

Single imputation for populations containing a large amount of zeroes

D. Haziza*, C-O. Nambu[†] and G. Chauvet[‡]

Abstract

Single imputation is often used in surveys to compensate for item nonresponse. In some cases, the variable requiring imputation contains a large amount of zeroes. This is especially frequent in business surveys that collect economic variables. In this paper, we study the properties of two imputation procedures frequently used in practice and show that they generally lead to biased estimators. Motivated by a mixture regression model, we then propose three imputation procedures and study their properties in terms of bias. For the proposed imputation procedures, we consider a jackknife variance estimator that is consistent for the true variance, provided the overall sampling fraction is negligible. Finally, we perform a simulation study to evaluate the performance of point and variance estimators in terms of relative bias and mean square error.

Key words: balanced imputation; deterministic imputation; item nonresponse; random imputation; variance estimation.

*Department of Mathematics and Statistics. University of Montreal, Montreal, Canada

[†]Business Survey Methods Division, Statistics Canada, Ottawa, Canada

[‡]Laboratoire de Statistique d'Enquête, CREST/ENSAI, Campus de Ker Lann, 35170 Bruz, France

1 Introduction

Imputation is often used in surveys to treat item nonresponse. It consists of replacing missing values with imputed values, which are constructed on the basis of auxiliary information recorded for all the sample units (respondents and nonrespondents). The main objective of imputation is to reduce the nonresponse bias, which may be important if the responding units and the nonresponding units show different characteristics with respect to the characteristics of interest. In some situations, the variable being imputed contains a large amount of zeroes. This situation is frequent in business surveys, which collect economic variables (revenue, expenses, etc.) For example, in the context of the Capital Expenditure Survey conducted at Statistics Canada, the main variables are Capital Machinery (CM) and Capital Construction (CC). In a given year, a large number of businesses have not invested any amount of money for new machinery or construction. As a result, the data file contains a large amount of zeroes for these two variables. Typically, the proportion of businesses reporting a value of zero to the variable CC is as high as 70 % whereas it is close to 50 % for the variable CM.

Consider a finite population U of size N . In this paper, we are interested in estimating the population mean of a variable of interest y , $\bar{Y} = N^{-1} \sum_{i \in U} y_i$. To that end, we select a sample s , of size n , according to a sampling design $p(s)$. Let I_i be a sample selection indicator for unit i such that $I_i = 1$ if $i \in s$ and $I_i = 0$, otherwise. In the absence of nonresponse, a complete data estimator of \bar{Y} is the expansion estimator given by

$$\hat{Y}_\pi = N^{-1} \sum_{i \in U} w_i y_i I_i,$$

where $w_i = 1/\pi_i$ denotes the design weight attached to unit i and π_i denotes its inclusion probability in the sample. The estimator \hat{Y}_π is p -unbiased for \bar{Y} ; that is, $E_p(\hat{Y}_\pi) = \bar{Y}$, where $E_p(\cdot)$ denotes the expectation with respect to the sampling design. In the presence of nonresponse, only a subset of s responds to item y . In this case, we define an imputed estimator of \bar{Y} given by

$$\hat{Y}_I = \frac{1}{N} \left[\sum_{i \in s} w_i r_i y_i + \sum_{i \in s} w_i (1 - r_i) y_i^* \right], \quad (1)$$

where r_i is a response indicator attached to unit i such that $r_i = 1$ if unit i responded to item y and $r_i = 0$, otherwise and y_i^* denotes the imputed value used to replace the missing value y_i . Further, the sets of respondents and nonrespondents to item y are denoted by s_r and s_m , respectively.

To study the properties of the imputed estimator (1), we use the standard decomposition of the total error of \hat{Y}_I :

$$\hat{Y}_I - \bar{Y} = (\hat{Y}_\pi - \bar{Y}) + (\hat{Y}_I - \hat{Y}_\pi). \quad (2)$$

The first term on the right-hand side of (2) is called the sampling error of \hat{Y}_I , whereas the second term, $\hat{Y}_I - \hat{Y}_\pi$, is called the nonresponse error. In this paper, we focus on the nonresponse error that can be expressed as

$$\hat{Y}_I - \hat{Y}_\pi = -\frac{1}{N} \left[\sum_{i \in s} w_i (1 - r_i) (y_i - y_i^*) \right]. \quad (3)$$

In the case of random imputation, the nonresponse error can be further decomposed into

$$\hat{Y}_I - \hat{Y}_\pi = (\tilde{Y}_I - \hat{Y}_\pi) + (\hat{Y}_I - \tilde{Y}_I), \quad (4)$$

where $\tilde{Y}_I = E_I(\hat{Y}_I | \mathbf{y}, \mathbf{I}, \mathbf{r})$ denote the imputed estimator that one would have obtained had a deterministic imputation been used with $\mathbf{y} = (y_1, \dots, y_N)'$,

$\mathbf{I} = (I_1, \dots, I_N)'$, $\mathbf{r} = (r_1, \dots, r_N)'$ and the subscript I denotes the random imputation mechanism.

To study the properties of the imputed estimator (1), we consider two distinct approaches for inference: (i) the Nonresponse Model (NM) approach and (ii) the Imputation Model (IM) approach.

- (i) In the NM approach, inference is made with respect to the joint distribution induced by the sampling design and the nonresponse model. The nonresponse model is a set of assumptions about the unknown distribution of the response indicators \mathbf{r} , called the nonresponse mechanism. Let $p_i = P(r_i = 1 | \mathbf{I}; I_i = 1)$ be the response probability of unit i . In this paper, we assume that the units respond independently of one another; that is $p_{ij} = P(r_i = 1, r_j = 1 | \mathbf{I}; I_i = 1, I_j = 1) = p_i p_j$ if $i \neq j$. In this approach, we assume that, after conditioning on s , the nonresponse mechanism is independent of all other variables involved in the imputed estimator. Under deterministic imputation, we define the conditional nonresponse bias of \hat{Y}_I as

$$B_q(\hat{Y}_I) = E_q(\hat{Y}_I - \hat{Y}_\pi | \mathbf{y}, \mathbf{I}),$$

where the subscript q denotes the unknown nonresponse mechanism. Under random imputation, the conditional nonresponse bias of \hat{Y}_I is defined as

$$B_{qI}(\hat{Y}_I) = E_q(\tilde{Y}_I - \hat{Y}_\pi | \mathbf{y}, \mathbf{I}).$$

When the nonresponse mechanism is uniform (i.e., $p_i = p$, say) we are in the case of the uniform nonresponse model (UNM) approach. In practice, it is customary to form imputation classes so that within

an imputation class, units have approximately the same probability of response. Imputation is then performed independently within each class.

