The Survey Statistician, 2023, Vol. 88, 75–85.

Ask the Experts

Doubly and Multiply Robust Procedures for Missing Survey Data

Sixia Chen[1] **and David Haziza**[2]

[1]University of Oklahoma Health Sciences Center, U. S. A., Sixia-Chen@ouhsc.edu
[2]University of Ottawa, Canada, dhaziza@uottawa.ca

Abstract

Missing data are ubiquitous in surveys. Unadjusted estimators may be substantially biased as the set of respondents is generally a non-representative subset of the original sample. Item nonresponse, which most often treated by some form of imputation, may eliminate the potential nonresponse bias if the first moment of the imputation model is correctly specified. However, the resulting estimators may suffer from appreciable bias if the model is misspecified. In this paper, we review doubly and multiply robust imputation procedures. These procedures, that combine multiple nonresponse and imputation models, may provide some protection against model miss-specification.

Keywords: Deterministic imputation; imputation model; missing data; nonresponse model; random imputation.

1 Introduction

As response rates have declined sharply over the past two decades, reducing the nonresponse bias has become an important issue for survey statisticians. Unadjusted estimators tend to exhibit significant bias as the behaviour of the respondents typically differ from that of the nonrespondents. In the absence of non-sampling errors, bias is generally not an issue as customary point estimators (e.g., the Horvitz-Thompson estimator and calibration estimators) are design-unbiased or asymptotically design-unbiased. In this ideal setup, survey statisticians would typically opt for an estimator that exhibit a small variance. In the presence of missing values, bias is the main issue. Reducing the nonresponse bias as much as possible requires the availability of powerful auxiliary information. The nonresponse treatment stage involves a modeling task that puts an additional burden on the survey statistician's shoulders, as heavily biased estimators will lead to misleading inferences. In this article, we focus on item nonresponse, most often treated by some form of imputation, the first step of which is to postulate an imputation model describing the relationship between the survey variable Y requiring imputation and a set of fully observed variables $\mathbf{x}$. The modeling task involves the selection of variables that are predictive of the survey variable Y, and the specification of a suitable model for the relationship between Y and $\mathbf{x}$.

If the data are Missing At Random (MAR, Rubin, 1976), the estimators obtained after imputation will exhibit a negligible bias if the first moment of the imputation model is correctly specified. Otherwise, the bias may be significant.

This led researchers to develop imputation/estimation procedures that provide some robustness against model misspecification. This is where doubly and multiply robust procedures come into play. The concept of double robustness in the context of missing data is attributed to Robins et al. (1994) for their article published in the *Journal of the American Statistical Association*. However, it is worth pointing out that, in the same issue of the journal, Kott (1994) also independently introduced the concept of double robustness for missing survey data. In fact, the origin of doubly robust procedures can be traced back in the 1970s to the survey sampling literature on the generalized regression (GREG) estimator and, more generally, on model-assisted estimation procedures; see, e. g., Cassel et al. (1976), Särndal (1980), Särndal and Wright (1984) and Särndal et al. (1992). For instance, the GREG estimator of a population total, whose construction is assisted by a linear regression model, possesses the double robustness property: it is model-unbiased if the model is correctly specified, but remains design-consistent even if the model is misspecified.

In the context of missing data, doubly robust procedures combine two models. The first, called the imputation model or the outcome regression model, describes the relationship between the survey variable Y and a vector of explanatory variables. The second, called the nonresponse model or the propensity score model, describes the relationship between the response indicator R and a set of explanatory variables. If the data are MAR, doubly robust procedures remain consistent if either the nonresponse model or the imputation model is correctly specified. This is an attractive property closely related to the philosophy of model-assisted estimation in survey sampling. The literature on doubly robust procedures is very rich; see e. g., Robins et al. (1994), Scharfstein et al. (1999), Bang and Robins (2005), Haziza and Rao (2006), Tan (2006), Kang and Schafer (2008), Cao et al. (2009), Kim and Haziza (2014), Boistard et al. (2016) and Seaman and Vansteelandt (2018). However, doubly robust procedures have been criticized because the resulting estimators have been shown to exhibit poor performances if both models are (slightly) misspecified; e. g., see Kang and Schafer (2008).

Multiple robustness can be viewed as an extension of the concept of double robustness. Instead of postulating a single imputation model and a single nonresponse model, one rather postulates multiple imputation models and/or multiple nonresponse models. Each model may be based on a different link function and a different set of explanatory variables. An imputation procedure is called multiple robust if the resulting estimator remains consistent if anyone of the postulated models is correctly specified; see e. g., Han and Wang (2013), Chan and Yam (2014), Han, (2014; 2016), Chen and Haziza (2017), Duan and Yin (2017), Chen and Haziza (2019) and Han et al. (2019). Therefore, these procedures provide some type of insurance against a single misspecified model. Multiply robust procedures belong to the class of ensemble or aggregation procedures as the goal is to construct a set of imputed values that can be viewed as a suitable aggregate of the information contained in the multiple models.

