Institute for Economic Studies, Keio University

Keio-IES Discussion Paper Series

偏りのあるサンプリングにおける母集団モーメントと母数の二重にロバストな推定法の提案

星野　崇宏、清水　祐弥

2019 年 2 月 11 日
DP2019-006
https://ies.keio.ac.jp/publications/10961/

Institute for Economic Studies, Keio University
2-15-45 Mita, Minato-ku, Tokyo 108-8345, Japan
ies-office@adst.keio.ac.jp
11 February, 2019

偏りのあるサンプリングにおける母集団モーメントと母数の二重にロバストな推定法の提案

星野崇宏、清水祐弥

IES Keio DP2019-006
2019 年 2 月 11 日
JEL Classification: C13, C18, C83
キーワード: 外部情報; 偏りのあるサンプリング; 欠測データ; 傾向スコア; 二重にロバストな推定量

【要旨】
一般的に考えられているbiased samplingの欠測では、関心のある変数のみが欠測しているが、本研究はデータの一部が共変量も含めたユニット単位で欠測するbiased samplingデータを扱う。このデータから母集団モーメントやパラメータを推定するための二重にロバストな推定量を考える。バイアスの補正には、共変量の分布やそのモーメントといった外部情報を利用する。

星野　崇宏
慶應義塾大学経済学部
〒108-8345
東京都港区三田2-15-45
hoshino@econ.keio.ac.jp

清水　祐弥
慶應義塾大学大学院経済学研究科
〒108-8345
東京都港区三田2-15-45
yuya_shimizu@z7.keio.jp

謝辞：本研究は日本学術振興会科学研究費補助金（18H03209）による研究成果である。ここに記して謝意を表したい。

Doubly Robust-type Estimation of Population Moments and Parameters in Biased Sampling

Yuya Shimizu [†] Takahiro Hoshino [‡]

February 11, 2019

Abstract

We propose an estimation method of population moments or population parameters in "biased sampling data" in which for some units of data, not only the variable of interest but also the covariates, have missing observations and the proportion of "missingness" is unknown. We use auxiliary information such as the distribution of covariates or their moments in random sampling data in order to correct the bias. Moreover, with additional assumptions, we can correct the bias even if we have only the moment information of covariates. The main contribution of this paper is the development of a doubly robust-type estimator for biased sampling data. This method provides a consistent estimator if either the regression function or the assignment mechanism is correctly specified. We prove the consistency and semi-parametric efficiency of the doubly robust estimator. Both the simulation and empirical application results demonstrate that the proposed estimation method is more robust than existing methods.

Key words: Auxiliary information, Biased sampling, Missing data, Propensity score, Doubly robust estimator

---

This research was supported by JSPS KAKENHI Grant Number 18H03209.

[†]Keio University, Graduate School of Economics
[‡]Keio University, Faculty of Economics / RIKEN AIP

1 INTRODUCTION

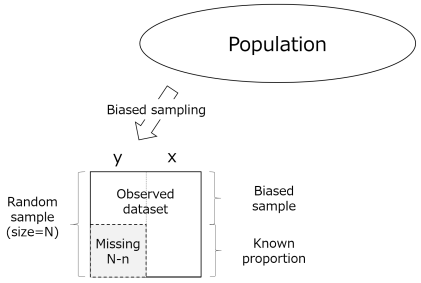

Figure 1: Missing response problem

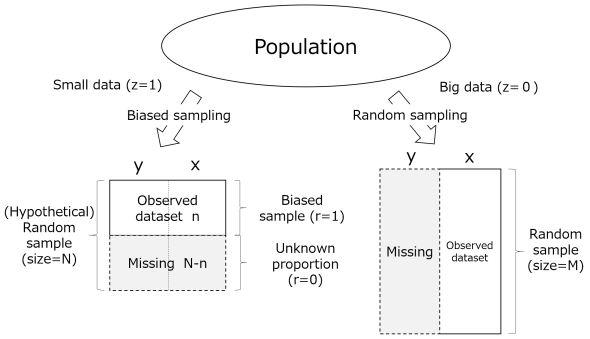

Figure 2: Biased sampling problem considered in this papaer

We propose an estimation method of population moments or population parameters in biased sampling data in which for some units of data, not only the variable of interest but also the covariates, have missing observations and the proportion of missingness is unknown. We use auxiliary information such as the distribution of covariates or their moments in random sampling data or moment information of them in order to correct the bias. In the analysis of survey sampling data, there is often sampling bias due to missing observations.

Let $\mathbf{y}$ be variables of interest in their relevant parameters or their population moments $\boldsymbol{\theta}$ and $\mathbf{x}$ be covariates. Figure 1 shows the missing response problem in which the researcher can correct bias by using the method for selection bias such as Heckman's probit selection model (Heckman, 1974, 1979), inversed probability weighting (Rubin, 1985), partial linear model (Robinson, 1988; Speckman, 1988), or the empirical likelihood approach (Qin, 1993; Qin *et al.*, 2002), when the missingness depends on the covariates. However, in order to use those methods, the researcher has to specify one model and those methods never provide a consistent estimator if the model is misspecified. To deal with this problem, the researcher can use the doubly robust estimator method, which has consistency even if either a model for the missing mechanism or a model for the distribution of the covariates is correctly specified (Bang and Robins, 2005). There are numerous studies about this type estimator (Robins *et al.*, 1994; Lipsitz *et al.*, 1999; Hoshino, 2007; Wang *et al.*, 2010). Qin *et al.* (2008) studied the empirical likelihood approach, which can conduct the doubly robust-type imputation for missing data. In this setting, the researcher has to assume no missingness in covariates $\mathbf{x}$. Other studies dealt with the missing covariates problem (see e.g., Liang *et al.*, 2004, JASA, 357-367; Ibrahim *et al.*, 1999, JRSSB, 173-190) in which there is no missingness in variables of interest $\mathbf{y}$.

In this paper, however, we consider biased sampling datasets as Figure 2, in which biased samples of n units are only observed. The observed biased sample consists of n units, but the total hypothetical sample size of the random sample is unknown ($= N$). In other words, the resulting biased sample with sample size n can be considered a subsample of the random sample with unknown sample size N (see e.g., Lee and Berger, 2001, JASA, 1397-1409; Qin, 2017). The setting here would be more natural than the setting of Figure 1 because, for applied research, the dataset contains the variables of concern, most of which are measured but obtained under non-random sampling, whereas the auxiliary population information of a part of variables is often available such as public statistics or databases. For example, in medicine, all receipt data are gathered by the government but it does not contain treatment results ($= \mathbf{y}$). We call these data "Big Data". Let z be an indicator for biased small data ($z = 1$) or big data ($z = 0$) and r be an indicator that indicates whether the unit is observed ($r = 1$) or missing ($r = 0$) in the small data. A similar setting could be seen in the research of positive and unlabeled data in the field of machine learning (Elkan and Noto, 2008) and presence-only data in biostatistics (Ward

et al., 2009); however, $\mathbf{y}$ is restricted to be binary in these studies. We propose a consistent doubly robust estimator for this set of bias, that can obtain a consistent estimator from biased sampling data even if one of the two models is erroneously set.

Hirano *et al.* (2001) and Nevo (2003) use auxiliary information in order to correct the bias of missing data. Both studies considered panel data examples with attrition by non-ignorable selection mechanism. Hirano *et al.* (2001) dealt with a two-period panel dataset. They proposed an estimation method that combined panel datasets with the auxiliary information of refreshment samples and proved the identifiability of their method. Nevo (2003) used the identification result of Hirano *et al.* (2001) and proposed a GMM-type estimator. In our setup, shown in Figure 2, although we can conduct inversed probability weighting using the propensity score, which can be estimated by the method of Nevo (2003), there is no previous research that has developed a doubly robust estimator.

The greatest novelty of the paper is that we consider the doubly robust estimator for missing units that occur not only in the dependent variables but also in the covariates.

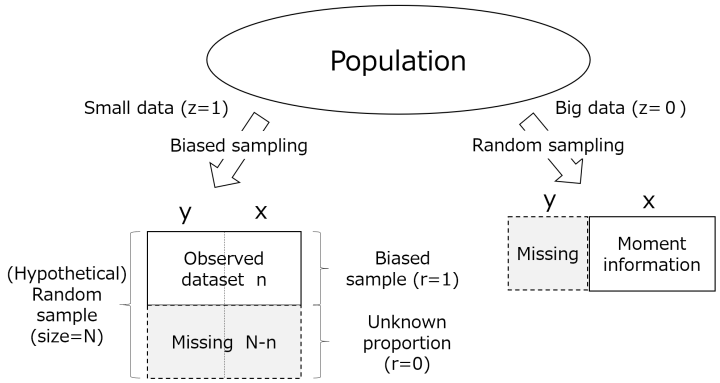

Figure 3: The case in which auxiliary information is moment information

Moreover, our method can be extended to the situation represented by Figure 3. The difference between this situation and the setting in Figure 2 is that available data are not unit-level datasets $\{\mathbf{x}_m\}_{m=1}^{M}$, but moment information or statistics such as averages or proportions. The government can extract random samples from the citizens or run censuses, and companies also have all their customers' transaction data; however, because unit-level (eg., individual-level) data usually contain sensitive private information, the government or companies may not provide full individual-level data but only moment information like the sample mean. Our method can correct sampling bias even if the researcher has only moment information with additional assumptions.

This paper is organized as follows. Section 2 introduces our basic setup and doubly robust estimator. Section 3 presents the proof of consistency and semi-parametric efficiency of the doubly robust estimator in our setup. Section 4 reports the simulation results and Section 5 provides an empirical application. Section 6 extends our method to the case where the auxiliary information is only moment information.

2 DOUBLY ROBUST ESTIMATOR

In this section, we explain the doubly robust estimator under the sampling bias. The bias can be corrected by (1) weighting by the inverse of the propensity score and (2) the marginal estimating equation using the regression model. With the doubly robust estimator, a consistent estimator can be obtained from biased data even if one of the two models is erroneously set.

2.1 Estimating Equation

Let $\mathbf{y} \in \mathbb{R}^p$ be variables of interest, $\mathbf{x} \in \mathbb{R}^s$ be covariates, $\boldsymbol{\theta} \in \mathbb{R}^q$ be population moments or population parameters, and $\psi(\mathbf{y}|\boldsymbol{\theta}) \in \mathbb{R}^q$ be an estimating function for the unbiased estimating equation such that the expectation about $\mathbf{y}$ will be zero with true parameter $\boldsymbol{\theta}_0$,

$$\psi(\mathbf{y}|\boldsymbol{\theta}) \qquad s.t. \quad E[\psi(\mathbf{y}|\boldsymbol{\theta})_0] = \mathbf{0}. \tag{1}$$

For example, if $E[\mathbf{y}]$ is the moment of interest, then $\boldsymbol{\theta} = E[\mathbf{y}]$ and $\psi(\mathbf{y}|\boldsymbol{\theta}) = \mathbf{y} - \boldsymbol{\theta}$. In the case in which random N sampling data can be obtained, the solution of equation (1), $\hat{\boldsymbol{\theta}}$, is a consistent estimator for $\boldsymbol{\theta}$. In the setting of this research, however, the units with $z = 1$ and $r = 1$ can only be obtained; thus, the sample mean of $\psi(\mathbf{y}|\boldsymbol{\theta})$ will become the expectation with respect to $p(\mathbf{y}, \mathbf{x}|z = 1, r = 1)$ as n goes to infinity, $E(\psi(\mathbf{y}|\boldsymbol{\theta})|z = 1, r = 1) \neq 0$.

