A semi-parametric estimation of mean functionals with non-ignorable missing data

Jae Kwang Kim Cindy Long Yu *

Abstract

Parameter estimation with non-ignorable missing data is a challenging problem in statistics. The fully parametric approach for joint modeling of the response model and the population model can produce results that are quite sensitive to the failure of the assumed model. We propose a more robust modeling approach by considering the model for the nonresponding part as an exponential tilting of the model for the responding part. The exponential tilting model can be justified under the assumption that the response probability can be expressed as a semi-parametric logistic regression model.

In this paper, based on the exponential tilting model, we propose a semi-parametric estimation method of mean functionals with non-ignorable missing data. A semi-parametric logistic regression model is assumed for the response probability and a non-parametric regression approach for missing data discussed in Cheng (1994) is used in the estimator. By adopting nonparametric components for the model, the estimation method can be made robust. Variance estimation is also discussed and results from a simulation study are presented. The proposed method is applied to real income data from the Korean Labor and Income Panel Survey.

Key Words: Exponential tilting; Not missing at random; Nonparametric regression.

*Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A.

1 INTRODUCTION

Missing data is frequently encountered in many areas of statistics. Statistical analysis in the presence of missing data has been an area of considerable interest because ignoring the missing data often destroys the representativeness of the remaining sample and is likely to lead to biased parameter estimates. To account for the possible bias associated with missing data, statistical modeling is used to predict the missing part of the data. This type of modeling is challenging because it often depends on unverifiable assumptions. Finding a good prediction model is a crucial part of the missing data analysis.

In practice, the prediction model depends on an auxiliary variable. We assume that the auxiliary variable, x, is observed for the entire sample and only the study variable, y, is subject to missingness. In this setup, the usual approach is to find the best prediction model for y in terms of x. The prediction model can be used to predict the missing data if the response mechanism is ignorable in the sense that the relationship between y and x in the respondents also holds for the non-responding part of the sample. Nonresponse is ignorable if the study variable, y, is independent of the response status variable, r, conditional on the auxiliary variable x. Hence, it follows that nonresponse is non-ignorable if the probability of y being missing depends on y itself, even after controlling for x. This situation exists, for example, in surveys of income, where the nonresponse rates tend to be higher for low socio-economic groups. If nonresponse is non-ignorable, standard nonresponse adjustments such as stratification, reweighting, and imputation assuming an ignorable response mechanism will fail to correct for the bias due to nonresponse.

Parameter estimation for non-ignorable nonresponse data is a challenging problem because the response mechanism is generally unknown and the parameters of the response probabilities need to be estimated. In the likelihood-based method, the fully parametric approach involves joint modeling of the outcome and the response mechanism. Greenlees et al (1982) and Diggle and Kenward (1994) used explicit models for the response probability to estimate the parameters. Baker and Laird (1988) and Ibrahim et al (1999) discussed maximum likelihood estimation of the parameters under non-ignorable missing data based on the expectation-maximization algorithm. Molenberghs and Kenward (2007) provided a comprehensive overview of the fully parametric approaches to the analysis of non-ignorable missing data. When the response mechanism is unknown, the identifiability of the parameters in the response mechanism is difficult to check. Chen (2001) and Tang et al (2003) discussed identifiability conditions only under some limited situations. Furthermore, the fully parametric approach is very sensitive to failure of the assumed parametric models (Little, 1985).

In this paper, we propose a novel approach for modeling non-ignorable nonresponse based on the exponential tilting model, where the missing part of the data is modeled as an exponential tilt of the model for the responding part. The tilting parameter, which characterizes the tilt, determines the amount of departure from the ignorability of the response mechanism. The exponential tilting model for non-ignorable nonresponse is similar in spirit to the stratified Cox proportional hazards model considered in Scharfstein et al (1999), which was used to model non-ignorable drop-out in the analysis of longitudinal data. A semi-parametric logistic regression model with the tilting parameter is assumed for the response probability. The behavior of the non-responding part is estimated by using the nonparametric regression approach for missing data discussed by Cheng (1994). By adopting nonparametric parts for the model, the estimation method can be made more robust. Unlike Scharfstein et al (1999), we also consider the case where the tilting parameter is estimated, rather than known. Asymptotic normality, including the $\sqrt{n}$-consistency, of the proposed estimator is derived for the cases when the tilting parameter is estimated as well as known.

In Section 2, a basic setup is introduced. In Section 3, we propose a nonparametric estimation method with known tilting parameters and discuss some asymptotic properties. In Section 4, a semi-parametric estimation method using parametric estimates of tilting parameters is discussed. A simulation study and a case study are given in Section 5 and 6, respectively. Concluding remarks are made in Section 7.

2 BASIC SETUP

Let $(x_i, y_i), i = 1, 2, \cdots, n$, be n independent realizations of continuous random variables (X, Y) from a distribution with joint distribution $F(x, y)$, where x_i is always observed and y_i is subject to missingness. We are interested in estimating $\theta = E(Y)$. Let r_i be the original response indicator for y_i, where $r_i = 1$ if y_i is observed and $r_i = 0$ otherwise. We assume that the response mechanism is

$$r_i \mid (x_i, y_i) \sim Bernoulli(\pi_i),$$

where $\pi_i = \pi(x_i, y_i)$, and r_i is independent of r_j for any $i \neq j$. If π_i does not depend on the value of y_i, then the response mechanism is called ignorable.

Under the ignorable response mechanism or missing at random (MAR) condition,

$$Pr(y_i \in B \mid x_i, r_i = 0) = Pr(y_i \in B \mid x_i, r_i = 1), \tag{1}$$

for any measurable set B. Thus, under MAR, the conditional distribution of y_i given x_i among the nonrespondents is the same as the conditional distribution among the respondents. Let $f_1(y_i \mid x_i)$ be the conditional density of y_i given x_i and $r_i = 1$, and let $f_0(y_i \mid \mathbf{x}_i)$ be the conditional density of y_i given x_i and $r_i = 0$. Under MAR, we have $f_1(y_i \mid x_i) = f_0(y_i \mid x_i)$, and a consistent estimator of θ can be obtained by

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n}\{r_i y_i + (1 - r_i)\hat{m}(x_i)\}, \tag{2}$$

where $\hat{m}(x_i)$ is a consistent estimator of $m(x_i) = E(y_i \mid x_i)$. The consistency of the estimator (2) can be justified under the MAR condition.

If the MAR condition does not hold, then (1) does not hold and the estimator $\hat{\theta}_1$ in (2) is biased. Instead, one can use

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}\{r_i y_i + (1 - r_i)\hat{m}_0(x_i)\},$$

where $\hat{m}_0(x_i)$ is a consistent estimator of $m_0(x_i) = E(y_i \mid x_i, r_i = 0)$. In the absence of the MAR condition, estimation of $m_0(x_i)$ is difficult because y_i is not observed in the set of nonrespondents.