2 Doubly robust procedures

Consider a finite population U of size N. Our goal is to estimate the population total of a survey variable Y, $t_y = \sum_{k \in U} y_k$. A sample S, of size n, is selected from U according to a sampling design with first-order inclusion probabilities π_k. Let R_k be a response indicator attached to unit k such that $R_k = 1$ if Y is observed and $R_k = 0$, otherwise. Let S_r be the set of respondents to item Y, of size n_r; that is the subset of S for which $R_k = 1$, and let $S_m = S - S_r$ be the set of nonrespondents. We assume that the data are MAR; i.e., the conditional distribution of Y given $\mathbf{x}$ observed among the respondents

is identical to the conditional distribution of Y given $\mathbf{x}$ observed among the nonrespondents, where $\mathbf{x}$ denotes a vector of fully observed variables. Under MAR, one can safely estimate the relationship between Y and $\mathbf{x}$ from the set of respondents S_r, and "extrapolate" from this relationship to construct a set of imputed values.

We assume that the relationship between R and $\mathbf{x}$ can be described by the following nonresponse model:

$$\mathbb{E}\left(R_k \mid \mathbf{x}_k\right) = p(\mathbf{x}_k; \boldsymbol{\alpha}), \tag{1}$$

where $p(\cdot; \boldsymbol{\alpha})$ is a prespecifed function and $\boldsymbol{\alpha}$ is a vector of unknown coefficients. We assume that the relationship between Y and $\mathbf{x}$ can be described by the following imputation model:

$$\mathbb{E}\left(y_k \mid \mathbf{x}_k\right) = m(\mathbf{x}_k; \boldsymbol{\beta}), \tag{2}$$

where $m(\cdot; \boldsymbol{\beta})$ is a prespecifed function and $\boldsymbol{\beta}$ is a vector of unknown coefficients. For simplicity, we assume that the first component of the $\mathbf{x}$-vector is 1 for all k and that $\mathbb{V}(y_k \mid \mathbf{x}_k) = \sigma^2$.

Doubly robust imputation can be described as follows:

(i) We obtain an estimator, $\widehat{\boldsymbol{\alpha}}$, of $\boldsymbol{\alpha}$ by solving, for example, the following estimating equations:

$$\sum_{k \in S} \pi_k^{-1} \frac{R_k - p(\mathbf{x}_k; \alpha)}{p(\mathbf{x}_k; \alpha)\left\{1 - p(\mathbf{x}_k; \alpha)\right\}} \frac{\partial p(\mathbf{x}_k; \alpha)}{\partial \alpha} = \mathbf{0}. \tag{3}$$

In the case of a logistic regression model, $p(\mathbf{x}_k; \alpha) = \exp\left(\mathbf{x}_k^\top \alpha\right)/\{1 + \exp\left(\mathbf{x}_k^\top \alpha\right)\}$, Expression (3) reduces to the customary estimating equations

$$\sum_{k \in S} \pi_k^{-1} \{R_k - p(\mathbf{x}_k; \alpha)\} \mathbf{x}_k = \mathbf{0}.$$

Let $p(\mathbf{x}_k; \widehat{\boldsymbol{\alpha}})$ denote the resulting estimated response probability attached to unit k.

(ii) We obtain an estimator, $\widehat{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$, by solving the estimating equations

$$\sum_{k \in S_r} \pi_k^{-1} \frac{1 - p(\mathbf{x}_k; \widehat{\boldsymbol{\alpha}})}{p(\mathbf{x}_k; \widehat{\boldsymbol{\alpha}})} \left\{y_k - m(\mathbf{x}_k; \boldsymbol{\beta})\right\} \mathbf{x}_k = \mathbf{0}.$$

Let $m(\mathbf{x}_k; \widehat{\boldsymbol{\beta}})$ denote the predicted value attached to unit k. If $m(\mathbf{x}_k; \boldsymbol{\beta}) = \mathbf{x}_k^\top \boldsymbol{\beta}$, the estimator $\widehat{\boldsymbol{\beta}}$ reduces to the weighted least squares estimator of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{k \in S_r} \pi_k^{-1} \frac{1 - p(\mathbf{x}_k; \widehat{\boldsymbol{\alpha}})}{p(\mathbf{x}_k; \widehat{\boldsymbol{\alpha}})} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S_r} \pi_k^{-1} \frac{1 - p(\mathbf{x}_k; \widehat{\boldsymbol{\alpha}})}{p(\mathbf{x}_k; \widehat{\boldsymbol{\alpha}})} \mathbf{x}_k y_k.$$

(iii) The imputed y-value for $k \in S_m$ is given by

$$y_k^* = m(\mathbf{x}_k; \widehat{\boldsymbol{\beta}}).$$

It follows that an estimator of t_y after imputation is given by

$$\widehat{t}_{y,DR} = \sum_{k \in S_r} \pi_k^{-1} y_k + \sum_{k \in S_m} \pi_k^{-1} m(\mathbf{x}_k; \widehat{\boldsymbol{\beta}}).$$

The estimator $\widehat{t}_{y,DR}$ is doubly robust in the sense that it remains consistent if either the nonresponse

model (1) or the imputation model (2) is correctly specified. To see why this is the case, note that $\widehat{t}_{y,DR}$ can be expressed in the following two forms:

$$\widehat{t}_{y,DR} = \widehat{t}_{y,F} - \sum_{k\in S_m} \pi_k^{-1}\left\{y_k - m(\mathbf{x}_k;\widehat{\boldsymbol{\beta}})\right\} \quad (4)$$

$$= \widehat{t}_{y,PSA} - \sum_{k\in S} \pi_k^{-1}\left(\frac{R_k}{p(\mathbf{x}_k;\widehat{\boldsymbol{\alpha}})} - 1\right) m(\mathbf{x}_k;\widehat{\boldsymbol{\beta}}), \quad (5)$$

