A Double Robust Approach for Non-Monotone Missingness in Multi-Stage Data

Shenshen Yang*

This Draft: August 6, 2024

Abstract

Multivariate missingness with a non-monotone missing pattern is complicated to deal with in empirical studies. The traditional Missing at Random (MAR) assumption is difficult to justify in such cases. Previous studies have strengthened the MAR assumption, suggesting that the missing mechanism of any variable is random when conditioned on a uniform set of fully observed variables. However, empirical evidence indicates that this assumption may be violated for variables collected at different stages. This paper proposes a new MAR-type assumption that fits non-monotone missing scenarios involving multi-stage variables. Based on this assumption, we construct an Augmented Inverse Probability Weighted GMM (AIPW-GMM) estimator. This estimator features an asymmetric format for the augmentation term, guarantees double robustness, and achieves the closed-form semiparametric efficiency bound. We apply this method to cases of missingness in both endogenous regressor and outcome, using the Oregon Health Insurance Experiment as an example. We check the correlation between missing probabilities and partially observed variables to justify the assumption. Moreover, we find that excluding incomplete data results in a loss of efficiency and insignificant estimators. The proposed estimator reduces the standard error by more than 50% for the estimated effects of the Oregon Health Plan on the elderly.

JEL Numbers: C13, C14, C26, C83

Keywords: Non-monotone missingness; AIPW; GMM; double robustness.

*Ma Yinchu School of Economics, Tianjin University. Email: shenshenyang@tju.edu.cn. I am very grateful to Jason Abrevaya and Sukjin Han for their guidance and support, and I would like thank Brendan Kline, Haiqing Xu, Stephen Donald, Vasiliki Skreta, Daniel Ackerberg, Dean Spears, Stephen Trejo, Isaiah Andrews, James Brand, Shaofei Jiang, Xue Li, Jiangang Zeng, and the participants at the UT Austin writing seminar, 2021 Asian Meeting of the Econometrics Society, 2021 European Winter Meeting of the Econometric Society for their helpful comments and suggestions.
1 Introduction

Missingness within multiple variables is common in practice. When there are two or more missing variables, the missing patterns can be divided into monotone missingness and (strictly) non-monotone missingness\(^1\). Taking the case of two missing variables as an example, monotone missingness occurs when the absence of one variable indicates the absence of the other. In contrast, a non-monotone missing pattern allows for one variable to be observed even when the other is missing. Monotone missingness can be justified by sub-sampling strategies and missing by attrition, both of which frequently occur in multi-stage data collection. However, even when data is collected sequentially at different stages, the missing pattern can still exhibit non-monotone feature. In such cases, the incompleteness of data is not by design, but relates to self-selection. One example is longitudinal survey data where multi-stage missingness occurs because participants skip certain surveys or questions.

The presence of non-monotone missing pattern in multi-stage data introduces more difficulties in identification. A prevalent approach to handling this issue is to drop all observations associated with incomplete data, known as Complete Case (CC) analysis. This approach often yields inefficient estimators. More crucially, when the missing mechanism is related to any endogenous variable, CC analysis can be biased, as noted by Little and Rubin [2002] and Qi and Sun [2014]. Another popular approach in the missing data literature is to assume the Missing at Random (MAR) assumption. This assumption posits that the missing mechanism is independent of the missing value itself, conditional on the observed variables. The MAR assumption is a relaxation of the Missing Completely at Random (MCAR) assumption\(^2\). Under MAR, the propensity of missingness can be easily identified and used to re-weight the sample to avoid bias from proportional missingness. However, the traditional MAR assumption fails with non-monotonically missing variables, as it involves simultaneous dependence between incomplete variables (Robins and Gill [1997], Robins [1997], Little and Rubin [2002], Sun and Tchetgen Tchetgen [2018]). To address this challenge, Chaudhuri and Guilkey [2016] propose a stronger version of MAR, wherein the missing mechanisms of different variables are random conditional on the same set of fully observed variables. This assumption guarantees identification and produces a closed-form efficient influence function. An Augmented Inverse Propensity Score Weighted (AIPW) estimator can be constructed.

Despite the usefulness and power of the above assumption in many common scenarios,

---

\(^1\)In some literature, non-monotone missingness refers to a more general missing pattern that includes both monotone and strictly non-monotone, and also univariate missing patterns (Van Buuren [2018]). Other literature considers only the strictly non-monotone missing case (Chaudhuri and Guilkey [2016]). Later in this paper, we restrict the definition of “non-monotone missingness” to strictly non-monotone missing patterns.

\(^2\)These two assumptions can be interpreted as strong ignorability and conditional ignorability, respectively.
such as multiple missing IVs (Chaudhuri and Guilkey [2016]), this assumption can easily fail when multiple missing variables are collected sequentially. The intuition is that when both variables are missing simultaneously, their missing mechanisms are likely to exhibit a symmetric feature. However, when missingness occurs at different stages, the missing mechanisms can differ, and later stage missingness can even depend on the incomplete variable from the previous stage. We can think of the monotone missing pattern with multi-stage missing variables, where the missing mechanisms are often assumed to follow an updating feature (Chaudhuri [2020]). The difference here is that we allow for a non-monotone missing pattern with such an updating feature. We observe this pattern in survey data associated with the Oregon Health Insurance Experiment, where there is a strong evidence that the missing mechanism of outcome variable is highly correlated with the partially observed treatment variable collected at the previous stage.

The aim of this paper is to provide a new MAR-type assumption that fits such scenario with multi-stage missing variables but allows for non-monotone missing pattern. Under this assumption, we propose an estimator that maintains the desired statistical properties. We show that under a mild assumption of additive separability on the moment function, this new assumption still provides a closed-form efficient influence function. The resulting estimator retains the AIPW form but with asymmetric augmentation terms for different stages. The AIPW estimator maintains the double robust property, which is challenging to achieve with non-monotone missing pattern.

This paper contributes to the literature in the following aspects: First, it contributes to the literature on missing data with a non-monotone missing pattern. Existing literature has established approaches that can be easily applied on simultaneously missing variables with a non-monotone pattern, as well as to the multi-stage missingness with a monotone pattern. However, sequentially collected data with a non-monotone missing pattern, despite being frequently encountered in practice, has not drawn enough attention. This paper proposes a MAR-type assumption and a corresponding AIPW estimator for this specific case and thereby fills the gap. Our framework allows for the presence of later-stage variables even when the corresponding earlier-stage variables are missing, and allow the later stage missing mechanism to differ across observability status of the first stage variable. Therefore, a one-way dependence relationship is allowed between the two partially missing variables. This more flexible assumption aligns well with many observational data sets, particularly in longitudinal surveys where non-monotone missing patterns are commonly seen over multiple stages of data collection.

Second, it is well known that with non-monotone missingness, it is challenging to achieve double robustness and closed-form efficient influence function. Chaudhuri and Guilkey [2016]
accomplished this goal under a new MAR assumption, and they discussed the key condition for achieving a closed-form efficient influence function and efficiency bound is that all missing mechanisms are independent with any missing variable, conditional on the same set of fully observed variables. However, in our setting, this condition does not hold anymore. This paper proposes an estimator that maintains the desired properties even with asymmetric missing mechanisms. This estimator performs better than those derived from CC analysis and IPW moment functions. Moreover, its performance surpasses that of the AIPW estimator designed for monotone missingness, underscoring the importance of considering the non-monotone missing part of the data set.

Third, we justify the new MAR type assumption and the AIPW approach with an empirical example. This approach is applied to the longitudinal survey data collected in the Oregon Health Insurance Experiment. Through simple regression, we observe a significant correlation between the first stage missing variable and later stage missing mechanism, implying that the MAR assumption previously used in the literature is likely invalid. Moreover, by applying a proxy variable collected from the administrative data, we justify the asymmetry of the later stage missing mechanism, depending on the realization of first stage missingness. We focus on the effect of the Oregon Health Program (OHP) for the elderly above 60 years old, and find that simply dropping incomplete observations results in insignificant effects with loss of information. However, with the AIPW approach, we obtain coefficients very close to the CC analysis, but significant at 99% significance level, indicating robust and significant effects of the Medicaid program on the health-related outcomes for the elderly.

The rest of this paper is organized as follows. Section 2 introduces the model and the key assumptions; Section 3 proposes the AIPW moment condition and GMM estimator based on the assumptions introduced in the previous sections; Section 4 discusses the statistical properties of the AIPW estimator; Section 5 illustrates the performance of the AIPW-GMM estimator through the Monte Carlo simulation results; Section 6 offers an empirical example using the Oregon Health Insurance Experiment data. All the proofs are listed in the appendix.

2 Model and Assumptions

To demonstrate the idea, we consider a simple two-stage model. We assume that $X \equiv (X_1, X_2)$ represents the incomplete variables collected at subsequent stages. $W$ is defined to be a vector of the fully observed variables in the data set, and the parameter $\beta^0$ is determined
by the moment condition:

\[ \beta = \beta^0 \text{ if and only if } E[g(X,W;\beta)] = 0, \text{ with} \]

\[ g(X,W;\beta) \equiv g_1(X_1,W;\beta) + g_2(X_2,W;\beta). \]  

(2.1)

The individual subscripts are suppressed for simplicity. We impose additive separability assumption for the moment function for identification purpose under a more complex missing mechanism (to be introduced later). For notational simplicity, we focus on the case with one missing variable at each stage.

Our primary objective is to identify and consistently estimate the parameter \( \beta \). We use \( R_1 \) and \( R_2 \) to denote the observability of \( X_1 \) and \( X_2 \) correspondingly, formally defined as

\[
R_1 = \begin{cases} 
1 & \text{X}_1 \text{ is observed} \\
0 & \text{X}_1 \text{ is missing} 
\end{cases} 
\]  

(2.2)

\[
R_2 = \begin{cases} 
1 & \text{X}_2 \text{ is observed} \\
0 & \text{X}_2 \text{ is missing} 
\end{cases} 
\]  

(2.3)

For later calculations, we further define

\[ p_1(L) = \Pr [R_1 = 1|L] \]

where \( L \) denotes a vector of variables, and the conditional and joint probabilities are defined as:

\[ p_{r_2|r_1}(L) = \Pr [R_2 = r_2|L, R_1 = r_1] \]

\[ p_{r_1,r_2}(L) = \Pr [R_1 = r_1, R_2 = r_2|L], \forall (r_1,r_2) \in \{0,1\}^2. \]

In the subsequent sections, we introduce the patterns of these missing values and explore the potential dependencies between \( R_1, R_2 \) and other associated variables.

### 2.1 Patterns of Missingness

First, we capture the missing mechanism and introduce the assumptions needed to identify the propensity of missingness. The traditional MAR assumption posits that the probability of each missing pattern, conditional on observed variables under that pattern, is independent of the missing values themselves.
The presence of multiple missing values complicates the missing mechanism. Multiple missingness is typically categorized into monotone missingness and non-monotone missingness. A monotone missing pattern characterizes situations where the occurrence of missing values progresses in a step-wise manner, meaning that the missingness of one variable indicates the missingness of subsequent variables. In the context of our framework, this implies that $X_2$ is missing when $X_1$ is missing. Formally, the missing mechanism is defined as monotone if $(1 - R_1)R_2 = 0$ almost surely. Under monotone missingness, each phase can be conceptualized as a subsample of the preceding one. A common cause for this pattern is designed sub-sampling, while another frequently seen reason is data attrition. For instance, when participants exit a survey due to demise or permanent relocation, they do not appear in subsequent assessments.

In many scenarios, even though data collection follows a sequential procedure, the missing patterns still exhibit a non-monotone feature. In survey data, participants may exit a survey at one stage but return later, or they may skip certain questions while answering surveys. Non-monotone missingness encompasses a broader spectrum of missing data patterns, and allows $X_2$ to be observed even when $X_1$ is missing. We adopt a narrow definition of non-monotone missingness, also referred to as strict non-monotone missingness. To provide a more illustrative comparison, we capture different missing patterns in Figure 1.

In the presence of a non-monotone missing pattern, the MAR assumption becomes harder to justify (Robins and Gill [1997], Robins [1997], Little and Rubin [2002], Tsiatis [2007], Tchetgen et al. [2018]), and inherently complicates identification. The challenge was first introduced by Robins and Gill [1997] when they considered a scenario with no fully observed variable and two partially missing variables. In such a scenario, they showed that MAR assumption implicitly implies the MCAR assumption in a logistic model. Similarly, when additional fully observed variables exist, the MAR assumption must be strengthened such that both missing mechanisms depend on the same set of fully observed variables for estimation purposes. Dependence on partially observed variables is not allowed. Following the definition and notations in Robins and Gill [1997], the traditional MAR assumption implies

---

3The statistical literature has focused more on challenge of estimation with non-monotone missingness under the traditional MAR assumption. In contrast, we focus more on the challenge of identification and show that some features allowed by the previous MAR assumption must be excluded to identify the missing mechanism.
Figure 1: Illustration of Different Missing Patterns

(a) Univariate Missing: $X_1$

(b) Univariate Missing: $X_2$

(c) Monotone Missing

(d) Non-monotone Missing
that

\[
\begin{align*}
\Pr \left[ R_1 = 0, R_2 = 0 \mid X_1, X_2, W \right] &= \pi_{00}(W) \\
\Pr \left[ R_1 = 1, R_2 = 0 \mid X_1, X_2, W \right] &= \pi_{10}(W, X_1) \\
\Pr \left[ R_1 = 0, R_2 = 1 \mid X_1, X_2, W \right] &= \pi_{01}(W, X_2) \\
\Pr \left[ R_1 = 1, R_2 = 1 \mid X_1, X_2, W \right] &= \pi_{11}(W, X_1, X_2)
\end{align*}
\]

for some functions \( \pi_{00}, \pi_{10}, \pi_{01} \) and \( \pi_{11} \).

