

## Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics

### Statistical Inference for a Finite Population Mean with Machine Learning-Based Imputation for Missing Survey Data

by Mehdi Dagdoug and David Haziza

Release date: September 8, 2025



Statistics  
Canada

Statistique  
Canada

Canada

# Statistical Inference for a Finite Population Mean with Machine Learning-Based Imputation for Missing Survey Data <sup>1</sup>

Mehdi Dagdoug and David Haziza<sup>2</sup>

## Abstract

National statistical offices have increasingly adopted machine learning (ML) for its potential to improve survey estimates. ML techniques offer significant advantages, notably the ability to manage high-dimensional data and to capture complex, nonlinear relationships, thereby enhancing the overall quality of survey statistics. In this article, following the approach of Chernozhukov et al. (2018), we describe a double debiased machine learning framework that enables valid statistical inference when imputed estimators are derived from ML procedures. Simulation results suggest that the proposed framework performs well in a wide range of scenarios.

Key Words: Cross-fitting; Debiased machine learning; Doubly robust estimation; Item nonresponse; Variance estimation.

## 1. Introduction

Machine learning (ML) has gained significant attention in national statistical offices for its potential to enhance the quality of survey estimates. ML techniques offer several advantages, including the ability to handle high-dimensional data and robustness to complex, nonlinear relationships. In particular, ML procedures may prove useful for the treatment of missing data. This is the focus of this article. Despite these strengths, the literature on statistical inference when the imputed values are derived from ML procedures remains relatively sparse. A notable exception is Dagdoug et al. (2025), who studied the problem of imputation through regression trees and random forests.

In this paper, we are interested in estimating the finite population mean of a survey variable  $Y$ , which is subject to missing values. If a parametric method (e.g., linear regression) is used to derive a set of imputed values, the resulting imputed estimator of the mean is root- $n$  consistent provided that the first moment of the imputation model is correctly specified. In the presence of misspecification of the first moment, the imputed estimator may exhibit significant bias, even asymptotically.

Turning to ML methods, while they are robust to model misspecification, imputed estimators obtained from these approaches do not typically achieve the parametric rate of convergence  $1/\sqrt{n}$ . Many ML methods, including those based on regression trees, neural networks, or kernel methods, converge at rates slower than  $1/\sqrt{n}$ , often at rates that depend on the smoothness and dimensionality of the function being estimated. This slow convergence rate is the price to pay for making weaker assumptions. As a result, developing statistical inference tools may prove more challenging. Indeed, point estimators derived from ML procedures may suffer from small sample bias in the presence of overfitting. Also, deriving variance estimators that account for sampling and nonresponse for arbitrary ML procedures may prove very challenging. Finally, since imputed estimators converge at slower rates, they do not naturally satisfy the conditions for asymptotic normality required by the Central Limit Theorem. This makes it difficult to justify  $100(1 - \alpha)\%$  normal-based confidence intervals in the same way as for estimators derived from parametric methods.

---

<sup>1</sup>This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

<sup>2</sup>Mehdi Dagdoug, McGill University, Montreal, Quebec; David Haziza, University of Ottawa, Ottawa, Ontario

In a seminal paper, Chernozhukov et al. (2018) showed that the challenges associated with ML can be overcome when these methods are combined with techniques such as double machine learning. This is where doubly robust (DR) estimators play a crucial role. DR estimators offer two opportunities to mitigate potential nonresponse bias by relying on two separate models: (i) the outcome regression model, which predicts the survey variable based on a set of fully observed covariates, and (ii) the propensity score model, which estimates the probability of response given the covariates. A key advantage of doubly robust estimators is that they remain consistent if at least one of these models is correctly specified see, e.g., Robins et al. (1994), Haziza and Rao (2006), Kang and Schafer (2008), Cao et al. (2009), Kim and Haziza (2014). Moreover, when combined with ML procedures, DR estimators exhibit an additional appealing feature: they can achieve the parametric convergence rate, despite the slower rates typically observed with ML-based estimators. This property is a consequence of Neyman orthogonality—a key ingredient highlighted in Chernozhukov et al. (2018) in the context of causal effect estimation—which ensures that estimation errors from the ML-based nuisance functions (namely, the outcome and propensity score models) have only a second-order effect on the point estimator of a finite population parameter. In other words, as long as the ML models converge at a sufficiently fast rate, the DR estimator retains the root- $n$  consistency necessary for valid inference. The combination of DR estimation and cross-fitting further reduces overfitting bias, thereby allowing for asymptotically normal inference.

## 2. Statistical inference based on a single imputation model

Consider a finite population  $U = \{1, \dots, k, \dots, N\}$  of size  $N$ . We are interested in estimating the finite population mean of a survey variable  $Y$ ,  $\mu = N^{-1} \sum_{k \in U} y_k$ . We select a sample  $S$ , of size  $n$ , according to a sampling design  $\mathcal{F}(\mathbf{I} \mid \mathbf{Z})$ , where  $\mathbf{Z}$  denotes the matrix of design information available prior to sampling for all the population units, and  $\mathbf{I} = (I_1, \dots, I_k, \dots, I_N)^\top$  denotes the  $N$ -vector of sample selection indicators such that  $I_k = 1$  if  $k \in S$  and  $I_k = 0$ , otherwise. The first-order and second-order inclusion probabilities are defined as  $\pi_k = \mathbb{E}(I_k \mid \mathbf{Z})$  and  $\pi_{k\ell} = \mathbb{E}(I_k I_\ell \mid \mathbf{Z})$ , respectively. The full sample estimator of  $\mu$  is the Horvitz–Thompson estimator given by  $\hat{\mu}_F = N^{-1} \sum_{k \in S} w_k y_k$ , where  $w_k := \pi_k^{-1}$  denotes the design weight attached to unit  $k$ .

In practice, the variable  $Y$  is prone to missing values. Let  $r_k$  be a response indicator for unit  $k \in U$  such that  $r_k = 1$  if  $y_k$  is observed, and  $r_k = 0$  if  $y_k$  is missing. Let  $\mathbf{r}$  be the  $N$ -vector containing  $r_k$  in its  $k$ -th element. We define the set of respondents to item  $Y$  as  $S_r := \{k \in S ; r_k = 1\}$ , and the set of nonrespondents as  $S_m := \{k \in S ; r_k = 0\}$ .

Let  $\mathbf{x}_k$  denote a vector of fully observed variables associated with  $k \in U$ . We restrict our attention to non-informative sampling designs; that is,  $\mathbb{P}[I_k = 1 \mid \mathbf{x}_k, y_k] = \mathbb{P}[I_k = 1 \mid \mathbf{x}_k]$ , see e.g., Pfeiffermann and Sverchkov (2009). To make the sampling design non-informative, it suffices to include the design variables  $\mathbf{Z}$  in the imputation model, if appropriate. The vectors  $[r_k, y_k, \mathbf{x}_k^\top]_{k \in U}$  are assumed to be independent and identically distributed (i.i.d.). We further assume that the nonresponse model satisfies: (i) The positivity assumption, i.e., there exists  $\delta > 0$  such that  $\mathbb{P}[r_k = 1 \mid \mathbf{x}_k] > \delta$ , almost surely; (ii) The Missing At Random (MAR) assumption (Rubin, 1976), i.e.,  $\mathbb{P}[r_k = 1 \mid y_k, \mathbf{x}_k] = \mathbb{P}[r_k = 1 \mid \mathbf{x}_k] := p(\mathbf{x}_k)$  where  $p$  denotes the propensity score function; (iii) The strong invariance property (Beaumont and Haziza, 2016), which implies that  $\mathbb{P}[r_k = 1 \mid \mathbf{x}_k, \mathbf{I}] = \mathbb{P}[r_k = 1 \mid \mathbf{x}_k]$ .