where $\widehat{t}_{y,F} = \sum_{k\in S} \pi_k^{-1} y_k$ denotes the (unfeasible) full sample estimator of t_y, and

$$\widehat{t}_{y,PSA} = \sum_{k\in S_r} \pi_k^{-1} \frac{y_k}{p(\mathbf{x}_k;\widehat{\boldsymbol{\alpha}})}$$

corresponds to the propensity score estimator of t_y. If the imputation model is correctly specified, we have $\mathbb{E}\left\{y_k - m(\mathbf{x}_k;\widehat{\boldsymbol{\beta}})\right\} \approx 0$ and the second term on the right hand-side of (4) is, on average, approximately equal to 0. We are left with the full sample estimator $\widehat{t}_{y,F}$, which is consistent for t_y. Next, if the nonresponse model is correctly specified, we have $\mathbb{E}\left(\frac{R_k}{p(\mathbf{x}_k;\widehat{\boldsymbol{\alpha}})} - 1\right) \approx 0$, and the second term on the right hand-side of (5) is approximately equal to 0. We are left with the propensity score adjusted estimator $\widehat{t}_{y,PSA}$, which is consistent for t_y since the nonresponse model is correctly specified.

The imputed values (2) belong to the class of deterministic imputation procedures. We can define a doubly robust random version (Haziza and Rao, 2006) as follows:

$$y_k^* = m(\mathbf{x}_k;\widehat{\boldsymbol{\beta}}) + e_k^*,$$

where e_k^* is selected at random with replacement from the set of standardized residuals observed among the respondents. That is,

$$e_k^* = e_\ell, \quad \ell \in S_r, \text{ with probability } \frac{\phi_\ell}{\sum_{t\in S_r}\phi_t},$$

where

$$\phi_\ell = \pi_\ell^{-1}\frac{1 - p(\mathbf{x}_\ell;\widehat{\boldsymbol{\alpha}})}{p(\mathbf{x}_\ell;\widehat{\boldsymbol{\alpha}})} \text{ and } e_\ell = y_k - m(\mathbf{x}_k;\widehat{\boldsymbol{\beta}}).$$

Donor imputation is often used in practice as the imputed values are necessarily eligible values observed among the respondents, which is often deemed a desirable feature when, for instance, the variable requiring imputation is categorical. A doubly robust procedure random hot-deck imputation procedure can be described as follows. We first obtain the scores $m(\mathbf{x}_k;\widehat{\boldsymbol{\beta}})$ and $p(\mathbf{x}_\ell;\widehat{\boldsymbol{\alpha}})$. Then, using a classification algorithm (e. g., the K-means algorithm), we create C homogeneous cells with respect to both $m(\mathbf{x}_k;\widehat{\boldsymbol{\beta}})$ and $p(\mathbf{x}_\ell;\widehat{\boldsymbol{\alpha}})$. Within each cell, a missing value is imputed using the y-value of a donor selected at random and with replacement from the set of donors belonging to the same cell.

3 Multiply robust imputation procedures

We consider two classes of parametric models: The first, $\mathcal{M}_1$, consists of H imputation models; i.e., $\mathcal{M}_1 = \left\{m^{(h)}(\mathbf{x}_k^{(h)};\boldsymbol{\beta}^{(h)}), h = 1,2,\ldots,H\right\}$ and, the second, $\mathcal{M}_2$, consists of J nonresponse models; i.e., $\mathcal{M}_2 = \left\{p^{(j)}(\mathbf{x}_k^{(j)};\boldsymbol{\alpha}^{(j)}), j = 1,2,\ldots,J\right\}$. The models in both classes may be based on different functionals and/or different sets of explanatory variables. Two methods for constructing a set of imputed values are: (i) Aggregation through calibration (Han and Wang, 2013; Han, 2014, 2016;

Chen and Haziza, 2017) and (ii) aggregation through refitting (Duan and Ying, 2017, Chen and Haziza, 2019). Although we focus on deterministic multiply robust imputation in the sequel, a random version as well as a random hot-deck version can be obtained using approaches similar to those described in Section 2 for doubly robust imputation.

Regardless of the aggregation approach, the first step is to fit each of the $H + J$ models in classes $\mathcal{M}_1$ and $\mathcal{M}_2$. For each $k \in S$, we can then construct 2 vectors: (i) The vector $\widehat{\mathbf{m}}$, of size H, given by $\widehat{\mathbf{m}} = \left(m^{(1)}(\mathbf{x}_k^{(1)}; \widehat{\boldsymbol{\beta}}^{(1)}), \cdots, m^{(H)}(\mathbf{x}_k^{(H)}; \widehat{\boldsymbol{\beta}}^{(H)})\right)^\top$. (ii) The vector $\widehat{\mathbf{p}}$, of size J, given by $\widehat{\mathbf{p}} = \left(p^{(1)}(\mathbf{x}_k^{(1)}; \widehat{\boldsymbol{\alpha}}^{(1)}), \cdots, p^{(J)}(\mathbf{x}_k^{(J)}; \widehat{\boldsymbol{\alpha}}^{(J)})\right)^\top$. The estimators $\widehat{\boldsymbol{\beta}}^{(1)}, \cdots, \widehat{\boldsymbol{\beta}}^{(H)}$, $\widehat{\boldsymbol{\alpha}}^{(1)}, \cdots, \widehat{\boldsymbol{\alpha}}^{(J)}$, denote suitable estimators (e. g., maximum likelihood estimators or weighted least squares estimators) for their corresponding parameters.

