Published in final edited form as:

Metron. 2017 December ; 75(3): 333–343. doi:10.1007/s40300-017-0128-9.

Multiply robust imputation procedures for zero-inflated distributions in surveys

Sixia Chen* and David Haziza†

*Department of Biostatistics and Epidemiology, University of Oklahoma, Oklahoma City, OK 73104, U.S.A

†Department of mathematics and statistics, Université de Montréal, Montreal, Canada, H3C 3J7

Abstract

Item nonresponse in surveys is usually treated by some form of single imputation. In practice, the survey variable subject to missing values may exhibit a large number of zero-valued observations. In this paper, we propose multiply robust imputation procedures for treating this type of variable. Our procedures may be based on multiple imputation models and/or multiple nonresponse models. An imputation procedure is said to be multiply robust if the resulting estimator is consistent when all models but one are misspecified. The variance of the imputed estimators is estimated through a generalized jackknife variance estimation procedure. Results from a simulation study suggest that the proposed procedures perform well in terms of bias, efficiency and coverage rate.

Keywords

Double robustness; Item nonresponse; Imputation; Multiple robustness; Variance estimation; Zero-valued observations

1 Introduction

Item nonresponse in surveys occurs in surveys when some, but not all, survey variables are missing. Estimators based on the complete cases are generally biased unless the data are missing completely at random (Rubin, 1976). Reducing the potential bias due to nonresponse requires the availability of fully observed variables that are recorded for both the responding and nonresponding units. For more information about different approaches for handling missing values, see Little and Rubin (2002) and Kim and Shao (2013), among others. Item nonresponse is usually treated by some form of single imputation, whereby a missing value is replaced by a single plausible imputed value constructed on the basis of fully observed variables. With imputation, a complete data file is produced, allowing data users to obtain point estimates through complete data estimation procedures. A number of single imputation procedures are used in practice; e.g., mean, ratio and regression imputation, nearest-neighbour imputation and random hot-deck imputation, among others; see Haziza (2009) for a detailed description of commonly used imputation procedures.

In practice, the survey variable subject to missingness may be prone to a large number of zero-valued observations. This situation arises, for example, in the context of the Industrial Consumption of Energy survey (ICE) conducted by Statistics Canada. ICE provides

estimates of energy consumption by manufacturing establishments in Canada. It collects data on the consumption of various energy commodities such as electricity, natural gas, propane, diesel, wood and steam. For variables such as propane consumptions, the data contain a large number of zero-valued observations as many Canadian establishments do not use propane. Imputing this type of survey variable requires a zero-inflated imputation model to reflect the mixture of zero-valued and nonzero-valued observations. The imputation model is a set of assumptions about the survey variable requiring imputation. Most often, the imputed values are derived from the assumed imputation model. However, being based on a single model, the resulting point estimators are vulnerable to misspecification of the imputation model. Protection against misspecification of the model may be achieved by assuming a nonresponse model, which is a set of assumptions about the unknown nonresponse mechanism. The imputed values are derived by using both models. The resulting estimator is doubly robust in the sense that it remains consistent if either the imputation model or the nonresponse model is correctly specified; see Haziza and Rao (2006), Kim and Park (2006) and Kim and Haziza (2014), among others. Haziza et al. (2014) proposed doubly robust imputation procedures for finite population means in the presence of a large number of zero-valued observations. Gelein (2016) extended the results of Haziza et al. (2014) to the case of random regression imputation and considered the estimation of the finite population distribution function. Although doubly robust methods provide some protection, some empirical investigations have shown that they tend to have poor numerical performances in terms of bias and efficiency if both models are misspecified; e.g., Kang and Schafer (2007), Han (2014b) and Chen and Haziza (2017a).

In recent years, multiply robust estimation procedures have attracted some attention; see Han and Wang (2013), Han (2014a, 2014b), Chan and Yam (2014) and Chen and Haziza (2017a, b, c) among others. The concept of multiple robustness can be viewed as an extension of double robustness. Unlike doubly robust procedures, multiply robust procedures make use of multiple nonresponse models and/or multiple imputation models. An estimation procedure is multiply robust if the resulting estimator is consistent when all models but one are misspecified. In the context of single imputation, Chen and Haziza (2017a) proposed multiply robust imputation procedures for complex survey data. Multiply robust multiple imputation procedures were studied in Chen and Haziza (2017c). When all the models are misspecified, a number of empirical investigations available in the literature suggest that multiply robust estimation procedures perform much better than doubly robust procedures based on the same working models, in terms bias and efficiency; see e.g., Han (2014b) and Chen and Haziza (2017a). In this paper, we extend the approach of Chen and Haziza (2017a) to the case of survey variables exhibiting a large number of zero-valued observations and propose multiply robust imputation procedures for treating this type of variable.

