Nonparametric Regression with Missing Outcomes Using Weighted Kernel Estimating Equations

Lu Wang, Andrea Rotnitzky and Xihong Lin *

Abstract

We consider nonparametric regression of a scalar outcome on a covariate when the outcome is missing at random (MAR) given the covariate and other observed auxiliary variables. We propose a class of augmented inverse probability weighted (AIPW) kernel estimating equations for nonparametric regression under MAR. We show that AIPW kernel estimators are consistent when the probability that the outcome is observed, i.e., the selection probability, is either known by design or estimated under a correctly specified model. In addition, we show that a specific AIPW kernel estimator in our class that employs the fitted values from a model for the conditional mean of the outcome given covariates and auxiliaries is double-robust, i.e. it remains consistent if this model is correctly specified even if the selection probabilities are modeled or specified incorrectly. Furthermore, when both models happen to be right, this double-robust estimator attains the smallest possible asymptotic variance of all AIPW kernel estimators and maximally extracts the information in the auxiliary variables. We also describe a simple correction to the AIPW kernel estimating equations that while preserving double-robustness it ensures efficiency improvement over non-augmented IPW estimation when the selection model is correctly specified regardless of the validity of the second model used in the augmentation term. We perform simulations to evaluate the finite sample performance of the proposed estimators, and apply the methods to the analysis of the AIDS Costs and Services Utilization Survey data. Technical proofs are available online.

Key Words: Asymptotics; Augmented kernel estimating equations; Double robustness; Efficiency; Inverse probability weighted kernel estimating equations; Kernel smoothing

*Lu Wang (email: luwang@umich.edu) is Assistant Professor, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109; Andrea Rotnitzky (email: andrea@utdt.edu) is Professor, Department of Economics, Di Tella University, Buenos Aires, 1425, Argentina and Adjunct Professor, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115 and Xihong Lin (email: xlin@hsph.harvard.edu) is Professor, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115. Wang and Lin's research is partially supported by a grant from the National Cancer Institute (R37-CA–76404). Rotnitzky's research is partially supported by grants R01-GM48704 and R01-AI051164 from the National Institutes of Health.

1 INTRODUCTION

The existing missing data literature mainly focuses on estimation methods in parametric regression models, i.e. models for the conditional mean of an outcome given covariates indexed by finite dimensional regression parameters. However, the functional form of the dependence of an outcome on a covariate is often unknown in advance and can be complicated (Hastie and Tibshirani 1990; Wand and Jones 1994). For example, Zhang et al. (2000) found that the profile of progesterone level during a menstrual cycle follows a nonlinear pattern which is hard to fit using standard parametric models and is best fitted by non-parametric smoothing techniques. Likewise, Harezlak, et al. (2007) found that the protein intensities from mass spectrometry are very complex and need to be fit using nonparametric smoothing methods. Limited literature is available for nonparametric regression in the presence of missing data.

Our work is motivated by the AIDS Costs and Services Utilization Survey (ACSUS) (Berk et al. 1993). The ACSUS sampled subjects with AIDS in 10 randomly selected U.S. cities with the highest AIDS rates. A question of interest in this study is how the risk of hospital admission one year after study enrollment is related to the baseline CD4 counts. Although it is known that a lower CD4 count is associated with a higher risk of hospitalization, the functional form of dependence is unknown and expected to be nonlinear with a potential threshold. We are hence interested in modeling this relationship nonparametrically. However, about 40% of the patients did not have the first year hospital admission data available. As shown in Section 4, naive nonparametric regression using complete data only could yield an inconsistent estimator of the mean curve if the missing is not completely at random, a likely situation in this problem. It is therefore of interest to develop flexible nonparametric regression methods to estimate the effect of baseline CD4 counts on the risk of hospitalization that adequately adjust for outcomes missing at random (MAR), i.e. missing depending on observed data (Little and Rubin 2002). In addition, because the fraction of missing outcomes is large, it is also important that the methodology maximally exploits the information in available auxiliary variables. The methods we develop in this paper are also useful for nonparametric regression estimation in two-stage studies (Pepe 1992), where the second-stage outcome is not observed for all study units and the probability of observing the outcome depends on the first-stage auxiliaries and covariates, but is independent of the outcome, i.e. it is MAR.

Limited work has been done on nonparametric regression in the presence of missing data. Wang et. al. (1998) considered estimation of a non-parametric regression curve with missing covariates. Liang et. al. (2004) considered estimation of a partially linear model with missing covariates

and described inverse probability weighted (IPW) estimation of the non-parametric component of the model. Chen et. al. (2006) studied local quasi-likelihood estimation with missing outcomes when missingness depends only on the regression covariate. None of these articles considered, as we do here, the possibility that always observed auxiliaries are available, a case that arises often in practice. Our work differs in that we propose augmented inverse probability weighted (AIPW) kernel estimators that exploit the information in the auxiliary variables while at the same time allowing for the possibility that missingness may depend on them, thus making the MAR assumption more plausible.

In this paper we generalize kernel estimating equation methods (Wand and Jones 1995; Fan and Gibjels 1996; Carroll et al. 1998) to accommodate outcomes missing at random in a similar spirit to IPW and AIPW methods for parametric regression (Robins et al. 1995; Rotnitzky and Robins 1995; Robins et al. 1994; Rotnitzky et al. 1997; Robins 1999). After studying the properties of naive kernel estimating equations based on complete cases, we propose the IPW kernel estimating equations and a class of AIPW kernel estimating equations. We present the asymptotic properties of the solutions to these weighted kernel estimating equations and compare them in terms of asymptotic biases and variances. We argue that clever choices of the augmentation term can yield important efficiency gains over the IPW kernel estimators. The proposed IPW and AIPW kernel estimators are consistent under MAR if the missingness mechanism is known by design or can be parametrically modeled. Indeed, with one specific choice of the augmentation term, the AIPW kernel estimator confers some protection against model misspecification in that it remains consistent even if the model for the missingness probabilities is misspecified provided that a parametric model for the conditional mean of the outcome given the covariates and auxiliaries is correctly specified, a property known as double-robustness.

2 THE GENERALIZED NONPARAMETRIC MODEL WITH MISSING OUTCOMES

We consider a generalized nonparametric mean model when the outcome may be missing at random. Specifically, suppose the study design calls for a vector of variables $(Y_i, Z_i, \boldsymbol{U}_i)$ to be measured in each subject i of a random sample of n subjects from a population of interest. The variable Y_i denotes the outcome which may not be observed in all subjects and the variable Z_i denotes a scalar covariate that is always observed. We assume that the mean of Y_i depends on Z_i through

a generalized nonparametric model

$$g(\mu_i) = \theta(Z_i), \tag{1}$$

where $g(\cdot)$ is a known monotonic link function (McCullagh and Nelder, 1989) with a continuous first derivative, $\mu_i = E\left(Y_i|Z_i\right)$, and $\theta(z) = g\left\{E\left(Y|Z=z\right)\right\}$ is an unknown smooth function of z that we wish to estimate. The variables $\boldsymbol{U}_i$, which we assume are always observed, are recorded in the dataset for secondary analyses. However, for our purposes they are regarded as auxiliary variables as we are not interested in estimation of $E\left(Y_i|Z_i, \boldsymbol{U}_i\right)$, but rather in estimation of $E\left(Y_i|Z_i\right)$. The covariates $\boldsymbol{U}_i$ are nevertheless useful in that they can both help explain the missing mechanism and improve the efficiency with which we estimate the nonparametric function $\theta(\cdot)$.

We assume that outcomes are missing at random (MAR) (Little and Rubin 2002), which in our setting amounts to assuming that

$$Pr(R_i = 1|Z_i, \boldsymbol{U}_i, Y_i) = Pr(R_i = 1|Z_i, \boldsymbol{U}_i) \tag{2}$$

where $R_i = 1$ if Y_i is observed and $R_i = 0$ otherwise. That is, we assume the probability that the outcome is missing may depend on the observed data, i.e. covariates and auxiliaries, but is independent of the outcome given the observed data. This assumption automatically holds in two stage sampling designs (Pepe 1992; Reilly and Pepe 1995) with covariates and auxiliaries measured at the first stage and outcomes measured on a subsample at the second stage. Using probabilities of selection into the second stage that depend on the variables collected at the first stage can help improve the efficiency with which one estimates the regression of Y on Z (Breslow and Cain 1988).

3 THE KERNEL ESTIMATING EQUATIONS FOR MISSING OUTCOMES AT RANDOM

In the absence of missing data, local polynomial kernel estimating equations have been proposed by Carroll et al. (1998) as an extension of local likelihood estimation. When the data are not fully observed, one naive estimation approach is to simply solve the local polynomial kernel estimating equations using only completely observed units. However, as we show in Theorem 1 in Section 4, the resulting estimator $\widehat{\theta}_{naive}\left(z\right)$ is generally inconsistent under MAR, except when: a) the conditional mean of $E(Y|Z, \boldsymbol{U})$ depends at most on Z or, b) the selection probability $Pr(R = 1|Z, \boldsymbol{U})$ depends at most on Z. This result is not surprising once we connect our inferential problem to causal inference objectives and relate it to well known facts in causality. The MAR assumption (2) is equivalent to the assumption of no unmeasured confounding (Robins et al.

1999) or ignorability (Rubin, 1976) for the potential outcome under treatment $R = 1$ in the subpopulation with $Z = z$. This assumption stipulates that, conditional on $Z = z$, $\boldsymbol{U}$ are the only variables that can simultaneously be i) correlates of the outcome within treatment level and ii) predictors of treatment $R = 1$. When a) or b) holds, either i) or ii) is violated. In such case, the effect of $R = 1$ on Y is unconfounded and consequently naive conventional, i.e. unadjusted, estimators of the association of Y with $R = 1$ conditional on $Z = z$ are consistent estimators of the causal estimand of interest. In fact, when b) holds but a) is false, the naive estimator will be consistent but inefficient because it fails to exploit the information about $E(Y|Z = z)$ in the auxiliary variables $\boldsymbol{U}$. Thus, even in such setting it is desirable to develop alternative, more efficient, estimation procedures. The Augmented Inverse Probability Weighted (AIPW) kernel estimators developed in this paper address this issue.

When the outcomes are missing at random, Robins et al. (1995) and Rotnitzky and Robins (1995) proposed an inverse probability weighted (IPW) estimating equation for parametric regression, i.e. when $\theta(\cdot)$ is parametrically modeled as $\theta(\cdot;\boldsymbol{\nu})$ indexed by a finite dimensional parameter vector $\boldsymbol{\nu}$, where $\boldsymbol{\nu} \in \boldsymbol{R}^k$. Robins and Rotnitzky (1995) showed that one can improve the efficiency of the IPW estimator by adding to the IPW estimating function a parametric augmentation term. We extend their idea and propose a class of AIPW kernel estimating equations for estimating the non-parametric function $\theta(\cdot)$. We weight the units with complete data by either the inverse of the true selection probability $\pi_{i0} = Pr(R_i = 1|Z_i, \boldsymbol{U}_i)$ (if known, for instance as in two-stage sampling designs) or the inverse of an estimator of it, and add an adequately chosen augmentation term. We show that, just as for estimation of a parametric model for $\theta(\cdot)$, inclusion of the augmentation term can lead to efficiency improvement for estimation of the nonparametric regression function $\theta(\cdot)$. Unlike parametric regression, the augmentation term depends on a kernel function.

Specifically, let $K_h(s) = h^{-1}K(s/h)$, where $K(\cdot)$ is a mean-zero density function. Without loss of generality, we here focus on local linear kernel estimators. For any scalar x, define $\boldsymbol{G}(x) = (1, x)^T$ and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^T$. For any target point z, the local linear kernel estimator approximates $\theta(Z_i)$ in the neighborhood of z by a linear function $\boldsymbol{G}(Z_i - z)^T\boldsymbol{\alpha}$. Let $\mu(\cdot) = g^{-1}(\cdot)$. Suppose we postulate a working variance model $var(Y_i|Z_i) = V[\mu\{\theta(Z_i)\};\boldsymbol{\zeta}]$, where $\boldsymbol{\zeta} \in \boldsymbol{R}^r$ is an unknown finite dimensional parameter and $V(\cdot,\cdot)$ is a known working variance function. To estimate π_{i0} we postulate a parametric model

$$\pi_{i0} = \pi(Z_i, \boldsymbol{U}_i; \boldsymbol{\tau}), \tag{3}$$

where $\pi(Z, \boldsymbol{U}; \boldsymbol{\tau})$ is a known smooth function of an unknown finite dimensional parameter vector

$\boldsymbol{\tau} \in \boldsymbol{R}^k$. For example, we can assume a logistic model $\text{logit}(\pi_{i0}) = \tau_1 + \tau_2 Z_i + \boldsymbol{\tau}_3^T \boldsymbol{U}_i$, where $\boldsymbol{\tau} = (\tau_1, \tau_2, \boldsymbol{\tau}_3^T)^T$. We compute $\widehat{\boldsymbol{\tau}}$, the maximum likelihood estimator of $\boldsymbol{\tau}$ under model (3) and then we estimate π_{i0} with $\widehat{\pi}_i = \pi(Z_i, \boldsymbol{U}_i; \widehat{\boldsymbol{\tau}})$. Then we define the augmented inverse probability weighted (AIPW) kernel estimating equations as

$$\sum_{i=1}^{n} \{U_{IPW,i}(\alpha) - A_i(\alpha)\} = 0, \tag{4}$$

where

$$\begin{aligned} U_{IPW,i}(\alpha) &= \frac{R_i}{\widehat{\pi}_i} K_h(Z_i - z) \mu_i^{(1)} V_i^{-1} \boldsymbol{G}(Z_i - z) \left[Y_i - \mu\{\boldsymbol{G}(Z_i - z)^T \boldsymbol{\alpha}\} \right] \\ A_i(\alpha) &= \left(\frac{R_i}{\widehat{\pi}_i} - 1 \right) K_h(Z_i - z) \mu_i^{(1)} V_i^{-1} \boldsymbol{G}(Z_i - z) \left[\delta(Z_i, \boldsymbol{U}_i) - \mu\{\boldsymbol{G}(Z_i - z)^T \boldsymbol{\alpha}\} \right] \end{aligned} \tag{5}$$

with $\mu_i^{(1)}$ is the first derivative of $\mu(\cdot)$ evaluated at $\boldsymbol{G}(Z_i - z)^T \boldsymbol{\alpha}$, $\delta(Z_i, \boldsymbol{U}_i)$ is any arbitrary, user-specified, possibly data-dependent, function of Z_i and $\boldsymbol{U}_i$, and $V_i = V[\mu\{\boldsymbol{G}(Z_i - z)^T \boldsymbol{\alpha}\}; \boldsymbol{\zeta}]$. As $\boldsymbol{\zeta}$ is unknown in practice, we estimate it using the inverse probability weighted moment equations $\sum_{j=1}^{n} R_j \widehat{\pi}_j^{-1} V_j^{(1)} \left[\{Y_j - \hat{\alpha}_{0,j}(\boldsymbol{\zeta})\}^2 - V\{\hat{\alpha}_{0,j}(\boldsymbol{\zeta}), \boldsymbol{\zeta}\} \right] = 0$, where $V_j^{(1)} = \partial V\{\hat{\alpha}_{0,j}(\boldsymbol{\zeta}); \boldsymbol{\zeta}\} / \partial \boldsymbol{\zeta}$, and $\hat{\boldsymbol{\alpha}}_j(\boldsymbol{\zeta}) = \{\hat{\alpha}_{0,j}(\boldsymbol{\zeta}), \hat{\alpha}_{1,j}(\boldsymbol{\zeta})\}^T$ solve (4) with $z = Z_j, j = 1, ..., n,$. Denote the resulting estimator by $\widehat{\boldsymbol{\zeta}}$. The AIPW estimator of $\theta(z)$ is $\widehat{\theta}_{AIPW}(z) = \widehat{\alpha}_{0,AIPW}(\widehat{\boldsymbol{\zeta}})$ where $\widehat{\boldsymbol{\alpha}}_{AIPW} = \left\{ \widehat{\alpha}_{0,AIPW}(\widehat{\boldsymbol{\zeta}}), \widehat{\alpha}_{1,AIPW}(\widehat{\boldsymbol{\zeta}}) \right\}$ solves (4) with V_i replaced by $V[\mu\{\boldsymbol{G}(Z_i - z)^T \boldsymbol{\alpha}\}; \widehat{\boldsymbol{\zeta}}]$.

