Published in final edited form as:

Biometrics. 2010 March ; 66(1): 115–122. doi:10.1111/j.1541-0420.2009.01231.x.

Adjustment for Missingness Using Auxiliary Information in Semiparametric Regression

Donglin Zeng^{1,*} and Qingxia Chen^{2,**}

¹Department of Biostatistics, University of North Carolina, U.S.A.

²Department of Biostatistics, Vanderbilt University, U.S.A.

Summary

In this article, we study the estimation of mean response and regression coefficient in semiparametric regression problems when response variable is subject to nonrandom missingness. When the missingness is independent of the response conditional on high-dimensional auxiliary information, the parametric approach may misspecify the relationship between covariates and response while the nonparametric approach is infeasible because of the curse of dimensionality. To overcome this, we study a model-based approach to condense the auxiliary information and estimate the parameters of interest nonparametrically on the condensed covariate space. Our estimators possess the double robustness property, i.e., they are consistent whenever the model for the response given auxiliary covariates or the model for the missingness given auxiliary covariate is correct. We conduct a number of simulations to compare the numerical performance between our estimators and other existing estimators in the current missing data literature, including the propensity score approach and the inverse probability weighted estimating equation. A set of real data is used to illustrate our approach.

Keywords

Auxiliary covariate; High-dimensional data; Kernel estimation; Missing at random; Semiparametric regression

1. Introduction

In many survey studies and medical studies, one fundamental statistical problem is to estimate the mean of an interested outcome variable for a given population. For example, in the National Population Health Survey in Canada (NPHSC), one specific interest is the average of the derived health status index in the surveyed population. Other examples include the mean social economic score of the residents in a census tract, the mean arsenic level in the blood of people living in a mine area, and the mean medical cost of hospitalized patients with cardiovascular disease. The estimator would be as simple as the average of all the observations if these observations were actually obtained. However, this is usually impossible in practice, because some surveyed subjects may not respond in many survey studies. Furthermore, the nonresponse or incompleteness in the study population is often not missing completely at random (MCAR) (Little and Rubin, 2002), meaning that the actual

outcomes among the subjects with missing observations are systematically different from the observed values. Therefore, the complete case (CC) method, which simply uses the average of the observed outcome values as the estimator of the population mean, would produce possibly large bias.

In addition to the derived health status index, NPHSC survey also collects information regarding subject's demographics and medical/life history. These include variables such as age, gender, household income, number visits to health provider, a derived type of smoker, a derived depression variable, a summary measure of the chronic conditions, activity restrictions, as well as a derived body mass index (BMI). Likely, such information can be used to explain the difference between subjects with and without complete observation. Because the additional information is not our main interest, it is often termed auxiliary information or auxiliary covariates. If the auxiliary covariates are available for study subjects, it is plausible to adjust for the difference among the complete observations and the missing observations as long as the estimation is performed conditioning on the auxiliary information. For example, suppose that the health status index only varies among groups depending on ages, gender, and BMI but is not related to any other covariates, or suppose that subjects do not respond only because they have different household income; then it is clear that within the subpopulation with the same age, gender, BMI, and household income, the dependence between the health status index and the non-response is minimal. Thus, the average of the actual observations in this group well reflects the true mean outcome in this subpopulation. Consequently, the weighted average across all the subpopulation partitioned based on age, gender, BMI, and household income should provide a reasonably accurate estimator of the whole population mean.

As in the NPHSC study, many studies collect a large amount of auxiliary information. Sometimes it is reasonable to assume that conditional on all the auxiliary information, subjects' outcome and their missingness are independent, i.e., the auxiliary information can be used to adjust for originally nonrandom missingness so that the missing at random (MAR, Little and Rubin, 2002) assumption is plausible. However, the dimension of the auxiliary information is so large that the way described above, which is based on computing the average values within each subpopulation stratified by the auxiliary information, breaks down, simply because there are few subjects in most of the stratified subpopulation. We are faced with a dilemma: we must conduct the partition based on large-dimension auxiliary information to adjust for nonrandom missingness but we do not have enough subjects in each subpopulation to ensure estimation accuracy. Therefore, seeking a way to address this difficulty is the first goal of this article.

Another fundamental statistical problem motivated by our survey data is to estimate the effects of some important variables on the outcome such like the association between the BMI and the health status index in the NPHSC study. When the missingness is nonrandom for the subjects with the same body index, regression analysis based on the CCs leads to biased estimation. Again, to adjust for nonrandom missingness, high-dimensional auxiliary information can be used. However, this poses the same dilemma as described before. Our article will also address this regression problem as its second goal.

Formally, the above two estimation problems can be described using the following notations. Let Y denote the outcome of interest and V denote important predictors. We also denote X as the high-dimensional auxiliary covariates and denote R as the indicator of missing status. $R_i = 1$ if Y_i is observed, and $R_i = 0$ if Y_i is missing. Then the observed data from n independent and identically distributed subjects consist of $\{R_i, R_i Y_i, L_i = (X_i, V_i)\}$, $i = 1, \dots, n$. Assume that R and Y are independent given L , in other words, Y is MAR. Then the first problem is to derive an estimator of the expectation of Y , denoted by $\mu = E[Y]$, and the

second problem is to estimate the effect of V on Y , which is usually specified as the parameter α in the following equation, $E[Y | V] = \alpha^T V$.