To compute the conditional distribution given $r_i = 0$, we use the following relationship:

$$\begin{aligned} & Pr(y_i \in B \mid x_i, r_i = 0) \\ = \;& Pr(y_i \in B \mid x_i, r_i = 1) \times \frac{Pr(r_i = 0 \mid x_i, y_i \in B) / Pr(r_i = 1 \mid x_i, y_i \in B)}{Pr(r_i = 0 \mid x_i) / Pr(r_i = 1 \mid x_i)}. \end{aligned}$$

Thus, we can write the conditional distribution of the missing data given x as

$$f_0(y_i \mid x_i) = f_1(y_i \mid x_i) \times \frac{O(x_i, y_i)}{E\{O(x_i, Y_i) \mid x_i, r_i = 1\}}, \tag{3}$$

where

$$O(x_i, y_i) = \frac{Pr(r_i = 0 \mid x_i, y_i)}{Pr(r_i = 1 \mid x_i, y_i)} \tag{4}$$

is the conditional odds of nonresponse. The expression (3) is a basis for computing the conditional expectation, $m_0(x_i) = E(y_i \mid x_i, r_i = 0)$.

Assume that the response probability model is a logistic regression model

$$\pi(x_i, y_i) \equiv Pr(r_i = 1 \mid x_i, y_i) = \frac{\exp\{g(x_i) + \phi y_i\}}{1 + \exp\{g(x_i) + \phi y_i\}} \tag{5}$$

for some function $g(\cdot)$ and parameter ϕ. The response probability model (5) is a semi-parametric model in the sense that the component associated with x_i, $g(x_i)$, is completely unspecified and only the component associate with y_i is parametrically modeled as ϕy_i with parameter ϕ. Under the response model (5), the odd function (4) can be written as $O(x_i, y_i) = \exp\{-g(x_i) - \phi y_i\}$ and the expression (3) can be simplified to

$$f_0(y_i \mid x_i) = f_1(y_i \mid x_i) \times \frac{\exp(\gamma y_i)}{E\{\exp(\gamma Y_i) \mid x_i, r_i = 1\}}, \tag{6}$$

where $\gamma = -\phi$. Model (6) states that the density for the nonrespondents is an exponential tilting of the density for the respondents. The parameter γ is the tilting parameter that determines the amount of departure from the ignorability of the response mechanism. In risk theory literature (Gerber and Shiu, 1994), transformation (6) is often called Esscher transformation of $f_1(y_i \mid x_i)$ indexed by parameter γ.

In (3), we need two models to compute the conditional distribution of the nonrespondent: $f_1(y_i \mid x_i)$ and $Pr(r_i = 1 \mid x_i, y_i)$. A consistent estimate of $f_1(y_i \mid x_i)$, denoted by $\hat{f}_1(y_i \mid x_i)$, can be non-parametrically estimated using a kernel estimator. Thus, in the exponential tilting model (6), the only parametric component that needs to be estimated is $\gamma^* = -\phi^*$, where ϕ^* is the true value of ϕ in $Pr(r_i = 1 \mid x_i, y_i; \phi^*)$ in (5). In some cases, such as with planned missingness or a sensitivity analysis as described in Rotnitzky et al (1998), the parameter γ^* is assumed to be known. In the other cases, the parameter γ^* has to be estimated. To estimate the parameter, we often utilize a follow-up study where a further attempt is made to obtain responses in a subset of the nonrespondents. In Section 3, the theory is developed when γ^* is known. In Section 4, we consider the case when γ^* is estimated.

3 Nonparametric estimation

We first briefly discuss a nonparametric regression method for estimating $m_1(x) = E(y \mid x, r = 1)$. Let $K(\cdot)$ be a symmetric density function on the real line and let $h = h_n$ be a smoothing bandwidth such that $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. The nonparametric regression estimator of $m_1(x) = E(y \mid x, r = 1)$ can be obtained by finding $\hat{m}(x)$ that minimizes

$$\frac{\sum_{i=1}^n r_i K_h(x_i, x)\{y_i - m(x)\}^2}{\sum_{i=1}^n r_i K_h(x_i, x)}, \tag{7}$$

where $K_h(u, x) = h^{-1}K\{(u - x)/h\}$. Note that (7) estimates the following quantity

$$E\left[r\{y - m(x)\}^2 \mid x\right] = E\left[E\left\{(y - m(x))^2 \mid x, r = 1\right\} \mid x\right].$$

The function that minimizes (7) is

$$\hat{m}_1(x) = \sum_{i=1}^n w_{i1}(x)\, y_i, \tag{8}$$

where

$$w_{i1}(x) = \frac{r_i K_h(x_i, x)}{\sum_{j=1}^n r_j K_h(x_j, x)}.$$

The weight $w_{i1}(x)$ in (8) represents the point mass assigned to y_i when $m_1(x)$ is approximated by $\sum_{i=1}^n w_{i1}(x) y_i$. The result of Devroye and Wagner (1980) can be used to show that, under some regularity conditions,

$$p \lim_{n\to\infty} \sum_{i=1}^n w_{i1}(x) y_i = \frac{E(rY \mid x)}{E(r \mid x)} = E(Y \mid x, r=1). \tag{9}$$

Cheng (1994) proved $\sqrt{n}$-consistency of $\hat{\theta}_1$ in (2) with $\hat{m}(x) = \hat{m}_1(x)$ using the Kernel-based regression estimator $\hat{m}_1(x)$ in (8) under ignorable missing data.

Under the non-ignorable missing setup described in Section 2 with an exponential tilting model (6), if the true value γ^* were known, the nonparametric regression estimator of $m_0(x) = E(y \mid x, r=0)$ would be

$$\hat{m}_0(x;\gamma^*) = \sum_{i=1}^n w_{i0}(x;\gamma^*) y_i, \tag{10}$$

where the weight

$$w_{i0}(x;\gamma^*) = \frac{r_i K_h(x, x_i)\exp(\gamma^* y_i)}{\sum_{j=1}^n r_j K_h(x, x_j)\exp(\gamma^* y_j)} = \frac{w_{i1}(x)\exp(\gamma^* y_i)}{\sum_{j=1}^n w_{j1}(x)\exp(\gamma^* y_j)}$$

represents the point mass assigned to y_i when $m_0(x)$ is approximated by $\sum_{i=1}^n w_{i0}(x;\gamma^*) y_i$. By the same argument for (9) ,

$$\begin{aligned}
p \lim_{n\to\infty} \sum_{i=1}^n w_{i0}(x) y_i &= \frac{E\{rY\exp(\gamma^* Y) \mid x\}}{E\{r\exp(\gamma^* Y) \mid x\}} \\
&= \frac{E\{\pi(x,Y) Y \exp(\gamma^* Y) \mid x\}}{E\{\pi(x,Y)\exp(\gamma^* Y) \mid x\}} \\
&= \frac{E\{(1-\pi(x,Y)) Y \mid x\}}{E\{1-\pi(x,Y) \mid x\}} \\
&= E(Y \mid x, r=0),
\end{aligned}$$

where $\pi(x,y) = Pr(r=1 \mid x,y)$. Using the nonparametric estimator $\hat{m}_0(x;\gamma^*)$ in (10), a nonparametric estimator of $\theta = E(y)$ is computed by

$$\hat{\theta}_{NP} = \frac{1}{n}\sum_{i=1}^{n}\left\{r_i y_i + (1-r_i)\,\hat{m}_0(x_i;\gamma^*)\right\}. \tag{11}$$

The following theorem, which is similar to Theorem 2.1 of Cheng (1994), presents some asymptotic properties of the estimator in (11). A sketch of the proof is presented in Appendix A.