- (ii) In the IM approach, inference is made with respect to the joint distribution induced by the imputation model, the sampling design, and the nonresponse model. The imputation model is a set of assumptions about the unknown distribution \mathbf{y} . In this approach, we assume that the distribution of the model errors $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_N)'$ does not depend on \mathbf{I} and \mathbf{r} , after conditioning on the appropriate auxiliary information. In the presence of zeroes to item y , the population U can be viewed as the mixture of two-subpopulations: $U_0 \subset U$, of size N_0 , which denotes the subpopulation of units for which the item of interest y is equal to zero and $U_1 \subset U$, of size N_1 , denotes the subpopulation of units for which y is positive. We have $U = U_0 \cup U_1$ et $N = N_0 + N_1$. In the case of regression imputation, the underlying imputation model is the following mixture regression model:

$$m : y_i = \delta_i(\mathbf{z}_i^\top \boldsymbol{\beta} + \epsilon_i) + (1 - \delta_i) \times 0, \quad (5)$$

where \mathbf{z} is a vector of auxiliary variables available for all the sample units (respondents and nonrespondents), $\boldsymbol{\beta}$ is a vector of unknown parameters and

$$\delta_i = \begin{cases} 1 & \text{if } i \in U_1 \\ 0 & \text{if } i \in U_0. \end{cases}$$

Let $\phi_i = P(\delta_i = 1)$ be the probability of unit i belonging to U_1 . We make the following assumptions: $E(\epsilon_i | \delta_i = 1) = 0$, $E(\epsilon_i \epsilon_j | \delta_i = 1, \delta_j =$

$1, i \neq j) = 0$ and $V(\epsilon_i|\delta_i = 1) = \sigma^2 c_i$, where $c_i = \boldsymbol{\lambda}^\top \mathbf{z}_i$, where $\boldsymbol{\lambda}$ is a vector of known constants. It follows that

$$\begin{aligned} E_m(y_i) &= \phi_i \mathbf{z}_i^\top \boldsymbol{\beta} + E(\epsilon_i|\delta_i = 1)P(\delta_i = 1) \\ &= \phi_i \mathbf{z}_i^\top \boldsymbol{\beta} \end{aligned}$$

and

$$\begin{aligned} V_m(y_i) &= \phi_i V(y_i|\delta_i = 1) + \phi_i(1 - \phi_i)E(y_i|\delta_i = 1)^2 \\ &= \sigma^2 \phi_i c_i + \phi_i(1 - \phi_i)(\mathbf{z}_i^\top \boldsymbol{\beta})^2, \end{aligned}$$

where the subscript m denotes the imputation model (5). A special case of (5) is the mixture mean model, which is obtained from (5) by setting $\mathbf{z}_i = 1$ and $c_i = 1$. Under deterministic imputation, we define the conditional nonresponse bias of \hat{Y}_I as

$$B_{qm}(\hat{Y}_I) = E_q E_m(\hat{Y}_I - \hat{Y}_\pi | \mathbf{I}, \mathbf{r}).$$

Under random imputation, the conditional nonresponse bias of \hat{Y}_I is defined as

$$B_{qmI}(\hat{Y}_I) = E_q E_m(\tilde{Y}_I - \hat{Y}_\pi | \mathbf{I}, \mathbf{r}).$$

In this paper, we seek an imputation procedure that satisfies simultaneously the following three criteria:

- (a) It leads to an asymptotically unbiased estimator under either the NM or the IM approach.
- (b) It leads to realistic imputed values in the sense that it reflects the nature of the data.

(c) It is fully efficient.

Imputation procedures satisfying the criterion (a) are often called doubly robust procedures; e.g. Haziza and Rao (2006) and Kim and Haziza (2010). The criterion (b) suggests that, since the population contains a large amount of zeroes, the imputed values must reflect the nature of the data. That is, realistic imputed values would consist of a mixture of null and positive values. Finally, the criterion (c) is satisfied when the resulting imputed estimator does not suffer from the extra variability due to the random selection of imputed values when a random imputation procedure is used; see, e.g., Kim and Fuller (2004). Deterministic imputation procedures are thus automatically fully efficient.

The paper is organized as follows: in Section 2, we study the properties of two imputation procedures often used in practice. In Section 3, we study the properties of three imputation procedures that are all motivated by a mixture regression model. We show that, unlike the usual imputation procedures used in practice, the proposed methods lead to asymptotically unbiased estimators of population means. For the latter procedures, we propose, in Section 4, a jackknife variance estimator, which is consistent for the true variance, provided the overall sampling fraction is negligible. In Section 5, we conduct a limited simulation study to investigate the performance of the proposed imputation procedures in terms of bias and relative efficiency. We also study the properties of the jackknife variance estimator. Finally, we conclude in Section 6.

2 Usual imputation procedures

In this section, we describe two imputation procedures that are often used in practice and study their properties in terms of bias under both the NM approach and the IM approach.

2.1 Imputation using the positive respondents only

In this section, we study the properties of the imputed estimator (1) when the imputed values are constructed on the basis of the responding units that belong to $s_1 = s \cap U_1$. In other words, the units with a zero value attached to item y are deleted and a regression model is fitted using the positive observations. In this case, the imputed values are given by

$$y_i^* = \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}, \quad (6)$$

where

$$\hat{\mathbf{B}}_{r_1} = \left(\sum_{i \in s_1} w_i r_i c_i^{-1} \mathbf{z}_i \mathbf{z}_i^\top \right)^{-1} \left(\sum_{i \in s_1} w_i r_i c_i^{-1} \mathbf{z}_i y_i \right). \quad (7)$$

We call this imputation deterministic positive regression (DPR) imputation. If the imputed values (6) are used in (1), it can be shown that the asymptotic conditional nonresponse bias of \hat{Y}_I under the NM approach is given by

$$B_q \left(\hat{Y}_I \right) \doteq -\frac{1}{N} \sum_{i \in s} w_i (1 - p_i) (y_i - \mathbf{z}_i^\top \hat{\mathbf{B}}_{p_1}), \quad (8)$$

where $\hat{\mathbf{B}}_{p_1} = \left(\sum_{i \in s_1} w_i p_i c_i^{-1} \mathbf{z}_i \mathbf{z}_i^\top \right)^{-1} \left(\sum_{i \in s_1} w_i p_i c_i^{-1} \mathbf{z}_i y_i \right)$. The asymptotic bias in (8) is not equal to zero, in general unless $p_i = 1$ for all i (i.e., the complete data case). Note that the bias (8) does not vanish, even under

uniform nonresponse, $p_i = p$ (i.e., UNM approach). Now, assuming model (5), the conditional nonresponse bias of \hat{Y}_I under the IM approach is given by

$$B_{qm}(\hat{Y}_I) = \frac{1}{N} \sum_{i \in s} w_i (1 - p_i) (1 - \phi_i) \mathbf{z}_i^\top \boldsymbol{\beta}. \quad (9)$$

Once again, the bias given by (9) does not vanish unless (i) $p_i = 1$ for all i (complete data case) or $\phi_i = 1$ (i.e., $U_0 = \emptyset$). As we show in Section 5, the imputed values (6) may lead to substantial bias.

2.2 Imputation based on all the responding units

In this section, we study the properties of \hat{Y}_I under deterministic regression (DR) imputation, which is, without a doubt, the most popular imputation procedure used in the case of populations containing a large amount of zeroes. In this case, the imputed values are given by

$$y_i^* = \mathbf{z}_i^\top \hat{\mathbf{B}}_r, \quad (10)$$

where $\hat{\mathbf{B}}_r = (\sum_{i \in s} w_i r_i c_i^{-1} \mathbf{z}_i \mathbf{z}_i^\top)^{-1} (\sum_{i \in s} w_i r_i c_i^{-1} \mathbf{z}_i y_i)$. Under the UNM approach, it is easily seen that $B_q(\hat{Y}_I) \doteq 0$. Therefore, as long as the units have identical response probabilities, the imputed estimator \hat{Y}_I is asymptotically unbiased for \bar{Y}_I , regardless of the shape of the relationship between y and \mathbf{z} .

