

Fully efficient estimation of coefficients of correlation in the presence of imputed data

Guillaume Chauvet* David Haziza †

January 3, 2011

Abstract

Marginal imputation that consists of imputing items separately, generally leads to biased estimators of population coefficients of correlation. To overcome this problem, two main approaches have been considered in the literature: the first consists of using marginal imputation and use a bias-adjusted estimator; Skinner and Rao (2002). The second consists of using an imputation method, which preserves the relationship between variables; Shao and Wang (2002), who proposed a joint random regression imputation method that succeeds in preserving the relationships between two variables. One drawback of the Shao-Wang method is that it introduces an additional amount of variability (called the imputation variance) due to the random selection of residuals. As a result, it could lead to inefficient estimators. Following Chauvet, Deville and Haziza (2010), we propose a balanced joint random regression imputation that preserves the coefficient of correlation between two variables, while virtually eliminating the imputation variance. Results of a simulation study support our findings.

Key words: Balanced imputation; coefficient of correlation; imputation; bootstrap variance estimation.

*G. Chauvet, Laboratoire de Statistique d'Enquête, CREST/ENSAI, Campus de Ker Lann, 35170 Bruz, France

†Département de mathématiques et de statistique, Université de Montréal, Montréal, Canada

1 Introduction

Single imputation, which consists of replacing a missing value by an artificial value, is often used in statistical agencies for treating item nonresponse. The main objective of imputation is to reduce the nonresponse bias, which requires powerful auxiliary information available for all the sample units (respondents and nonrespondents). Most often, some form of marginal imputation is used. That is, items requiring imputation are treated separately. For univariate parameters such as population totals (or means), marginal imputation leads to asymptotically unbiased estimators provided the assumed model is correctly specified; e.g., Haziza (2009). In practice, many surveys use deterministic or random regression imputation, which includes mean imputation, ratio imputation and random hot-deck imputation as special cases.

Sometimes, the interest lies in estimating bivariate parameters such as regression and correlation coefficients. While marginal imputation is appropriate for univariate parameters, it may lead to considerably biased estimators of bivariate parameters. Despite the practical importance of regression and correlation coefficients, the literature on this topic is quite limited. Santos (1981) studied the bias of several marginal imputation methods on regression coefficients in the case of simple random sampling without replacement. Skinner and Rao (2002) proposed to adjust for the bias at the estimation stage. They considered the case of common donor imputation (see Section 2) and simple random sampling without replacement and studied the properties of the bias-adjusted estimator under the so-called nonresponse approach. Shao and Wang (2002) proposed a joint random regression imputation, which succeeds in preserving the coefficient of correlation between two items.

In this paper, we focus on the finite population coefficient of correlation between two study variables x and y :

$$\rho_{xy} = \frac{t_{11} - t_{10}t_{01}/N}{(t_{20} - (t_{10})^2/N)^{1/2} (t_{02} - (t_{01})^2/N)^{1/2}},$$

where $t_{kl} = \sum_{i \in U} x_i^k y_i^l$, $(k, l) \in \{(1, 0), (2, 0), (1, 1), (0, 1), (0, 2)\}$ and U denotes the finite population of size N . For example, $t_{10} = \sum_{i \in U} x_i$ and $t_{11} = \sum_{i \in U} x_i y_i$.

A sample s is selected from U according to a sampling design $p(s)$. Let $w_i = 1/\pi_i$ be the sampling weight attached to unit i , where $\pi_i = P(i \in s)$ denotes its first-order inclusion probability in the sample. A complete data estimator of ρ_{xy} is the plug-in estimator given by

$$\hat{\rho}_{xy\pi} = \frac{\hat{t}_{11,\pi} - \hat{t}_{10,\pi}\hat{t}_{01,\pi}/\hat{N}_\pi}{\left(\hat{t}_{20,\pi} - (\hat{t}_{10,\pi})^2/\hat{N}_\pi\right)^{1/2} \left(\hat{t}_{02,\pi} - (\hat{t}_{01,\pi})^2/\hat{N}_\pi\right)^{1/2}}, \quad (1)$$

where $\hat{t}_{kl,\pi} = \sum_{i \in s} w_i x_i^k y_i^l$ and $\hat{N}_\pi = \sum_{i \in s} w_i$ denote the design-unbiased (or p -unbiased) expansion estimators of t_{kl} and N , respectively. Under mild regularity conditions (e.g., Deville, 1999), the estimator $\hat{\rho}_{xy\pi}$ is asymptotically design-unbiased for ρ_{xy} .

In practice, both x and y may be potentially missing and require some form of imputation. We adopt the following notation: let r_{xi} be a response indicator attached to unit i such that $r_{xi} = 1$ if i responds to item x and $r_{xi} = 0$, otherwise. Similarly, let $r_{yi} = 1$ if i responds to item y and $r_{yi} = 0$, otherwise. Let x_i^* be the imputed value used to replace the missing x_i and y_i^* be the imputed value corresponding to the missing y_i . Finally, let $\tilde{x}_i = x_i$ if $r_{xi} = 1$ and $\tilde{x}_i = x_i^*$ if $r_{xi} = 0$. Similarly, let $\tilde{y}_i = y_i$ if $r_{yi} = 1$ and $\tilde{y}_i = y_i^*$ if $r_{yi} = 0$. An imputed estimator of ρ_{xy} , based on observed and imputed

values, is defined as

$$\hat{\rho}_{xyI} = \frac{\hat{t}_{11,I} - \hat{t}_{10,I}\hat{t}_{01,I}/\hat{N}_\pi}{\left(\hat{t}_{20,I} - (\hat{t}_{10,I})^2/\hat{N}_\pi\right)^{1/2} \left(\hat{t}_{02,I} - (\hat{t}_{01,I})^2/\hat{N}_\pi\right)^{1/2}}, \quad (2)$$

where $\hat{t}_{kl,I} = \sum_{i \in s} w_i \tilde{x}_i^k \tilde{y}_i^l$. Thus, obtaining an asymptotically unbiased estimator of ρ_{xy} requires determining an asymptotically unbiased estimator of each term, $t_{kl} = \sum_{i \in U} x_i^k y_i^l$, $(k, l) \in \{(1, 0), (2, 0), (1, 1), (0, 1), (0, 2)\}$. For the terms t_{10} and t_{01} (i.e., the marginal first moments), this can be achieved by using an appropriate deterministic or random imputation method, whereas the terms t_{20} and t_{02} (i.e., the marginal second moments) require a random imputation method because deterministic methods tend to distort the second moment of the variables being imputed. The main difficulty lies in obtaining an asymptotically unbiased estimator of the cross-product term, t_{11} , which can be viewed as a measure of the relationship between the study variables x and y . Marginal imputation, which consists of imputing x and y separately, generally distorts the relationship between variables so the resulting estimator of t_{11} is generally biased. To overcome this difficulty, two main approaches may be used: (i) Use marginal imputation (or a variant, see Section 2) and use a bias-adjusted estimator; see Skinner and Rao (2002). (ii) Use a tailor-made imputation method and use the imputed estimator (1); see Shao and Wang (2002).