Our paper is organized as follows. The notation and the basic setup is introduced in Section 2. In Section 3, we describe two multiply robust imputation procedures tailored for survey variables subject to missingness and exhibiting a large number of zeros. Jackknife variance estimation is discussed in Section 4. The results of a simulation study, comparing several estimation procedures, in terms of bias, efficiency and coverage rate, are presented in Section 5. Section 6 concludes the article with some discussion.

2 Basic setup

Consider a finite population $\mathcal{P}$ of size N . We are interested in estimating the finite population mean $\bar{Y} = t_y/N$, where $t_y = \sum_{i \in \mathcal{P}} y_i$, denotes the finite population total corresponding to the survey variable y . A sample S is selected from $\mathcal{P}$ according to a given sampling design $p(S)$ with first-order inclusion probabilities π_i and second-order inclusion probabilities π_{ij} , i, j . A complete data estimator of $\bar{Y}$ is the Hájek estimator (Hájek, 1971):

$$\hat{\bar{Y}} = \frac{1}{\hat{N}} \sum_{i \in S} w_i y_i, \quad (1)$$

where $w_i = \pi_i^{-1}$ is the design (or basic) weight attached to unit i and $\hat{N} = \sum_{i \in S} w_i$ is a design-unbiased estimator of the population size N . When the y -variable is subject to missing values, the complete data estimator (1) cannot be computed.

We assume the y -variable obeys the following zero-inflated model:

$$y_i = \delta_i \{m(x_i; \beta) + \varepsilon_i\} + (1 - \delta_i) \times 0, \quad (2)$$

where $m(\cdot)$ is an unknown function, δ_i is a zero/non-zero indicator such that $\delta_i = 1$ if $y_i \neq 0$ and $\delta_i = 0$ otherwise, x_i is a vector of fully observed variables associated with unit i and β is a vector of unknown parameters. We assume that δ_i is independent of ε_i after accounting for x_i . We also assume $E(\varepsilon_i / x_i) = 0$, $E(\varepsilon_i \varepsilon_j / x_i) = 0$ for $i \neq j$ and $V(\varepsilon_i / x_i) = \sigma^2$, where σ^2 is an unknown parameter. Further, we assume that δ_i follows a Bernoulli distribution with unknown probability function

$$\Pr(\delta_i = 1 | x_i, y_i) = \Pr(\delta_i = 1 | x_i) \equiv q(x_i; \gamma), \quad (3)$$

where $q(x_i; \gamma)$ is a unknown function with parameter γ . Combining (2) and (3), we have $E(y_i / x_i) = q(x_i; \gamma) m(x_i; \beta) \equiv M(x_i; \beta, \gamma)$. Model (2) is called an imputation model. Note that for the imputation model to be correctly specified, both $m(x; \beta)$ and $q(x; \gamma)$ must be correctly specified.

Let r_i be the response indicator such that $r_i = 1$ if y_i is observed and $r_i = 0$ if y_i is missing. We make the Missing At Random (MAR) assumption (Rubin, 1976); i.e.,

$$\Pr(r_i = 1 | x_i, y_i) = \Pr(r_i = 1 | x_i) \equiv p(x_i; \alpha), \quad (4)$$

where $p(x_i; \alpha)$ is a unknown function with parameter α . Model (4) is called a nonresponse model.

If one places a complete reliance on the nonresponse model (4), a natural estimator of $\bar{Y}$ is the propensity score weighted estimator

$$\hat{\bar{Y}}_{PS} = \frac{1}{\hat{N}} \sum_{i \in S} w_i \frac{r_i}{p(x_i; \hat{\alpha})} y_i, \quad (5)$$

where $\hat{\alpha}$ is a consistent estimator of α .

If one places a complete reliance on the imputation model (2), a natural estimator of $\bar{Y}$ is the imputed estimator

$$\hat{\bar{Y}}_I = \frac{1}{\hat{N}} \left\{ \sum_{i \in S} w_i r_i y_i + \sum_{i \in S} w_i (1 - r_i) M(x_i; \hat{\beta}, \hat{\gamma}) \right\}, \quad (6)$$

where $\hat{\gamma}$ and $\hat{\beta}$ denote consistent estimators of γ and β , respectively.

Alternatively, one may combine both the imputation model (2) and the nonresponse model (4) to obtain the doubly robust estimator

$$\hat{\bar{Y}}_{DR} = \frac{1}{\hat{N}} \left[\sum_{i \in S} w_i \frac{r_i}{p(x_i; \hat{\alpha})} y_i + \sum_{i \in S} w_i \left\{ 1 - \frac{r_i}{p(x_i; \hat{\alpha})} \right\} M(x_i; \hat{\beta}, \hat{\gamma}) \right]. \quad (7)$$

Because $V(\varepsilon_i / x_i) = \sigma^2 \lambda^\top x_i$, the estimator $\hat{\bar{Y}}_{DR}$ can be written as a weighted sum of observed and imputed values; see, e.g., Haziza and Rao (2006) and Haziza et al. (2014).