Under the IM approach, the asymptotic conditional nonresponse bias of \hat{Y}_I is given by

$$B_{qm}(\hat{Y}_I) \doteq \frac{1}{N} \sum_{i \in s} w_i \mathbf{z}_i^\top (\hat{\mathbf{T}}_p^{-1} \hat{\mathbf{T}}_{p\phi} - \phi_i \mathbf{I}_q) \boldsymbol{\beta}, \quad (11)$$

where $\hat{\mathbf{T}}_p = \sum_{i \in s} w_i p_i c_i^{-1} \mathbf{z}_i \mathbf{z}_i^\top$, $\hat{\mathbf{T}}_{p\phi} = \sum_{i \in s} w_i p_i \phi_i c_i^{-1} \mathbf{z}_i \mathbf{z}_i^\top$ and \mathbf{I}_q is the identity matrix of order q . To better understand (11), consider the mixture mean

model, which is a special case of (5) with $\mathbf{z}_i = 1$ and $c_i = 1$ for all i . In this case, the expression (11) reduces to

$$B_{qm}(\hat{Y}_I) \doteq \hat{P}^{-1} \beta \sum_{i \in s} w_i (p_i - \hat{P})(\phi_i - \hat{\Phi}), \quad (12)$$

where $\hat{P} = \sum_{i \in s} w_i p_i / \sum_{i \in s} w_i$ et $\hat{\Phi} = \sum_{i \in s} w_i \phi_i / \sum_{i \in s} w_i$. The bias in (12) vanishes if the p_i 's and the ϕ_i 's are unrelated, which occurs, for example, if $p_i = p$ (uniform nonresponse mechanism) or if $\phi_i = \phi$ (i.e., when the probability of being strictly positive is constant). Also, note that the bias decreases as the average of the response probabilities increases, as expected. Similarly, the bias in (11) vanishes when $p_i = p$ or when $\phi_i = \phi$. As we show in Section 5, the imputed estimator \hat{Y}_I is biased if the p_i 's and ϕ_i 's are related.

3 Proposed imputation procedures

In this section, we propose three imputation procedures that are motivated by the mixture regression model (5) and study their properties under both the NM and IM approaches. In Sections 3.1-3.3, we assume that the ϕ_i 's are known. The estimation of the ϕ_i 's is discussed in Section 3.4.

3.1 Deterministic imputation

Motivated by the mixture regression model (5), we first propose to use the imputed values

$$y_i^* = \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}, \quad i \in s_m, \quad (13)$$

where $\hat{\mathbf{B}}_{r_1}$ is given by (7). Using the imputed values (13) in (1) leads to

$$\hat{Y}_I = \frac{1}{N} \left[\sum_{i \in s} w_i r_i y_i + \sum_{i \in s} w_i (1 - r_i) \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} \right]. \quad (14)$$

We call this imputation procedure deterministic ϕ -regression (DR- ϕ) imputation. Using a first-order Taylor expansion, one can show that \hat{Y}_I given by (14) is asymptotically unbiased under either the UNM approach or the IM approach. The proofs are sketched in the Appendix. As result, this imputation method satisfies the criterion (a). The criterion (c) is also satisfied since we are in presence of a deterministic imputation method. However since $0 < y_i^* \leq \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}$, the criterion (b) is not met since this imputation cannot provide a mixture of zero and positive values.

3.2 Random imputation

The imputed values given by (13) do not satisfy the criterion (b). To overcome the problem, we propose the following random imputation procedure, for which the imputed values are given by

$$y_i^* = \begin{cases} \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} & \text{with probability } \phi_i \\ 0 & \text{with probability } 1 - \phi_i, \end{cases} \quad (15)$$

for $i \in s_m$, where $\hat{\mathbf{B}}_{r_1}$ is given by (7). We call this imputation procedure random ϕ -regression (RR- ϕ) imputation. Noting that $E_I(y_i^* | \mathbf{y}, \mathbf{I}, \mathbf{r}) = \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}$, it follows that \tilde{Y}_I reduces to (14). As a result, the imputed estimator \hat{Y}_I obtained by replacing y_i^* by (15) in (1), is asymptotically unbiased under either the UNM approach or the IM approach. Also, from a micro data point of view, the imputed values (15) are preferable to (13) because they lead

to realistic imputed values. Hence, this imputation procedure satisfies both the criteria (a) and (b). However, the criterion (c) is not satisfied since this imputation procedure suffers from an extra variability due to the random selection of the imputed values. In fact, noting that

$$V_I(y_i^*|\mathbf{y}, \mathbf{I}, \mathbf{r}) = \phi_i(1 - \phi_i)(\mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1})^2,$$

the imputation variance of \hat{Y}_I is given by

$$EV_I(\hat{Y}_I|\mathbf{r}) = \frac{1}{N^2} \left[E \left(\sum_{i \in s} w_i^2 (1 - r_i) \phi_i (1 - \phi_i) (\mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1})^2 \right) \right], \quad (16)$$

where $E(\cdot)$ in (16) stands for $E_p E_q(\cdot)$ in the context of the NM approach and $E_m E_q E_p(\cdot)$ in the context of the IM approach. In some cases, the contribution of the imputation variance (16) to the total variance may be important, leading to a potentially inefficient estimator.

3.3 Balanced random imputation

Following Chauvet, Deville et Haziza (2010), we consider a balanced random imputation procedure, which consists of selecting the imputed values y_i^* so that the imputation error, $\tilde{Y}_I - \hat{Y}_I$, is equal to zero. If the imputation error is eliminated, so is the imputation variance. Here, the goal is to select the imputation values y_i^* so that the following constraint is satisfied:

$$\hat{Y}_I - \tilde{Y}_I = \frac{1}{N} \left[\sum_{i \in s} w_i (1 - r_i) (y_i^* - \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}) \right] = 0. \quad (17)$$

If the balancing equation (17) is exactly satisfied, then the imputation variance is completely eliminated. Hence, in addition to the criteria (a) and

(b), the proposed imputation procedure, which we call balanced random ϕ -regression (BRR- ϕ) imputation, satisfies the criterion (c).

We now turn to the algorithm for selecting the imputed values y_i^* under constraint (17). From s_m , we select a subset s_* with inclusion probabilities ϕ_i using the balancing variable

$$x_i = w_i \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}.$$

For $i \in s_m$, we set $y_i^* = \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}$ if $i \in s_*$ and $y_i^* = 0$, otherwise. Balanced sampling on the x -variable ensures that

$$\sum_{i \in s_*} \frac{x_i}{\phi_i} = \sum_{i \in s_m} x_i,$$

which is equivalent to

$$\sum_{i \in s_*} w_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} = \sum_{i \in s} w_i (1 - r_i) \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}. \quad (18)$$

Now, note that

$$\begin{aligned} \sum_{i \in s} w_i (1 - r_i) y_i^* &= \sum_{i \in s_m} w_i y_i^* \\ &= \sum_{i \in s_*} w_i y_i^* + \sum_{i \in s_m \setminus s_*} w_i y_i^* \\ &= \sum_{i \in s_*} w_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1}, \end{aligned} \quad (19)$$

where $y_i^* = 0$ if $i \in s_m \setminus s_*$. Using the equations (18) and (19), we obtain

$$\sum_{i \in s} w_i (1 - r_i) y_i^* = \sum_{i \in s} w_i (1 - r_i) \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1},$$

which is equivalent to (17). As a result, the imputation variance is eliminated and this imputation procedure satisfies the criteria (a)-(c).