However, such a joint distribution function can only be identified under the strengthened assumption that all missing mechanisms depend only on \( W \). To see this, we rewrite the above probabilities:

\[
\Pr \left[ R_1 = 0, R_2 = 0 \mid X_1, X_2, W \right] = \left( 1 - p_{1|0}(X_1, X_2, W) \right) (1 - p_1(X_1, X_2, W)) = \pi_{00}(W).
\]

This implies that:

\[
\begin{align*}
p_{1|0}(X_1, X_2, W) &= p_{1|0}(W) \\
p_1(X_1, X_2, W) &= p_1(W).
\end{align*}
\]

Then,

\[
\Pr \left[ R_1 = 1, R_2 = 0 \mid X_1, X_2, W \right] = \left( 1 - p_{1|0}(X_1, X_2, W) \right) p_1(X_1, X_2, W) = \pi_{10}(X_1, W)
\]

only holds when \( \pi_{10}(X_1, W) = \pi_{10}(W) \), and similar arguments hold for \( \pi_{10} \) and \( \pi_{11} \). The above statement holds when \((R_1, R_2)\) are at least mean independent of \((X_1, X_2)\) conditional on \( W \).

However, when the variables are sequentially collected, the strengthened MAR assumption above is frequently violated. Later in Section 6, we show the empirical example with evidence that when \( R_1 = 1 \), \( R_2 \) is highly correlated with \( X_1 \) conditional on \( W \). In such case, though it is still possible that \( R_2 \) is independent with \( X_1 \) on the aggregated unconditional level, but it is difficult to justify from empiricist’s perspective. Following this motivation, we propose an asymmetric dependence relationship between missing mechanism and partially observed variables.
2.2 A Sequential Missing at Random Assumption

In this section, we propose a new MAR-type assumption that integrates the “dynamic updating” feature introduced in the monotone missing literature (Chaudhuri [2020]) but allows for a non-monotone missing pattern. To accommodate both features, we impose asymmetric MAR assumption for different stages. For the first stage missing, we assume that $R_1$ is independent of any unobserved values when conditional on the completely observed variables $W$. For the second stage of the missing mechanism $R_2$, we make different assumptions depending on realization of $R_1$. More specifically, we assume that $R_2$ is independent of only $X_2$ when conditioned on $W$ and partially observed $X_1$ when $R_1 = 1$, indicating a dependency between the second stage missing mechanism and first stage partially observed variable. When $R_1 = 0$, we assume that $R_2$ is independent of both $X_1$ and $X_2$ when conditioned on $W$.

The difference between this assumption and previous assumptions used in the literature is that we allow the partially observed variable to correlate with the missing mechanism. By taking advantage of the sequential feature, this dependency is one-way and avoids the identification problem discussed in the previous section. We refer to this assumption as the Sequential MAR (SMAR) assumption, which captures the updating feature. The formal structure is as follows:

**Assumption SMAR.**

\[
\begin{align*}
R_1 & \perp (X_1, X_2) \mid W \\
R_2 & \perp (X_1, X_2) \mid W, R_1 = 0 \\
R_2 & \perp X_2 \mid X_1, W, R_1 = 1.
\end{align*}
\] (2.4)

The SMAR assumption relaxes the previous MAR assumption in the following perspectives: first, it allows for asymmetric ignorability assumptions for different missing patterns; second, it permits $R_2$ to correlate with $X_1$ when $X_1$ is observed, and the independence between $X_2$ and $R_2$ is asserted conditional on a finer set of variables. More information is thus included in capturing the missing mechanism of the second stage. The two key features to be justified are: (1) $R_2$ correlates with $X_1$ only when $X_1$ is observed; (2) $R_1$ does not correlate with $X_2$ conditional on $W$. These two features are justified later in Section 6 with real life data, and now we first provide two examples to justify this assumption.

**Example 1.** In the literature of policy evaluation, the treatment variable and outcome variable are usually collected in two stages to ensure enough time for the policy to take effect. For example, let $X_2$ be a health outcome and $X_1$ be enrollment in a public health program. The realization of $X_1$ always precedes realization of $X_2$, and sometimes $X_1$ is collected before
Consequently, the later-realized health outcome does not influence the enrollment status or its missing mechanism, conditional on past health history. However, the missing status and the observed treatment status can affect $R_2$. One reason could be that researchers’ efforts to reach out to the initial survey participants might differ based on the previous period’s missing status and realized value (e.g., the treatment group receives more attention). Additionally, if the enrollment status is missing because it was not realized for certain participants during collection, it might affect participants’ enthusiasm for continuing participation in subsequent surveys.

**Example 2.** Consider a panel data set where the same measurement variable $X_t$ is collected at different stages for $t = 1, 2$. For instance, let $X_t$ represents a child’s test score, reported by a parent. If the parents report the real grades as long as they have the information, missingness in the first period indicates the parent’s unawareness of the child’s academic performance. In this scenario, missingness is due to a lack of information by participants, and therefore $X_1$ does not affect parents’ motivation to learn the second stage grade. However, for those who report $X_1$, indicating awareness of their child’s performance, their attention to the child’s score in the following semester can vary depending on realized value of $X_1$.

Under Assumption SMAR, the distribution of four different missing patterns is identified, following directly from selection on observables:

\[
\begin{align*}
\Pr [R_1 = 0, R_2 = 0 | X_1, X_2, W] &= (1 - p_{1|0}(W)) \cdot (1 - p_1(W)) \\
\Pr [R_1 = 1, R_2 = 0 | X_1, X_2, W] &= (1 - p_{1|1}(X_1, W)) \cdot p_1(W) \\
\Pr [R_1 = 0, R_2 = 1 | X_1, X_2, W] &= p_{1|0}(W) \cdot (1 - p_1(W)) \\
\Pr [R_1 = 1, R_2 = 1 | X_1, X_2, W] &= p_{1|1}(X_1, W) \cdot p_1(W).
\end{align*}
\]

### 3 The AIPW-GMM Estimator

With identified missing mechanisms, the IPW estimator has been widely applied in the missing data literature to identify target parameters (Rosenbaum and Rubin [1983], Wooldridge [2007], Seaman and White [2013]). Its essence lies in re-weighting the sample to magnify underrepresented subsamples due to missingness, thereby delivering a consistent estimator. Compared to the imputation method (e.g., expectation-maximization, multiple imputations), the IPW estimator sidesteps intricacies of making structural assumptions to compute missing values and is computationally easier to implement. However, one drawback of the IPW approach is that while it eliminates bias, it does not address the loss of efficiency. To enhance
efficiency, one can incorporate an augmentation term, transforming the IPW approach into an AIPW approach.

The AIPW estimator is well-known for maintaining the double robust property and can achieve semiparametric efficiency bound with the right choice of augmentation term (Robins et al. [1994]; Robins [1997]; Carpenter et al. [2006]; Tsiatis [2007]; Chen et al. [2008]; Glynn and Quinn [2010]). However, constructing such an estimator with non-monotone missing data is challenging, especially when the missing mechanisms depend on different variables (Tsiatis [2007], Chaudhuri and Guilkey [2016]). We propose an AIPW estimator with an asymmetrically designed augmentation term, ensuring the estimator still maintains the desired properties.

### 3.1 AIPW Moment Condition

Recall that we impose additive separability on the moment function in 2.1 such that $X_1$ and $X_2$ enter the moment condition via $g_1$ and $g_2$ separately. For notational simplicity, we define $p_1 \equiv p_1(W)$, $p_{11} \equiv p_{11}(X_1, W)$, $p_{01} = p_{01}(W)$, and denote the set of observed variables as $O \equiv (R_1, R_2, R_1X_1, R_2X_2, W)$. We construct the AIPW moment function $g_{aipw}(\beta)$ as:

$$
g_{aipw}(\beta) = \frac{R_1R_2}{p_{11}}g(X,W; \beta) + \phi(O; \beta).
$$

The moment function consists of an IPW moment function and an augmentation term $\phi$, defined as:

$$
\phi(O; \beta) = \left(1 - \frac{R_1R_2}{p_{11}}\right)E[g(X,W; \beta)|W]
+ \left(\frac{R_1}{p_1} - \frac{R_1R_2}{p_{11}}\right)(g_1(X_1, W; \beta) - E[g_1(X_1, W; \beta)|W])
+ p_1 \cdot \left(\frac{R_1}{p_1} - \frac{R_1R_2}{p_{11}}\right)(E[g_2(X_2, W; \beta)|X_1, W] - E[g_2(X_2, W; \beta)|W])
+ (1 - p_1) \cdot \left(\frac{(1-R_1)R_2}{p_{01}} - \frac{R_1R_2}{p_{11}}\right)(g_2(X_2, W; \beta) - E[g_2(X_2, W; \beta)|W]). \quad (3.1)
$$

The augmentation terms above utilize separability feature of the original moment function. With this feature, we separately impute the portions associated with $X_1$ and $X_2$ using different variable sets. The last two terms crucially capture different information sets across missing status of $X_1$ when imputing $X_2$. However, $X_2$ cannot be used to reversely impute $X_1$ because when they are both observed, $X_1$ correlates with $X_2$’s observability.
If we reorganize the augmentation term above, we can also write it as:

\[
\phi(O; \beta) = \left(1 - \frac{R_1R_2}{p_{11}}\right) E[g(X, W; \beta)|W] \\
+ \left(\frac{R_1}{p_1} - \frac{R_1R_2}{p_{11}}\right) (g(X, W; \beta) - E[g(X, W; \beta)|W]) \\
+ (1 - p_1) \left\{ \left(\frac{1 - R_1}{p_{01}} - \frac{R_1R_2}{p_{11}}\right) (g_2(X_2, W; \beta) - E[g_2(X_2, W; \beta)|W]) \\
- \left(\frac{R_1}{p_1} - \frac{R_1R_2}{p_{11}}\right) (E[g_2(X_2, W; \beta)|X_1, W] - E[g_2(X_2, W; \beta)|W]) \right\}.
\]

Note that in the above expression, the first two components are the augmentation terms for a monotone missing pattern, as introduced in Chaudhuri [2020]. The last two components externally take into consideration observations with Assumption SMAR, Overlap and M hold, if \(\beta = 1\).

Nest, we introduce an overlap assumption and the standard assumptions in the GMM method.

**Assumption Overlap.** \(p_1, p_{10} > c\) almost surely in \(W\), \(p_{11} > c\) almost surely in \((W, X_1)\) for some \(c > 0\).

**Assumption M.** (1) \((X_{1,i}, X_{2,i}, W_i, R_{1,i}, R_{2,i}, R_{1,i} X_{1,i}, R_{2,i} X_{2,i})\) are i.i.d;
(2) \(E[g_{aipw}(\beta)]\) is differentiable with respect to \(\beta \in \text{int}(B)\);
(3) Define \(G(\beta) = \frac{\partial}{\partial \beta} E[g_{aipw}], G(\beta)\) has full rank at \(\beta = \beta^0\);
(4) \(\text{Var}(g_{aipw})\) is bounded and positive semidefinite.

The augmentation term encompasses two sets of auxiliary parameters: the missing mechanism of missing data \(p_1, p_{01}, p_{11}\) and the imputed values for missing data: \(E[g_1(X_1, W; \beta)|W], E[g_2(X_2, W; \beta)|W], E[g_2(X_2, W; \beta)|X_1, W]\). Identification of the missing mechanisms has been shown in previous section. The second set of auxiliary parameters can be identified by:

\[
E[g_1(X_1, W; \beta)|W] = E[g_1(X_1, W; \beta)|W, R_1 = 1] \\
E[g_2(X_2, W; \beta)|W] = E[g_2(X_2, W; \beta)|W, R_1 = 0, R_2 = 1] \\
E[g_2(X_2, W; \beta)|X_1, W] = E[g_2(X_2, W; \beta)|X_1, W, R_1 = 1, R_2 = 1].
\]

**Theorem 3.1.** When Assumption SMAR, Overlap and M hold, \(E[g_{aipw}(\beta)] = 0\) if and only if \(\beta = \beta^0\).