There are two models, the missing mechanism and the regression model, used to obtain the consistent estimator. Moreover, we can develop the doubly robust estimator by combining these two models.

In this paper, we assume following two conditions:

Assumption 1. *strong ignorability (Rosenbaum and Rubin, 1983)*
When $\mathbf{x}$ *is conditioned,* r *and* $\mathbf{y}$ *are independent;*

$$r \perp\!\!\!\perp \mathbf{y}|\mathbf{x}, \qquad 0 < Pr(r = 1|\mathbf{x}) < 1 \ for \ all \ \mathbf{x}. \tag{2}$$

In other words, the missing or not depends on the covariates, but not on the dependent variables: $p(r = 1|\mathbf{y}, \mathbf{x}) = p(r = 1|\mathbf{x})$.

Assumption 2. *Auxiliary random sampling data* $\{\mathbf{x}_i\}_{i=N+1}^{N+M}$ *of size* M *has been obtained* [1].

2.2 Model 1: Missing Mechanism

$E[\psi(\mathbf{y}|\boldsymbol{\theta})]$ can be transformed as follows:

$$\begin{aligned} E[\psi(\mathbf{y}|\boldsymbol{\theta})] &= \int \psi(\mathbf{y}|\boldsymbol{\theta}) p(\mathbf{y}, \mathbf{x}) d\mathbf{x} d\mathbf{y} = \int \psi(\mathbf{y}|\boldsymbol{\theta}) \frac{p(\mathbf{y}, \mathbf{x}|z = 1, r = 1) p(z = 1, r = 1)}{p(z = 1, r = 1|\mathbf{y}, \mathbf{x})} d\mathbf{x} d\mathbf{y} \\ &= \int \psi(\mathbf{y}|\boldsymbol{\theta}) \frac{p(\mathbf{y}, \mathbf{x}|z = 1, r = 1) p(z = 1, r = 1)}{p(z = 1, r = 1|\mathbf{x})} d\mathbf{x} d\mathbf{y} = \mathbf{0}. \end{aligned}$$

Therefore, by using the method of moment estimator,

$$\frac{1}{N} \sum_{i=1}^{N} \frac{z_i r_i p(z = 1, r = 1)}{p(z = 1, r = 1|\mathbf{x}_i)} \psi(\mathbf{y}_i|\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{n} \frac{p(z = 1, r = 1)}{p(z = 1, r = 1|\mathbf{x}_i)} \psi(\mathbf{y}_i|\boldsymbol{\theta}) = \mathbf{0}. \tag{3}$$

Here, $p(z = 1, r = 1)$ is constant; then, both sides of equation (3) can be divided by it:

$$\sum_{i=1}^{N} \frac{z_i r_i}{p(z = 1, r = 1|\mathbf{x}_i)} \psi(\mathbf{y}_i|\boldsymbol{\theta}) = \mathbf{0}. \tag{4}$$

The solution of equation (4), $\hat{\boldsymbol{\theta}}$, is a consistent estimator for $\boldsymbol{\theta}_0$.

Here, N and $N - n$ are unknown, thus, it is not possible to estimate the propensity score $p(z = 1, r = 1|\mathbf{x})$ with the $(z = 1)$ dataset directly. We use the method of Nevo (2003) in order to estimate it. Let $\bar{\mathbf{h}}(\mathbf{x})$ be a J-dimensional function. Nevo (2003) defined $\mathbf{h}(\mathbf{x}) = \bar{\mathbf{h}}(\mathbf{x}) - E[\bar{\mathbf{h}}(\mathbf{x})]$ and noted that $E[\mathbf{h}(\mathbf{x})] = \mathbf{0}$. For example, if the researcher knows the first moment of covariates, then, $\bar{\mathbf{h}}(\mathbf{x}) = \mathbf{x}$ and $\mathbf{h}(\mathbf{x}) = \mathbf{x} - E[\mathbf{x}]$. In addition, $E[\bar{\mathbf{h}}(\mathbf{x})]$ can be replaced with the sample mean of random sampling data $\frac{1}{M} \sum_{m=1}^{M} \bar{\mathbf{h}}(\mathbf{x}_m)$.

[1] We also denote this auxiliary random sampling data as $\{\mathbf{x}_m\}_{m=1}^{M}$ if necessary.

Let the logit of the propensity score be a linear equation about $\mathbf{h}(\mathbf{x})$:

$$logit[p(r=1|\mathbf{x})] = \mathbf{h}(\mathbf{x})^t\boldsymbol{\alpha}, \tag{5}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^J$ are the parameters for logit [2].

Then, with the biased data x_i, by solving simultaneous equations:

$$\begin{cases} \sum_{i=1}^{n} \dfrac{1}{p(r=1|\mathbf{x}_i,\boldsymbol{\alpha})} = \dfrac{n}{\hat{p}(r=1)} & (6) \\ \sum_{i=1}^{n} \dfrac{\mathbf{h}(\mathbf{x}_i)}{p(r=1|\mathbf{x}_i,\boldsymbol{\alpha})} = 0 \quad , & (7) \end{cases}$$

$\hat{\boldsymbol{\alpha}}$ can be estimated and the propensity score using the estimated value can be calculated, where, $\hat{p}(r=1) = \frac{1}{M}\sum_{m=1}^{M} p(r=1|\mathbf{x}_m,\boldsymbol{\alpha})$. Nevo (2003) proposed this GMM method when the sample is not random and auxiliary information is obtained, and proved that $\hat{\boldsymbol{\alpha}}$ is consistent and has asymptotic normality. Equation (7) means that $\boldsymbol{\alpha}$ is estimated as to satisfy the condition that the weighted sample analog of $E[\mathbf{h}(\mathbf{x}_i)]$ equals to 0. We need equation (6) in order to make the sample mean of weight equal to its population value $\frac{1}{p(r=1)}$.

We calculate the inverse probability weighting (IPW) estimator using weighting by the inverse of the estimated propensity score $p(r=1|\mathbf{x}_i,\hat{\boldsymbol{\alpha}})$.

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{p(r=1|\mathbf{x}_i,\hat{\boldsymbol{\alpha}})}\psi(\mathbf{y}_i|\boldsymbol{\theta}) = \mathbf{0}. \tag{8}$$

The solution $\hat{\boldsymbol{\theta}}$ is a consistent estimator for the parameter $\boldsymbol{\theta}_0$ only when the missing mechanism is correctly set.

2.3 Model 2: Regression Model

From the assumption that $\mathbf{y}$ and r are independent when $\mathbf{x}$ is conditioned, $p(\mathbf{y}|r=1,\mathbf{x},\boldsymbol{\beta}) = p(\mathbf{y}|\mathbf{x},\boldsymbol{\beta})$. $\boldsymbol{\beta} \in \mathbb{R}^s$ is the parameter vector.

Therefore, the regression model estimated from biased data coincides with that estimated from random sampling data.

Monte Carlo integration with the estimated regression model can be conducted.

$$\frac{1}{M}\sum_{m=1}^{M}\sum_{l=1}^{L} \psi(\mathbf{y}_{ml}|\boldsymbol{\theta}) = \mathbf{0}, \tag{9}$$

where, $\mathbf{y}_{ml} \sim p(\mathbf{y}|r=1,\mathbf{x}_m,\hat{\boldsymbol{\beta}})$ $l=1,\ldots,L$, $m=1,\ldots,M$ and $\hat{\boldsymbol{\beta}}$ is a consistent estimator.

The solution $\hat{\boldsymbol{\theta}}$ is a consistent estimator for the parameter $\boldsymbol{\theta}_0$ only when the regression model is correctly set.

2.4 Doubly Robust Estimator

If at least one of the two models is set correctly, then the solution $\hat{\boldsymbol{\theta}}$ of the following equation (10) will be a consistent estimator for the parameter $\boldsymbol{\theta}_0$:

$$\begin{aligned} \frac{1}{N+M}\sum_{i=1}^{N+M}\Big\{ \frac{z_i r_i}{p(r=1|\mathbf{x}_i,\boldsymbol{\alpha})}\psi(\mathbf{y}_i|\boldsymbol{\theta}) + z_i\left(1-\frac{r_i}{p(r=1|\mathbf{x}_i,\boldsymbol{\alpha})}\right)E_{\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\beta}}[\psi(\mathbf{y}_i|\boldsymbol{\theta})] \\ + (1-z_i)E_{\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\beta}}[\psi(\mathbf{y}_i|\boldsymbol{\theta})]\Big\} = \mathbf{0}, \end{aligned} \tag{10}$$

[2] In the case in which $\mathbf{x}$ contains a constant as a first element, $\mathbf{h}(\mathbf{x})$ may contain the element, $constant - E[constant] = 0$. In order to contain a constant in $\mathbf{h}(\mathbf{x})$, this zero element can be replaced by 1.

where, $E_{\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\beta}}[\psi(\mathbf{y}_i|\boldsymbol{\theta})]$ is a conditional expectation of $\psi(\mathbf{y}_i|\boldsymbol{\theta})$ over a probability distribution function $p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\beta})$[3].

By the definition of indicator, $z_i r_i = 1$ if and only if the unit i is an observed unit in biased sampling dataset. Equation (10) can be transformed as follows:

$$\frac{1}{N+M}\sum_{i=1}^{n}\left[\frac{1}{p(r=1|\mathbf{x}_i,\boldsymbol{\alpha})}\left\{\psi(\mathbf{y}_i|\boldsymbol{\theta})-E_{\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\beta}}[\psi(\mathbf{y}_i|\boldsymbol{\theta})]\right\}\right] + \frac{1}{N+M}\sum_{i=1}^{N+M}E_{\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\beta}}[\psi(\mathbf{y}_i|\boldsymbol{\theta})]=\mathbf{0}.$$

We need to estimate $N, p(r=1|\mathbf{x}_i,\boldsymbol{\alpha})$, and $p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\beta})$; therefore, replace these by estimated value $\hat{N}$ and probability density functions $p(r=1|\mathbf{x}_i,\hat{\boldsymbol{\alpha}})$, $p(\mathbf{y}_i|\mathbf{x}_i,\hat{\boldsymbol{\beta}})$. The meaning of $p(r=1)$ is the proportion of observed units on a hypothetical random sample $p(r=1)=\frac{n}{N}$; hence, $\hat{N}=\frac{n}{\hat{p}(r=1)}=\frac{n}{\frac{1}{M}\sum_{m=1}^{M}p(r=1|\mathbf{x}_m,\hat{\boldsymbol{\alpha}})}$. Because N is unknown, the second term has to be replaced with $\frac{1}{M}\sum_{m=1}^{M}E_{\mathbf{y}_m|\mathbf{x}_m,\hat{\boldsymbol{\alpha}}}[\psi(\mathbf{y}_m|\boldsymbol{\theta})]$. After replacing, the estimating equation will be:

$$\frac{1}{\hat{N}+M}\sum_{i=1}^{n}\left[\frac{1}{p(r=1|\mathbf{x}_i,\hat{\boldsymbol{\alpha}})}\left\{\psi(\mathbf{y}_i|\boldsymbol{\theta})-E_{\mathbf{y}_i|\mathbf{x}_i,\hat{\boldsymbol{\beta}}}[\psi(\mathbf{y}_i|\boldsymbol{\theta})]\right\}\right] + \frac{1}{M}\sum_{m=1}^{M}E_{\mathbf{y}_m|\mathbf{x}_m,\hat{\boldsymbol{\beta}}}[\psi(\mathbf{y}_m|\boldsymbol{\theta})]=\mathbf{0}. \tag{11}$$