Theorem 1 *Assume that the response mechanism satisfies the semi-parametric response model (5) with known parameter value ϕ^*. Under the regularity conditions described in Appendix A, the nonparametric estimator $\hat{\theta}_{NP}$ in (11) with $\gamma^* = -\phi^*$ satisfies*

$$\sqrt{n}\left\{\hat{\theta}_{NP} - \theta\right\} \to N\left(0, \sigma_1^2\right) \tag{12}$$

where $\sigma_1^2 = V(\eta_i)$ and

$$\eta_i = m_0(x_i) + \frac{r_i}{\pi(x_i,y_i)}\left\{y_i - m_0(x_i)\right\}. \tag{13}$$

By Theorem 1, since $\eta_i = y_i + \{r_i/\pi(x_i,y_i) - 1\}\{y_i - m_0(x_i)\}$, we have

$$\sigma_1^2 \;=\; V(Y) + E\left[\left\{\frac{1}{\pi(X,Y)} - 1\right\}(Y - m_0(X))^2\right] \tag{14}$$

and the increase in variance due to missing data is

$$V\left(\hat{\theta}_{NP}\right) - V\left(\hat{\theta}_n\right) \;=\; n^{-1}E\left[\left\{\pi(X,Y)^{-1} - 1\right\}(Y - m_0(X))^2\right] \geq 0,$$

where $\hat{\theta}_n = n^{-1}\sum_{i=1}^{n} y_i$. The variance increase is determined by two factors: the inverse of the response probability and the squared error term $\{Y - m_0(X)\}^2$. If the response probabilities for some units are quite small, the variance increase can be quite large. If $\pi(X,Y)$ does not depend on Y, σ_1^2 in (14) reduces to

$$\sigma_1^2 = V(Y) + E\left[\left\{\frac{1}{\pi(X)} - 1\right\}V(Y \mid X)\right] = V\{E(Y \mid X)\} + E\left\{\frac{1}{\pi(X)}V(Y \mid X)\right\}, \tag{15}$$

which is equal to the result of Cheng (1994). Thus, Theorem 1 is an extension of the result of Cheng (1994) to non-ignorable missing data. Wang and Rao (2002) also derived a result similar to (15).

To estimate the variance of the nonparametric estimator $\hat{\theta}_{NP}$, we need to estimate σ_1^2 in (14). A consistent estimator of σ_1^2 is

$$\hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\eta}_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\eta}_i\right)^2, \tag{16}$$

where

$$\hat{\eta}_i = \hat{m}_0\left(x_i;\gamma^*\right) + \frac{r_i}{\hat{\pi}_i}\left\{y_i - \hat{m}_0\left(x_i;\gamma^*\right)\right\}, \tag{17}$$

and $\hat{\pi}_i$ is the estimated response probability of (5) with known γ^*. Writing

$$\alpha\left(x;\gamma^*\right) = O\left(x,y\right)/\exp\left(\gamma^* y\right) = \left(\pi(x,y)^{-1} - 1\right)\exp\left(-\gamma^* y\right),$$

where $O\left(x,y\right)$ is defined in (4), we have

$$E\left\{rO(x,Y) \mid x\right\} = E\left(1 - r \mid x\right) = \alpha(x;\gamma^*)E\left\{r\exp\left(\gamma^* y\right) \mid x\right\}.$$

Thus, under the semi-parametric logistic regression model (5) with known parameter $\gamma^* = -\phi^*$, a non-parametric estimator of $\pi_i = \pi(x_i, y_i)$ can be obtained by $\hat{\pi}_i = \hat{\pi}_i(\gamma^*)$, where

$$\hat{\pi}_i(\gamma) = \left\{1 + \hat{\alpha}\left(x_i;\gamma\right)\exp\left(\gamma y_i\right)\right\}^{-1}, \tag{18}$$

and

$$\hat{\alpha}\left(x_i;\gamma\right) = \frac{\sum_{j=1}^{n}(1 - r_j)K_h(x_i, x_j)}{\sum_{j=1}^{n} r_j \exp\left(\gamma y_j\right) K_h\left(x_i, x_j\right)}.$$

The non-parametric estimator $\hat{\pi}_i$ in (18) can be used to compute the pseudo-value $\hat{\eta}_i$ in (17). Note that the use of $\hat{\eta}_i = \hat{m}_0\left(x_i;\gamma^*\right) + r_i\left\{y_i - \hat{m}_0\left(x_i;\gamma^*\right)\right\}$ is equivalent to the naive variance estimator, which is well known to underestimate the variance. The inflation factor $\hat{\pi}_i^{-1}$ in the residual part of $\hat{\eta}_i$ properly reflects the increase of variance due to missing data. The pseudo-values in (17) bear the same form as those in Carpenter, Kenward and Vansteelandt (2006) under ignorable missing, and are also used for variance estimation in Kim and Rao (2009).

4 Semi-parametric estimation

In many cases, tilting parameter γ^* is unknown and has to be estimated. We now consider a semi-parametric estimator of θ in the sense that we use a parametric component $\hat{\gamma}$ for the nonparametric estimation of $m_0(x;\gamma) = E(Y \mid x, r=0;\gamma)$. We consider two scenarios. The first scenario is the case when the parameter estimate for γ^* is computed from an independent survey. The second scenario is when the parameter estimate is obtained from a validation sample, which is a subsample of the nonrespondents.

In either case, the resulting semi-parametric estimator of θ is

$$\hat{\theta}_{SP} = \frac{1}{n}\sum_{i=1}^{n}\left\{r_i y_i + (1-r_i)\,\hat{m}_0(x_i;\hat{\gamma})\right\}, \tag{19}$$

where $\hat{m}_0(x_i;\gamma)$ is defined in (10). We first consider the scenario where $\hat{\gamma}$ is estimated from an independent survey. The following theorem presents some asymptotic properties of the proposed estimator in (19) for this scenario. A sketch of the proof is in Appendix B.

Theorem 2 *Assume that the conditions of Theorem 1 hold, except that ϕ^* in the response model (5) is known. Let $\hat{\theta}_{SP}$ be the semi-parametric estimator constructed in (19) for the marginal mean of y with $\hat{\gamma}$ satisfying*

$$\sqrt{n}\,(\hat{\gamma}-\gamma^*) \to N(0, V_\gamma), \tag{20}$$

and assume that $\hat{\gamma}$ is independent of $\hat{\theta}_{NP}$ in (11).

Then, we have

$$\sqrt{n}(\hat{\theta}_{SP}-\theta) \to N(0,\sigma_2^2), \tag{21}$$

where

$$\sigma_2^2 = \sigma_1^2 + H^2 V_\gamma, \tag{22}$$

$$H = E\left\{(1-r)\,(Y - m_0(X))^2\right\},$$

and σ_1^2 is defined in (14).