Some significant work has been done toward the estimation of μ in the past years. Little and Rubin (2002) discussed the idea of using propensity score (PS), which is $P(R = 1 | L)$. By stratifying the subjects based on the PS, the average of the observed Y values can be calculated for each stratum, and the estimator of μ is the weighted average of these values, where weights are the fractions of the strata. Additionally, a further stratification can be based on the so-called prediction mean score, which is the mean values of $E[Y | R = 1, L]$ in a linear model of Y given L (Little, 1986). Thus, a new weighted average based on the refined stratification can be calculated. Little (1986) concluded that the estimator based on the combined stratifications could reduce the variation of estimator. However, either the method of using PS or the combined stratification method requires consistent estimation of the PS $P(R = 1 | L)$, so their estimators may be biased when the model for estimating the PS is misspecified. In a different context, Lawless, Kalbfleisch, and Wild (1999) derived a semi-parametric likelihood method for missing covariates assuming that the probability of missingness depends only on which of a finite number of strata and that the stratum membership is observed for every unit. Qin, Leung, and Shao (2002) proposed a likelihood based estimation with survey data for nonignorable (Little and Rubin, 1987) nonresponse or informative sampling, where the correct model for the response given the covariates needs to be specified. Additionally, Hu et al. (2007) considered a pseudoscore-based estimation for estimating the distribution parameters of the MAR response. Chen, Zeng, and Ibrahim (2007) proposed a semiparametric likelihood estimation based on fully nonparametric distribution for MAR covariates.

Recently, Scharfstein, Rotnitzky, and Robins (1999) and Robins, Rotnitzky, and van der Laan (2000) proposed an estimator for μ by solving an explicit estimating equation, named the inverse probability weighted estimating equation. Their estimator is a modification of the well-known Thompson–Horvitz estimator and is given as

$$\hat{\mu}_{RR} = \frac{1}{n} \sum_{i=1}^n \left[\frac{R_i Y_i}{\hat{P}(R_i=1|L_i)} - \frac{R_i - \hat{P}(R_i=1|L_i)}{\hat{P}(R_i=1|L_i)} \hat{E}(Y_i|L_i) \right],$$
 where $\hat{P}(R_i=1|L_i)$ is the estimated nonmissing probability based on a model for $[R | L]$ and $\hat{E}(Y_i|L_i)$ is the estimated conditional mean based on another model for $[Y | L]$. Such an estimator was shown to have double robustness property meaning that it is consistent if either the model for $[R | L]$ or the model for $[Y | L]$ is correct. Therefore, the estimator is robust to the misspecification of either of the two models, but not both. The estimator of Scharfstein et al. (1999) thus allows the case that the PS is not consistently estimated. Furthermore, Scharfstein et al. (1999) constructed an estimator for α in our second purpose by solving

$$\sum_{i=1}^n \left[\frac{R_i}{\hat{P}(R_i=1|L_i)} V_i (Y_i - \alpha^T V_i) - \frac{R_i - \hat{P}(R_i=1|L_i)}{\hat{P}(R_i=1|L_i)} V_i (\hat{E}(Y_i|L_i) - \alpha^T V_i) \right] = 0.$$
 It was shown that when either the model $[R | L]$ or the model $[Y | L]$ is correct, the above equation is an estimating equation for estimating α . Other related work in the weighted estimating equation literature is under the missing covariates setting. Among others, Wang and Chen (2001) suggested an augmented weighted estimating equation imposing a parametric model for conditional distribution of the missing covariates for proportional hazards regression. Instead, Qi, Wang, and Prentice (2005) estimated weights nonparametrically, and hence only applied it to low-dimensional covariate setting.

In this article, we propose a new approach to estimate μ and α in the presence of nonrandom missingness. Our idea is very intuitive: we first use two working models $[Y | L]$ and $[R | L]$, respectively to condense the high-dimensional auxiliary information; we then estimate μ and α by optimizing a simple objective function in the covariate space only consisting of the

condensed covariates. Thus, our approach generalizes the idea of the PS approach and the predicted mean score approach. Furthermore, our estimator will be shown to possess the double robustness property. However, as compared to Scharfstein et al.'s method, the estimation in our approach only relies on optimizing some pseudolikelihood functions, so it can be easily generalized to other missing data problems.

The rest of the article consists of the following parts. Section 2 describes the details of estimating $E[Y]$ using our approach and moreover, we give an intuitive explanation why the double robustness holds for the proposed estimator. Section 3 gives the estimation of α as our second goal. In Section 4, we perform a number of simulation studies to compare the numerical performance of our estimator with other existing methods, and a subset of the health data from the 1994 National Population Health Survey in Canada is analyzed using our approach. Most of the technical proofs are given in the appendix available on the website.

2. Inference Procedure for Estimating Mean Outcome

To condense the high-dimensional auxiliary information, we propose two working models. We tentatively assume that the model $[R|L]$ is a generalized linear model with linear predictor $\gamma^T L$, i.e., L predicts R through $\gamma^T L$. We also tentatively assume that the conditional density of Y given $L = l$, denoted as $p(y|l)$, is a parametric density with mean $\beta^T l$. Thus, from the two working models, we obtain a two-dimension vector $(\beta^T L, \gamma^T L)$. In practice, two new parameters (β, γ) introduced in these working models need to be estimated. By performing the generalized linear regression for R given L , we obtain the estimator of γ , denoted as $\hat{\gamma}_n$. Equivalently, $\hat{\gamma}_i$ is obtained by maximizing

$\prod_{i=1}^n [P(R_i=0|L_i)^{1-R_i} P(R_i=1|L_i)^{R_i}]$, where $P(R=1|L=l)$ is the probability of $R=1$ given $L=l$ and $P(R=0|L=l) = 1 - P(R=1|L=l)$. At the same time, the estimator $\hat{\beta}_n$ of β can be

acquired by maximizing $\prod_{i=1}^n p(Y_i|L_i)^{R_i}$ with $p(y|l)$ substituted by the working density. For the convenience of illustration, we particularly assume logit $(P(R=1|L=l)) = \gamma^T l$, and $p(y|l) = (2\pi)^{-1/2} \exp\{-(y - \beta^T l)^2/2\sigma^2\}$. However, our results apply to many commonly used parametric models. Obviously, the following result holds for $\hat{\beta}_n$ and $\hat{\gamma}_n$

Lemma 1: There exist two constants β^* and γ^* such that $\hat{\beta}_n$ and $\hat{\gamma}_n$ converge to β^* and γ^* in

probability, respectively. Moreover, $\sqrt{n}(\hat{\beta}_n - \beta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^n l_\beta(R_i, L_i, R_i Y_i; \beta^*) + o_p(1)$,