The relationship between the survey variable and the  $\mathbf{x}$ -variables may be described as

$$y_k = m(\mathbf{x}_k) + \epsilon_k, \tag{1}$$

where  $m(\mathbf{x}) := \mathbb{E}[Y \mid \mathbf{x}]$  denotes the unknown regression function. The errors  $\epsilon_k$  in (1) are assumed to be i.i.d. random variables, satisfying  $\mathbb{E}[\epsilon_k \mid \mathbf{x}_k] = 0$  and  $\mathbb{V}(\epsilon_k \mid \mathbf{x}_k) = \sigma_k^2$ . In Sections 2.1-2.3, we revisit the problem of statistical inference for two commonly encountered imputation procedures: Linear Regression Imputation (LRI) and Regression Tree Imputation (RTI).

## 2.1. Point estimation

An estimator of  $\mu$  after imputation is defined as

$$\hat{\mu}_I = \frac{1}{N} \sum_{k \in S} w_k \tilde{y}_k, \quad (2)$$

where  $\tilde{y}_k = r_k y_k + (1 - r_k) \hat{m}(\mathbf{x}_k)$  with  $\hat{m}(\mathbf{x}_k)$  denoting the imputed value used to replace the missing  $y_k$ . The form (2) is convenient for data users because  $\hat{\mu}_I$  is expressed as a weighted mean of observed and imputed values, making it readily computable in practice.

For LRI, the imputed values  $\hat{m}(\mathbf{x}_k)$  are given by  $\hat{m}_{LR}(\mathbf{x}_k) = \mathbf{x}_k^\top \hat{\beta}_r$ , for  $k \in S_m$ , where

$$\hat{\beta}_r = \left( \sum_{k \in S} w_k r_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S} w_k r_k \mathbf{x}_k y_k$$

is the ordinary least squares estimator of  $\beta$  based on the responding units. Under mild regularity conditions, the imputed estimator (2) based on LRI is root- $n$  consistent for  $\mu$ , provided that the first moment of the model is correctly specified.

Turning to RTI, the prediction at a point  $\mathbf{x}$ , denoted by  $\hat{m}_{tree}(\mathbf{x})$ , is obtained by partitioning the predictor space—spanned by the  $\mathbf{x}$ -variables observed on the responding units—into disjoint terminal nodes according to a specified criterion. The imputed value for  $k \in S_m$  belonging to the terminal node  $A$  is then computed as the average of the  $y$ -values of the responding units belonging to the same node:

$$\hat{m}_{tree}(\mathbf{x}) := \frac{1}{n_{rA}} \sum_{k \in S_r: \mathbf{x}_k \in A} y_k, \quad \mathbf{x} \in A,$$

where  $n_{rA}$  denotes the number of respondents belonging to the terminal node  $A$ . Dagdoug et al. (2025) showed that the estimator  $\hat{\mu}_I$ , based on either RTI or random forests (RF) is mean square consistent for  $\mu$ . However, while they did not provide an explicit rate of convergence; it is generally slower than the parametric rate of  $1/\sqrt{n}$ . The same holds true for most ML methods.

## 2.2. Rate of convergence: Empirical study

To illustrate the slow convergence of nonparametric imputation, we conducted a limited simulation study. We generated 20 independent predictors  $X_1, \dots, X_{20}$ , each from a Gamma distribution with shape parameter equal to 1 and scale parameter equal to 10. We then repeated  $R = 10,000$  iterations of the following process:

- 1) A finite population  $U$  of size  $N = 20,000$  was generated, consisting of 2 survey variables,  $Y_1$  and  $Y_2$ . These variables were generated according to two models: a linear model and a nonlinear model. More specifically,

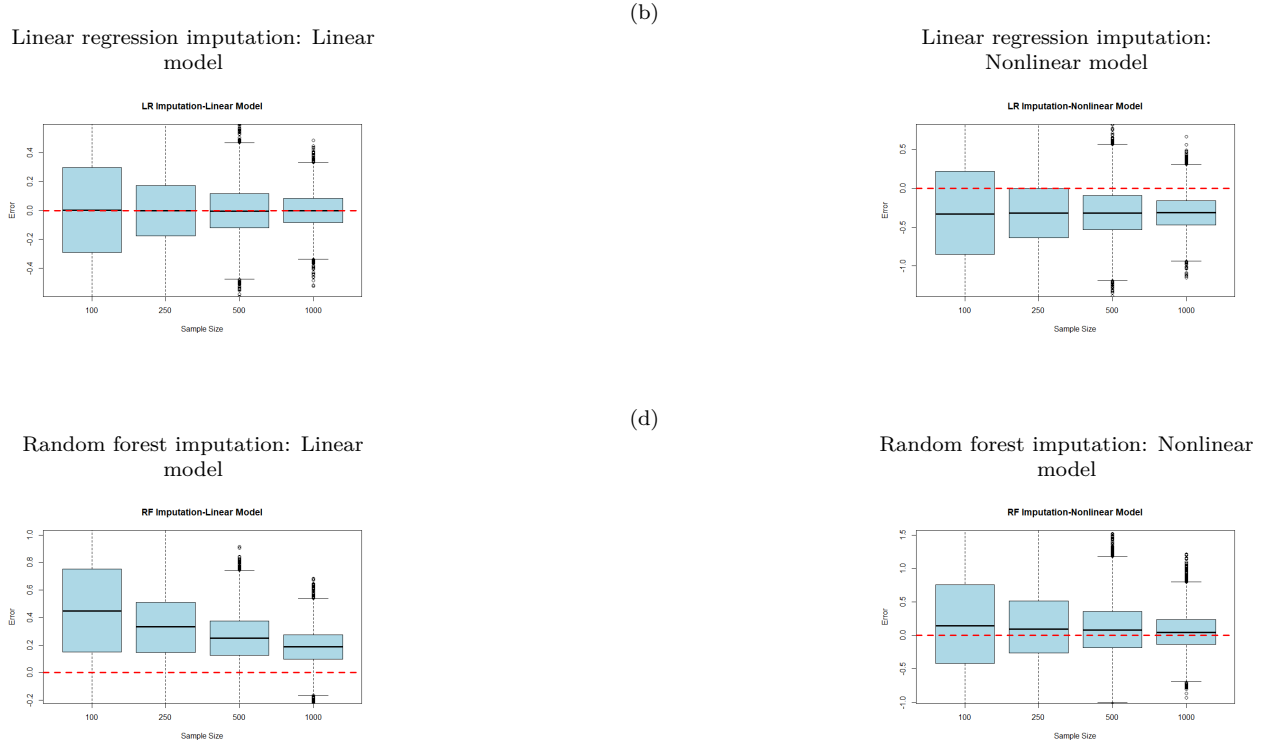
$$Y_1 = 0.15 \{(x_1 - 10) + (x_2 - 10) + (x_3 - 10)\} + \mathcal{N}(0, 4); \quad (\text{Linear model})$$

$$Y_2 = 5 \mathbb{I}(x_1 > 10) + 3 \sin(5(x_2 - 10)) + 0.02((x_1 - 10)^2 - 100) + 2 \mathbb{I}(x_2 > 10) \cos(7(x_3 - 10)) + \mathcal{N}(0, 4); \quad (\text{Nonlinear model})$$

In both models, the coefficient of determination  $R^2$  was set to 0.5.