In the AIPW kernel estimating equations (4), the term $U_{IPW,i}(\alpha)$ is zero for subjects with missing outcomes and for those with observed outcomes it is simply equal to their usual contribution to the local kernel regression estimating equations weighted by the inverse of their probability of observing the outcome given their auxiliaries and covariates. The term $A_i(\alpha)$, which is often referred to as an augmentation term, differs from that used in parametric regression (eq.38 and eq.39, Robins et al. 1994) in that it additionally includes the kernel function $K_h(\cdot)$, and in that it approximates $\mu\{\theta(Z_i)\} = g^{-1}\{\theta(Z_i)\}$ by the local polynomial $\mu\{\boldsymbol{G}(Z_i - z)^T \boldsymbol{\alpha}\}$.

Two key properties, formally proved in Section 4, make the AIPW kernel estimating equation methodology appealing, namely: (1) exploitation of the information in the auxiliary variables of subjects with missing outcomes and (2) double robustness.

Informally, property (1) is seen because both the subjects with complete data and those with missing outcomes in a local neighborhood of $Z = z$ have a non-negligible contribution to the AIPW kernel estimating equations. Consider the alternative IPW kernel estimator $\hat{\theta}_{IPW}(z)$, which is obtained by simply solving the IPW kernel estimating equations $\sum_i U_{IPW,i}(\alpha) = 0$, i.e. ignoring the augmentation term in the estimating equations (4). Although $\hat{\theta}_{IPW}(z)$ depends on the

auxiliary variables $\boldsymbol{U}$ of the units with missing outcomes through the estimators $\widehat{\boldsymbol{\tau}}$ that define the $\widehat{\pi}_i$'s, this information is asymptotically negligible. Specifically, in Theorem 2, we show that when the support of Z is compact, under regularity conditions, the asymptotic distribution of $\widehat{\theta}_{IPW}(z)$ as $h \to 0, n \to \infty$ and $nh \to \infty$ is the same regardless of whether one uses the true π_{i0} (and hence do not use auxiliary data of incomplete units) or the fitted value $\widehat{\pi}_i$ computed under a correctly specified parametric model (3). This is different from inference under a parametric regression model for $E(Y|Z)$ where, as noted by Robins et al. (1994, 1995), estimation of the missingness probabilities helps improve the efficiency in estimation of regression coefficients. The reason is that the convergence of the ML estimator of π_{i0} under a parametric model is at the $\sqrt{n}$-rate while non-parametric estimation of $\theta(z)$ is at a slower rate. To see this note that only the $O(nh)$ units that have values of Z in a neighborhood of z of width $O(h)$ contribute to the IPW kernel estimating equations for $E(Y|Z=z)$, so only the auxiliary variables of these units are relevant. However, as $n \to \infty$, the data of these units could not enter into the IPW kernel estimating equations via the estimation of π_{i0} through the estimation of the finite dimensional parameter $\boldsymbol{\tau}$. This is so because for computing $\widehat{\boldsymbol{\tau}}$ parametrically all n units are used and the contribution of the $O(nh)$ relevant units is asymptotically negligible. The above discussions suggest that compared to the IPW kernel estimator, the AIPW kernel estimator of $\theta(z)$ can better explore the information in the auxiliary variables of subjects with missing outcomes.

To construct AIPW estimators with property (2), the double-robustness, we specify a parametric model

$$E(Y_i|Z_i, \boldsymbol{U}_i) = \delta(Z_i, \boldsymbol{U}_i; \boldsymbol{\eta}), \tag{6}$$

where $\boldsymbol{\eta}$ is an unknown finite dimensional parameter vector, and we estimate $\boldsymbol{\eta}$ using the method of moments estimator $\widehat{\boldsymbol{\eta}}$ based on data from completely observed units. Under the MAR assumption (2), $\widehat{\boldsymbol{\eta}}$ is $\sqrt{n}-$ consistent for $\boldsymbol{\eta}$, provided model (6) is correctly specified (Little and Rubin 2002). We then compute $\widehat{\theta}_{AIPW}(z)$ using $\delta(Z_i, \boldsymbol{U}_i) = \delta(Z_i, \boldsymbol{U}_i; \widehat{\boldsymbol{\eta}})$. In Theorem 3 in Section 4, we show that such estimator $\widehat{\theta}_{AIPW}(z)$ is doubly robust, that is, it is consistent when either model (3) for π_{i0} is correct or model (6) for $E(Y_i|Z_i, \boldsymbol{U}_i)$ is correct, but not necessarily both. The practical consequence of double-robustness is that it gives data analysts two opportunities of carrying out valid inference about $\theta(z)$, one for each of the possibly correctly specified models (6) or (3). In contrast, as shown in Theorem 1 in Section 4, consistency of the IPW kernel estimator $\widehat{\theta}_{IPW}(z)$ requires that the selection probability model (3) for π_{i0} must be correctly specified. One may question the possibility that the fully parametric model (6) for $E(Y_i|Z_i, \boldsymbol{U}_i)$ is correct when in fact

the model of scientific interest for $E(Y_i|Z_i)$ is left fully non-parametric precisely because of the lack of knowledge about the dependence of the mean of Y on Z. This valid concern is dissipated when it is understood that model (6) is only a working model that simply serves to enhance the chances of getting nearly correct (and indeed, nearly efficient) inference. Aside from this, it should also be noted that it is possible that data analysts may have refined knowledge of the conditional dependence of Y on Z within level of $\boldsymbol{U}$, but not marginally over $\boldsymbol{U}$.

In addition, in Section 4 we show that the preceding double-robust estimator $\widehat{\theta}_{AIPW}(z)$ has an additional desirable property. Specifically, if model (6) is correctly specified then the double-robust estimator $\widehat{\theta}_{AIPW}(z)$ has the smallest asymptotic variance among all estimators solving AIPW kernel estimating equations with π_{i0} either known or estimated from a correctly specified parametric model (3). That is, the asymptotic variance of the resulting double-robust estimator $\widehat{\theta}_{AIPW}(z)$ that uses $\delta(Z_i, \boldsymbol{U}_i) = \delta(Z_i, \boldsymbol{U}_i; \widehat{\boldsymbol{\eta}})$ with $\widehat{\boldsymbol{\eta}}$ a $\sqrt{n}-$consistent estimator of $\boldsymbol{\eta}$ under a correct model (6), is less than or equal to that of an AIPW kernel estimator using any other arbitrary function $\delta(Z_i, \boldsymbol{U}_i)$ when the selection probability model (3) is correct.

Remark: Our estimators $\widehat{\theta}_{AIPW}(z)$ use the IPW method of moments estimator of the variance parameter $\boldsymbol{\zeta}$. Although one could construct an AIPW method of moments estimator of $\boldsymbol{\zeta}$, this is unnecessary because improving the efficiency in estimation of the parameters $\boldsymbol{\zeta}$ does not help improve the efficiency in estimation of the nonparametric function $\theta(z)$. This is in accordance to estimation of parametric regression models for $E(Y|Z)$, where it is well known that the efficiency of two-stage weighted least squares is unaffected by the choice of $\sqrt{n}-$ consistent estimator of $var(Y|Z)$ at the first stage. In fact, Theorem 3 in Section 4 asserts that the efficiency with which $\theta(z)$ is estimated is unaltered even if the working model for $var(Y|Z)$ is incorrectly specified. This is in contrast to parametric regression models where incorrect modeling of $var(Y|Z)$ results in inefficient estimators of the regression parameters. The reason is that nonparametric regression is local and variability in a diminishing neighbor of z is constant asymptotically.

4 ASYMPTOTIC PROPERTIES

4.1 Asymptotic properties of the proposed estimators

In this section, we investigate the asymptotic properties of the AIPW local linear kernel estimator introduced in the preceding section and compare it with the naive and IPW nonparametric estimators. In our developments we make the following assumptions: I) $n \to \infty$, $h \to 0$, and $nh \to \infty$; II) z is in the interior of the support of Z; and III) The regularity conditions (i) and (ii) stated at

the beginning of the web Appendix hold.

Denote by $\tilde{\theta}_{naive}(z)$, $\tilde{\theta}_{IPW}(z)$, $\tilde{\theta}_{AIPW}(z)$ the asymptotic limits of $\hat{\theta}_{naive}(z)$, $\hat{\theta}_{IPW}(z)$, $\hat{\theta}_{AIPW}(z)$. The AIPW kernel estimator $\hat{\theta}_{AIPW}(z)$ solves (4). The IPW kernel estimator $\hat{\theta}_{IPW}(z)$ solves $\sum_{i=1}^{n} U_{IPW,i}(\boldsymbol{\alpha}) = 0$, where $U_{IPW,i}(\boldsymbol{\alpha})$ is defined in (5). The naive estimator $\hat{\theta}_{naive}(z)$ is the standard kernel estimator using only the complete data and solves a kernel estimating equation similar to the IPW kernel estimating equation $\sum_{i=1}^{n} U_{IPW,i}(\boldsymbol{\alpha}) = 0$ except that $\hat{\pi}_i$ is set to be 1 for all units. Standard arguments on the convergence of solutions to kernel estimating equations imply that under assumptions I)-III) there should exist a sequence of solutions $(\widehat{\alpha}_{0,naive}, \widehat{\alpha}_{1,naive})$ of the naive kernel estimating equations at z such that as the sample size $n \to \infty$, the sequence converges in probability to a vector $(\widetilde{\alpha}_{0,naive}, \widetilde{\alpha}_{1,naive})$ with the first component $\widetilde{\alpha}_{0,naive}$, throughout denoted as $\widetilde{\theta}_{naive}(z)$, satisfying

$$E\left[R\mu^{(1)}\{\widetilde{\theta}_{naive}(z)\}V^{-1}\{\widetilde{\theta}_{naive}(z);\widetilde{\boldsymbol{\zeta}}\}\left[Y-\mu\{\widetilde{\theta}_{naive}(z)\}\right]|Z=z\right]=0 \tag{7}$$

where $\widetilde{\boldsymbol{\zeta}}$ is the probability limit of $\widehat{\boldsymbol{\zeta}}$.

Likewise, the IPW kernel estimating equations should have a sequence of solutions $(\widehat{\alpha}_{0,IPW}, \widehat{\alpha}_{1,IPW})$ that converge in probability to a vector $(\widetilde{\alpha}_{0,IPW}, \widetilde{\alpha}_{1,IPW})$ with the first component $\widetilde{\alpha}_{0,IPW}$, throughout denoted as $\widetilde{\theta}_{IPW}(z)$, satisfying

$$E\left[\frac{R}{\widetilde{\pi}}\mu^{(1)}\{\widetilde{\theta}_{IPW}(z)\}V^{-1}\{\widetilde{\theta}_{IPW}(z);\widetilde{\boldsymbol{\zeta}}\}\left[Y-\mu\{\widetilde{\theta}_{IPW}(z)\}\right]|Z=z\right]=0, \tag{8}$$

where $\widetilde{\pi}=\pi(Z,\boldsymbol{U};\widetilde{\boldsymbol{\tau}})$, and $\widetilde{\boldsymbol{\tau}}$ is the probability limit of $\widehat{\boldsymbol{\tau}}$**.**

Similarly, the AIPW kernel estimating equations (4) should have a sequence of solutions $(\widehat{\alpha}_{0,AIPW}, \widehat{\alpha}_{1,AIPW})$ that converge in probability to a vector $(\widetilde{\alpha}_{0,AIPW}, \widetilde{\alpha}_{1,AIPW})$ with the first component $\widetilde{\alpha}_{0,AIPW}$, throughout denoted as $\widetilde{\theta}_{AIPW}(z)$, satisfying

$$\begin{aligned}&E\left[\frac{R}{\widetilde{\pi}}\mu^{(1)}\{\widetilde{\theta}_{AIPW}(z)\}V^{-1}\{\widetilde{\theta}_{AIPW}(z);\widetilde{\boldsymbol{\zeta}}\}\left[Y-\mu\{\widetilde{\theta}_{AIPW}(z)\}\right]|Z=z\right]+E\left\{\left(\frac{R}{\widetilde{\pi}}-1\right)\right.\\&\left.\times\mu^{(1)}\{\widetilde{\theta}_{AIPW}(z)\}V^{-1}\{\widetilde{\theta}_{AIPW}(z);\widetilde{\boldsymbol{\zeta}}\}\left[\widetilde{\delta}(Z,\boldsymbol{U})-\mu\{\widetilde{\theta}_{AIPW}(z)\}\right]|Z=z\right\}=0\end{aligned} \tag{9}$$

where $\widetilde{\delta}(Z,\boldsymbol{U})=\delta(Z,\boldsymbol{U};\widetilde{\boldsymbol{\eta}})$, and $\widetilde{\boldsymbol{\eta}}$ is the probability limit of $\widehat{\boldsymbol{\eta}}$.

Throughout we assume that such sequences exist. Theorem 1 exploits the form of (7), (8), and (9) to derive concise expressions for the probability limits of $\widehat{\theta}_{naive}(z)$, $\widehat{\theta}_{IPW}(z)$, and $\widehat{\theta}_{AIPW}(z)$ under MAR.