$\sqrt{n}(\hat{\gamma}_n - \gamma^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^n l_\gamma(R_i, L_i, R_i Y_i; \gamma^*) + o_p(1)$, where $l_\beta(R, L, RY; \beta^*) =$

$$\{E[RL L^T]\}^{-1} RL(Y - \beta^{*T} L), \quad l_\gamma(R, L, RY; \gamma^*) = \left\{ E \left[\frac{LL^T e^{\gamma^{*T} L}}{(1 + e^{\gamma^{*T} L})} \right] \right\}^{-1} \left(RL - \frac{Le^{\gamma^{*T} L}}{1 + e^{\gamma^{*T} L}} \right),$$

and they are the influence functions associated with $\hat{\beta}_n$ and $\hat{\gamma}_n$ respectively. In addition, if the working model $[Y|L]$ is correct, then β^* is the correct constant for the parameter β in the working model $[Y|L]$; if the working model $[R|L]$ is correct, then γ^* is the correct constant for the parameter γ in the working model $[R|L]$.

The proof is straightforward and given in the appendix. Furthermore, the following lemma shows that when either working model is correct, the two-dimensional covariates obtained in Lemma 1, $Z^* = (\beta^{*T} L, \gamma^{*T} L)$, are truly the condensed covariates meaning that only these two covariates instead of the high-dimensional L are sufficient to explain the dependence between the outcome and the missingness.

Lemma 2: If one of the two working models is correct, then R and Y are independent given Z^* .

The proof is clear from the following arguments. Note $P(R = 1, Y < y | Z^*) = E[P(R = 1, Y < y | L) | Z^*] = E[P(R = 1 | L)P(Y < y | L) | Z^*]$. Thus, if Y depends on L via $\beta^{*T}L$, then $P(Y < y | L)$ is a function of Z^* , and we can get $P(R = 1, Y < y | Z^*) = P(Y < y | Z^*)P(R = 1 | Z^*)$. Similarly, the above equation holds if R depends on L via $\gamma^{*T}L$. We obtain Lemma 2.

From now on, we suppose that one of the two working models is correct. From Lemma 2, we replace the L observations by the observations of Z^* to reduce the dimension of the covariates and obtain the reduced data $(R_i, Y_i, Z_i^*), i = 1, \dots, n$. Because R and Y are independent given Z^* , the observed log-likelihood function for the reduced data is

$$\sum_{i=1}^n \{R_i \ln p(Y_i | Z_i^*) + R_i \ln P(R=1 | Z_i^*) + (1 - R_i) \ln P(R=0 | Z_i^*) + f_{Z^*}(Z_i^*)\}. \quad (1)$$

We propose to estimate $E[Y]$ based on equation (1).

To this end, we first estimate the conditional distribution of Y given $Z^* = z$ by maximizing the local log-likelihood function (Tibshirani and Hastie, 1987) of equation (1) as

$\sum_{i=1}^n R_i K\left(\frac{Z_i^* - z}{h_n}\right) \ln p(Y_i | Z_i^* = z)$, where $K(\cdot)$ is a symmetric kernel function in R^2 and h_n is a bandwidth. This gives that the distribution of Y given $Z^* = z$ is an empirical function at the

observed Y_i $i = 1, \dots, n$, and the mass at Y_i is $R_i K\left(\frac{Z_i^* - z}{h_n}\right) / \sum_{k=1}^n R_k K\left(\frac{Z_k^* - z}{h_n}\right)$. Hence, $E[Y]$

can be estimated by $n^{-1} \sum_{j=1}^n \left[\sum_{i=1}^n Y_i R_i K\left(\frac{Z_j^* - Z_i^*}{h_n}\right) / \sum_{k=1}^n R_k K\left(\frac{Z_j^* - Z_k^*}{h_n}\right) \right]$. We further replace the unknown parameters (β^*, γ^*) in the above expression by their estimators $(\hat{\beta}_n, \hat{\gamma}_n)$ and obtain the final estimator for $E[Y]$ as

$$\hat{\mu}_n = \frac{1}{n} \sum_{j=1}^n \frac{\sum_{i=1}^n Y_i R_i K\left(\frac{\hat{Z}_i - \hat{Z}_j}{h_n}\right)}{\sum_{k=1}^n R_k K\left(\frac{\hat{Z}_j - \hat{Z}_k}{h_n}\right)}, \quad (2)$$

where $\hat{Z}_i = (\hat{\beta}_n^T L_i, \hat{\gamma}_n^T L_i)$.

Theorem 1: Under Assumptions (A.1)–(A.5) (see Web Appendix A for the assumptions), if either working model is correct, $\sqrt{n}(\hat{\mu}_n - E[Y])$ weakly converges to a normal distribution with mean zero. Moreover, $\hat{\mu}_n$ is an asymptotic linear estimator of $E[Y]$ with influence function

$$\begin{aligned}
& E[Y|Z^*] - E[Y] \\
& + \frac{R(Y - E[Y|Z^*])}{E[R|Z^*]} \\
& + \nabla_{\gamma}|_{\gamma=\gamma^*} E \left[\frac{RY}{E[R|\gamma^T L, \beta^{*T} L]} \right] l_{\gamma}(R, L, RY; \gamma^*) \\
& + \nabla_{\beta}|_{\beta=\beta^*} E \left[\frac{RY}{E[R|\gamma^{*T} L, \beta^T L]} \right] l_{\beta}(R, L, RY; \beta^*).
\end{aligned} \tag{3}$$

When both working models are correct, the term $E[RY/E[R|\gamma^T L, \beta^{*T} L]]$ is equal to $E[Y]$, so it is independent of γ . Similarly, the term $E[RY/E[R|\gamma^{*T} L, \beta^T L]]$ is independent of β . Thus, the influence function becomes $E[Y|Z^*] - E[Y] + R(Y - E[Y|Z^*])/E[R|Z^*]$, which can be easily shown to be the efficient influence function (3) of μ_0 . Thus, we conclude that when both working models are correct, the asymptotic variance of $\widehat{\mu}_n$ is the same as the semiparametric efficiency bound. Some literatures name this property as local efficiency. However, this local efficiency should be more accurately called single-point efficiency. In practice, the kernel function $K(\cdot, \cdot)$ usually has little effect on the estimator $\widehat{\mu}$ and we use the Gaussian kernel function in the subsequent numerical studies. However, the choice of the bandwidth h_n is critical and we suggest the Silverman's rule of thumb (Silverman, 1986, Section 3.4.2) in practice. An example of the choice of kernel function $K(\cdot, \cdot)$ and bandwidth h_n is given in Section 4.