- 2) A sample  $S$  of size  $n \in \{100; 250; 500; 1000\}$  was selected from  $U$  according to simple random sampling without replacement.
- 3) In each sample, the response indicators  $r_k$  were generated independently from a Bernoulli distribution

**Figure 2.1-1**  
**Distribution of the error of the imputed estimator across 10, 000 iterations**



with probability

$$p(\mathbf{x}_k) = \frac{1}{2} \left\{ 1 + \frac{0.48 + \sin\left(\frac{x_1-10}{10}\right) + \sin\left(\frac{x_2-10}{10}\right) + \sin\left(\frac{x_3-10}{10}\right)}{3} \right\}, \quad k \in U. \quad (3)$$

The parameters in (3) were set to obtain a proportion of missing values approximately equal to 50% in each sample.

- 4) In each sample, we computed the imputed estimator  $\hat{\mu}_I$  given by (2) based on LRI and RF. For RF, the minimal number of observations in each terminal node and the number of trees were set to  $n_0 = 10$  and  $B = 250$ , respectively. The full set of predictors  $X_1, \dots, X_{20}$  were used in the model fitting procedures.
- 5) In each sample, we computed the error  $\hat{\mu}_I - \mu$  for each value of  $n$  under both LR and RF. The distribution of  $\hat{\mu}_I - \mu$  for the different scenarios is displayed in Figure 2.1-1.

From Figure 2.1-1, when the true model is linear, LR imputation yields nearly unbiased estimators, with errors centered around zero and variance decreasing as the sample size increases. In contrast, RF imputation exhibits bias under the linear model, especially for smaller sample sizes. Both the bias and the variance decrease as the sample size increases, suggesting that the imputed estimator under RF imputation is consistent. However, we note that the rate of convergence appears to be rather slow. When the true model is nonlinear, RF imputation performs well, producing nearly unbiased results with decreasing variance as the sample size increases, whereas LR imputation consistently underestimates, with a bias that remains roughly constant and a variance that shrinks with larger samples. This is not surprising, as the imputed estimator under LR imputation is inconsistent in the nonlinear setting due to misspecification of the first moment.

### 2.3. Variance estimation

Several approaches have been proposed in the literature for estimating the variance of  $\hat{\mu}_I$ ; see Haziza and Vallée (2020) for an overview of variance estimation procedures for imputed survey data. In this article, we consider the so-called reverse approach originally proposed by Fay (1991) and Shao and Steel (1999). The total variance of  $\hat{\mu}_I$  can be expressed as

$$\mathbb{V}_{tot}(\hat{\mu}_I) = \mathbb{E}\mathbb{V}(\hat{\mu}_I | \mathbf{X}, \mathbf{r}, \mathbf{y}) + \mathbb{E}\mathbb{V}\{\mathbb{E}(\hat{\mu}_I - \mu | \mathbf{X}, \mathbf{r}, \mathbf{y}) | \mathbf{X}, \mathbf{r}\}, \quad (4)$$

where  $\mathbf{y}$  is the  $N$ -vector whose  $k$ -th element is  $y_k$ , and  $\mathbf{X}$  is the matrix whose  $k$ -th row is  $\mathbf{x}_k^\top$ . Under mild regularity conditions, the contribution to the total variance of the second term on the right hand-side of (4), is negligible if the sampling fraction  $n/N$  is negligible (Shao and Steel, 1999). In this section, we assume that  $n/N$  is negligible so that we omit the computation of the second term on the right hand-side of (4). Using a first-order Taylor expansion, an estimator of  $\mathbb{E}\mathbb{V}(\hat{\mu}_I | \mathbf{X}, \mathbf{r}, \mathbf{y})$ , denoted by  $\hat{\mathbb{V}}_1(\hat{\mu}_I)$ , is given by

$$\hat{\mathbb{V}}_1(\hat{\mu}_I) := \frac{1}{N^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\hat{\phi}_k}{\pi_k} \frac{\hat{\phi}_\ell}{\pi_\ell}, \quad (5)$$

where  $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell$ . The pseudo-value  $\hat{\phi}_k, k \in S$ , in (5) is given by

$$\hat{\phi}_k := \hat{m}(\mathbf{x}_k) + r_k \hat{C}_k \{y_k - \hat{m}(\mathbf{x}_k)\}, \quad k \in S, \quad (6)$$

where  $\hat{C}_k$  is equal to

$$\hat{C}_k = 1 + \left( \sum_{\ell \in S} w_\ell (1 - r_\ell) \mathbf{x}_\ell \right)^\top \left( \sum_{\ell \in S} w_\ell r_\ell \mathbf{x}_\ell \mathbf{x}_\ell^\top \right)^{-1} \mathbf{x}_k, \quad k \in S,$$

for LRI, and equal to

$$\hat{C}_k = \frac{\hat{N}(\mathbf{x}_k, S)}{\hat{N}(\mathbf{x}_k, S_r)}, \quad k \in S,$$

for RTI, where  $\hat{N}(\mathbf{x}, S) := \sum_{\ell \in S} \mathbf{1}_{\{\mathbf{x}_\ell \in A(\mathbf{x})\}} \pi_\ell^{-1}$  and  $\hat{N}(\mathbf{x}_k, S_r)$  is defined similarly.

We make the following remarks: (i) For RTI, the partition of the predictor space (that varies from one sample to another) was treated as fixed in the first-order Taylor expansion procedure. (ii) The variance estimator (5) corresponds the customary full sample Horvitz–Thompson variance estimator applied to the pseudo-values  $\hat{\phi}_k$ . (iii) The term  $\hat{C}_k$  in (6) may be interpreted as an implicit estimate of the response propensity  $p(\mathbf{x}_k)$ , since under both LRI and RTI, the imputed estimator  $\hat{\mu}_I$  given by (2) can be written in the form of a propensity score adjusted estimator, i.e.,  $\hat{\mu}_I = \sum_{k \in S_r} (w_k / \hat{C}_k) y_k$ .