2.5 Monte Carlo Integration

The doubly robust estimator of estimate equation (11) contains the conditional expectations. In the case where $\psi(\mathbf{y}|\boldsymbol{\theta})$ is a continuous function, we have to solve the integral, which is often difficult to. If the integral cannot be solved analitically, then Monte Carlo integration can be used. Because the second term of equation (11) is about the random sample, it can be replaced with the Monte Carlo integral on random sampling dataset ($z=0$). On the other hand, because the first term is about the biased sample, the Monte Carlo integration should be done on biased sampling dataset ($z=1$).

$$\frac{1}{\hat{N}+M}\sum_{i=1}^{n}\left[\frac{1}{p(r=1|\mathbf{x}_i,\hat{\boldsymbol{\alpha}})}\left\{\psi(\mathbf{y}_i|\boldsymbol{\theta})-\frac{1}{L}\sum_{l=1}^{L}\psi(\mathbf{y}_{il}|\boldsymbol{\theta})\right\}\right]+\frac{1}{M}\sum_{m=1}^{M}\frac{1}{L}\sum_{l=1}^{L}\psi(\mathbf{y}_{ml}|\boldsymbol{\theta})=\mathbf{0}, \tag{12}$$

where, $\mathbf{y}_{il}\sim p(\mathbf{y}|\mathbf{x}_i,\hat{\boldsymbol{\beta}})$, $l=1,\ldots,L$, $i=1,\ldots,n$, and $\mathbf{y}_{ml}\sim p(\mathbf{y}|\mathbf{x}_m,\hat{\boldsymbol{\beta}})$, $l=1,\ldots,L, m=1,\ldots,M$.

3 ASYMPTOTIC PROPERTIES

In this section, we provide two theorems, their proofs, and the asymptotic distribution. First, we show that our estimator is consistent if one of two models is correctly specified. This proof verifies the estimator is a doubly robust-type. Second, we prove that our estimator is the most efficient when both models are correctly specified. We denote $p(r=1|\mathbf{x},\hat{\boldsymbol{\alpha}})=\pi$ and $E_{\mathbf{y}|\mathbf{x},\hat{\boldsymbol{\beta}}}[\psi(\mathbf{y}|\boldsymbol{\theta})]=E_{\mathbf{y}|\mathbf{x}}[\psi(\mathbf{y}|\boldsymbol{\theta})]$ in this section. Finally, the asymptotic distribution for our estimator is shown.

[3]In the case where the researcher knows $E_{\mathbf{y}|\mathbf{x}}[\psi(\mathbf{y}|\boldsymbol{\theta})]$, the regression model and consistent estimator for $\boldsymbol{\beta}$ are not necessary and the researcher can use $E_{\mathbf{y}|\mathbf{x}}[\psi(\mathbf{y}|\boldsymbol{\theta})]$ directly.

3.1 Proof of Consistency

Theorem 1. *If either the missing mechanism or the regression model is set correctly, then the doubly robust estimator has consistency.*

Proof. We prove this theorem by showing that the doubly robust estimator is M-estimator, which has consistency.

(i) When the propensity score $p(r=1|\mathbf{x},\hat{\boldsymbol{\alpha}})=\pi$ is correctly specified

In this case, regression model $p(\mathbf{y}|\mathbf{x},\boldsymbol{\beta})$ may be misspecified. This implies $E_{\mathbf{y}|\mathbf{x}}[\psi(\mathbf{y}|\boldsymbol{\theta})]$ may also be misspecified. We denote this expectation as $\tilde{E}_{\mathbf{y}|\mathbf{x}}[\psi(\mathbf{y}|\boldsymbol{\theta})]$.
As $M\to\infty, n\to\infty$ and $\frac{M}{n}\to k$ (*constant*),

$$
\begin{aligned}
&\frac{1}{N+M}\sum_{i=1}^{N+M}\left\{\frac{z_i r_i}{\pi_i}\psi(\mathbf{y}_i|\boldsymbol{\theta})+z_i\left(1-\frac{r_i}{\pi_i}\right)\tilde{E}_{\mathbf{y}_i|\mathbf{x}_i}[\psi(\mathbf{y}_i|\boldsymbol{\theta})]+(1-z_i)\tilde{E}_{\mathbf{y}_i|\mathbf{x}_i}[\psi(\mathbf{y}_i|\boldsymbol{\theta})]\right\}\\
=&\frac{1}{N+M}\sum_{i=1}^{N+M}\left\{\frac{r_i}{\pi_i}\psi(\mathbf{y}_i|\boldsymbol{\theta})+\frac{\pi_i-r_i}{\pi_i}\tilde{E}_{\mathbf{y}_i|\mathbf{x}_i}[\psi(\mathbf{y}_i|\boldsymbol{\theta})]\right\}\\
\to&E_{\mathbf{x}}\left[E_{r|\mathbf{x}}\left[\frac{r}{\pi}\right]E_{\mathbf{y}|\mathbf{x}}\left[\psi(\mathbf{y}|\boldsymbol{\theta})\right]+E_{r|\mathbf{x}}\left[\frac{r-\pi}{\pi}\right]\tilde{E}_{\mathbf{y}|\mathbf{x}}[\psi(\mathbf{y}|\boldsymbol{\theta})]\right]\\
=&E_{\mathbf{x}}\left[E_{\mathbf{y}|x}\left[\psi(\mathbf{y}|\boldsymbol{\theta})\right]\right]=0\quad\left(\because E_{r|\mathbf{x}}\left[\frac{r-\pi}{\pi}\right]=\frac{\pi-\pi}{\pi}\right)\\
=&E\left[\psi(\mathbf{y}|\boldsymbol{\theta})\right].
\end{aligned}
$$

Therefore, as $M\to\infty, n\to\infty$ and $\frac{M}{n}\to k$ (*constant*), $\hat{\boldsymbol{\theta}}\xrightarrow{\mathrm{p}}\boldsymbol{\theta}_0$.

(ii) When the regression model $p(\mathbf{y}|\mathbf{x},\boldsymbol{\beta})$ is correctly specified

In this case, missing mechanism $p(r=1|\mathbf{x},\boldsymbol{\alpha})=\pi$ may be misspecified. This implies that the estimated $\hat{N}$ may be incorrect. We denote these π and N as $\tilde{\pi}$ and $\tilde{N}$.
Equation (11) can be represented as follows:

$$
\frac{n}{\tilde{N}+M}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{r_i}{\tilde{\pi}}\left\{\psi(\mathbf{y}_i|\boldsymbol{\theta})-E_{\mathbf{y}_i|\mathbf{x}_i}[\psi(\mathbf{y}_i|\boldsymbol{\theta})]\right\}\right]+\frac{1}{M}\sum_{m=1}^{M}E_{\mathbf{y}_m|\mathbf{x}_m}[\psi(\mathbf{y}_m|\boldsymbol{\theta})]=\mathbf{0}. \tag{13}
$$

As $M\to\infty, n\to\infty$ and $\frac{M}{n}\to k$,

$\frac{n}{\tilde{N}+M}\to\frac{\hat{p}(r=1)}{1+k\hat{p}(r=1)}=C$ (*constant*) and,

$$
\begin{aligned}
&\frac{n}{\tilde{N}+M}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{r_i}{\tilde{\pi}}\left\{\psi(\mathbf{y}_i|\boldsymbol{\theta})-E_{\mathbf{y}_i|\mathbf{x}_i}[\psi(\mathbf{y}_i|\boldsymbol{\theta})]\right\}\right]+\frac{1}{M}\sum_{m=1}^{M}E_{\mathbf{y}_m|\mathbf{x}_m}[\psi(\mathbf{y}_m|\boldsymbol{\theta})]\\
\to&C\cdot E_{\mathbf{x}}\left[E_{r|\mathbf{x}}\left[\frac{r}{\tilde{\pi}}\right]E_{\mathbf{y}|\mathbf{x}}\left[\psi(\mathbf{y}|\boldsymbol{\theta})-\psi(\mathbf{y}|\boldsymbol{\theta})\right]\right]+E_{\mathbf{x}}\left[E_{\mathbf{y}|\mathbf{x}}[\psi(\mathbf{y}|\boldsymbol{\theta})]\right]\\
=&E_{\mathbf{x}}\left[E_{\mathbf{y}|\mathbf{x}}\left[\psi(\mathbf{y}|\boldsymbol{\theta})\right]\right]\\
=&E\left[\psi(\mathbf{y}|\boldsymbol{\theta})\right].
\end{aligned}
$$

Therefore, as $M\to\infty, n\to\infty$ and $\frac{M}{n}\to k$ (*constant*), $\hat{\boldsymbol{\theta}}\xrightarrow{\mathrm{p}}\boldsymbol{\theta}_0$.

The proposed estimator has consistency in both cases. □

3.2 Proof of Efficiency

Theorem 2. *When both the missing mechanism and the regression model are set correctly, the doubly robust estimator has minimum variance in specific models with an augment term (semi-parametric efficiency).*

Proof. The proof is similar to Robins *et al.* (1994). Consider augment term $A(\boldsymbol{\phi})$.

If all the data of z = 1 were observed, let $D^F(\boldsymbol{\theta}) = z\psi(\mathbf{y}|\boldsymbol{\theta})$,

$$\frac{1}{N+M}\sum_{i=1}^{N+M}\left\{D^F(\boldsymbol{\theta}) - (1-z_i)E_{\mathbf{y}_i|\mathbf{x}_i}[\psi(\mathbf{y}_i|\boldsymbol{\theta})]\right\} = \mathbf{0}, \tag{14}$$

then, the solution $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ will be a consistent estimator. However, there is missingness in this setting; therefore, we use $D(\boldsymbol{\theta},\boldsymbol{\phi}) = \frac{r}{\pi}D^F(\boldsymbol{\theta}) - A(\boldsymbol{\phi})$ and estimate $\boldsymbol{\theta}$ by solving

$$\frac{1}{N+M}\sum_{i=1}^{N+M}\left\{D_i(\boldsymbol{\theta},\boldsymbol{\phi}) - (1-z_i)E_{\mathbf{y}_i|\mathbf{x}_i}[\psi(\mathbf{y}_i|\boldsymbol{\theta})]\right\} = \mathbf{0}. \tag{15}$$

where, $\pi = p(r=1|\mathbf{x}), A(\boldsymbol{\phi}) = \frac{r-\pi}{\pi}\boldsymbol{\phi}$, and

$$D(\boldsymbol{\theta},\boldsymbol{\phi}) = D^F(\boldsymbol{\theta}) + \frac{r-\pi}{\pi}\left\{D^F(\boldsymbol{\theta}) - \boldsymbol{\phi}\right\}. \tag{16}$$

$D(\boldsymbol{\theta},\boldsymbol{\phi}) - (1-z)E_{\mathbf{y}|\mathbf{x}}[\psi(\mathbf{y}|\boldsymbol{\theta})]$ of equation (15) is equal to

$$D^F(\boldsymbol{\theta}) + (1-z)E_{\mathbf{y}|\mathbf{x}}[\psi(\mathbf{y}|\boldsymbol{\theta})] + \frac{r-\pi}{\pi}\left\{D^F(\boldsymbol{\theta}) - \boldsymbol{\phi}\right\}. \tag{17}$$