The intuition behind Theorem 1 is clear. Note that $E[Y|Z^*] = \int yp(y|Z^*)dy$, where $p(y|Z^*)$ is the conditional density of Y given Z^* . By Lemma 2, when either working model is correct, such conditional density is also equal to the conditional density of Y given Z^* among the CCs (i.e., $R = 1$). Because the latter can be estimated consistently nonparametrically using the method we gave, we conclude that $\widehat{\mu}_n$ should have double robustness property.

Finally, to estimate the asymptotic variance of $\widehat{\mu}_n$, we can either estimate the explicit expression of the influence function in equation (3) using the empirical observations, or adopt the bootstrap method. In our experience, the latter appears to be more accurate with small sample size so will be used in the subsequent numerical studies.

3. Inference Procedure in Semiparametric Regression

In this section, we consider estimating the regression coefficient α in model $E[Y|V] = \alpha^T V$. As before, we introduce the same working models to condense the high-dimensional covariates and try to estimate α using the reduced data of $(R_i, \widehat{Z}_i, V_i, R_i Y_i)$, $i = 1, \dots, n$. The pseudolikelihood function for the reduced data is given as $\prod_{i=1}^n \left\{ p(Y_i|V_i, \widehat{Z}_i)^{R_i} p(V_i, \widehat{Z}_i) \right\}$.

One approach of estimating α is to maximize the likelihood function of the reduced data by considering $E[Y|V] = \alpha^T V$ as a constraint in the model. Either empirical likelihood approach (Owen, 2001) or sieve estimation can be used. However, both approaches require the estimation of the conditional density of Y given V and $Z = (\beta^T L, \gamma^T L)$, and therefore still suffer from the curse of dimensionality when V is not low dimensional.

We provide an alternative estimator for α in this article. Our idea is based on the estimation of mean outcome in the previous section and the fact that for any $\mathcal{V}$ in the support of

V , $E[Y|V \in \mathcal{V}] = \frac{E[I(V \in \mathcal{V})E[Y|V]]}{P(V \in \mathcal{V})} = \alpha^T E[V|V \in \mathcal{V}]$. Therefore, if we can partition the whole support of V into $v_1, \dots, v_m$ for a fixed number m , we can use the observations

associated with each partition to estimate the mean outcome within each partition, denoted by $\widehat{\mu}_1, \dots, \widehat{\mu}_m$. Particularly, for

$$k=1, \dots, m, \widehat{\mu}_k = \left\{ \sum_{i=1}^n I(V_i \in \mathcal{V}_k) \right\}^{-1} \sum_{i=1}^n I(V_i \in \mathcal{V}_k) \frac{\sum_{j=1}^n R_j Y_j I(V_j \in \mathcal{V}_k) K\left(\frac{\widehat{Z}_i - \widehat{Z}_j}{h_n}\right)}{\sum_{j=1}^n R_j I(V_j \in \mathcal{V}_k) K\left(\frac{\widehat{Z}_i - \widehat{Z}_j}{h_n}\right)}.$$

As a result, α can be estimated as the least square estimator of $\{\widehat{\mu}_1, \dots, \widehat{\mu}_m\}$ regressing on

$\{\widehat{v}_1, \dots, \widehat{v}_m\}$, where $\widehat{v}_k$ is the estimator of $E[V \in \mathcal{V}_k]$ and is given by $\widehat{v}_k = \frac{\sum_{i=1}^n V_i I(V_i \in \mathcal{V}_k)}{\sum_{i=1}^n I(V_i \in \mathcal{V}_k)}$. That is,

$$\widehat{\alpha}_n = \left[\sum_{k=1}^m \widehat{v}_k^T \widehat{v}_k \right]^{-1} \left[\sum_{k=1}^m \widehat{v}_k^T \widehat{\mu}_k \right].$$

Obviously, $\widehat{\alpha}_n$ is sensible only if $\sum_{k=1}^m \widehat{v}_k^T \widehat{v}_k$ is nonsingular. In fact, such condition is guaranteed by the following lemma.

Lemma 3: Under (A.1)–(A.5), there exists a partition $v_1, \dots, v_m$ such that $E[V | V \in v_1], \dots, E[V | V \in v_m]$ spans the real space $\mathbb{R}^d$ where d is the dimension of α .

The proof of Lemma 3 is straightforward. Because the probability of V being colinear is less than 1, we can find $u_1, \dots, u_d$ in the support of V such that $u_1, \dots, u_d$ spans the space $\mathbb{R}^d$, i.e., $\min |t_1|^2 + \dots + |t_d|^2 = 1 \|t_1 u_1 + \dots + t_d u_d\| > 0$. By continuity, we can further find a neighborhood of $u_1, \dots, u_d$ denoted by $\tilde{u}_1, \dots, \tilde{u}_d$ such that $\min_{k=1, \dots, d} P(V \in \tilde{u}_k) > 0$ and

$\min_{|t_1|^2 + \dots + |t_d|^2 = 1} \|t_1 \tilde{u}_1 + \dots + t_d \tilde{u}_d\| > 0$ hold for any $\widehat{u}_1 \in \tilde{u}_1, \dots$, and $\widehat{u}_d \in \tilde{u}_d$. Thus, any partition of the support of V including $u_1, \dots, u_d$ satisfies Lemma 3.