In the next section, we introduce multiply robust imputation procedures that allow for multiple models for $p(x; \alpha)$ and/or $m(x; \beta)$ and/or $q(x; \gamma)$, leading to an imputed estimator that remains consistent if at least one of the imputation models (which entails correctly specifying both $m(x; \beta)$ and $q(x; \gamma)$) or nonresponse models is correctly specified.

3 Proposed method

Let $\mathcal{C}_1 = \{p^j(x; \alpha^j); j = 1, \dots, J\}$ be a class of J nonresponse models, $\mathcal{C}_2 = \{m^k(x; \beta^k); k = 1, \dots, K\}$ be the class of K models for $m(x; \beta)$ and $\mathcal{C}_3 = \{q^l(x; \gamma^l); l = 1, \dots, L\}$ be the class of L models for $q(x; \gamma)$.

Estimators of γ^l are obtained by solving the following survey weighted estimating equations

$$S_{\gamma}^l(\gamma^l) = \sum_{i \in S} w_i r_i \frac{\delta_i - q^l(x_i; \gamma^l)}{q^l(x_i; \gamma^l) \{1 - q^l(x_i; \gamma^l)\}} \frac{\partial q^l(x_i; \gamma^l)}{\partial \gamma^l} = 0, \quad l = 1, \dots, L.$$

Define $\hat{U}_{qi} = q^1(x_i; \hat{\gamma}^1), \dots, q^L(x_i; \hat{\gamma}^L)^\top$. If one of the models in $\mathcal{C}_3$ is correctly specified, then a consistent estimator of $q(x_i; \gamma)$ is $\hat{q}_i = \hat{U}_{qi}^\top \times (\hat{\eta}_q^2 / \hat{\eta}_q^\top \hat{\eta}_q)$, where

$$\hat{\eta}_q = \left(\sum_{i \in S} w_i r_i \hat{U}_{qi} \hat{U}_{qi}^\top \right)^{-1} \sum_{i \in S} w_i r_i \hat{U}_{qi} \delta_i.$$

Here and throughout the paper, if $\mathbf{a} = (a_1, \dots, a_\ell)$ is a ℓ -vector, $\mathbf{a}^2$ denotes the column vector $(a_1^2, \dots, a_\ell^2)$. Using the normalized scores rather than the customary prediction $\hat{U}_{qi}^\top \hat{\eta}_q$ ensures that the weighted average lies in the appropriate range; see Duan and Yin (2017). Note that $\hat{\eta}_q$ is the weighted least square estimator obtained regressing δ_i on $\hat{U}_{qi}$ based on the responding units. The prediction $\hat{q}_i$ can be viewed as a score that compresses the model information contained in the L models for $q(x; \gamma)$. Due to the nature of the least squares procedure, $\hat{q}_i$ reproduces the true fitted probability if one of the models in $\mathcal{C}_3$ is correctly specified. Suppose that the first model, $q^1(x; \gamma^1)$ in $\mathcal{C}_3$ is correctly specified. Then, it can be shown that $\hat{q}_i = q^1(x_i; \gamma^1) + O_p(n^{-1/2})$; see Duan and Yin (2017).

The estimators $\hat{\alpha}^j$ and $\hat{\beta}^k$ can be obtained by solving the following survey weighted estimating equations

$$S_\alpha^j(\alpha^j) = \sum_{i \in S} w_i \frac{r_i - p^j(x_i; \alpha^j)}{p^j(x_i; \alpha^j) \{1 - p^j(x_i; \alpha^j)\}} \frac{\partial p^j(x_i; \alpha^j)}{\partial \alpha^j} = 0,$$

and

$$S_\beta^k(\beta^k) = \sum_{i \in S} w_i r_i \delta_i \{y_i - m^k(x_i; \beta^k)\} \frac{\partial m^k(x_i; \beta^k)}{\partial \beta^k} = 0.$$

Define

$$\hat{U}_{pi} = (p^1(x_i; \hat{\alpha}^1), \dots, p^J(x_i; \hat{\alpha}^J)), \quad \hat{U}_{mi} = (m^1(x_i; \hat{\beta}^1), \dots, m^K(x_i; \hat{\beta}^K)).$$

To summarize the working models information, we regress r_i on $\hat{U}_{pi}$ and y_i on $\hat{U}_{mi}$, which leads to the weighted least square regression coefficients

$$\hat{\eta}_p = \left(\sum_{i \in S} w_i \hat{U}_{pi} \hat{U}_{pi}^\top \right)^{-1} \sum_{i \in S} w_i \hat{U}_{pi} r_i$$

and

$$\hat{\eta}_m = \left(\sum_{i \in S} w_i r_i \delta_i \hat{U}_{mi} \hat{U}_{mi}^\top \right)^{-1} \sum_{i \in S} w_i r_i \delta_i \hat{U}_{mi} y_i.$$