In this paper, the properties of estimators are studied under two distinct approaches: (i) the Nonresponse Model (NM) approach and (ii) the Imputation Model (IM) approach. Before describing both approaches, we introduce further notation. Let $\mathbf{x} = (x_1, \dots, x_N)'$ and $\mathbf{y} = (y_1, \dots, y_N)'$, where

x_i and y_i denote the i -th value corresponding to items x and y , respectively. Let $\boldsymbol{\delta} = (\delta_1, \dots, \delta_N)'$ be the vector of sample selection indicators, where $\delta_i = 1$ if unit i is selected in the sample and $\delta_i = 0$, otherwise. Finally, let $\mathbf{r} = (r_{x1}, \dots, r_{xN}, r_{y1}, \dots, r_{yN})'$ be the vector of response indicators. The NM and IM approaches are described below:

(i) The NM approach: we make explicit assumptions (called the nonresponse model) about the unknown nonresponse mechanism. Inferences are made with respect to the joint distribution induced by the sampling design and the assumed nonresponse model. Let $p_{rr} = P(r_{xi} = 1, r_{yi} = 1)$, $p_{rm} = P(r_{xi} = 1, r_{yi} = 0)$, $p_{mr} = P(r_{xi} = 0, r_{yi} = 1)$ and $p_{mm} = P(r_{xi} = 0, r_{yi} = 0)$. If $p_x = p_{rr} + p_{rm}$ denotes the probability of response to item x and $p_y = p_{rr} + p_{mr}$ denotes the probability of response to item y , note that $p_{rr} \neq p_x p_y$, in general. Also, we assume that the sample units respond independently of one another.

Under the NM approach and random imputation, we define the conditional nonresponse bias of $\hat{\rho}_{xyI}$ as $B_{qI}(\hat{\rho}_{xyI}) = E_q E_I ((\hat{\rho}_{xyI} - \hat{\rho}_{xy\pi} | \mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r}) | \mathbf{x}, \mathbf{y}, \boldsymbol{\delta})$, where the subscripts q and I denote respectively the unknown nonresponse mechanism and the imputation mechanism. Conditionally given $\boldsymbol{\delta}$, note that the complete data estimator $\hat{\rho}_{xy\pi}$ is a fixed quantity. To simplify the notation, we write $E_I(\hat{\rho}_{xyI} | \mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r}) \equiv \tilde{\rho}_{xyI}$ in the remainder of the paper.

(ii) The IM approach: explicit assumptions about the distributions of the items of interest, called the imputation model, are made. Unlike the NM approach, the underlying nonresponse mechanism is not explicitly specified,

except that it is assumed to be unconfounded; e.g., Rubin (1976). In this paper, we consider (deterministic and random) regression imputation, which is motivated by the following bivariate imputation model:

$$m : \begin{aligned} y_i &= \mathbf{z}'_i \boldsymbol{\beta} + \sqrt{v_i} \epsilon_i \\ x_i &= \mathbf{z}'_i \boldsymbol{\gamma} + \sqrt{u_i} \eta_i, \end{aligned} \quad (3)$$

where \mathbf{z}_i is a q -vector of auxiliary variables available for all $i \in s$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are two q -vectors of parameters, $v_i = v(\mathbf{z}_i)$ and $u_i = u(\mathbf{z}_i)$ are two known functions and ϵ_i (respectively η_i) is a random term independent of \mathbf{z}_i with mean 0 and variance σ_ϵ^2 (respectively σ_η^2). Note that ϵ_i and η_i are not independent in general, and their covariance is denoted as $\sigma_{\epsilon\eta}$.

Under the IM approach and random imputation, we define the conditional nonresponse bias of $\hat{\rho}_{xyI}$ as $B_{mI}(\hat{\rho}_{xyI}) = E_m(\tilde{\rho}_{xyI} - \hat{\rho}_{xy\pi} | \boldsymbol{\delta}, \mathbf{r})$, where the subscript m denotes the imputation model (3).

The paper is organized as follows: in Section 2, three versions of the customary random-hot deck imputation procedure are first presented. Then, the resulting imputed estimator of the coefficient of correlation between x and y is shown to be asymptotically biased for all three versions under either the NM approach or the IM approach. Finally, following Skinner and Rao (2002), we consider a doubly robust bias-adjusted estimator. However, obtaining a bias-adjusted estimator for more complex imputation procedures may prove to be difficult. To overcome this difficulty, Shao and Wang (2002) proposed a joint random regression imputation, which preserves the coefficient of correlation between x and y . It is described in Section 3.1. In Section 3.2, following Chauvet, Deville and Haziza (2010), we propose a fully effi-

cient version of the Shao-Wang procedure, which we call balanced random regression imputation. Under this procedure, the imputation variance arising from the random selection of residuals, is eliminated. The algorithm for performing the proposed balanced imputation procedure is presented in Section 3.3. In Section 4, motivated by the reverse framework of Fay (1991) and Shao and Steel (1999), we propose a bootstrap variance estimator which does not require re-imputation within each bootstrap sample, unlike the method advocated by Shao and Sitter (1996). In Section 5, we first compare the performance of several imputation procedures in terms of relative bias and relative efficiency. Furthermore, the performance of the proposed bootstrap variance estimator is studied in terms of relative bias and coverage rate. Finally, we conclude in Section 6.

2 Bias under random hot-deck imputation

In this section, we study the bias of the imputed estimator $\hat{\rho}_{xyI}$ under random hot-deck imputation. We consider three versions of random hot-deck imputation, which are described in Section 2.1. In Section 2.2, we show that all these methods lead to biased estimators of a coefficient of correlation, in general.

2.1 Random hot-deck imputation

In this section, we describe three versions of random hot-deck imputation: common donor random hot-deck imputation (CDI), which was studied by Skinner and Rao (2002), marginal random hot-deck imputation (MI) and hybrid random hot-deck imputation (HI). We introduce further notation: let

$s_{rr} = \{i \in s : r_{xi} = 1 \text{ and } r_{yi} = 1\}$, $s_{rm} = \{i \in s : r_{xi} = 1 \text{ and } r_{yi} = 0\}$,
 $s_{mr} = \{i \in s : r_{xi} = 0 \text{ and } r_{yi} = 1\}$ and $s_{mm} = \{i \in s : r_{xi} = 0 \text{ and } r_{yi} = 0\}$
 with respective sizes n_{rr} , n_{rm} , n_{mr} and n_{mm} such that $n_{rr} + n_{rm} + n_{mr} + n_{mm} \equiv n$, the overall sample size.

Common Donor Random Hot-Deck Imputation:

for $i \in s_{mr}$, missing x_i is imputed by $x_i^* = x_j, j \in s_{rr}$, such that

$$P(x_i^* = x_j) = \frac{w_j}{\sum_{k \in s} w_k r_{xk} r_{yk}};$$

for $i \in s_{rm}$, missing y_i is imputed by $y_i^* = y_j, j \in s_{rr}$ such that

$$P(y_i^* = y_j) = \frac{w_j}{\sum_{k \in s} w_k r_{xk} r_{yk}};$$

for $i \in s_{mm}$, missing (x_i, y_i) is imputed by $(x_i^*, y_i^*) = (x_j, y_j), j \in s_{rr}$ such that

$$P[(x_i^*, y_i^*) = (x_j, y_j)] = \frac{w_j}{\sum_{k \in s} w_k r_{xk} r_{yk}}.$$