The following theorem gives the asymptotic property of $\widehat{\alpha}_n$.

Theorem 2: Suppose that one of the two working models is correct. Under assumptions (A.1)–(A.5), $\widehat{\alpha}_n$ is consistent with true parameter α . Moreover, $\sqrt{n}(\widehat{\alpha}_n - \alpha)$ converges in distribution to a normal distribution with mean zero.

The asymptotic variance of $\sqrt{n}(\widehat{\alpha}_n - \alpha)$ is given in Web Appendix B and similar to the previous section, it can be explicitly estimated. The details are given in the appendix.

We start to discuss the practical implementation of our approach. First, we note that when V is discrete, the partition obviously consists of each level of V . When V includes continuous covariates, the greater number of partitions will lead to more efficient estimators for α . This is because the variability of estimating $\widehat{\mu}_k$ within each partition will be smaller. However, the large number of partitions can make the number of observations in some partitioned domains sparse and leads to large bias in estimating μ 's in these domains. Thus, choosing a partition of the support of V is a tradeoff between bias and variation. As a rule of thumb, we suggest using the quartiles of V to construct the partition and this will be implemented in the subsequent numerical studies. The approach discussed in this section is motivated by low-dimensional V , especially low-dimensional continuous covariates in V . Furthermore, we recommend using bootstrap or jackknife method to estimate the sample variability of $\widehat{\alpha}_n$. The validity of the bootstrap method relies on the asymptotic linear expansion of $\widehat{\alpha}_n$, which is given in the proof of Theorem 2. In practice, when the auxiliary variables contain more

than a couple of discrete variables, the jackknife method is suggested because the bootstrap may yield some sparse cells.

4. Numerical Studies

4.1 Simulation Study I: Estimating Mean Outcome

We conduct extensive simulations to compare the performance of the proposed method with the PS method, the predicted mean stratification (PMS) method, where stratification is based on the predicted outcome model, and the inverse probability weighted estimating equation method (IPWEE). This section focuses on estimating the mean outcome. Especially, we consider two scenarios of generating data: [scenario I]: $Y \sim N(\beta\mathbf{X}, \sigma^2)$, $R \sim \text{Bernoulli}(1/(1 + \exp(-\phi\mathbf{X})))$; $+ \beta_5 X_2^2 \sigma^2) R \sim \text{Bernoulli}(1/(1 + \exp(\phi\mathbf{X} - \phi_5 X_2^2)))$, where $\beta = (\beta_0, \dots, \beta_4)$, $\phi = (\phi_0, \dots, \phi_4)$, $\mathbf{X} = (X_1, \dots, X_4)$, $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 1$, $\sigma^2 = 1$, $\phi_1 = -1$, $\phi_2 = 1$, $\phi_3 = -1$, $\phi_4 = 1$, and $\phi_5 = -0.5$. Thus, the second scenario allows an additional nonlinear effect from X_2 . The auxiliary information X_1 is simulated from $\text{Unif}(0.5, 1)$, X_2 is simulated from $N(1, 1)$, X_3 is from $\text{Bernoulli}(0.3)$, and X_4 is from $\text{Exp}(1)$. We choose specific values of ϕ_0 to obtain either low or high missing percentages. The true mean values for Y are 4.05 and 6.05 in these two scenarios, respectively. In the simulation, the working models of Y and R are parametric models and they are misspecified by ignoring the important covariate X_4 in the first scenario, and by excluding the quadratic term of X_2 in the second scenario.

We apply all four methods to estimate the mean outcome for each simulated sample. Especially, for the PS and PMS methods, the stratification is based on the quartiles of the propensity and prediction mean scores, respectively. To implement our approach, the Gaussian kernel function and the bandwidth following the Silverman's rule of thumb (Silverman 1986, Section 3.4.2) are used in the simulation study. Equivalently, we let $K = (z_1, z_2) = \exp\left[-(z_1^2 + z_2^2)/2\right] (2\pi)$ and $h_n = 0.9 \min(\hat{\sigma}, \text{IQR}/1.34) n^{-1/5}$, where $\hat{\sigma}$ and IQR are the respective standard deviation and the interquartile range of the variable in the kernel estimation.

Table 1 gives the simulation results from 1000 replicates with sample size 250. The table reports the bias (Bias), empirical standard errors (SEE), average of the estimated standard errors using 100 bootstrapped samples (ESE), the coverage rate for the 95% confidence intervals (95%), and the mean square error (MSE) of the estimates using all four methods. As expected, the PS method yields unbiased estimator when the missing mechanism model (model $[R/L]$) is correctly specified but produces biased estimator if it is misspecified. The opposite is true for the PMS method. The simulation results also confirm the double robustness property of both the IPWEE and our method. In all these settings, the estimators from our approach have smaller bias and MSEs than the ones from either PS method or PMS method. Furthermore, when at least $[Y/L]$ is correct, our approach performs as well as the IPWEE. However, when $[Y/L]$ is misspecified and $[R/L]$ is correct, our approach tends to produce estimators with smaller mean square errors than the IPWEE and this is even more transparent when missing rate is higher and the model misspecification is due to ignoring nonlinear effects. For some simulated data with high missing rate, the IPWEE produces extremely large standard errors due to computational instability. Interestingly, we also note that even if both models are misspecified, our approach outperforms all the other methods.