Let $\hat{p}_i = \hat{U}_{pi}^\top \times (\hat{\eta}_p^2 / \hat{\eta}_p^\top \hat{\eta}_p)$ and $\hat{m}_i = \hat{U}_{mi}^\top \times (\hat{\eta}_m^2 / \hat{\eta}_m^\top \hat{\eta}_m)$. As for $\hat{q}_i$, the scores $\hat{p}_i$ and $\hat{m}_i$ compress respectively the information contained in the J nonresponse models and the K models for $m(x; \beta)$. It can be shown that $\hat{p}_i$ is a consistent estimator of $p(x_i; \alpha)$ if one of the models in $\mathcal{C}_1$ is correctly specified and $\hat{m}_i$ is a consistent estimator of $m(x_i; \beta)$ if one of the models in $\mathcal{C}_2$ is correctly specified.

Using the compressed scores $\hat{q}_i$ and $\hat{m}_i$, we define $\hat{M}_i = \hat{q}_i \hat{m}_i^\top$. Following Chen and Haziza (2017a), we consider the imputed value

$$y_i^* = h_i^\top \hat{\tau} \quad (8)$$

for the missing y_i , where $h_i = (1, \hat{M}_i)^\top$ with

$$\hat{\tau} = \left\{ \sum_{i \in S} w_i r_i (\hat{p}_i^{-1} - 1) h_i h_i^\top \right\}^{-1} \sum_{i \in S} w_i r_i (\hat{p}_i^{-1} - 1) h_i y_i.$$

The resulting imputed estimator is given by

$$\hat{Y}_{MR} = \frac{1}{\hat{N}} \left\{ \sum_{i \in S} w_i r_i y_i + \sum_{i \in S} w_i (1 - r_i) h_i^\top \hat{\tau} \right\}. \quad (9)$$

The imputation procedure (8) is referred to as a deterministic multiply robust imputation procedure. Since the first component of h_i is 1, we have

$$\sum_{i \in S} r_i w_i (\hat{p}_i^{-1} - 1) (y_i - h_i^\top \hat{\tau}) = 0$$

and (9) can be rewritten as

$$\hat{Y}_{MR} = \frac{1}{\hat{N}} \left\{ \sum_{i \in S} \frac{r_i}{\hat{p}_i} y_i + \sum_{i \in S} \left(1 - \frac{r_i}{\hat{p}_i} \right) h_i^\top \hat{\tau} \right\}.$$

Using similar techniques to those used in Chen and Haziza (2017a), it can be shown that $\widehat{Y}_{MR}$ is multiply robust.

To better reflect the mixture of zero-valued and nonzero-valued observations, we consider the following random version of (8):

$$y_i^* = \begin{cases} \hat{q}_i^{-1} h_i^\top \hat{\tau} & \text{with probability } \hat{q}_i \\ 0 & \text{with probability } 1 - \hat{q}_i. \end{cases} \quad (10)$$

The imputation procedure (10) is referred to as a random multiply robust imputation procedure. We denote by $\widehat{Y}_{MR}^*$ the resulting imputed estimator. Let E_I denote the expectation with respect to the random imputation mechanism. Noting that $E_I(y_i^*) = h_i^\top \hat{\tau}$, it follows that the estimator $E_I(\widehat{Y}_{MR}^*) = \widehat{Y}_{MR}$, where $\widehat{Y}_{MR}$ is given by (9). Therefore, $\widehat{Y}_{MR}^*$ is also multiply robust.

4 Variance estimation

It is well known that treating the imputed values as observed leads to an underestimation of the variance of the imputed estimators and ultimately to confidence intervals that are too narrow. In the case of negligible sampling fractions, there exist several variance estimation procedures taking nonresponse and imputation into account; see, e.g., Rao and Shao (1992) and Shao and Sitter (1996). In this section, we consider the generalized jackknife procedure of Berger (2007) for estimating the variance of $\widehat{Y}_{MR}$ and $\widehat{Y}_{MR}^*$. In the complete data situation, Berger (2007) showed that his jackknife procedure leads to a consistent variance estimator provided that the sampling design has a high entropy and that the sampling fraction n/N is negligible. Examples of high entropy sampling designs include conditional Poisson sampling design (Hájek, 1964), the Rao-Sampford design (Rao, 1965; Sampford, 1967), randomized systematic sampling (Tillé, 2006, Section 7.2) and sequential Poisson sampling (Ohlsson, 1998).