Because under the condition of $\mathbf{x}$, $\mathbf{y}$ and r are independent, the first and second terms do not correlate and the variance of equation (17) is

$$Var\left(D^F(\boldsymbol{\theta}) + (1-z)E_{\mathbf{y}|\mathbf{x}}[\psi(\mathbf{y}|\boldsymbol{\theta})]\right) + E_{\mathbf{x}}\left[\frac{1-\pi}{\pi}E_{\mathbf{y}|\mathbf{x}}\left[\left\{D^F(\boldsymbol{\theta}) - \boldsymbol{\phi}\right\}^{\otimes 2}\right]\right]^{4}. \tag{18}$$

Following $\boldsymbol{\phi}^*$ minimize (18) and according to Proposition2.2 of Robins *et al.* (1994), the $\boldsymbol{\phi}^*$ minimizing the variance of equation (17) minimizes the asymptotic variance of $\hat{\boldsymbol{\theta}}$,

$$\boldsymbol{\phi}^* = E_{\mathbf{y},z|\mathbf{x}}\left[D^F(\boldsymbol{\theta})\right] = zE_{\mathbf{y}|\mathbf{x}}\left[\psi(\mathbf{y}|\boldsymbol{\theta})\right]. \tag{19}$$

Then, equation (17) will be

$$\frac{zr}{\pi}\psi(\mathbf{y}|\boldsymbol{\theta}) + z\left(1-\frac{r}{\pi}\right)E_{\mathbf{y}|\mathbf{x}}[\psi(\mathbf{y}|\boldsymbol{\theta})] + (1-z)E_{\mathbf{y}|\mathbf{x}}[\psi(\mathbf{y}|\boldsymbol{\theta})]. \tag{20}$$

This is the estimating equation of the doubly robust estimator. □

3.3 Asymptotic Distribution

We have shown that our doubly robust estimator is M-estimator. The asymptotic distribution of our estimator is the following normal distribution:

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim N(0, \mathbf{V}(\boldsymbol{\theta}_0)), \tag{21}$$

$\mathbf{V}(\boldsymbol{\theta}_0)$ can be estimated by $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{A}}(\hat{\boldsymbol{\theta}})^{-1}\hat{\mathbf{B}}(\hat{\boldsymbol{\theta}})\{\hat{\mathbf{A}}(\hat{\boldsymbol{\theta}})^{-1}\}^t$,
where

$$\hat{\mathbf{A}}(\hat{\boldsymbol{\theta}}) = -\frac{1}{N+M}\sum_{i=1}^{N+M}\frac{\partial}{\partial\boldsymbol{\theta}^t}\mathbf{m}(\mathbf{y}_i,\mathbf{x}_i,z_i,r_i,\hat{\boldsymbol{\theta}}),$$

$$\hat{\mathbf{B}}(\hat{\boldsymbol{\theta}}) = \frac{1}{N+M}\sum_{i=1}^{N+M}\mathbf{m}(\mathbf{y}_i,\mathbf{x}_i,z_i,r_i,\hat{\boldsymbol{\theta}})\mathbf{m}(\mathbf{y}_i,\mathbf{x}_i,z_i,r_i,\hat{\boldsymbol{\theta}})^t,$$

$$\mathbf{m}(\mathbf{y}_i,\mathbf{x}_i,z_i,r_i,\boldsymbol{\theta}) = \frac{z_ir_i}{\pi_i}\psi(\mathbf{y}_i|\boldsymbol{\theta}) + z_i\left(1-\frac{r_i}{\pi_i}\right)E_{\mathbf{y}_i|\mathbf{x}_i}[\psi(\mathbf{y}_i|\boldsymbol{\theta})] + (1-z_i)E_{\mathbf{y}_i|\mathbf{x}_i}[\psi(\mathbf{y}_i|\boldsymbol{\theta})].$$

[4]Define $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}'$.

4 SIMULATION STUDY

In the simulation study, we generate 1,000 datasets in which the total sample size $(M + N)$ is 3,000. All statistical computations in this paper are carried out by R version 3.5.1. There are two covariates in the dataset, x_{1i}, x_{2i} following Student's t-distributions with freedom of degree $\nu = 5$ and with parallel transport so as to the sample mean equals to 1. We arbitrarily set the regression model and missing mechanism. y is calculated by the regression model, $p(y_i|x_{1i}, x_{2i}, \boldsymbol{\beta})$. Then, 2,000 $(= M)$ observations are divided as random sampling data among the 3,000 observations $\{x_{1m}, x_{2m}\}_{m=1}^{2000}$ and the remaining 1,000 $(= N)$ observations are divided as candidates of biased sampling data. Whether each unit is observed in the biased sampling data is determined by the following probabilistic missing mechanism (logistic function), $p(r = 1|x_{1i}, x_{2i}, \boldsymbol{\alpha})$[5]. Suppose we are interested in the population mean of y, ie., $\theta = E[y]$ and $\psi(y|\theta) = y - \theta$.

To investigate the properties of the proposed doubly robust-type estimator, we conduct simulations in the following two conditions:

(i) the case in which the regression model is misspecified but the missing mechanism is correctly specified,

(ii) the case in which the missing mechanism is misspecified but the regression model is correctly specified.

First, we consider the case in which the regression model is misspecified. So as to generate the simulation datasets, we set the regression model as a quadratic function $y = \mathbf{x}^t\mathbf{B}\mathbf{x} + \varepsilon$ and the logit of missing mechanism as a linear function $p(r = 1|\mathbf{x}, \boldsymbol{\alpha}) = \mathbf{h}(\mathbf{x})^t\boldsymbol{\alpha} + \varepsilon$ [6], where $\mathbf{h}(\mathbf{x}) = (1, x_1 - \frac{1}{2000}\sum_{m=1}^{2000} x_{1m}, x_2 - \frac{1}{2000}\sum_{m=1}^{2000} x_{2m})^t, \varepsilon \sim N(0, 1)$.

We compared the following three methods:

- REG: Monte Carlo integration based on MLE, where the regression function is *misspecified* as liner.
- IPW: inverse probability weighting estimator based on the propensity score by the method of Nevo (2003), where the missing mechanism is *correctly specified* as linear.
- PROP-DR: the proposed doubly robust estimation method, where the regression function is *misspecified* as liner.

Table 1: Average of estimates $\theta = E[y]$ when the regression function is misspecified

	True value	REG	IPW	PROP-DR
mean	0.000	-0.127	0.003	-0.087
SD		0.119	0.116	0.105
MSE		0.026	0.011	0.015
MSEratio		170.81	70.52	100.00

Table 1 reports the simulation results when the regression model is misspecified as linear; however, in the data-generating model, the regression function is set to be quadratic. "True value" is the sample mean of y in random sampling data. The row "mean" shows the average of estimation results from 1,000 datasets and the row "SD" shows the standard deviations. "MSE" means mean squared error and "MSEratio" shows the ratio of MSE of PROP-DR and MSE of another method when the "MSEratio" of PROP-DR is set as 100. We compare three types of estimator that use only biased sampling data and the covariates x_1, x_2 in random sampling data

[5] Because the missing mechanism is not a function of y, Assumption 1 for strong ignorability is satisfied.

[6] We set $\mathbf{B} = \begin{pmatrix} -3.0 & 0.8 & 1.2 \\ 0.4 & -0.2 & -0.2 \\ 0.8 & 0.0 & 0.2 \end{pmatrix}$ and $\boldsymbol{\alpha} = \begin{pmatrix} 0.4 & -0.4 & 0.8 \end{pmatrix}^t$.

as auxiliary information. Although the IPW estimator provides the best estimates since the missing mechanism is correctly specified, proposed estimator also works well even though the regression model is misspecified.

Second, we consider the case in which the missing mechanism is misspecified. To generate the simulation datasets, we set the logit of missing mechanism as a quadratic function $p(r = 1|\mathbf{x}, \boldsymbol{\alpha}) = \mathbf{h}(\mathbf{x})^t\boldsymbol{\alpha} + \varepsilon$ and the regression model as a linear function $y = \mathbf{x}^t\boldsymbol{\beta} + \varepsilon$ [7], where $\mathbf{h}(\mathbf{x}) = (1, x_1 - \frac{1}{2000}\sum_{m=1}^{2000} x_{1m}, x_2 - \frac{1}{2000}\sum_{m=1}^{2000} x_{2m}, x_1^2 - \frac{1}{2000}\sum_{m=1}^{2000} x_{1m}^2, x_2^2 - \frac{1}{2000}\sum_{m=1}^{2000} x_{2m}^2)^t, \varepsilon \sim N(0, 1)$.

We compared the following three methods:

- REG: Monte Carlo integration based on MLE, where the regression function is *correctly specified* as linear.
- IPW: inverse probability weighting estimator based on the propensity score by the method of Nevo (2003), where the missing mechanism is *misspecified* as liner.
- PROP-DR: the proposed doubly robust estimation method, where the missing mechanism is *misspecified* as liner.

Table 2: Average of estimates $\theta = E[y]$ when the Assignment mechanism is misspecified

	True value	REG	IPW	PROP-DR
mean	0.000	0.000	-0.023	0.000
SD		0.057	0.0084	0.057
MSE		0.0025	0.0070	0.0026
MSEratio		97.02	272.15	100.00

Table 2 reports the simulation results when the logit of the missing mechanism is misspecified as linear; however, in the data-generating model, the logit is set to be quadratic. The true value is the sample mean of y in random sampling data. Although REG provides the best estimates because the regression model is correctly specified, the mean squared error (MSE) of the proposed estimator is almost the same as REG even though the missing mechanism is misspecified.

5 EMPIRICAL APPLICATION

We apply our method to marketing data in which the target of inference is the population mean of the purchasing interval of drink products. We consider the case where we have two datasets, the biased sampling dataset (purchasing interval and covariate information obtained in a specific store chain) and the random sampling dataset (only covariate information obtained from all the competing stores in our dataset). In this case, y in Figure 2 is the purchasing interval of drink products, θ is the population mean of purchasing intervals $E[y]$ and $\mathbf{x}$ in that figure are covariates as stated below. Small data ($z = 1$) in Figure 2 corresponds to the biased sampling dataset of a specific store chain and big data ($z = 0$) in that figure is equivalent to covariates data of all stores. The purchasing interval represents purchasing frequency, which is a quite important index in marketing. We show that in our marketing data, the sample mean of the biased sampling dataset is smaller than the (true) sample mean of the random sampling dataset. Our doubly robust estimator allows us to correct that bias by using the auxiliary information, ie., covariates of the random sampling dataset.

[7] We set $\boldsymbol{\alpha} = \begin{pmatrix} 0.4 & 0.1 & 0.1 & -0.4 & 0.6 \end{pmatrix}$ and $\boldsymbol{\beta} = \begin{pmatrix} 0 & 0.8 & -0.8 \end{pmatrix}^t$.

5.1 Data

We use the *SCI* data by Intage Inc. in Japan. The SCI is scanner panel data and is one of the most popular purchase panel datasets in Japanese marketing. The SCI data contains purchase incidences, purchase products, number of products the consumer purchases, amounts, prices, and purchase stores with date and time. Additionally, the SCI has records about covariates of consumers such as age, gender, income, and so on.