The above theorem establishes the identification of the parameter \(\beta\) with the AIPW moment condition

\[
E[g_{aipw}(\beta)] = 0.
\]
The estimation follows two steps. First, we construct appropriate estimators for the nuisance parameters, which include the missing mechanisms and imputed values. The estimation strategy for the nuisance parameters highly depends on the researcher’s prior beliefs about the model structures. When the specification is not clear, a nonparametric estimation approach can be employed to avoid overly restrictive constraints. We apply sieve estimation with series and spline basis for the subsequent simulation and empirical application, and we provide asymptotic properties based on this approach.

After estimating the nuisance parameters, they can be incorporated back into the GMM estimation equation, we then solve for:

\[
\hat{\beta}_{aipw} = \arg\min_{\beta} \hat{g}_{aipw}(\beta) \hat{\Omega} \hat{g}_{aipw}(\beta)
\]

with the estimated chosen weight \(\hat{\Omega}\) and

\[
\hat{g}_{aipw} = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{R_{1,i} R_{2,i}}{p_{11,i}} \hat{g}(X_{1,i}, X_{2,i}, W_i; \beta) + \hat{\phi}_i \right).
\]

4 Statistical Properties

4.1 Double Robustness

The AIPW estimator is well-known for its double robust property. With this property, the estimator remains unbiased if at least one set of nuisance parameters is correctly specified. The AIPW estimator we propose above maintains the double robust property under the relaxed SMAR assumption. The result is formally stated below and proved in the appendix.

**Theorem 4.1.** Under Assumption SMAR, Overlap and M, \(E[g_{aipw}(\beta)] = 0\) when either:

(a) the missing mechanisms \((p_1, p_{11}, p_{01})\), or
(b) the imputed values \((E[g_1(X_1, W; \beta) | W], E[g_2(X_2, W; \beta) | W], E[g_2(X_2, W; \beta) | X_1, W])\) are correctly specified.

As a direct result of the double robust property, in the two-step semiparametric estimation procedure, the variance of the first stage estimation does not affect the efficiency of estimating the target parameter of interest (Newey [1994]). The efficiency of the second stage estimator can be guaranteed once we find the efficient influence function.
4.2 Efficient Estimator

This section calculates the closed-form efficiency bound under the moment condition 3.1 and Assumption SMAR. We first calculate the variance of the moment function.

Proposition 4.1. Let $V$ denote $V(g_{aipw}(\beta))$ evaluated at $\beta = \beta^0$. Under the assumptions SMAR, Overlap and M,

$$V = E \left[ \frac{1}{p_{11}} Var \left( g - E[g|W] | X_1, W \right) \right] + Var \left( E[g|X_1, W] \right) - \Delta,$$

$\Delta$ is equivalent to

$$\Delta = E \left[ \left( \frac{1}{p_{11}} - \frac{1}{p_1} \right) Var \left( g_1|W \right) \right] + E \left[ p_1(2-p_1) \left( \frac{1}{p_{11}} - \frac{1}{p_1} \right) Var \left( E(g_2|X_1, W)|W \right) \right] + E \left[ (1-p_1) \left( \frac{1+p_1}{p_{11}} - \frac{1-p_1}{p_{01}} \right) Var \left( g_2|W \right) \right] + E \left[ 2 \left( \frac{1}{p_{11}} - 1 \right) Cov \left( g_1, E[g_2|X_1, W]|W \right) \right].$$

The moment function $g_{aipw}$ has the form of an efficient influence function, and therefore, the corresponding estimator reaches the lower bound of asymptotic variance.

Theorem 4.2. Suppose the assumptions 2.2, 3.1 and 3.1 hold, then $\sqrt{N}(\hat{\beta} - \beta^0)$ reaches the lower bound of asymptotic variance of any regular estimator, given by

$$\Omega = (G'\tilde{V}^{-1}G)^{-1}.$$

An estimator with an asymptotic variance that is equal to $\Omega$ has the following asymptotic linear representation:

$$\sqrt{N}(\hat{\beta} - \beta^0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(W_i, R_{1,i}, R_{2,i}, X_{1,i}, X_{2,i}),$$

where

$$\psi(W_i, R_{1,i}, R_{2,i}, X_{1,i}, X_{2,i}) = -\Omega^{-1}G'\tilde{V}^{-1}g_{aipw}(W, R_{1}, R_{2}, X_1, X_2; \beta^0).$$

Achieving a closed-form semiparametric efficiency bound is difficult with non-monotone missing mechanisms, especially when multiple missing mechanisms depend on different sets of variables. The above theorem asserts that one key condition for obtaining the closed-form
semiparametric influence function is that two different missing variables affect the moment condition via two separate terms. When this key condition holds, the desired statistical properties still hold.

Let $n$ denote the sample size. Under certain regularity conditions, $\sqrt{n}$ convergence can be maintained with a non-parametric estimation in the first stage. We assume the well-established regularity conditions for Sieve basis functions that are presented in the literature (Newey [1994]; Chen et al. [2008]; Cattaneo [2010]; Chaudhuri and Guilkey [2016]), and we construct $\sqrt{n}$ normality.

**Theorem 4.3.** Let $\hat{E}(w)$ denote a vector of sieve estimation of first-stage nuisance estimators. For each component $e \in E$, suppose $e$ is a function of $d_e$ elements and is $s_e$ times differentiable. Let $\eta = 1$ for power series basis; and $\eta = \frac{1}{2}$ for spline basis. Let $K$ denote the terms in the series estimator, $n$ denote the sample size, and $K = n^\nu$ such that:

$$4\eta + 2 < \frac{1}{\nu} < 4\frac{s_e}{d_e} - 6\eta$$

then, under Assumption SMAR, Overlap, and $M$,

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, (G^\prime \Omega G)^{-1}G^\prime \Omega V \Omega G (G^\prime \Omega G)^{-1}).$$

Furthermore, if $\Omega = V^{-1}$,

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, (G^\prime V^{-1} G)^{-1}).$$

The regularity conditions used here are no different from the conditions used in the work of Cattaneo [2010] and Chaudhuri and Guilkey [2016]. These conditions ensure that the estimation of nuisance parameters converge fast enough so as not to affect the convergence rate of the second-step estimation.

5 Simulation

For both the simulation and application sections, we run examples where the endogenous regressor and the outcome variables are missing at two stages. This is a frequent problem encountered in empirical studies. When the missing regressor is endogenous, it implies that the propensity of observing both variables should correlate with an endogenous variable, and therefore a simple complete case analysis will result in a biased estimator. The challenge of identification in such scenarios has been overlooked in previous literature. Moreover, the
Table 1: Monte Carlo Simulation with Different Values for $corr(\epsilon, u)$

|                  | $\alpha = 0.3$ |                  | $\beta = 0.5$ |                  |
|------------------|----------------|------------------|----------------|------------------|
|                  | $\hat{\alpha}$ | Mean Bias        | RMSE           | $\hat{\beta}$  | Mean Bias        | RMSE           |
| Complete Case    | 0.1715         | 0.1285           | 0.1666         | 0.0075          | 0.1131           |
| IPW              | 0.3013         | 0.0013           | 0.1196         | 0.0037          | 0.1384           |
| AIPW (monotone)  | 0.3025         | 0.0025           | 0.1197         | -0.0050         | 0.1422           |
| AIPW             | 0.3012         | 0.0012           | 0.0858         | 0.0008          | 0.0824           |
| N = 1000, R = 500, $corr(\epsilon, u) = 0.8$ |
| Complete Case    | 0.2219         | 0.0781           | 0.1344         | 0.0054          | 0.1191           |
| IPW              | 0.3015         | 0.0015           | 0.1238         | 0.0011          | 0.1421           |
| AIPW (monotone)  | 0.3025         | 0.0025           | 0.1241         | -0.0007         | 0.1511           |
| AIPW             | 0.3006         | 0.0006           | 0.0889         | 0.0001          | 0.0875           |
| N = 1000, R = 500, $corr(\epsilon, u) = 0.5$ |
| Complete Case    | 0.2550         | 0.0450           | 0.1191         | 0.0039          | 0.1213           |
| IPW              | 0.3016         | 0.0016           | 0.1250         | -0.0004         | 0.1427           |
| AIPW (monotone)  | 0.3020         | 0.0020           | 0.1259         | 0.0018          | 0.1534           |
| AIPW             | 0.3002         | 0.0002           | 0.0908         | 0.00007         | 0.0887           |

commonly observed partial linear relationship between the outcome and the regressor aligns with our additive separability restriction on the moment function.

The previous results suggest that the AIPW-GMM approach yields consistent and efficient results. This section provides numerical evidence of these properties. We consider the following model:

$$X_{2i} = \alpha X_{1i} + \beta W_{2i} + \epsilon_i \equiv 0.3X_{1i} + 0.5W_{2i} + \epsilon_i$$

$$X_{1i} = 1(0.1 + 0.3W_{1i} + 0.1W_{2i} \geq u_i)$$

where $X_{1i}$ and $X_{2i}$ are the partially observed regressors representing treatment and the outcome variable. $W_i \equiv (W_{1i}, W_{2i})$ are the fully observed variables, with $W_{1i}$ being an instrument variable. $\epsilon_i$ and $u_i$ are jointly normally distributed with $\gamma = corr(\epsilon_i, u_i)$. The missing statuses are determined via the binary model stated below:

$$R_1 = 1(0.2 + 0.2W_{2i} + 0.3W_{1i} \geq u_1)$$

$$R_2 = 1(0.3 - 0.05W_{2i} + 0.2W_{1i} + 0.3R_1X_{1i} \geq u_2).$$

Table 1 shows the simulation results with different estimation strategies. We include the
Table 2: Monte Carlo Simulation with Misspecified Imputed Values

|                        | $\alpha = 0.3$ | $\beta = 0.5$ |
|------------------------|----------------|---------------|
|                        | $\hat{\alpha}$ | Mean Bias | RMSE | $\hat{\beta}$ | Mean Bias | RMSE |
| N = 1000, R = 500, misspecified $E[X_2|W], E[X_2|X_1,W]$ | Complete Case | 0.1840 | 0.1160 | 0.1523 | 0.4904 | 0.0096 | 0.1253 |
|                        | IPW            | 0.3016 | 0.0016 | 0.1250 | 0.4996 -0.0004 | 0.1427 |
|                        | AIPW           | 0.3002 | 0.0002 | 0.0907 | 0.4990 | 0.0010 | 0.0870 |
| N = 1000, R = 500, misspecified $E[X_1|W]$ | Complete Case | 0.2550 | 0.0450 | 0.1191 | 0.4961 | 0.0039 | 0.1213 |
|                        | IPW            | 0.3016 | 0.0016 | 0.1250 | 0.4996 | -0.0004 | 0.1427 |
|                        | AIPW           | 0.3001 | 0.0001 | 0.0908 | 0.5003 | 0.0003 | 0.0886 |
| N = 1000, R = 500, misspecified $p_1$ | Complete Case | 0.2550 | 0.0450 | 0.1191 | 0.4961 | 0.0039 | 0.1213 |
|                        | IPW            | 0.3039 | 0.0039 | 0.1278 | 0.4948 | -0.0052 | 0.1485 |
|                        | AIPW           | 0.3002 | 0.0002 | 0.0907 | 0.5005 | 0.0005 | 0.0886 |
| N = 1000, R = 500, misspecified $p_{11}$ | Complete Case | 0.2550 | 0.0450 | 0.1191 | 0.4961 | 0.0039 | 0.1213 |
|                        | IPW            | 0.2535 | 0.0465 | 0.1193 | 0.4980 | -0.0020 | 0.1411 |
|                        | AIPW           | 0.3001 | 0.0001 | 0.0908 | 0.5005 | 0.0005 | 0.0883 |

CC estimator, the estimators from the IPW moment function, the AIPW moment function ignoring the non-monotone missing feature, and the AIPW estimator proposed in this paper. We also vary the correlation between $\epsilon$ and $u$, which can be seen as the endogeneity level of $X_1$. Since $R_2$ is allowed to correlate with $X_1$, the higher the correlation, the more endogenous the missing mechanism. Therefore, when $corr(\epsilon, u)$ is high, the CC estimator induces higher bias, while the other estimators remain unbiased. Since the positive correlation between $\epsilon$ and $u$ implies a negative correlation between $\epsilon$ and $X_1$, the bias has a negative sign. The other finding is that in all exercises, the AIPW estimators have smaller RMSE compared to the other estimation strategies. For both the IPW and AIPW type estimators, we apply series estimation with the polynomial order set to 5 for all nuisance parameters. The IPW estimators have higher RMSE than the CC estimator in some cases. This might result from the fact that IPW estimator also drops all incomplete observations, but runs a nonparametric estimation for nuisance parameters with limited data. We also show results from the other AIPW type estimator, which only allows for a monotone missing pattern (i.e., we ignore the last two components in the augmentation term), and we find its statistical performance no better than the IPW estimator in this case. This demonstrates the importance of considering observations with $R_1 = 0$ and $R_2 = 1$ when faced with a non-monotone missing scenario.