Note that if $\hat{\gamma}$ is exactly estimated, then $V_\gamma = 0$ and σ_2^2 is equal to σ_1^2. Thus, the second term in (22), the increase in variance, is the cost from estimating γ. A consistent estimator of σ_2^2 is

$$\hat{\sigma}_2^2 = \hat{\sigma}_1^2 + \hat{H}^2 \hat{V}_\gamma,$$

where $\hat{\sigma}_1$ is computed using (16), $\hat{V}_\gamma$ is a consistent estimator of $nV(\hat{\gamma})$, and

$$\hat{H} = \frac{1}{n} \sum_{i=1}^{n} (1 - r_i)\, \hat{\sigma}_0^2(x_i),$$

with

$$\hat{\sigma}_0^2(x_i) = \frac{\sum_{j=1}^{n} r_j K_h(x_i, x_j) \exp(\hat{\gamma} y_j) (y_j - \hat{m}_0(x_j))^2}{\sum_{j=1}^{n} r_j K_h(x_i, x_j) \exp(\hat{\gamma} y_j)}.$$

A consistent estimator of σ_1^2 using (16) can be computed by using the pseudo values

$$\hat{\eta}_i = \hat{m}_0(x_i; \hat{\gamma}) + \frac{r_i}{\hat{\pi}_i(\hat{\gamma})} \{y_i - \hat{m}_0(x_i; \hat{\gamma})\},$$

where $\hat{\pi}_i(\gamma)$ is defined in (18).

We now consider the case when a validation sample is randomly selected from the set of nonrespondents and the responses are obtained for all the elements in the validation sample. A consistent estimator $\hat{\gamma}$ of γ^* can be obtained by solving

$$\sum_{i=1}^{n} (1 - r_i)\, \delta_i \{y_i - \hat{m}_0(x_i; \gamma)\} = 0, \tag{23}$$

for γ, where δ_i is an indicator function that takes the value one if unit i belongs to the follow-up sample and takes the value zero otherwise, and $\hat{m}_0(x_i; \gamma)$ is defined in (10).

Using the estimated tilting parameter $\hat{\gamma}$ obtained from (23), one can construct $\hat{\theta}_{SP}$ in (19). The following theorem presents some asymptotic properties of the estimators using the estimated tilting parameter obtained from (23). A sketch of the proof is in Appendix C.

Theorem 3 *Assume that the conditions of Theorem 1 hold, except for the semi-parametric response model in (5). Assume that the solution $\hat{\gamma}$ to (23) exists almost everywhere. Let*

$\hat{\theta}_{SP}$ be the semi-parametric estimator constructed in (19) for the marginal mean of y using $\hat{\gamma}$ obtained by solving (23). Then, we have

$$\sqrt{n}(\hat{\theta}_{SP} - \theta) \to N(0, \sigma_3^2) \tag{24}$$

where $\sigma_3^2 = V(\eta_{2i})$,

$$\eta_{2i} = \tilde{m}(x_i; \gamma_0) + \left\{ \frac{\delta_i}{\nu}(1 - r_i) + r_i \right\} \{y_i - \tilde{m}(x_i; \gamma_0)\},$$

$\tilde{m}(x; \gamma) = p \lim_{n\to\infty} \hat{m}_0(x; \gamma)$, $\nu = E(\delta \mid r = 0)$ and γ_0 is the probability limit of $\hat{\gamma}$.

In Theorem 3, the response model (5) is not needed to show the result (24). The variance σ_3^2 can be written

$$\sigma_3^2 = V(Y) + \left(\nu^{-1} - 1\right) E\left[(1 - r)\{y - \tilde{m}(x; \gamma_0)\}^2\right].$$

Thus, in the extreme case of $\nu = 1$, we have $\sigma_3^2 = V(Y)$.

Note that

$$\tilde{m}(x; \gamma) = p \lim_{n\to\infty} \frac{\sum_{j=1}^n K_h(x, x_j) r_j \exp(\gamma y_j) y_j}{\sum_{j=1}^n K_h(x, x_j) r_j \exp(\gamma y_j)} = \frac{E\{r \exp(\gamma Y) Y \mid x\}}{E\{r \exp(\gamma Y) \mid x\}}.$$

Thus, if the response model (5) is true, then $\gamma_0 = \gamma^*$ and, by (24),

$$\tilde{m}(x; \gamma_0) = \frac{E\{r \exp(\gamma^* Y) Y \mid x\}}{E\{r \exp(\gamma^* Y) \mid x\}} = \frac{E\{(1 - r)Y \mid x\}}{E\{(1 - r) \mid x\}} = E(Y \mid x, r = 0) = m_0(x).$$

Since

$$E\left[(1 - r)\{y - \tilde{m}(x; \gamma_0)\}^2\right] \geq E\left[(1 - r)\{y - m_0(x)\}^2\right],$$

the variance σ_3^2 in (24) is minimized when the assumed response model (5) is true. Thus, the validity of the proposed estimator does not depend on the assumed response model and the role of the response model (5) is to improve the efficiency.

For variance estimation, a consistent estimator of σ_3^2 is

$$\hat{\sigma}_3^2 = \frac{1}{n}\sum_{i=1}^n \hat{\eta}_{2i}^2 - \left(\frac{1}{n}\sum_{i=1}^n \hat{\eta}_{2i}\right)^2,$$

where

$$\hat{\eta}_{2i} = \hat{m}_0\left(x_i;\hat{\gamma}\right) + \left\{\frac{\delta_i}{\nu}\left(1-r_i\right) + r_i\right\}\left\{y_i - \hat{m}_0\left(x_i;\hat{\gamma}\right)\right\}.$$

Instead of using $\hat{\theta}_{SP}$ in (19), one can use the observed values y_i from both the respondents and the follow-up samples directly to get

$$\hat{\theta}_{SP2} = \frac{1}{n}\sum_{i=1}^{n}\left\{r_i y_i + (1-r_i)\delta_i y_i + (1-r_i)\left(1-\delta_i\right)\hat{m}_0\left(x_i;\hat{\gamma}\right)\right\}. \tag{25}$$

By (23), we have

$$\hat{\theta}_{SP2} = \frac{1}{n}\sum_{i=1}^{n}\left\{r_i y_i + (1-r_i)\hat{m}_0\left(x_i;\hat{\gamma}\right)\right\} = \hat{\theta}_{SP}.$$

Thus, the extra information in the follow-up sample is fully incorporated in $\hat{\theta}_{SP}$ and there is no efficiency gain in using $\hat{\theta}_{SP2}$.

5 Simulation Study

To test our theory, we performed a simulation study. In the simulation, we considered two models for generating (x_i, y_i). In model A, the sample of (x_i, y_i) is generated from $x_i \sim N(2,1)$ and $y_i = 1 + 0.7x_i + e_i$ where $e_i \sim N(0,1)$. In model B, (x_i, e_i) are the same as in model A but $y_i = 1 + 0.5(x-2.5)^2 + e_i$. In addition to (x_i, y_i), we also generated r_i, the response indicator variable, from Bernoulli distributions with probability π_i. We considered eight response models for π_i:

(M1): (Linear Ignorable)

$$\pi_i = \frac{\exp(\phi_0 + \phi_1 x_i)}{1 + \exp(\phi_0 + \phi_1 x_i)},$$

where $(\phi_0, \phi_1) = (-1.5, 1.0)$ for both models.

(M2): (Linear Non-ignorable)

$$\pi_i = \frac{\exp(\phi_0 + \phi_1 x_i + \phi_2 y_i)}{1 + \exp(\phi_0 + \phi_1 x_i + \phi_2 y_i)},$$

where $(\phi_0, \phi_1, \phi_2) = (-0.85, 0.3, 0.3)$ for model A and $(\phi_0, \phi_1, \phi_2) = (-1.58, 0.5, 0.7)$ for model B.