Marginal Random Hot-Deck Imputation:

for $i \in s_{mr}$, missing x_i is imputed by $x_i^* = x_j, j \in s_{rr} \cup s_{rm}$ such that

$$P(x_i^* = x_j) = \frac{w_j}{\sum_{k \in s} w_k r_{xk}};$$

for $i \in s_{rm}$, missing y_i is imputed by $y_i^* = y_j, j \in s_{rr} \cup s_{mr}$ such that

$$P(y_i^* = y_j) = \frac{w_j}{\sum_{k \in s} w_k r_{yk}};$$

for $i \in s_{mm}$, missing (x_i, y_i) is imputed by $(x_i^*, y_i^*) = (x_j, y_l)$, $j \in s_{rr} \cup s_{rm}$ and $l \in s_{rr} \cup s_{mr}$ such that

$$P[(x_i^*, y_i^*) = (x_j, y_l)] = \frac{w_j}{\sum_{k \in s} w_k r_{xk}} \frac{w_l}{\sum_{k \in s} w_k r_{yk}}.$$

Hybrid Random Hot-Deck Imputation:

for $i \in s_{mr}$, missing x_i is imputed by $x_i^* = x_j$, $j \in s_{rr} \cup s_{rm}$ such that

$$P(x_i^* = x_j) = \frac{w_j}{\sum_{k \in s} w_k r_{xk}};$$

for $i \in s_{rm}$, missing y_i is imputed by $y_i^* = y_j$, $j \in s_{rr} \cup s_{mr}$ such that

$$P(y_i^* = y_j) = \frac{w_j}{\sum_{k \in s} w_k r_{yk}};$$

for $i \in s_{mm}$, missing (x_i, y_i) is imputed by $(x_i^*, y_i^*) = (x_j, y_j)$, $j \in s_{rr}$ such that

$$P[(x_i^*, y_i^*) = (x_j, y_j)] = \frac{w_j}{\sum_{k \in s} w_k r_{xk} r_{yk}}.$$

Note that HI may be seen as compromise between CDI and MI.

2.2 Bias under the NM approach

In this section, we study the conditional nonresponse bias of $\hat{\rho}_{xyI}$ under the NM approach. The relative conditional nonresponse bias of $\hat{\rho}_{xyI}$, $RB_{qI}(\hat{\rho}_{xyI}) = B_{qI}(\hat{\rho}_{xyI}) / \hat{\rho}_{xy\pi}$, can be approximated by

$$RB_{qI}(\hat{\rho}_{xyI}) \doteq -(1 - h), \quad (4)$$

provided $\hat{\rho}_{xy\pi} \neq 0$, where

$$h = \begin{cases} p_{rr} + p_{mm} & \text{for CDI and HI} \\ p_{rr} & \text{for MI} \end{cases}$$

If $\hat{\rho}_{xy\pi} = 0$ (i.e, the variables x and y are unrelated), the imputed estimator $\hat{\rho}_{xyI}$ is asymptotically qI -unbiased for $\hat{\rho}_{xy\pi}$, as expected. Expression (4) shows that the asymptotic bias is always negative. That is, the three versions of random hot-deck imputation attenuate the relationship between both items. Also, the bias increases as the response probability to both items, p_{rr} , decreases. For CDI and HI, the asymptotic bias vanishes if $p_{rm} = p_{mr} = 0$, or equivalently, if $s_{rm} = s_{mr} = \emptyset$. In this case, CDI and HI preserve the relationships between items x and y since both methods use a single donor when both items are missing, unlike MI. For MI, the asymptotic bias does not vanish, even if $s_{rm} = s_{mr} = \emptyset$.

Denote the imputed estimator $\hat{\rho}_{xyI}$ under CDI or HI by $\hat{\rho}_{(CDI-HI)}$ and by $\hat{\rho}_{(MI)}$ under MI. From (4), we obtain

$$\left| \frac{RB_{qI}(\hat{\rho}_{(CDI-HI)})}{RB_{qI}(\hat{\rho}_{(MI)})} \right| = 1 - \frac{p_{mm}}{p_{mm} + p_{rm} + p_{mr}} \leq 1. \quad (5)$$

Expression (5) shows that the bias occurring under MI is always greater or equal to the bias occurring under CDI or HI. Hence, MI distorts the relationship between x and y to a greater extent than CDI or HI.

2.3 Bias under the IM approach

In this section, we study the conditional nonresponse bias of $\hat{\rho}_{xyI}$ under the IM approach. All three versions of random hot-deck imputation are motivated by (3) with $\mathbf{z}_i = 1$ and $v_i = u_i = 1$ for all i . The relative conditional

nonresponse bias of $\hat{\rho}_{xyI}$, $RB_{mI}(\hat{\rho}_{xyI}) = B_{mI}(\hat{\rho}_{xyI})/E_m(\hat{\rho}_{xyI})$, can be approximated by

$$RB_{mI}(\hat{\rho}_{xyI}) \doteq -\left(1 - \hat{h}\right), \quad (6)$$

where

$$\hat{h} = \begin{cases} \hat{p}_{rr} + \hat{p}_{mm} & \text{for CDI and HI} \\ \hat{p}_{rr} & \text{for MI} \end{cases}$$

provided $\sigma_{\epsilon\eta} \neq 0$, where $\hat{p}_{rr} = \sum_{i \in s_{rr}} w_i / \sum_{i \in s} w_i$ and $\hat{p}_{mm} = \sum_{i \in s_{mm}} w_i / \sum_{i \in s} w_i$. If $\sigma_{\epsilon\eta} = 0$ (i.e, the variables x and y are unrelated), the imputed estimator $\hat{\rho}_{xyI}$ is mI -unbiased for $\hat{\rho}_{xy\pi}$, as expected. Expression (6) follows from noting that $E_m(\tilde{\rho}_{xyI} | \boldsymbol{\delta}, \mathbf{r}) \doteq \hat{h} \frac{\sigma_{\epsilon\eta}}{\sigma_{\epsilon}\sigma_{\eta}}$ and $E_m(\hat{\rho}_{xy\pi} | \boldsymbol{\delta}) \doteq \frac{\sigma_{\epsilon\eta}}{\sigma_{\epsilon}\sigma_{\eta}}$. Note that expression (6) is a function of the observed response rates \hat{p}_{rr} and \hat{p}_{mm} unlike (4), which depends on the true response probabilities.

2.4 A bias-adjusted estimator

Expressions (4) and (6) suggest the following simple bias-adjusted estimator denoted by $\hat{\rho}_{xyI}^a$:

$$\hat{\rho}_{xyI}^a = \min\left(1, \hat{h}^{-1} \hat{\rho}_{xyI}\right). \quad (7)$$

The bias-adjusted estimator (7) is similar to the bias-adjusted estimator proposed by Skinner and Rao (2002). If $\hat{\rho}_{xyI}^a \leq 1$ for all samples and sets of respondents, then it is asymptotically unbiased for ρ_{xy} under either the NM approach or the IM approach. Therefore, it is doubly robust since it can be justified under either approach. For secondary analysts however, computing the bias-adjusted estimator may not be an easy task because of its non-standard form. Also, the bias-adjusted estimator given by (7) was obtained

in a relatively straightforward fashion for random hot-deck imputation. For more general imputation methods (e.g., random regression imputation), obtaining a bias-adjusted estimator may prove to be much more complex. To overcome these issues, Shao and Wang (2002) proposed a joint random regression procedure, which is presented next.

3 Joint random imputation procedures

We first describe the Shao-Wang procedure (Shao and Wang, 2002) in Section 3.1. In Section 3.2, we present a fully efficient version of the Shao-Wang procedure. The details of the algorithm for performing balanced imputation are provided in Section 3.3.