Table 2 shows the simulation results with misspecified nuisance parameters and confirms the double robustness property. We maintain the same DGP with $corr(\epsilon, u) = 0.3$.
for these exercises. For the first two blocks, we apply misspecified imputed values for $E[X_2|W], E[X_2|X_1,W]$ and $E[X_1|W]$ when constructing the AIPW estimator. The performance of the CC and IPW estimators remains the same because these imputed values do not enter the construction of these two estimators. The performance of the AIPW estimator stays almost the same, maintaining the smallest RMSE among the three estimators. For the next two blocks, we apply misspecified missing mechanisms. The CC estimator remains the same, while bias from the IPW estimator starts to increase because these missing mechanisms are key components of the IPW moment function. The AIPW estimator maintains desired statistical property, confirming the double robustness property.

6 An Empirical Example

To better justify the new assumption and AIPW approach proposed above, we apply the approach on the Oregon Health Insurance Experiment (OHIE), a large-scale social experiment. In 2008, a group of low-income individuals was randomly selected for the opportunity to apply for the Oregon Health Plan (OHP) Standard. The OHP Standard is a Medicaid extension program that covers low-income adults who are not eligible for the OHP Plus, which only covers children, pregnant women and families enrolled in the Temporary Assistance for Needy Families Program. The OHP Standard program was not open for applications until 2008 during the OHIE and faced a large number of lottery participants. People registered for the lottery were randomly assigned to win, conditional on the number of household members on the waiting list. Once they won the lottery, winners were required to return an application form. However, only 60.82% of the winners returned the form in time, and among these individuals, some failed the qualification screening. As such, it induced an endogeneity problem.

The data sets used in this paper are composed of three parts. The first data set contains participants’ basic information and administrative data on the lotteries. This data set was recorded by the researchers, and most variables are completely observed. The researchers conducted three follow-up surveys to collect information on health insurance, healthcare needs, expenses and costs. There were three rounds of surveys in total: the initial follow-up survey (0m) was conducted right after the experiment, from June to November 2008, covering 58,405 participants; the intermediate survey (6m) was conducted six months after the experiment for a subsample of 11,756 people; the final round survey (12m) was conducted a year after the experiment for the same group that participated in the initial survey. We include data from the 0m and 12m surveys, with treatment status extracted from the 0m survey and health outcomes taken from the 12m survey. This time gap allows the Medicaid
program to have enough time to take effect.

We merged the three data sets collected from different time periods and the variables selected are summarized in Table 3. We apply same notations as in Section 5 in the table. 50.66% of the survey participants were randomly selected to win the lottery, and selection into the lottery is used as the instrument variable to construct moment condition. The number of household members is a key control variable to guarantee exogeneity of the instrument variable. For this variable, 1 represents a household with a single member, while 2 and 3 represent households with two and more than two members. We select some other variables for control variables too. The age varies from 20 to 63, and this guarantees that the influence of Medicare is excluded. The gender ratio is balanced, with approximately 55% female. Most of the respondents are from the metropolitan statistical area, and less than 10% of participants required a non-English questionnaire.

We record the partially missing variables: endogenous regressor (treatment) and outcomes in the second block. To overall evaluate the effect of the Medicaid program, we choose enrollment in the OHP program, including both the OHP Standard and Plus, as the treatment variable. There are 23,140 participants with observed treatment statuses. We select four outcome variables from the final stage survey to measure the health status of participants: physical activities compared to the same age, depression, whether participants got all needed medical care, and whether they received all needed dental care in the last six months. The first two variables, referred to as “Physical Activities” and “Depression”, measure the physical and mental health of participants, respectively. “Physical Activity” is rated on a three-point scale: 1 (more active), 2 (same) and 3 (less active). “Depression” is an index variable constructed from two questions: “How often have you been disinterested in doing things in the past two weeks?” and “How often have you felt depressed in the past two weeks?”. The latter two outcome variables are binary and measure the medical services participants received. “Got all needed medical care” indicates participants’ satisfaction with the medical care they received, and “Got all needed dental care” measures the usage of medical services outside primary care.

We first show the missing pattern. The main reason for missingness in this data set is non-response to surveys, with the response rates for 0m and 12m surveys shown in Figure 2. For both the 0m and 12m surveys, the response rates are less than 50%. Among those participants, 16.85% only participated in the 0m survey, and 12.32% responded to the 12m but not the 0m survey.

Another important source of missingness is non-response to survey questions among responders. Figure 3 shows the missing rates for the treatment variable and selected outcome variables for the participants who returned both surveys. Among the 16,579 participants
Table 3: Summary Statistics in the OHIE Data

| Variable | Count | Mean  | Sd    | Min | Max |
|----------|-------|-------|-------|-----|-----|
| $W_1$: Selected in the lottery | 58405 | 0.5066 | 0.5000 | 0   | 1   |
| $W_2$: Number of people in household | 58405 | 1.2926 | 0.4608 | 1   | 3   |
| $W_2$: Age | 58405 | 39.8863 | 12.1226 | 20  | 63  |
| $W_2$: Female | 58405 | 0.5464 | 0.4978 | 0   | 1   |
| $W_2$: Zip code in a metropolitan statistical area | 58405 | 0.7658 | 0.4235 | 0   | 1   |
| $W_2$: Individual requested English-language materials | 58405 | 0.9072 | 0.2902 | 0   | 1   |
| $X_1$: Currently have OHP insurance | 23140 | 0.1734 | 0.3786 | 0   | 1   |
| $X_2$: Physical Activities | 23308 | 2.1977 | 0.7335 | 1   | 3   |
| $X_2$: Depression | 23181 | 2.0164 | 0.9863 | 1   | 4   |
| $X_2$: Got All Needed Medical Care | 22940 | 0.6191 | 0.4856 | 0   | 1   |
| $X_2$: Got All Needed Dental Care | 23172 | 0.4102 | 0.4919 | 0   | 1   |

Observations: 58405

Figure 2: Response Rates to the Surveys
who returned both the 0m and 12m surveys, we could not confirm self reported treatment status for 8.84% after correction of the variable. For the chosen outcome variables, the non-response rates vary from 1.91% to 3.60% among responders to both surveys.

The non-response to surveys and survey questions results in non-monotone missing patterns. For each variable, we observe four different missing patterns, consistent with the definition of non-monotone missingness. For each outcome variable, the rate of observing the outcome but not the treatment is about 14%-15%, comprising a relatively large portion of the total participants. We take the outcome variable “Physical Activities” as an example to show its missing patterns in Figure 4, the other outcome variables present very similar proportions of each missing pattern.

Next, we show evidence of a violation of the MAR assumption in this example. Under the strengthened MAR assumption, there should be no correlation between the missingness of the outcome variables and the partially observed treatment variable. We run a regression of observing the outcome variable on the treatment variable for those whose treatment is observed, as a simple test on the correlation between them. As shown in Table 4, when the first stage treatment is observed, there is a significant correlation between the second stage missing mechanism and the first stage partially observed variable. Though statistically, it is

---

4We correct the OHP enrollment status to be 1, regardless of whether it was originally missing, if the survey participants mentioned that they had been successfully enrolled in the OHP Standard and the administrative data showed consistent enrollment status.
still possible that $R_2$ is independent with $X_1$ on the aggregated unconditional level, this case should be rare and lacks a good economic interpretation. It is also necessary to confirm that the first stage missing mechanism does not depend on the second stage partially observed variable. We run regressions for each outcome variable as a simple test, finding very small and insignificant correlations between the first stage missing propensity and any second stage outcome variables when these outcomes are observed. The results are recorded in Table 5.

The other key feature of the SMAR assumption is the asymmetric relation between the first stage missing variable $X_1$ and the second stage missing mechanism $R_2$, conditional on the missing status of $X_1$. More specifically, the first stage partially observed variable only affects the second stage missing mechanism when it is observed. However, there is no easy way to test correlation between $X_1$ and $R_2$ when $X_1$ is unobserved. To better justify this assumption, we borrow the fully observed Medicaid enrollment status from the administrative data, collected by Finkelstein et al. [2012] as a proxy for the treatment variable in our data set, and test its correlation with $R_2$ conditional on different observability status of the treatment variable. From Table 6, we observe a very similar pattern of correlation between $R_2$ and the proxy variable when the treatment variable is observed. On the other hand, when the treatment variable is not observed, we do not observe significant correlation between $R_2$ and the proxy variable, as shown in Table 7. These two tables externally justify the SMAR assumption in
Table 4: Regression of $R^2$ on $X_1$ when $X_1$ is observed

|                                | Physical Activities | Depression | Got all Needed Medical Care | Got all Needed Dental Care |
|--------------------------------|---------------------|------------|----------------------------|---------------------------|
| Currently have OHP insurance   | -0.0373***          | -0.0371*** | -0.0388***                 | -0.0391***                |
|                                 | (0.00855)           | (0.00858)  | (0.00864)                  | (0.00860)                 |
| Selected in the lottery         | -0.0218***          | -0.0226*** | -0.0222***                 | -0.0195**                |
|                                 | (0.00647)           | (0.00651)  | (0.00655)                  | (0.00655)                 |
| Number of people in household   | 0.00648             | 0.0104     | 0.0142*                    | 0.0163*                  |
|                                 | (0.00698)           | (0.00700)  | (0.00705)                  | (0.00702)                 |
| Female                          | 0.0446***           | 0.0424***  | 0.0413***                  | 0.0465***                |
|                                 | (0.00637)           | (0.00639)  | (0.00643)                  | (0.00640)                 |
| Age                             | 0.00516***          | 0.00492*** | 0.00444***                 | 0.00475***               |
|                                 | (0.000258)          | (0.000258) | (0.000260)                 | (0.000259)               |
| Zip code in a metropolitan      | -0.00366            | -0.00737   | -0.000473                  | -0.00450                 |
| statistical area                | (0.00720)           | (0.00722)  | (0.00727)                  | (0.00724)                 |
| Individual requested English    | 0.0427***           | -0.0190    | -0.0175                    | -0.0163                  |
| language materials              | (0.0126)            | (0.0126)   | (0.0127)                   | (0.0126)                 |
| Constant                        | 0.370***            | 0.434***   | 0.435***                   | 0.423***                 |
|                                 | (0.0216)            | (0.0217)   | (0.0218)                   | (0.0217)                 |
| Observations                    | 23140               | 23140      | 23140                      | 23140                    |

Standard errors in parentheses

* $p < 0.05$,  ** $p < 0.01$,  *** $p < 0.001$
Table 5: Regression of $R_1$ on $X_2$ when $X_2$ is observed

|                          | Physical Activities | Depression | Got all Needed Medical Care | Got all Needed Dental Care |
|--------------------------|---------------------|------------|------------------------------|---------------------------|
| [item] Physical activities | -0.000825           |            | 0.0114                       |                           |
|                          | (0.00429)           |            | (0.00657)                    |                           |
| Depression               | -0.00557            |            |                              |                           |
|                          | (0.00323)           |            |                              |                           |
| Got all needed medical care |                   |           |                              |                           |
|                          | 0.00578             |            | 0.00578                      |                           |
|                          | (0.00645)           |            | (0.00645)                    |                           |
| Selected in the lottery  | -0.00670            | -0.00769   | -0.0109                      | -0.00830                  |
|                          | (0.00628)           | (0.00630)  | (0.00636)                    | (0.00631)                 |
| Number of people in household | 0.00317             | 0.00324    | 0.00263                      | 0.00462                   |
|                          | (0.00695)           | (0.00699)  | (0.00700)                    | (0.00696)                 |
| Female                   | 0.0465***           | 0.0443***  | 0.0468***                    | 0.0477***                 |
|                          | (0.00639)           | (0.00639)  | (0.00643)                    | (0.00413***               |
| Age                      | 0.00418***          | 0.00421*** | 0.00410***                   | 0.00413***                |
|                          | (0.0002578)         | (0.000261) | (0.000261)                   | (0.000259)                |
| Zip code in a            | -0.00770            | -0.0111    | -0.00759                     | -0.00939                  |
| metropolitan statistical area | -0.00723         | (0.00725)  | (0.00732)                    | (0.00727)                 |
| Individual requested     | 0.107***            | 0.0989***  | 0.0907***                    | 0.0957***                 |
| English-language materials | 0.0120             | (0.0117)   | (0.0118)                     | (0.0118)                  |
| Constant                 | 0.351***            | 0.363***   | 0.366***                     | 0.348***                  |
|                          | (16.60)             | (16.45)    | (16.00)                      | (0.0213)                  |
| Observations             | 23308               | 23181      | 22940                        | 23172                     |

$t$ statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Table 6: Regression of $R^2$ on proxy of $X_1$ when $X_1$ is observed

|                                  | Physical Activities | Depression | Got all Needed Medical Care | Got all Needed Dental Care |
|----------------------------------|---------------------|------------|-----------------------------|---------------------------|
| Currently have OHP insurance     | -0.0233**           | -0.0241*** | -0.0253***                  | -0.0272***                |
|                                  | (0.00725)           | (0.00727)  | (0.00732)                   | (0.00729)                 |
| Selected in the lottery          | -0.0214**           | -0.0218**  | -0.0214**                  | -0.0181**                |
|                                  | (0.00672)           | (0.00674)  | (0.00679)                   | (0.00676)                 |
| Number of people in household    | 0.00640             | 0.0102     | 0.0140*                    | 0.0160*                  |
|                                  | (0.00699)           | (0.00701)  | (0.00706)                   | (0.00703)                 |
| Female                           | 0.0447***           | 0.0425***  | 0.0414***                  | 0.0468***                |
|                                  | (0.00637)           | (0.00639)  | (0.00643)                   | (0.00640)                 |
| Age                              | 0.00519***          | 0.00494*** | 0.00447***                 | 0.00477***               |
|                                  | (0.000257)          | (0.000258) | (0.000260)                 | (0.000259)               |
| Zip code in a metropolitan statistical area | -0.00376 | -0.00748 | -0.000581 | -0.000462 |
|                                  | (0.00720)           | (0.00722)  | (0.00727)                   | (0.00724)                 |
| Individual requested English-language materials | 0.0437*** | -0.0181 | -0.0165 | -0.0152 |
|                                  | (0.0126)            | (0.0126)   | (0.0127)                    | (0.0126)                 |
| Constant                         | 0.369***            | 0.433***   | 0.434***                   | 0.423                    |
|                                  | (0.0216)            | (0.0217)   | (0.0218)                    | (0.0217)                 |
| Observations                     | 23140               | 23140      | 23140                       | 23140                    |

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The AIPW approach is important in enhancing estimation efficiency, particularly in empirical studies with a small sample size. This increase efficiency is crucial as it affects the significance levels of the estimated parameters. To better show the approach’s performance in the OHP data, we focus on a sample with 3,036 elderly individuals over 60 years, to assess the effect of enrolling in the OHP on various health-related outcomes. The nuisance parameters were estimated using a sieve approach with B-spline basis functions, with the number of knots set to three.