3.4 Estimation of ϕ_i

In practice, the ϕ_i 's are unknown and need to be estimated. We assume that

$$\phi_i = f(\mathbf{u}_i; \boldsymbol{\alpha}),$$

for some function $f(\cdot)$, where \mathbf{u}_i is a vector of variables recorded for all the sample units (respondents and nonrespondents) attached to unit i and $\boldsymbol{\alpha}$ is a vector of unknown parameters. Parametric methods such as logistic regression require the specification of $f(\cdot)$. Alternatively, one could use nonparametric methods such as kernel-type smoothing methods or local polynomial regression (see, e.g., Wand and Jones, 1995). Unlike parametric methods, the nonparametric methods do not require the specification of $f(\cdot)$ and are thus more robust to model misspecification. The appropriate selection of auxiliary variables \mathbf{u} to include in the model is important. A careful modeling exercise should then be conducted in order to select the auxiliary variables that explain the variable δ well. If a parametric method is used, an estimate of ϕ_i is given by $\hat{\phi}_i = f(\mathbf{u}_i; \hat{\boldsymbol{\alpha}})$, where $\hat{\boldsymbol{\alpha}}$ is any consistent estimator of $\boldsymbol{\alpha}$.

4 Jackknife variance estimation

In this section, we discuss the problem of variance estimation in the presence of imputed data. It is well known that treating the imputed values as real values leads to an underestimation of the variance of the imputed estimators. Several variance estimation methods, taking the nonresponse and imputation into account, have been proposed in the literature. In particular, resampling methods have been adapted in order to provide consistent

variance estimators. For example, Rao and Shao (1992) proposed a jackknife variance estimator, which is computed in the usual way except that the imputed values are adjusted whenever a responding unit is deleted. Rao and Shao (1992) showed that the resulting jackknife variance estimator is consistent, provided the overall sampling fraction is negligible. In this section, we consider an alternative jackknife variance estimator that does not require adjusting the imputed values. As for the Rao-Shao jackknife variance estimator, it is asymptotically unbiased and consistent for the true variance, provided the sampling fraction n/N is negligible. To motivate our jackknife procedure, we use the reverse framework to express the total variance of \hat{Y}_I ; see Fay (1991) and Shao and Steel (1999). In the case of deterministic imputation and the NM approach, the total variance of \hat{Y}_I can be written as

$$V_T^{NM} = V_1^{NM} + V_2^{NM},$$

where $V_1^{NM} = E_q V_p \left(\hat{Y}_I | \mathbf{y}, \mathbf{r} \right)$ and $V_2^{NM} = V_q E_p \left(\hat{Y}_I | \mathbf{y}, \mathbf{r} \right)$.

Under the IM approach, the total variance of \hat{Y}_I is given by

$$V_T^{IM} = V_1^{IM} + V_2^{IM},$$

where $V_1^{IM} = E_m E_q V_p \left(\hat{Y}_I - \bar{Y} | \mathbf{y}, \mathbf{r} \right)$ and $V_2^{IM} = E_q V_m E_p \left(\hat{Y}_I - \bar{Y} | \mathbf{y}, \mathbf{r} \right)$.

Under mild regularity conditions, the contribution of the term V_2^{NM} (respectively, V_2^{IM}) to the total variance, V_2^{NM}/V_T^{NM} (respectively, V_2^{IM}/V_T^{IM}) is of order $O(n/N)$. Thus, when the sampling fraction n/N is negligible, the contribution of the term V_2^{NM} (respectively V_2^{IM}) to the total variance is negligible and, as a result, can be omitted from the variance calculations. It remains to estimate the term V_1^{NM} (respectively V_1^{IM}) consistently, which requires obtaining a consistent estimator of $V_p \left(\hat{Y}_I | \mathbf{y}, \mathbf{r} \right)$. Conditionally given

\mathbf{y} and \mathbf{r} , the estimator \hat{Y}_I can be expressed as a smooth function of estimated totals. The problem of estimating $V_p\left(\hat{Y}_I|\mathbf{y}, \mathbf{r}\right)$ reduces to the classical problem of estimating the sampling variance of a smooth function of estimated totals conditionally given \mathbf{y} and \mathbf{r} . Hence, any complete data variance estimation method can thus be used (e.g., Taylor linearization, jackknife or bootstrap).

Under random imputation and the NM approach, the total variance of \hat{Y}_I can be expressed as

$$V_T^{NM} = \tilde{V}_1^{NM} + \tilde{V}_2^{NM} + \tilde{V}_3^{NM},$$

where $\tilde{V}_1^{NM} = E_q V_p\left(\tilde{Y}_I|\mathbf{y}, \mathbf{r}\right)$, $\tilde{V}_2^{NM} = V_q E_p\left(\tilde{Y}_I|\mathbf{y}, \mathbf{r}\right)$ and $\tilde{V}_3^{NM} = E_q E_p\left(V_I\left(\hat{Y}_I|\mathbf{y}, \mathbf{I}, \mathbf{r}\right)|\mathbf{y}, \mathbf{r}\right)$.

Under the IM approach, the total variance of \hat{Y}_I is given by

$$V_T^{IM} = \tilde{V}_1^{IM} + \tilde{V}_2^{IM} + \tilde{V}_3^{IM},$$

where $\tilde{V}_1^{IM} = E_m E_q V_p\left(\tilde{Y}_I - \bar{Y}|\mathbf{y}, \mathbf{r}\right)$, $\tilde{V}_2^{IM} = E_q V_m E_p\left(\tilde{Y}_I - \bar{Y}|\mathbf{y}, \mathbf{r}\right)$ and $\tilde{V}_3^{IM} = E_m E_q E_p\left(V_I\left(\hat{Y}_I - \bar{Y}|\mathbf{y}, \mathbf{I}, \mathbf{r}\right)|\mathbf{y}, \mathbf{r}\right)$.

Note that \tilde{V}_3^{NM} (respectively \tilde{V}_3^{IM}) represents the imputation variance under the NM approach (respectively the IM approach), which is due to the random selection of imputed values. As with the case of deterministic imputation, a consistent estimator of \tilde{V}_1^{NM} (respectively \tilde{V}_1^{IM}) is obtained by estimating consistently $V_p\left(\tilde{Y}_I|\mathbf{y}, \mathbf{r}\right)$. Since \tilde{Y}_I can be expressed as a smooth function of totals, estimating $V_p\left(\tilde{Y}_I|\mathbf{y}, \mathbf{r}\right)$ reduces to the problem of estimating the sampling variance of a smooth function of totals. Once again, we can omit the term V_2^{NM} (respectively V_2^{IM}) from the computations when the sampling fraction n/N is negligible.

Now, let $\hat{\phi}_{i(j)}$ denote the estimated probability for unit i when unit j has been deleted. Also, let $w_{i(j)}$ be the so-called jackknife weights given by

$$w_{i(j)} = \begin{cases} w_i \frac{n}{n-1} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

For DR- ϕ imputation, a consistent estimator of either V_1^{NM} or V_1^{IM} is given by

$$\hat{V}_J = \frac{n-1}{n} \sum_{i \in s} (\hat{Y}_{I(j)} - \hat{Y}_I)^2, \quad (20)$$

where \hat{Y}_I is given by (14),

$$\hat{Y}_{I(j)} = \frac{1}{\sum_{i \in s} w_{i(j)}} \left[\sum_{i \in s} w_{i(j)} r_i y_i + \sum_{i \in s} w_{i(j)} (1 - r_i) \hat{\phi}_{i(j)} \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1(j)} \right]$$

and $\hat{\mathbf{B}}_{r_1(j)}$ is computed the same way as $\hat{\mathbf{B}}_{r_1}$ but with the jackknife weights $w_{i(j)}$ instead of the original weights w_i . Since the sampling fraction n/N is assumed to be negligible, the estimator \hat{V}_J in (20) is a consistent estimator of the total variance of \hat{Y}_I under either the NM approach or the IM approach. That is, we use $\hat{V}_T = \hat{V}_J$.