As illustrated in Dagdoug et al. (2025) in the context of regression trees and random forests, the estimator  $\hat{\mathbb{V}}_1(\hat{\mu}_I)$  may substantially underestimate the true variance,  $\mathbb{E}\mathbb{V}(\hat{\mu}_I | \mathbf{X}, \mathbf{r}, \mathbf{y})$ , in the presence of overfitting. Overfitting can produce artificially small residuals,  $y_k - \hat{m}(\mathbf{x}_k)$ , leading to variance estimators that are biased downward. In the case of LRI, overfitting may occur when the number of predictors is large relative to the sample size. In the case of RTI, overfitting may occur when the tree grows too deep or makes too many splits relative to the available data. To address this issue, the residuals  $y_k - \hat{m}(\mathbf{x}_k)$  are replaced with “honest residuals” computed on a separate test set. Dagdoug et al. (2025) proposed constructing these honest residuals by using an  $L$ -fold cross-validation procedure, described as follows:

1. Divide the treated sample  $S$  into  $L$  mutually exclusive folds,  $\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_L$ . Each fold  $\mathcal{F}_j$  will serve as the test set once,  $j = 1, \dots, L$ .
2. For each fold  $j = 1, \dots, L$ , define the training set as  $S_{(-j)} = S \setminus \mathcal{F}_j$ . Use the data in  $S_{(-j)}$  to fit the outcome regression model.

3. For each unit  $k$  in the test set  $\mathcal{F}_j$ , obtain the prediction  $\widehat{m}_{(-j)}(\mathbf{x}_k)$ .
4. Compute the honest residuals for the units in the test set as  $e_{k,cv} = y_k - \widehat{m}_{(-j)}(\mathbf{x}_k)$ , for  $k \in \mathcal{F}_j$ .

Dagdoug et al. (2025) showed empirically that, in the case of imputation through regression trees, the estimator  $\widehat{\mathbb{V}}_1(\widehat{\mu}_I)$  based on the pseudo-values

$$\widehat{\phi}_{k,cv} := \widehat{m}(\mathbf{x}_k) + r_k \widehat{C}_k e_{k,cv}, \quad k \in S, \quad (7)$$

led to a substantial improvement in terms of bias, and achieved coverage rates of normal-based confidence intervals close to the nominal level. A drawback of the approach of Dagdoug et al. (2025) is that it appears difficult to generalize to arbitrary machine learning methods. The cross-validation procedure described above is closely related to cross-fitting (e.g., Chernozhukov et al., 2018; Wager and Athey, 2018); see Section 3.3.

### 3. A double debiased imputation framework

In this section, we extend the double debiased machine learning framework of Chernozhukov et al. (2018) to survey sampling.

#### 3.1. Oracle doubly robust estimation

We begin by considering an oracle (unfeasible) doubly robust estimator (or AIPW estimator) of  $\mu$  based on the true quantities  $m(\mathbf{x}_k)$  and  $p(\mathbf{x}_k)$ . It is defined as

$$\begin{aligned} \widehat{\mu}_{aipw}(m, p) &= \frac{1}{N} \left\{ \sum_{k \in S} w_k m(\mathbf{x}_k) + \sum_{k \in S_r} w_k \frac{y_k - m(\mathbf{x}_k)}{p(\mathbf{x}_k)} \right\}, \\ &= \frac{1}{N} \sum_{k \in S} w_k \eta_k, \end{aligned} \quad (8)$$

where

$$\eta_k = m(\mathbf{x}_k) + \frac{r_k}{p(\mathbf{x}_k)} \{y_k - m(\mathbf{x}_k)\}, \quad \text{for } k \in S. \quad (9)$$

It is easily seen that  $\widehat{\mu}_{aipw}(m, p)$  is unbiased for  $\mu$ , in the sense  $\mathbb{E}\{\widehat{\mu}_{aipw}(m, p) - \mu\} = 0$ , where  $\mathbb{E}(\cdot)$  denotes the expectation with respect to the joint distribution induced by the outcome regression model, the sampling design, and the propensity score model. Moreover, under mild regularity conditions,  $\widehat{\mu}_{aipw}(m, p)$  is root- $n$  consistent for  $\mu$ . Now, using the variance decomposition (4), the total variance of  $\widehat{\mu}_{aipw}(m, p)$  can be expressed as

$$\begin{aligned} \mathbb{V}_{tot}(\widehat{\mu}_{aipw}(m, p)) &:= \mathbb{E} \left\{ \frac{1}{N^2} \sum_{k \in U} \sum_{\ell \in U} \Delta_{k\ell} \frac{\eta_k}{\pi_k} \frac{\eta_\ell}{\pi_\ell} \right\} + \frac{1}{N^2} \sum_{k \in U} \frac{1 - p(\mathbf{x}_k)}{p(\mathbf{x}_k)} \sigma_k^2 \\ &\equiv \mathbb{V}_1(\widehat{\mu}_{aipw}(m, p)) + \mathbb{V}_2(\widehat{\mu}_{aipw}(m, p)). \end{aligned} \quad (10)$$

Under mild regularity conditions, the first term on the right hand-side of (10) is  $O(1/n)$ , whereas the second term is  $O(1/N)$ . Therefore, the contribution of the second term to the total variance is  $O(n/N)$ , which is negligible when the sampling fraction  $n/N$  is negligible. Assuming that  $\mathbb{V}(\epsilon_k | \mathbf{x}_k) = \sigma_k^2$  is known, an unbiased estimator of  $\mathbb{V}_{tot}(\widehat{\mu}_{aipw}(m, p))$  can be obtained by estimating each term in (10) without bias, leading to

$$\widehat{\mathbb{V}}_{tot}(\widehat{\mu}_{aipw}(m, p)) := \frac{1}{N^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\eta_k}{\pi_k} \frac{\eta_\ell}{\pi_\ell} + \frac{1}{N^2} \sum_{k \in S_r} w_k \frac{1 - p(\mathbf{x}_k)}{\{p(\mathbf{x}_k)\}^2} \sigma_k^2. \quad (11)$$

Although both (8) and (11) are unfeasible in practice, they serve as the foundation for the developments in Sections 3.2 and 3.3. Finally, under some regularity conditions (e.g., Chen and Rao, 2007), it can be shown

that

$$\left\{ \widehat{\mathbb{V}}_{tot}(\widehat{\mu}_{aipw}(m, p)) \right\}^{-1/2} (\widehat{\mu}_{aipw}(m, p) - \mu) \rightarrow \mathcal{N}(0, 1).$$

### 3.2. Doubly robust estimation based on estimated nuisance functions

As an estimator of  $\mu$ , we consider a DR estimator of the form

$$\begin{aligned} \widehat{\mu}_{aipw}(\widehat{m}, \widehat{p}) &= \frac{1}{N} \left\{ \sum_{k \in S} w_k \widehat{m}(\mathbf{x}_k) + \sum_{k \in S_r} w_k \frac{y_k - \widehat{m}(\mathbf{x}_k)}{\widehat{p}(\mathbf{x}_k)} \right\}, \\ &= \frac{1}{N} \sum_{k \in S} w_k \widehat{\eta}_k, \end{aligned} \quad (12)$$

where

$$\widehat{\eta}_k = \widehat{m}(\mathbf{x}_k) + \frac{r_k}{\widehat{p}(\mathbf{x}_k)} \{y_k - \widehat{m}(\mathbf{x}_k)\}, \quad \text{for } k \in S, \quad (13)$$

with  $\widehat{m}(\mathbf{x}_k)$  denoting the predicted value of the missing  $y_k, k \in S$ , based on some ML method, and  $\widehat{p}(\mathbf{x}_k)$  the estimated response probability for  $k \in S_r$  based on another ML method. Since neither  $\widehat{m}(\mathbf{x})$  nor  $\widehat{p}(\mathbf{x})$  is of primary interest, they can be viewed as nuisance functions, serving as intermediate steps toward the ultimate goal of estimating  $\mu$ .

