), C_2(\eta) < \infty$.

Lemma 7

Under assumptions 1–8,

$$\mathcal{F} = \bigcup_{N=0}^{\infty} \{T_N(\eta)(h_1) | \eta \in \mathcal{E}, h_1 \in H_{10}\} \cup \{T(\eta)(h_1) | \eta \in \mathcal{E}, h_1 \in H_{10}\}$$

is a P_0 -Donsker class.

Proof of Theorem 1

Let $\tilde{\eta} = (\tilde{\beta}, \tilde{\theta})$. By definition, $P_n T(\tilde{\eta})(h_1) = 0$ for $h_1 \in H_{10}$. From Lemma 7, $\{T(\eta)(h_1) | \eta \in \mathcal{E}, h_1 \in H_{10}\}$ is a P_0 -Donsker class with bounded envelope function, and thus a P_0 -Glivenko-Cantelli class. For a convergent point of a subsequence of $\tilde{\eta}$, denoted by $\eta_0^* = (\beta_0^*, \theta^*)$, it follows from the continuity of $T(\eta)$ in a neighborhood of η_0 (Lemma B.1 (a)) that $P_0 T(\eta_0^*)(h_1) = 0$. Note that $P_0 T(\eta_0)(h_1) = 0$, where $\eta_0 = (\beta_0, \theta^*)$. Since $T(\eta)(h_1)$ is differentiable with respect to β in a neighborhood of β_0 in $L^2(P_{\eta_0})$ at η_0 and the derivative is continuous at η_0 in η , by the mean value theorem,

$$P_0\{T(\beta_0^*, \theta^*)(h_1) - T(\beta_0, \theta^*)(h_1)\} = P_0 \left[\frac{\partial T}{\partial \beta} \{\eta_0 + \lambda(\eta_0^* - \eta_0)\}(h_1)(\beta_0^* - \beta_0) \right]$$

for some $0 \leq \lambda \leq 1$ and all $h_1 \in H_{10}$. From assumptions 8 and 9, we can conclude that $\beta_0^* = \beta_0$ locally. Since $\tilde{\theta}$ converges (Assumption 2) and β varies in a compact set, which implies each convergent subsequence converges to the same limit, β locally converges to β_0 almost surely.

Since $E_0\{T(\tilde{\eta})(h_1) - T(\eta_0)(h_1)\}^2 \rightarrow 0$ uniformly for $h_1 \in H_{10}$ and $\{T(\eta)(h_1) | \eta \in \mathcal{E}, h_1 \in H_{10}\}$ is a P_0 -Donsker class with bounded envelope function, it follows (van der Vaart & Wellner, 1996, Lemma 3.3.5 on page 311) that $\sqrt{n}(P_n - P_0)\{T(\tilde{\eta})(h_1) - T(\eta_0)(h_1)\} = o_p(1)$, which implies that $-\sqrt{n}P_0 T(\tilde{\eta})(h_1) = \sqrt{n}(P_n - P_0)T(\eta_0)(h_1) + o_{P_0}(1)$. Note that

$$\sqrt{n}P_0 \left\{ T(\tilde{\eta})(h_1) - T(\beta_0, \tilde{\theta})(h_1) \right\} = P_0 \left\{ \frac{\partial T}{\partial \beta}(\beta_0, \theta^*)(h_1) \sqrt{n}(\tilde{\beta} - \beta_0) \right\} + o_{P_0} \left(\left\| \sqrt{n}(\tilde{\beta} - \beta_0) \right\| \right),$$

and that $P_0 T(\beta_0, \tilde{\theta})(h_1) = P_0 T(\beta_0, \theta^*)(h_1) = 0$. It now follows that

$$-P_0 \left[\frac{\partial T}{\partial \beta}(\eta_0)(h_1) \left\{ \sqrt{n}(\tilde{\beta} - \beta_0) \right\} \right] = \sqrt{n}(P_n - P_0)T(\eta_0)(h_1) + o_{P_0}(1 + \left\| \sqrt{n}(\tilde{\beta} - \beta_0) \right\|).$$

By replacing h_1 in the foregoing equation by $Q_0^{-1}h_1$, it follows that

$$\langle \sqrt{n}(\tilde{\beta} - \beta_0), h_1 \rangle_{H_1} = \sqrt{n}P_n T(\eta_0) \left\{ -Q_0^{-1}(h_1) \right\} + o_{P_0}(1 + \left\| \sqrt{n}(\tilde{\beta} - \beta_0) \right\|),$$

which implies that $\langle \sqrt{n}(\tilde{\beta} - \beta_0), h_1 \rangle_{H_1} = O_{P_0}(1)$ and that $\langle \sqrt{n}(\tilde{\beta} - \beta_0), h_1 \rangle_{H_1} \rightarrow N(0, V(h_1))$ uniformly on $h_1 \in H_{10}$, where $V(h_1) = E_0 \left[T(\eta_0) \left\{ -Q_0^{-1}(h_1) \right\} \right]^2$.