THEOREM 1 *Under the MAR assumption (2), the following results hold:*

(I) The probability limit $\widetilde{\theta}_{naive}(z)$ *of the naive kernel estimator defined in (7) satisfies* $\widetilde{\theta}_{naive}(z) = \mu^{-1}\left[\mu\{\theta(z)\} + cov\left(R, Y|Z=z\right)/E\left(R|Z=z\right)\right]$ *;*

(II) The probability limit $\widetilde{\theta}_{IPW}(z)$ *of the IPW kernel estimator defined in (8) satisfies* $\widetilde{\theta}_{IPW}(z) = \theta(z)$ *when* $\widehat{\pi}_i$ *is either computed under a correctly specified model (3) or is replaced by the true* π_{i0} *in the IPW kernel estimating function (5);*

(III) The probability limit $\widetilde{\theta}_{AIPW}(z)$ *of the AIPW kernel estimator defined in (9) satisfies* $\widetilde{\theta}_{AIPW}(z) = \theta(z)$ *when the AIPW kernel estimating equations (4) use either i) the true* π_{i0} *or* $\widehat{\pi}_i$ *computed under a correctly specified model (3); or ii)* $\delta(Z, \boldsymbol{U}) = E(Y|Z, \boldsymbol{U})$*, or* $\delta(Z, \boldsymbol{U}) = \delta(Z, \boldsymbol{U}; \widehat{\boldsymbol{\eta}})$ *with* $\widehat{\boldsymbol{\eta}}$ *calculated under a correctly specified model (6).*

The proof of Theorem 1 is given in web Appendix A.1. It follows from Theorem 1 that $\hat{\theta}_{naive}(z)$ is generally inconsistent for $\theta(z)$ except when R and Y are conditionally uncorrelated given Z. In particular, this implies that when missingness depends on variables $\boldsymbol{U}$ other than Z which further predict Y, $\hat{\theta}_{naive}(z)$ is inconsistent. However, if either of the following two conditions hold, then $cov\left(R, Y|Z=z\right) = 0$ and therefore $\hat{\theta}_{naive}(z)$ is consistent for $\theta(z)$. Specifically,

Condition a: The missing indicator R *depends on the covariate* Z *but given* Z *it is conditionally independent of auxiliary variables* $\boldsymbol{U}$*.*

Condition b: The conditional mean of Y *given* Z *and* $\boldsymbol{U}$ *depends only on* Z*.*

Theorem 1, part (III) shows that the AIPW kernel estimator $\widehat{\theta}_{AIPW}(z)$ has the remarkable double-robustness property alluded to in the preceding section: its consistency requires the correct specification of either a model for π_{i0} or a model for $E\left(Y|Z, \boldsymbol{U}\right)$, but not necessarily both.

In what follows, we study the asymptotic distributions of the proposed estimators. Theorem 2 and Theorem 3 provide the asymptotic bias and variance of $\widehat{\theta}_{IPW}(z)$ and $\widehat{\theta}_{AIPW}(z)$ respectively under MAR. Corollaries following these theorems show that in the class of AIPW kernel estimating equations that use either the true π_{i0} or a consistent estimate of π_{i0}, the optimal AIPW kernel estimating equation that yields a solution with the smallest asymptotic variance is obtained by setting $\delta(Z_i, \boldsymbol{U}_i) = E(Y_i|Z_i, \boldsymbol{U}_i)$ or $\delta(Z_i, \boldsymbol{U}_i) = E(Y_i|Z_i, \boldsymbol{U}_i; \widehat{\boldsymbol{\eta}})$ with $\widehat{\boldsymbol{\eta}}$ a $\sqrt{n}-$ consistent estimator of $\boldsymbol{\eta}$ computed under a correctly specified model (6). In addition, the solution of the optimal AIPW kernel estimating equations is at least as efficient as that of the IPW kernel estimating equations. A sketch of the proofs of Theorems 2 and 3 is given in web Appendix A.2 and web Appendix A.3 respectively. In what follows, $f_Z(\cdot)$ stands for the density function of Z, $b_K(z) \equiv \int K^2(s)ds/[\mu^{(1)}\{\theta(z)\}]^2 f_Z(z)$, $c_2(K) \equiv \int s^2 K(s)ds$, and $\pi_0(Z, \boldsymbol{U})$ denotes the true probability of $R = 1$ given $(Z, \boldsymbol{U})$.

THEOREM 2 *Suppose* $\widehat{\pi}_i$ *is computed under a correctly specified model (3) or is replaced by its true value. Suppose* $\Pr(R=1|Z,\boldsymbol{U}) > c > 0$ *for some constant* c *with probability 1 in a neighborhood of* $Z=z$*. Then, under the MAR assumption (2) and assumptions I)-III) above, we have that*

$$\sqrt{nh}\left\{\hat{\theta}_{IPW}(z) - \theta(z) - \frac{1}{2}h^2\theta''(z)c_2(K) + o(h^2)\right\} \rightarrow N\left\{0, W_{IPW}(z)\right\} \tag{10}$$

where

$$\begin{aligned} W_{IPW}(z) &\equiv b_K(z)\, E\left[\left[\frac{R}{\pi_0(Z,\boldsymbol{U})}\left(Y - \mu\{\theta(Z)\}\right)\right]^2 \middle| Z=z\right] \\ &= b_K(z)\, E\left[\frac{var(Y|Z,\boldsymbol{U}) + [E(Y|Z,\boldsymbol{U}) - \mu\{\theta(Z)\}]^2}{\pi_0(Z,\boldsymbol{U})} \middle| Z=z\right]. \end{aligned}$$

Theorem 2 shows that the asymptotic bias of $\widehat{\theta}_{IPW}(z)$ is of order $O(h^2)$, and the variance of $\widehat{\theta}_{IPW}(z)$ is of order $O(1/nh)$ and does not depend on the working variance $V(\cdot)$ in the IPW kernel estimating equations. This result indicates that, in contrast to parametric regression estimation, misspecification of the working variance $V(\cdot)$ of $Y|Z$ does not affect the asymptotic variance of $\widehat{\theta}_{IPW}(z)$. Theorem 2 also shows that to this order the bias and variance do not depend on whether the selection probabilities are known or estimated parametrically.

THEOREM 3 *Suppose that in the AIPW kernel estimating equations (4), (a)* $\widehat{\pi}_i$ *is computed under a model (3) or it is replaced by fixed probabilities* $\pi_i^* \equiv \pi^*(Z_i, \boldsymbol{U}_i)$ *and (b)* $\delta(Z,\boldsymbol{U})$ *is a fixed and known function or it is replaced by the function* $\delta(Z,\boldsymbol{U};\widehat{\boldsymbol{\eta}})$ *with* $\widehat{\boldsymbol{\eta}}$*, a method of moments estimator of* $\boldsymbol{\eta}$ *under model (6) based on units with observed outcomes. Suppose* $\Pr(R=1|Z,\boldsymbol{U}) > c > 0$ *for some constant* c *with probability* 1 *in a neighborhood of* $Z=z$*, and the MAR assumption (2) and assumptions I)-III) above hold. Consider additional conditions:*

i) model (3) is correct or, $\pi^*(Z,\ \boldsymbol{U}) = \pi_0(Z,\boldsymbol{U})$ *when* π_i^* *is used instead of* $\widehat{\pi}_i$ *in (4), or*

ii) model (6) is correct when $\delta(Z,\boldsymbol{U};\widehat{\boldsymbol{\eta}})$ *replaces* $\delta(Z,\boldsymbol{U})$ *in (4) or* $\delta(Z,\boldsymbol{U})$ *is equal to the true conditional expectation* $E(Y|Z,\boldsymbol{U})$ *otherwise.*

If either i) or ii) (but not necessarily both) hold, then

$$\sqrt{nh}\left\{\hat{\theta}_{AIPW}(z) - \theta(z) - \frac{1}{2}h^2\theta''(z)c_2(K) + o(h^2)\right\} \rightarrow N\left\{0, W_{AIPW}(z)\right\} \tag{11}$$

where

$$W_{AIPW}(z) = b_K(z)\, E\left[\left[\frac{R}{\widetilde{\pi}(Z,\boldsymbol{U})}\left(Y - \mu\{\theta(Z)\}\right) - \left(\frac{R}{\widetilde{\pi}(Z,\boldsymbol{U})} - 1\right)\left(\widetilde{\delta}(Z,\boldsymbol{U}) - \mu\{\theta(Z)\}\right)\right]^2 \middle| Z=z\right] \tag{12}$$

$\widetilde{\pi}(Z,\boldsymbol{U})$ denotes $\pi^*(Z,\boldsymbol{U})$ if π_i^* is used, or it denotes the probability limit of $\widehat{\pi}(Z,\boldsymbol{U})$ if $\widehat{\pi}_i$ is used, and $\widetilde{\delta}(Z,\boldsymbol{U})$ denotes $\delta(Z,\boldsymbol{U})$ if $\delta(Z,\boldsymbol{U})$ is used, or the probability limit of $\delta(Z,\boldsymbol{U};\widehat{\boldsymbol{\eta}})$ if $\delta(Z,\boldsymbol{U};\widehat{\boldsymbol{\eta}})$ is used.

Theorem 3 shows that the leading term of the asymptotic bias of $\hat{\theta}_{AIPW}(z)$ is the same as that of $\hat{\theta}_{IPW}(z)$ when the model for the selection probability is correctly specified. Furthermore, it remains the same even when the model for the selection probability is wrong, as long as the model for the conditional mean of the outcome given covariates and auxiliaries is correctly specified. Display (12) provides the general form of the asymptotic variance of $\hat{\theta}_{AIPW}(z)$ when either model (3) or model (6) is correctly specified. If model (6) is correctly specified, then (12) simplifies to $b_K(z)\,E\left[\pi_0(Z,\boldsymbol{U})/\widetilde{\pi}^2(Z,\boldsymbol{U})var(Y|Z,\boldsymbol{U}) + [E(Y|Z,\boldsymbol{U}) - \mu\{\theta(Z)\}]^2\,|Z=z\right]$.

On the other hand, if model (3) for the selection probability is correctly specified, the following corollary explores the properties of $W_{AIPW}(z)$ and it establishes that among the AIPW kernel estimating equations, the one that uses $\delta(Z_i,\boldsymbol{U}_i) = \delta(Z_i,\boldsymbol{U}_i;\widehat{\boldsymbol{\eta}})$ with $\widehat{\boldsymbol{\eta}}$ estimated under a correctly specified model (6) has a solution with the smallest asymptotic variance.

COROLLARY 1 *Under the assumptions of Theorem 3, if the selection probability model (3) is correctly specified, then*

$$\begin{aligned} W_{AIPW}(z) =& b_K(z)\,E\left[\frac{1}{\pi_0(Z,\boldsymbol{U})}var(Y|Z,\boldsymbol{U}) + [E(Y|Z,\boldsymbol{U}) - \mu\{\theta(Z)\}]^2\right. \\ &\left. + \left(\frac{1}{\pi_0(Z,\boldsymbol{U})} - 1\right)\left\{E(Y|Z,\boldsymbol{U}) - \widetilde{\delta}(Z,\boldsymbol{U})\right\}^2 \middle| Z=z\right]. \end{aligned} \tag{13}$$

$W_{AIPW}(z)$ is minimized at $\widetilde{\delta}(Z,\boldsymbol{U}) = E(Y|Z,\boldsymbol{U})$. Consequently, when model (3) is correct, the estimator $\widehat{\theta}_{AIPW}(z)$ that uses $\delta(Z,\boldsymbol{U}) = \delta(Z,\boldsymbol{U};\widehat{\boldsymbol{\eta}})$ from a correctly specified model for $E(Y|Z,\boldsymbol{U})$, throughout denoted as $\hat{\theta}_{opt,AIPW}(z)$, has the smallest asymptotic variance among all AIPW estimators $\widehat{\theta}_{AIPW}(z)$. The asymptotic variance of $\hat{\theta}_{opt,AIPW}(z)$ is equal to

$$W_{opt,AIPW}(z) = b_K(z)\,E\left\{\frac{var(Y|Z,\boldsymbol{U})}{\pi_0(Z,\boldsymbol{U})} + [E(Y|Z,\boldsymbol{U}) - \mu\{\theta(Z)\}]^2 \middle| Z=z\right\}.$$

Note that it follows from (13) that $W_{AIPW}(z)$ agrees with $W_{IPW}(z)$ when $\widetilde{\delta}(Z,\boldsymbol{U}) = \mu\{\theta(Z)\}$. This implies that, under correct specification of the selection probability model, the AIPW estimators that use $\delta(Z,\boldsymbol{U})$ equal to the fitted value $\delta(Z;\widehat{\boldsymbol{\omega}})$ from a parametric model $\delta(Z;\omega)$ for $E(Y|Z)$, rather than the fitted value from a parametric model for $E(Y|Z,\boldsymbol{U})$, are asymptotically equivalent to IPW estimators.

A direct comparison of the asymptotic variance of $\hat{\theta}_{opt,AIPW}(z)$ to that of $\hat{\theta}_{IPW}(z)$ in Theorem 2 immediately gives that the optimal AIPW kernel estimator is always at least as efficient as the IPW kernel estimator when indeed model (6) is correctly specified, as the next corollary establishes.

COROLLARY 2 *Suppose that $\hat{\theta}_{opt,AIPW}(z)$ and $\hat{\theta}_{IPW}(z)$ solve respectively the optimal AIPW and IPW kernel estimating equations that use the true π_{i0} or $\widehat{\pi}_i$ estimated under a correctly specified model (3). Then $\hat{\theta}_{opt,AIPW}(z)$ is at least as efficient as $\hat{\theta}_{IPW}(z)$ asymptotically, and the reduction in the asymptotic variance conferred by $\hat{\theta}_{opt,AIPW}(z)$ is*

$$W_{IPW}(z) - W_{opt,AIPW}(z) = b_K(z)\, E\left[\left(\frac{1}{\pi_0(Z,\boldsymbol{U})} - 1\right)\left[E(Y|Z,\boldsymbol{U}) - \mu\{\theta(Z)\}\right]^2 \middle| Z = z\right].$$

When $\Pr\left[\pi_0(Z,\boldsymbol{U}) < 1\right] > 0$, the difference $W_{IPW}(z) - W_{opt,AIPW}(z)$ is 0 only when $E(Y|Z = z,\boldsymbol{U}) - E(Y|Z = z) = 0$, i.e. when $\boldsymbol{U}$ does not predict Y in addition to Z. When $\boldsymbol{U}$ predicts Y above and beyond Z, as is expected for covariates $\boldsymbol{U}$ usually recorded in epidemiological studies, $W_{IPW}(z) - W_{opt,AIPW}(z)$ is strictly positive. Thus $\hat{\theta}_{opt,AIPW}(z)$ is usually more efficient than $\hat{\theta}_{IPW}(z)$.