Assuming that the sampling fraction  $n/N$  is negligible, an estimator of  $\mathbb{V}_{tot}(\widehat{\mu}_{aipw}(\widehat{m}, \widehat{p}))$ , denoted by  $\widehat{\mathbb{V}}_1(\widehat{\mu}_{aipw}(\widehat{m}, \widehat{p}))$ , is given by the first term on the right hand-side of (11) with  $\eta_k$  replaced by  $\widehat{\eta}_k$  given by (13). However,  $\widehat{\mu}_{aipw}(\widehat{m}, \widehat{p})$  in (12) and its associated variance estimator,  $\widehat{\mathbb{V}}_1(\widehat{\mu}_{aipw}(\widehat{m}, \widehat{p}))$ , suffer from a significant drawback: They both involve the residuals,  $y_k - \widehat{m}(\mathbf{x}_k), k \in S_r$ , making them vulnerable to small-sample bias in the presence of overfitting. To address this issue, it is advisable to combine DR estimation with cross-fitting for both point and variance estimation. This is discussed in the next section.

### 3.3. Doubly robust estimation based on cross-fitting

Cross-fitting proceeds according to the following steps:

1. Divide the sample  $S$  into  $L$  mutually exclusive folds,  $\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_L$ . Each fold  $\mathcal{F}_j$  will serve as the test set once,  $j = 1, \dots, L$ .
2. For each fold  $j = 1, \dots, L$ , define the training set as  $S_{(-j)} = S \setminus \mathcal{F}_j$ . Use the data in  $S_{(-j)}$  to estimate the outcome regression model  $m$  with  $\widehat{m}_{(-j)}$  and the propensity score model  $p$  with  $\widehat{p}_{(-j)}$ .
3. For each unit  $k$  in the test set  $\mathcal{F}_j$ , obtain the cross-fitted predictions  $\widehat{m}_{(-j)}(\mathbf{x}_k)$  and  $\widehat{p}_{(-j)}(\mathbf{x}_k)$ .

A cross-fitted version of  $\widehat{\mu}_{aipw}(\widehat{m}, \widehat{p})$  is given by

$$\begin{aligned} \widehat{\mu}_{aipw.cf}(\widehat{m}, \widehat{p}) &= \frac{1}{N} \left\{ \sum_{k \in S} w_k \widehat{m}_{(-j)}(\mathbf{x}_k) + \sum_{k \in S_r} w_k \frac{y_k - \widehat{m}_{(-j)}(\mathbf{x}_k)}{\widehat{p}_{(-j)}(\mathbf{x}_k)} \right\} \\ &= \frac{1}{N} \sum_{k \in S} w_k \widehat{\eta}_{k,cf}, \end{aligned} \quad (14)$$

where

$$\widehat{\eta}_{k,cf} = \widehat{m}_{(-j)}(\mathbf{x}_k) + \frac{r_k}{\widehat{p}_{(-j)}(\mathbf{x}_k)} \{y_k - \widehat{m}_{(-j)}(\mathbf{x}_k)\}, \quad \text{for } k \in S. \quad (15)$$

Unlike the estimator (2) derived from a single ML method, the doubly robust estimator  $\widehat{\mu}_{aipw.cf}(\widehat{m}, \widehat{p})$  achieves root- $n$  consistency, even if both  $\widehat{m}(\mathbf{x})$  and  $\widehat{p}(\mathbf{x})$  converge at a slower rate. For instance, root- $n$  consistency is achieved if both  $\widehat{m}(\mathbf{x})$  and  $\widehat{p}(\mathbf{x})$  converge to the corresponding targets at a rate faster than  $n^{-1/4}$ . Theorem 1 below shows that the cross-fitted estimator (14) is asymptotically equivalent to the oracle

estimator (8), provided that the estimated nuisance functions  $\widehat{m}(\mathbf{x}_k)$  and  $\widehat{p}(\mathbf{x}_k)$  converge to their target "fast enough". As a result, the cross-fitted estimator  $\widehat{\mu}_{aipw,cf}(\widehat{m}, \widehat{p})$  inherits the properties of its oracle version. In particular, the  $\widehat{\mu}_{aipw,cf}(\widehat{m}, \widehat{p})$  is asymptotically normal with a variance asymptotically equal to  $\mathbb{V}_{tot}(\widehat{\mu}_{aipw}(m, p))$ ; see Expression (10).

To establish the theoretical properties of  $\widehat{\mu}_{aipw,cf}(\widehat{m}, \widehat{p})$ , we consider the asymptotic framework of Isaki and Fuller (1982). We consider an increasing sequence of finite populations  $\{U_v\}_{v \in \mathbb{N}}$ , of sizes  $\{N_v\}_{v \in \mathbb{N}}$ , such that  $U_v \subset U_{v+1}$ , for all  $v \in \mathbb{N}$ . From  $U_v$ , a sample  $S_v$  is selected according to the sampling design  $\mathcal{P}_v$  to estimate  $\mu_v$ . The first and second-order inclusion probabilities of  $\mathcal{P}_v$  are denoted by  $\{\pi_{k,v}\}_{k \in U_v}$  and  $\{\pi_{kl,v}\}_{k \neq l \in U_v}$ , respectively. For ease of notation, the subscript  $v$  will be omitted whenever possible. We consider the following assumptions on the sampling design.

(H1) The sampling fraction satisfies  $\lim_{v \rightarrow \infty} n_v/N_v := \pi_* > 0$ .

(H2) There exists  $\lambda > 0$  such that, for all  $v \in \mathbb{N}$ ,  $\min_{k \in U_v} \pi_{k,v} > \lambda$ .

(H3) The sampling covariances satisfy

$$\limsup_{v \rightarrow \infty} n_v \max_{\substack{k, l \in U_v \\ k \neq l}} |\pi_{kl,v} - \pi_{k,v} \pi_{l,v}| < \infty.$$

**Theorem 1.** *Consider sequences  $\{\widehat{\mu}_{aipw,cf}(\widehat{m}_v, \widehat{p}_v)\}_{v \in \mathbb{N}}$  and  $\{\widehat{\mu}_{aipw,v}(m, p)\}_{v \in \mathbb{N}}$  of cross-fitted and oracle DR estimators, respectively. We assume the following conditions.*

(H1) *The sequence of estimators  $\{\widehat{m}_v\}_{v \in \mathbb{N}}$  and  $\{\widehat{p}_v\}_{v \in \mathbb{N}}$  satisfy the following rate condition:*

$$\lim_{v \rightarrow \infty} \sqrt{n_v} \times \mathbb{E} \left[ \frac{1}{|\mathcal{F}_j|} \sum_{k \in \mathcal{F}_j} (\widehat{m}_{(-j)}(\mathbf{x}_k) - m(\mathbf{x}_k))^2 \right]^{1/2} \mathbb{E} \left[ \frac{1}{|\mathcal{F}_j|} \sum_{k \in \mathcal{F}_j} (\widehat{p}_{(-j)}(\mathbf{x}_k) - p(\mathbf{x}_k))^2 \right]^{1/2} = 0, \quad (16)$$

for all  $j = 1, \dots, L$ .