(M3): (Non-linear Non-ignorable: quadratic in x)

$$\pi_i = \frac{\exp(\phi_0 + \phi_1 x_i + \phi_2 x_i^2 + \phi_3 y_i)}{1 + \exp(\phi_0 + \phi_1 x_i + \phi_2 x_i^2 + \phi_3 y_i)},$$

where $(\phi_0, \phi_1, \phi_2, \phi_3) = (-2.0, 0.3, 0.3, 0.3)$ for model A and $(\phi_0, \phi_1, \phi_2, \phi_3) = (-2.72, 2.72, -0.68, 0.7)$ for model B.

(M4): (Jump Non-ignorable)

$$\begin{aligned} \pi_i &= 0.5 \quad \text{if } y_i \leq c \\ &= 1.0 \quad \text{if } y_i > c \end{aligned},$$

where $c = 3.4$ for model A and $c = 2.5$ for model B.

(M5): (Non-linear Non-ignorable: quadratic in y)

$$\pi_i = \frac{\exp(\phi_0 + \phi_1 x_i + \phi_2 y_i + \phi_3 y_i^2)}{1 + \exp(\phi_0 + \phi_1 x_i + \phi_2 y_i + \phi_3 y_i^2)},$$

where $(\phi_0, \phi_1, \phi_2, \phi_3) = (-0.65, 0.1, 0.1, 0.1)$ for model A and $(\phi_0, \phi_1, \phi_2, \phi_3) = (-0.85, 0.1, 0.1, 0.3)$ for model B.

(M6): (Probit Non-ignorable)

$$\pi_i = \Phi(\phi_0 + \phi_1 x_i + \phi_2 y_i),$$

where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution, $(\phi_0, \phi_1, \phi_2) = (-0.64, 0.1, 0.3)$ for model A and $(\phi_0, \phi_1, \phi_2) = (-0.53, 0.1, 0.4)$ for model B.

(M7): (Complementary log-log Non-ignorable)

$$\pi_i = 1 - \exp\left\{-\exp(\phi_0 + \phi_1 x_i + \phi_2 y_i)\right\},$$

where $(\phi_0, \phi_1, \phi_2) = (-1.4, 0.3, 0.3)$ for model A and $(\phi_0, \phi_1, \phi_2) = (-1.15, 0.3, 0.3)$ for model B.

(M8): (Non-linear Non-ignorable: interaction)

$$\pi_i = \frac{\exp(\phi_0 + \phi_1 x_i + \phi_2 y_i + \phi_3 x_i y_i)}{1 + \exp(\phi_0 + \phi_1 x_i + \phi_2 y_i + \phi_3 x_i y_i)},$$

where $(\phi_0, \phi_1, \phi_2, \phi_3) = (-1.4, 0.1, 0.1, 0.3)$ for model A and $(\phi_0, \phi_1, \phi_2, \phi_3) = (-0.15, 0.1, 0.1, 0.1)$ for model B.

The missing scenarios considered above include one ignorable missing case (M1) and seven different kinds of non-ignorable missing cases (M2)-(M8). The response rates are about 60% in every combination of the two models and eight response mechanisms. Scenarios (M1)-(M3) satisfy the response probability assumption in (5). Missing mechanisms (M4)-(M8), which do not satisfy (5), were included to examine the robustness of our semi-parametric estimators against failure of the assumed missing mechanism.

For each combination of the two models and eight missing scenarios above, Monte Carlo samples of size $n = 200$ were independently generated $B = 2,000$ times. In each of the sixteen samples, we computed four point estimators:

1. $\hat{\theta}_n = n^{-1} \sum_{i=1}^n y_i$: sample mean of y. Note that $\hat{\theta}_n$ is not used in practice because y_i is not available for $r_i = 0$.

2. $\hat{\theta}_{NA} = n^{-1} \sum_{i=1}^n \{r_i y_i + (1 - r_i)\tilde{m}_0(x_i)\}$: a naive estimator where

$$\tilde{m}_0(x) = \frac{\sum_{i=1}^n (1 - r_i)\delta_i K_h(x, x_i) y_i}{\sum_{i=1}^n (1 - r_i)\delta_i K_h(x, x_i)},$$

   using only the follow-up data.

3. $\hat{\theta}_1$: Cheng's estimator in (2) with $\hat{m}(x) = \hat{m}_1(x)$ in (8). $\hat{\theta}_1$ assumes that missing data from the response mechanism are ignorable.

4. $\hat{\theta}_{SP}$: the semi-parametric estimator in (19) using the estimated tilting parameter $\hat{\gamma}$ obtained in (23). The follow-up rate used is 15%.

The nonparametric Kernel regression estimator was computed using a Gaussian kernel function with bandwidth $h = \hat{\sigma}_x n^{-1/5}$, where $\hat{\sigma}_x$ is the estimated standard deviation of x_i in the sample. The estimated tilting parameter $\hat{\gamma}$ was computed by solving the equation (23) using a Newton-Raphson method.

$<$ Table 1 around here. $>$

$<$ Table 2 around here. $>$

Table 1 and Table 2 present the Monte Carlo relative biases and variances of the four point estimators computed from the Monte Carlo samples of size $B = 2,000$ for missing cases (M1)-(M4) and (M5)-(M8) separately. Mean squared errors (MSE) are reported as well.

Comparing our semi-parametric estimator $\hat{\theta}_{SP}$ to Cheng's estimator $\hat{\theta}_1$, we found that (i) under the ignorable missing mechanism (M1), although the relative biases in $\hat{\theta}_{SP}$ are smaller than those in Cheng's estimator, Cheng's estimator has better performance in terms of MSE since the missing mechanism is correctly specified; (ii) under all the non-ignorable missing mechanisms (M2)-(M8), Cheng's estimator as expected is much more seriously biased than our semi-parametric estimator because Cheng's estimator incorrectly assumes that the response mechanism is ignorable. Although our semi-parametric estimator loses some efficiency due to estimating $\hat{\gamma}$, the serious biases in Cheng's lead to much bigger MSE under all the non-ignorable missing cases.

When comparing our semi-parametric estimator $\hat{\theta}_{SP}$ to the naive estimator $\hat{\theta}_{NA}$, we found that under all the circumstances our semi-parametric estimator performs better than the naive estimator in terms of efficiency and MSE. The efficiency gain in the semi-parametric estimator may be ascribed to the fact that in our semi-parametric estimator the respondent data is used for estimating $m_0(x)$, while the naive estimator utilizes only the follow-up data to estimate $m_0(x)$. It is also noteworthy that our semi-parametric estimator consistently performs reasonably well even in situations when the assumed response probability models are wrong, i.e. (M4)-(M8). This robustness property is consistent with our finding in Theorem 3 that the validity of the proposed estimator does not depend on the assumed response model.

6 Empirical Study

In this section, the proposed semi-parametric estimators are applied to the Korea Labor and Income Panel Survey (KLIPS). A brief description of the panel survey can be found at http://www.kli.re.kr/klips/en/about/introduce.jsp. The data consist of $n = 2,506$ regular wage earners from the year 2008 sample. The study variable (y) is the average monthly income for the current year and the auxiliary variable (x) is the average monthly income for the previous year. The sample mean of (x, y) is $(1.6643, 1.8504) \times 10^6$ Korean Won and the sample correlation between x and y is 0.8144.