The regression results presented in Table 8 and 9 demonstrate that the AIPW approach significantly improves estimation efficiency. The estimated coefficients obtained by CC and IPW methods are only significant for the outcome variable “Depression”, at 95% signifi-
Table 7: Regression of $R^2$ on proxy of $X_1$ when $X_1$ is not observed

|                                           | Physical Activities | Depression | Got all Needed Medical Care | Got all Needed Dental Care |
|-------------------------------------------|---------------------|------------|-----------------------------|---------------------------|
| Currently have OHP insurance              | 0.00528             | 0.00491    | 0.00492                     | 0.00499                   |
|                                           | (0.00543)           | (0.00542)  | (0.00541)                   | (0.00543)                 |
| Selected in the lottery                   | -0.00936*           | -0.0798    | -0.00538                    | -0.00646                  |
|                                           | (0.00467)           | (0.00466)  | (0.00465)                   | (0.00467)                 |
| Number of people in household             | 0.0208***           | 0.0215***  | 0.0245***                   | 0.0238***                 |
|                                           | (0.00504)           | (0.00502)  | (0.00502)                   | (0.00503)                 |
| Female                                    | 0.0448***           | 0.0457***  | 0.0430***                   | 0.0441***                 |
|                                           | (0.00455)           | (0.00454)  | (0.00453)                   | (0.00455)                 |
| Age                                       | 0.00356***          | 0.00343*** | 0.00334***                  | 0.00343***                |
|                                           | (0.000191)          | (0.000190)| (0.000190)                  | (0.000191)                |
| Zip code in a metropolitan statistical area | -0.0241***       | -0.0215*** | -0.0222***                  | -0.0220***                |
|                                           | (0.00548)           | (0.00547)  | (0.00546)                   | (0.00548)                 |
| Individual requested                      | 0.0280***           | 0.0123     | 0.0210**                    | 0.0165*                   |
| English-language materials                | (0.00747)           | (0.00745)  | (0.00744)                   | (0.00746)                 |
| Constant                                  | 0.0502***           | 0.0636***  | 0.0546***                   | 0.0584***                 |
|                                           | (0.0142)            | (0.0142)   | (0.0142)                    | (0.0142)                  |
| Observations                              | 35265               | 35265      | 35265                       | 35265                     |

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
cance level, while the AIPW method provided estimators significant at 99% level. For the other outcome variables, the CC and IPW estimators are not significant, while the AIPW estimators are significant at 99% level. These results confirm that the OHP program has a positive impact on the health of the elderly, significantly improving physical activity ability and reducing depression levels. Moreover, the program increased the probability of receiving adequate primary and dental care by 0.269 and 0.296, respectively. The change in significance level indicates severe loss of efficiency when incomplete observations are dropped\textsuperscript{5}.

In the above empirical analysis, we do not find evidence suggesting that CC estimator diverges from the AIPW estimator due to bias. One possible explanation is that the dependence between $R_2$ and $X_1$ is not as strong as in the simulation exercises. An interesting observation is that for some outcomes, such as “Depression”, the estimates from CC estimator are closer to those from the AIPW approach, compared to the IPW. Though the difference is omittable, and the testing suggests equivalence between all the estimators, the small deviation might be attributed to inclusion of estimated nuisance parameters. However, the AIPW estimator maintains the double robustness property, and thereby less affected by the bias in estimating the nuisance parameters.

7 Conclusion

This paper addresses the issue of non-monotone missingness at two stages and seeks to develop an estimator that is both unbiased and efficient. Under the SMAR assumption, we can derive a closed-form semiparametric efficient influence function. This allows us to propose an AIPW estimator that achieves the closed-form efficiency bound. The AIPW estimator has an asymmetric form for different stage missingness, and therefore maintains double robustness property though the missing mechanisms depend on different sets of variables.

We provide an empirical example to support the SMAR assumption. In this example, we find that when the first stage treatment variable is observed, it affects the missing mechanism of outcome variables collected at the second stage. Conversely, when the first stage treatment variable is not observed, there is no evidence of correlation between the second stage missing mechanism and a proxy for the first stage treatment. Such asymmetric dependence does not hold between first stage missing mechanism and second stage partially observed variables. In the empirical example, the AIPW approach reduces the standard error by approximately 50% across all four exercises, compared to the commonly used CC estimator. Such improvement directly alters the significance level, indicating a significantly positive effect of enrolling

\textsuperscript{5}Estimation with the full sample presents a similar pattern in standard error reduction, but due to the large sample size, the significance level does not change as dramatically.
### Table 8: Regression Results: Health Status

|                  | Physical Activities | Depression |
|------------------|---------------------|-------------|
|                  | (1) CC GMM          | (2) IPW GMM | (3) AIPW GMM | (1) CC GMM | (2) IPW GMM | (3) AIPW GMM |
| OHP              | -0.460 (0.271)     | -0.457 (0.270) | -0.500*** (0.134) | -0.932** (0.338) | -0.912** (0.335) | -0.931*** (0.181) |
| Female           | 0.0540 (0.0471)    | 0.0506 (0.0471) | 0.0124 (0.0229) | -0.00906 (0.0581) | -0.0132 (0.0581) | 0.0455 (0.0313) |
| Number of Household Members | -0.0786 (0.0558) | -0.0860 (0.0559) | -0.0431 (0.0267) | -0.194** (0.0654) | -0.201** (0.0666) | -0.193*** (0.0352) |
| Age              | -0.0314 (0.206)    | -0.0333 (0.0206) | -0.0340*** (0.0101) | -0.0601* (0.0257) | -0.0634* (0.0257) | -0.101*** (0.0133) |
| MSA              | 0.0335 (0.0492)    | 0.0275 (0.0492) | 0.00230 (0.0234) | -0.0197 (0.0623) | -0.0241 (0.0625) | 0.0415 (0.0319) |
| English-Speaking | 0.126 (0.102)      | 0.118 (0.102) | 0.0150 (0.0498) | -0.00293 (0.123) | -0.0183 (0.125) | -0.175* (0.0702) |
| Constant         | 4.026** (1.301)    | 4.169** (1.300) | 4.337*** (0.627) | 6.012*** (1.616) | 6.237*** (1.619) | 8.642*** (0.836) |
| Observations     | 1222               | 1220         | 3036           | 1215          | 1213          | 3036          |

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
|                          | Got All Needed Medical Care |                       | Got All Needed Dental Care |                       |
|--------------------------|-----------------------------|-----------------------|---------------------------|-----------------------|
|                          | (1) CC GMM                 | (2) IPW GMM           | (3) AIPW GMM              | (1) CC GMM             | (2) IPW GMM | (3) AIPW GMM |
| OHP                      | 0.268                      | 0.264                 | 0.268***                  | 0.224                 | 0.238       | 0.296***     |
|                          | (0.156)                    | (0.158)               | (0.0814)                  | (0.169)               | (0.174)     | (0.0882)     |
| Female                   | -0.0398                    | -0.0408               | -0.0245                   | 0.000966              | 0.000643    | -0.0290      |
|                          | (0.0279)                   | (0.0280)              | (0.0139)                  | (0.0305)              | (0.0306)    | (0.0151)     |
| Number of Household Members | 0.0735*                      | 0.0717*               | 0.121***                  | 0.0916**              | 0.0944**    | 0.0729***    |
|                          | (0.0306)                   | (0.0319)              | (0.0156)                  | (0.0352)              | (0.0356)    | (0.0175)     |
| Age                      | 0.0217                     | 0.0217                | 0.0303***                 | 0.0214                | 0.0227      | 0.0235***    |
|                          | (0.0126)                   | (0.0127)              | (0.00611)                 | (0.0133)              | (0.0133)    | (0.00657)    |
| MSA                      | 0.0332                     | 0.0349                | 0.0193                    | 0.00802               | 0.00894     | -0.0103      |
|                          | (0.0300)                   | (0.0302)              | (0.0147)                  | (0.0322)              | (0.0323)    | (0.0157)     |
| English-Speaking         | 0.0215                     | 0.0223                | 0.00684                   | -0.187**              | -0.182**    | -0.115***    |
|                          | (0.0636)                   | (0.0640)              | (0.0312)                  | (0.0638)              | (0.0643)    | (0.0340)     |
| Constant                 | -0.777                     | -0.775                | -1.374***                 | -0.814                | -0.902      | -0.962*      |
|                          | (0.795)                    | (0.801)               | (0.382)                   | (0.835)               | (0.836)     | (0.411)      |
| Observations             | 1177                       | 1175                  | 3031                      | 1203                  | 1201        | 3031         |

Standard errors in parentheses

* p < 0.05, ** p < 0.01, *** p < 0.001
in a Medicaid program on the health of the elderly. Simulation exercises further support the desirable properties of the AIPW estimator, confirming its efficiency and robustness in practical applications.

A Proofs

A.1 Proof of Theorem 3.1

To prove the theorem, we first show that \(E[m_{aipw}(\beta)] = E[m(\beta)], \forall \beta \in B\).

We break down the proof into individual terms. For any fixed \(\beta\), firstly, we show that

\[
E \left[ \frac{R_1 R_2}{p_{11}} g(X, W; \beta) \right] = E \left[ g(X, W; \beta) \right].
\]

Following the law of iterated expectation,

\[
E \left[ \frac{R_1 R_2}{p_{11}} g(X, W; \beta) \right] = E \left\{ E \left[ \frac{R_1 R_2}{p_{11}} g(X, W; \beta) | X, W \right] \right\}
= E \left\{ g(X, W; \beta) \frac{1}{p_{11}} E \left[ R_2 | X, W, R_1 = 1 \right] \cdot E \left[ R_1 | X, W \right] \right\}
= E \left\{ g(X, W; \beta) \frac{p_{11}(X_1, W) \cdot p_1(W)}{p_{11}} \right\}
= E \left[ g(X, W; \beta) \right].
\]

The third equality sign follows the SMAR assumption. Next, we show the augmentation term equals to zero. This follows from:

\[
E \left[ \frac{1 - R_1 R_2}{p_{11}} | X, W \right] = 0
\]
\[
E \left[ \frac{R_1}{p_1} - \frac{R_1 R_2}{p_{11}} | X, W \right] = 0
\]
\[
E \left[ \frac{(1 - R_1) R_2}{p_{01}} - \frac{R_1 R_2}{p_{11}} | X, W \right] = 0.
\]

The above equalities hold directly from the SMAR assumption. By the law of iterated expectation,

\[
E \left[ (1 - \frac{R_1 R_2}{p_{11}}) E[g|W] \right] = E \left[ E[g|W] \cdot E \left[ 1 - \frac{R_1 R_2}{p_{11}} | X, W \right] \right] = 0.
\]
Similarly,

\[
E \left[ \left( \frac{R_1}{p_1} - \frac{R_1 R_2}{p_{11}} \right) (g_1(X_1, W; \beta) - E[g_1(X_1, W; \beta) | W]) \right]
\]

\[
= E \left[ (g_1(X_1, W; \beta) - E[g_1(X_1, W; \beta) | W]) E\left[ \frac{R_1}{p_1} - \frac{R_1 R_2}{p_{11}} X, W \right] \right] = 0,
\]

\[
E \left[ p_1 \left( \frac{R_1}{p_1} - \frac{R_1 R_2}{p_{11}} \right) (E[g_2|X_1, W] - E[g_2|W]) \right]
\]

\[
= E \left[ p_1 \left( E[g_2|X_1, W] - E[g_2|W] \right) E\left[ \frac{R_1}{p_1} - \frac{R_1 R_2}{p_{11}} X, W \right] \right] = 0,
\]

\[
E \left[ (1 - p_1) \left( \frac{(1 - R_1) R_2}{p_{01}} - \frac{R_1 R_2}{p_{11}} \right) (g_2 - E[g_2|W]) \right]
\]

\[
= E \left[ (1 - p_1) \left( g_2 - E[g_2|W] \right) E\left[ \frac{(1 - R_1) R_2}{p_{01}} - \frac{R_1 R_2}{p_{11}} X, W \right] \right] = 0.
\]