3.1 Aggregation through calibration

Aggregation through calibration proceeds as follows:

(i) We start by obtaining a calibrated weighting system $\{w_1, w_2, \ldots, w_{n_r}\}$, where $w_k, k \in S_r$, is a scalar summary of the information contained in the H imputation models and the J nonresponse models. For simplicity, we consider the generalized chi-square distance (Deville and Särndal, 1992). We seek a weighting system $\{w_1, w_2, \ldots, w_{n_r}\}$ such that

$$\sum_{k \in S_r} \pi_k (w_k - \pi_k^{-1})^2 / 2$$

is minimum subject to the $H + J + 1$ calibration constraints

$$\sum_{k \in S_r} w_k = \sum_{k \in S} \pi_k^{-1},$$

$$\sum_{k \in S_r} w_k m^{(h)}(\mathbf{x}_k^{(h)}; \widehat{\boldsymbol{\beta}}^{(h)}) = \sum_{k \in S} \pi_k^{-1} m^{(h)}(\mathbf{x}_k^{(h)}; \widehat{\boldsymbol{\beta}}^{(h)}), h = 1, \ldots, H,$$

and

$$\sum_{k \in S_r} w_k \frac{1}{p^{(j)}(\mathbf{x}_k^{(j)}; \widehat{\boldsymbol{\alpha}}^{(j)})} = \sum_{k \in S} \pi_k^{-1} \frac{1}{p^{(j)}(\mathbf{x}_k^{(j)}; \widehat{\boldsymbol{\alpha}}^{(j)})}, j = 1, \ldots J.$$

The resulting weights w_k are given by

$$w_k = \pi_k^{-1} \times (1 + \widehat{\boldsymbol{\lambda}}^\top \mathbf{h}_k), \tag{6}$$

where $\widehat{\boldsymbol{\lambda}}$ is a vector of estimated Lagrange multipliers of size $H + J + 1$ and

$$\mathbf{h}_k = (1, \mathbf{h}_{1k}^\top, \mathbf{h}_{2k}^\top)^\top$$

with

$$\mathbf{h}_{1k} = \left(m^{(1)}(\mathbf{x}_k^{(1)}; \widehat{\boldsymbol{\beta}}^{(1)}) - \overline{m}^{(1)}, \ldots, m^{(H)}(\mathbf{x}_k^{(H)}; \widehat{\boldsymbol{\beta}}^{(H)}) - \overline{m}^{(H)}\right)^\top$$

and

$$\mathbf{h}_{2k} = \left(p^{(1)}(\mathbf{x}_k^{(1)}; \widehat{\boldsymbol{\alpha}}^{(1)}) - \overline{p}^{(1)}, \ldots, p^{(J)}(\mathbf{x}_k^{(J)}; \widehat{\boldsymbol{\alpha}}^{(J)}) - \overline{p}^{(J)}\right)^\top,$$

with $\overline{m}^{(h)} = \sum_{k \in S} \pi_k^{-1} m^{(h)}(\mathbf{x}_k^{(h)}; \widehat{\boldsymbol{\beta}}^{(h)}) / \sum_{k \in S} \pi_k^{-1}$ and $\overline{p}^{(j)} = \sum_{k \in S} \pi_k^{-1} p^{(j)}(\mathbf{x}_k^{(j)}; \widehat{\boldsymbol{\alpha}}^{(j)})^{-1} / \sum_{k \in S} \pi_k^{-1}$. Distance functions other than the generalized chi-square distance can be used; see Chen and Haziza (2017). To better understand the rationale behind this type of aggregation, we define the standardized version of $\widehat{\boldsymbol{\lambda}}$ as $\widehat{\boldsymbol{\lambda}}^2 / \widehat{\boldsymbol{\lambda}}^\top \widehat{\boldsymbol{\lambda}}$, where $\widehat{\boldsymbol{\lambda}}^2 \equiv (\widehat{\lambda}_0^2, \cdots, \widehat{\lambda}_{J+H}^2)^\top$. It follows that the

standardized version of the term $\widehat{\boldsymbol{\lambda}}^{\top}\mathbf{h}_k$ on the right hand-side of (6) can be expressed as

$$\begin{aligned}\frac{\widehat{\boldsymbol{\lambda}}^2}{\widehat{\boldsymbol{\lambda}}^{\top}\widehat{\boldsymbol{\lambda}}}\mathbf{h}_k &= \delta_0 + \delta_1\left\{m^{(h)}(\mathbf{x}_k^{(h)};\widehat{\boldsymbol{\beta}}^{(H)})\right\} + \cdots + \delta_H\left\{m^{(H)}(\mathbf{x}_k^{(H)};\widehat{\boldsymbol{\beta}}^{(H)})\right\} \\ &\quad + \cdots + \delta_{H+1}\left\{p^{(1)}(\mathbf{x}_k^{(1)};\widehat{\boldsymbol{\alpha}}^{(1)}) - \overline{p}^{(1)}\right\} + \cdots + \delta_{H+J}\left\{p^{(J)}(\mathbf{x}_k^{(J)};\widehat{\boldsymbol{\alpha}}^{(J)}) - \overline{p}^{(J)}\right\}, \quad (7)\end{aligned}$$

where $\delta_h = \widehat{\lambda}_h^2/\sum_{h=0}^{J+H}\widehat{\lambda}_h^2, h = 0,\ldots,H+J$. The aggregation weights δ_h sum to 1, which makes (7) a convex combination of the individual predictions obtained from each of the $H+J$ models. Therefore, the calibration weight in (6) can be viewed as an aggregate score or a scalar summary of the information contained in the $H+J$ models. If one of the models in either class is correctly specified, we expect the associated aggregation weight δ_h to be large and the other weights to be small.