We analyze the purchasing interval of drink products like soft drinks and alcoholic beverages, coffee, tea etc. There are covariates: Age Class (in five-year increments), Gender, Occupation Code, logarithms for Individual Annual Income, Living with Spouse or Not, Living with Child (17-years-old or younger), Owning a Car or Not, and Educational Background Code. The data period is from July 2015 to June 2016. We use the purchasing interval of drink products and the covariates of a specific supermarket store chain, that is popular in Japan and with stores throughout the country, as biased sampling dataset[8]. We regard the covariate information from the whole SCI dataset as a random sampling dataset [9]. The sample size of biased sampling dataset is $n = 3870$ and that of random sampling one is $M = 55620$. In this setup, we can compute the sample mean of purchasing intervals of full SCI data and call this value "true" value. The dataset of consumers in that store chain can be thought as a biased sampling dataset because the sample mean of a biased sampling dataset is smaller than the true sample mean of a random sampling dataset as Table 3.

5.2 Result

We use the following three methods:

- REG: Monte Carlo integration based on MLE, where the regression model is an exponential regression, linear about the parameter[10].
- IPW: inverse probability weighting estimator based on the propensity score by the method of Nevo (2003), where the logit function is liner.
- PROP-DR: the proposed doubly robust estimation method.

Table 3: Difference between the estimation result for mean of purchasing interval θ

	True	Sample mean of biased data	REG	IPW	PROP-DR
θ	7.993	6.774	8.074	7.208	7.974
s.e.		0.137	0.008	0.348	0.100

Table 3 reports the results of this analysis. *Ture* means the sample mean of purchasing intervals in the random sampling dataset, *Sample mean of biased data* is computed by just taking the average of those in the biased sampling dataset, and *s.e.* represents the standard error[11] and Bias/SE is computed by $|True - \hat{\theta}|/SE$. All three methods can correct the bias between the true and sample means of biased data. Especially, we can correct the bias greatly by the proposed method.

[8]Purchasing interval of each observation is average of purchasing intervals of each consumer. The purchasing intervals of each consumer are intervals between their purchases from any store in the SCI data.

[9]In both biased and random sampling datasets, we use records of consumer who have bought drink products more than two times from any stores in the period.

[10]For example, see Cameron and Trivedi (2005) for a review of exponential regression and estimation method for such non-linear models.

[11]The standard errors are computed using asymptotic variance with the assumption that both the regression model and the missing mechanism are correctly specified. The row of SE shows that the standard error of REG is quite small because in REG calculation we use $M = 55620$ observations.

6 USE ONLY MOMENT INFORMATION AS AUXILIARY information

In previous sections, we use a unit-level covariates dataset as auxiliary information. In this section, we also develop a method that needs only moment information as auxiliary information (like Figure 3).

6.1 Setup

In this section, we assume Assumption 1 (strong ignorability) and the following two conditions:

Assumption 3. *Let $\bar{\mathbf{h}}(\mathbf{x})$ be an J-dimensional function and $\mathbf{m}(\mathbf{x})$ be a K-dimensional function ($\bar{\mathbf{h}}(\mathbf{x})$ and $\mathbf{m}(\mathbf{x})$ are allowed to contain some elements that are bijections from one to the others).*

Suppose that $E[\bar{\mathbf{h}}(\mathbf{x})]$ and $E[\mathbf{m}(\mathbf{x})]$ are obtained as the auxiliary moment information. Then, at least one of the following conditions is satisfied:

(i) A missing mechanism $p(r=1|\mathbf{x})$ can be correctly specified with $\mathbf{h}(\mathbf{x})$:
$logit[p(r=1|\mathbf{x})] = \mathbf{h}(\mathbf{x})^t\boldsymbol{\alpha}$, where, $\mathbf{h}(\mathbf{x}) = \bar{\mathbf{h}}(\mathbf{x}) - E[\bar{\mathbf{h}}(\mathbf{x})]$ and $\boldsymbol{\alpha}$ is a parameter vector.

(ii) A regression model $p(\mathbf{y}|\mathbf{x})$ can be correctly specified with $\mathbf{m}(\mathbf{x})$:
$p(\mathbf{y}|\mathbf{x}) = \mathbf{m}(\mathbf{x})^t\boldsymbol{\beta}$, where, $\boldsymbol{\beta}$ is a parameter vector.

Assumption 4. *Let $\boldsymbol{\beta}$ be a parametor vector of the regression model:*
$E_{\mathbf{x}}[E_{\mathbf{y}|\mathbf{x},\boldsymbol{\beta}}[\psi(\mathbf{y}|\boldsymbol{\theta})]]$ is a function of $E[\mathbf{m}(\mathbf{x})], \boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

Let $\mathbf{h}(\mathbf{x})$ and $\mathbf{m}(\mathbf{x})$ be represented together by $\mathbf{g}(\mathbf{x})$. We need Assumption 3 in order to develop models in the case where only moment information can be used as the auxiliary information. This assumption means that if the researchers want to develop a complex model with higher-degree covariates, then they have to obtain higher-degree moments. For example, if the researchers want to develop a quadratic regression model, they must obtain up to not the first moment, but up to the second moment.

6.2 Doubly Robust Estimator

Model 1: missing mechanism

A missing mechanism is similar to previous sections. Let the logit of the propensity score be a linear equation about $\mathbf{h}(\mathbf{x})$,

$$logit[p(r=1|\mathbf{x})] = \mathbf{h}(\mathbf{x})^t\boldsymbol{\alpha}, \tag{22}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^J$ is parameters for logit [12].

Then, with the biased data x_i and by solving simultaneous equations:

$$\begin{cases} \displaystyle\sum_{i=1}^{n} \frac{1}{p(r=1|\mathbf{x}_i,\boldsymbol{\alpha})} = \frac{n}{\hat{p}(r=1)} & (23) \\ \displaystyle\sum_{i=1}^{n} \frac{\mathbf{h}(\mathbf{x}_i)}{p(r=1|\mathbf{x}_i,\boldsymbol{\alpha})} = 0 \quad , & (24) \end{cases}$$

$\hat{\boldsymbol{\alpha}}$ can be estimated and the propensity score can be calculated using the estimated value. In previous sections, we compute $\hat{p}(r=1)$ by using a unit-level covariates dataset; however, in

[12] In the case in which $\mathbf{x}$ contains a constant as a first element, $\mathbf{h}(\mathbf{x})$ may contain the element $constant - E[constant] = 0$. To contain a constant in $\mathbf{h}(\mathbf{x})$, this zero element can be replaced by 1.

the setting of this section we do not know the datasets. We only know the moment information of the covariates. Because $p(r=1)=\frac{n}{N}$, we can compute $\hat{p}(r=1)$ as follows:

$$\hat{p}(r=1)=\frac{E[\bar{\mathbf{h}}(\mathbf{x}_i)_j]}{\frac{1}{n}\sum_{i=1}^{n}\frac{\bar{\mathbf{h}}(\mathbf{x}_i)_j}{p(r=1|\mathbf{x},\boldsymbol{\alpha})}}, \tag{25}$$

where $\bar{\mathbf{h}}(\mathbf{x}_i)_j$ is any j-th component of $\bar{\mathbf{h}}(\mathbf{x}_i)$. This equation is derived from $E[\bar{\mathbf{h}}(\mathbf{x}_i)_j]=\frac{1}{N}\sum_{i=1}^{n}\frac{r_i\bar{\mathbf{h}}(\mathbf{x}_i)_j}{p(r=1|\mathbf{x},\boldsymbol{\alpha})}$.

Calculate the IPW estimator using weighting by the inverse of the estimated propensity score $p(r=1|\mathbf{x}_i,\hat{\boldsymbol{\alpha}})$.

Model 2: regression model

From the assumption that $\mathbf{y}$ and r are independent when $\mathbf{x}$ is conditioned, $p(\mathbf{y}|r=1,\mathbf{x},\boldsymbol{\beta})=p(\mathbf{y}|\mathbf{x},\boldsymbol{\beta})$. Therefore, the regression model estimated from biased data coincides with that estimated from random sampling data.

By assumption 4, $E_{\mathbf{x},\mathbf{y}|\boldsymbol{\beta}}[\psi(\mathbf{y}|\boldsymbol{\theta})]=E_{\mathbf{x}}[E_{\mathbf{y}|\mathbf{x},\boldsymbol{\beta}}[\psi(\mathbf{y}|\boldsymbol{\theta})]]$ can be represented with $E[\mathbf{m}(\mathbf{x})],\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The solution $\hat{\boldsymbol{\theta}}$ of the following equation (26) is a consistent estimator for the parameter $\boldsymbol{\theta}_0$ if the regression model is correctly set:

$$E_{\mathbf{x},\mathbf{y}|\hat{\boldsymbol{\beta}}}[\psi(\mathbf{y}|\boldsymbol{\theta})]=\mathbf{0}, \tag{26}$$

where $\hat{\boldsymbol{\beta}}$ is a consistent estimator for the coefficient of the regression model.

Doubly robust estimator

If at least one of the two models is set correctly, then the solution $\hat{\boldsymbol{\theta}}$ of the following equation (27) will be a consistent estimator for the parameter $\boldsymbol{\theta}_0$.

$$\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{r_i}{p(r=1|\mathbf{x}_i,\boldsymbol{\alpha})}\psi(\mathbf{y}_i|\boldsymbol{\theta})+\left(1-\frac{r_i}{p(r=1|\mathbf{x}_i,\boldsymbol{\alpha})}\right)E_{\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\beta}}[\psi(\mathbf{y}_i|\boldsymbol{\theta})]\right\}=\mathbf{0}, \tag{27}$$

When N is enough large, $\frac{1}{N}\sum_{i=1}^{N}E_{\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\beta}}[\psi(\mathbf{y}_i|\boldsymbol{\theta})]=E_{\mathbf{x},\mathbf{y}|\boldsymbol{\beta}}[\psi(\mathbf{y}|\boldsymbol{\theta})]$. By definition of r_i, equation (27) will be

$$\frac{1}{N}\sum_{i=1}^{n}\left\{\frac{1}{p(r=1|\mathbf{x}_i,\boldsymbol{\alpha})}\left(\psi(\mathbf{y}_i|\boldsymbol{\theta})-E_{\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\beta}}[\psi(\mathbf{y}_i|\boldsymbol{\theta})]\right)\right\}+E_{\mathbf{x},\mathbf{y}|\boldsymbol{\beta}}[\psi(\mathbf{y}|\boldsymbol{\theta})]=\mathbf{0}. \tag{28}$$

We need to estimate $N, p(r=1|\mathbf{x}_i,\boldsymbol{\alpha})$ and $p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\beta})$, so replace these by the estimated value $\hat{N}$ and probability density functions $p(r=1|\mathbf{x}_i,\hat{\boldsymbol{\alpha}})$, $p(\mathbf{y}_i|\mathbf{x}_i,\hat{\boldsymbol{\beta}})$. We can estimate $\hat{N}$ by $\hat{N}=\frac{\sum_{i=1}^{n}\frac{\bar{\mathbf{h}}(\mathbf{x}_i)_j}{p(r=1|\mathbf{x},\hat{\boldsymbol{\alpha}})}}{E[\bar{\mathbf{h}}(\mathbf{x}_i)_j]}$. After replacing, the estimating equation will be

$$\frac{1}{\hat{N}}\sum_{i=1}^{n}\left\{\frac{1}{p(r=1|\mathbf{x}_i,\hat{\boldsymbol{\alpha}})}\left(\psi(\mathbf{y}_i|\boldsymbol{\theta})-E_{\mathbf{y}_i|\mathbf{x}_i,\hat{\boldsymbol{\beta}}}[\psi(\mathbf{y}_i|\boldsymbol{\theta})]\right)\right\}+E_{\mathbf{x},\mathbf{y}|\hat{\boldsymbol{\beta}}}[\psi(\mathbf{y}|\boldsymbol{\theta})]=\mathbf{0}. \tag{29}$$

The second term of equation (29) can be calculated when we satisfy assumption 4.