From the sample described above, we created artificial missing data by deliberately deleting some of the y values according to the eight response models defined in Section 5. Specifically, we used $(\phi_0, \phi_1) = (-1.13, 1.0)$ for (M1), $(\phi_0, \phi_1, \phi_2) = (-1.5, 0.5, 0.7)$ for (M2), $(\phi_0, \phi_1, \phi_2, \phi_3) = (-2.15, 0.2, 0.5, 0.7)$ for (M3), $c = 2.5$ for (M4), $(\phi_0, \phi_1, \phi_2, \phi_3) = (-0.65, 0.1, 0.1, 0.2)$ for (M5), $(\phi_0, \phi_1, \phi_2) = (-0.41, 0.1, 0.3)$ for (M6), $(\phi_0, \phi_1, \phi_2) = (-1.42, 0.1, 0.7)$ for (M7), and $(\phi_0, \phi_1, \phi_2, \phi_3) = (-0.78, 0.1, 0.1, 0.3)$ for (M8). Each of the eight response mechanisms with the specified parameter values above produced about 60% response rates. Among the nonrespondents, 15% were randomly selected for follow-up samples. Thus, from the original data with sample size of $n = 2,506$, we have about 1,504 respondents and 150 people who responded to the follow-up. Cheng's estimator $\hat{\theta}_1$ and our semi-parametric estimator $\hat{\theta}_{SP}$ were computed using the real data with the artificial missing values for each response probability model.

$<$ Table 3 around here. $>$

Table 3 reports the differences between each point estimate and the "true" sample mean

$\widehat{\theta}_n = 1.8504$ and the estimated standard errors of the point estimators for each of the missing mechanisms. We used the variance estimation formula for $\hat{\theta}_{SP}$ from Section 4 and the estimated variance of $\hat{\theta}_1$ was computed following the approach in Cheng (1994). The estimated mean errors $\widehat{\theta}_1 - \widehat{\theta}_n$ based on Cheng's estimator are consistently larger in magnitude than $\widehat{\theta}_{SP} - \widehat{\theta}_n$ based on our semi-parametric estimator under all the missing scenarios, including the missing ignorable case (M1). This case study demonstrates the empirical effectiveness of our semi-parametric estimator.

7 Conclusion

In the presence of missing data, estimation of $\theta = E(y)$ involves computing the conditional expectation $E(y_i \mid x_i, r_i = 0)$. When the response mechanism is ignorable, Cheng (1994) considered using a nonparametric estimator, $m_1(x_i) = E(y_i \mid x_i, r_i = 1)$, for $E(y_i \mid x_i, r_i = 0)$. If the response mechanism is not ignorable, then the exponential tilting model (6) can be used to derive a consistent estimator of $m_0(x_i) = E(y_i \mid x_i, r_i = 0)$. If the tilting parameter γ^* is known in advance, a non-parametric estimator of $m_0(x_i)$ can be obtained by $\hat{m}_0(x_i; \gamma^*)$ in (10). When the tilting parameter γ^* is unknown, an estimating equation (23) can be used to obtain $\hat{\gamma}$, which can be used to construct semi-parametric estimators in (19) and in (25). The asymptotic properties and the simulation and empirical results presented in this paper show that the semi-parametric estimator provides satisfactory performances in general. Extension to other parameters, such as the population variance, can follow naturally. Extension of the theorems to cases where x is a d-dimensional vector, which is not discussed in this paper, can also be made by choosing the bandwidth $nh^d \to \infty$ instead of $nh \to \infty$. However, as with any nonparametric kernel method, the proposed semi-parametric method can show poor performance for the samples with small sizes or with some extreme missing data patterns.

When a validation subsample is used to estimate the tilting parameters, we assume complete response among the elements in the validation subsample. If there is still missingness in the validation subsample, the estimating equation (23) cannot be used to estimate γ^* and the proposed method is not applicable. In this case, a prior belief about γ^* can be used, as (20), using a Bayesian argument. Investigation of alternative methods for estimating γ^*, including the Bayesian approach, is a topic for future research.

Acknowledgments

We thank two anonymous referees and the associate editor for very helpful comments. The research was partially supported by Cooperative Agreement No. 68-3A75-4-122 between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University.

References

Baker, S. G. and N. M. Laird (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association* **83**, 62-69.

Carpenter, J.R., Kenward, M. G. and Vansteelandt, S. (2006). A Comparison of Multiple Imputation and Doubly Robust Estimation for Analyses with Missing Data. *Journal of the Royal Statistical Society, Series A* **169**, 571-584.

Chen, K. (2001). Parametric models for response-biased sampling. *Journal of the Royal Statistical Society, Series B* **63**, 775-789.

Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association* **89**, 81-87.

Devroye, L.P. and Wagner, T.J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimations, *The Annals of Statistics* **8**, 231-239.

Diggle, P. and Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis (with discussion), *Applied Statistics* **42**, 105-115.

Gerber, H.U. and Shiu, E.S.W. (1994). Option Pricing by Esscher Transforms, *Transactions of the Society of Actuaries*, **46**, pp. 99191.

Greenlees, W.S., Reece, J.S., and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed, *Journal of the American Statistical Association* **77**, 251–261.

Ibrahim, J.G., Lipsitz, S.R., and Chen, M. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable, *Journal of the Royal Statistical Society: Series B* **61**, 173-190.

Kim, J.K. and Rao, J.N.K. (2009). Unified approach to linearization variance estimation from survey data after imputation for item nonresponse, *Biometrika*, **96**, 917-932.

Little, R.J.A. (1985). A note about models for selectivity bias, *Econometrica* **53**, 1469-1474.

Molenberghs, G. and Kenward, M.G. (2007). *Missing data in clinical studies*, John Wiley & Sons: Chichester.

Randles, R. H. (1982). On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics* **10**, 462-474.

Rotnitzky, A., Robins, J. and Scharfstein, D. (1998). Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association* **93**, 1321-1339.

Scharfstein D.O., Rotnitzky A., Robins J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association* **94**, 1096-1120.

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

Wang, Q. and Rao, J.N.K. (2002). Empirical likelihood-based inference under imputation for missing response data. *Annals of Statistics* **30**, 896-924.

Appendix

A: Proof of Theorem 1

Before deriving the asymptotic limit of $\hat{\theta}$, we need the following regularity conditions.

(C.1) The kernel function $K(w)$ is a probability density function such that

(i) it is bounded and has compact support;

(ii) it is symmetric with $\sigma_k^2 = \int w^2 K(w)dw < \infty$;

(iii) $K(w) \geq d_1$ for some $d_1 > 0$ in some closed interval centered at zero.

(C.2) $nh \to \infty$ and $nh^4 \to 0$.

(C.3) $E[y^2]$ and $E[\exp(2\gamma^* y)]$ are finite.

(C.4)

(i) $\pi(x, y) > d_2 > 0$ and $p(x) = E[\pi(x, y)|x] \neq 1$ almost surely.

(ii) The density of X decays exponentially fast.

(iii) $m_0(x)$ has bounded second derivative and satisfies

$$E[\exp(\gamma^* y)|m_0'(x)\alpha'(x) + 0.5 m_0''(x)\alpha(x)|] < \infty,$$

where $\alpha(x) = O(x, y)/\exp(\gamma^* y)$.