4.2 Simulation Study II: Estimating Regression Coefficients

We further compare the proposed method to IPWEE method when we are interested in estimating important regression coefficients in a semiparametric regression model. Two similar scenarios as the previous study are constructed where Y and R are generated from the

distributions. [Scenario I]: $Y \sim N(\beta\mathbf{X} + \theta_1 V, \sigma^2)$, $R \sim \text{Bernoulli}(1/\{1 + \exp(-\phi\mathbf{X})\})$;

[Scenario II]: $Y \sim N(\beta\mathbf{X} + \beta_5 X_2^2 + \theta_1 V, \sigma^2)$, $R \sim \text{Bernoulli}(1/\{1 + \exp(\phi\mathbf{X} - \phi_5 X_2^2)\})$, where $\beta = (\beta_0, \dots, \beta_4)$, $\phi = (\phi_0, \dots, \phi_4)$, $\mathbf{X} = (X_1, \dots, X_4)$, $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 1$, $\sigma^2 = 1$, $\phi_1 = -1$, $\phi_2 = 1$, $\phi_3 = -1$, $\phi_4 = -1$, $\phi_5 = -0.5$, and $\theta_1 = 1$. We generate X_1 from $\text{Unif}(0.5, 1)$, X_2 from $N(1, 1)$, X_3 from $\text{Bernoulli}(0.5)$, and X_4 from $\text{Exp}(1)$. The covariate of interest, V , is from $N(1.5, 1)$. Clearly, both scenarios imply $E[Y|V] = \theta_0 + \theta_1 V$. The true values of (θ_0, θ_1) are $(4.25, 1)$ and $(6.25, 1)$ in these two scenarios, respectively. The simulation study allows the missing percentages to vary from 20% to 60%. Similar to first simulation, the models of Y and R are misspecified by ignoring the important covariate X_4 in the first scenario, and by excluding the quadratic term of X_2 in the second scenario.

We apply the IPWEE and our approach to estimate θ 's. Particularly, to calculate the estimators using our approach, we first partition the data into four subsets, each representing a possible combination of $R = 0$ or 1, and $X_3 = 1$ or 1. We then combine the four subsets according to the quartiles of V calculated based on the full dataset and each partition of the support V may contain different numbers of the four possible combinations. Table 2 reports the summarized results from 1000 replicates with sample size 250. As seen in Table 2, our proposed method produces double robust estimators to the misspecification of $[R/L]$ or $[Y|L]$ model. The comparison between our approach and the IPWEE indicates similar conclusions as observed in Table 1. In general, our approach performs equally well as the IPWEE when $[Y/L]$ model is correct, but it can be better in turns of the smaller MSEs if $[Y|L]$ model is misspecified. The latter is more significant when missing percentage is higher and nonlinear effect is ignored in the working model. We also implemented our method for V containing two continuous covariates. The sample size to achieve reliable estimation of α will increase as the dimension of V increases due to the partition of the support of V . For multivariate V , we suggest conducting the partition stepwise. For example, when there are two continuous covariates in $V = (V_1, V_2)$, we suggest conducting the partition based on the quantile of V_1 first; then at each subset of dataset created by V_1 quantile, calculate the quantile of V_2 and do further partition based on that. Web Tables 1 and 2 present results of additional simulation when the two covariates are independent or correlated, respectively. The simulation is based on 1000 replicates with 500 sample size. It is observed that our estimators usually have smaller MSEs than IPWEE estimators, although the coverage rate of the bootstrap-based 95% confidence intervals tend to be greater than the nominal level for the regression coefficients of V .

4.3 Application

In this section, we apply the proposed method to a subset of the NPHSC study. The data contain 2389 persons who were aged 20–65 years, lived in a private household of the prairie provinces, and were not pregnant during the survey. Our interested outcome is the derived health status index (hst), which is a continuous variable with higher value as healthier status. The nonresponse rate in the data is 30%. The covariates collected in the survey include both the subject's demographic information such as age (agegrp), gender (sex), household income (houinc), and the subject's medical or life history such as the number visits (visits) to health provider, a derived type of smoker (smoke), a derived depression variable (dvpp), a summary measure of the chronic conditions and activity restrictions (numchron), as well as a derived BMI.

Our primary interest is to estimate the mean of the derived hst. The working model for the missing data mechanism is a logistics regression model, where the covariates contain *agegrp*, *sex*, *houinc*, *dvpp*, *numchron*, *visits*, and the interaction between *sex* and *numchron*. The covariates entering and staying in the model were determined by maximizing the value of area under curve (AUC), which measures goodness of fit for logistic regression. The

AUC value for the final working model is 0.908, implying a good fit of the working model. On the other hand, the working model for the outcome, health status index, is a linear regression model with the covariates *agegrp*, *sex*, *houinc*, *BMI*, five dummy variables indicating smoking status, *dvpp*, *numchron*, *visits*, the two-way interactions between *agegrp* and *sex*, between *agegrp* and *numchron*, between *numchron* and *visits*, and three-way interactions among *agegrp*, *numchron* and *visits*. The covariates kept in the model were chosen by maximizing adjusted *R*-square of the linear regression model.

Especially, we estimate the mean health status index using the CC analysis, PS, PMS, IPWEE, and the proposed method. In Figure 1, we plot the means and the 95% confidence intervals of the derived health status index using CC, PS, PMS, IPWEE, and the proposed method. We find that among the five methods, CC estimator is higher than all the other four methods, although the difference may not be significant at the 0.05 level. Our proposed method yields similar estimate as PS and PMS but with slightly larger standard error. Comparatively, the IPWEE produces the smallest mean estimator but the largest standard error.

For illustration, we also study the relationship between the derived health status index and the BMI. Particularly, we assume a marginal linear model for the health status index given the body mass index. Table 3 shows the results of estimates, standard errors and the associated *P*-values from the CC method, the IPWEE method, and our method. Similar conclusions can be drawn using all three methods that the higher BMI is associated with the lower health status index. However, our method gives a more significant result.

5. Discussion

We have proposed a new approach of estimating the mean outcome and the coefficient in a marginal regression model in the presence of nonrandom missingness. Our estimators have double robustness property and perform as well as or sometimes better than other competitive methods.

The two working models for $[Y | L]$ and $[R / L]$ are useful and critical to condense high-dimensional auxiliary information, which make our estimation feasible. The linear working models used in this article are mainly used for practical convenience. Other general working models can be used, for example, generalized additive models, single-index models, etc. Clearly, the more general models we use, the more likely that our estimators are consistent. However, too general working models may result in moderate size condensed covariates not being estimated well, and that the constructed estimators may not converge at the $\sqrt{n}$ -rate. We are currently studying a reasonable criterion to compare across a wide class of working models.