3.1 The Shao-Wang procedure

Shao and Wang (2002) showed that marginal random regression imputation does not preserve the coefficient of correlation between the study variables x and y . Motivated by (3), they proposed a joint random regression imputation procedure, which can be described as follows:

- (i) for $r_{xi} = 0$ and $r_{yi} = 1$ (x_i missing, y_i observed), we use the imputed values

$$x_i^* = \mathbf{z}_i' \hat{\boldsymbol{\gamma}}_r + \sqrt{u_i} \eta_i^*,$$

where

$$\hat{\boldsymbol{\gamma}}_r = \left(\sum_{i \in s} w_i r_{xi} u_i^{-1} \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \sum_{i \in s} w_i r_{xi} u_i^{-1} \mathbf{z}_i x_i \quad (8)$$

and

$$\eta_i^* = \tilde{m}_\eta + \tilde{\eta}_i^*,$$

where

$$\tilde{m}_\eta = \frac{\hat{\sigma}_{\epsilon\eta}}{\sqrt{v_i}\hat{\sigma}_\epsilon^2} \left(y_i - \mathbf{z}'_i \hat{\boldsymbol{\beta}}_r \right),$$

$$\hat{\boldsymbol{\beta}}_r = \left(\sum_{i \in s} w_i r_{yi} v_i^{-1} \mathbf{z}_i \mathbf{z}'_i \right)^{-1} \sum_{i \in s} w_i r_{yi} v_i^{-1} \mathbf{z}_i y_i \quad (9)$$

and, given the observed data, the $\tilde{\eta}_i^*$'s are independent random variables with mean 0 and variance $\tilde{\sigma}_\eta^2 = \hat{\sigma}_\eta^2 - \hat{\sigma}_{\epsilon\eta}^2 / \hat{\sigma}_\epsilon^2$ with

$$\hat{\sigma}_\epsilon^2 = \frac{1}{\sum_{j \in s} w_j r_{xj} r_{yj}} \sum_{j \in s} w_j r_{xj} r_{yj} v_j^{-1} \left(y_j - \mathbf{z}'_j \hat{\boldsymbol{\beta}}_r \right)^2, \quad (10)$$

$$\hat{\sigma}_\eta^2 = \frac{1}{\sum_{j \in s} w_j r_{xj} r_{yj}} \sum_{j \in s} w_j r_{xj} r_{yj} u_j^{-1} \left(x_j - \mathbf{z}'_j \hat{\boldsymbol{\gamma}}_r \right)^2 \quad (11)$$

and

$$\hat{\sigma}_{\epsilon\eta} = \frac{1}{\sum_{j \in s} w_j r_{xj} r_{yj}} \sum_{j \in s} w_j r_{xj} r_{yj} u_j^{-1/2} v_j^{-1/2} \left(x_j - \mathbf{z}'_j \hat{\boldsymbol{\gamma}}_r \right) \left(y_j - \mathbf{z}'_j \hat{\boldsymbol{\beta}}_r \right). \quad (12)$$

(ii) for $r_{xi} = 1$ and $r_{yi} = 0$ (x_i observed, y_i missing), we use the imputed values

$$y_i^* = \mathbf{z}'_i \hat{\boldsymbol{\beta}}_r + \sqrt{v_i} \epsilon_i^*,$$

where

$$\epsilon_i^* = \tilde{m}_\epsilon + \tilde{\epsilon}_i^*,$$

$$\tilde{m}_\epsilon = \frac{\hat{\sigma}_{\epsilon\eta}}{\sqrt{u_i \hat{\sigma}_\eta^2}} (x_i - \mathbf{z}'_i \hat{\boldsymbol{\gamma}}_r)$$

and, given the observed data, the $\tilde{\epsilon}_i^*$'s are independent random variables with mean 0 and variance $\tilde{\sigma}_\epsilon^2 = \hat{\sigma}_\epsilon^2 - \hat{\sigma}_{\epsilon\eta}^2 / \hat{\sigma}_\eta^2$.

(iii) For $r_{xi} = 0$ and $r_{yi} = 0$ (both x_i and y_i missing), we use the imputed values

$$\begin{aligned} x_i^* &= \mathbf{z}'_i \hat{\boldsymbol{\gamma}}_r + \sqrt{u_i} \eta_i^* \\ y_i^* &= \mathbf{z}'_i \hat{\boldsymbol{\beta}}_r + \sqrt{v_i} \epsilon_i^*, \end{aligned}$$

where (ϵ_i^*, η_i^*) 's are independently distributed with mean 0 and covariance matrix

$$\hat{\Sigma}_1 = \begin{pmatrix} \hat{\sigma}_\epsilon^2 & \hat{\sigma}_{\epsilon\eta} \\ \hat{\sigma}_{\epsilon\eta} & \hat{\sigma}_\eta^2 \end{pmatrix}$$

There exist numerous ways for generating the $\tilde{\epsilon}_i^*$'s and $\tilde{\eta}_i^*$'s. Under the IM approach, Shao and Wang (2002) showed that this joint regression imputation method leads to asymptotically unbiased estimators of coefficients of correlation, provided the $\tilde{\epsilon}_i^*$'s and $\tilde{\eta}_i^*$'s are independently selected from any distribution with appropriate mean and variance. They argued that the random residuals should be generated from the respondents' residuals if other non-linear parameters such as quantiles are of interest. In any case, the procedures described by Shao and Wang (2002) all suffer from an extra variability, called the imputation variance, due to the random selection of the residuals. As a result, these procedures are not fully efficient, a term coined by Kim and Fuller (2004).

3.2 Balanced imputation procedure

To develop a fully efficient procedure, we first express the total error of $\hat{\rho}_{xyI}$ as

$$\hat{\rho}_{xyI} - \rho_{xy} = (\hat{\rho}_{xy\pi} - \rho_{xy}) + (\tilde{\rho}_{xyI} - \hat{\rho}_{xy\pi}) + (\hat{\rho}_{xyI} - \tilde{\rho}_{xyI}). \quad (13)$$

The first term on the right hand side of (13) represents the sampling error, whereas the second and the third terms represent the nonresponse error and the imputation error, respectively. In order to eliminate the imputation variance, we suggest selecting the residuals ϵ_i^* and η_i^* at random so that the imputation error, $\hat{\rho}_{xyI} - \tilde{\rho}_{xyI}$, is equal to zero. In other words, we select the residuals so that the following constraints are satisfied:

$$\hat{t}_{kl,I} - \tilde{t}_{kl,I} = 0 \quad (14)$$

for $(k, l) \in \{(1, 0), (2, 0), (1, 1), (0, 1), (0, 2)\}$, where $\tilde{t}_{kl,I} = E_I(\hat{t}_{kl,I} | \mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r})$. An imputation procedure satisfying (14) has been called a balanced imputation procedure by Chauvet, Deville and Haziza (2010).