Proof: Write

$$\hat{\theta}_{NP} = A + B + C \tag{A.1}$$

where

$$A = n^{-1} \sum_{i=1}^{n} \{r_i m_1(x_i) + (1-r_i) m_0(x_i)\} \tag{A.2}$$

$$B = n^{-1} \sum_{i=1}^{n} r_i \{y_i - m_1(x_i)\} \tag{A.3}$$

$$C = n^{-1} \sum_{i=1}^{n} (1-r_i) \{\hat{m}_0(x_i) - m_0(x_i)\}.$$

By the classical central limit theorem, $\sqrt{n}(A-\theta)$ converges to the normal distribution with mean 0 and variance $V\{rm_1(x) + (1-r)m_0(x)\}$. The term $\sqrt{n}B$ converges to the normal distribution with mean 0 and variance

$$E\left\{r(Y - m_1(X))^2\right\} = E\left\{\pi(X,Y)(Y - m_1(X))^2\right\}.$$

For the C term, note that we can write

$$C = n^{-1} \sum_{i=1}^{n} (1-r_i) \frac{\sum_{j=1}^{n} r_j \exp(\gamma^* y_j) K_h(x_j, x_i) \{y_j - m_0(x_i)\}}{\sum_{j=1}^{n} r_j \exp(\gamma^* y_j) K_h(x_j, x_i)} \tag{A.4}$$

and

$$\begin{aligned} p \lim_{n\to\infty} \frac{\sum_{j=1}^{n} r_j \exp(\gamma^* y_j) K_h(x_j, x)}{\sum_{j=1}^{n} K_h(x_j, x)} &= E\{r \exp(\gamma^* Y) \mid x\} \\ &= E[rO(x,Y) \exp\{g(x)\} \mid x] \\ &= \{1 - p(x)\} \exp\{g(x)\}, \end{aligned}$$

where $O(x,y)$ is defined in (4) and $p(x) = E(r \mid x)$. Thus,

$$p \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} r_j \exp(\gamma^* y_j) K_h(x_j, x) = f(x)\{1 - p(x)\} \exp\{g(x)\},$$

where $f(x)$ is the marginal density of X. Using the same argument for Theorem 2.1 of Cheng (1994), it can be shown that

$$\sqrt{n}(C - C^*) = o_p(1), \tag{A.5}$$

where

$$C^* = n^{-1} \sum_{j=1}^{n} r_j \exp\left(\gamma^* y_j\right) \left\{y_j - m_0\left(x_j\right)\right\} \alpha\left(x_j\right)$$

and

$$\alpha\left(x_j\right) \;=\; E\left[(1-r)\, K_h(x, x_j) \left\{1 - p\left(x\right)\right\}^{-1} \exp\left\{-g(x)\right\} f^{-1}(x) \mid x_j\right] = \exp\left\{-g(x_j)\right\}.$$

Thus, we can write

$$\begin{aligned} C^* &= n^{-1} \sum_{j=1}^{n} r_j O\left(x_j, y_j\right) \left\{y_j - m_0\left(x_j\right)\right\} \\ &= n^{-1} \sum_{j=1}^{n} r_j \left\{\frac{1}{\pi(x_j, y_j)} - 1\right\} \left\{y_j - m_0\left(x_j\right)\right\}. \end{aligned}$$

Thus, inserting (A.5) into (A.1), we have

$$\sqrt{n} \left\{\hat{\theta}_{NP} - \left(A + B + C^*\right)\right\} = o_p\left(1\right).$$

Note that $A+B+C^* = \bar{\eta}_n$, where $\bar{\eta}_n = n^{-1}\sum_{i=1}^{n} \eta_i$ with η_i in (13). Since η_i are independently and identically distributed with mean $E(\eta_i) = \theta$, we have

$$\sqrt{n}\left(\bar{\eta}_n - \theta\right) \to N\left(0, \sigma_1^2\right)$$

and (12) follows by the Slutsky theorem.

B: Proof of Theorem 2

Proof: Using the argument similar to the proof of Theorem 1,

$$\hat{\theta}_{SP} = A + B + C^* + W + o_p\left(n^{-1/2}\right), \tag{A.6}$$

where A, B, and C^* are defined in (A.2), (A.3), and (A.6), respectively, and $W = n^{-1}\sum_{i=1}^{n}(1-r_i)[\hat{m}_0(x_i;\hat{\gamma}) - \hat{m}_0(x_i;\gamma^*)]$. By a Taylor expansion,

$$W = (\hat{\gamma} - \gamma^*)' \frac{1}{n} \sum_{i=1}^{n} (1 - r_i) \frac{\partial \hat{m}_0(x_i;\gamma_1)}{\partial \gamma},$$

where

$$\frac{\partial \hat{m}_0(x_i;\gamma)}{\partial \gamma} = \frac{\sum_{j=1}^n r_j K_h(x_i,x_j)\exp(\gamma y_j) y_j^2}{\sum_{j=1}^n r_j K_h(x_i,x_j)\exp(\gamma y_j)} - \hat{m}_0^2(x_i;\gamma),$$

and γ_1 is in the line segment between $\hat{\gamma}$ and γ^*. Standard arguments used to derive the asymptotic equivalence (A.5) can also be used to show that, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^n (1-r_i)\frac{\partial \hat{m}_0(x_i;\gamma_1)}{\partial \gamma} \to_p E[(1-r)\{y-m_0(x)\}^2|x]. \tag{A.7}$$

Hence, W is asymptotically equivalent to $(\hat{\gamma}-\gamma)E[(1-r)\{y-m_0(x)\}^2]$ and $\sqrt{n}W$ converges to $N(0, H^2 V_\gamma)$. Due to the independence of $\hat{\gamma}$, W is uncorrelated with (A, B, C^*) and the result (21) follows.

C: Proof of Theorem 3

Proof: Writing

$$\hat{\theta}_{SP}(\gamma) = \frac{1}{n}\sum_{i=1}^n \{r_i y_i + (1-r_i)\hat{m}_0(x_i;\gamma)\} + \frac{1}{n}\sum_{i=1}^n (1-r_i)\frac{\delta_i}{\nu}\{y_i - \hat{m}_0(x_i;\gamma)\},$$

we have $\hat{\theta}_{SP}(\hat{\gamma}) = \hat{\theta}_{SP}$ and

$$E\left\{\frac{\partial}{\partial\gamma}\hat{\theta}_{SP}(\gamma) \mid \gamma = \gamma_0\right\} = 0 \tag{A.8}$$

where γ_0 is the probability limit of $\hat{\gamma}$. According to Randles (1982), using $\sqrt{n}(\hat{\gamma}-\gamma_0) = O_p(1)$, we have

$$\hat{\theta}_{SP}(\hat{\gamma}) = \hat{\theta}_{SP}(\gamma_0) + o_p(n^{-1/2}). \tag{A.9}$$