To prove the consistency of the variance estimate, let $\tilde{\eta}_h = (\tilde{\beta} + \frac{h}{\sqrt{n}}, \tilde{\theta})$ for a fixed $h \in H_1$. That $\{T(\tilde{\eta}_h)(h_1) - T(\tilde{\eta})(h_1)/h_1 \in H_{10}\}$ is a P_0 -Donsker class implies that

$\sqrt{n}(P_n - P_0) \{T(\tilde{\eta}_h) - T(\tilde{\eta})\}(h_1) = o_{P_0}(1)$. It follows that, apart from a $o_{P_0}(1)$ term,

$$\begin{aligned} \sqrt{n}P_n \{T(\tilde{\eta}_h) - T(\tilde{\eta})\}(h_1) &= \sqrt{n}P_0 \{T(\tilde{\eta}_h) - T(\tilde{\eta})\}(h_1) \\ &= P_0 \left\{ \frac{\partial T(\eta_0)}{\partial \beta}(h)(h_1) \right\} = \langle Q_0 h, h_1 \rangle_{H_1}. \end{aligned}$$

Since $\{T^2(\tilde{\eta})(h_1)/h_1 \in H_{10}\}$ is a Glivenko-Cantelli class, $P_n \{T(\tilde{\eta})(h_1)\}^2 = P_0 \{T(\eta_0)(h_1)\}^2 + o_{P_0}(1)$ uniformly in $h_1 \in H_{10}$. It can now be seen that the asymptotic variance of $\langle \sqrt{n}(\tilde{\beta} - \beta_0), h_1 \rangle_{H_1}$ can be consistently estimated by $P_n \{T(\tilde{\eta})(h)\}^2$, where $h \in H_1$ solves the equation $\langle h_1, h_1 \rangle_{H_1} = \sqrt{n}P_n \{T(\tilde{\eta}_h)(h_1) - T(\tilde{\eta})(h_1)\}$. Note that, from the previous derivation, for any fixed $h_1 \in H_{10}$, h thus defined converges in probability (P_0) to $Q_0^{-1}(h_1)$ uniformly over $h_1 \in H_{10}$.

When both the missing data mechanism model and the nuisance model for the full data are correctly specified, $\theta^* = \theta_0$. That is, $P_{\eta_0} = P_0$. It follows from

$E_{(\beta_0+h/\sqrt{n}, \theta_0)} T(\beta_0+h/\sqrt{n}, \theta_0)(h_1) = 0$ for any $h, h_1 \in H_{10}$ that

$$-E_0 \left\{ \frac{\partial T}{\partial \beta}(\eta_0)(h)(h_1) \right\} = E_0 \{T(\beta_0, \theta_0)(h_1) A_{1(\beta_0, \theta_0)}(h)\} = E_0 \{T(\beta_0, \theta_0)(h_1) T(\beta_0, \theta_0)(h)\}.$$

This implies that $\tilde{\beta}$ is asymptotically efficient.

Proof of Theorem 2

For any convergent point of $\tilde{\beta}_N$, denoted by β_N , let $\hat{\eta}_N = (\tilde{\beta}_N, \hat{\theta})$ and $\eta_N = (\beta_N, \hat{\theta})$. By definition, $P_n T_N(\hat{\eta}_N)(h_1) = 0$. Note that $\mathcal{F}$ (see Lemma 7 for its definition) is a P_{η_0} -Donsker class with bounded envelop function, and thus a P_{η_0} -Glivenko-Cantelli class. Since dP_{η_0}/dP_0 is bounded away from 0 and ∞ , $\mathcal{F}$ is a P_0 -Glivenko-Cantelli class. Since $P_0\{T_N(\hat{\eta}_N)(h_1) - T_N(\eta_N)(h_1)\}^2 \rightarrow 0$ uniformly over η_N and $h_1 \in H_{10}$ for any fixed N as $n \rightarrow \infty$, it follows that $(P_n - P_0)\{T_N(\hat{\eta}_N)(h_1) - T_N(\eta_N)(h_1)\} = o_{P_0}(1)$. Hence, $P_n T_N(\eta_N)(h_1) = o_{P_0}(1)$, which implies that $P_0 T_N(\eta_N)(h_1) = 0$ for all $h_1 \in H_{10}$. Note that