To fix ideas, we first consider the case $(k, l) = (1, 0)$. We have

$$\begin{aligned} \hat{t}_{10,I} &= \sum_{i \in s} w_i \tilde{x}_i \\ &= \sum_{i \in s} w_i r_{xi} x_i + \sum_{i \in s} w_i (1 - r_{xi}) r_{yi} x_i^* + \sum_{i \in s} w_i (1 - r_{xi}) (1 - r_{yi}) x_i^* \\ &= \sum_{i \in s} w_i r_{xi} x_i + \sum_{i \in s} w_i (1 - r_{xi}) \mathbf{z}'_i \hat{\gamma}_r \\ &+ \sum_{i \in s} w_i (1 - r_{xi}) r_{yi} \sqrt{u_i} \eta_i^* + \sum_{i \in s} w_i (1 - r_{xi}) (1 - r_{yi}) \sqrt{u_i} \eta_i^*. \end{aligned}$$

It follows that $\hat{t}_{10,I} - \tilde{t}_{10,I} = 0$ if the balancing equations

$$\begin{aligned} \sum_{i \in s} w_i (1 - r_{xi}) r_{yi} \sqrt{u_i} (\eta_i^* - \tilde{m}_\eta) &= \sum_{i \in s_{mr}} w_i \sqrt{u_i} (\eta_i^* - \tilde{m}_\eta) \\ &= 0 \end{aligned} \quad (15)$$

and

$$\begin{aligned} \sum_{i \in s} w_i (1 - r_{xi}) (1 - r_{yi}) \sqrt{u_i} \eta_i^* &= \sum_{i \in s_{mm}} w_i \sqrt{u_i} \eta_i^* \\ &= 0 \end{aligned} \quad (16)$$

are satisfied. Similar balancing equations may be derived for any of the constraints in (14). After some algebra, we obtain that the constraints in (14) are satisfied if the imputation procedure is such that (i) the four balancing equations (corresponding to imputation on s_{mr})

$$\begin{aligned} \sum_{i \in s_{mr}} w_i \sqrt{u_i} (\eta_i^* - \tilde{m}_\eta) [1, \mathbf{z}'_i \hat{\gamma}_r, y_i] &= 0 \\ \sum_{i \in s_{mr}} w_i u_i [\eta_i^{*2} - (\tilde{m}_\eta^2 + \tilde{\sigma}_\eta^2)] &= 0 \end{aligned} \quad (17)$$

are satisfied; (ii) the four balancing equations (corresponding to imputation on s_{rm})

$$\begin{aligned} \sum_{i \in s_{rm}} w_i \sqrt{v_i} (\epsilon_i^* - \tilde{m}_\epsilon) [1, \mathbf{z}'_i \hat{\beta}_r, x_i] &= 0 \\ \sum_{i \in s_{rm}} w_i v_i [(\epsilon_i^{*2} - (\tilde{m}_\epsilon^2 + \tilde{\sigma}_\epsilon^2))] &= 0 \end{aligned} \quad (18)$$

are satisfied and (iii) the seven balancing equations (corresponding to impu-

tation on s_{mm})

$$\begin{aligned}
\sum_{i \in s_{mm}} w_i \left[\mathbf{z}'_i \hat{\boldsymbol{\gamma}}_r, \mathbf{z}'_i \hat{\boldsymbol{\beta}}_r \right] \left[\sqrt{u_i} \eta_i^*, \sqrt{v_i} \epsilon_i^* \right]' &= 0 \\
\sum_{i \in s_{mm}} w_i \left[u_i (\eta_i^{*2} - \hat{\sigma}_\eta^2), v_i (\epsilon_i^{*2} - \hat{\sigma}_\epsilon^2) \right] &= 0 \\
\sum_{i \in s_{mm}} w_i \sqrt{u_i v_i} [\eta_i^* \epsilon_i^* - \hat{\sigma}_{\epsilon\eta}] &= 0
\end{aligned} \tag{19}$$

are satisfied. In other words, we perform the imputations separately on each of the sub-samples s_{mr} , s_{rm} and s_{mm} . The algorithm for performing the random selection of residuals while satisfying (17)-(19) is described in Section 3.3.

3.3 Balanced imputation algorithm

We assume that ϵ_i^* and η_i^* are generated from the set of respondent residuals.

We introduce further notation. Let

$$\begin{aligned}
e_{xj} &= u_j^{-1/2} (x_j - \mathbf{z}'_j \hat{\boldsymbol{\gamma}}_r), \\
e_{yj} &= v_j^{-1/2} (y_j - \mathbf{z}'_j \hat{\boldsymbol{\beta}}_r)
\end{aligned}$$

and

$$(\tilde{e}_{xj}, \tilde{e}_{yj})' = \hat{\Sigma}_2^{-1/2} \left(x_j - \mathbf{z}'_j \hat{\boldsymbol{\gamma}}_{rr}, y_j - \mathbf{z}'_j \hat{\boldsymbol{\beta}}_{rr} \right)'$$

with

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{rr} &= \left(\sum_{i \in s} w_i r_{xi} r_{yi} v_i^{-1} \mathbf{z}_i \mathbf{z}'_i \right)^{-1} \sum_{i \in s} w_i r_{xi} r_{yi} v_i^{-1} \mathbf{z}_i y_i, \\
\hat{\boldsymbol{\gamma}}_{rr} &= \left(\sum_{i \in s} w_i r_{xi} r_{yi} u_i^{-1} \mathbf{z}_i \mathbf{z}'_i \right)^{-1} \sum_{i \in s} w_i r_{xi} r_{yi} u_i^{-1} \mathbf{z}_i x_i,
\end{aligned}$$

and

$$\hat{\Sigma}_2 = \begin{pmatrix} \hat{\sigma}_{\epsilon,rr}^2 & \hat{\sigma}_{\epsilon\eta} \\ \hat{\sigma}_{\epsilon\eta} & \hat{\sigma}_{\eta,rr}^2 \end{pmatrix},$$

where

$$\begin{aligned} \hat{\sigma}_{\epsilon,rr}^2 &= \frac{1}{\sum_{j \in s} w_j r_{xj} r_{yj}} \sum_{j \in s} w_j r_{xj} r_{yj} \left(y_j - \mathbf{z}'_j \hat{\boldsymbol{\beta}}_{rr} \right)^2, \\ \hat{\sigma}_{\eta,rr}^2 &= \frac{1}{\sum_{j \in s} w_j r_{xj} r_{yj}} \sum_{j \in s} w_j r_{xj} r_{yj} \left(x_j - \mathbf{z}'_j \hat{\boldsymbol{\gamma}}_{rr} \right)^2. \end{aligned}$$

In order to select the random residuals, we extend the balanced imputation procedure described in Chauvet, Deville and Haziza (2010). It is presented next:

Step 1: (x_i is missing but y_i is observed)

Build a $n_{mr} \times (n_{rm} + n_{rr})$ table, where the cell (i, j) is given the probability of selection $\phi_{ij} = w_j / \sum_{l \in s} w_l r_{xl}$, and the values

$$\mathbf{t}_{ij}^0 = w_i \phi_{ij} \sqrt{u_i} e_{xj} [1, \mathbf{z}'_i \hat{\boldsymbol{\gamma}}_r, y_i, \sqrt{u_i} e_{xj}]$$

and

$$\mathbf{t}_{ij}^1 = (t_{ij}^1, \dots, t_{ij}^{n_{mr}})',$$

where $t_{ij}^k = \phi_{ij} a_{ki}$ such that $a_{ki} = 1$ if $k = i$ and $a_{ki} = 0$, otherwise, $k = 1 \dots n_{mr}$. A random sample of cells s_{mr}^* is selected with inclusion probabilities ϕ_{ij} and balancing variables \mathbf{t}_{ij}^0 and \mathbf{t}_{ij}^1 . If the cell (i, j) is selected, then we let $\tilde{\eta}_i^* = \tilde{\sigma}_\eta e_{xj}$.