(ii) The imputed values y_k^* are obtained by fitting a weighted linear regression with Y as the dependent variable, and $\mathbf{h}_k$ as the vector of explanatory variables. The regression weights are given by $\phi_k = \pi_k^{-1}\left\{(1+\widehat{\lambda}_r^{\top}\mathbf{h}_k) - 1\right\}$. This leads to

$$y_k^* = \mathbf{h}_k^{\top}\widehat{\boldsymbol{\gamma}}, \quad k \in S_m,$$

where

$$\widehat{\boldsymbol{\gamma}} = \left(\sum_{k\in S_r}\phi_k\mathbf{h}_k\mathbf{h}_k^{\top}\right)^{-1}\left(\sum_{k\in S_r}\phi_k\mathbf{h}_k y_k\right).$$

It follows that an estimator of t_y after imputation is given by

$$\widehat{t}_{y,MR} = \sum_{k\in S_r}\pi_k^{-1}y_k + \sum_{k\in S_m}\pi_k^{-1}\mathbf{h}_k^{\top}\widehat{\boldsymbol{\gamma}}. \quad (8)$$

The estimator $\widehat{t}_{y,MR}$ given by (8) is multiply robust in the sense that it remains consistent if all but one of the $H+J$ models are misspecified.

3.2 Aggregation through refitting

Aggregation through refitting proceeds as follows:

(i) Fit a linear regression model based on $k \in S_r$ with Y as the dependent variable and $\widehat{\mathbf{m}}$ as the vector of explanatory variables. The vector of estimated regression coefficients is denoted as $\widehat{\boldsymbol{\eta}} = (\widehat{\eta}_1,\cdots,\widehat{\eta}_H)^{\top}$. Define the standardized version of $\widehat{\boldsymbol{\eta}}$ as $\widehat{\boldsymbol{\eta}}^2/\widehat{\boldsymbol{\eta}}^{\top}\widehat{\boldsymbol{\eta}}$, where $\widehat{\boldsymbol{\eta}}^2 \equiv (\widehat{\eta}_1^2,\cdots,\widehat{\eta}_K^2)^{\top}$. The aggregate or compressed score attached to unit $k \in S_r$ is defined as

$$\widehat{m}_k = \sum_{h=1}^{H}\omega_h m^{(h)}(\mathbf{x}_k^{(h)};\widehat{\boldsymbol{\beta}}^{(h)}), \quad (9)$$

where $\omega_h = \widehat{\eta}_h^2/\sum_{h=1}^{H}\widehat{\eta}_h^2$. The aggregation weights ω_h sum to 1. Therefore, the aggregate score $\widehat{m}_k, k \in S_r$, can be viewed as a convex combination of the individual predictions obtained from each of the H imputation models.

(ii) Fit a linear regression model based on $k \in S$ with R as the dependent variable and $\widehat{\mathbf{p}}$ as the vector of explanatory variables. The vector of estimated regression coefficients is denoted as $\widehat{\boldsymbol{\tau}} = (\widehat{\tau}_1,\cdots,\widehat{\tau}_J)^{\top}$. Define the standardized version of $\widehat{\boldsymbol{\tau}}$ as $\widehat{\boldsymbol{\tau}}^2/\widehat{\boldsymbol{\tau}}^{\top}\widehat{\boldsymbol{\tau}}$, where $\widehat{\boldsymbol{\tau}}^2 \equiv (\widehat{\tau}_1^2,\cdots,\widehat{\tau}_J^2)^{\top}$.

The aggregate score attached to unit $k \in S$ is defined as

$$\widehat{p}_k = \sum_{j=1}^{J} \phi_j p^{(j)}(\mathbf{x}_k^{(j)}; \widehat{\boldsymbol{\alpha}}^{(j)}), \tag{10}$$

where $\phi_j = \widehat{\tau}_j^2 / \sum_{j=1}^{J} \widehat{\tau}_j^2$. The aggregation weights ϕ_j sum to 1. Therefore, the aggregate score $\widehat{p}_k, k \in S$, can be viewed as a convex combination of the individual predictions obtained from each of the J nonresponse models. This ensures that the aggregate score $\widehat{p}_k$ lies between 0 and 1.

(iii) The imputed values $y_k^*, k \in S_m$ is given by

$$y_k^* = \mathbf{h}_k^\top \widehat{\boldsymbol{\gamma}}^*, \quad k \in S_m,$$

where $\mathbf{h}_k = (1, \widehat{m}_k)^\top$ and

$$\widehat{\boldsymbol{\gamma}}^* = \left(\sum_{k \in S_r} \pi_k^{-1} \frac{1 - \widehat{p}_k}{\widehat{p}_k} \mathbf{h}_k \mathbf{h}_k^\top \right)^{-1} \sum_{k \in S_r} \pi_k^{-1} \frac{1 - \widehat{p}_k}{\widehat{p}_k} \mathbf{h}_k y_k.$$