(H2) *The predicted probabilities are bounded away from 0, that is, there exist  $\rho > 0$  such that, for all  $v \in \mathbb{N}$ ,  $\widehat{p}_v(\mathbf{x}_k) > \rho$ .*

Then,

$$\sqrt{n_v} \{\widehat{\mu}_{aipw,cf,v}(\widehat{m}_v, \widehat{p}_v) - \widehat{\mu}_{aipw,v}(m, p)\} = o_p(1).$$

Theorem 1 suggests that the variability associated with the estimated nuisance functions  $\widehat{m}(\mathbf{x}_k)$  and  $\widehat{p}(\mathbf{x}_k)$  can be safely ignored. As a result, an estimator of  $\mathbb{V}_1(\widehat{\mu}_{aipw,cf}(\widehat{m}, \widehat{p}))$  in (10) is obtained from the first term on the right hand-side of (11) by replacing  $\eta_k$  with  $\widehat{\eta}_{k,cf}$  given by (15). We denote the resulting variance estimator by  $\widehat{\mathbb{V}}_1(\widehat{\mu}_{aipw,cf}(\widehat{m}, \widehat{p}))$ . Notice that  $\widehat{\eta}_{k,cf}$  involves the inverse of the cross-fitted estimated response propensities,  $\widehat{p}_{(-j)}(\mathbf{x}_k)$ . Very small values of these probabilities may affect the performance of  $\widehat{\mathbb{V}}_1(\widehat{\mu}_{aipw,cf}(\widehat{m}, \widehat{p}))$ . In such cases, trimming could be applied to mitigate the impact of extremely small  $\widehat{p}_{(-j)}(\mathbf{x}_k)$ -values. However, more research is needed to determine an appropriate cut-off threshold. Alternatively, one could employ propensity stratification—also known as the score method (e.g., Little, 1986; Lunceford and Davidian, 2004; Haziza and Beaumont, 2007)—to construct  $C$  homogeneous cells with respect to  $\widehat{p}_{(-j)}(\mathbf{x}_k)$  and use the response rate within each cell in place of  $\widehat{p}_{(-j)}(\mathbf{x}_k)$  in (15). Again, this requires further investigation.

To obtain an estimator of  $\mathbb{V}_2(\widehat{\mu}_{aipw,cf}(\widehat{m}, \widehat{p}))$  in (10), we proceed as follows:

- (1) Start with the cross-fitted residuals,  $y_k - \widehat{m}_{(-j)}(\mathbf{x}_k) \equiv e_{k,cf}$  in (15).
- (2) Fit the ML procedure used to fit the outcome regression model with  $\log(e_{k,cf}^2)$  as the dependent variable and  $\mathbf{x}_k$  a the vector of predictors. Obtain the predicted values.

(3) Exponentiate the fitted values obtained in Step 2. Call them  $\tilde{e}_{k,cf}^2$ .

An estimator of  $\mathbb{V}_2(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}))$  in (10) is given by

$$\hat{\mathbb{V}}_2(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})) = \frac{1}{N^2} \sum_{k \in S_r} w_k \frac{1 - \hat{p}_{(-j)}(\mathbf{x}_k)}{\{\hat{p}_{(-j)}(\mathbf{x}_k)\}^2} \tilde{e}_{k,cf}^2. \quad (17)$$

Finally, an estimator of the total variance,  $\mathbb{V}_{tot}(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}))$ , is given by

$$\hat{\mathbb{V}}_{tot}(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})) = \hat{\mathbb{V}}_1(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})) + \hat{\mathbb{V}}_2(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})).$$

**Remark 1.** *If the response propensity  $p(\mathbf{x}_k)$  were known, we could construct a cross-fitted DR estimator of the form*

$$\hat{\mu}_{aipw,cf}(\hat{m}, p) = \frac{1}{N} \left\{ \sum_{k \in S} w_k \hat{m}_{(-j)}(\mathbf{x}_k) + \sum_{k \in S_r} w_k \frac{y_k - \hat{m}_{(-j)}(\mathbf{x}_k)}{p(\mathbf{x}_k)} \right\}, \quad (18)$$

where  $\hat{m}(\mathbf{x}_k)$  is obtained using any ML method. In this case, it is easily seen that  $\mathbb{E}(\hat{\mu}_{aipw}(\hat{m}, p) - \mu) = 0$ , regardless of the ML method used to obtain the predictions  $\hat{m}(\mathbf{x}_k)$ . Moreover, the variance of  $\hat{\mu}_{aipw,cf}(\hat{m}, p)$  is asymptotically equal to (10), as long as  $\hat{m}(\mathbf{x}_k)$  is consistent for  $m(\mathbf{x}_k)$ . In other words, no rate of convergence is required.

### 3.4. Implementation

The estimator  $\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})$  is not expressed in the convenient form (2). To address this issue, we propose the following implementation. For the responding units (i.e., the units for which  $r_k = 1$ ), we report the observed values  $y_k$  in the data file, whereas we report the modified imputed values

$$\hat{m}^*(\mathbf{x}_k) = \hat{m}_{(-j)}(\mathbf{x}_k) + \frac{1}{\sum_{\ell \in S_m} w_\ell} \sum_{\ell \in S_r} w_\ell \left\{ \frac{1 - \hat{p}_{(-j)}(\mathbf{x}_\ell)}{\hat{p}_{(-j)}(\mathbf{x}_\ell)} \right\} \{y_\ell - \hat{m}_{(-j)}(\mathbf{x}_\ell)\} \quad (19)$$

for the nonresponding units (i.e., the units for which  $r_k = 0$ ). With the modified imputed values (19), it is easy to verify that

$$\frac{1}{N} \left[ \sum_{k \in S} w_k \{r_k y_k + (1 - r_k) \hat{m}^*(\mathbf{x}_k)\} \right] = \hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}). \quad (20)$$

As a result, data users can readily obtain the estimator  $\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})$  using customary point estimation procedures.

### 3.5. Summary of the proposed approach

To apply the proposed approach, we proceed as follows:

- **Step 1:** Select a machine learning procedure for the outcome and another for the propensity score, optimize the hyperparameters of each method (using, for example, cross-validation). Optimizing the hyperparameters is an important step to ensure that Condition  $(H_1)$  given by (16) is satisfied.
- **Step 2:** Compute

$$\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}) = \frac{1}{N} \left[ \sum_{k \in S} w_k \hat{m}_{(-j)}(\mathbf{x}_k) + \sum_{k \in S_r} w_k \frac{y_k - \hat{m}_{(-j)}(\mathbf{x}_k)}{\hat{p}_{(-j)}(\mathbf{x}_k)} \right].$$

- **Step 3:** If needed, obtain the modified imputed values  $\hat{m}^*(\mathbf{x}_k)$ .