For RR- ϕ imputation, the imputation variance must be taken into account. An estimator of either \tilde{V}_3^{NM} or \tilde{V}_3^{IM} is given by

$$\hat{V}_I = \frac{1}{N^2} \left[\sum_{i \in s} w_i^2 (1 - r_i) \hat{\phi}_i (1 - \hat{\phi}_i) (\mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1})^2 \right]. \quad (21)$$

Noting that \tilde{Y}_I is given by (14), a consistent estimator of the total variance of \hat{Y}_I under either the NM approach or the IM approach is given by

$$\hat{V}_T = \hat{V}_J + \hat{V}_I, \quad (22)$$

where \hat{V}_J is given by (20) and \hat{V}_I is given by (21).

Finally, for BRR- ϕ imputation, we can omit \tilde{V}_3^{NM} (respectively \tilde{V}_3^{IM}) since the imputation variance was eliminated under this procedure. As a result, a consistent estimator of the total variance of \hat{Y}_I under either the NM approach or the IM approach is also given by (20).

5 Simulation Study

5.1 Performance of point estimators

We performed a limited simulation study to assess the performance of the proposed imputation procedures in terms of relative bias and mean square error. We generated three finite populations of size $N = 1000$, each consisting of a variable of interest y and an auxiliary variable z . First, the z -values were independently generated from a Gamma distribution with shift parameter $\alpha = 4$ and scale parameter $\beta = 25$. Then, the y -values were generated according to ratio model

$$y = 2z + \epsilon \tag{23}$$

where the ϵ_i 's were generated from a Normal distribution with mean 0 and variance σ^2 , which value was set so that we obtained a coefficient of determination R^2 equal to 0.36 for population 1, equal to 0.5 for population 2 and equal to 0.7 for population 3. In each population, we generated either 25 % or 50 % of zeroes according to given mechanisms, which are described next. For unit i , we then generated the indicator δ_i from a Bernoulli distribution with parameter ϕ_i . More specifically, the zeroes to the variable y were gen-

erated according to three distinct mechanisms (called ϕ -mechanisms), which are described below:

(1) ϕ -mechanism 1: it is the uniform ϕ -mechanism. That is, we set $\phi_i = 0.50$ (respectively $\phi_i = 0.75$) for all $i \in U$.

(2) ϕ -mechanism 2: the probability ϕ_i attached to unit i is defined as

$$\log \left(\frac{\phi_i}{1 - \phi_i} \right) = \lambda_0 + \lambda_1 z_i, \quad (24)$$

where the parameters λ_0 and λ_1 were chosen so that the average of the ϕ_i 's was approximately equal to 0.5% (respectively, 0.75%).

(3) ϕ -mechanism 3: the probability ϕ_i is obtained from (24), by replacing z_i with y_i .

When we obtained $\delta_i = 0$, we set $y_i = 0$, whereas we used the original y -value generated from (23) when we obtained $\delta_i = 1$. We were interested in estimating the population mean, $\bar{Y} = \sum_{i \in U} y_i / N$.

From each population, we selected $R = 10,000$ simple random samples without replacement of size $n = 200$. In each sample, we generated the response indicator r_i from a Bernoulli distribution with parameter p_i . Specifically, nonresponse to item y was generated according to three nonresponse mechanisms (called p -mechanisms), which are described below:

(1) p -mechanism 1: it is the uniform nonresponse mechanism. That is, we set $p_i = 0.7$ for all i .

(2) p -mechanism 2: the response probability p_i attached to unit i is defined as

$$\log \left(\frac{p_i}{1 - p_i} \right) = \eta_0 + \eta_1 z_i \quad (25)$$

where the parameters η_0 and η_1 were chosen so that the average of the p_i 's was approximately equal to 0.7%.

- (3) p -mechanism 3: the probability ϕ_i is obtained from (25), by replacing z_i with y_i .

We were interested in comparing the following five imputation methods:

- (i) DPR imputation, for which the imputed values by (6) with $\mathbf{z}_i = z_i$ and $c_i = z_i$.
- (ii) DR imputation, for which the imputed values by (10) with $\mathbf{z}_i = z_i$ and $c_i = z_i$.
- (iii) DR- ϕ imputation, for which the imputed values are given by (13) with $\mathbf{z}_i = z_i$ and $c_i = z_i$.
- (iv) RR- ϕ imputation, for which the imputed values are given by (15) with $\mathbf{z}_i = z_i$ and $c_i = z_i$.
- (v) BRR- ϕ imputation, for which the imputed values are given by (15) with $\mathbf{z}_i = z_i$ and $c_i = z_i$, while satisfying the balancing equation (17).

For the imputation methods (iii)-(v), we estimated the ϕ_i 's using a logistic regression model:

$$\phi_i = \exp(\boldsymbol{\mu}_i^\top \boldsymbol{\alpha}) / (1 + \exp(\boldsymbol{\mu}_i^\top \boldsymbol{\alpha})), \quad (26)$$

where $\boldsymbol{\mu}_i$ denotes a vector of auxiliary variables and $\boldsymbol{\alpha}$ is a vector of unknown parameters. When the zeroes were generated according to the ϕ -mechanism 1, we estimated the ϕ_i 's using (26) with $\boldsymbol{\mu}_i = 1$ and $\boldsymbol{\alpha} = \alpha_0$, whereas we

used $\boldsymbol{\mu}_i = (1, z_i)^\top$ and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^\top$ for the ϕ -mechanism 2 and for the ϕ -mechanism 3.

Finally, we computed the imputed estimator \hat{Y}_I given by (1). As a measure of bias of \hat{Y}_I , we use the monte carlo percent relative bias (in %) given by

$$RB_{MC}(\hat{Y}_I) = \frac{1}{R} \sum_{r=1}^R \frac{(\hat{Y}_{I(r)} - \bar{Y})}{\bar{Y}} * 100\%, \quad (27)$$

where $\hat{Y}_{I(r)}$ denotes the imputed estimator \hat{Y}_I in the r -th sample. As measure of variability, we used the monte carlo mean square error given by

$$EQM_{MC}(\hat{Y}_I) = \frac{1}{R} \sum_{r=1}^R (\hat{Y}_{I(r)} - \bar{Y})^2. \quad (28)$$

In the first part of the simulation, we compared the imputation methods DPR, RR- ϕ and RR in terms of relative bias and relative efficiency, defined as

$$RE = \frac{EQM_{MC}(\hat{Y}_I^{(.)})}{EQM_{MC}(\hat{Y}_I^{(DR)})}, \quad (29)$$

where \hat{Y}_I^{DR} denotes the imputed estimator under DR and $(.)$ denote either DPR or RR- ϕ . The results are shown in Tables 1-3.