Furthermore, it is also important conducting the model checking for the two working models. For the linear model $[Y | L]$, one may use the model residuals to construct the goodness of fit, such as the adjusted *R*-square in our real data analysis, whereas for the model $[R / L]$, one goodness-of-fit measure can be the area under the curve. When *L* is very high dimensional, some methods of dimension reduction or variable selection can also be of use.

Our construction of the double robust estimator is not the only way. In fact, it is intuitive to see that the kernel estimation of $[Y | Z^*]$ can be treated as one way of imputing the missing outcome using the observed values in the subpopulation stratified by Z^* . Obviously, there exists a number of alternative ways for imputation, for instance, histogram-type estimator, Monte Carlo imputation, multiple imputation, etc. Because all the imputation is

approximately conditional on Z^* , it is not difficult to show that the derived estimators should also have the double robustness property. However, the numerical performance of all these possible estimators remain unknown and is currently under investigation.

The partition idea in this type of regression problem is new, to the best of our knowledge. The optimal partition that balances the bias and variation is an open problem. Practically, the idea of the classification tree may be used to derive an “optimal” partition. In particular, the leave-one-out crossvalidation can be used to calculate the error of one particular partition:

given a partition $v_1, \dots, v_m$, we calculate $\hat{\alpha}_n^{(-1)}, \dots, \hat{\alpha}_n^{(-m)}$, where $\hat{\alpha}_n^{(-k)}$ is the least square estimator using the partition except v_k , $k = 1, \dots, m$. The crossvalidation error is defined as

$CV(\mathcal{V}_1, \dots, \mathcal{V}_m) = \frac{1}{m} \sum_{k=1}^m \{\hat{\mu}_k - \hat{\alpha}_n^{(-k)} v_k\}^2$. An additional penalty like the pruning tree can be used to avoid overpartition of V -space. The formalization of this method will be reported in separate work.

Finally, both the idea of estimating mean outcome and the idea of estimation in regression are ready to be generalized with many other missing data problems, including missing covariates, correlated data, intermittent missing data, etc.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to the editor, associate editor, and two anonymous referees for very useful comments that improved the presentation of the article. The research of DZ was supported by the National Institutes of Health grants R01 CA082659-10 and R01 HL57444-11.

References

- Chen Q, Zeng D, Ibrahim JG. Sieve maximum likelihood estimation for regression models with covariates missing at random. *Journal of the American Statistical Association*. 2007; 102:1309–1317.
- Hu XJ, Schroeder RJ, Wang WC, Boyett JM. Pseudoscore-based estimation from biased observations. *Statistics in Medicine*. 2007; 26:2836–2852. [PubMed: 17133648]
- Lawless JF, Kalbfleisch JD, Wild CJ. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society, Series B*. 1999; 61:413–438.
- Little RJA. Survey nonresponse adjustments. *International Statistical Review*. 1986; 54:139–157.
- Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. 2nd edition. John Wiley & Sons; New York: 2002.
- Mack Y, Silverman B. Weak and strong uniform consistency of kernel regression estimates. *Probability Theory and Related Fields*. 1982; 60:405–415.
- Owen, AB. *Empirical Likelihood*. Chapman & Hall; New York: 2001.
- Qi L, Wang CY, Prentice RL. Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association*. 2005; 100:1250–1263.
- Qin J, Leung D, Shao J. Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Association*. 2002; 97:193–200.
- Robins JM, Rotnitzky A, Van Der Laan M. Comment on “On profile likelihood”. *Journal of the American Statistical Association*. 2000; 95:477–482. by Murphy, S. and van der Vaart, A. W.
- Scharfstein DO, Rotnitzky A, Robins JM. Rejoinder to adjusting for non-ignorable drop-out using semiparametric non-response models. *Journal of the American Statistical Association*. 1999; 94:1135–1146.
- Silverman, BW. *Density Estimation*. Chapman and Hall; New York: 1986.

- Tibshirani R, Hastie T. Local likelihood estimation. *Journal of the American Statistical Association*. 1987; 82:559–567.
- Van Der Vaart, AW.; Wellner, JA. *Weak Convergence and Empirical Processes*. Springer-Verlag; New York: 1996.
- Wang CY, Chen HY. Augmented inverse probability weighted estimator for Cox regression missing covariate regression. *Biometrics*. 2001; 57:414–419. [PubMed: 11414564]
- Wong WH. Theory of partial likelihood. *Annals of Statistics*. 1986; 14:88–123.

Biometrics. Author manuscript; available in PMC 2011 August 2.

Table 1

Results from simulations on estimating mean response

$c\%$	Y	R	PS			PMS			IPWEE			Our approach		
			Bias	CP	MSE	Bias	CP	MSE	Bias	CP	MSE	Bias	CP	MSE
1 ^a	19	T	0.037	92.3	0.017	0.041	92.7	0.017	0.005	93.8	0.016	0.010	94.0	0.016
		F	0.139	79.4	0.036	0.041	92.7	0.017	0.006	93.5	0.015	0.013	93.8	0.016
		T	0.037	92.3	0.017	0.140	79.2	0.036	0.005	93.6	0.016	-0.001	93.9	0.016
		F	0.139	79.4	0.036	0.140	79.2	0.036	0.125	81.7	0.033	0.129	82.1	0.033
60		T	0.093	91.1	0.047	0.124	86.8	0.044	0.013	94.3	0.032	0.045	91.5	0.032
		F	0.601	22.2	0.408	0.124	86.8	0.044	0.012	94.7	0.027	0.053	90.7	0.032
		T	0.093	91.1	0.047	0.604	16.2	0.410	0.029	92.2	0.051	0.056	92.0	0.035
		F	0.601	22.2	0.408	0.604	16.2	0.410	0.579	23.1	0.387	0.585	20.3	0.392
24		T	-0.103	92.1	0.065	-0.108	92.4	0.063	0.010	95.2	0.056	-0.027	94.2	0.056
		F	-0.215	82.2	0.097	-0.108	92.4	0.063	0.010	95.4	0.055	-0.023	94.0	0.055
		T	-0.103	92.1	0.065	-0.135	89.9	0.071	-0.013	94.4	0.146	-0.054	93.5	0.058
		F	-0.215	82.2	0.097	-0.135	89.9	0.071	-0.225	81.2	0.104	-0.080	92.6	0.061
41		T	-0.120	90.9	0.081	-0.129	90.8	0.074	0.017	95.5	0.072	-0.043	94.8	0.058
		F	-0.287	74.8	0.140	-0.129	90.8	0.074	0.013	95.6	0.057	-0.039	94.7	0.058
		T	-0.120	90.9	0.081	-0.173	87.4	0.087	0.030	94.1	0.094 ^c	-0.086	93.2	0.065
		F	-0.287	74.8	0.140	-0.173	87.4	0.087	-0.308	72.6	0.151	-0.125	91.3	0.073