4.2 An improved Estimator

A warning is appropriate at this stage. Our results show that using the optimal augmentation term we improve upon the efficiency of the IPW estimator. However, it is not guaranteed that any augmentation term in the AIPW kernel estimating equation leads to efficiency gains over the IPW method. In practice, one often does not know whether model (6) is correct, and hence is uncertain that $\hat{\theta}_{AIPW}(z)$ is more efficient than $\hat{\theta}_{IPW}(z)$. Nevertheless we can follow a strategy proposed by Tan (2006) for estimation of the marginal mean of an outcome and remedy this problem. Specifically, the following simple modification results in an AIPW kernel estimating function that yields double-robust estimators guaranteed to be at least as efficient as the IPW estimator $\hat{\theta}_{IPW}(z)$ and as the optimal AIPW estimator $\hat{\theta}_{opt,AIPW}(z)$ when model (3) holds for the selection probability. Let $M_{1i}(\alpha) = R_i\widehat{\pi}_i^{-1}V_i^{-1}K_h(Z_i - z)[Y_i - \mu\{\boldsymbol{G}(Z_i - z)^T\boldsymbol{\alpha}\}]$, $M_{2i}(\alpha) = (R_i\widehat{\pi}_i^{-1} - 1)V_i^{-1}K_h(Z_i - z)[\delta(Z_i,\boldsymbol{U}_i) - \mu\{\boldsymbol{G}(Z_i - z)^T\boldsymbol{\alpha}\}]$, $M_{3i}(\alpha) = R_i\widehat{\pi}_i^{-1}(\widehat{\pi}_i^{-1} - 1)V_i^{-2}K_h(Z_i - z)^2[\delta(Z_i,\boldsymbol{U}_i) - \mu\{\boldsymbol{G}(Z_i - z)^T\boldsymbol{\alpha}\}]^2$ and $\hat{\kappa}(\boldsymbol{\alpha}) = \{\sum_{i=1}^n M_{1i}(\alpha)M_{2i}(\alpha)\} / \{\sum_{i=1}^n M_{3i}(\alpha)\}$. Let $\widehat{\boldsymbol{\alpha}}_{\text{mod}} = \{\widehat{\alpha}_{\text{mod},0}, \widehat{\alpha}_{\text{mod},1}\}$ solve

$$\sum_{i=1}^n \left\{ \frac{R_i}{\widehat{\pi}_i} K_h(Z_i - z)\mu_i^{(1)}V_i^{-1}\boldsymbol{G}(Z_i - z)\left[Y_i - \mu\{\boldsymbol{G}(Z_i - z)^T\boldsymbol{\alpha}\}\right] \right. \tag{14}$$
$$\left. -\hat{\kappa}(\boldsymbol{\alpha})\left(\frac{R_i}{\widehat{\pi}_i} - 1\right)K_h(Z_i - z)\mu_i^{(1)}V_i^{-1}\boldsymbol{G}(Z_i - z)\left[\delta(Z_i,\boldsymbol{U}_i) - \mu\{\boldsymbol{G}(Z_i - z)^T\boldsymbol{\alpha}\}\right]\right\} = 0,$$

where V_i^{-1} is evaluated at $\widehat{\boldsymbol{\zeta}}$. The proposed modified estimator is $\widehat{\theta}_{\text{mod}}(z) = \widehat{\alpha}_{\text{mod},0}$. Note that (14) is just like the AIPW equation (4) except that the contribution to the augmentation term of each subject is multiplied by the factor $\hat{\kappa}(\boldsymbol{\alpha})$. Remarkably, this modification ensures that the new estimator $\widehat{\theta}_{\text{mod}}(z)$ is at least as efficient as the IPW estimator $\widehat{\theta}_{IPW}(z)$ and as the optimal AIPW estimator $\hat{\theta}_{opt,AIPW}(z)$ when model (3) holds and at the same time is double-robust. To see this, first note that multiplication by the factor $\hat{\kappa}(\boldsymbol{\alpha})$ in the augmentation term implies that the solution $\widehat{\theta}_{\text{mod}}(z)$ to the modified AIPW estimating equations converges in probability to the solution of a population equation just like (9) except that the second term in the left hand side of that equation is multiplied by

$$\kappa = \frac{E\left[\frac{R}{\widetilde{\pi}(Z,\boldsymbol{U})}\left(Y - \mu\{\theta(Z)\}\right)\left(\frac{R}{\widetilde{\pi}(Z,\boldsymbol{U})} - 1\right)\left(\widetilde{\delta}(Z,\boldsymbol{U}) - \mu\{\theta(Z)\}\right)|Z = z\right]}{E\left[\frac{R}{\widetilde{\pi}(Z,\boldsymbol{U})}\left(\frac{1}{\widetilde{\pi}(Z,\boldsymbol{U})} - 1\right)\left(\widetilde{\delta}(Z,\boldsymbol{U}) - \mu\{\theta(Z)\}\right)^2|Z = z\right]}$$

When model (3) is correct, then $\widetilde{\pi}(Z,\boldsymbol{U}) = \pi_0(Z,\boldsymbol{U})$ and the second term of the left hand side of (9) is zero, regardless of whether it is evaluated at the true $\theta(z)$ or not and regardless whether or not it is multiplied by the constant κ while the first term is unaffected by the modification and remains equal to zero when evaluated at $\theta(z)$. Thus $\widehat{\theta}_{\text{mod}}(z)$ is consistent for $\theta(z)$ when model (3) is correctly specified. On the other hand, when model (6) is correct, then $\widetilde{\delta}(Z,\boldsymbol{U}) = E(Y|Z,\boldsymbol{U})$ and a straightforward calculation shows that $\kappa = 1$ regardless of whether or not $\widetilde{\pi}(Z,\boldsymbol{U})$ is equal to $P(R = 1|Z,\boldsymbol{U})$, thus implying that $\widehat{\theta}_{\text{mod}}(z)$ is consistent for $\theta(z)$ since, as we argued earlier, $\theta(z)$ solves equation (9). This shows that $\widehat{\theta}_{\text{mod}}(z)$ is double-robust. To show that $\widehat{\theta}_{\text{mod}}(z)$ is at least as efficient as $\widehat{\theta}_{opt,AIPW}(z)$ and as $\widehat{\theta}_{IPW}(z)$ when model (3) is correctly specified, we can argue as in the proof of Theorem 3 and show that $\widehat{\theta}_{\text{mod}}(z)$ has the same limiting distribution as $\widehat{\theta}_{AIPW}(z)$, except that the asymptotic variance $W_{AIPW}(z)$ is replaced by

$$W_{\text{mod}}(z) = b_K(z)\, E\left[\left\{\frac{R}{\pi_0(Z,\boldsymbol{U})}\left[Y - \mu\{\theta(Z)\}\right] - \kappa \times \left(\frac{R}{\pi_0(Z,\boldsymbol{U})} - 1\right)\left[\widetilde{\delta}(Z,\boldsymbol{U}) - \mu\{\theta(Z)\}\right]\right\}^2 \middle| Z = z\right]$$

A straightforward calculation yields that the denominator of κ is equal to

$$E\left[\left\{\frac{R}{\pi_0(Z,\boldsymbol{U})} - 1\right\}^2\left[\widetilde{\delta}(Z,\boldsymbol{U}) - \mu\{\theta(Z)\}\right]^2|Z = z\right].$$

Thus, $W_{\text{mod}}(z)$ is equal to $b_K(z)$ times the residual variance from the population regression of $Y^* = R\left[Y - \mu\{\theta(Z)\}\right]/\pi_0(Z,\boldsymbol{U})$ on $X^* = \{R/\pi_0(Z,\boldsymbol{U}) - 1\}\left[\widetilde{\delta}(Z,\boldsymbol{U}) - \mu\{\theta(Z)\}\right]$. Since the residual variance $E\left[(Y^* - \kappa X^*)^2\right]$ minimizes the mean squared error $E\left[(Y^* - aX^*)^2\right]$ over all $a \in \mathbf{R}$, then we conclude that $W_{\text{mod}}(z) = b_K(z)\, E\left[(Y^* - \kappa X^*)^2\right]$ is less than or equal

to $W_{IPW}(z) = b_K(z)\, E\left[Y^{*2}\right]$ and to $W_{opt,AIPW}(z) = b_K(z)\, E\left[(Y^* - X^*)^2\right]$, where $\widetilde{\delta}(Z, \boldsymbol{U}) = E(Y|Z, \boldsymbol{U})$. Consequently, $\widehat{\theta}_{\text{mod}}(z)$ is at least as efficient as $\widehat{\theta}_{IPW}(z)$ and as $\widehat{\theta}_{opt,AIPW}(z)$ when $\widehat{\pi}_i$ is computed under a correctly specified model for the selection probabilities.

4.3 Bandwidth Selection

Choosing an appropriate bandwidth parameter h is important in nonparametric regression. From Theorems 2 and 3, the asymptotic optimal bandwidths $h_{IPW,opt}$ and $h_{AIPW,opt}$ can be chosen by minimizing the corresponding asymptotic weighted mean integrated squared errors, respectively. Specifically, the asymptotically optimal bandwidth for estimating $\widehat{\theta}_{IPW}(z)$ is given by $h_{IPW,opt} = \left[\left\{4\int W_{IPW}(z)dz\right\} / \left\{c_2(K)\int \theta''(z)dz\right\}\right]^{\frac{1}{5}} n^{-\frac{1}{5}}$ and the asymptotically optimal bandwidth for estimating $\widehat{\theta}_{AIPW}(z)$ is given by $h_{AIPW,opt} = \left[\left\{4\int W_{AIPW}(z)dz\right\} / \left\{c_2(K)\int \theta''(z)dz\right\}\right]^{\frac{1}{5}} n^{-\frac{1}{5}}$.

To choose h in practice, we can easily generalize the empirical bias bandwidth selection (EBBS) method of Ruppert (1997) to derive a data-driven bandwidth selection approach for nonparametric regression with missing data. Specifically, one calculates the empirical mean squared errors $EMSE\{z; h(z)\}$ of $\widehat{\theta}(z)$, where $EMSE\{z; h(z)\} = \widehat{bias}\left\{\widehat{\theta}(z)\right\}^2 + \widehat{var}\left\{\widehat{\theta}(z)\right\}$, at a series of z and $h(z)$ and chooses $h(z)$ to minimize $EMSE\{z; h(z)\}$. Note $h(z)$ is choosen to vary with z, and thus is local. Here $\widehat{bias}\left\{\widehat{\theta}(z)\right\}$ is the empirical bias, and $\widehat{var}\left\{\widehat{\theta}(z)\right\}$ is the Sandwich variance estimator. For example, the Sandwich variance estimator of the IPW kernel estimator $\hat{\theta}_{IPW}(z)$ can be calculated as the (1,1) element of the matrix $(\boldsymbol{A}_{IPW})^{-1}\boldsymbol{B}_{IPW}(\boldsymbol{A}_{IPW})^{-1}$, where

$$\boldsymbol{B}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{R_i}{\widehat{\pi}_i}K_h(Z_i - z)\mu_i^{(1)}V_i^{-1}\left[Y_i - \mu\{\boldsymbol{G}(Z_i - z)^T\boldsymbol{\alpha}\}\right]\right\}^2 \boldsymbol{G}(Z_i - z)\boldsymbol{G}(Z_i - z)^T$$

and

$$\boldsymbol{A}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\widehat{\pi}_i}K_h(Z_i - z)\left\{\mu_i^{(1)}\right\}^2 V_i^{-1}\boldsymbol{G}(Z_i - z)\boldsymbol{G}(Z_i - z)^T.$$

The Sandwich variance estimator of the naive kernel estimator $\widehat{\theta}_{naive}(z)$, and of the AIPW kernel estimator $\widehat{\theta}_{AIPW}(z)$ can be constructed in a similar way.

5 SIMULATIONS

In this section, we conduct simulation studies to evaluate the finite-sample performance of the AIPW kernel estimator $\hat{\theta}_{AIPW}(z)$, and compare it with the naive kernel estimator $\hat{\theta}_{naive}(z)$ and the IPW kernel estimator $\hat{\theta}_{IPW}(z)$. Our simulation mimics the observed data generating process of a two stage study design, in which $\boldsymbol{U}$ and Z are measured at the first stage on all study subjects,

but Y is measured at the second stage only on a subset of the study participants. The second-stage validation subset is selected with selection probabilities that may depend on the first stage variables. We consider two situations, where the outcome Y is either normal or binary respectively. We generate a random sample of size n of (Z, U, Y, R) for each replication. Z is generated from a $uniform(0, 1)$ distribution, U is generated from a $uniform(0, 6)$ independently of Z, and the mean of the outcome Y has the general form

$$g\left\{E\left(Y|Z,U\right)\right\} = m(Z) + \beta_1 U, \tag{15}$$

In case one, $g\left(x\right) = x$ and the outcome Y is generated from a normal distribution with mean $E\left(Y|Z,U\right)$ and variance $\sigma^2 = 3$, where $\beta_1 = 1.3$, $m(x) = 2 \cdot F_{8,8}(x)$ and $F_{p,q}(x) = \Gamma(p+q)\{\Gamma(p)\Gamma(q)\}^{-1}x^{p-1}(1-x)^{q-1}$, a unimodal function. In case two, $g\left(x\right) =$logit(x) where logit$(x) = \log\left\{x/\left(1+x\right)\right\}$ and the outcome Y is generated from a Bernoulli distribution with mean $E\left(Y|Z,U\right)$, where $\beta_1 = 0.32$, and $m(x) = 1.2 \cdot \Phi\left(8 \times x - 4\right) + 0.4$. In both situations, We generate R, the selection indicator, according to the probability model

$$\text{logit}\{\pi(Z_i, U_i)\} = \tau_0 + \tau_1 \cdot (U_i - a_1)I(a_1 < U_i \leq a_2) + \tau_1 \cdot (a_2 - a_1)I(U_i > a_2) \tag{16}$$

where $\pi(Z_i, U_i) = P\left(R_i = 1|Z_i, U_i\right)$ is the probability that subject i is selected to the second stage, $a_1 = 0.5$ and $a_2 = 6$. τ_0 and τ_1 are selected so that the Monte Carlo median missing percentage of the outcome Y is around 50% for the normal case and about 30% for the bernoulli case. Since the selection probability depends on U only, the missing is at random.

Our primary interest lies in estimating the marginal mean curve of the outcome Y given the scalar covariate Z, i.e., $\mu\left\{\theta(z)\right\}$, which is $E(Y|Z) = E\left[E(Y|Z,U)|Z\right]$. We generated 500 datasets with sample size $n = 500$ or 300. For each simulated dataset, we computed the naive, IPW and AIPW estimates of $\theta(z)$, in the first case under the model $\mu_i = \theta(Z_i)$ and in the second case under model $logit(\mu_i) = \theta(Z_i)$. We use the generalized EBBS method as described in section 4.3 to choose the optimal local bandwidth.