Monte Carlo integration

The doubly robust estimator of estimate equation (29) contains the conditional expectations $E_{\mathbf{y}_i|\mathbf{x}_i,\hat{\boldsymbol{\beta}}}[\cdot]$. In the case where $\psi(\mathbf{y}|\boldsymbol{\theta})$ is a continuous function, we have to solve the integral,

which is often difficult to do. If the researchers cannot solve the integral analytically, than thay can replace it with Monte Carlo integration.

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{p(r=1|\mathbf{x}_i,\hat{\boldsymbol{\alpha}})}\left(\psi(\mathbf{y}_i|\boldsymbol{\theta})-\frac{1}{L}\sum_{l=1}^{L}\psi(\mathbf{y}_{il}|\boldsymbol{\theta})\right)\right\}+E_{\mathbf{x},\mathbf{y}|\hat{\boldsymbol{\beta}}}[\psi(\mathbf{y}|\boldsymbol{\theta})]=\mathbf{0}. \tag{30}$$

where, $\mathbf{y}_{il}\sim p(\mathbf{y}|\mathbf{x}_i,\hat{\boldsymbol{\beta}})$, $l=1,\ldots,L,\ i=1,\ldots,n$.

Example

We provide an example for estimating the population moment that shows when Assumption 3 and 4 are satisfied. We consider the case where we want to estimate the population moment of y (one dimension) and we have two covariates $\mathbf{x}=(x_1,x_2)^t$. There are biased sampling dataset and auxiliary first moment information $E[x_1]$ and $E[x_2]$. In this case, $\theta=E[y]$ and $\psi(y)=y-\theta$. Suppose that the true regression model is $p(y|\mathbf{x},\boldsymbol{\beta})=\beta_0+\beta_1x_1+\beta_2x_2+\varepsilon=m(\mathbf{x})^t\boldsymbol{\beta}+\varepsilon$, where $m(\mathbf{x})=(1,x_1,x_2)^t$ and the true missing mechanism is $p(r=1|\mathbf{x},\boldsymbol{\alpha})=h_1(\mathbf{x})\boldsymbol{\alpha_1}+h_2(\mathbf{x})\boldsymbol{\alpha_2}+\varepsilon=\mathbf{h}(\mathbf{x})^t\boldsymbol{\alpha}+\varepsilon$, where $h_1(\mathbf{x})=(x_1,x_2)^t-E[(x_1,x_2)^t]$ and $h_2(\mathbf{x})=(x_1^2,x_2^2)^t-E[(x_1^2,x_2^2)^t]$. We estimate $\theta=E[y]$ by using the doubly robust estimator with the regression model correctly specified as $p(y|\mathbf{x},\boldsymbol{\beta})=m(\mathbf{x})^t\boldsymbol{\beta}+\varepsilon$ and the missing mechanism misspecified as $p(r=1|\mathbf{x},\boldsymbol{\alpha})=h_1(\mathbf{x})\boldsymbol{\alpha_1}+\varepsilon$ (we cannot contain a term of $h_2(\mathbf{x})$ in the missing mechanism because we do not know the second moment $E[x_1^2]$ and $E[x_2^2]$). Although a missing mechanism is misspecified, Assumption 3 is satisfied because we correctly specified a regression model and we have $E[m(\mathbf{x})]=(E[x_1],\ E[x_2])^t$ as auxiliary information. This case satisfies Assumption 4 because $E_{\mathbf{x}}[E_{y|\mathbf{x},\boldsymbol{\beta}}[\psi(y|\theta)]]=E_{\mathbf{x}}[E_{y|\mathbf{x},\boldsymbol{\beta}}[y]]-\theta=E_{\mathbf{x}}[m(\mathbf{x})]^t\boldsymbol{\beta}-\theta$. Moreover, Assumption 4 can be satisfied if the regression model is linear about $\boldsymbol{\beta}$. Therefore, if we can assume Assumption 1, then the doubly robust estimator provides a consistent estimator.

Asymptotic Properties

Proof of Consistency

We denote $p(r=1|\mathbf{x},\hat{\boldsymbol{\alpha}})=\pi$, $E_{\mathbf{y}|\mathbf{x},\hat{\boldsymbol{\beta}}}[\psi(\mathbf{y}|\boldsymbol{\theta})]=E_{\mathbf{y}|\mathbf{x}}[\psi(\mathbf{y}|\boldsymbol{\theta})]$, and $E_{\mathbf{x},\mathbf{y}|\hat{\boldsymbol{\beta}}}[\psi(\mathbf{y}|\boldsymbol{\theta})]=E_{\mathbf{x},\mathbf{y}}[\psi(\mathbf{y}|\boldsymbol{\theta})]$ in the proof.

Theorem 3. *If either the missing mechanism or the regression model is set correctly, then the doubly robust estimator has consistency.*

Proof. We prove this theorem by showing that the doubly robust estimator is M-estimator, which has consistency.

(i) When propensity score $p(r=1|\mathbf{x},\hat{\boldsymbol{\alpha}})=\pi$ is correctly specified

In this case, regression model $p(\mathbf{y}|\mathbf{x},\boldsymbol{\beta})$ may be misspecified. This implies $E_{\mathbf{y}|\mathbf{x}}[\psi(\mathbf{y}|\boldsymbol{\theta})]$ may also be misspecified. We denote this expectation as $\tilde{E}_{\mathbf{y}|\mathbf{x}}[\psi(\mathbf{y}|\boldsymbol{\theta})]$.
As $n\to\infty$ (and $\frac{n}{N}=p(r=1)>0$),

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{r_i}{\pi}\psi(\mathbf{y}_i|\boldsymbol{\theta})+\left(1-\frac{r_i}{\pi}\right)\tilde{E}_{\mathbf{y}_i|\mathbf{x}_i}[\psi(\mathbf{y}_i|\boldsymbol{\theta})]\right\}\\
\to&E_{\mathbf{x}}\left[E_{r|\mathbf{x}}\left[\frac{r}{\pi}\right]E_{\mathbf{y}|\mathbf{x}}\left[\psi(\mathbf{y}|\boldsymbol{\theta})\right]+E_{r|\mathbf{x}}\left[\frac{r-\pi}{\pi}\right]\tilde{E}_{\mathbf{y}|\mathbf{x}}[\psi(\mathbf{y}|\boldsymbol{\theta})]\right]\\
=&E_{\mathbf{x}}\left[E_{\mathbf{y}|x}\left[\psi(\mathbf{y}|\boldsymbol{\theta})\right]\right]\quad\left(\because E_{r|\mathbf{x}}\left[\frac{r-\pi}{\pi}\right]=\frac{\pi-\pi}{\pi}=0\right)\\
=&E\left[\psi(\mathbf{y}|\boldsymbol{\theta})\right].
\end{aligned}$$

Therefore, as $n\to\infty$ (and $\frac{n}{N}=p(r=1)>0$), $\hat{\boldsymbol{\theta}}\xrightarrow{\mathrm{p}}\boldsymbol{\theta}_0$.

(ii) When the regression model $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta})$ is correctly specified

In this case, the missing mechanism $p(r = 1|\mathbf{x}, \boldsymbol{\alpha}) = \pi$ may be misspecified. This implies that estimated $\hat{N}$ may be incorrect. We denote these π and N as $\tilde{\pi}$ and $\tilde{N}$.

Equation (29) can be represented as follows:

$$\frac{1}{\tilde{N}} \sum_{i=1}^{n} \left\{ \frac{1}{\tilde{\pi}} \left(\psi(\mathbf{y}_i|\boldsymbol{\theta}) - E_{\mathbf{y}_i|\mathbf{x}_i}[\psi(\mathbf{y}_i|\boldsymbol{\theta})] \right) \right\} + E_{\mathbf{x},\mathbf{y}}[\psi(\mathbf{y}|\boldsymbol{\theta})] = \mathbf{0}. \tag{31}$$

As $n \to \infty$ (and $\frac{n}{\tilde{N}} = \hat{p}(r = 1) > 0$),

$$\begin{aligned}
&\frac{n}{\tilde{N}} \cdot \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{1}{\tilde{\pi}} \left(\psi(\mathbf{y}_i|\boldsymbol{\theta}) - E_{\mathbf{y}_i|\mathbf{x}_i}[\psi(\mathbf{y}_i|\boldsymbol{\theta})] \right) \right\} + E_{\mathbf{x},\mathbf{y}}[\psi(\mathbf{y}|\boldsymbol{\theta})] \\
\to &\hat{p}(r = 1) \cdot E_{\mathbf{x}} \left[E_{r|\mathbf{x}} \left[\frac{r}{\tilde{\pi}} \right] E_{\mathbf{y}|\mathbf{x}} \left[\psi(\mathbf{y}|\boldsymbol{\theta}) - \psi(\mathbf{y}|\boldsymbol{\theta}) \right] \right] + E_{\mathbf{x},\mathbf{y}}[\psi(\mathbf{y}|\boldsymbol{\theta})] \\
= &E_{\mathbf{x},\mathbf{y}}[\psi(\mathbf{y}|\boldsymbol{\theta})] \\
= &E\left[\psi(\mathbf{y}|\boldsymbol{\theta}) \right].
\end{aligned}$$

Therefore, as $n \to \infty$ (and $\frac{n}{\tilde{N}} = \hat{p}(r = 1) > 0$), $\hat{\boldsymbol{\theta}} \xrightarrow{\mathrm{p}} \boldsymbol{\theta}_0$.

The proposed estimator has consistency in both cases. □

Proof of Efficiency

Theorem 4. *When both the missing mechanism and the regression model are set correctly, the doubly robust estimator has minimum variance in specific models with the augment term (semi-parametric efficiency).*

Proof. If all the samples were observed, then let $D^F(\boldsymbol{\theta}) = \psi(\mathbf{y}|\boldsymbol{\theta})$,

$$\frac{1}{N} \sum_{i=1}^{N} \left\{ D^F(\boldsymbol{\theta}) \right\} = \mathbf{0}, \tag{32}$$

then, the solution $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ will be a consistent estimator. However, there is missingness in this setting; therefore, we use $D(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{r}{\pi} D^F(\boldsymbol{\theta}) - A(\boldsymbol{\phi})$ and estimate $\boldsymbol{\theta}$ by solving

$$\frac{1}{N} \sum_{i=1}^{N} \left\{ D_i(\boldsymbol{\theta}, \boldsymbol{\phi}) \right\} = \mathbf{0}. \tag{33}$$

where $\pi = p(r = 1|\mathbf{x})$, $A(\boldsymbol{\phi}) = \frac{r-\pi}{\pi} \boldsymbol{\phi}$, and

$$D(\boldsymbol{\theta}, \boldsymbol{\phi}) = D^F(\boldsymbol{\theta}) + \frac{r - \pi}{\pi} \left\{ D^F(\boldsymbol{\theta}) - \boldsymbol{\phi} \right\}. \tag{34}$$

Because under the condition of $\mathbf{x}$, $\mathbf{y}$ and r are independent, the first and second terms do not correlate and the variance of $D(\boldsymbol{\theta}, \boldsymbol{\phi})$ is

$$Var\left(D^F(\boldsymbol{\theta}) \right) + E_{\mathbf{x}} \left[\frac{1 - \pi}{\pi} E_{\mathbf{y}|\mathbf{x}} \left[\left\{ D^F(\boldsymbol{\theta}) - \boldsymbol{\phi} \right\}^{\otimes 2} \right] \right]. \tag{35}$$