^aThe misspecification means that covariate X_4 is left out of the model.^bThe misspecification means nonlinear term X_2^2 is left out of the model.^cObtained after excluding one extreme sample; however, if included, the value becomes 3.623.

Table 2

Results from simulations on estimating regression coefficient

c%	Y	R	Par.	IPWEE					Our approach				
				Bias	SEE	ESE	95%	MSE	Bias	SEE	ESE	95%	MSE
I	22	T	θ_0	0.011	0.244	0.239	93.1	0.060	0.010	0.239	0.251	95.2	0.057
			θ_1	-0.008	0.136	0.135	93.9	0.019	0.003	0.134	0.143	97.6	0.018
		F	θ_0	0.007	0.238	0.235	93.3	0.057	0.016	0.238	0.248	95.1	0.057
			θ_1	-0.007	0.133	0.133	94.1	0.018	0.002	0.135	0.141	97.8	0.018
	44	T	θ_0	0.012	0.260	0.255	92.8	0.068	0.012	0.243	0.253	95.4	0.059
			θ_1	-0.010	0.141	0.1410	93.2	0.020	0.002	0.136	0.144	97.9	0.019
		F	θ_0	0.148	0.254	0.248	87.4	0.086	0.012	0.243	0.253	95.4	0.059
			θ_1	-0.008	0.143	0.140	93.7	0.020	-0.001	0.146	0.150	97.9	0.021
	64	T	θ_0	0.070	0.483	0.433	88.5	0.238	0.087	0.408	0.418	93.9	0.174
			θ_1	-0.049	0.278	0.247	88.2	0.080	-0.000	0.224	0.240	98.4	0.050
		F	θ_0	0.068	0.407	0.375	90.4	0.170	0.111	0.402	0.404	93.8	0.174
			θ_1	-0.049	0.226	0.208	88.4	0.054	-0.002	0.221	0.233	97.6	0.049
II	27	T	θ_0	0.067	0.695	0.637	87.1	0.488	0.135	0.416	0.431	93.9	0.192
			θ_1	-0.040	0.354	0.323	89.5	0.127	0.001	0.230	0.248	98.6	0.053
		F	θ_0	0.678	0.493	0.460	60.3	0.702	0.652	0.484	0.483	70.6	0.660
			θ_1	-0.045	0.284	0.255	89.8	0.082	-0.005	0.277	0.276	96.6	0.077
	45	T	θ_0	0.014	0.449	0.448	95.5	0.202	-0.110	0.457	0.451	94.1	0.221
			θ_1	-0.008	0.249	0.249	93.8	0.062	0.014	0.262	0.258	97.1	0.069
		F	θ_0	0.002	0.436	0.4360	95.1	0.190	-0.103	0.455	0.448	93.4	0.217
			θ_1	-0.001	0.242	0.239	93.9	0.059	0.013	0.259	0.256	97.0	0.067
	63	T	θ_0	-0.019	0.601	0.571	94.6	0.361	-0.142	0.458	0.452	92.5	0.230
			θ_1	-0.003	0.350	0.323	95.1	0.122	0.013	0.262	0.258	97.2	0.069
		F	θ_0	-0.252	0.434	0.423	89.5	0.252	-0.168	0.458	0.448	92.4	0.238
			θ_1	-0.001	0.242	0.236	94.8	0.059	0.013	0.262	0.256	97.1	0.069
	81	T	θ_0	0.058	0.490	0.487	94.3	0.243	-0.137	0.473	0.477	92.9	0.243
			θ_1	-0.031	0.287	0.273	93.7	0.083	0.007	0.270	0.273	97.5	0.073

c%		Y		R	Par.	IPWEE					Our approach				
						Bias	SEE	ESE	95%	MSE	Bias	SEE	ESE	95%	MSE
T	F	θ_0				0.034	0.461	0.459	94.7	0.214	-0.129	0.472	0.473	92.7	0.239
						-0.018	0.256	0.256	94.6	0.066	0.005	0.270	0.270	97.1	0.073
F	T	θ_0				-0.031	0.593	0.635	94.4	0.353 ^a	-0.197	0.472	0.478	92.3	0.261
						-0.001	0.370	0.387	94.6	0.137 ^b	0.009	0.271	0.273	97.3	0.074
F	F	θ_0				-0.314	0.457	0.452	88.2	0.308	-0.234	0.473	0.472	91.2	0.279
						-0.013	0.255	0.254	95.0	0.065	0.009	0.272	0.269	97.0	0.074

^a Obtained after excluding one extreme sample; however, if included, the value becomes 0.643.

^b Obtained after excluding one extreme sample; however, if included, the value becomes 0.431.

Table 3

Regression coefficients

Method	Coefficient	Estimate	Standard error	P-value
CC	(Intercept)	0.950	0.016	<0.001
	BMI	−0.002	0.001	0.018
IPWEE	(Intercept)	0.974	0.034	<0.001
	BMI	−0.003	0.001	0.033
Our approach	(Intercept)	0.987	0.028	<0.001
	BMI	−0.003	0.001	0.003