The empirical average of the estimated nonparametric curves $\hat{\theta}(\cdot)$ over the 500 replications, using the naive, IPW and AIPW estimators are displayed in Figure 1. The plot in the left panel shows the estimators of $\theta(z)$ in case 1 (identity link) and the plot in the right panel shows the estimators in case 2 (logit link). The same trend was observed for both plots. The IPW and AIPW kernel estimates are close to the true curve $\theta(\cdot)$, while the naive approach yields a biased estimate. Figure 2 illustrates the empirical point-wise variances of $\hat{\theta}_{IPW}(\cdot)$ and $\hat{\theta}_{AIPW}(\cdot)$ when

$n = 500$, the top panel for the identity link case and the bottom panel for the logit link case. The figure shows that the AIPW estimator has a smaller point-wise variance than the IPW estimator.

Table 1 summarizes the performance of each nonparametric estimate using the integrated relative bias, the integrated empirical standard error (S.E.), the integrated estimated S.E., and the integrated empirical mean integrated squared error (MISE), over the support of Z. As predicted by theory, the naive kernel estimate has a much larger relative bias than the IPW and AIPW kernel estimates. Furthermore, the corresponding AIPW kernel estimate has a smaller variance and a smaller MISE than the IPW kernel estimate. For example in the identity link case, the AIPW kernel estimate has about 52% gain in MISE efficiency compared to the IPW kernel estimate when $n = 500$. In the logit link case, the MISE efficiency gain is about 7%. The increased efficiency gain of AIPW over IPW in case 1 (identity link) compared to case 2 (logit link) can be explained by the fact that in case 1 the auxiliary variable U is highly correlated with the outcome Y while in case 2, the correlation between U and Y is much lower.

To check the double-robustness property of the AIPW estimator, we computed $\hat{\theta}_{AIPW}(\cdot)$ using i) estimates of π_{i0}'s under an incorrectly specified model with U_i replaced by $U_i^* = exp(U_i)$ in the right hand side of (16) but with δ_{i0}'s computed under a correctly specified model (15), ii) δ_{i0}'s computed under an incorrectly specified model with U_i replaced by U_i^* in the right hand side of (15) but with estimates of π_{i0}'s under the correctly specified model (16), and iii) both $\widehat{\pi}_i$ and δ_i computed under incorrectly specified models, with U_i replaced by U_i^* in the right hand side of (16) and (15) respectively. The simulation results in Table 2 and Figure 3 show that the AIPW kernel estimate is still close to the true $\theta(z)$ when either the model of $\pi(Z, U)$ or the model of $E(Y|Z, U)$ is correctly specified. In contrast, the IPW estimate with a misspecified model of $\pi(Z, U)$ is further away from the true $\theta(z)$, as well as the AIPW estimate when both the model of $\pi(Z, U)$ and the model of $E(Y|Z, U)$ are not correctly specified.

6 APPLICATION TO ACSUS DATA

We applied the IPW kernel estimating equation and the AIPW kernel estimating equation, as well as the naive kernel estimating equation, to analyze the ACSUS data described in Section 1. In this illustrative example, our main interest is to investigate the effect of the baseline CD4 counts on the risk of hospitalization during the first year since enrollment into the study. Since the risk of hospitalization depends on various covariates, such as HIV status, treatments, race, and gender, but we only consider a marginal nonparametric mean model of the risk of hospital admission on

baseline CD4 counts, we restricted our analysis to a subset of homogeneous subjects for illustrative purpose. Specifically, we limited our analysis to 219 white patients, who were between 25 and 45 years old at entry. They were HIV infected or had AIDS and were treated with antiretroviral drugs but not admitted to hospital at entry. The CD4 counts ranged from 4 to 1716 among this study cohort, with median equal to 186, and inter-quartile-range (70, 315). Health care records were used to determine hospitalization during the first year after study enrollment. Although lower CD4 counts are expected to be associated with a higher risk of hospitalization, the functional form of this association is unknown and might be nonlinear. As discussed in Section 1, about 40% of the patients did not have the first year hospital admission data available. If missing outcomes induced selection bias, the patients who have the first year hospitalization information may not represent the original study cohort and may lead to biased estimation.

Because the distribution of CD4 counts is highly skewed, we took a log transformation and define $Z = log($ baseline CD4 count). The missing data model was fit using a logistic regression with Z as well as the other covariates in Table 3, which are binary. The coefficient estimates and their SEs are shown in Table 3. Having insurance and help with transportation enhance the chance of remaining in the study, while use of other medical practitioners, psychological counseling, having help at home and lower CD4 count are significantly associated with a higher chance of dropping out.

We fit the generalized nonparametric model (1) using $logit(\mu_i) = \theta(Z_i)$ to investigate the dependence pattern of the first-year risk of hospitalization on baseline CD4 counts. The bandwidth was selected using the generalized EBBS method described in section 4.3. The estimates of the curve $\theta(z)$ using the naive kernel estimating equations, the IPW kernel estimating equations and the AIPW kernel estimating equations are shown in Figure 4. Point-wise Wald CIs centered at the naive, IPW and AIPW kernel estimates and with standard error estimated using the Sandwich formulae described in section 4.3, are also presented. For computing the AIPW estimate, we fit parametric models for δ. Exploration of the data shows that the regression function with a quadratic term in logcd4 and the other covariates in Table 3 fits the data well. Residual plot shows no patterns.

Since only very few patients had log CD4 count lower than 3, the kernel estimates are not stable when log CD4 count is less than 3. We focus our discuss on the estimates of the curve when log CD4 count is greater than 3. The IPW and AIPW estimates are similar, while the naive one underestimates the risk of hospitalization for most of the range of CD4 in our study cohort.

Since patients having help at home are more likely to drop out and these patients are likely to be sicker patients, the patients who have the first-year hospital admission information available are actually a biased sample of the whole study population. Therefore, the naive approach using the complete cases directly leads to a biased estimate of the nonparametric function $\theta(z)$ and underestimates the risk of hospitalization. Our analysis using the IPW and the AIPW kernel estimating equations indicates that the risk of hospitalization decreases nonlinearly as CD4 count increases with a change point. Specifically, when CD4 count is relatively low (CD4 count < 90), the risk of being admitted to hospitals remains fairly stable at about 25%. As the CD4 count exceeds this threshold, the risk of hospitalization decreases quickly as CD4 count goes up.

7 DISCUSSION

In this paper we proposed local polynomial kernel estimation methods for nonparametric regression when outcomes are missing at random. We showed that the naive local polynomial kernel estimator is generally inconsistent except for special cases. We proposed IPW and AIPW kernel estimating equations to correct for potential selection bias, with the ultimate goal of maximally exploiting the information in the observed data. Unlike parametric regression, the augmentation term in the AIPW kernel estimating equations incorporates a kernel function. We showed that both the IPW and AIPW kernel estimators are consistent when the selection probabilities are known by design or consistently estimated. When the model for the selection probabilities is misspecified, the IPW kernel estimating equation fails to yield a consistent estimator. However, the AIPW kernel estimator still yields consistent estimators of the regression function if a model for $E(Y|Z, \boldsymbol{U})$ is correctly specified. This double robustness property of the AIPW approach provides the investigators two chances to make a valid inference. The AIPW kernel estimating equation also has the potential to enhance the efficiency with which we estimate the nonparametric regression function. We have shown that within the AIPW estimating equation family, the optimal estimator is obtained by using the true selection probability or its consistent estimates and the augmentation term estimated from a correctly specified model for $E(Y|Z, \boldsymbol{U})$. It is of future research interest to study whether this estimator is optimal in a bigger class of estimators. Another interesting topic of future investigation is the possibility of enhancing the efficiency of the IPW estimator via estimation of the missingness probabilities at non-parametric rates, for example, under generalized additive models rather than under parametric models.

The IPW and AIPW kernel estimating equations provide consistent estimators when the se-

lection probability model π is correctly specified and is bounded away from 0. In finite samples, when some π's are close to 0, the IPW and AIPW estimators might not perform well. This is not surprising, as very large weights associated with these very small π's dramatically inflate a few observations especially when the sample size is moderate, and cause results unstable. Special caution is hence needed when applying the proposed methods to studies when the selection probability is very small for some sample units.

We have focused in this paper on nonparametric regression on a single scalar covariate when the outcome is missing at random. The proposed method can be extended to semiparametric regression, where some covariates are modeled parametrically and some covariates are modeled nonparametrically. The proposed methods can also be easily generalized to higher order local polynomial kernel regression and nonparametric regression with multiple covariates, e.g., using generalized additive models. Extension of our work to these settings will be reported in a separate paper.

8 Supplemental Materials

Technical Proofs: Regularity conditions and proofs for Theorems 1, 2, and 3 in Section 4.

References

Berk, M., Maffeo, C., and Schur, C. (1993), *Research Design and Analysis Objectives. AIDS Cost and Services Utilization Survey (ACSUS) Reports, No. 1.*, AHCPR Publication No. 93-0019. Rockville, MD: Agency for Health Care Policy and Research.

Breslow, N. and Cain, K. (1988), "Logistic regression for two-stage case-control data." *Biometrika*, 75, 11–20.

Carroll, R. J., Ruppert, D., and Welsh, A. H. (1998), "Local Estimating Equations," *Journal of the American Statistical Association*, 93, 214–227.

Chen, J., Fan, J., Li, K.-H., and Zhou, H. (2006), "Local quasi-likelihood estimation with data missing at random," *Statistica Sinica*, 16, 1044–1070.

Fan, J. and Gijbels, I. (1995), "Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation," *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 371–394.

— (1996), *Local Polynomial Modelling and Its Applications*, Chapman. & Hall, London.

Fan, J., Heckman, N. E., and Wand, M. P. (1995), "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions," *Journal of the American Statistical Association*, 90, 141–150.

Harezlak, J., Wang, M., Christiani, D., and Lin, X. (2007), "Quantitative quality-assessment techniques to compare fractionation and depletion methods in SELDI-TOF mass spectrometry experiments," *Bioinformatics*, 23, 2441.

Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall/CRC.

Liang, H., Wang, S., Robins, J. M., and Carroll, R. J. (2004), "Estimation in Partially Linear Models With Missing Covariates," *Journal of the American Statistical Association*, 99, 357.

Little, R. (1982), "Models for nonresponse in sample surveys," *Journal of the American Statistical Association*, 77, 237–50.

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

McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models*, Chapman and Hall, London.

Pepe, M. S. (1992), "Inference Using Surrogate Outcome Data and a Validation Sample," *Biometrika*, 79, 355–365.

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

Robins, J. M. (1999), "Robust estimation in sequentially ignorable missing data and causal inference models," *Proceedings of the American Statistical Association Sectionon Bayesian Statistical Science, ASA, Alexandria, VA*, n/a, 6–10.

Robins, J. M. and Rotnitzky, A. (1995), "Semiparametric Efficiency in Multivariate Regresion Models with Missing Data," *Journal of the American Statistical Association*, 90, 122–129.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866.

— (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90, 106–121.

Rotnitzky, A., Holcroft, C., and Robins, J. M. (1997), "Efficiency comparisons in multivariate multiple regression with missing outcomes," *Journal of Multivariate Analysis*, 61, 102–128.

Rotnitzky, A. and Robins, J. (1995), "Semiparametric regression estimation in the presence of dependent censoring," *Biometrika*, 82(4), 805–820.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581.

Ruppert, D. (1997), "Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation," *Journal of the American Statistical Association*, 92, 1049.

Tan, Z. (2006), "A Distributional Approach for Causal Inference Using Propensity Scores," *Journal of the American Statistical Association*, 101, 1619–1637.

Wand, M. and Jones, M. (1995), *Kernel Smoothing*, Chapman & Hall, London.

Wang, C. Y., Wang, S., Gutierrez, R. G., and Carroll, R. J. (1998), "Local Linear Regresion for Generalized Linear Models with Missing Data," *The Annals of Statistics*, 26, 1028.

Table 1: Simulation results of relative biases, S.E.s and MISEs of the naive, IPW and AIPW estimates of $\theta(z)$ based on 500 replications. (In parenthesis are the Monte Carlo S.E.s)

	$n = 500$				$n = 300$			
	Relative bias[1]	EMP S.E.[2]	EST S.E.[3]	EMP MISE[4]	Relative bias	EMP S.E.	EST S.E.	EMP MISE
Normal Case (Identity Link)								
no missing	0.017	0.336	0.326	0.130	0.017	0.434	0.431	0.207
	(0.002)	(0.011)	(0.001)	(0.008)	(0.004)	(0.005)	(0.003)	(0.011)
naive	0.234	0.437	0.431	1.713	0.233	0.578	0.573	1.843
	(0.003)	(0.014)	(0.003)	(0.032)	(0.006)	(0.020)	(0.002)	(0.079)
IPW	0.034	0.645	0.642	0.451	0.044	0.843	0.839	0.770
	(0.004)	(0.013)	(0.002)	(0.020)	(0.012)	(0.017)	(0.006)	(0.043)
AIPW	0.018	0.443	0.438	0.215	0.019	0.579	0.567	0.356
	(0.003)	(0.012)	(0.004)	(0.012)	(0.005)	(0.012)	(0.005)	(0.013)
Logistic Case (Logit Link)								
no missing	0.048	0.220	0.213	0.049	0.074	0.254	0.249	0.067
	(0.024)	(0.005)	(0.001)	(0.002)	(0.016)	(0.013)	(0.001)	(0.006)
naive	0.662	0.229	0.223	0.075	0.667	0.267	0.260	0.099
	(0.048)	(0.007)	(0.001)	(0.004)	(0.051)	(0.011)	(0.001)	(0.008)
IPW	0.058	0.239	0.234	0.058	0.095	0.283	0.278	0.084
	(0.021)	(0.007)	(0.001)	(0.003)	(0.022)	(0.009)	(0.001)	(0.006)
AIPW	0.054	0.234	0.231	0.054	0.099	0.276	0.270	0.080
	(0.016)	(0.096)	(0.001)	(0.003)	(0.025)	(0.007)	(0.001)	(0.005)

1. Relative bias is defined as $\int |\widehat{bias}\{\hat{\theta}(z)\}/\theta(z)|dF(z)$.
2. EMP S.E. is the empirical S.E., defined as $\int \widehat{SE}_{EMP}\{\hat{\theta}(z)\}dF(z)$, where $\widehat{SE}_{EMP}\{\hat{\theta}(z)\}$ is the sampling S.E. of the replicated $\hat{\theta}(z)$.
3. EST S.E. is the estimated S.E., defined as $\int \widehat{SE}_{EST}\{\hat{\theta}(z)\}dF(z)$, where $\widehat{SE}_{EST}\{\hat{\theta}(z)\}$ is the sampling average of the replicated sandwich estimates $\widehat{SE}\{\hat{\theta}(z)\}$.
4. EMP MISE is the empirical MISE, defined as $\int\{\hat{\theta}(z) - \theta(z)\}^2 dF(z)$

Zhang, D., Lin, X., and Sowers, M. (2000), "Semiparametric regression for periodic longitudinal hormone data from multiple menstrual cycles," *Biometrics*, 56, 31–39.