Following $\boldsymbol{\phi}^*$ minimize (35) and according to Proposition2.2 of Robins *et al.* (1994), the $\boldsymbol{\phi}^*$ minimizing the variance of $D(\boldsymbol{\theta}, \boldsymbol{\phi})$ minimizes the asymptotic variance of $\hat{\boldsymbol{\theta}}$,

$$\boldsymbol{\phi}^* = E_{\mathbf{y}|\mathbf{x}} \left[D^F(\boldsymbol{\theta}) \right] = E_{\mathbf{y}|\mathbf{x}} \left[\psi(\mathbf{y}|\boldsymbol{\theta}) \right]. \tag{36}$$

Then, $D(\boldsymbol{\theta}, \boldsymbol{\phi})$ will be

$$\frac{r}{\pi} \psi(\mathbf{y}|\boldsymbol{\theta}) + \left(1 - \frac{r}{\pi} \right) E_{\mathbf{y}|\mathbf{x}}[\psi(\mathbf{y}|\boldsymbol{\theta})]. \tag{37}$$

This is the estimating equation of the doubly robust estimator. □

Asymptotic Distribution

We have shown that our doubly robust estimator is M-estimator. The asymptotic distribution of our estimator is the following normal distribution:

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim N(0, \mathbf{V}(\boldsymbol{\theta}_0)), \tag{38}$$

$\mathbf{V}(\boldsymbol{\theta}_0)$ can be estimated by $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{A}}(\hat{\boldsymbol{\theta}})^{-1}\hat{\mathbf{B}}(\hat{\boldsymbol{\theta}})\{\hat{\mathbf{A}}(\hat{\boldsymbol{\theta}})^{-1}\}^t$,
where

$$\hat{\mathbf{A}}(\hat{\boldsymbol{\theta}}) = -\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial \boldsymbol{\theta}^t}\mathbf{m}(\mathbf{y}_i, \mathbf{x}_i, r_i, \hat{\boldsymbol{\theta}}),$$

$$\hat{\mathbf{B}}(\hat{\boldsymbol{\theta}}) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{m}(\mathbf{y}_i, \mathbf{x}_i, r_i, \hat{\boldsymbol{\theta}})\mathbf{m}(\mathbf{y}_i, \mathbf{x}_i, r_i, \hat{\boldsymbol{\theta}})^t,$$

$$\mathbf{m}(\mathbf{y}_i, \mathbf{x}_i, r_i, \boldsymbol{\theta}) = \frac{r_i}{\pi}\psi(\mathbf{y}_i|\boldsymbol{\theta}) + \left(1 - \frac{r_i}{\pi}\right)E_{\mathbf{y}_i|\mathbf{x}_i}[\psi(\mathbf{y}_i|\boldsymbol{\theta})].$$

Simulation

We conduct two types of simulation studies when auxiliary information is moment information. One is the estimation of population moment and the other is the estimation of the population parameters for the logistic model.

Estimation of population moment

In this simulation study, the goal is to estimate population moment. We generate 1,000 datasets in which the total sample size $(M+N)$ is 3,000. There are two covariates in the dataset, x_{1i}, x_{2i}, following Student's t-distributions with freedom of degree $\nu = 3$ and with parallel transport so that the sample mean equals to 1. We arbitrarily set the regression model and missing mechanism. y is calculated by the regression model, $p(y_i|x_{1i}, x_{2i}, \boldsymbol{\beta})$. Then, 2,000 $(= M)$ observations are divided as random sampling data among the 3,000 observations $\{x_{1m}, x_{2m}\}_{m=1}^{2000}$ and we use their sample mean as the auxiliary information. The remaining 1,000 $(= N)$ observations are divided as candidates of biased sampling data. Whether each unit is observed in the biased sampling data is determined by the following probabilistic missing mechanism (logistic function): $p(r = 1|x_{1i}, x_{2i}, \boldsymbol{\alpha})$. Suppose that we are interested in a population mean of y, ie., $\theta = E[y]$ and $\psi(y|\theta) = y - \theta$.

To investigate the properties of the proposed doubly robust-type estimator, we conduct simulations in the following two conditions:

(i) the case in which the regression model is misspecified but the missing mechanism is correctly specified,

(ii) the case in which the missing mechanism is misspecified but the regression model is correctly specified.

First, we consider the case in which the regression model is misspecified. To generate the simulation datasets, we set the regression model as a quadratic function $y = \mathbf{x}^t\mathbf{B}\mathbf{x} + \varepsilon$ and the logit of the missing mechanism as a linear function $p(r = 1|\mathbf{x}, \boldsymbol{\alpha}) = \mathbf{h}(\mathbf{x})^t\boldsymbol{\alpha} + \varepsilon$ [13], where $\mathbf{h}(\mathbf{x}) = (1, x_1 - \frac{1}{2000}\sum_{m=1}^{2000}x_{1m}, x_2 - \frac{1}{2000}\sum_{m=1}^{2000}x_{2m})^t, \varepsilon \sim N(0, 1)$.

We compared the following three methods:

- REG: Calculate $E[y]$ based on MLE, where the regression function is *misspecified* as liner.

[13] We set $\mathbf{B} = \begin{pmatrix} -3.0 & 0.8 & 1.2 \\ 0.4 & -0.2 & -0.2 \\ 0.8 & 0.0 & 0.2 \end{pmatrix}$ and $\boldsymbol{\alpha} = \begin{pmatrix} 0.4 & -0.4 & 0.8 \end{pmatrix}^t$.

- IPW: inverse probability weighting estimator based on the propensity score by the method of Nevo (2003), where the missing mechanism is *correctly specified* as linear.
- PROP-DR: the proposed doubly robust estimation method, where the regression function is *misspecified* as liner.

Table 4: Average of estimates $\theta = E[y]$ when the regression function is misspecified

	True value	REG	IPW	PROP-DR
mean	0.003	-0.328	0.006	-0.045
SD		0.85	0.33	0.29
MSE		0.89	0.20	0.17
MSEratio		510.37	116.01	100.00

Table 4 reports the simulation results when the regression model is misspecified as linear; however, in the data-generating model, the regression function is set to be quadratic. "True value" is the sample mean of y in random sampling data. The row "mean" shows the average of estimation results from 1,000 datasets and the row "SD" shows their standard deviations. "MSE" means mean squared error and "MSEratio" shows the ratio of MSE of PROP-DR and MSE of another method when the "MSEratio" of PROP-DR is set as 100. We compare three types of estimator that use only biased sampling data and the first moment of the covariates x_1, x_2 as auxiliary information. Although the missing mechanism is correctly specified, the proposed estimator provides the minimum MSE even though the regression model is misspecified.

Second, we consider the case in which the missing mechanism is misspecified. To generate the simulation datasets, we set the logit of the missing mechanism as a quadratic function $p(r = 1|\mathbf{x}, \boldsymbol{\alpha}) = \mathbf{h}(\mathbf{x})^t\boldsymbol{\alpha} + \varepsilon$ and the regression model as a linear function $y = \mathbf{x}^t\boldsymbol{\beta} + \varepsilon$ [14], where $\mathbf{h}(\mathbf{x}) = (1, x_1 - \frac{1}{2000}\sum_{m=1}^{2000} x_{1m}, x_2 - \frac{1}{2000}\sum_{m=1}^{2000} x_{2m}, x_1^2 - \frac{1}{2000}\sum_{m=1}^{2000} x_{1m}^2, x_2^2 - \frac{1}{2000}\sum_{m=1}^{2000} x_{2m}^2)^t, \varepsilon \sim N(0, 1)$.

We compared the following three methods:

- REG: Calculate $E[y]$ based on MLE, where the regression function is *correctly specified* as linear.
- IPW: inverse probability weighting estimator based on the propensity score by the method of Nevo (2003), where the missing mechanism is *misspecified* as liner.
- PROP-DR: the proposed doubly robust estimation method, where the missing mechanism is *misspecified* as liner.

Table 5: Average of estimates $\theta = E[y]$ when the assignment mechanism is misspecified

	True value	REG	IPW	PROP-DR
mean	-0.002	-0.003	-0.046	-0.004
SD		0.065	0.16	0.068
MSE		0.0027	0.026	0.0032
MSEratio		85.71	820.19	100.00

Table 5 reports the simulation results when the logit of the missing mechanism is misspecified as linear; however, in the data-generating model, the logit is set to be quadratic. The true value is the sample mean of y in random sampling data. Although REG provides the best estimates because the regression model is correctly specified, the mean squared error (MSE) of the proposed estimator is almost the same as REG even though the missing mechanism is misspecified.

[14] We set $\boldsymbol{\alpha} = \begin{pmatrix} 0.4 & 0.1 & 0.1 & -0.4 & 0.6 \end{pmatrix}$ and $\boldsymbol{\beta} = \begin{pmatrix} 0 & 0.8 & -0.8 \end{pmatrix}^t$.

Estimation of population parameters: logistic function

We also conduct a simulation study to estimate the population parameters of logistic functions. Let be $\mathbf{y} = (y_1, y_2)^t$, $y_1 = (0, 1)$, $y_2 \in \mathbb{R}$. The goal is to estimate population parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)^t$ for logistic function $Pr(y_1 = 1|y_2) = \frac{\exp(\theta_1+\theta_2 y_2)}{1+\exp(\theta_1+\theta_2 y_2)}$ and the estimating equation is the score function of logistic regression $\psi(\mathbf{y}|\boldsymbol{\theta}) = \left(y_1 - \frac{1}{1+\exp(-\theta_1-\theta_2 y_2)}\right)\mathbf{y_2}$, where $\mathbf{y_2} = (1, y_2)^t$. Let $x \in \mathbb{R}$ be a one-dimensional covariate. We consider the case in which we have biased sampling data $\{x_i, \mathbf{y}_i\}_{i=1}^n$ and auxiliary moment information $E[y_1] = Pr(y_1 = 1), E[x|y_1 = 0], E[x^2|y_1 = 0], E[x|y_1 = 1], E[x^2|y_1 = 1]$[15]. We generate 1,000 datasets in which the total sample size $(M + N)$ is 8,000 with an arbitrarily set regression model and missing mechanism. Then, 5,000 $(= M)$ observations are divided as random sampling data among the 8,000 observations and we use their moments as the auxiliary information. The remaining 3,000 $(= N)$ observations are divided as candidates of biased sampling data. Whether each unit is observed in the biased sampling data is determined by the following probabilistic missing mechanism (logistic function). The detail of data generation is explained below.

To investigate the properties of the proposed doubly robust-type estimator, we conduct simulations in the following two conditions:

(i) the case in which the regression model is misspecified but the missing mechanism is correctly specified.

(ii) the case in which the missing mechanism is misspecified but the regression model is correctly specified.

First, we consider the case in which the regression model is misspecified[16]. We generate the data by $p(y_2, x|y1 = j) = p(x|y_2, y_1 = j)p(y_2|y_1 = j), for\ j = 0, 1$,[17] where $p(y_2|y_1 = j)$ are normal distributions $N(4, 1)$ and $N(3, 1)$. The regression models $p(x|y_2, y_1 = j), for\ j = 0, 1$ are set as the same quadratic function $y_2 = (1, x, x^2)\boldsymbol{\beta} + \varepsilon$ and the logit of the missing mechanism is set as a linear function $p(r = 1|x, \boldsymbol{\alpha}) = \mathbf{h}(x)^t\boldsymbol{\alpha} + \varepsilon$ [18], where $\mathbf{h}(x) = (1, x - \frac{1}{5000}\sum_{m=1}^{5000} x_m)^t$, $\varepsilon \sim N(0, 1)$.