We start by describing the jackknife procedure for $\widehat{Y}_{MR}$. Let $w_{\lambda(j)}$ be the jackknife weights such that $w_{\lambda(j)} = n(n-1)^{-1} w_i$ if $i \neq j$ and $w_{\lambda(j)} = 0$ if $i = j$. Let $\widehat{Y}_{MR(j)}$ be the estimator of $\widehat{Y}_{MR}$ based on the data with the j -th unit excluded,

$$\widehat{Y}_{MR(j)} = \frac{1}{\hat{N}_{(j)}} \left\{ \sum_{i \in S} w_{i(j)} r_i y_i + \sum_{i \in S} w_{i(j)} (1 - r_i) y_{i(j)}^* \right\},$$

where $\hat{N}_{(j)} = \sum_{i \in S} w_{\lambda(j)}$, $y_{i(j)}^* = h_{i(j)}^\top \hat{\tau}_{(j)}$ with $h_{i(j)}^\top$ and $\hat{\tau}_{(j)}$ computed in the same way as $h_i^\top$ and $\hat{\tau}$ in Section 3 but with the jackknife weights $w_{\lambda(j)}$ instead of the original weights w_i . A jackknife variance estimator is given by

$$\hat{V}_J = \frac{n}{n-1} \sum_{i \in S} (1 - \pi_i) \left(u_i - \sum_{k \in S} \psi_k u_k \right)^2, \quad (11)$$

where $u_i = (1 - w_i)(\hat{Y}_{MR} - \hat{Y}_{MR(i)})$ and $\psi_i = c_i / \sum_{k \in S} c_k$ with $c_i = n(n-1)(1 - \pi_i)$. The consistency of $\hat{V}_J$ can be established through the reverse approach of Shao and Steel (1999); see also Fay (1991) and Kim and Rao (2009). It is worth noting that $\hat{V}_J$ remains consistent for the true variance of $\hat{Y}_{MR}$ even if all the models in $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ are misspecified. A $(1 - \zeta\%)$ confidence interval for $\bar{Y}$ is

$$\hat{Y}_{MR} \pm z_{\zeta/2} \sqrt{\hat{V}_J}, \quad (12)$$

where $z_{\zeta/2}$ is the upper $(1 - \zeta/2\%)$ critical value for the standard normal distribution. The confidence interval (12) is multiply robust in the sense that its coverage rate is close to nominal rate if all models but one are misspecified.

Turning to $\hat{Y}_{MR}^*$, we express its variance as

$$\begin{aligned} V(\hat{Y}_{MR}^*) &= V(\hat{Y}_{MR} + \hat{Y}_{MR}^* - \hat{Y}_{MR}) \\ &= V(\hat{Y}_{MR}) + V(\hat{Y}_{MR}^* - \hat{Y}_{MR}) \\ &= V(\hat{Y}_{MR}) + E \left\{ V_I \left(\hat{Y}_{MR}^* - \hat{Y}_{MR} \right) \right\} \\ &= V(\hat{Y}_{MR}) + E \left[V_I \left\{ \frac{1}{\hat{N}} \sum_{i \in S} w_i (1 - r_i) (y_i^* - h_i^\top \hat{\tau}) \right\} \right] \\ &= V(\hat{Y}_{MR}) + E \left\{ \frac{1}{\hat{N}^2} \sum_{i \in S} w_i^2 (1 - r_i) \left(\hat{q}_i^{-1} - 1 \right) \left(h_i^\top \hat{\tau} \right)^2 \right\}, \end{aligned} \quad (13)$$

where $V_I(\cdot)$ denotes the variance with respect to the random imputation mechanism. It follows that a consistent variance estimator is given by

$$\hat{V}_J^* = \hat{V}_J + \hat{V}_I^*, \quad (14)$$

where $\hat{V}_J$ is given by (11) and

$$\hat{V}_I^* = \frac{I}{\hat{N}^2} \sum_{i \in S} w_i^2 (1 - r_i) \left(\hat{q}_i^{-1} - 1 \right) \left(h_i^\top \hat{\tau} \right)^2.$$

A $(1 - \zeta\%)$ confidence interval for $\bar{Y}$ is

$$\hat{Y}_{MR}^* \pm z_{\zeta/2} \sqrt{\hat{V}_I^*}.$$

5 Simulation study

We conducted a simulation study to assess the performance of several estimation procedures in terms of bias and efficiency. We considered the simulation setup of Haziza et al. (2014). We generated a finite population of size $N = 10,000$ consisting of two variables: a survey variable y and an auxiliary variable x . First, the x -values were generated from a Gamma distribution with shift parameter 2 and scale parameter 5. Given the x -values, the population y -values were generated according to the model

$$y_i = \delta_i (15 + 1.5x_i + \varepsilon_i),$$

where the errors ε_i 's were generated from a normal distribution with mean equal to zero and variance equal to σ^2 , whose value was set so as to lead to a coefficient of determination between x and y of 70% for units with $\delta_i = 1$ (i.e., the non-zero part of the population). The δ_i 's were generated from a Bernoulli distribution with probability $q(x_i; \gamma)$ such that

$$\log \left(\frac{q(x_i; \gamma)}{1 - q(x_i; \gamma)} \right) = \gamma_0 + \gamma_1 x_i,$$

where γ_0 and γ_1 were set so that the proportion of zero-valued observations was approximately equal to 50%.