$$P_0\{T_N(\beta_N, \hat{\theta})(h_1) - T_N(\beta_0, \hat{\theta})(h_1)\} = P_0\left\{\frac{\partial T_N}{\partial \beta}(\beta_0, \theta^*)(h_1)(\beta_N - \beta_0)\right\} + o_{P_0}(\|\beta_N - \beta_0\|_{H_1}).$$

It follows from the definition of $\mathcal{Q}_{0N}$ that, for all $h_1 \in H_{10}$,

$$\langle \beta_N - \beta_0, h_1 \rangle_{H_1} = P_0\{T_N(\beta_N, \hat{\theta})(-Q_{0N}^{-1}h_1) - T_N(\beta_0, \hat{\theta})(-Q_{0N}^{-1}h_1)\} + o_{P_0}(\|\beta_N - \beta_0\|_{H_1}).$$

Next, note that $\sqrt{n}(P_n - P_0)\{T_N(\tilde{\beta}_N, \hat{\theta})(h_1) - T_N(\beta_N, \hat{\theta})(h_1)\} = o_{P_0}(1)$ uniformly over $h_1 \in H_{10}$. It follows that, for all $h_1 \in H_{10}$,

$$\begin{aligned} -\sqrt{n}P_n T_N(\beta_N, \hat{\theta})(h_1) &= \sqrt{n}P_0\{T_N(\tilde{\beta}_N, \hat{\theta})(h_1) - T_N(\beta_N, \hat{\theta})(h_1)\} + o_{P_0}(1) \\ &= P_0\left\{\frac{\partial T}{\partial \beta}(\beta_N, \theta^*)\sqrt{n}(\tilde{\beta}_N - \beta_N)(h_1)\right\} + o_{P_0}(\sqrt{n}\|\tilde{\beta}_N - \beta_N\|_{H_1} + 1). \end{aligned}$$

From assumption 10 and the definition of $\mathcal{Q}_N$, it follows that

$$\begin{aligned} \langle \sqrt{n}(\tilde{\beta}_N - \beta_N), h_1 \rangle_{H_1} &= \sqrt{n}P_0\{T_N(\beta_N, \hat{\theta})(-Q_N^{-1}h_1) - T_N(\beta_N, \theta^*)(-Q_N^{-1}h_1)\} \\ &\quad + \sqrt{n}(P_n - P_0)T_N(\beta_N, \theta^*)(-Q_N^{-1}h_1) + o_{P_0}(\|\sqrt{n}(\tilde{\beta}_N - \beta_N)\|_{H_1} + 1). \end{aligned}$$

Note that, apart from a $o_{P_0}(1)$ term,

$$\sqrt{n}P_0\{T_N(\beta_N, \hat{\theta})(h_1) - T_N(\beta_N, \theta^*)(h_1)\} = \sum_{i=1}^n P_0\left\{T_N(\beta_N, \theta^* + \frac{U_i}{\sqrt{n}})(h_1) - T_N(\beta_N, \theta^*)(h_1)\right\}$$

because T_N is Fréchet differentiable at (β_N, θ^*) with respect to θ along the path $\theta(\gamma)$ in $L^2(P_0)$ and $\hat{\theta} - \theta^* = \frac{1}{n} \sum_{i=1}^n U_i + o_{P_0}(\frac{1}{\sqrt{n}})$, where $E_0(U) = 0$. It follows that, uniformly over $h_1 \in H_{10}$, $\sqrt{n}(\tilde{\beta}_N - \beta_N)(h_1) \rightarrow N(0, V_N(h_1))$ where

$$V_N(h_1) = E_0 \left[T_N(\beta_N, \theta^*)(-Q_N^{-1}h_1) + P_0 \left\{ \frac{\partial T_N(\beta_N, \theta^*)}{\partial \theta}(u)(-Q_N^{-1}h_1) \right\}_{u=U} \right]^2.$$

The consistency of the variance estimate can be shown by virtually identical statements as those given in the previous theorem.