In the second part of the simulation study, we focussed on the proposed imputation methods: DR- ϕ , RR- ϕ and BRR- ϕ for a subset of the scenarios considered in the first part. We restricted ourselves to populations that contained 50% of zeroes and that showed a coefficient of determination R^2 equal to 0.5. Also, we only considered the scenarios for which the zeroes were generated according to ϕ -mechanism 1 and ϕ -mechanism 2. As a measure of

RE, we used (29), using RR- ϕ imputation as the reference. That is, we used (29), where the denominator was replaced by $\hat{Y}_I^{(\text{RR}-\phi)}$, which denotes the imputed estimator \hat{Y}_I under RR- ϕ . The results are presented in Table 4.

We first discuss the results shown in Tables 1-3. It is clear from the the results that DPR imputation led to positively heavily biased estimators in all the scenarios, as expected. Also, the bias increased as the proportion of zeroes increased.

We now turn to the case of DR imputation. In the case of ϕ -mechanism 1, the imputed estimator showed negligible bias under p -mechanism 1 or p -mechanism 2. This can be explained by the fact that, in this case, the ϕ_i 's and the p_i 's were unrelated. For p -mechanism 3, the imputed estimator showed appreciable bias. This result is not surprising since p -mechanism 3 is nonignorable, the probability of response depending explicitly on the variable being imputed. In the case of ϕ -mechanism 2, the imputed estimator showed negligible bias under p -mechanism 1 since the ϕ_i 's and the p_i 's were, once again, unrelated. Under p -mechanism 2, we note that the imputed estimator was biased. For example, in Table 1, the estimator showed a RB of 7.87% for a proportion of zeroes of 50%. In this case, the ϕ_i 's and the p_i 's were related since both p_i and ϕ_i depended on z_i . Finally, in the case of ϕ -mechanism 3, the imputed estimator showed negligible bias under p -mechanism 1 for the same reasons as above. In the case of p -mechanism 2, the imputed estimator was biased. Once again, this can be explained by the fact that the ϕ_i 's and the p_i 's were related. Also, the bias increased as the coefficient of determination, R^2 , increased. For example, for a proportion of zeroes of 50%, the RB was equal to 3.93% for $R^2 = 0.36$, whereas it equalled 7.93% for $R^2 = 0.7$. Moreover,

the RB increased as the proportion of zeroes increased. For example, we note from Table 3 that the RB was equal to 3.76% for a proportion of zeroes for 25%, whereas it was equal to 7.70% for a proportion of zeroes of 50%. For p -mechanism 3, the imputed estimator showed appreciable bias. The bias decreased as the R^2 increased. This result is not surprising, since the imputation model became richer as the R^2 increased thus achieving a greater bias reduction.

We now discuss the case of RR- ϕ imputation. Under ϕ -mechanism 1, we note that the imputed estimator was approximately unbiased (except under p -mechanism 3, as expected). In the case of ϕ -mechanism 2, the imputed estimator showed negligible bias under both p -mechanism 1 and p -mechanism 2, unlike DR, which led to biased estimators under p -mechanism 2. For example, when the proportion of zeroes was equal to 50%, we note from Table 3 that, under p -mechanism 2, RR- ϕ imputation showed a RB of 0.80%, whereas it equalled 7.70% for DR imputation. Under p -mechanism 3, the imputed estimator was biased for both DR and RR- ϕ imputation. However, the bias under RR- ϕ imputation was significantly smaller than that under DR imputation. For example, for a proportion of zeroes equal to 50%, we note from Table 3 that DR imputation showed a RB of 10.28%, whereas we obtained 2.93% for RR- ϕ imputation. Thus, the mixture regression model (5), motivating RR- ϕ imputation, seems to provide a better description of the true model than the usual regression model, motivating DR imputation. As a result, we were able to achieve a greater bias reduction with RR- ϕ imputation. Finally in the case of ϕ -mechanism 3, we obtained similar results than those obtained under ϕ -mechanism 2, except that, when $R^2 = 0.36$, the

imputed estimator showed a slight bias under RR- ϕ imputation.

We now turn to RE. For DPR imputation, the values of RE were all greater than 1. This was mainly due to bias, which dominated the MSE. When both DR and RR- ϕ imputation led to negligible bias, we note that the values of RE were all greater than 1, showing that DR imputation was more efficient than RR- ϕ . This can be explained by the fact that, unlike DR imputation, RR- ϕ suffered from the imputation variance due to the random selection of the imputed values.

Finally, we discuss the results presented in Table 4. The three methods DR- ϕ , RR- ϕ and BRR- ϕ all led to negligible bias, except under p -mechanism 3, as expected. In terms of RE, both DR- ϕ and BRR- ϕ imputation showed values of RE smaller than 1, which is not surprising since RR- ϕ suffers from the additional variability due to the random selection of the imputed values, unlike DR- ϕ and BRR- ϕ imputation. Also, both DR- ϕ and BRR- ϕ imputation showed almost identical results, illustrating that the imputation variance was eliminated under BRR- ϕ imputation.

5.2 Performance of jackknife variance estimators

We performed a limited simulation study to assess the performance of the proposed jackknife variance estimator in terms of relative bias. We generated a population of size $N = 5000$ consisting of two variables: a variable of interest y and an auxiliary variable z . We first generated z according to a Gamma distribution with shift parameter $\alpha = 4$ and scale parameter $\beta = 25$. The y -values are then generated according to (23) such that the coefficient of determination R^2 was equal to 0.5.

Then, the zero values to y were generated according to ϕ -mechanism 1 and ϕ -mechanism 2 (see Section 5.1) so that the proportion of zeroes was equal to 10% and 25%, respectively.

In each population, we selected $R = 10,000$ simple random samples without replacement, of size $n = 150$. Note that the sampling fraction n/N is equal to 0.03, which can be considered negligible. Nonresponse to item y was generated according to an uniform nonresponse mechanism with probability 0.7.

As a measure of bias of a variance estimator \hat{V} , we used the monte carlo percent relative bias given by

$$RB_{MC}(\hat{V}) = 100 \times \frac{E_{MC}(\hat{V}) - V_{MC}(\hat{Y}_I)}{V_{MC}(\hat{Y}_I)} \quad (30)$$

where $E_{MC}(\hat{V}) = \frac{1}{R} \sum_{r=1}^R \hat{V}_{(r)}$ and $\hat{V}_{(r)}$ denotes the estimator \hat{V} in the r -th sample and $V_{MC}(\hat{Y}_I) = E_{MC}(\hat{V} - E_{MC}(\hat{Y}_I))^2$. As a measure of efficiency of a variance estimator \hat{V} , we used the mean square error given by:

$$EQM_{MC}(\hat{V}) = E_{MC}(\hat{V} - V_{MC}(\hat{Y}_I))^2. \quad (31)$$

For DR- ϕ and BRR- ϕ imputation, we used (20) as an estimator of the total variance, whereas we used (22) for RR- ϕ imputation.

Table 5 shows the monte carlo percent relative bias (RB) and mean square error of the jackknife variance estimators. It is clear from Table 5 that the proposed estimators performed well in all the scenarios with an absolute relative bias less than 6%. In terms of MSE, we note that the jackknife variance estimator corresponding to BRR- ϕ was more stable than the one corresponding to RR- ϕ . This is most likely because the variance estimator

corresponding to BRR- ϕ does not contain the additional term in the variance resulting from the random selection of the imputed values as in the case of RR- ϕ imputation.