Writing

$$\tilde{m}(x;\gamma) = p\lim_{n\to\infty}\hat{m}_0(x;\gamma) = \frac{E\{r\exp(\gamma Y)Y \mid x\}}{E\{r\exp(\gamma Y) \mid x\}},$$

we have

$$\hat{\theta}_{SP}(\gamma_0) \;=\; n^{-1}\sum_{i=1}^n \{r_i y_i + (1-r_i)\tilde{m}(x;\gamma_0)\} + n^{-1}\sum_{i=1}^n (1-r_i)\frac{\delta_i}{\nu}\{y_i - \tilde{m}(x;\gamma_0)\} + U(\gamma_0),$$

where

$$U(\gamma_0) = n^{-1} \sum_{i=1}^{n} (1 - r_i)(1 - \delta_i/\nu) \{\hat{m}_0(x_i;\gamma_0) - \tilde{m}(x_i;\gamma_0)\}.$$

Because

$$\hat{m}_0(x_i;\gamma_0) - \tilde{m}(x_i;\gamma) = \frac{\sum_{j=1}^{n} r_j K_h(x_i,x_j)\exp(\gamma_0 y_j)\{y_j - \tilde{m}(x_i;\gamma_0)\}}{\sum_{j=1}^{n} r_j K_h(x_i,x_j)\exp(\gamma_0 y_j)},$$

we can apply the same argument for (A.5) to the last term of $U(\gamma_0)$ to get

$$\sqrt{n}\{U(\gamma_0) - U^*(\gamma_0)\} = o_p(1), \tag{A.10}$$

where

$$U^*(\gamma_0) = n^{-1} \sum_{i=1}^{n} r_i \exp(\gamma_0 y_i)\{y_i - \tilde{m}(x_i;\gamma_0)\}\alpha^*(x_i)$$

and

$$\alpha^*(x_i) = \frac{E\{(1-r)(1-\delta/\nu)K_h(x,x_i) \mid x_i\}}{f(x_i)E\{r\exp(\gamma_0 Y) \mid x_i\}}.$$

Because $E\{\delta \mid r = 0, x\} = \nu$, $\alpha^*(x_i) = 0$ and (A.10) reduces to $U(\gamma_0) = o_p(n^{-1/2})$. Therefore, (A.9) reduces to

$$\hat{\theta}_{SP} = n^{-1} \sum_{i=1}^{n} \{r_i y_i + (1 - r_i)\tilde{m}(x;\gamma_0)\} + n^{-1} \sum_{i=1}^{n} (1 - r_i) \frac{\delta_i}{\nu} \{y_i - \tilde{m}(x;\gamma_0)\} + o_p(n^{-1/2}),$$

which proves (24).

Table 1: Monte Carlo relative biases, Monte Carlo variances, and Monte Carlo mean squared errors of the four point estimators for missing scenarios (M1)-(M4) in the simulation study.

Missing Mechanism	Model	Estimates	$\hat{\theta}_n$	$\hat{\theta}_{NA}$	$\hat{\theta}_1$	$\hat{\theta}_{SP}$
$(M1)$	A	Relative Bias	0.0006	-0.0019	0.0157	-0.0008
		Var.	0.0074	0.0220	0.0119	0.0201
		MSE	0.0074	0.0220	0.0133	0.0201
	B	Relative Bias	0.0011	-0.0202	-0.0459	0.0006
		Var.	0.0089	0.0247	0.0129	0.0234
		MSE	0.0089	0.0258	0.0185	0.0234
$(M2)$	A	Relative Bias	-0.0008	-0.0016	0.0542	0.0002
		Var.	0.0075	0.0223	0.0107	0.0194
		MSE	0.0075	0.0223	0.0276	0.0194
	B	Relative Bias	0.0007	-0.0174	0.1479	0.0070
		Var.	0.0088	0.0219	0.0117	0.0176
		MSE	0.0088	0.0227	0.0694	0.0177
$(M3)$	A	Relative Bias	-0.0002	-0.0007	0.0641	0.0011
		Var.	0.0074	0.0217	0.0117	0.0181
		MSE	0.0074	0.0217	0.0354	0.0181
	B	Relative Bias	-0.0001	-0.0351	0.1204	0.0010
		Var.	0.0086	0.0240	0.0114	0.0201
		MSE	0.0086	0.0273	0.0497	0.0201
$(M4)$	A	Relative Bias	0.0005	0.0021	0.0827	0.0011
		Var.	0.0071	0.0174	0.0117	0.0162
		MSE	0.0071	0.0174	0.0511	0.0162
	B	Relative Bias	-0.0001	-0.0066	0.1085	0.0008
		Var.	0.0086	0.0185	0.0132	0.0172
		MSE	0.0086	0.0186	0.0443	0.0172

Table 2: Monte Carlo relative biases, Monte Carlo variances, and Monte Carlo mean squared errors of the four point estimators for missing scenarios (M5)-(M8) in the simulation study.

Missing Mechanism	Model	Estimates	$\hat{\theta}_n$	$\hat{\theta}_{NA}$	$\hat{\theta}_1$	$\hat{\theta}_{SP}$
$(M5)$	A	Relative Bias	0.0008	-0.0007	0.0865	0.0012
		Var.	0.0078	0.0199	0.0111	0.0171
		MSE	0.0078	0.0199	0.0542	0.0171
	B	Relative Bias	-0.0005	-0.0103	0.1727	0.0033
		Var.	0.0086	0.0185	0.0125	0.0164
		MSE	0.0086	0.0188	0.0913	0.0164
$(M6)$	A	Relative Bias	0.0002	-0.0002	0.0836	0.0026
		Var.	0.0078	0.0208	0.0106	0.0173
		MSE	0.0077	0.0208	0.0509	0.0173
	B	Relative Bias	0.0001	-0.0153	0.1437	0.0082
		Var.	0.0091	0.0223	0.0117	0.0180
		MSE	0.0091	0.0229	0.0662	0.0182
$(M7)$	A	Relative Bias	-0.0004	0.0018	0.0798	0.0032
		Var.	0.0075	0.0215	0.0111	0.0178
		MSE	0.0075	0.0215	0.0478	0.0179
	B	Relative Bias	-0.0012	-0.0180	0.0972	0.0003
		Var.	0.0087	0.0222	0.0123	0.0191
		MSE	0.0087	0.0230	0.0372	0.0191
$(M8)$	A	Relative Bias	0.0002	0.0011	0.0979	0.0064
		Var.	0.0079	0.0200	0.0111	0.0163
		MSE	0.0079	0.0200	0.0664	0.0165
	B	Relative Bias	-0.0013	-0.0277	0.0525	-0.0020
		Var.	0.0087	0.0252	0.0114	0.0223
		MSE	0.0087	0.0272	0.0187	0.0223

Table 3: Estimated mean errors and estimated standard errors for the Cheng's estimator ($\hat{\theta}_1$) and our semi-parametric estimator ($\hat{\theta}_{SP}$) in the case study.

Missing	$\widehat{\theta}_1 - \widehat{\theta}_n$	$\widehat{SE}(\widehat{\theta}_1)$	$\widehat{\theta}_{SP} - \widehat{\theta}_n$	$\widehat{SE}(\widehat{\theta}_{SP})$
(M1)	0.0213	0.0219	0.0084	0.0256
(M2)	0.0570	0.0219	0.0089	0.0227
(M3)	0.0640	0.0224	0.0003	0.0223
(M4)	0.0568	0.0231	0.0255	0.0236
(M5)	0.0600	0.0224	-0.0018	0.0230
(M6)	0.0488	0.0222	-0.0133	0.0266
(M7)	0.0864	0.0222	0.0183	0.0229
(M8)	0.0465	0.0223	-0.0118	0.0231