In conclusion, we prove that \( E[g_{aipw}(\beta)] = E[g(\beta)] \), \( \forall \beta \). Since \( E[g(\beta)] = 0 \) if and only if \( \beta = \beta_0 \), \( E[g_{aipw}(\beta)] = 0 \) if and only if \( \beta = \beta_0 \).

\[\square\]

### A.2 Proof of Theorem 4.1

The proof of Theorem 3.1 demonstrates that \( E[g_{aipw}(\beta)] = 0 \) under condition (a). Now we proceed to show that \( E[g_{aipw}(\beta)] = 0 \) under condition (b).
We first reorganize the AIPW moment function into:

\[ g_{aipw}(X, W; \beta) = \left( \frac{R_1 R_2}{p_{11}} g(X, W; \beta) + \left( 1 - \frac{R_1 R_2}{p_{11}} \right) E[g(X, W; \beta)|W] \right) \]

\[ + \left( \frac{R_1}{p_1} - \frac{R_1 R_2}{p_{11}} \right) \left( E[g(X, W; \beta)|X_1, W] - E[g(X, W; \beta)|W] \right) \]

\[ + (1 - p_1) \left\{ \left( \frac{1 - R_1}{p_{01}} - \frac{R_1 R_2}{p_{11}} \right) (g_2(X_2, W; \beta) - E[g_2(X_2, W; \beta)|W]) \right. \]

\[ - \left. \left( \frac{R_1}{p_1} - \frac{R_1 R_2}{p_{11}} \right) (E[g_2(X_2, W; \beta)|X_1, W] - E[g_2(X_2, W; \beta)|W]) \right\} . \]

Note that the sum of the first three terms can be rewritten as

\[ (1) + (2) + (3) = E[g(X, W; \beta)|W] + \frac{R_1 R_2}{p_{11}} (g(X, W; \beta) - E[g(X, W; \beta)|X_1, W]) \]

\[ + \frac{R_1}{p_1} (E[g(X, W; \beta)|X_1, W] - E[g(X, W; \beta)|W]) . \]

When \( E[g(X, W; \beta)|W] \) and \( E[g(X, W; \beta)|X_1, W] \) are correctly specified,

\[ E[E[g(X, W; \beta)|W]] = E[g(X, W; \beta)] = 0. \]

\[ E \left[ \frac{R_1 R_2}{p_{11}} (g(X, W; \beta) - E[g(X, W; \beta)|X_1, W]) \right] \]

\[ = E \left\{ E \left[ \frac{R_1 R_2}{p_{11}} (g(X, W; \beta) - E[g(X, W; \beta)|X_1, W]) |X_1, W \right] \right\} \]

\[ = E \left\{ \frac{R_1 R_2}{p_{11}} E[g(X, W; \beta) - E[g(X, W; \beta)|X_1, W] |X_1, W] \right\} = 0. \]

The second equality follows that under the assumption SMAR, \( R_1 R_2 \) is independent with \( X_2 \) conditional on \( (X_1, W) \). Similarly,

\[ E \left[ \frac{R_1}{p_1} (E[g(X, W; \beta)|X_1, W] - E[g(X, W; \beta)|X_1, W]) \right] = 0 \]
The last term in the above expression of \( g_{aipw}(X, W; \beta) \) can be rewritten as

\[
(4) = -(1 - p_1) \frac{R_1}{p_1} \left( E\left[ g_2(X_2, W; \beta) \right| X_1, W] - E\left[ g_2(X_2, W; \beta) \right| W] \right) \\
- (1 - p_1) \frac{R_1 R_2}{p_{11}} \left( g_2(X_2, W; \beta) - E\left[ g_2(X_2, W; \beta) \right| X_1, W] \right) \\
+ (1 - p_1) \left( 1 - R_1 \right) \frac{R_2}{p_{01}} \left( g_2(X_2, W; \beta) - E\left[ g_2(X_2, W; \beta) \right| W] \right).
\]

Following analogous analysis, under the assumption SMAR, \( R_1 \perp X_1 | W \), \( R_1 R_2 \perp X_2 | (X_1, W) \), and \( (1 - R_1) R_2 \perp X_2 | W \). When \( E\left[ g_2(X_2, W; \beta) \right| X_1, W] \) and \( E\left[ g_2(X_2, W; \beta) \right| W] \) are correctly specified, according to law of iterated expectation, expected values of the above terms in (4) are all equivalent to 0, and therefore \( E\left[ g_{aipw}(X, W; \beta) \right] = 0 \) when the imputed values are correctly specified.

\[ \square \]

A.3 Proof of Proposition 4.1

We proceed calculation of \( V \) term by term. For notational simplicity, we use \( g \) to represent \( g(X, W; \beta) \), and \( g_1, g_2 \) for \( g_1(X_1, W; \beta) \) and \( g_2(X_2, W; \beta) \) respectively.

We decompose \( g_{aipw}(\beta) \) into:

\[
g_{aipw}(\beta) = \frac{R_1 R_2}{p_{11}} (g - E\left[ g \right| W]) + E\left[ g \right| W] \\
+ \left( \frac{R_1}{p_1} - \frac{R_1 R_2}{p_{11}} \right) (g_1 - E\left[ g_1 \right| W]) \\
+ p_1 \left( \frac{R_1}{p_1} - \frac{R_1 R_2}{p_{11}} \right) (E\left[ g_2 \right| X_1, W] - E\left[ g_2 \right| W]) \\
+ (1 - p_1) \left( \frac{(1 - R_1) R_2}{p_{01}} - \frac{R_1 R_2}{p_{11}} \right) (g_2 - E\left[ g_2 \right| W]) \quad \text{(A.1)}
\]
We first calculate $Var \((1)\)$

$$Var \((1)\) = E \left[ Var \left( \frac{R_1 R_2}{p_{11}} (g - E [g|W] |X_1, W) \right) \right] + Var \left( E \left[ \frac{R_1 R_2}{p_{11}} (g - E [g|W]) |X_1, W) \right] \right)$$

$$= E \left[ \frac{1}{p_{11}} Var \left( g - E [g|W] |X_1, W) \right) \right] + Var \left( E [g|X_1, W] - E [g|W]) \right)$$

$$= E \left[ \frac{1}{p_{11}} Var \left( g - E [g|W] |X_1, W) \right) \right] + Var \left( E [g|X_1, W] \right) - Var \left( E [g|W]) \right).$$

The last equality holds by $Cov \left( E [g|X_1, W], E [g|W]\right) = Var \left( E [g|W]\right)$. We also note that $Var \left( E [g|W]\right) = Var \left( (2)\right)$, and

$$Cov \left( (1), (2)\right) = E \left[ \frac{R_1 R_2}{p_{11}} (g - E [g|W]) E [g|W]\right] = E \left[ (g - E [g|W]) E [g|W]\right] = 0.$$

Therefore, $Var \left( (1) + (2)\right) = E \left[ \frac{1}{p_{11}} Var \left( g - E [g|W] |X_1, W) \right) \right] + Var \left( E [g|X_1, W] \right)$.

Now we proceed with the left terms. Firstly, we notice that the covariance between (2) and any other terms are zero, the proof is similar to the one above. Then, we combine the left terms with same components, and get:

$$Var \left( (3)\right) + 2Cov \left( (1), (3)\right) = -E \left[ \left( \frac{1}{p_{11}} - \frac{1}{p_1}\right)^2 (g_1 - E [g_1|W])^2 \right]$$

$$- 2E \left[ \left( \frac{1}{p_{11}} - \frac{1}{p_1}\right) (g_1 - E [g_1|W]) (g_2 - E [g_2|W]) \right]$$

$$= -E \left[ \left( \frac{1}{p_{11}} - \frac{1}{p_1}\right) Var \left( g_1|W) \right] \right]$$

$$- 2E \left[ \left( \frac{1}{p_{11}} - \frac{1}{p_1}\right) Cov \left( g_1, E [g_2|X_1, W]|W\right) \right].$$

$$Var \left( (4)\right) + Cov \left( (1), (4)\right) = -E \left[ p_1 (1 - p_1) \left( \frac{1}{p_{11}} - \frac{1}{p_1}\right)^2 \right]$$

$$- E \left[ 2p_1 \left( \frac{1}{p_{11}} - \frac{1}{p_1}\right) (g_1 - E [g_1|W]) (E [g_2|X_1, W] - E [g_2|W]) \right]$$

$$= -E \left[ p_1 (2 - p_1) \left( \frac{1}{p_{11}} - \frac{1}{p_1}\right) Var \left( E(g_2|X_1, W)|W\right) \right]$$

$$- 2E \left[ p_1 \left( \frac{1}{p_{11}} - \frac{1}{p_1}\right) Cov \left( g_1, E [g_2|X_1, W]|W\right) \right].$$
\[ Var((5)) + Cov((1), (5)) = -E \left[ (1 - p_1) \left( \frac{1 - p_1}{p_{01}} - \frac{1 + p_1}{p_{11}} \right) (g_2 - E[g_2|W])^2 \right] \]
\[ - E \left[ \frac{p_1}{p_{11}} (g_1 - E[g_1|W]) (g_2 - E[g_2|W]) \right] \]
\[ = -E \left[ (1 - p_1) \left( \frac{1 + p_1}{p_{11}} - \frac{1 - p_1}{p_{01}} \right) Var(g_2|W) \right] \]
\[ - 2E \left[ \frac{p_1}{p_{11}} Cov(g_1, E[g_2|X_1, W]|W) \right]. \]

The last terms in the above three equations, as well as \( Cov((3), (4)) \) and \( Cov((3), (5)) \) can all be written in terms of \((g_1 - E[g_1|W]) (E[g_2|X_1, W] - E[g_2|W])\). We combine them and get

\[ E \left[ 2 \left( 1 - \frac{1}{p_{11}} \right) Cov(g_1, E[g_2|X_1, W]|W) \right]. \]

We proved that

\[ \Delta = E \left[ \left( \frac{1}{p_{11}} - \frac{1}{p_1} \right) Var(g_1|W) \right] \]
\[ + E \left[ p_1(2 - p_1)(\frac{1}{p_{11}} - \frac{1}{p_1})Var(E(g_2|X_1, W)|W) \right] \]
\[ + E \left[ (1 - p_1) \left( \frac{1 + p_1}{p_{11}} - \frac{1 - p_1}{p_{01}} \right) Var(g_2|W) \right] \]
\[ + E \left[ 2 \left( \frac{1}{p_{11}} - 1 \right) Cov(g_1, E[g_2|X_1, W]|W) \right]. \]

\[ \square \]

A.4 Proof of Theorem 4.2

We proceed the proof the efficiency following the classical three steps. The logic of this proof and notations used are closely related to Chaudhuri and Guilkey [2016], despite different data structure. We appreciate the clear and inspiring rationale provided by Chaudhuri and Guilkey [2016] to finish this proof.

Step 1 We first construct a set of fully observed variables in the data set as \( O = (W, R_1, R_2, R_1X_1, R_2X_2) \). We consider a class of parametric submodels indexed by \( \theta \) such
that the distribution of $O$ can be expressed as

\[
\begin{align*}
  f_\theta(O) &= \left[p_{\theta}^{11}(W, X_1) f_\theta(X, X_2|W)\right]^{R_1 R_2} \times \left[p_{\theta}^{10}(W, X_1) f_\theta(X_1|W)\right]^{R_1 (1-R_2)} \\ 
  &\quad \times \left[p_{\theta}^{01}(W) f_\theta(X_2|W)\right]^{(1-R_1) R_2} \times \left[p_{\theta}^{00}(W)\right]^{(1-R_1)(1-R_2)} f_\theta(W),
\end{align*}
\]

where

\[
\begin{align*}
  p_{\theta}^{11}(W, X_1) &= \text{Pr}[R_1 = 1, R_2 = 1|W, X] = \text{Pr}[R_1 = 1, R_2 = 1|W, X_1] \\
  p_{\theta}^{10}(W, X_1) &= \text{Pr}[R_1 = 1, R_2 = 0|W, X_1] \\
  p_{\theta}^{01}(W) &= \text{Pr}[R_1 = 0, R_2 = 1|W, X_2] = \text{Pr}[R_1 = 1, R_2 = 1|W, X_1] \\
  p_{\theta}^{00}(W) &= \text{Pr}[R_1 = 0, R_2 = 1|W].
\end{align*}
\]

The second and fourth equality holds under assumption SMAR. The score function is defined as:

\[
S_\theta(O) = s_\theta(W) + R_1 R_2 s_\theta(X_1, X_2|W) + R_1 (1-R_2) s_\theta(X_1|W) + (1-R_1) R_2 s_\theta(X_2|W)
\]

\[
+ \left(R_1 R_2 \frac{\hat{p}_{\theta}^{11}(W, X_1)}{p_{\theta}^{11}(W, X_1)} + R_1 (1-R_2) \frac{\hat{p}_{\theta}^{10}(W, X_1)}{p_{\theta}^{10}(W, X_1)} + (1-R_1) R_2 \frac{\hat{p}_{\theta}^{01}(W)}{p_{\theta}^{01}(W)} + (1-R_1)(1-R_2) \frac{\hat{p}_{\theta}^{00}(W)}{p_{\theta}^{00}(W)}\right)
\]

where $s_\theta(\cdot) \equiv \frac{\partial}{\partial \theta} f_\theta(\cdot)$, and $\hat{p}_{\theta}^{ij}(\cdot) = \frac{\partial}{\partial \theta} p_{\theta}^{ij}(\cdot)$ for $i, j \in \{0, 1\}^2$.