Proof of Theorem 3

Let β_{N_n} denote the solution to the equation $P_n T_{N_n}(\beta_{N_n}, \hat{\theta})(h_1) = 0$ for all $h_1 \in H_1$. From the Donsker class result and the uniform convergence of T_N to T in $L^2(P_0)$, it follows that

$\sqrt{n}(P_n - P_0) \{T_{N_n}(\beta_{N_n}, \hat{\theta})(h_1) - T(\beta_{N_n}, \hat{\theta})(h_1)\} = o_{P_0}(1)$, uniformly over $h_1 \in H_{10}$. The equations imply that

$$\sqrt{n}P_n \{T(\beta_{N_n}, \hat{\theta})(h_1)\} = \sqrt{n}P_0 \{T_{N_n}(\beta_{N_n}, \hat{\theta})(h_1) - T(\beta_{N_n}, \hat{\theta})(h_1)\} + o_{P_0}(1)$$

uniformly for $h_1 \in H_{10}$. Since $|\sqrt{n}P_0 \{T_{N_n}(\beta_{N_n}, \hat{\theta})(h_1) - T(\beta_{N_n}, \hat{\theta})(h_1)\}| \leq K \sqrt{n}(1 - \sigma)^{N_n}$, and that $\log(n)/N_n \rightarrow 0$ implies $K \sqrt{n}(1 - \sigma)^{N_n} \rightarrow 0$ as $n \rightarrow \infty$, it follows that

$\sqrt{n}P_n \{T(\beta_{N_n}, \hat{\theta})(h_1)\} = o_{P_0}(1)$, uniformly over $h_1 \in H_{10}$. From the definition of $\tilde{\beta}$, it follows that

$\sqrt{n}P_n \{T(\beta_{N_n}, \hat{\theta})(h_1) - T(\tilde{\beta}, \hat{\theta})(h_1)\} = o_{P_0}(1)$, uniformly over $h_1 \in H_{10}$. From the fact that $\mathcal{F}$ is a P_0 -Donsker class and that T is continuous in $L^2(P_0)$ at η_0 , it follows that $P_0 \{T(\beta_{N_n}, \hat{\theta})(h_1) - T(\tilde{\beta}, \hat{\theta})(h_1)\}^2 \rightarrow 0$, uniformly over $h_1 \in H_{10}$ and thus that

$\sqrt{n}(P_n - P_0) \{T(\beta_{N_n}, \hat{\theta})(h_1) - T(\tilde{\beta}, \hat{\theta})(h_1)\} = o_{P_0}(1)$, uniformly over $h_1 \in H_{10}$. It follows that

$\sqrt{n}P_0 \{T(\beta_{N_n}, \hat{\theta})(h_1) - T(\tilde{\beta}, \hat{\theta})(h_1)\} = o_{P_0}(1)$, uniformly over $h_1 \in H_{10}$. It now follows from

Lemma 2 and assumption 9 that $\sqrt{n}(\beta_{N_n} - \tilde{\beta}, h_1)_{H_1} = o_{P_0}(1)$, uniformly over $h_1 \in H_{10}$

For the variance estimate, let $\eta_{N_n h} = (\tilde{\beta}_{N_n} + \frac{h}{\sqrt{n}}, \hat{\theta})$ for a fixed $h \in H_1$. Again, from the Donsker class result and the uniform convergence of T_N to T in $L^2(P_0)$, it follows that

$\sqrt{n}(P_n - P_0) \{T_{N_n}(\eta_{N_n h})(h_1) - T_{N_n}(\eta_{N_n})(h_1)\} = o_{P_0}(1)$, uniformly over $h_1 \in H_{10}$. It follows from

the foregoing equation that $\sqrt{n}P_n \{T_{N_n}(\eta_{N_n h})(h_1) - T_{N_n}(\eta_{N_n})(h_1)\}$ can be rewritten as

$\sqrt{n}P_0 \{T_{N_n}(\eta_{N_n h})(h_1) - T(\eta_{N_n h})(h_1)\} + \sqrt{n}P_0 \{T_{N_n}(\eta_{N_n})(h_1) - T(\eta_{N_n})(h_1)\}$ plus