From the finite population, we selected $B = 1,000$ samples, of size $n = 200$, according to simple random sampling without replacement. In each sample, the response indicators r_i were generated from a Bernoulli distribution with probability $p(x_i; \alpha)$ such that

$$\log \left(\frac{p(x_i; \alpha)}{1 - p(x_i; \alpha)} \right) = \alpha_0 + \alpha_1 x_i,$$

where α_0 and α_1 were set so that the overall response rate was approximately equal to 70%.

We were interested in estimating the finite population mean of the y -values, $\bar{Y}$. Because the proposed imputation procedures may be based on different combinations of the models, we

used six digits between parentheses to distinguish estimators constructed using different models. The first two digits correspond to the correct and incorrect specification of $p(x; \alpha)$, respectively, the third and fourth digits corresponds to the correct and incorrect specification of $m(x; \beta)$, respectively, and the last two digits correspond to the correct and incorrect specification of $q(x; \gamma)$, respectively. For instance, the estimator $\hat{Y}(10|10|10)$ is based on correct specification of $p(x; \alpha)$, $m(x; \beta)$ and $q(x; \gamma)$, whereas the estimator $\hat{Y}(11|11|11)$ is based on all the (correct and incorrect) models for $p(x; \alpha)$, $m(x; \beta)$ and $q(x; \gamma)$. In our experiments, the incorrectly specified models for $p(x; \alpha)$, $m(x; \beta)$ and $q(x; \gamma)$ only included the intercept; that is, the x -variable was omitted from each of the three models.

We computed the following estimators: (i) The complete data estimator, $\hat{Y}_{COM}$, given by (1); (ii) The imputed estimator $\hat{Y}_I$, given by (6) based on correct and incorrect specification of $m(x; \beta)$ and $q(x; \gamma)$; (iii) The multiply robust estimator, $\hat{Y}_{MR}$, given by (9) based on correct and incorrect specification of $p(x; \alpha)$, $m(x; \beta)$ and $q(x; \gamma)$.

For each point estimator, we computed the Monte Carlo percent relative bias (RB), the relative standard error (RSE) and the relative root mean square error (RRMSE). The results are presented in Table 1. Before discussing the results, it is worth noting that, in our simulation set-up, some pairs of estimators were identical. This is due to the misspecified nonresponse model $p(x; \alpha)$ that contains only the intercept. As a result, the estimated response probability based on the incorrect model was equal to the overall response rate for all the units. In this case, only one estimator of each pair is shown in Table 1. Pairs of identical estimators include $\hat{Y}_{MR}(10|11|11)$ and $\hat{Y}_{MR}(11|11|11)$, $\hat{Y}_{MR}(10|11|01)$ and $\hat{Y}_{MR}(11|11|01)$ and $\hat{Y}_{MR}(11|01|01)$ and $\hat{Y}_{MR}(10|01|01)$.

From Table 1, the complete data estimator showed negligible bias and was the most efficient estimator in all the scenarios, as expected. We now turn to the estimators based on a single imputation model, which is typically what would be done in practice. The estimator $\hat{Y}_I(00|10|10)$, based on the correct imputation model only, showed a negligible bias (approximately 0.77%) and a value of RRMSE of 8.65. This estimator can be viewed as the gold standard in terms of bias and efficiency. On the other hand, when either $m(x; \beta)$ and/or $q(x; \gamma)$ was misspecified, the resulting estimator was highly biased. For example, the estimator $\hat{Y}_I(00|01|10)$ showed a value of RB of 10.64%, whereas the estimator $\hat{Y}_I(00|10|01)$ showed a value of RB equal to -29.54%. Finally, under the misspecification of both $m(x; \beta)$ and $q(x; \gamma)$, the estimator $\hat{Y}_I(00|01|01)$ exhibited a bias approximately equal to -26.52%. Therefore, the estimators based on a single misspecified imputation model (i.e., either $m(x; \beta)$ and/or $q(x; \gamma)$ was incorrectly misspecified) showed poor performances in terms of bias, which is not surprising.

Turning to the MR estimators based on a single nonresponse model and a single imputation model, they showed a small bias, when either the model for $p(x; \alpha)$ or the imputation model (which entails the correct specification of both $m(x; \beta)$ and $q(x; \gamma)$) was correctly specified.