We compared the following three methods:

- REG: Because conditional probability distributions $p(x|y_1 = j)$ are normal distributions and we have moment information $E[y_1], E[x|y_1 = j], E[x^2|y_1 = j], for\ j = 0, 1$, we can draw $x|y_1 = j$ randomly from $p(x|y_1 = j)$. Calculate $\boldsymbol{\beta}$ based on MLE, where the regression functions $p(y_2|x, y_1 = j) for\ j = 0, 1$ is *misspecified* as linear. Using the estimated regression model, we can conduct Monte Carlo integration for $E[\psi(\mathbf{y}|\boldsymbol{\theta})] = \int \sum_{j=0,1} \{\psi(y_1 = j, y_2|\boldsymbol{\theta})p(y_2|x, y_1 = j)p(x|y_1 = j)Pr(y_1 = j)\}\, dx$ and estimate $\boldsymbol{\theta}$ by solving this equation.

- IPW: inverse probability weighting estimator based on the propensity score by the method of Nevo (2003), where the missing mechanism is *correctly specified* as linear[19].

- PROP-DR: the proposed doubly robust estimation method, where the regression function is *misspecified* as liner.

[15] In this simulation, the conditional expectations $E[x|y_1 = 0], E[x^2|y_1 = 0], E[x|y_1 = 1], E[x^2|y_1 = 1]$ are supposed to be given. If this situation is strange in practical data, then the population parameters of logit can be estimated with auxiliary information $Pr(y_1 = 1)$ and unit-level random sampling covariate x_i.

[16] While we generate data with a quadratic regression, we estimate the parameter with a linear model.

[17] If we generate data by $p(y_2, x|y_1 = j) = p(y_2|x, y_1 = j)p(x|y_1 = j), for\ j = 0, 1$, then $p(y_2|y_1 = j)$ are not always normal distributions and the relationship between y_1 and y_2 is not always a true model. Therefore, we generate data not by $p(y_2|x, y_1 = j)p(x|y_1 = j)$ but by $p(x|y_2, y_1 = j)p(y_2|y_1 = j)$.

[18] We set $\boldsymbol{\beta} = \begin{pmatrix} -0.8 & -1.6 & 0.8 \end{pmatrix}^t$ and $\boldsymbol{\alpha} = \begin{pmatrix} 0.4 & -0.4 \end{pmatrix}^t$.

[19] We calculate $E[x] = Pr(y_1 = 0)E[x|y_1 = 0] + Pr(y_1 = 1)E[x|y_1 = 1]$ and $E[x^2] = Pr(y_1 = 0)E[x^2|y_1 = 0] + Pr(y_1 = 1)E[x^2|y_1 = 1]$ for the method of Nevo (2003).

Table 6: Average of estimates $\boldsymbol{\theta} = (\theta_1, \theta_2)^t$ when the regression model is misspecified

	True value	REG	IPW	PROP-DR
θ_1	3.502	2.778	3.533	3.017
SD		0.124	0.394	0.176
MSE		0.532	0.168	0.256
MSEratio		208.12	65.69	100.00
θ_2	-1.000	-0.779	-1.010	-0.849
SD		0.035	0.126	0.050
MSE		0.0050	0.017	0.025
MSEratio		201.72	68.62	100.00

Table 6 reports simulation results when the regression model is misspecified as linear; however, in the data-generating model, the regression function is set to be quadratic. "True value" is estimated in random sampling data. The row "mean" shows the average of estimation results from 1,000 replications, and the row "SD" shows their standard deviations. "MSE" means mean squared error and "MSEratio" shows the ratio of MSE of PROP-DR and MSE of another method when the "MSEratio" of PROP-DR is set as 100. We compare three types of estimators. Although the missing mechanism is correctly specified, the proposed estimator does not provide much larger MSE than IPW. The MSE of the proposed estimator is twice as small as that for REG.

Second, we consider the case in which the missing mechanism is misspecified[20]. $(x, y_2|y_1 = 0)$ and $(x, y_2|y_1 = 1)$ are separately generated from $N\left(\begin{pmatrix}2\\4\end{pmatrix}, \begin{pmatrix}1 & 1/4\\1/4 & 1\end{pmatrix}\right)$ and $N\left(\begin{pmatrix}1\\3\end{pmatrix}, \begin{pmatrix}1 & \frac{1}{4}\\\frac{1}{4} & 1\end{pmatrix}\right)$. This data-generation makes $p(y_2|y_1 = j) for\ j = 0, 1$ to be a normal distribution and $p(y_1|y_2)$ to be a logistic regression. Moreover, the regression models $p(y_2|x, y_1 = j) for\ j = 0, 1$ are linear. To decide whether a unit is missing or observed, we set the logit of the missing mechanism as a quadratic function $p(r = 1|x, \boldsymbol{\alpha}) = \mathbf{h}(x)^t\boldsymbol{\alpha} + \varepsilon$ [21], where $\mathbf{h}(x) = (1, x - \frac{1}{5000}\sum_{m=1}^{5000} x_m, x^2 - \frac{1}{5000}\sum_{m=1}^{5000} x_m^2)^t, \varepsilon \sim N(0, 1)$.

We compared the following three methods:

- REG: Because conditional probability distributions $p(x|y_1 = j)$ are normal distributions and we have moment information $E[y_1], E[x|y_1 = j], E[x^2|y_1 = j] for\ j = 0, 1$, we can draw $x|y_1 = j$ randomly from $p(x|y_1 = j)$. Calculate $\boldsymbol{\beta}$ based on MLE, where the regression functions $p(y_2|x, y_1 = j) for\ j = 0, 1$ is *correctly specified* as linear. Using the estimated regression model, we can conduct Monte Carlo integration for $E[\psi(\mathbf{y}|\boldsymbol{\theta})] = \int \sum_{j=0,1} \{\psi(y_1 = j, y_2|\boldsymbol{\theta}) p(y_2|x, y_1 = j) p(x|y_1 = j) Pr(y_1 = j)\}\, dx$ and estimate $\boldsymbol{\theta}$ by solving this equation.

- IPW: inverse probability weighting estimator based on the propensity score by the method of Nevo (2003), where the missing mechanism is *misspecified* as linear[22].

- PROP-DR: the proposed doubly robust estimation method, where the missing mechanism is *misspecified* as linear.

Table 7 reports the simulation results when the logit of the missing mechanism is misspecified as linear; however, in the data-generating model, the logit is set to be quadratic. The true value is estimated in random sampling data. Although REG provides the best estimates because

[20] While we generate data with quadratic missing mechanism, we estimate the parameter with the linear missing mechanism.

[21] We set $\boldsymbol{\alpha} = \begin{pmatrix} -0.4 & 0.2 & 0.2 \end{pmatrix}^t$.

[22] We calculate $E[x] = Pr(y_1 = 0)E[x|y_1 = 0] + Pr(y_1 = 1)E[x|y_1 = 1]$ and $E[x^2] = Pr(y_1 = 0)E[x^2|y_1 = 0] + Pr(y_1 = 1)E[x^2|y_1 = 1]$ for the method of Nevo (2003).

Table 7: Average of estimates $\boldsymbol{\theta} = (\theta_1, \theta_2)^t$ when the missing mechanism is misspecified

	True value	REG	IPW	PROP-DR
θ_1	3.502	3.508	4.000	3.508
SD		0.253	0.259	0.255
MSE		0.076	0.328	0.077
MSEratio		98.83	426.12	100.00
θ_2	-1.001	-1.002	-1.055	-1.002
SD		0.071	0.072	0.072
MSE		0.006	0.009	0.006
MSEratio		98.78	149.90	100.00

the regression model is correctly specified, the mean squared error (MSE) of the proposed estimator is almost the same as for REG in both θ_1 and θ_2 even though the missing mechanism is misspecified.

References

[1] Bang, Heejung, and James M. Robins. "Doubly robust estimation in missing data and causal inference models." Biometrics 61.4 (2005): 962-973.

[2] Cameron, A. Colin, and Pravin K. Trivedi. Microeconometrics: methods and applications. Cambridge university press, 2005.

[3] Elkan, Charles, and Keith Noto. "Learning classifiers from only positive and unlabeled data." Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008.

[4] Heckman, James. "Shadow prices, market wages, and labor supply." Econometrica: Journal of the econometric society (1974): 679-694.

[5] Heckman, James J. "Sample Selection Bias as a Specification Error." Econometrica: Journal of the Econometric Society (1979): 153-161.

[6] Hirano, Keisuke, et al. "Combining panel data sets with attrition and refreshment samples." Econometrica 69.6 (2001): 1645-1659.

[7] Hoshino, Takahiro. "Doubly robust-type estimation for covariate adjustment in latent variable modeling." Psychometrika 72.4 (2007): 535-549.

[8] Ibrahim, Joseph G., Stuart R. Lipsitz, and M- H. Chen. "Missing covariates in generalized linear models when the missing data mechanism is non- ignorable." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61.1 (1999): 173-190.

[9] Lee, Jaeyong, and James O. Berger. "Semiparametric Bayesian analysis of selection models." Journal of the American Statistical Association 96.456 (2001): 1397-1409.

[10] Liang, Hua, et al. "Estimation in partially linear models with missing covariates." Journal of the American Statistical Association 99.466 (2004): 357-367.

[11] Nevo, Aviv. "Using weights to adjust for sample selection when auxiliary information is available." Journal of Business & Economic Statistics 21.1 (2003): 43-52.

[12] Qin, Jing. "Empirical likelihood in biased sample problems." The Annals of Statistics (1993): 1182-1196.

[13] Qin, Jing. Biased Sampling, Over-identified Parameter Problems and Beyond. Singapore: Springer, 2017.

[14] Qin, Jing, Denis Leung, and Jun Shao. "Estimation with survey data under nonignorable nonresponse or informative sampling." Journal of the American Statistical Association 97.457 (2002): 193-200.

[15] Qin, Jing, Jun Shao, and Biao Zhang. "Efficient and doubly robust imputation for covariate-dependent missing responses." Journal of the American Statistical Association 103.482 (2008): 797-810.

[16] Robins, James M., 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 (1994): 846-866.

[17] Robinson, Peter M. "Root-N-consistent semiparametric regression." Econometrica: Journal of the Econometric Society (1988): 931-954.

[18] Rosenbaum, Paul R., and Donald B. Rubin. "The central role of the propensity score in observational studies for causal effects." Biometrika 70.1 (1983): 41-55.

[19] Rubin, Donald B. "The use of propensity scores in applied Bayesian inference." Bayesian statistics 2 (1985): 463-472.

[20] Speckman, Paul. "Kernel smoothing in partial linear models." Journal of the Royal Statistical Society. Series B (Methodological) (1988): 413-436.

[21] Wang, Lu, Andrea Rotnitzky, and Xihong Lin. "Nonparametric regression with missing outcomes using weighted kernel estimating equations." Journal of the American Statistical Association 105.491 (2010): 1135-1146.

[22] Ward, Gill, et al. "Presence-only data and the EM algorithm." Biometrics 65.2 (2009): 554-563.