$\sqrt{n}P_0 \{T(\eta_{N_n h})(h_1) - T(\eta_{N_n})(h_1)\} + o_{P_0}(1)$. Since

$$|\sqrt{n}P_0 \{T_{N_n}(\eta_{N_n h})(h_1) - T(\eta_{N_n h})(h_1)\}| \leq \sqrt{n}K(1 - \sigma)^{N_n} \rightarrow 0$$

and $|\sqrt{n}P_0 \{T_{N_n}(\eta_{N_n})(h_1) - T(\eta_{N_n})(h_1)\}| \leq \sqrt{n}K(1 - \sigma)^{N_n} \rightarrow 0$ as $n \rightarrow \infty$ when $\log(n)/N_n \rightarrow 0$, it follows that

$$\sqrt{n}P_n \{T_{N_n}(\eta_{N_n h})(h_1) - T_{N_n}(\eta_{N_n})(h_1)\} = \sqrt{n}P_0 \{T(\eta_{N_n h})(h_1) - T(\eta_{N_n})(h_1)\} + o_{P_0}(1).$$

Note that $\{T_{N_n}^2(\eta_{N_n})(h_1) | \forall n, h_1 \in H_{10}\}$ is a Glivenko-Cantelli class. It follows that

$$P_n\{T_{N_n}(\eta_{N_n})(h_1)\}^2 = P_0\{T(\eta_0)(h_1)\}^2 + o_{P_0}(1),$$

uniformly in $h_1 \in H_{10}$. Those two results combined with the proof of consistency of the variance estimate in Theorem 1 imply the consistency of the variance estimate.

Table 1
Simulation results for the logistic regression model with missing covariates and auxiliary information.

CM	MDM	Method	β_0			β_1			β_2		
			Bias	Var	Vest	Bias	Var	Vest	Bias	Var	Vest
R	R	CC	0.29	0.071		-0.01	0.132		0.08	0.043	
		SW	0.03	0.062		0.02	0.154		0.03	0.052	
		AW	0.01	0.038		0.00	0.088		0.02	0.021	
		LR-20	0.01	0.034	0.034	0.00	0.075	0.071	0.01	0.019	0.017
		LR-10	0.01	0.034	0.034	0.00	0.075	0.071	0.01	0.019	0.017
		ML	0.01	0.034		0.00	0.075		0.01	0.019	
R	W	CC	0.50	0.068		0.18	0.126		0.11	0.037	
		SW	-0.08	0.061		0.22	0.131		0.22	0.039	
		AW	0.00	0.036		0.01	0.083		0.01	0.018	
		LR-20	0.01	0.034	0.034	0.01	0.076	0.070	0.01	0.017	0.017
		LR-10	0.01	0.034	0.034	0.01	0.076	0.070	0.01	0.017	0.017
		ML	0.01	0.033		0.00	0.075		0.01	0.017	
W	R	CC	0.26	0.070		0.00	0.145		0.02	0.044	
		SW	0.04	0.063		-0.02	0.163		0.03	0.054	
		AW	0.02	0.046		-0.01	0.100		0.02	0.030	
		LR-20	0.02	0.036	0.036	-0.01	0.083	0.079	0.02	0.023	0.021
		LR-10	0.03	0.036	0.036	-0.03	0.082	0.078	0.03	0.023	0.021
		ML	0.10	0.040		-0.17	0.089		0.07	0.025	
W	W	CC	0.36	0.050		0.38	0.118		0.01	0.033	
		SW	-0.42	0.032		0.37	0.117		0.03	0.038	
		AW	-0.03	0.033		0.08	0.087		0.03	0.025	
		LR-20	-0.03	0.032	0.030	0.07	0.077	0.070	0.00	0.020	0.020
		LR-10	-0.04	0.032	0.030	0.08	0.077	0.070	0.00	0.020	0.020

CM	β_0			β_1			β_2		
	Method	Bias	Var	Vest	Bias	Var	Vest	Bias	Var
	ML	-0.07	0.030		0.15	0.080		-0.02	0.020

CM: covariate model used in generating the data; R = linear model without the interaction term; W =linear model with the interaction term. MDM: missing data model used in generating the data; R =logistic model without interaction; W = logistic model with interaction. CC: complete-case analysis. SW: simple weighted estimating equation approach. AW: augmented weighted estimating equation approach. LR-20: Neumann approximation approach with $N = 20$. LR-10: Neumann approximation approach with $N = 10$. ML: the maximum likelihood estimator. Bias: average of estimated values minus the true value. Var: variance of the estimate computed from the simulation replicates. Vest: average of the estimated variances.