Table 2: Simulation results of the relative biases, S.E.s and MISEs of the IPW and AIPW estimates of $\theta(\cdot)$ using $\widehat{\pi}$ inconsistent for π_0 and/or $\delta = \widehat{E}(Y|Z,\boldsymbol{U})$ inconsistent for $E(Y|Z,\boldsymbol{U})$, based on 500 replications. (In parenthesis are the Monte Carlo S.E.s)

	n=500				n=300			
	Relative bias[1]	EMP S.E.[2]	EST S.E.[3]	EMP MISE[4]	Relative bias	EMP S.E.	EST S.E.	EMP MISE
Normal Case:								
AIPW (π wrong)	0.017	0.481	0.475	0.251	0.018	0.635	0.629	0.423
	(0.004)	(0.016)	(0.002)	(0.017)	(0.005)	(0.020)	(0.008)	(0.028)
AIPW ($E[Y\|Z,U]$ wrong)	0.023	0.640	0.636	0.442	0.021	0.832	0.825	0.728
	(0.004)	(0.018)	(0.007)	(0.021)	(0.010)	(0.034)	(0.019)	(0.042)
AIPW (both wrong)	0.068	0.641	0.638	0.835	0.066	0.841	0.837	1.125
	(0.003)	(0.012)	(0.004)	(0.019)	(0.004)	(0.022)	(0.011)	(0.065)
IPW (π wrong)	0.105	0.471	0.462	0.522	0.108	0.632	0.629	0.723
	(0.004)	(0.011)	(0.003)	(0.027)	(0.006)	(0.021)	(0.003)	(0.044)
Logistic Case:								
AIPW (π wrong)	0.052	0.254	0.251	0.066	0.102	0.295	0.289	0.092
	(0.021)	(0.007)	(0.001)	(0.003)	(0.026)	(0.012)	(0.001)	(0.008)
AIPW ($E[Y\|Z,U]$ wrong)	0.056	0.236	0.233	0.057	0.095	0.276	0.271	0.080
	(0.022)	(0.006)	(0.001)	(0.003)	(0.027)	(0.007)	(0.001)	(0.005)
AIPW (both wrong)	0.975	0.249	0.250	0.111	0.978	0.286	0.281	0.136
	(0.058)	(0.008)	(0.001)	(0.005)	(0.064)	(0.010)	(0.001)	(0.008)
IPW (π wrong)	0.662	0.229	0.223	0.075	0.667	0.267	0.263	0.099
	(0.047)	(0.007)	(0.001)	(0.004)	(0.051)	(0.011)	(0.001)	(0.008)

1. Relative bias is defined as $\int |\widehat{bias}\{\hat{\theta}(z)\}/\theta(z)|dF(z)$.
2. EMP S.E. is the empirical S.E., defined as $\int \widehat{SE}_{EMP}\{\hat{\theta}(z)\}dF(z)$, where $\widehat{SE}_{EMP}\{\hat{\theta}(z)\}$ is the sampling S.E. of the replicated $\hat{\theta}(z)$.
3. EST S.E. is the estimated S.E., defined as $\int \widehat{SE}_{EST}\{\hat{\theta}(z)\}dF(z)$, where $\widehat{SE}_{EST}\{\hat{\theta}(z)\}$ is the sampling average of the replicated sandwich estimates $\widehat{SE}\{\hat{\theta}(z)\}$.
4. EMP MISE is the empirical MISE, defined as $\int\{\hat{\theta}(z)-\theta(z)\}^2dF(z)$

Table 3: Estimates of the logistic regression coefficients of the probability of being observed by the end of the first year in the ACSUS data

Covariates	Estimate	S.E.	P-Value
Intercept	-2.62	0.85	0.002
Has help at home	-0.65	0.36	0.063
Has private health insurance only	0.53	0.45	0.241
Has both private and public health insurance	2.13	0.83	0.010
Has public health insurance only	-0.11	0.47	0.819
Use other medical practitioners	-0.95	0.49	0.053
Use psychological counseling	-0.80	0.35	0.022
Log CD4 count	0.64	0.14	<0.001
Has help with transportation	2.39	0.94	0.011

Figure 1: Simulation results of the estimated nonparametric functions using naive, IPW and AIPW kernel methods based on 500 replications with sample size $n = 500$. The left panel is for case 1 (identity link), while the right panel is for case 2 (logit link): —— true $\theta(z)$, – · – · the naive kernel estimator, · · · · the IPW kernel estimator, and – – – the AIPW kernel estimator.

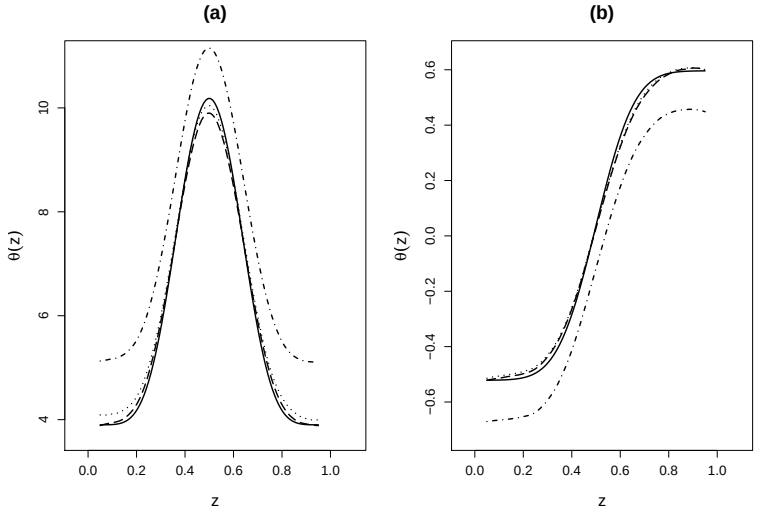


Figure 2: Empirical point-wise variances of the IPW and AIPW estimates of $\theta(\cdot)$, based on 500 replications with sample size $n = 500$. The top panel is for case 1 (identity link), while the bottom panel is for case 2 (logit link): —— the IPW kernel estimate, – – – the AIPW kernel estimate, and · · · · · the kernel estimate when there is no missing data.

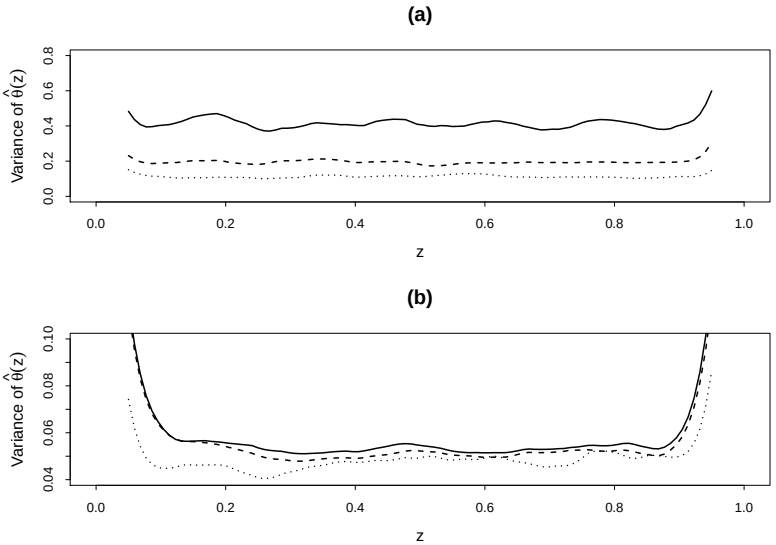


Figure 3: Simulation results of the IPW and AIPW estimates of $\theta(\cdot)$ using an incorrectly specified π model and/or an incorrectly specified $\delta = E(Y|Z, \boldsymbol{U})$ model, based on 500 replications with sample size $n = 500$. The left panel is for case 1 (identity link) and the right panel is for case 2 (logit link): —— the true $\theta(z)$, – – – the AIPW kernel estimator when the model for $\pi(Z, U)$ is misspecified, – · – · the AIPW kernel estimator when the model for $E[Y|Z, U]$ is misspecified, - - - the AIPW kernel estimator when both models are misspecified, and · · · · the IPW kernel estimator when the model for $\pi(Z, U)$ is misspecified.

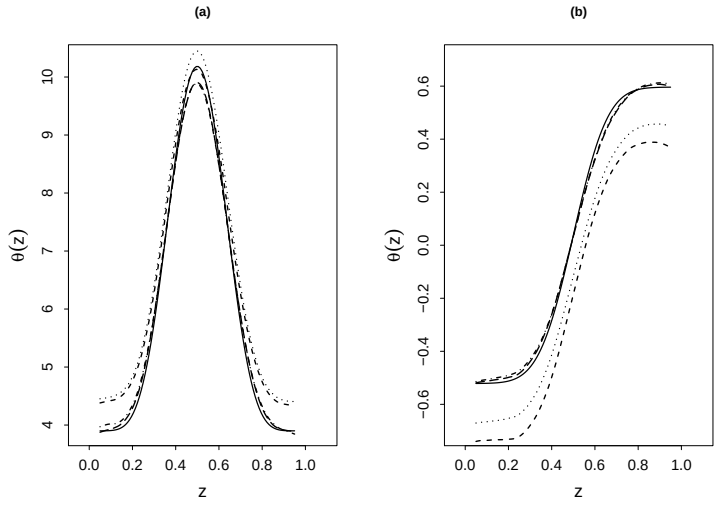
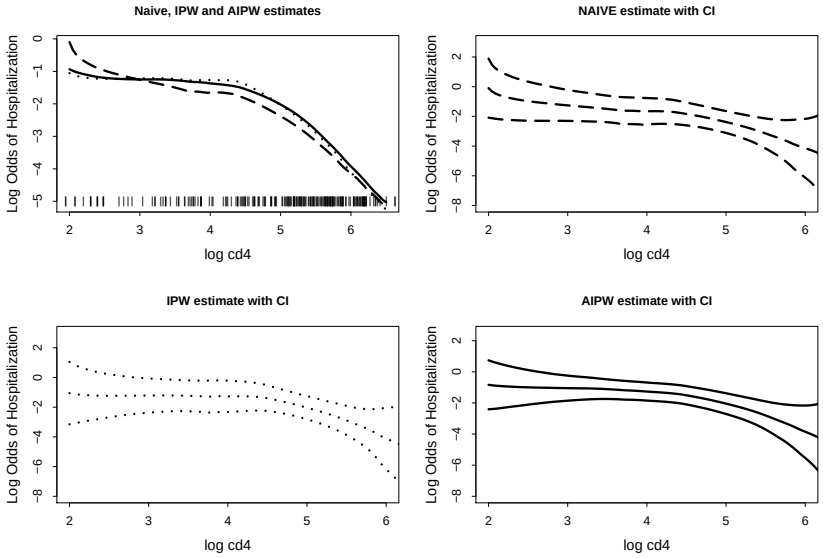


Figure 4: The naive, IPW and AIPW estimates of θ(log CD4 counts) on the log odds of one-year hospitalization in the ACSUS study. The upper left panel displays three estimates: – – – the naive kernel estimate, · · · · the IPW kernel estimate, —— the AIPW kernel estimate. Each vertical ticker along the x-axis stands for one observation. The other three panels display each estimate separately together with point-wise CIs.



Appendix

Throughout the appendix, we assume that $h = h(n)$ is a sequence such that as $n \to \infty$, $h \to 0$, and $nh \to \infty$. We also assume that z is an interior point of the support of Z. We assume the following regularity conditions:

i) $\theta(\cdot)$ and $f_Z(\cdot)$ satisfy the smoothness assumptions of Fan, et. al. (1995);

ii) The estimating functions in the right hand side of naive kernel estimating equations, IPW kernel estimating equations, and AIPW kernel estimating equations are twice continuously differentiable with respect to $\boldsymbol{\alpha}$ at a target point z, and the second derivatives are uniformly bounded.

A.1 Sketch of the Proof of Theorem 1

If $\mu^{(1)}\{\widetilde{\theta}_{naive}(z)\} \neq 0$, simple calculations show that the solution of equation (7) for $\widetilde{\theta}_{naive}(z)$ is $\mu\{\widetilde{\theta}_{naive}(z)\} = E(RY|Z=z)/E(R|Z=z)$, which is equal to $cov(R,Y|Z=z)/E(R|Z=z) + \mu\{\theta(z)\}$. This gives the expression for $\widetilde{\theta}_{naive}(z)$ stated in the theorem.

Next study the expression of $\widetilde{\theta}_{IPW}(z)$. The left hand side of (8) is equal to

$$E\left[\frac{E(R|Y,Z,\boldsymbol{U})}{\widetilde{\pi}}\mu^{(1)}\{\widetilde{\theta}_{IPW}(z)\}V^{-1}\{\widetilde{\theta}_{IPW}(z);\widetilde{\boldsymbol{\zeta}}\}\left[Y-\mu\{\widetilde{\theta}_{IPW}(z)\}\right]\middle| Z=z\right]$$

by taking a double expectation given Y, Z and $\boldsymbol{U}$**.** If model (3) of π is correctly specified, then $\widetilde{\pi} = E(R|Z,\boldsymbol{U})$. Also under MAR, $E(R|Y,Z,\boldsymbol{U}) = E(R|Z,\boldsymbol{U})$. Therefore the above quantity equals to $E[\mu^{(1)}\{\widetilde{\theta}_{IPW}(z)\} \times V^{-1}\{\widetilde{\theta}_{IPW}(z);\widetilde{\boldsymbol{\zeta}}\}[Y-\mu\{\widetilde{\theta}_{IPW}(z)\}]|Z=z]$. If $\mu^{(1)}\{\widetilde{\theta}_{IPW}(z)\} \neq 0$, solving for $\widetilde{\theta}_{IPW}(z)$ yields $\mu\{\widetilde{\theta}_{IPW}(z)\} = E[Y|Z=z] = \mu\{\theta(z)\}$. Therefore, $\widehat{\theta}_{IPW}(z)$ is a consistent estimator of $\theta(z)$ when model (3) of π is correctly specified or π_0 is known by design.