It follows that an estimator of t_y after imputation is given by

$$\widehat{t}_{y,MR} = \sum_{k \in S_r} \pi_k^{-1} y_k + \sum_{k \in S_m} \pi_k^{-1} \mathbf{h}_k^\top \widehat{\boldsymbol{\gamma}}^*. \tag{11}$$

The estimator $\widehat{t}_{y,MR}$ given by (11) is multiply robust in the sense that it remains consistent if all but one of the $H + J$ models are misspecified. This can be explained as follows: if the class $\mathcal{M}_1$ contains the true imputation model, say $m^{(1)}(\mathbf{x}_k^{(1)}; \boldsymbol{\beta}^{(1)})$, we expect the aggregation weight ω_1 associated with the prediction $m^{(1)}(\mathbf{x}_k^{(1)}; \widehat{\boldsymbol{\beta}}^{(1)})$ to be close to 1, and the other aggregation weights $\omega_h, h = 2, \ldots, H$, to be close to 0. Similarly, if the class $\mathcal{M}_2$ contains the true nonresponse model, say $p^{(1)}(\mathbf{x}_k^{(1)}; \boldsymbol{\alpha}^{(1)})$, we expect the aggregation weight ϕ_1 associated with the prediction $p^{(1)}(\mathbf{x}_k^{(1)}; \widehat{\boldsymbol{\alpha}}^{(1)})$ to be close to 1, and the other aggregation weights $\phi_j, j = 2, \ldots, J$, to be close to 0. This is illustrated in Section 4.1

4 Empirical investigation

In this section, we present two limited empirical investigation: the first examines the distribution of the weights involved in the aggregation procedures, whereas the second compares the performance of several estimators in terms of bias and efficiency in the case of data Not Missing At Random (NMAR).

4.1 Distribution of the aggregation weights

We generated $1,000$ finite populations, each of size $N = 10,000$. In each population, we generated 4 explanatory variables X_1-X_4 independently from a standard normal distribution. The y-values were then generated according to $y_k = 1 + x_{1k} + x_{2k} + x_{3k} + x_{4k} + \epsilon_k$, $k = 1, 2, \ldots, N$, where the ϵ_k's were independently generated from a standard normal distribution. In each population, we selected a sample S, of size $n = 800$, according to inclusion probability proportional-to-size systematic sampling with size variable $s_k = 0.5v_k + 1$, where v_k was generated from a standard chi-square distribution with one degree of freedom. In each sample the response indicators R_k were independently generated from a Bernoulli distribution with probability $logit(p(\mathbf{x}_k; \alpha)) = 0.5 + x_{1k} + x_{2k} + x_{3k} + x_{4k}$. This led to an overall response rate of about 56%.

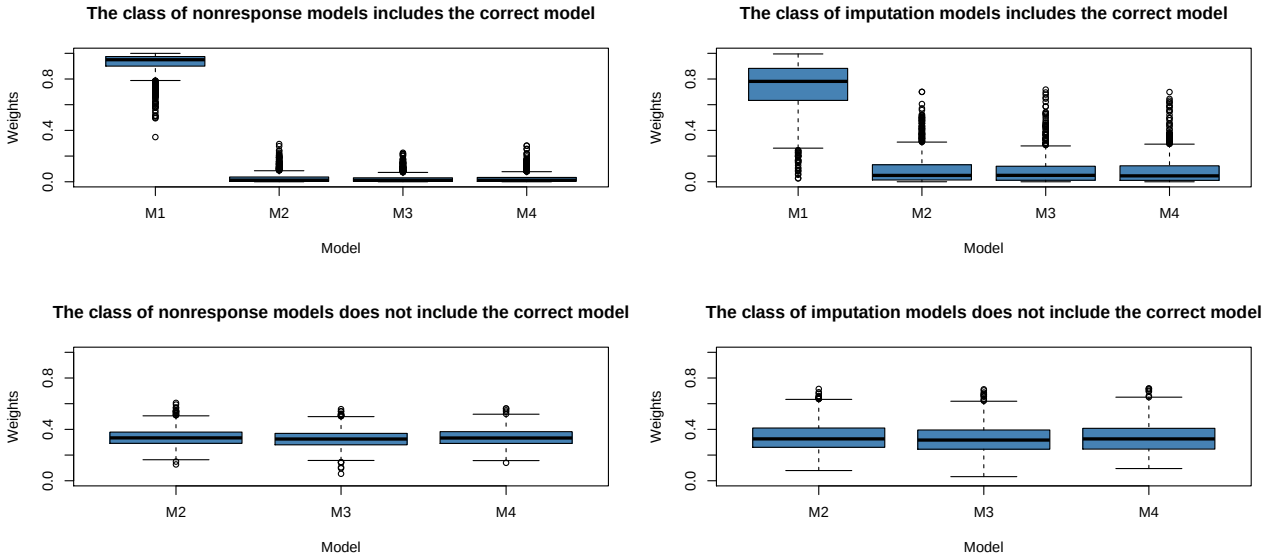

Figure 1: Distribution of the aggregation weights δ_h for the aggregation through calibration

The class of imputation models, $\mathcal{M}_1$, consisted of 4 imputation models: the correct imputation model based on X_1-$X4$ (M1); an incorrect linear regression model based on X_2 only (M2); an incorrect linear regression model based on X_3 only (M3); an incorrect linear regression model based on x_4 only (M4). The class of nonresponse models, $\mathcal{M}_2$, also consisted of 4 nonresponse models: the correct nonresponse model based on X_1-X_4 (M1); an incorrect logistic regression model based on X_2 only (M2); an incorrect logistic regression model based on X_3 only (M3); an incorrect logistic regression model based X_4 only (M4). In each class, each of the 4 models was fitted and the predictions were aggregated using both aggregation through calibration (see Section 3.1) and aggregation through refitting (see Section 3.2). For the aggregation through calibration procedure, we computed, in each sample, the aggregation weights δ_h in (7). For the aggregation through refitting procedure, we computed, in each sample, the aggregation weights ω_h in (9) and the aggregation weights ϕ_j in (10).