Step 2: (y_i is missing but x_i is observed)

Build a $n_{rm} \times (n_{mr} + n_{rr})$ table, where the cell (i, j) is given the probability of selection $\phi_{ij} = w_j / \sum_{l \in s} w_l r_{yl}$, and the values

$$\mathbf{t}_{ij}^0 = w_i \phi_{ij} \sqrt{v_i} e_{yj} [1, \mathbf{z}'_i \hat{\boldsymbol{\beta}}_r, x_i, \sqrt{v_i} e_{yj}]$$

and

$$\mathbf{t}_{ij}^1 = (t_{ij}^1, \dots, t_{ij}^{n_{rm}})',$$

where $t_{ij}^k = \phi_{ij} a_{ki}$, $k = 1 \dots n_{rm}$. A random sample of cells s_{rm}^* is selected with inclusion probabilities ϕ_{ij} and balancing variables \mathbf{t}_{ij}^0 and \mathbf{t}_{ij}^1 . If the cell (i, j) is selected, then we let $\tilde{\epsilon}_i^* = \tilde{\sigma}_\epsilon e_{yj}$.

Step 3: (both x_i and y_i are missing)

Build a $n_{mm} \times n_{rr}$ table, where the cell (i, j) is given the probability of selection $\phi_{ij} = w_j / \sum_{l \in s} w_j r_{xl} r_{yl}$, and the values

$$\mathbf{t}_{ij}^0 = w_i \phi_{ij} \left[\begin{array}{c} \sqrt{u_i}(\mathbf{z}'_i \hat{\gamma}_r) \tilde{e}_{xj}, \sqrt{u_i}(\mathbf{z}'_i \hat{\beta}_r) \tilde{e}_{xj}, \sqrt{v_i}(\mathbf{z}'_i \hat{\gamma}_r) \tilde{e}_{yj}, \\ \sqrt{v_i}(\mathbf{z}'_i \hat{\beta}_r) \tilde{e}_{yj}, u_i \tilde{e}_{xj}^2, v_i \tilde{e}_{yj}^2, \sqrt{u_i v_i} \tilde{e}_{xj} \tilde{e}_{yj} \end{array} \right]$$

and

$$\mathbf{t}_{ij}^1 = (t_{ij}^1, \dots, t_{ij}^{n_{mm}})',$$

where $t_{ij}^k = \phi_{ij} a_{ki}$, $k = 1 \dots n_{mm}$. A random sample of cells s_{mm}^* is selected with inclusion probabilities ϕ_{ij} and balancing variables \mathbf{t}_{ij}^0 and \mathbf{t}_{ij}^1 . If the cell (i, j) is selected, then we let $[\epsilon_i^*, \eta_i^*]' = \hat{\Sigma}_1^{1/2} [\tilde{e}_{yj}, \tilde{e}_{xj}]'$.

The balancing conditions on variables \mathbf{t}_{ij}^1 ensure that exactly one residual will be selected for each nonrespondent. In steps 1-3, the balancing conditions on variables \mathbf{t}_{ij}^0 enable us to satisfy the sets of balancing equations (17)-(19). In this case, the imputation variance is eliminated and the imputation procedure is fully efficient. Even if the balancing conditions hold only approximately, we expect the imputation variance to be significantly reduced. This is illustrated in Section 5.

4 Bootstrap variance estimation

In this section, we discuss the problem of variance estimation under the proposed balanced imputation procedure. For the joint random regression procedure described in Section 3, Shao and Wang (2002) proposed an adjusted jackknife variance estimator, which requires re-imputation within each jackknife replicate. Their variance estimator can be viewed as an extension of the adjusted jackknife variance estimator proposed by Rao and Shao (1992). The Shao and Wang jackknife variance estimator is asymptotically unbiased and consistent provided the sampling fraction n/N is negligible. Alternatively, one could extend the bootstrap procedure proposed by Shao and Sitter (1996), which also requires re-imputing within each bootstrap sample. Once again, the consistency of the resulting bootstrap variance estimator can be established, provided the sampling fraction n/N is negligible. Here, assuming that the sampling fraction n/N is negligible, we consider a bootstrap procedure that does not require re-imputation within each bootstrap sample and that can be performed using a complete data bootstrap software, which is attractive from a data user point of view. In other words, no specialized variance estimation software is needed.

To introduce our bootstrap procedure, we start by expressing the total variance of $\hat{\rho}_{xyI}$ as

$$V_T = V_1 + V_2 + V_3,$$

where

$$V_1 = E_q V_p(\tilde{\rho}_{xyI} | \mathbf{x}, \mathbf{y}, \mathbf{r}),$$

$$V_2 = E_q E_p(V_I(\hat{\rho}_{xyI} | \mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r}) | \mathbf{x}, \mathbf{y}, \mathbf{r}),$$

and

$$V_3 = V_q E_p(\tilde{\rho}_{xyI} | \mathbf{x}, \mathbf{y}, \mathbf{r});$$

see Fay (1991) and Shao and Steel (1999). Note that the term V_2 denotes the imputation variance arising solely from the random selection of residuals. In the case of the proposed balanced imputation procedure, the residuals are randomly selected so that the imputation variance is eliminated. Hence, the term V_2 is (approximately) equal to zero and can be omitted from the variance calculations. Also, under mild regularity conditions, the contribution of V_3 to the total variance, V_3/V_T , is of order $O(n/N)$. Thus, when the sampling fraction n/N is negligible, the contribution of the term V_3 to the total variance is negligible and, as a result, can also be omitted from the variance calculations. It remains to estimate the term V_1 consistently. To that end, it suffices to estimate $V_p(\tilde{\rho}_{xyI} | \mathbf{x}, \mathbf{y}, \mathbf{r})$ consistently. Noting that $E_I(\tilde{\epsilon}_i^* | \mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r}) = E_I(\tilde{\eta}_i^* | \mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r}) = 0$, it follows that $\tilde{\rho}_{xyI}$ can be written as a smooth function of estimated totals since it corresponds to the estimator of ρ_{xy} that we would have obtained under the deterministic version of the Shao-Wang procedure. The problem of estimating $V_p(\tilde{\rho}_{xyI} | \mathbf{x}, \mathbf{y}, \mathbf{r})$ reduces to the classical problem of estimating the sampling variance of a smooth function of estimated totals conditionally given \mathbf{x}, \mathbf{y} and \mathbf{r} . Any complete data variance estimation method can thus be used (e.g., Taylor linearization, jackknife or bootstrap). For the proposed balanced imputation procedure, resampling methods such as the jackknife and the bootstrap are more appealing because a Taylor linearization procedure would involve very messy calculations. In this paper, we focus on bootstrap variance estimation.

For simplicity, we consider a special case of the joint random regression

imputation procedure described in Section 3.1, which can be called joint random ratio imputation. It is obtained from (i)-(iii) in Section 3.1 by letting $\mathbf{z}_i = z_i$ (a scalar) and $u_i = v_i = z_i$. Assuming that the sample size n is large, we can approximate $\tilde{\rho}_{xyI}$ by

$$\tilde{\rho}_{xyI} \doteq \frac{\tilde{t}_{11,I} - \tilde{t}_{10,I}\tilde{t}_{01,I}/\hat{N}_\pi}{\left(\tilde{t}_{20,I} - (\tilde{t}_{10,I})^2/\hat{N}_\pi\right)^{1/2} \left(\tilde{t}_{02,I} - (\tilde{t}_{01,I})^2/\hat{N}_\pi\right)^{1/2}}.$$