- **Step 4:** Estimate the variance of  $\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})$  by

$$\begin{aligned}\widehat{V}_{tot}(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})) &= \widehat{V}_1(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})) + \widehat{V}_2(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})) \\ &= \frac{1}{N^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\hat{\eta}_{k,cf}}{\pi_k} \frac{\hat{\eta}_{\ell,cf}}{\pi_\ell} + \frac{1}{N^2} \sum_{k \in S_r} w_k \frac{1 - \hat{p}_{(-j)}(\mathbf{x}_k)}{\{\hat{p}_{(-j)}(\mathbf{x}_k)\}^2} \tilde{e}_{k,cf}^2.\end{aligned}$$

- **Step 5:** Compute a  $1 - \alpha\%$  confidence interval for  $\mu$ :

$$\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}) \pm z_{\alpha/2} \sqrt{\widehat{V}_{tot}(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}))},$$

where  $z_{\alpha/2}$  is the  $1 - \alpha/2$  -quantile of the standard normal distribution.

## 4. Simulation study

To assess the performance of the proposed methods in terms of bias, efficiency and coverage rate, we conducted a simulation study. First, we generated 5 mutually independent predictors  $X_1, \dots, X_5$ , according to a standard normal distribution. We then repeated  $R = 10,000$  iterations of the following process:

- 1) Given the predictors  $X_1, \dots, X_5$ , we generated two finite populations with different values of  $N$ : 2,000 and 16,000. Each population consisted of two survey variables,  $Y_1$  and  $Y_2$  that were generated according to

$$\begin{aligned}Y_1 &= 1 + 2x_1 - 3x_2^2 + 1.5x_3^3 + 0.5x_4^2 - x_5^3 + \mathcal{N}(0, 1); \\ Y_2 &= 20 \times \mathbf{1}_{x_1 > 0} + 10 \sin(x_2) + 5x_3^2 + \mathcal{N}(0, 1);\end{aligned}$$

- 2) A sample  $S$  was selected from each population according to simple random sampling without replacement with a sampling fraction  $n/N = 10\%$ , which led to the following values of  $n$ : 200 and 1,600.
- 3) In each sample, the missing values to the survey variable  $Y_1$  were generated independently from a Bernoulli distribution with probability

$$p(\mathbf{x}_k) = 0.05 + 0.95 \frac{\exp\{(x_1 - 0.5) + (x_2 - 0.5) + (x_3 - 0.5) + (x_4 - 0.5) + (x_5 - 0.5)\}}{1 + \exp\{(x_1 - 0.5) + (x_2 - 0.5) + (x_3 - 0.5) + (x_4 - 0.5) + (x_5 - 0.5)\}}, \quad k \in S. \quad (21)$$

The missing values to the survey variable  $Y_2$  were generated independently from a Bernoulli distribution with probability

$$p(\mathbf{x}_k) = \frac{1}{6} \{0.5 + \sin(x_1) + \sin(x_2) + \sin(x_3)\}, \quad k \in S. \quad (22)$$

In both (21) and (22), the parameters were set to obtain a proportion of missing values approximately equal to 50% in each sample.

- 4) In each sample, we computed the doubly robust estimators  $\hat{\mu}_{aipw}(\hat{m}, \hat{p})$  or  $\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})$  given by (12) and (14), respectively based on different combinations of ML methods. For the outcome regression model, we used RF (Breiman, 2001) with 15 observations in each terminal node and  $B = 250$  trees; Cubist (Quinlan, 1992); and XGboost (Chen and Guestrin, 2016) with a learning rate set to 1, a depth set to 5, and  $B = 50$  trees. For the propensity score model, we used either regression trees with 40 observations in each terminal node and RF with 20 observations per terminal node and  $B = 250$  trees; see Tables 4-1 and 4-2 for the different combinations of outcome regression models and propensity score models.

- 5) In each sample, we computed 95%-based confidence intervals for  $\mu$  of the form  $\hat{\mu} \pm 1.96 \sqrt{\widehat{V}_{tot}(\hat{\mu})}$ , where the pair  $[\hat{\mu}, \widehat{V}_{tot}(\hat{\mu})]$  was either  $[\hat{\mu}_{aipw}(\hat{m}, \hat{p}), \widehat{V}_{tot}(\hat{\mu}_{aipw}(\hat{m}, \hat{p}))]$  or  $[\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}), \widehat{V}_{tot}(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}))]$ .

As a measure of bias of a generic point estimator  $\hat{\mu}$ , we computed its Monte Carlo percent relative bias defined as  $\text{RB}(\hat{\mu}) = 100 \times \mathbb{E}_{MC}(\hat{\mu} - \mu)/\mu$ , where  $\mathbb{E}_{MC}(\cdot)$  denotes the Monte Carlo average across the  $R$  iterations. As a measure of relative efficiency of  $\hat{\mu}$ , we computed  $\text{RE}(\hat{\mu}) = 100 \times \{\text{MSE}_{MC}(\hat{\mu})/\text{MSE}_{MC}(\hat{\mu}_F)\}$ , where  $\hat{\mu}_F$  denotes the full sample estimator (see Section 2). As a measure of the bias of a generic variance estimator  $\hat{V}(\hat{\mu})$ , we computed its Monte Carlo percent relative bias defined as

$$\text{RB}(\hat{V}(\hat{\mu})) = 100 \times \frac{\mathbb{E}_{MC}(\hat{V}(\hat{\mu})) - \mathbb{V}_{MC}(\hat{\mu})}{\mathbb{V}_{MC}(\hat{\mu})},$$

where  $\mathbb{V}_{MC}(\cdot)$  denotes the Monte Carlo variance. Finally, we computed the Monte Carlo coverage rate of 95% normal-based confidence intervals. The results are displayed in Tables 4-1 and 4-2. In both tables, note that, for a given sample size  $n$  and a given point estimator, the first row represents the Monte Carlo RB, whereas the Monte Carlo RE is reported between parentheses on the second row. Also, for a given sample size  $n$  and a given variance estimator, the first row represents the Monte Carlo RB, whereas the Monte Carlo coverage rate is reported between parentheses on the second row.

We begin by discussing the performance of the point estimators  $\hat{\mu}_{aipw}(\hat{m}, \hat{p})$  and its cross-fitted version  $\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})$ . Both exhibited negligible bias in all but one scenario. Specifically, when RF was used to estimate  $m(\mathbf{x}_k)$  and  $p(\mathbf{x}_k)$ ,  $\hat{\mu}_{aipw}(\hat{m}, \hat{p})$  and  $\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})$  showed slight bias in the case of  $Y_2$ , with values of RB equal to -2.7% and 2.6%, respectively. In terms of RE, the estimator  $\hat{\mu}_{aipw}(\hat{m}, \hat{p})$  was generally slightly more efficient than its cross-fitted version  $\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p})$  although the differences tended to vanish as the sample size increased. Turning to the performance of the variance estimators, we note that  $\hat{V}_{tot}(\hat{\mu}_{aipw}(\hat{m}, \hat{p}))$  exhibited negative bias in all scenarios. In several cases, the bias was significant with values of absolute RB exceeding 15%. In these scenarios, the coverage rate of normal-based confidence intervals fell short of the nominal rate. This poor performance is most likely due to the problem of overfitting. On the other hand, the variance estimator  $\hat{V}_{tot}(\hat{\mu}_{aipw,cf}(\hat{m}, \hat{p}))$  performed well across all scenarios, leading to coverage rates close to the nominal rate.