Figure 1 and Figure 2 display the distribution of the aggregation weights for the aggregation through calibration and the aggregation through refitting, respectively. When the class $\mathcal{M}_1$ or $\mathcal{M}_2$ included the correct model (M1), we note that both aggregation procedures put most of the weight on the correct model (M1). The incorrect models received a much smaller weight. This suggests that both aggregation procedures perform some type of implicit of model selection. When the classes $\mathcal{M}_1$ and $\mathcal{M}_2$ did not include the correct model, each of the models (M2)-(M4) contributed almost equally to the resulting predictions. In other words, a prediction was essentially defined as the average of the predictions generated from each of the models.

4.2 Data Not Missing At Random

We evaluated the performance of several estimators in terms of bias and efficiency in the context of NMAR. We used a simulation setup similar to that of Chen and Haziza (2021). We generated $B = 1,000$ finite populations, each of size $N = 10,000$. In each population, we generated 4 explanatory variables X_1-X_4 independently from a standard normal distribution. The y-values were then generated according to $y_k = 210 + 27.4x_{1k} + 13.7(x_{2k} + x_{3k} + x_{4k}) + \epsilon_k,\ k = 1, 2, \ldots, N$, where the ϵ_k's were independently generated from a standard normal distribution. In each population, we selected a sample S, of size $n = 800$, according to the same sampling design described in Section 4.1. In each sample, the response indicators R_k were independently generated from a

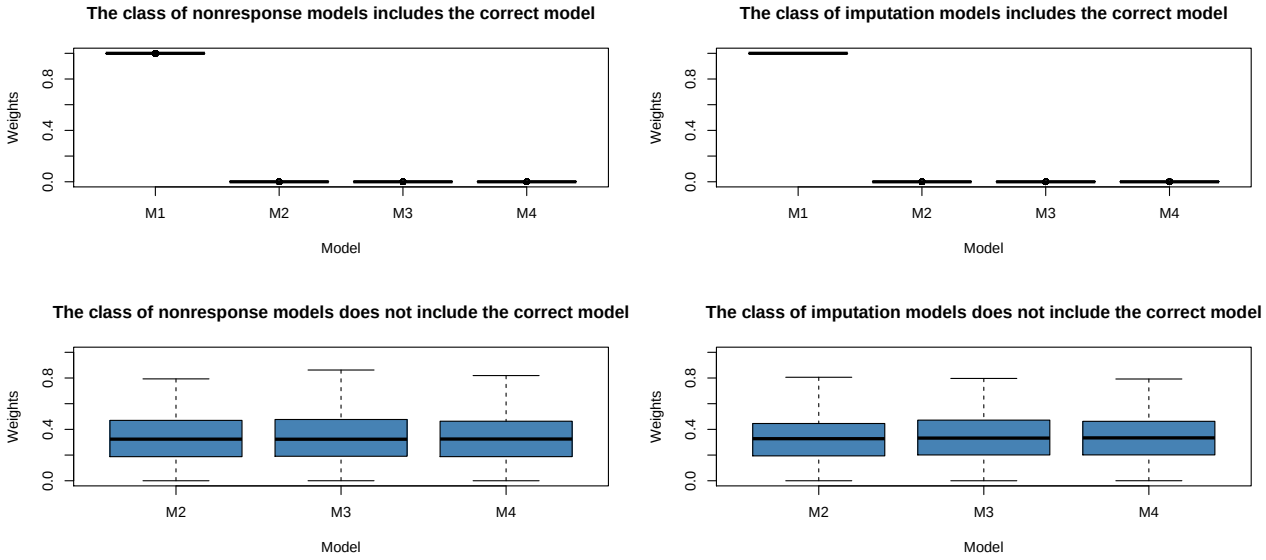

Figure 2: Distribution of the aggregation weights ω_h and ϕ_j for the aggregation through refitting

Bernoulli distribution with probability $logit(p_k(\alpha)) = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \alpha_3 x_{3,i} + \alpha_4 x_{4,i} + \alpha_5 y_i^{1/4}$ with $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (-2.4, -1, 0.5, -0.25, -0.1, 0.5)$, which corresponds to a response rate of about 40% and $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (-1.3, -1, 0.5, -0.25, -0.1, 0.5)$, which corresponds to a response rate of about 60%. We assumed that only the transformed variables Z_1-Z_4 of X_1-X_4, were available to the imputer, where $z_{1k} = \exp(x_{1k}/2)$, $z_{2k} = x_{2k}\{1 + \exp(x_{1k})\}^{-1} + 10$, $z_{3k} = (x_{1k}x_{3k}/25 + 0.6)^3$, and $z_{4k} = (x_{2k} + x_{4k} + 20)^2$. In other words, the imputer did not have access to the variables X_1-X_4. Kang and Schafer (2008) used a similar setup.