Now study the expression of $\widetilde{\theta}_{AIPW}(z)$ from (9). Under the MAR assumption (2), the left hand side of (9) can be rewritten as

$$\begin{aligned} & E\left[\mu^{(1)}\{\widetilde{\theta}_{AIPW}(z)\}V^{-1}\{\widetilde{\theta}_{AIPW}(z);\widetilde{\boldsymbol{\zeta}}\}\left[Y-\mu\{\widetilde{\theta}_{AIPW}(z)\}\right]|Z=z\right] \\ + \quad & E\left[\left(\frac{R}{\widetilde{\pi}}-1\right)\mu^{(1)}\{\widetilde{\theta}_{AIPW}(z)\}V^{-1}\{\widetilde{\theta}_{AIPW}(z);\widetilde{\boldsymbol{\zeta}}\}\left[Y-\widetilde{\delta}(Z,\boldsymbol{U})\right]|Z=z\right] = 0. \quad \text{(A.1)} \end{aligned}$$

If model (3) for π is correctly specified, i.e., $\tilde{\pi} = E(R|Z,\boldsymbol{U})$, or model (6) for $\delta(\cdot)$ is correctly specified, i.e., $\widetilde{\delta}(Z,\boldsymbol{U}) = E(Y|Z,\boldsymbol{U})$, one can easily see that the second term of (A.1) is 0. Hence (A.1) is equal to

$$E\left[\mu^{(1)}\{\widetilde{\theta}_{AIPW}(z)\}V^{-1}\{\widetilde{\theta}_{AIPW}(z);\widetilde{\boldsymbol{\zeta}}\}\left[Y-\mu\{\widetilde{\theta}_{AIPW}(z)\}\right]|Z=z\right] = 0.$$

It follows that if $\mu^{(1)}\{\widetilde{\theta}_{AIPW}(z)\} \neq 0$, we have $\widetilde{\theta}_{AIPW}(z) = \theta(z)$, i.e., $\widehat{\theta}_{AIPW}(z)$ is a consistent estimator of $\theta(z)$.

A.2 Proof of Theorem 2: Asymptotic Bias and Variance of the IPW Estimator

We first assume that π_0 is known by design and prove that the asymptotic distribution of $\hat{\theta}_{IPW}(z)$ is given in (10). We also assume that the variance parameter $\boldsymbol{\zeta}$ in the working variance V is known. We will then extend the results when π and $\boldsymbol{\zeta}$ are estimated. For any interior point z, reparameterize $\boldsymbol{\alpha}$ as $\left\{\theta\left(z\right), h\theta'\left(z\right)\right\}^T$ and denote by $\theta_0(z)$ the true value of $\theta(z)$, $\boldsymbol{\alpha}_0 = \{\theta_0(z), h\theta_0'(z)\}^T$ and $\hat{\boldsymbol{\alpha}}_{IPW}(z)$ the solution of the local linear IPW kernel estimating equations. A Taylor expansion of the local linear IPW kernel estimating equations gives

$$\sqrt{nh}\{\boldsymbol{\alpha}_{IPW}(z) - \boldsymbol{\alpha}_0\} = -\sqrt{nh}\left\{\boldsymbol{\Gamma}_n(\boldsymbol{\alpha}_*)\right\}^{-1}\boldsymbol{\Lambda}_n(\boldsymbol{\alpha}_0),$$

where $\boldsymbol{\alpha}_*$ is between $\hat{\boldsymbol{\alpha}}_{IPW}(z)$ and $\boldsymbol{\alpha}_0$, and

$$\boldsymbol{\Lambda}_n(\boldsymbol{\alpha}) = n^{-1}\sum_{i=1}^{n} R_i\pi_{i0}^{-1}(Z_i, \boldsymbol{U}_i)K_h(Z_i - z)\mu_i^{(1)}(z, \boldsymbol{\alpha})V_i^{-1}(z, \boldsymbol{\alpha})\boldsymbol{G}(Z_i - z)[Y_i - \mu\{\boldsymbol{G}(Z_i - z)^T\boldsymbol{\alpha}\}],$$

where $\mu_i^{(1)}(z, \boldsymbol{\alpha}) = \mu^{(1)}\{\boldsymbol{G}(Z_i - z)^T\boldsymbol{\alpha}\}$ and $V_i(z, \boldsymbol{\alpha}) = V[\mu\{\boldsymbol{G}(Z_i - z)^T\boldsymbol{\alpha}\}; \boldsymbol{\zeta}_0]$, $\boldsymbol{\Gamma}_n(\boldsymbol{\alpha}) = \partial\boldsymbol{\Lambda}_n(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}^T$.

Using the results in Appendix A.1, we have $\hat{\boldsymbol{\alpha}}_{IPW}(z) \to \boldsymbol{\alpha}_0$ in probability. Therefore, $\boldsymbol{\alpha}_* \xrightarrow{P} \boldsymbol{\alpha}_0$. Under the MAR assumption (2), simple calculations show that

$$\begin{aligned}
\boldsymbol{\Gamma}_n(\boldsymbol{\alpha}_*) &= -E\left[K_h(Z - z)\left\{\mu^{(1)}(z, \boldsymbol{\alpha}_0)\right\}^2 V^{-1}(z, \boldsymbol{\alpha}_0)\boldsymbol{G}(Z - z)\boldsymbol{G}(Z - z)^T\right] + o_p(1) \\
&= -f_Z(z)\left(\mu^{(1)}\{\theta(z)\}\right)^2 V^{-1}\{\theta(z)\}\boldsymbol{D}(K) + o_p(1)
\end{aligned}$$

where $\boldsymbol{D}\left(K\right)$ is a 2×2 matrix with the (j,k)th element $c_{j+k-2}(K) \times h^{(j+k-2)}$, and $c_r(K) = \int s^r K(s)ds$. It follows that

$$\sqrt{nh}\{\hat{\boldsymbol{\alpha}}_{IPW}(z) - \boldsymbol{\alpha}_0\} = \left\{f_Z(z)\left[\mu^{(1)}\{\theta(z)\}\right]^2 V^{-1}\{\theta(z)\}\boldsymbol{D}(K)\right\}^{-1}\sqrt{nh}\boldsymbol{\Lambda}_n(\boldsymbol{\alpha}_0) + o_p(1). \quad \text{(A.2)}$$

Now write $\boldsymbol{\Lambda}_n(\boldsymbol{\alpha}_0) = \boldsymbol{\Lambda}_{1n}(\boldsymbol{\alpha}_0) + \boldsymbol{\Lambda}_{2n}(\boldsymbol{\alpha}_0)$, where

$$\begin{aligned}
\boldsymbol{\Lambda}_{1n}(\boldsymbol{\alpha}_0) &= n^{-1}\sum_{i=1}^{n} R_i\pi_{i0}^{-1}(Z_i, \boldsymbol{U}_i)K_h(Z_i - z)\mu_i^{(1)}(z, \boldsymbol{\alpha}_0)V_i^{-1}(z, \boldsymbol{\alpha}_0)\boldsymbol{G}(Z_i - z)[Y_i - \mu\{\theta(Z_i)\}] \\
\boldsymbol{\Lambda}_{2n}(\boldsymbol{\alpha}_0) &= n^{-1}\sum_{i=1}^{n} R_i\pi_{i0}^{-1}(Z_i, \boldsymbol{U}_i)K_h(Z_i - z)\mu_i^{(1)}(z, \boldsymbol{\alpha}_0)V_i^{-1}(z, \boldsymbol{\alpha}_0)\boldsymbol{G}(Z_i - z)[\mu\{\theta(Z_i)\} - \mu\{\boldsymbol{G}(Z_i - z)^T\boldsymbol{\alpha}_0\}].
\end{aligned}$$

One can easily show that $\boldsymbol{\Lambda}_{1n}\left(\boldsymbol{\alpha}_0\right)$ is asymptotically normal with mean zero and asymptotic variance

$$\begin{aligned}
var\{\boldsymbol{\Lambda}_{1n}(\boldsymbol{\alpha}_0)\} &= \frac{1}{n}E\left[K_h^2(Z - z)\left\{\mu^{(1)}(z, \boldsymbol{\alpha}_0)\right\}^2 V^{-2}(z, \boldsymbol{\alpha}_0)\left(\frac{R\left[Y - \mu\left\{\theta(Z)\right\}\right]}{\pi_0(Z, \boldsymbol{U})}\right)^2 \boldsymbol{G}(Z - z)\boldsymbol{G}(Z - z)^T\right] \\
&= \frac{1}{nh}f_Z(z)\left(\mu^{(1)}\{\theta(z)\}\right)^2 V^{-2}\{\theta(z)\}E\left[\left(\frac{R\left[Y - \mu\left\{\theta(Z)\right\}\right]}{\pi_0(Z, \boldsymbol{U})}\right)^2 \middle| Z = z\right]\boldsymbol{D}(K^2) + \boldsymbol{o}(\frac{1}{nh}),
\end{aligned}$$

where $\boldsymbol{D}(K^2)$ is defined similarly to $\boldsymbol{D}(K)$ with K replaced by K^2.

Now study $\boldsymbol{\Lambda}_{2n}$, which contributes to the leading bias term. One can easily show under MAR, we have

$$\begin{aligned} bias\left\{\boldsymbol{\Lambda}_{2n}(\boldsymbol{\alpha}_0)\right\} &= E\left\{K_h(Z-z)\mu^{(1)}(z,\boldsymbol{\alpha}_0)V^{-1}(z,\boldsymbol{\alpha}_0)\left[\mu\left\{\theta(Z)\right\}-\mu\left\{\boldsymbol{G}(Z-z)^T\boldsymbol{\alpha}_0\right\}\right]\boldsymbol{G}(Z-z)\right\}+o_p(1) \\ &= \frac{1}{2}\theta''(z)\left[\mu^{(1)}\{\theta(z)\}\right]^2 V^{-1}\{\theta(z)\}f_Z(z)\boldsymbol{H}(K)+o(h^2), \end{aligned} \tag{A.3}$$

where $\boldsymbol{H}(K)$ is a 2×1 vector with the kth element $c_{k+1}(K)\times h^{(k+1)}$. Note that the asymptotic variance of $\boldsymbol{\Lambda}_{2n}$ is of order $\boldsymbol{o}(1/nh)$ and is asymptotically negligible compared to $\boldsymbol{\Lambda}_{1n}$, and the asymptotic covariance of $\boldsymbol{\Lambda}_{1n}$ and $\boldsymbol{\Lambda}_{2n}$ is $\mathbf{0}$. Applying these results to (A.2), simple calculations show that the asymptotic distribution of the IPW estimator $\hat{\theta}_{IPW}(z;\pi)$, the first element of $\hat{\boldsymbol{\alpha}}_{IPW}$, is given in (10).

We next study the distribution of $\hat{\theta}_{IPW}\{z;\pi(\hat{\boldsymbol{\tau}})\}$ when π_0 is estimated consistently at the $\sqrt{n}$-rate, i.e. $\sqrt{n}(\hat{\boldsymbol{\tau}}-\boldsymbol{\tau}_0)=O_p(1)$, where $\boldsymbol{\tau}_0$ is the true value of $\boldsymbol{\tau}$**.** Suppose under some regularity conditions, $\partial\hat{\theta}_{IPW}\{z;\pi(\boldsymbol{\tau})\}/\partial\boldsymbol{\tau}^T$ is bounded in the neighborhood of the $\boldsymbol{\tau}_0$, *i.e.*,

$$\partial\hat{\theta}_{IPW}\{z;\pi(\boldsymbol{\tau})\}/\partial\boldsymbol{\tau}^T|_{\boldsymbol{\tau}\in\mathcal{N}(\boldsymbol{\tau}_0)}=O_p(1),$$

where $\mathcal{N}(\boldsymbol{\tau}_0)\supset\{\boldsymbol{\tau}:||\boldsymbol{\tau}-\boldsymbol{\tau}_0||<||\hat{\boldsymbol{\tau}}-\boldsymbol{\tau}_0||\}$. We have

$$\begin{aligned} &\sqrt{nh}[\hat{\theta}_{IPW}\{z;\pi(\hat{\boldsymbol{\tau}})\}-\theta(z)] \\ &=\sqrt{nh}[\hat{\theta}_{IPW}\{z;\pi(\hat{\boldsymbol{\tau}})\}-\hat{\theta}_{IPW}\{z;\pi(\tau_0)\}]+\sqrt{nh}[\hat{\theta}_{IPW}\{z;\pi(\tau_0)\}-\theta(z)] \\ &=\sqrt{h}\left[\frac{\partial\hat{\theta}_{IPW}\{z;\pi(\boldsymbol{\tau})\}}{\partial\boldsymbol{\tau}^T}|_{\boldsymbol{\tau}^*}\right]\sqrt{n}(\hat{\boldsymbol{\tau}}-\boldsymbol{\tau}_0)+\sqrt{nh}[\hat{\theta}_{IPW}\{z;\pi(\tau_0)\}-\theta(z)] \end{aligned} \tag{A.4}$$

for some $\boldsymbol{\tau}^*\in\{\boldsymbol{\tau}:||\boldsymbol{\tau}-\boldsymbol{\tau}_0||<||\hat{\boldsymbol{\tau}}-\boldsymbol{\tau}_0||\}$. Note $\sqrt{n}(\hat{\boldsymbol{\tau}}-\boldsymbol{\tau}_0)=O_p(1)$, $\partial\hat{\theta}_{IPW}\left\{z;\pi(\boldsymbol{\tau})\right\}/\partial\boldsymbol{\tau}^T|_{\boldsymbol{\tau}^*}=O_p(1)$, and $h\to 0$ as $n\to\infty$, the first term in (A.4) is $o_p(1)$. Therefore, the asymptotic distribution of $\hat{\theta}_{IPW}\{z;\pi(\hat{\boldsymbol{\tau}})\}$ when $\boldsymbol{\tau}$ is estimated consistently at $\sqrt{n}$-rate is the same as that of $\hat{\theta}_{IPW}(z;\pi_0)$ when π_0 is known. Similar argument shows that the asymptotic distribution of $\hat{\theta}_{IPW}\{z\}$ remains the same if $\boldsymbol{\zeta}$ is estimated at the $\sqrt{n}$-rate.