## 5. Final remarks

In this article, we have used a single ML method to model the survey variable  $Y$  and another ML method for the propensity score model. An alternative approach would be to use aggregation methods, which combine multiple ML algorithms to enhance predictive performance and robustness. One such method is the Super Learner algorithm, which builds an ensemble of candidate models and optimally weights them using cross-validation (van der Laan et al., 2007). This approach can be applied to both the outcome and propensity score models, potentially improving efficiency and reducing bias. Under mild conditions, the SuperLearner achieves an asymptotic rate of convergence that is at least as fast as the best candidate learner in the library of models. Investigating the properties of aggregation methods in the context of imputation for missing survey data remains an area for future research.

**Table 4-1**

Monte Carlo statistics associated with point and variance estimator for the survey variable  $Y_1$

	$\hat{\mu}_{aipw}$	$\widehat{V}(\hat{\mu}_{aipw})$	$\hat{\mu}_{aipw,cf}$	$\widehat{V}(\hat{\mu}_{aipw,cf})$
$[m(\mathbf{x}_k); p(\mathbf{x}_k)] = [\text{Cubist}; \text{Trees}]$				
$n = 200$	0.4 (162)	-4.0 (94.4)	0.4 (173)	4.0 (95.1)
$n = 1,600$	0.0 (158)	-3.7 (95.0)	0.0 (165)	-2.1 (95.8)
$[m(\mathbf{x}_k); p(\mathbf{x}_k)] = [\text{Cubist}; \text{RF}]$				
$n = 200$	0.3 (161)	-14.0 (92.8)	0.4 (166)	-2.0 (94.4)
$n = 1,600$	0.1 (148)	-12.1 (93.9)	0.1 (148)	-3.0 (94.9)
$[m(\mathbf{x}_k); p(\mathbf{x}_k)] = [\text{RF}; \text{RF}]$				
$n = 200$	0.1 (156)	-17.8 (93.2)	0.0 (160)	2.5 (95.5)
$n = 1,600$	0.1 (152)	-23.0 (91.8)	0.1 (152)	-6.4 (94.9)
$[m(\mathbf{x}_k); p(\mathbf{x}_k)] = [\text{XGboost}; \text{Trees}]$				
$n = 200$	-0.4 (203)	-53.0 (82.6)	0.2 (217)	-0.4 (95.1)
$n = 1,600$	-0.1 (195)	-50.4 (83.9)	0.0 (218)	0.4 (95.0)

**Table 4-2**

Monte Carlo statistics associated with point and variance estimator for the survey variable  $Y_2$

	$\hat{\mu}_{aipw}$	$\widehat{V}(\hat{\mu}_{aipw})$	$\hat{\mu}_{aipw,cf}$	$\widehat{V}(\hat{\mu}_{aipw,cf})$
$[m(\mathbf{x}_k); p(\mathbf{x}_k)] = [\text{Cubist}; \text{Trees}]$				
$n = 200$	0.1 (131)	-18.5 (92.1)	0.6 (144)	-0.9 (94.5)
$n = 1,600$	-0.1 (106)	-6.7 (94.0)	0.2 (110)	-3.6 (94.3)
$[m(\mathbf{x}_k); p(\mathbf{x}_k)] = [\text{Cubist}; \text{RF}]$				
$n = 200$	0.1 (113)	-25.9 (90.9)	0.4 (147)	-11.0 (93.7)
$n = 1,600$	-0.1 (106)	-4.8 (94.3)	0.1 (107)	-1.8 (94.6)
$[m(\mathbf{x}_k); p(\mathbf{x}_k)] = [\text{RF}; \text{RF}]$				
$n = 200$	2.7 (151)	-14.0 (91.1)	2.6 (156)	1.7 (93.6)
$n = 1,600$	-0.1 (112)	-8.8 (93.8)	-0.1 (113)	-0.8 (94.6)
$[m(\mathbf{x}_k); p(\mathbf{x}_k)] = [\text{XGboost}; \text{Trees}]$				
$n = 200$	-0.4 (121)	-20.2 (91.8)	0.2 (135)	-1.1 (94.7)
$n = 1,600$	-0.1 (116)	-11.3 (93.7)	0.0 (120)	0.0 (95.0)

## References

- Beaumont, J.-F. and Haziza, D. (2016), A note on the concept of invariance in two-phase sampling designs. *Survey Methodology*, **42**, pp. 319–323.
- Breiman, L. (2001), Random forests. *Machine Learning*, **45**, pp. 5–32.
- Cao, W., Tsiatis, A. A. and Davidian, M. (2009), Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, **96**, pp. 723–734.
- Chen, T. and Guestrin, C. (2016), XGBoost. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16*. ACM Press.
- Chen, J. and Rao, J.N.K. (2007), Asymptotic normality under two-phase sampling designs. *Statistica Sinica*, **17**, pp. 1047–1064.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2018), Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, **21**, pp. C1–C68.
- Dagdoug, M., Goga, C., and Haziza, D. (2023b), Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison. *Journal of Survey Statistics and Methodology*, **11**, pp. 141–188.
- Dagdoug, M., Goga, C., and Haziza, D. (2025), Statistical Inference in the presence of imputed survey data through random forests. To appear in *Scandinavian Journal of Statistics*.
- Haziza, D. and Beaumont, J. F. (2007), On the construction of imputation classes in surveys. *International Statistical Review*, **75**, pp. 25–43.
- Haziza, D. and Rao, J. N. K. (2006), A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology*, **32**, pp. 53–64.
- Haziza, D. and Vallée, A.-A. (2020). Variance estimation in the presence of singly imputed data: A critical review. *Japanese Journal of Statistics and Data Science*, **3**, pp. 583–623.
- Isaki, C.-T. and Fuller, W.-A. (1982), Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, **77**, pp. 49–61.
- Kang, J.D.Y. and Schafer, J.L. (2008). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, **22**, pp. 523–539.
- Kim, J. K. and Haziza, D. (2014), Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, **24**, pp. 375–394.
- van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007), Super learner. *Statistical Applications in Genetics and Molecular Biology*, **6**, pp. 1–23.
- Little, R. J. (1986), Survey nonresponse adjustments for estimates of means. *International Statistical Review*, **54**, pp. 139–157.
- Lunceford, J. K. and Davidian, M. (2004), Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, **23**, pp. 2937–2960.
- Pfeffermann, D. and Sverchkov, M. (2009), Inference under informative sampling. *In Handbook of Statistics*, vol. 29, pp. 455–487. Elsevier.
- Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1994), Estimation of regression coefficient when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, pp. 846–866.
- Rubin, D. B. (1976), Inference and missing data. *Biometrika*, **63**, pp. 581–592.

Shao, J. and Steel, P. (1999), Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, **94**, pp. 254–265.

Wager, S. and Athey, S. (2018), Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association*, **113**, pp. 1228–1242.