6 Conclusion

Populations containing a large amount of zeroes are frequent in surveys. In this context, we proposed several imputation procedures that were all motivated by a mixture regression model. The simulation results showed that the proposed procedures performed well in all the scenarios. In our view, BRR- ϕ is particularly attractive because it satisfies simultaneously the criteria (a)-(c). We also proposed a jackknife variance estimator, which does not require the imputed values to be imputed, unlike in the Rao-Shao procedure. The proposed jackknife variance estimator is asymptotically unbiased and consistent for the true variance under either the NM approach or the IM approach, provided the overall sampling fraction is negligible.

References

- Chauvet, G., Deville, J-C. and Haziza, D. (2010). On random balanced imputation in surveys. *Submitted for publication*.
- Fay, R.E. (1991). A Design-Based Perspective on Missing Data Variance. *Proceedings of the 1991 Annual Research Conference, US Bureau of the Census*, pp. 429-440.
- 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.
- Kim, J.K and Fuller, W.A. (2004). Fractional hot-deck imputation. *Biometrika*, 91, pp. 559-578.
- Kim, J.K and Haziza, D. (2010). Doubly robust inference with missing data in survey sampling. *Submitted for publication*.
- Rao, J.N.K. and Shao, J. (1992). On variance estimation under imputation for missing data. *Biometrika*, 79, pp. 811-822.
- 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.
- Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

Appendix

Bias of DR- ϕ imputation under the UNM approach

Under the UNM approach, the conditional nonresponse bias of \hat{Y}_I is given by

$$B_q(\hat{Y}_I) = E_q \left(\sum_{i \in s} w_i r_i y_i + \sum_{i \in s} w_i (1 - r_i) \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} - \sum_{i \in s} w_i y_i | \mathbf{I} \right).$$

Using a first-order Taylor expansion, we obtain

$$B_q(\hat{Y}_I) \doteq p \sum_{i \in s} w_i y_i + (1 - p) \sum_{i \in s} w_i \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_1 - \sum_{i \in s} w_i y_i,$$

where $\hat{\mathbf{B}}_1 = (\sum_{i \in s} w_i p_i c_i^{-1} \mathbf{z}_i \mathbf{z}_i^\top)^{-1} (\sum_{i \in s} w_i p_i c_i^{-1} \mathbf{z}_i y_i)$. Using the fact that

$$\boldsymbol{\lambda}^\top \left(\sum_{i \in s} w_i \delta_i \mathbf{z}_i \mathbf{z}_i^\top / (\boldsymbol{\lambda}^\top \mathbf{z}_i) \right) \hat{\mathbf{B}}_1 = \sum_{i \in s} w_i \delta_i \mathbf{z}_i^\top \hat{\mathbf{B}}_1 = \sum_{i \in s} w_i y_i,$$

we obtain

$$\begin{aligned} B_q(\hat{Y}_I) &\doteq p \sum_{i \in s} w_i \delta_i \mathbf{z}_i^\top \hat{\mathbf{B}}_1 + (1 - p) \sum_{i \in s} w_i \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_1 - \sum_{i \in s} w_i \delta_i \mathbf{z}_i^\top \hat{\mathbf{B}}_1 \\ &= \left((1 - p) \sum_{i \in s} w_i \phi_i \mathbf{z}_i^\top - (1 - p) \sum_{i \in s} w_i \delta_i \mathbf{z}_i^\top \right) \hat{\mathbf{B}}_1 \\ &= (1 - p) \sum_{i \in s} w_i (\phi_i - \delta_i) \mathbf{z}_i^\top \hat{\mathbf{B}}_1. \end{aligned}$$

Finally, under mild regularity conditions, we have $\sum_{i \in s} w_i \delta_i \mathbf{z}_i^\top - \sum_{i \in s} w_i \phi_i \mathbf{z}_i^\top \xrightarrow{p} 0$.

Bias of DR- ϕ imputation under the IM approach

Under the IM approach, the conditional nonresponse bias of \hat{Y}_I is given by

$$\begin{aligned}
 B_{qm}(\hat{Y}_I) &= E_q E_m \left(\sum_{i \in s} w_i r_i y_i + \sum_{i \in s} w_i (1 - r_i) y_i^* - \sum_{i \in s} w_i y_i | \mathbf{I}, \mathbf{r} \right) \\
 &= E_q E_m \left(\sum_{i \in s} w_i r_i y_i + \sum_{i \in s} w_i (1 - r_i) \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} - \sum_{i \in s} w_i y_i | \mathbf{I}, \mathbf{r} \right) \\
 &= E_q \left(\sum_{i \in s} w_i r_i \phi_i \mathbf{z}_i^\top \boldsymbol{\beta} | \mathbf{I} \right) + E_q E_m \left(\sum_{i \in s} w_i (1 - r_i) \delta_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} | \mathbf{I}, \mathbf{r} \right) \\
 &\quad - E_q \left(\sum_{i \in s} w_i \phi_i \mathbf{z}_i^\top \boldsymbol{\beta} | \mathbf{I} \right).
 \end{aligned}$$

Now noting that $E_m(\hat{\mathbf{B}}_{r_1} | \mathbf{I}, \mathbf{r}) = \boldsymbol{\beta}$, we obtain

$$E_m \left(\sum_{i \in s} w_i (1 - r_i) \phi_i \mathbf{z}_i^\top \hat{\mathbf{B}}_{r_1} | \mathbf{I}, \mathbf{r} \right) = \sum_{i \in s} w_i (1 - r_i) \phi_i \mathbf{z}_i^\top \boldsymbol{\beta}.$$

It follows that

$$\begin{aligned}
 B_{qm}(\hat{Y}_I) &= E_q \left(\sum_{i \in s} w_i r_i \phi_i \mathbf{z}_i^\top \boldsymbol{\beta} + \sum_{i \in s} w_i (1 - r_i) \phi_i \mathbf{z}_i^\top \boldsymbol{\beta} - \sum_{i \in s} w_i \phi_i \mathbf{z}_i^\top \boldsymbol{\beta} | s \right) \\
 &= 0.
 \end{aligned}$$

Table 1: Monte carlo percent RB and RE for $R^2 = 0.36$

| Imputation method | | | DPR | | DR | | RR- ϕ | | |
|-----------------------|--------------------|--------------------|-----------|-------|-----------|-------|------------|-------|------|
| ϕ - mechanism | proportion of 0 | p - mechanism | RB (in %) | RE | RB (in %) | RE | RB (in %) | RE | |
| 1 | 25% | 1 | 9.99 | 2.56 | -0.05 | 1 | -0.02 | 1.11 | |
| | | 2 | 6.20 | 1.62 | -0.93 | 1 | -1.19 | 1.04 | |
| | | 3 | 24.02 | 2.23 | 14.96 | 1 | 14.88 | 1.00 | |
| | 50% | 1 | 27.23 | 7.18 | 0.03 | 1 | 0.12 | 1.14 | |
| | | 2 | 19.15 | 4.32 | -0.06 | 1 | -0.74 | 1.03 | |
| | | 3 | 39.86 | 5.04 | 14.79 | 1 | 14.12 | 0.97 | |
| | 2 | 25% | 1 | 5.58 | 1.55 | 0.10 | 1 | 0.11 | 1.04 |
| | | | 2 | 5.73 | 1.48 | 2.45 | 1 | 0.30 | 0.91 |
| | | | 3 | 20.37 | 1.51 | 16.01 | 1 | 13.69 | 0.78 |
| 50% | | 1 | 18.99 | 4.19 | -0.08 | 1 | 0.00 | 1.01 | |
| | | 2 | 19.86 | 2.95 | 7.87 | 1 | 0.85 | 0.58 | |
| | | 3 | 33.14 | 2.88 | 17.40 | 1 | 10.20 | 0.52 | |
| 3 | 25% | 1 | 7.30 | 2.18 | -0.11 | 1 | -0.09 | 1.09 | |
| | | 2 | 3.89 | 1.33 | -0.22 | 1 | -2.89 | 1.29 | |
| | | 3 | 21.33 | 1.22 | 19.11 | 1 | 18.71 | 0.97 | |
| | 50% | 1 | 20.34 | 6.41 | 0.04 | 1 | 0.11 | 1.13 | |
| | | 2 | 15.39 | 3.50 | 3.93 | 1 | -1.97 | 0.87 | |
| | | 3 | 32.85 | 2.01 | 22.14 | 1 | 18.12 | 0.73 | |