Table 2

Simulation results for the linear regression model with missing covariates.

CM	MDM	Method	β_0			β_1			β_2		
			Bias	Var	Vest	Bias	Var	Vest	Bias	Var	Vest
R	R	CC	0.20	0.020		0.03	0.039		-0.04	0.021	
		SW	0.00	0.022		0.00	0.058		-0.01	0.030	
		AW	0.01	0.015		-0.01	0.032		0.00	0.015	
		LR-10	0.01	0.013	0.014	-0.01	0.028	0.029	-0.01	0.012	0.014
		LR-5	0.01	0.013	0.014	-0.01	0.028	0.029	-0.01	0.012	0.014
		ML	0.01	0.013		-0.01	0.028		-0.01	0.012	
R	W	CC	0.00	0.015		0.09	0.023		-0.10	0.014	
		SW	-0.32	0.018		0.06	0.041		-0.07	0.022	
		AW	0.01	0.013		-0.01	0.030		0.00	0.014	
		LR-10	0.01	0.011	0.013	-0.01	0.024	0.027	0.00	0.012	0.013
		LR-5	0.01	0.011	0.013	-0.01	0.024	0.027	0.00	0.012	0.013
		ML	0.01	0.011		-0.01	0.024		0.00	0.012	
W	R	CC	0.22	0.015		0.00	0.032		-0.03	0.009	
		SW	0.01	0.016		-0.02	0.045		-0.01	0.015	
		AW	0.02	0.012		-0.04	0.033		0.01	0.006	
		LR-10	0.02	0.012	0.014	-0.04	0.029	0.033	0.00	0.005	0.007
		LR-5	0.03	0.011	0.014	-0.05	0.028	0.032	0.01	0.005	0.007
		ML	0.08	0.012		-0.16	0.027		0.04	0.006	
W	W	CC	-0.03	0.017		0.22	0.028		-0.15	0.009	
		SW	-0.33	0.023		0.20	0.051		-0.13	0.016	
		AW	-0.04	0.014		0.07	0.050		0.02	0.009	
		LR-10	-0.02	0.011	0.011	0.02	0.028	0.030	-0.01	0.007	0.007
		LR-5	-0.02	0.011	0.011	0.03	0.028	0.030	-0.01	0.007	0.007

CM	MDM	Method	β_0			β_1			β_2		
			Bias	Var	Vest	Bias	Var	Vest	Bias	Var	Vest
		ML	-0.04	0.011		0.08	0.030		-0.02	0.007	

Abbreviations are defined similarly as in Table 1.

Table 3

Analysis of the hip fracture data

Variable	CC	Method			
		CC*	AW*	ML*	LR-10*
Etoh	1.39(0.39)	1.43(0.39)	1.26(0.30)	1.30(0.30)	1.29(0.29)
Smoke	0.93(0.40)	0.96(0.40)	0.76(0.29)	0.74(0.31)	0.73(0.29)
Dementia	2.51(0.72)	2.50(0.73)	1.57(0.46)	1.53(0.47)	1.48(0.47)
Antiseiz	3.31(1.06)	3.43(1.06)	2.66(0.62)	2.62(0.64)	2.62(0.62)
LevoT4	2.01(1.02)	2.10(1.03)	1.01(0.55)	0.91(0.61)	0.96(0.55)
AntiChol	-1.92(0.77)	-2.20(0.79)	-1.77(0.52)	-1.71(0.56)	-1.76(0.54)
BMI	-0.10(0.04)	-0.10(0.04)	-0.10(0.03)	-0.10(0.03)	-0.10(0.03)
logHGB	-2.60(1.20)	-2.21(1.19)	-2.92(0.95)	-2.98(0.94)	-2.99(0.94)
Albumin	-0.91(0.35)	-1.13(0.37)	-1.14(0.33)	-1.15(0.29)	-1.16(0.32)

* : Results are based on roundings of BMI, log(HGB), and Albumin to create 11, 11, and 10 categories respectively. Results from LR-20 are not shown because they are the same as those of LR-10. Standard deviation in the parentheses. Variances for CC and CC* estimators are estimated by information matrix; Variance of AW estimator is estimated by the sandwich approach with adjusting the fact that nuisance parameters are estimated. See Table 1 for the abbreviations.