We were interested in estimating the finite population mean of Y. In each sample, we computed the following estimators of the mean: (1). The complete data estimator that corresponds to the weighted mean of the sample y-values (COM); (2). The estimator based on doubly robust imputation, where the nonresponse model was a logistic regression model based on Z_1-Z_4 and the imputation model was a linear regression model based on Z_1-Z_4 (DR); (3). The estimator based on nearest-neighbor imputation using Z_1-Z_4 as matching variables (NN); (4). Five multiply robust estimators based on aggregation through calibration with the pseudo-empirical likelihood distance function. (MRC1)– Both the nonresponse model and the imputation model were based on Z_1-Z_4. (MRC2)– Both models in (MRC1) and their two-way, three-way, and four-way interaction terms; (MRC3)– Both models in (MRC1) and the additional models with $|Z|_1^{1/2}$, $|Z|_2^{1/2}$, $|Z|_3^{1/2}$, $|Z|_4^{1/2}$, and their two-way, three-way, and four-way interaction terms; (MRC4)– Both models in (MRC1) and the additional models with $\log|Z_1|$, $\log|Z_2|$, $\log|Z_3|$, $\log|Z_4|$ as explanatory variables, and their two-way, three-way, and four-way interaction terms; (MRC5)–Based on all the models used in (MRC1) to (MRC4); (5). Four multiply robust robust estimators based on aggregation through refitting: (MRP2)–Using the same models as in (MRC2); (MRP3)– Using the same models as in (MRC3); (MRP4)– Using the same models as in (MRC4); (MRP5)– Using the same models as in (MRC5).

For each estimator, we computed the following Monte Carlo measures: bias, standard error and root mean squared error. The results are shown in Table 1. As expected, the complete data estimator COM exhibited negligible bias and was the most efficient. Both the estimator DR and NN showed appreciable bias. Except for MRP2, the MRC and MRP estimators performed much better than DR and NN. The MRC estimators performed generally better than their MRP counterparts. This can be explained by the fact that the calibration procedure used in the aggregation through calibration provides some robustness against the presence of small estimated response probabilities.

Table 1: Monte Carlo Bias (BIAS), Standard Error (SE), and Root Mean Squared Error (RRMSE) for different estimation procedures.

	Response rate–40%			Response rate–60%		
Procedure	Bias	SE	RMSE	Bias	SE	RMSE
COM	- 0.05	1.39	1.39	0.01	1.34	1.34
DR	-12.69	44.49	46.26	-6.71	40.90	41.44
NN	-9.83	1.97	10.02	-5.30	1.52	5.52
MRC1	-2.70	1.95	3.33	-1.84	1.57	2.42
MRC2	-0.69	1.90	2.02	-0.69	1.57	1.72
MRC3	-0.83	1.65	1.84	-0.57	1.45	1.56
MRC4	-1.27	1.58	2.02	-0.67	1.40	1.55
MRC5	-1.02	2.26	2.48	-0.66	1.47	1.61
MRP2	-6.49	34.95	35.55	-3.98	17.51	17.95
MRP3	-1.62	4.27	4.56	-1.16	3.05	3.26
MRP4	-1.28	1.61	2.05	-0.72	1.47	1.64
MRP5	-1.08	2.16	2.42	-0.80	1.91	2.07

References

Bang, H. and Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–973.

Boistard, H., Chauvet, G. and Haziza, D. (2016). Doubly robust inference for the distribution function in the presence of missing survey data. *Scandinavian Journal of Statistics* **43**, 683—699.

Cassel, C. M., Särndal, C. E. and Wretman, J. H. (1977). *Foundations of Inference in Survey Sampling*. Wiley New York.

Chan, K. C. G. and Yam, S. C. P. (2014). Oracle, multiple robust and multipurpose calibration in a missing response problem. *Statistical Science* **29**, 380—396.

Chen, S. and Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika* **104**, 439—453.

Chen, S. and Haziza, D. (2019). On the nonparametric multiple imputation with multiply robustness. *Statistica Sinica* **29**, 2035–2053.

Chen, S. and D. Haziza, D. (2021). A review of multiply robust estimation with missing data. *In Modern Statistical Methods for Health Research*, 103–118.

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**, 723–734.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382.

Duan, X. and Yin, G. (2017). Ensemble approached to estimation of the population mean with missing responses. *Scandinavian Journal of Statistics* **44**, 899–917.

Han, P. (2014). A further study of the multiply robust estimator in missing data analysis. *Journal of Statistical Planning and Inference* **148**, 101–110.

Han, P. (2016). Intrinsic efficiency and multiple robustness in longitudinal studies with drop-out. *Biometrika* **103**, 683–700.

Han, P. and Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika* **100**, 417–430.

Han, P., Kong, L., Zhao, J. and Zhou, X. (2019). A general framework for quantile estimation with incomplete data. *Journal of the Royal Statistical Society B* **81**, 305–333.

Haziza, D. and Rao, J. N. K. (2006). A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology* **32**, 53–64.

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**, 523–539.

Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica* **24**, 375–394.

Kott, P. S. (1994). A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association* **89**, 693–696.

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**, 846–866.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model-Assisted Survey Sampling.* Springer-Verlag.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion and rejoinder). *Journal of the American Statistical Association* **94**, 1096–1146.

Seaman, S. R. and Vansteelandt, S. (2018). Introduction to double robust methods for incomplete data. *Statistical Science* **33**, 184–197.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association* **101**, 1619–1637.