A.3 Proof of Theorem 3: Asymptotic Bias and Asymptotic Variance of AIPW estimator

Following similar arguments as those in Appendix A.2, the asymptotic results hold when the parameters $(\boldsymbol{\tau},\boldsymbol{\eta})$ in π and δ are estimated at the $\sqrt{n}$-rate, or the probability limit of $(\widehat{\boldsymbol{\tau}},\widehat{\boldsymbol{\eta}})$ is used in the AIPW kernel estimating equations (4). Denote by $(\widetilde{\boldsymbol{\tau}},\widetilde{\boldsymbol{\eta}})$ the probability limit of $(\widehat{\boldsymbol{\tau}},\widehat{\boldsymbol{\eta}})$, and let $\widetilde{\pi}(Z_i,\boldsymbol{U}_i)=\pi(Z_i,\boldsymbol{U}_i;\widetilde{\boldsymbol{\tau}})$, $\widetilde{\delta}\left(Z_i,\boldsymbol{U}_i\right)=\delta\left(Z_i,\boldsymbol{U}_i;\widetilde{\boldsymbol{\eta}}\right)$. We focus our proof on assuming that $(\widetilde{\boldsymbol{\tau}},\widetilde{\boldsymbol{\eta}})$ are known. By a linear Taylor expansion of the AIPW estimating function (4) about $\boldsymbol{\alpha}_0$, the AIPW kernel estimator satisfies

$$\sqrt{nh}\{\hat{\boldsymbol{\alpha}}_{AIPW}(z)-\boldsymbol{\alpha}_0\}=-\sqrt{nh}\left\{\boldsymbol{\Gamma}_{n,\delta}(\boldsymbol{\alpha}_*)\right\}^{-1}\boldsymbol{\Lambda}_{n,\delta}(\boldsymbol{\alpha}_0),$$

where $\boldsymbol{\alpha}_*$ is between $\hat{\boldsymbol{\alpha}}_{AIPW}(z)$ and $\boldsymbol{\alpha}_0$,

$$\begin{aligned}\boldsymbol{\Lambda}_{n,\delta}(\boldsymbol{\alpha}) &= n^{-1}\sum_{i=1}^{n}\Big\{R_i\widetilde{\pi}^{-1}(Z_i,\boldsymbol{U}_i)K_h(Z_i-z)\mu_i^{(1)}(z,\boldsymbol{\alpha})V_i^{-1}(z,\boldsymbol{\alpha})\boldsymbol{G}(Z_i-z)\left[Y_i-\mu\left\{\boldsymbol{G}(Z_i-z)^T\boldsymbol{\alpha}\right\}\right]\\ &\quad-\left\{R_i\widetilde{\pi}^{-1}(Z_i,\boldsymbol{U}_i)-1\right\}K_h(Z_i-z)\mu_i^{(1)}(z,\boldsymbol{\alpha})V_i^{-1}(z,\boldsymbol{\alpha})\boldsymbol{G}(Z_i-z)\left[\widetilde{\delta}(Z_i,\boldsymbol{U}_i)-\mu\left\{\boldsymbol{G}(Z_i-z)^T\boldsymbol{\alpha}\right\}\right]\Big\},\end{aligned}$$

and $\boldsymbol{\Gamma}_{n,\delta}(\boldsymbol{\alpha})=\partial\boldsymbol{\Lambda}_{n,\delta}(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}^T$.

We consider the following two situations:

(1) When model (3) for the selection probability π_{i0} is correctly specified, i.e. $\widetilde{\pi}(Z_i,\boldsymbol{U}_i)=\pi_{i0}(Z_i,\boldsymbol{U}_i)$;

(2) When model (6) for $E(Y|Z,\boldsymbol{U})$ is correctly specified, i.e. $\widetilde{\delta}(Z_i,\boldsymbol{U}_i)=E(Y_i|Z_i,\boldsymbol{U}_i)$.

As shown in Appendix A.1, $\hat{\boldsymbol{\alpha}}_{AIPW}(z)$ converges to $\boldsymbol{\alpha}_0$ when either of the above conditions holds. Therefore, $\boldsymbol{\alpha}_*\xrightarrow{P}\boldsymbol{\alpha}_0$. We first show that under either of the above situations, we have

$$\boldsymbol{\Gamma}_{n,\delta}(\boldsymbol{\alpha}_*)\xrightarrow{P}-f_Z(z)\left[\mu^{(1)}\{\theta(z)\}\right]^2V^{-1}\{\theta(z)\}\boldsymbol{D}(K). \tag{A.5}$$

First consider situation (1), i.e., when $\widetilde{\pi}(Z_i,\boldsymbol{U}_i)=\pi_{i0}(Z_i,\boldsymbol{U}_i)$. The second term of $\boldsymbol{\Lambda}_{n,\delta}(\boldsymbol{\alpha})$, i.e. the augmentation term, has mean 0 under MAR . It follows that $\boldsymbol{\Lambda}_{n,\delta}(\boldsymbol{\alpha}_*)=\boldsymbol{\Lambda}_n(\boldsymbol{\alpha}_0)+\boldsymbol{o}_p(1)$, where $\boldsymbol{\Lambda}_n$ is defined in Appendix A.2. Hence $\boldsymbol{\Gamma}_{n,\delta}(\boldsymbol{\alpha}_*)=\boldsymbol{\Gamma}_n(\boldsymbol{\alpha}_0)+\boldsymbol{o}_p(1)$. Therefore $\boldsymbol{\Gamma}_{n,\delta}(\boldsymbol{\alpha}_*)$ has the same probability limit as $\boldsymbol{\Gamma}_n(\boldsymbol{\alpha}_*)$. As shown in Appendix A.2, the probability of limit of $\boldsymbol{\Gamma}_n(\boldsymbol{\alpha}_*)$ is exactly the right hand side of $(A.5)$, and thus $(A.5)$ holds for $\boldsymbol{\Gamma}_{n,\delta}(\boldsymbol{\alpha}_*)$ as well.

Next consider situation (2), i.e., when $\widetilde{\delta}(Z_i,\boldsymbol{U}_i)=E(Y_i|Z_i,\boldsymbol{U}_i)$. Rewrite $\boldsymbol{\Lambda}_{n,\delta}(\boldsymbol{\alpha})$ as

$$\begin{aligned}\boldsymbol{\Lambda}_{n,\delta}(\boldsymbol{\alpha}) &= n^{-1}\sum_{i=1}^{n}\Big\{R_i\widetilde{\pi}^{-1}(Z_i,\boldsymbol{U}_i)K_h(Z_i-z)\mu_i^{(1)}(z,\boldsymbol{\alpha})V_i^{-1}(z,\boldsymbol{\alpha})\boldsymbol{G}(Z_i-z)\left[Y_i-\widetilde{\delta}(Z_i,\boldsymbol{U}_i)\right]\\ &\quad+K_h(Z_i-z)\mu_i^{(1)}(z,\boldsymbol{\alpha})V_i^{-1}(z,\boldsymbol{\alpha})\boldsymbol{G}(Z_i-z)\left[\widetilde{\delta}(Z_i,\boldsymbol{U}_i)-\mu\left\{\boldsymbol{G}(Z_i-z)^T\boldsymbol{\alpha}\right\}\right]\Big\}.\end{aligned}$$

One can easily see the first term on the right hand side has mean 0. It follows that

$$\boldsymbol{\Lambda}_{n,\delta}(\boldsymbol{\alpha}_*)=n^{-1}\sum_{i=1}^{n}K_h(Z_i-z)\mu_i^{(1)}(z,\boldsymbol{\alpha}_0)V_i^{-1}(z,\boldsymbol{\alpha})\boldsymbol{G}(Z_i-z)\left[E\left(Y_i|Z_i,\boldsymbol{U}_i\right)-\mu\left\{\boldsymbol{G}(Z_i-z)^T\boldsymbol{\alpha}_0\right\}\right]+\boldsymbol{o}_p(1).$$

Differentiating it with respect to $\boldsymbol{\alpha}$ shows that $\boldsymbol{\Gamma}_{n,\delta}(\boldsymbol{\alpha}_*)=\boldsymbol{\Gamma}_n(\boldsymbol{\alpha}_0)+\boldsymbol{o}_p(1)$. Therefore, (A.5) still holds in this situation.

Therefore, when either the π or δ model is correctly specified, we have

$$\sqrt{nh}\{\hat{\boldsymbol{\alpha}}_{AIPW}(z)-\boldsymbol{\alpha}_0\}=\left\{f_Z(z)\left[\mu^{(1)}\{\theta(z)\}\right]^2V^{-1}\{\theta(z)\}\boldsymbol{D}(K)\right\}^{-1}\sqrt{nh}\boldsymbol{\Lambda}_{n,\delta}(\boldsymbol{\alpha}_0)+o_p(1). \tag{A.6}$$

Write $\boldsymbol{\Lambda}_{n,\delta}(\boldsymbol{\alpha}_0)=\boldsymbol{\Lambda}_{1n,\delta}(\boldsymbol{\alpha}_0)-\boldsymbol{\Lambda}_{2n,\delta}(\boldsymbol{\alpha}_0)+\boldsymbol{\Lambda}_{3n,\delta}(\boldsymbol{\alpha}_0)$, where

$$\boldsymbol{\Lambda}_{1n,\delta}(\boldsymbol{\alpha}_0)=n^{-1}\sum_{i=1}^{n}R_i\widetilde{\pi}^{-1}(Z_i,\boldsymbol{U}_i)K_h(Z_i-z)\mu_i^{(1)}(z,\boldsymbol{\alpha}_0)V_i^{-1}(z,\boldsymbol{\alpha}_0)\left[Y_i-\mu\{\theta(Z_i)\}\right]\boldsymbol{G}(Z_i-z),$$

$$\boldsymbol{\Lambda}_{2n,\delta}\left(\boldsymbol{\alpha}_0\right)=n^{-1}\sum_{i=1}^{n}\{R_i\widetilde{\pi}^{-1}(Z_i,\boldsymbol{U}_i)-1\}K_h(Z_i-z)\mu_i^{(1)}(z,\boldsymbol{\alpha}_0)V_i^{-1}(z,\boldsymbol{\alpha}_0)\left[\widetilde{\delta}(Z_i,\boldsymbol{U}_i)-\mu\{\theta(Z_i)\}\right]\boldsymbol{G}(Z_i-z),$$

and

$$\boldsymbol{\Lambda}_{3n,\delta}\left(\boldsymbol{\alpha}_0\right)=n^{-1}\sum_{i=1}^{n}K_h(Z_i-z)\mu_i^{(1)}(z,\boldsymbol{\alpha}_0)V_i^{-1}(z,\boldsymbol{\alpha}_0)\left[\mu\{\theta(Z_i)\}-\mu\{\boldsymbol{G}(Z_i-z)^T\boldsymbol{\alpha}_0\}\right]\boldsymbol{G}(Z_i-z).$$

One can easily see that $\boldsymbol{\Lambda}_{1n,\delta}(\boldsymbol{\alpha}_0)$ and $\boldsymbol{\Lambda}_{2n,\delta}(\boldsymbol{\alpha}_0)$ have mean 0 when either π or δ is correctly specified. The third term $\boldsymbol{\Lambda}_{3n,\delta}(\boldsymbol{\alpha}_0)$ is the leading bias term. When π_i or δ_i is correctly specified, simple calculations show that $E\left[\boldsymbol{\Lambda}_{3n,\delta}(\boldsymbol{\alpha}_0)\right]$ is equal to (A.3). It follows that

$$bias\{\hat{\boldsymbol{\alpha}}_{AIPW}(z)\}=\frac{1}{2}h^2\theta''(z)c_2(K)+o(h^2).$$

Now study $\boldsymbol{\Lambda}_{1n,\delta}-\boldsymbol{\Lambda}_{2n,\delta}$, which contributes to the leading variance and asymptotic normality. Note that the variance of $\boldsymbol{\Lambda}_{3n,\delta}\left(\boldsymbol{\alpha}_0\right)$ is of order $o(1/nh)$, and hence can be ignored asymptotically. Under MAR, we have $E[R|Y,Z,\boldsymbol{U}]=E[R|Z,\boldsymbol{U}]=\pi_0(Z,\boldsymbol{U})$, the true conditional mean of $[R|Z,\boldsymbol{U}]$. It follows that when either π or δ is correctly specified, $\boldsymbol{\Lambda}_{1n,\delta}(\boldsymbol{\alpha}_0)-\boldsymbol{\Lambda}_{2n,\delta}(\boldsymbol{\alpha}_0)$ is asymptotically normal with mean 0 and variance

$$var\left\{\boldsymbol{\Lambda}_{1n,\delta}(\boldsymbol{\alpha}_0)-\boldsymbol{\Lambda}_{2n,\delta}(\boldsymbol{\alpha}_0)\right\}=\frac{1}{n}\left[var\left\{\boldsymbol{\Lambda}_{1,2,\delta}(\boldsymbol{\alpha}_0)\right\}\right],$$

where

$$\begin{aligned}&\boldsymbol{\Lambda}_{1,2,\delta}(\boldsymbol{\alpha}_0)=K_h(Z-z)\mu^{(1)}(z,\boldsymbol{\alpha}_0)V^{-1}(z,\boldsymbol{\alpha}_0)\boldsymbol{G}(Z-z)\\&\times\left(\frac{R}{\widetilde{\pi}(Z,\boldsymbol{U})}\left[Y-\mu\{\theta(Z)\}\right]-\left\{\frac{R}{\widetilde{\pi}(Z,\boldsymbol{U})}-1\right\}\left[\widetilde{\delta}(Z,\boldsymbol{U})-\mu\{\theta(Z)\}\right]\right)\end{aligned}$$

Further calculations show that

$$\begin{aligned}&\frac{1}{n}var\left\{\boldsymbol{\Lambda}_{1,2,\delta}(\boldsymbol{\alpha}_0)\right\}=\frac{1}{n}E\left[K_h^2(Z-z)\left\{\mu^{(1)}(z,\boldsymbol{\alpha}_0)\right\}^2V^{-2}(z,\boldsymbol{\alpha}_0)\boldsymbol{G}(Z-z)\boldsymbol{G}(Z-z)^T\right.\\&\left.\times\left(\frac{R}{\widetilde{\pi}(Z,\boldsymbol{U})}\left[Y-\mu\{\theta(Z)\}\right]-\left\{\frac{R}{\widetilde{\pi}(Z,\boldsymbol{U})}-1\right\}\left[\widetilde{\delta}(Z,\boldsymbol{U})-\mu\{\theta(Z)\}\right]\right)^2\right]\end{aligned}$$

$$\begin{aligned}=\;&\frac{1}{nh}f_Z(z)\left[\mu^{(1)}\{\theta(z)\}\right]^2V^{-2}\{\theta(z)\}E\left[\left(\frac{R}{\widetilde{\pi}(Z,\boldsymbol{U})}\left[Y-\mu\{\theta(Z)\}\right]\right.\right.\\&\left.\left.-\left\{\frac{R}{\widetilde{\pi}(Z,\boldsymbol{U})}-1\right\}\left[\widetilde{\delta}(Z,\boldsymbol{U})-\mu\{\theta(Z)\}\right]\right)^2|Z=z\right]\boldsymbol{D}(K^2)+\boldsymbol{o}(\frac{1}{nh})\end{aligned}$$

Applying these results to (A.6) and Theorem 3 follows.