After some algebra, we obtain

$$\begin{aligned} \tilde{t}_{10,I} &= \sum_{i \in s} w_i r_{xi} x_i + \sum_{i \in s} w_i (1 - r_{xi}) r_{yi} \left[z_i \hat{\gamma}_r + \frac{\hat{\sigma}_{\epsilon\eta}}{\hat{\sigma}_\epsilon^2} (y_i - z_i \hat{\beta}_r) \right] \\ &\quad + \sum_{i \in s} w_i (1 - r_{xi}) (1 - r_{yi}) [z_i \hat{\gamma}_r], \\ \tilde{t}_{20,I} &= \sum_{i \in s} w_i r_{xi} x_i^2 \\ &\quad + \sum_{i \in s} w_i (1 - r_{xi}) r_{yi} \left[u_i V_I(\tilde{\eta}_i^* | \mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r}) + \left(z_i \hat{\gamma}_r + \frac{\hat{\sigma}_{\epsilon\eta}}{\hat{\sigma}_\epsilon^2} (y_i - z_i \hat{\beta}_r) \right)^2 \right] \\ &\quad + \sum_{i \in s} w_i (1 - r_{xi}) (1 - r_{yi}) [u_i V_I(\eta_i^* | \mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r}) + (z_i \hat{\gamma}_r)^2] \end{aligned}$$

and

$$\begin{aligned} \tilde{t}_{11,I} &= \sum_{i \in s} w_i r_{xi} r_{yi} x_i y_i \\ &\quad + \sum_{i \in s} w_i (1 - r_{xi}) r_{yi} y_i \left[z_i \hat{\gamma}_r + \frac{\hat{\sigma}_{\epsilon\eta}}{\hat{\sigma}_\epsilon^2} (y_i - z_i \hat{\beta}_r) \right] \\ &\quad + \sum_{i \in s} w_i (1 - r_{yi}) r_{xi} x_i \left[z_i \hat{\beta}_r + \frac{\hat{\sigma}_{\epsilon\eta}}{\hat{\sigma}_\eta^2} (x_i - z_i \hat{\gamma}_r) \right] \\ &\quad + \sum_{i \in s} w_i (1 - r_{xi}) (1 - r_{yi}) \left[\sqrt{u_i v_i} \text{Cov}_I(\eta_i^*, \epsilon_i^* | \mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r}) + (z_i \hat{\gamma}_r) (z_i \hat{\beta}_r) \right], \end{aligned}$$

where $\hat{\gamma}_r$, $\hat{\beta}_r$, $\hat{\sigma}_\epsilon^2$, $\hat{\sigma}_\eta^2$ and $\hat{\sigma}_{\epsilon\eta}$ are given by (8)-(12), respectively. The terms $\tilde{t}_{01,I}$ and $\tilde{t}_{02,I}$ may be obtained similarly. Due to the non-independence in the

selection of the random residuals for the proposed balanced imputation procedure, the terms $V_I(\tilde{\eta}_i^*|\mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r})$, $V_I(\eta_i^*|\mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r})$ and $Cov_I(\eta_i^*, \epsilon_i^*|\mathbf{x}, \mathbf{y}, \boldsymbol{\delta}, \mathbf{r})$ are difficult to compute exactly. One option consists of replacing these terms by an approximation; see Deville and Tillé (2005). Here, we propose to use a somewhat simpler, conservative approximation: the required terms are replaced with the values obtained under the original procedure of Shao and Wang (2002), where the random residuals are selected independently in each of the subsamples s_{mr} , s_{rm} and s_{mm} . This leads to the new set of equations

$$\begin{aligned}\tilde{t}_{10,I} &= \sum_{i \in s} w_i r_{xi} x_i + \sum_{i \in s} w_i (1 - r_{xi}) r_{yi} \left[z_i \hat{\gamma}_r + \frac{\hat{\sigma}_{\epsilon\eta}}{\hat{\sigma}_\epsilon^2} (y_i - z_i \hat{\beta}_r) \right] \\ &+ \sum_{i \in s} w_i (1 - r_{xi}) (1 - r_{yi}) [z_i \hat{\gamma}_r],\end{aligned}\quad (20)$$

$$\begin{aligned}\tilde{t}_{20,I} &\doteq \sum_{i \in s} w_i r_{xi} x_i^2 \\ &+ \sum_{i \in s} w_i (1 - r_{xi}) r_{yi} \left[u_i \left(\hat{\sigma}_\eta^2 - \frac{\hat{\sigma}_{\epsilon\eta}}{\hat{\sigma}_\epsilon^2} \right) + \left(z_i \hat{\gamma}_r + \frac{\hat{\sigma}_{\epsilon\eta}}{\hat{\sigma}_\epsilon^2} (y_i - z_i \hat{\beta}_r) \right)^2 \right] \\ &+ \sum_{i \in s} w_i (1 - r_{xi}) (1 - r_{yi}) [u_i \hat{\sigma}_\eta^2 + (z_i \hat{\gamma}_r)^2]\end{aligned}\quad (21)$$

and

$$\begin{aligned}\tilde{t}_{11,I} &\doteq \sum_{i \in s} w_i r_{xi} r_{yi} x_i y_i \\ &+ \sum_{i \in s} w_i (1 - r_{xi}) r_{yi} y_i \left[z_i \hat{\gamma}_r + \frac{\hat{\sigma}_{\epsilon\eta}}{\hat{\sigma}_\epsilon^2} (y_i - z_i \hat{\beta}_r) \right] \\ &+ \sum_{i \in s} w_i (1 - r_{yi}) r_{xi} x_i \left[z_i \hat{\beta}_r + \frac{\hat{\sigma}_{\epsilon\eta}}{\hat{\sigma}_\eta^2} (x_i - z_i \hat{\gamma}_r) \right] \\ &+ \sum_{i \in s} w_i (1 - r_{xi}) (1 - r_{yi}) \left[\sqrt{u_i v_i} \hat{\sigma}_{\epsilon\eta} + (z_i \hat{\gamma}_r) (z_i \hat{\beta}_r) \right].\end{aligned}\quad (22)$$

As a result, each of the terms $\tilde{t}_{kl,I}$ may be expressed as a smooth function of estimated totals. We can thus apply any complete data bootstrap procedure to estimate the term V_1 . To illustrate the method, we consider the special case of simple random sampling without replacement. For other sampling designs, any complete data bootstrap procedure leading to consistent variance estimation can be used. We consider the so-called bootstrap weight procedure proposed by Rao, Wu and Yue (1992), which is popular in practice. It can be described as follows:

1. Let n' be the bootstrap sample size, which may be different from n .
2. Draw a simple random sample s^* of size n' *with* replacement from s . Let m_i^* be the number of times unit i is selected in s^* . We have $n' = \sum_{i \in s} m_i^*$. For unit i , define the bootstrap weight as

$$w_i^* = \left[1 + \sqrt{C} \left(\frac{nm_i^*}{n'} - 1 \right) \right] w_i, \quad \text{for } i = 1, \dots, n,$$

where

$$C = \frac{n'(1 - \frac{n}{N})}{n - 1}.$$

Calculate

$$\tilde{\rho}_{xyI}^* = \frac{\tilde{t}_{11,I}^* - \tilde{t}_{10,I}^* \tilde{t}_{01,I}^* / \hat{N}_\pi^*}{\left(\tilde{t}_{20,I}^* - (\tilde{t}_{10,I}^*)^2 / \hat{N}_\pi^* \right)^{1/2} \left(\tilde{t}_{02,I}^* - (\tilde{t}_{01,I}^*)^2 / \hat{N}_\pi^* \right)^{1/2}},$$

where $\hat{N}_\pi^* = \sum_{i \in s} w_i^*$ and $\tilde{t}_{10,I}^*$, $\tilde{t}_{20,I}^*$ and $\tilde{t}_{11,I}^*$ are respectively obtained from (8)-(12) and (20)-(22) by replacing w_i with w_i^* . The terms $\tilde{t}_{01,I}^*$ and $\tilde{t}_{02,I}^*$ are obtained similarly.