Table 2: Monte carlo percent RB and RE for $R^2 = 0.5$

| Imputation method | | | DPR | | DR | | RR- ϕ | | |
|-----------------------|--------------------|--------------------|-----------|-------|-----------|-------|------------|-------|------|
| ϕ - mechanism | proportion of 0 | p - mechanism | RB (in %) | RE | RB (in %) | RE | RB (in %) | RE | |
| 1 | 25% | 1 | 9.90 | 2.69 | -0.01 | 1 | 0.00 | 1.11 | |
| | | 2 | 6.58 | 1.79 | -0.38 | 1 | -0.65 | 1.02 | |
| | | 3 | 19.08 | 2.37 | 11.01 | 1 | 10.88 | 0.99 | |
| | 50% | 1 | 31.34 | 8.83 | 0.04 | 1 | 0.05 | 1.16 | |
| | | 2 | 21.54 | 5.20 | -0.47 | 1 | 0.35 | 1.03 | |
| | | 3 | 37.52 | 6.54 | 10.50 | 1 | 11.15 | 1.09 | |
| | 2 | 25% | 1 | 6.13 | 1.73 | -0.02 | 1 | 0.01 | 1.03 |
| | | | 2 | 7.23 | 1.61 | 3.76 | 1 | 0.79 | 0.79 |
| | | | 3 | 17.36 | 1.58 | 13.07 | 1 | 10.04 | 0.69 |
| 50% | | 1 | 16.57 | 3.98 | -0.23 | 1 | -0.17 | 1.02 | |
| | | 2 | 17.80 | 2.80 | 7.53 | 1 | 0.89 | 0.58 | |
| | | 3 | 26.86 | 2.62 | 14.70 | 1 | 7.91 | 0.49 | |
| 3 | 25% | 1 | 6.87 | 2.06 | -0.06 | 1 | -0.05 | 1.08 | |
| | | 2 | 4.89 | 1.54 | 1.19 | 1 | -1.72 | 1.10 | |
| | | 3 | 18.12 | 1.28 | 15.74 | 1 | 14.80 | 0.91 | |
| | 50% | 1 | 19.99 | 6.12 | -0.09 | 1 | -0.01 | 1.12 | |
| | | 2 | 16.50 | 3.34 | 5.37 | 1 | -0.63 | 0.72 | |
| | | 3 | 30.11 | 2.13 | 19.50 | 1 | 14.81 | 0.66 | |

Table 3: Monte carlo percent RB and RE for $R^2 = 0.7$

| Imputation method | | | DPR | | RD | | RR- ϕ | | |
|-----------------------|--------------------|--------------------|-----------|-------|-----------|-------|------------|-------|------|
| ϕ - mechanism | proportion of 0 | p - mechanism | RB (in %) | RE | RB (in %) | RE | RB (in %) | RE | |
| 1 | 25% | 1 | 9.76 | 3.02 | -0.03 | 1 | -0.09 | 1.13 | |
| | | 2 | 7.34 | 2.22 | 0.43 | 1 | 0.06 | 1.03 | |
| | | 3 | 12.35 | 2.81 | 5.01 | 1 | 4.60 | 0.96 | |
| | 50% | 1 | 27.05 | 8.69 | -0.11 | 1 | -0.09 | 1.18 | |
| | | 2 | 19.82 | 5.30 | 0.56 | 1 | -0.23 | 1.01 | |
| | | 3 | 25.67 | 6.41 | 4.98 | 1 | 4.16 | 0.97 | |
| | 2 | 25% | 1 | 4.80 | 1.63 | -0.03 | 1 | -0.01 | 1.05 |
| | | | 2 | 6.44 | 1.57 | 3.76 | 1 | 1.16 | 0.75 |
| | | | 3 | 10.26 | 1.59 | 7.28 | 1 | 4.32 | 0.61 |
| 50% | | 1 | 18.15 | 4.98 | -0.11 | 1 | -0.06 | 1.04 | |
| | | 2 | 18.98 | 3.17 | 7.70 | 1 | 0.80 | 0.54 | |
| | | 3 | 22.36 | 3.11 | 10.28 | 1 | 2.93 | 0.46 | |
| 3 | 25% | 1 | 5.81 | 1.92 | 0.07 | 1 | 0.08 | 1.06 | |
| | | 2 | 6.14 | 1.61 | 3.35 | 1 | -0.40 | 0.79 | |
| | | 3 | 12.30 | 1.37 | 10.06 | 1 | 7.84 | 0.71 | |
| | 50% | 1 | 16.94 | 5.18 | -0.05 | 1 | -0.03 | 1.06 | |
| | | 2 | 16.99 | 2.77 | 7.93 | 1 | -0.14 | 0.48 | |
| | | 3 | 23.14 | 2.17 | 14.41 | 1 | 7.15 | 0.43 | |

Table 4: Monte carlo percent RB and RE for the proposed imputation procedures

| Sample description | | | RR- ϕ | | BRR- ϕ | | DR- ϕ | |
|--------------------|-----------------|----------------|------------|----|-------------|------|------------|------|
| ϕ -mechanism | proportion of 0 | p -mechanism | RB (in %) | RE | RB (in%) | RE | RB (in %) | RE |
| 1 | 50% | 1 | 0.05 | 1 | 0.08 | 0.87 | 0.08 | 0.87 |
| | | 2 | 0.35 | 1 | 0.28 | 0.91 | 0.29 | 0.91 |
| | | 3 | 11.15 | 1 | 11.16 | 0.95 | 11.18 | 0.95 |
| 2 | 50% | 1 | -0.17 | 1 | -0.15 | 0.95 | -0.15 | 0.94 |
| | | 2 | 0.89 | 1 | 0.89 | 0.96 | 0.89 | 0.95 |
| | | 3 | 7.91 | 1 | 7.89 | 0.97 | 7.88 | 0.96 |

Table 5: Monte carlo percent RB (in %) and MSE (in parentheses) of the jackknife variance estimator

| Sample description | | RB (MSE) | |
|--------------------|-----------------|-----------------|----------------|
| ϕ -mechanism | proportion of 0 | RR- ϕ | BRR- ϕ |
| 1 | 10% | 1.53 (871.87) | 0.85 (779.71) |
| | 25% | -2.02 (1042.74) | -3.16 (877.85) |
| 2 | 10% | 1.69 (846.26) | 1.67 (804.02) |
| | 25% | 5.23 (1004.93) | 5.27 (943.52) |