When both models were correctly specified, the estimator $\hat{Y}_{MR}(10|10|10)$ showed a value of RB of 0.86% and a value of RRMSE equal to 8.86. A comparison of $\hat{Y}_I(00|10|10)$ and $\hat{Y}_{MR}(10|10|10)$ suggests that the latter was almost as efficient as the former with a slight increase in RRMSE of 2.4% approximately. However, when both $p(x; \alpha)$ and $m(x; \beta)$ were misspecified, the MR estimators were significantly biased. For example, the RB of $\hat{Y}_{MR}(01|10|01)$ was equal to -5.97% . It is interesting to note that, despite being inconsistent, the estimator $\hat{Y}_{MR}(01|10|01)$ performed much better than $\hat{Y}_I(00|10|01)$ in terms of bias (-5.95% vs. -29.54%).

Finally, we discuss the performance of MR estimator based on more than one imputation model and/or more than one nonresponse model. All the multiply robust estimators showed small biases and were almost as efficient as $\hat{Y}_{MR}(00|10|10)$ in all the scenarios. For example, the estimator $\hat{Y}_{MR}(11|11|11)$ that was based all the models, showed a value of RB of 0.89% and a value of RRMSE of 8.79. Therefore, the loss of efficiency with respect to the gold standard $\hat{Y}_I(00|10|10)$ was only about 1.6%. These results tend to suggest that incorporating multiple models does not seem to affect the efficiency significantly.

In addition, we assessed the performance of the jackknife variance estimator given by (11) in terms of percent relative bias (RB) as well as the coverage rate (CR) of the confidence interval (12). Table 2 presents the results corresponding to five multiply robust estimators. Other estimators led to very similar results and hence, are not shown here. From Table 2, the jackknife variance estimator showed a small bias with a value of RB less than 7.6%. The coverage rates were close to the nominal rate (95%) in all the scenarios.

6 Discussion

We proposed multiply robust imputation procedures for complex survey data when the variable subject to missingness exhibits a large number of zero-valued observations. The proposed procedures showed good performances in terms of bias and efficiency. To better reflect the mixture of zero-valued and nonzero-valued observations, we suggested the use of

(10) rather than (8). However, the resulting estimator, $\hat{Y}_{MR}^*$, suffers from an additional variability arising from the random imputation mechanism. Eliminating this variance can be achieved through the use of a balanced random imputation procedure similar to the one described in Haziza et al. (2014). This is a topic of future research.

If the interest lies in estimating the finite population distribution function, a random noise may be added for the portion of sample units with $\delta_i = 1$. More specifically, the missing y_i is replaced by

$$y_i^* = \delta_i^* (\hat{m}_i + \hat{\sigma} \sqrt{c_i} \varepsilon_i^*),$$

where $\hat{\sigma}$ is an estimator of σ , $\hat{\delta}_i^* = 1$ with probability $\hat{q}_i$ and $\hat{\delta}_i^* = 0$ with probability $1 - \hat{q}_i$, $\hat{m}_i$ is defined in Section 3 and ε_i^* are selected independently and with replacement from the set of residuals $R_r = \{e_j; r_j = 1, \delta_j = 1\}$ with $e_j = (\hat{\sigma} \sqrt{c_i})^{-1} (y_j - \hat{m}_j)$ with probability proportional to w_j . Estimation of the finite population distribution function will be presented elsewhere.

Acknowledgments

The authors would like to sincerely thank the guest co-editors Mary Dickson, Jean Opsomer and Giovanna Ranalli for their invitation to contribute to this special issue of *Metron*. The first author wishes to acknowledge the partial funding provided by National Institutes of Health, National Institute of General Medical Sciences (Grant 1 U54GM104938), an IDeA-CTR to the University of Oklahoma Health Sciences Center. The second author wishes to acknowledge the support of grants from the Natural Sciences and Engineering Research Council and the Canadian Statistical Sciences Institute.