The tangent set is characterized by:

\[
T \equiv R_1 R_2 f_{11}(W, X_1, X_2) + R_1 (1-R_2) f_{10}(W, X_1) + (1-R_1) R_2 f_{01}(W, X_2) + f_{00}(W)
\]

\[
+ R_1 R_2 \frac{\hat{p}_{\theta}^{11}(W, X_1)}{p_{\theta}^{11}(W, X_1)} + R_1 (1-R_2) \frac{\hat{p}_{\theta}^{10}(W, X_1)}{p_{\theta}^{10}(W, X_1)} + (1-R_1) R_2 \frac{\hat{p}_{\theta}^{01}(W)}{p_{\theta}^{01}(W)} + (1-R_1)(1-R_2) \frac{\hat{p}_{\theta}^{00}(W)}{p_{\theta}^{00}(W)}
\]

where

\[
\begin{align*}
  f_{11}(W, X_1, X_2) &\in L_0^2 \left(F(X_1, X_2|W)\right), f_{10}(W, X_1) \in L_0^2 \left(F(X_1|W)\right) \\
  f_{01}(W, X_2) &\in L_0^2 \left(F(X_2|W)\right), f_{00}(W) \in L_0^2 \left(F(W)\right).
\end{align*}
\]

\[
\begin{align*}
  \hat{p}_{11}(W, X_1) + \hat{p}_{10}(W, X_1) + \hat{p}_{01}(W) + \hat{p}_{00}(W) &= 0 \\
  p_{11}(W, X_1) + p_{10}(W, X_1) + p_{01}(W) + p_{00}(W) &= 1, \forall (X_1, W).
\end{align*}
\]

**Step 2** For the second step, we make a conjecture on the efficient influence function and verify that it lies in the tangent set by showing pathwise differentiability of $\beta^0$. 

36
From the moment condition of the model, \( AE [g(W, X_1, X_2; \beta^0)] = 0 \) for any matrix \( A \) of size \( d_\beta \times d_g \), with \( d_\beta \) being the dimension of unknown parameters \( \beta \) and \( d_g \) being the number of moment conditions. The role of \( A \) is to convert an possible over identification into a just identified moment conditions system. Recall that in our framework, \( g(W, X; \beta) = g_1(W, X_1; \beta) + g(W, X_2; \beta) \).

\[
\frac{\partial}{\partial \theta} \beta^0(\theta_0) = -(AG)^{-1}AE \left[ g(W, X_1, X_2; \beta^0) \frac{\partial}{\partial \theta} \log f(\theta, X) \right] \\
= -(AG)^{-1}AE \left[ g(W, X_1, X_2; \beta^0) \left( s(W)' + s(X_1|W)' + s(X_2|W, X_1)' \right) \right]
\]

Then, we conjecture the efficient influence function \( \varphi \) to be

\[
\varphi(A, W, R_1X_1, R_2X_2, R_1, R_2; \beta^0) = -(AG)^{-1}Ag_{aipw}(W, R_1X_1, R_2X_2, R_1, R_2; \beta^0).
\]

It is easy to show \( \varphi(A, W, R_1X_1, R_2X_2, R_1, R_2; \beta^0) \in T \) since only available information is used under each observability pattern, and we want to show that pathwise differentiability holds such that

\[
E \left[ \varphi S^\prime_O \right] = E \left[ g_{aipw}(O; \beta^0) (s(W)' + s(X_1|W)' + s(X_2|W, X_1)') \right].
\]

Recall that from proof of Theorem 4.1, \( g_{aipw} \) can be reorganized into:

\[
g_{aipw}(O; \beta) = E \left[ g(X, W; \beta)|W \right] + \frac{R_1R_2}{p_{11}} (g(X, W; \beta) - E [g(X, W; \beta)|X_1, W])
\]

\[
= (1 - p_1) \frac{R_1}{p_1} (E [g_2(X_2, W; \beta)|X_1, W] - E [g_2(X_2, W; \beta)|W])
\]

\[
= (1 - p_1) \frac{R_1R_2}{p_{11}} (g_2(X, W; \beta) - E [g_2(X, W; \beta)|X_1, W])
\]

\[
+ (1 - p_1) \frac{(1 - R_1)R_2}{p_{01}} (g_2(X_2, W; \beta) - E [g_2(X_2, W; \beta)|W]).
\]

Then, we confirm term by term. We first separate \( S_O \) into two terms regarding the data distribution and the propensity scores:

\[
S_\theta(O) = s_\theta(W) + R_1R_2s_\theta(X_1, X_2|W) + R_1(1 - R_2)s_\theta(X_1|W) + (1 - R_1)R_2s_\theta(X_2|W) + q(R_1, R_2, W, X_1)
\]
where
\[
q(R_1, R_2, W, X_1) \equiv R_1 R_2 \frac{\hat{p}^{11}_0(W, X_1)}{p^1_0(W, X_1)} + R_1 (1 - R_2) \frac{\hat{p}^{10}_0(W, X_1)}{p^1_0(W, X_1)} + (1 - R_1) R_2 \frac{\hat{p}^{01}_0(W)}{p^0_0(W, X_1)} + (1 - R_1) (1 - R_2) \frac{\hat{p}^{00}_0(W)}{p^0_0(W, X_1)}.
\]

And we first show that \( E[g_{aipw}(O; \beta^0)q(R_1, R_2, W, X_1)] = 0 \). When we proceed term by term, by law of iterated expectation conditional on \((W, X_1)\),

\[
E[(1 \cdot q(R_1, R_2, W, X_1)] = E[E[g(X, W; \beta)|W](\hat{p}^{11}_1(W, X_1) + \hat{p}^{10}_1(W, X_1) + \hat{p}^{01}_1(W) + \hat{p}^{00}_1(W))] = 0.
\]

Moreover,

\[
E[(2) \cdot q(R_1, R_2, W, X_1)] = E\left[ \frac{R_1 R_2}{p^{11}_1} (g(X, W; \beta) - E[g(X, W; \beta)|X_1, W]) \frac{\hat{p}^{11}_1(W, X_1)}{p^1_0(W, X_1)} \right] = 0,
\]

\[
E[(4) \cdot q(R_1, R_2, W, X_1)] = E\left[ (1 - p_1) \frac{R_1 R_2}{p^{11}_1} \frac{(g_2(X_2, W; \beta) - E[g_2(X_2, W; \beta)|X_1, W])}{p^1_0(W, X_1)} \hat{p}^{11}_0(W, X_1) \right] = 0,
\]

\[
E[(5) \cdot q(R_1, R_2, W, X_1)] = E\left[ (1 - p_1) \frac{(1 - R_1) R_2}{p^{01}_1} (g_2(X_2, W; \beta) - E[g_2(X_2, W; \beta)|W]) \frac{\hat{p}^{01}_0(W)}{p^0_0(W)} \right] = 0.
\]

The only remaining term is

\[
E[(3) \cdot q(R_1, R_2, W, X_1)] .
\]

Note that the components containing \( R_1 \) in \( q(R_1, R_2, W, X_1) \) are:

\[
\frac{R_1 - p^1_0(W)}{p^1_0(W)(1 - p^1_0(W))} \hat{p}^1_0(W) + R_1 \frac{R_2 - \hat{p}^{11}_0(W, X_1)}{p^1_0(W, X_1)(1 - p^1_0(W, X_1))} \hat{p}^{11}_0(W, X_1).
\]

\( p(W) \) is defined as \( \Pr(R_1 = 1|W) \) and \( \hat{p}^{11}_0(W, X_1) \) denotes \( \Pr(R_2 = 1|W, X_1, R_1 = 1). \) Then

\[
E\left[ (1 - p_1) \frac{R_1}{p_1} (E[g_2(X_2, W; \beta)|X_1, W] - E[g_2(X_2, W; \beta)|W]) \frac{R_1 - p^1_0(W)}{p^1_0(W)(1 - p^1_0(W))} \hat{p}^1_0(W) \right] = 0
\]

taking expectations conditional on \( W \)

\[
E\left[ (1 - p_1) \frac{R_1}{p_1} (E[g_2(X_2, W; \beta)|X_1, W] - E[g_2(X_2, W; \beta)|W]) \frac{R_2 - \hat{p}^{11}_0(W, X_1)}{p^1_0(W, X_1)(1 - p^1_0(W, X_1))} \hat{p}^{11}_0(W, X_1) \right]
\]

Next, we calculate

\[
E\left[ g_{aipw}(s_\theta(W)' + R_1 R_2 s_\theta(X_1, X_2|W)' + R_1 (1 - R_2) s_\theta(X_1|W)' + (1 - R_1) R_2 s_\theta(X_2|W)') \right].
\]
For this part, we applied the original form of the AIPW augmentation.

\[
g_{\text{aipw}} = \frac{R_1R_2}{p_{11}} g(X, W; \beta) + \left(1 - \frac{R_1R_2}{p_{11}}\right) \left(1 - \frac{R_1R_2}{p_{11}}\right) E[g(X, W; \beta)|W] \\
+ \left(\frac{R_1}{p_1} - \frac{R_1R_2}{p_{11}}\right) (g_1(X_1, W; \beta) - E[g_1(X_1, W; \beta)|W]) \\
+ p_1 \cdot \left(\frac{R_1}{p_1} - \frac{R_1R_2}{p_{11}}\right) (E[g_2(X_2, W; \beta)|X_1, W] - E[g_2(X_2, W; \beta)|W]) \\
+ (1 - p_1) \cdot \left(\frac{(1 - R_1)R_2}{p_{01}} - \frac{R_1R_2}{p_{11}}\right) (g_2(X_2, W; \beta) - E[g_2(X_2, W; \beta)|W])
\]

We proceed term by term.

\[
E\left[(1) \left(s_\theta(W)' + R_1R_2s_\theta(X_1, X_2|W)' + R_1(1 - R_2)s_\theta(X_1|W)' + (1 - R_1)R_2s_\theta(X_2|W)'ight)\right] \\
= E\left[g(X, W; \beta) \left(s_\theta(W)' + s_\theta(X_1, X_2|W)'ight)\right] \\
= E\left[g(X, W; \beta) \left(s_\theta(W)' + s_\theta(X_1|W)' + s_\theta(X_2|X_1, W)'ight)\right].
\]

\[
E\left[(2) \left(s_\theta(W)' + R_1R_2s_\theta(X_1, X_2|W)' + R_1(1 - R_2)s_\theta(X_1|W)' + (1 - R_1)R_2s_\theta(X_2|W)'ight)\right] = 0.
\]

This equality holds directly from that \(E\left[1 - \frac{R_1R_2}{p_{11}}|W, X\right] = 0\).

\[
E\left[(3) \left(s_\theta(W)' + R_1R_2s_\theta(X_1, X_2|W)' + R_1(1 - R_2)s_\theta(X_1|W)' + (1 - R_1)R_2s_\theta(X_2|W)'ight)\right] \\
= E\left[(g_1(X_1, W; \beta) - E[g_1(X_1, W; \beta)|W]) \left(\frac{p_{11}}{p_1} - 1\right) s_\theta(X_1, X_2|W)' + \frac{p_{10}}{p_1} s_\theta(X_1|W)'ight] \\
= E\left[\frac{p_{10}}{p_1} (g_1(X_1, W; \beta) - E[g_1(X_1, W; \beta)|W]) \left(s_\theta(X_1|W)' - s_\theta(X_1, X_2|W)'ight)\right] \\
= -E\left[\frac{p_{10}}{p_1} (g_1(X_1, W; \beta) - E[g_1(X_1, W; \beta)|W]) s_\theta(X_2|W, X_1)'ight] = 0.
\]

The last equality holds from \(s_\theta(X_2|W, X_1) \in L^2_0(F(X|W, X_1))\). Similarly,

\[
E\left[(4) \left(s_\theta(W)' + R_1R_2s_\theta(X_1, X_2|W)' + R_1(1 - R_2)s_\theta(X_1|W)' + (1 - R_1)R_2s_\theta(X_2|W)'ight)\right] = 0.
\]
Lastly,  
\[
E \left[ (5) \left( s_\theta(W)' + R_1 R_2 s_\theta(X_1, X_2 W)' + R_1 (1 - R_2)s_\theta(X_1|W)' + (1 - R_1) R_2 s_\theta(X_2|W)'ight) \right] \\
= E \left[ (1 - p_1) (g_2(X_2, W; \beta) - E [g_2(X_2, W; \beta)|W]) (s_\theta(X_2|W)' - s_\theta(X_1, X_2|W)') \right] \\
= E \left[ (1 - p_1) (g_2(X_2, W; \beta) - E [g_2(X_2, W; \beta)|W]) s_\theta(X_1|W, X_2) \right] = 0
\]

following \( s_\theta(X_1|W, X_2) \in L^2_0 (F (X_1|W, X_2)) \).

By the above argument, we verify that any regular estimator for \( \beta^0 \) is asymptotic linear with the influence function \( -(AG)^{-1} A_{aipw} (W, R_1 X_1, R_2 X_2, R_1, R_2; \beta^0) \).

**Step 3** For any given weighting matrix \( A \), the variance of \( -(AG)^{-1} A_{aipw} (W, R_1 X_1, R_2 X_2, R_1, R_2; \beta^0) \) is \( (AG)^{-1} AVA' ( (AG)^{-1} )' \), where the variance \( V \) is as stated in Proposition 4.1. Optimal \( A^* \) is chosen at \( A^* = G' V^{-1} \) to minimize the variance, and the efficiency bound is reached by influence function \( -(A^* G)^{-1} A_{aipw} (W, R_1 X_1, R_2 X_2, R_1, R_2; \beta^0) \) and is equal to \( \Omega = (G' V G)^{-1} \).

\( \square \)

### A.5 Proof of Theorem 4.3

The sketch of the proof follows Cattaneo [2010], Chaudhuri and Guilkey [2016] closely, and most of the results have been proved in the proof of Proposition 2.3 in Chaudhuri and Guilkey [2016] and theorem 8 in Cattaneo [2010]. These results are proofed based on Newey [1994], Newey [1997].

We first apply \( p \) to stand for the elements in the missing mechanism set \( \{p_1, p_{11}, p_{01}\} \), and \( q \) denote the set of imputed values \( \{E [g_1|W], E [g_2|W], E [g_2|X_1, W]\} \). First, by the results showed in Theorem 5 and Theorem 8 in Cattaneo [2010], under the conditions listed in Theorem 4.3, \( \forall p, q \),

\[
\| \hat{p} - p \|_\infty = o_p(N^{-\frac{1}{2}}) \quad (A.2) \\
\| \hat{q} - q \|_\infty = o_p(N^{-\frac{1}{2}}). \quad (A.3)
\]

What left is to show the multiplication of the nuisance parameters converges to zero in probability. We refer to the decomposition in equation A.1 and first show the sketch of proof for term (3)

\[
\left( \frac{R_1}{p_1} - \frac{R_1 R_2}{p_{11}} \right) (g_1(X_1, W; \beta) - E [g_1(X_1, W; \beta)|W])
\]

and we use analogous conditions to the conditions (32) and (33) from the proof of proposition
under either the assumption SMAR, therefore, the condition A.5 reduces to:

\[
o_p(1) = \sqrt{N} \left( \xi_N(\beta, \hat{p}, \hat{q}(\beta)) - E[\xi_N(\beta, p, q(\beta))] \right)
\]

(A.4)

\[
o_p(1) = \sup_{|\beta - \beta_0| \leq \delta_N} \frac{\sqrt{N} |\xi_N(\beta, \hat{p}, \hat{q}(\beta)) - E[\xi_N(\beta, p, q(\beta))] - \xi_N(\beta, \hat{p}, \hat{q}(\beta))|}{1 + C\sqrt{N}|\beta - \beta_0|}
\]

(A.5)

for all positive sequences \(\delta_N = o(1)\) and a generic constant \(C > 0\), and

\[
\tilde{\xi}_N(\beta, \hat{p}, \hat{q}) = \frac{1}{N} \sum_{i=1}^{N} \hat{p}_i \hat{r}_i
\]

\[
\tilde{\xi}_N(\beta, \hat{p}, \hat{q}) = \frac{1}{N} \sum_{i=1}^{N} \nu_i \tau_i
\]

where

\[
\nu_i = \left( \frac{R_{1,i}}{p_{1,i}} - \frac{R_{1,i}R_{2,i}}{p_{11,i}} \right),
\]

\[
\tau_i = (g_1(X_{1,i}, W_i; \beta) - E[g_1(X_{1,i}, W_i; \beta)|W]).
\]

Condition A.4 holds follows the fact that \(\frac{1}{N} \sum_{i=1}^{N} |\nu_i|\) and \(\frac{1}{N} \sum_{i=1}^{N} |\tau_i|\) are \(O_p(1)\), which follows the conditions A.2 and A.3. The proof of the condition A.5 also follows the proof of the analogous condition in Chaudhuri and Guilkey [2016]. First, \(E[\xi_N(\beta, p, q(\beta))] = 0\) under either the assumption SMAR, therefore, the condition A.5 reduces to:

\[
o_p(1) = \sup_{|\beta - \beta_0| \leq \delta_N} \frac{\sqrt{N} |\tilde{\xi}_N(\beta, \hat{p}, \hat{q}(\beta)) - \tilde{\xi}_N(\beta, \hat{p}, \hat{q}(\beta))|}{1 + C\sqrt{N}|\beta - \beta_0|}
\]

\[
= \sup_{|\beta - \beta_0| \leq \delta_N} \frac{|\frac{1}{N} \sum_{i=1}^{N} \hat{p}_i (\hat{r}_i(\beta) - \hat{r}_i(\beta_0))|}{1 + C\sqrt{N}|\beta - \beta_0|}.
\]

(A.6)

Proof of equation A.6 can be found in proof of Proposition 2.3 and Proposition 2.4 in Chaudhuri and Guilkey [2016], by setting \(\omega_i = c\) for some constant \(c\) in their corresponding condition. The key condition used in the proof is that \(E[\tau_i(\beta)] = 0\) for any \(\beta\), and it holds from the definition of \(\tau_i(\beta)\).

The analogous proof sketch can be applied on the remaining terms corresponding to the term (4) and (5) imputing the case with \(R_2 = 1\), i.e.,

\[
p_1 \cdot \left( \frac{R_1}{p_1} - \frac{R_1R_2}{p_{11}} \right) (E[g_2(X_2, W; \beta)|X_1, W] - E[g_2(X_2, W; \beta)|W]) \quad \text{and}
\]

\[
(1 - p_1) \cdot \left( \frac{(1 - R_1)R_2}{p_{01}} - \frac{R_1R_2}{p_{11}} \right) (g_2(X_2, W; \beta) - E[g_2(X_2, W; \beta)|W])
\]
Next, we combine the term (1) and (2) into:

\[
\frac{R_1 R_2}{p_{11}} (g - E[g(X, W; \beta)|W]) + E[g(X, W; \beta)|W].
\] (A.7)

The proof of convergence of equation A.7 can be found in Theorem 8 in Cattaneo [2010]. □

References

Jason Abrevaya. Missing dependent variables in fixed-effects models. *Journal of econometrics*, 211(1):151–165, 2019.

Jason Abrevaya and Stephen G Donald. A gmm approach for dealing with missing data on regressors. *Review of Economics and Statistics*, 99(4):657–662, 2017.

Rotnitzky Andrea, Daniel Scharfstein, Ting-Li Su, and James Robins. Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics*, 57(1):103–113, 2001.

Joshua D Angrist and Guido W Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association*, 90(430):431–442, 1995.

Jorge F Balat and Sukjin Han. Multiple treatments with strategic interaction. *Available at SSRN 3182766*, 2018.

Jean-Louis Barnwell and Saraswata Chaudhuri. Efficient estimation in sub and full populations with monotonically missing at random data. Technical report, Technical report, McGill University, 2018.

Christopher F Baum, Mark E Schaffer, and Steven Stillman. Instrumental variables and gmm: Estimation and testing. *The Stata Journal*, 3(1):1–31, 2003.

Christoph Breunig and Peter Haan. Nonparametric regression with selectively missing covariates. *arXiv preprint arXiv:1810.00411*, 2018.

James R Carpenter, Michael G Kenward, and Stijn Vansteelandt. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):571–584, 2006.
Matias D Cattaneo. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155(2):138–154, 2010. 4.2, 4.2, A.5, A.5

Saraswata Chaudhuri. On efficiency gains from multiple incomplete subsamples. *Econometric Theory*, 36(3):488–525, 2020. 1, 2.2, 3.1

Saraswata Chaudhuri and David K Guilkey. Gmm with multiple missing variables. *Journal of Applied Econometrics*, 31(4):678–706, 2016. 1, 1, 3, 4.2, 4.2, A.4, A.5, A.5, A.5, A.5

Xiaohong Chen, Han Hong, Alessandro Tarozzi, et al. Semiparametric efficiency in gmm models with auxiliary data. *The Annals of Statistics*, 36(2):808–843, 2008. 3, 4.2

Philip E Cheng. Nonparametric estimation of mean functionals with data missing at random. *Journal of the American statistical association*, 89(425):81–87, 1994.

Federico Ciliberto and Elie Tamer. Market structure and multiple equilibria in airline markets. *Econometrica*, 77(6):1791–1828, 2009.

Amy Finkelstein, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group. The oregon health insurance experiment: evidence from the first year. *The Quarterly journal of economics*, 127(3):1057–1106, 2012. 6

Ahmed M Gad. A selection model for longitudinal data with non-ignorable non-monotone missing values. *Journal of Data Science*, 9(171-180), 2011.

Adam N Glynn and Kevin M Quinn. An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, pages 36–56, 2010. 3

Rolf HH Groenwold, A Rogier T Donders, Kit CB Roes, Frank E Harrell Jr, and Karel GM Moons. Dealing with missing outcome data in randomized trials and observational studies. *American journal of epidemiology*, 175(3):210–217, 2012.

Joel L Horowitz and Charles F Manski. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American statistical Association*, 95(449):77–84, 2000.

Roderick JA Little. Regression with missing x’s: a review. *Journal of the American statistical association*, 87(420):1227–1237, 1992.

Roderick JA Little and Donald B Rubin. Bayes and multiple imputation. *Statistical analysis with missing data*, pages 200–220, 2002. 1, 2.1
KW Newey and Daniel McFadden. Large sample estimation and hypothesis. *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, pages 2112–2245, 1994.

Whitney K Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382, 1994. 4.1, 4.2, A.5

Whitney K Newey. Convergence rates and asymptotic normality for series estimators. *Journal of econometrics*, 79(1):147–168, 1997. A.5

Whitney K Newey. Nonparametric instrumental variables estimation. *American Economic Review*, 103(3):550–56, 2013.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Li Qi and Yanqing Sun. Missing data approaches for probability regression models with missing outcomes with applications. *Journal of statistical distributions and applications*, 1(1):23, 2014. 1

James M Robins. Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in medicine*, 16(1):21–37, 1997. 1, 2.1, 3

James M Robins and Richard D Gill. Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in medicine*, 16(1):39–56, 1997. 1, 2.1

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994. 3

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. 3

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

Shaun Seaman, John Galati, Dan Jackson, and John Carlin. What is meant by” missing at random”? *Statistical Science*, pages 257–268, 2013.

Shaun R Seaman and Stijn Vansteelandt. Introduction to double robust methods for incomplete data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 33(2):184, 2018.

44
Shaun R Seaman and Ian R White. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295, 2013. 3

BaoLuo Sun and Eric J Tchetgen Tchetgen. On inverse probability weighting for nonmonotone missing at random data. *Journal of the American Statistical Association*, 113(521):369–379, 2018. 1

BaoLuo Sun, Neil J Perkins, Stephen R Cole, Ofer Harel, Emily M Mitchell, Enrique F Schisterman, and Eric J Tchetgen Tchetgen. Inverse-probability-weighted estimation for monotone and nonmonotone missing data. *American journal of epidemiology*, 187(3):585–591, 2018.

Elie Tamer. Incomplete simultaneous discrete response model with multiple equilibria. *The Review of Economic Studies*, 70(1):147–165, 2003.

Eric J Tchetgen Tchetgen, Linbo Wang, and BaoLuo Sun. Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. *Statistica Sinica*, 28(4):2069–2088, 2018. 2.1

Henri Theil. Principles of econometrics. Technical report, 1971.

Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007. 2.1, 3

Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2018. 1

Lu Wang, Andrea Rotnitzky, and Xihong Lin. Nonparametric regression with missing outcomes using weighted kernel estimating equations. *Journal of the American Statistical Association*, 105(491):1135–1146, 2010.

Jeffrey M Wooldridge. Inverse probability weighted estimation for general missing data problems. *Journal of econometrics*, 141(2):1281–1301, 2007. 3