References

- Berger YG. A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika*. 2007; 94:953–964.
- Chan KCG, Yam SCP. Oracle, multiple robust and multipurpose calibration in missing response problem. *Statistical Science*. 2014; 29:380–386.
- Chen S, Haziza D. Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*. 2017a; 102:439–453.
- Chen S, Haziza D. Jackknife empirical likelihood method for multiply robust estimation with missing data. 2017b Submitted.
- Chen S, Haziza D. Multiply robust nonparametric multiple imputation for the treatment of missing data. 2017c In revision for *Statistica Sinica*.
- Duan X, Yin G. Ensemble approaches to estimation of the population mean with missing responses. 2017 To appear in the *Scandinavian Journal of Statistics*.
- Fay, RE. A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference, US Bureau of the Census*; 1991. p. 429–440.
- Gelein B. Preserving the distribution function in surveys in case of imputation for zero inflated data. 2016 Submitted.
- Hájek J. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of the Institute of Statistical Mathematics*. 1964; 35:1491–1523.
- Hájek, J. Comment on “An essay on the logical foundations of survey sampling” by Basu, D. In: Godambe, VP., Sprott, DA., editors. *Foundations of Statistical Inference*. Toronto: Holt, Rinehart, and Winston; 1971. p. 236
- Han P, Wang L. Estimation with Missing Data: Beyond Double Robustness. *Biometrika*. 2013; 100:417–430.
- Han P. A further study of the multiply robust estimator in missing data analysis. *Journal of Statistical Planning and Inference*. 2014a; 148:101–110.
- Han P. Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*. 2014b; 109:1159–1173.
- Haziza, D. Imputation and inference in the presence of missing data. In: Pfeiffermann, D., Rao, CR., editors. *Handbook of Statistics, Volume 29A, Sample Surveys: Theory Methods and Inference*. Elsevier; 2009. p. 215–246.
- Haziza D, Rao JNK. A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology*. 2006; 32:53–64.
- Haziza D, Nambeu CO, Chauvet G. Doubly robust imputation procedures for finite population means in the presence of a large number of zeros. *The Canadian Journal of Statistics*. 2014; 42:650–669.

- Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*. 2007; 22:523–539.
- Kim JK, Haziza D. Doubly robust inference with missing data in survey sampling. *Statistica Sinica*. 2014; 24:375–394.
- Kim JK, Park H. Imputation using response probability. *Canadian Journal of Statistics*. 2006; 34:171–182.
- Kim JK, Rao JNK. Unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*. 2009; 96:917–932.
- Kim, JK., Shao, J. *Statistical Methods for Handling Incomplete Data*. London: Chapman and Hall/CRC; 2013.
- Little, RJA., Rubin, DB. *Statistical Analysis with Missing Data*. 2. John Wiley & Sons; Hoboken, NJ: 2002.
- Ohlsson E. Sequential Poisson sampling. *Journal of Official Statistics*. 1998; 14:149–162.
- Rao JNK. On two simple schemas of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*. 1965; 3:173–180.
- Rao JNK, Shao J. Jackknife variance estimation with survey data under hot-deck imputation. *Biometrika*. 1992; 79:811–822.
- Rubin DB. Inference and missing data. *Biometrika*. 1976; 63:581–592.
- Sampford MR. On sampling without replacement with unequal probabilities of selection. *Biometrika*. 1967; 54:499–513. [PubMed: 6064012]
- Shao J, Sitter RR. Bootstrap for imputed survey data. *Journal of the American Statistical Association*. 1996; 93:819–831.
- Shao J, Steel P. Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*. 1999; 94:254–265.
- Tillé, Y. *Sampling Algorithms*. Springer-Verlag; New York: 2006.

Table 1

Relative bias (RB), relative standard error (RSE) and relative root mean squared error (RRMSE) of all estimators

Estimators	RB(%)	RSE(%)	RRMSE(%)
$\hat{Y}_{COM}$	0.38	7.56	7.57
$\hat{Y}_I(00 10 10)$	0.77	8.62	8.65
$\hat{Y}_I(00 01 10)$	10.64	9.31	14.14
$\hat{Y}_I(00 10 01)$	-29.54	8.85	30.84
$\hat{Y}_I(00 01 01)$	-26.52	9.25	28.09
$\hat{Y}_{MR}(10 10 10)$	0.87	8.82	8.86
$\hat{Y}_{MR}(10 01 10)$	1.23	8.92	9.00
$\hat{Y}_{MR}(10 10 01)$	1.50	8.97	9.10
$\hat{Y}_{MR}(01 10 01)$	-5.97	10.41	12.00
$\hat{Y}_{MR}(01 11 11)$	0.65	8.64	8.66
$\hat{Y}_{MR}(11 01 11)$	1.19	8.92	9.00
$\hat{Y}_{MR}(11 10 11)$	0.89	8.75	8.79
$\hat{Y}_{MR}(11 11 11)$	0.89	8.75	8.79
$\hat{Y}_{MR}(01 11 01)$	-5.97	10.41	12.00
$\hat{Y}_{MR}(11 01 01)$	1.04	8.88	8.94
$\hat{Y}_{MR}(11 10 01)$	1.52	8.98	9.11
$\hat{Y}_{MR}(11 11 01)$	1.52	8.98	9.11

Table 2

Relative bias (RB) and Coverage rate (CR) of the jackknife variance estimator

Estimators	RB(%)	CR
$\hat{\bar{Y}}_{MR}(01 11 11)$	4.0	94.8
$\hat{\bar{Y}}_{MR}(10 11 11)$	5.7	95.0
$\hat{\bar{Y}}_{MR}(11 01 11)$	7.5	95.5
$\hat{\bar{Y}}_{MR}(11 10 11)$	7.6	95.5
$\hat{\bar{Y}}_{MR}(11 11 11)$	7.6	95.0