3. Repeat Step 2 a large number of times, B , to get $\tilde{\rho}_{xyI}^{*(1)}, \dots, \tilde{\rho}_{xyI}^{*(B)}$.

4. Estimate

$$V_p(\tilde{\rho}_{xyI}|\mathbf{x}, \mathbf{y}, \mathbf{r})$$

by

$$\hat{V}_B = \frac{1}{B-1} \sum_{b=1}^B \left[\tilde{\rho}_{xyI}^{*(b)} - \tilde{\rho}_{xyI}^{*(\cdot)} \right]^2, \quad (23)$$

where $\tilde{\rho}_{xyI}^{*(\cdot)} = \frac{1}{B} \sum_{b=1}^B \tilde{\rho}_{xyI}^{*(b)}$.

5 Simulation study

5.1 Performance of point estimators

We conducted a limited simulation study similar to that of Shao and Wang (2002) to investigate the performance of the proposed balanced imputation procedure in terms of bias and relative efficiency. We first generated 3 finite populations of size $N = 4\,000$, each containing two study variables, x and y , and one auxiliary variable z . The variable z was first generated independently from a Gamma distribution with mean 2 and variance 0.1. Then, given the z -values, bivariate data (x_i, y_i) 's were generated according to (3), with $\mathbf{z}_i = z_i$, $u_i = v_i = z_i$, and $\beta = \gamma = 1$. The error terms ϵ_i and η_i were independently generated according to

$$\epsilon_i = \kappa \chi_i + \nu_i$$

and

$$\eta_i = \kappa \chi_i + \varsigma_i,$$

where χ_i , ν_i and ς_i were independently generated according to a normal distribution with mean 0 and variance 1, and κ is a parameter whose value was set to $\kappa = 0.62$ for population 1, $\kappa = 0.97$ for population 2 and $\kappa = 1.51$

for population 3. Table 1 shows the population coefficient of correlation between x and y in each population.

Table 1: Coefficient of correlation for the three generated populations

Population	1	2	3
Correlation coefficient ρ_{xy}	0.29	0.49	0.70

From each population, we selected 1000 simple random samples without replacement of size $n = 200$. Then, in each generated sample, respondents were generated according to following nonresponse mechanisms:

$$\Pr[r_{xi} = 1|z_i] = \frac{\exp^{0.1+0.25z_i}}{1 + \exp^{0.1+0.25z_i}} \quad (24)$$

and

$$\Pr[r_{yi} = 1|z_i] = \frac{\exp^{0.1+0.25z_i}}{1 + \exp^{0.1+0.25z_i}} \quad (25)$$

with independent r_{xi} 's and r_{yi} 's. The marginal average of the response probabilities for both x and y was approximately equal to 65% .

In each sample, we computed $\hat{\rho}_{xyI}$ based on (i) marginal random ratio imputation (MRI), (ii) the Shao and Wang procedure (SW) and (iii) the proposed balanced random regression imputation (BRI). To measure the bias of $\hat{\rho}_{xyI}$, we used the Monte Carlo Percent Relative Bias (RB) given by

$$RB(\hat{\rho}_{xyI}) = \frac{E_{MC}(\hat{\rho}_{xyI}) - \rho_{xy}}{\rho_{xy}} \times 100, \quad (26)$$

where $E_{MC}(\hat{\rho}_{xyI}) = \frac{1}{1000} \sum_{r=1}^{1000} \hat{\rho}_{xyI}^{(r)}$ and $\hat{\rho}_{xyI}^{(r)}$ denotes the estimator $\hat{\rho}_{xyI}$ in the r -th sample, $r = 1, \dots, 1000$. To measure of variability of $\hat{\rho}_{xyI}$, we used the Monte Carlo Mean Square Error (MSE) given by

$$MSE(\hat{\rho}_{xyI}) = \frac{1}{1000} \sum_{r=1}^{1000} \left(\hat{\rho}_{xyI}^{(r)} - \rho_{xy} \right)^2. \quad (27)$$

Let $\hat{\rho}_{xyI}^{MRI}$, $\hat{\rho}_{xyI}^{SW}$ and $\hat{\rho}_{xyI}^{BRI}$ denote the estimator $\hat{\rho}_{xyI}$ under MRI, SW and BRI, respectively. In order to compare the relative efficiency of the imputed estimators, using $\hat{\rho}_{xyI}^{SW}$ as the reference, we used the following measure:

$$RE = \frac{MSE_{MC}(\hat{\rho}_{xyI}^{(\cdot)})}{MSE_{MC}(\hat{\rho}_{xyI}^{(SW)})}. \quad (28)$$

Table 2 shows the Monte Carlo Percent Relative Bias (RB) and the RE of the imputed estimator for each of the three imputation strategies: MRI, SW and BRI. It is clear that MRI did not preserve the relationship between x and y for all three populations, as expected. The RB is substantial (more than 50% in all the scenarios) and negative, which clearly illustrates that marginal imputation attenuates the relationship between variables. Unlike MRI, both SW and BRI showed a small bias for all three populations, which indicates that both imputation procedures succeeded in preserving the relationship between x and y . Turning to the RE, we note that the BRI is more efficient than SW with a value of RE equal to 0.81 for population 1, 0.79 for population 2 and 0.77 for population 3. Therefore, BRI is significantly more efficient than SW.

Table 2: Monte-Carlo Relative Bias of the imputed estimator and RE

		MRI	SW	BRI
$\kappa = 1$	$RB(\hat{\rho}_{xyI})$	-51.3	-1.6	-0.7
	RE	2.25	1	0.81
$\kappa = 2$	$RB(\hat{\rho}_{xyI})$	-57.2	-0.3	-0.5
	RE	9.45	1	0.79
$\kappa = 3$	$RB(\hat{\rho}_{xyI})$	-56.6	-0.3	-0.3
	RE	44.73	1	0.77

5.2 Performance of the Bootstrap variance estimator

We performed a limited simulation study to assess the performance of the proposed bootstrap variance estimator in terms of relative bias. We generated three populations of size $N = 10,000$ with two study variables, x and y , and one auxiliary variable z . We first generated z according to a Gamma distribution with mean 100 and standard deviation 50. Then, as in Section 5.1, bivariate data (x_i, y_i) 's were generated so that the population coefficient of correlation approximately equaled 0.3 for population 1, 0.5 for population 2 and 0.7 for population 3.

From each population, we selected $R = 1000$ simple random samples without replacement of size $n = 200$. Note that the sampling fraction n/N is equal to 0.02, which can be considered negligible. Then, in each generated sample, respondents to items x and y were generated independently according to (24) and (25).

We were interested in estimating the variance of $\hat{\rho}_{xyI}$ under the proposed balanced random regression imputation (BRI). In each sample (containing respondents and nonrespondents), we selected $B = 2,000$ bootstrap samples according to the bootstrap weight procedure in Section 4. To measure the bias of \hat{V}_B (see 23), we used the Monte Carlo percent relative bias given by

$$RB_{MC}(\hat{V}_B) = 100 \times \frac{E_{MC}(\hat{V}_B - V_{MC}(\hat{\rho}_{xyI}))}{V_{MC}(\hat{\rho}_{xyI})}, \quad (29)$$

where $E_{MC}(\hat{V}_B) = \sum_{r=1}^{1000} \hat{V}_B^{(r)} / 1000$ with $\hat{V}_B^{(r)}$ denoting the estimator \hat{V}_B in the r -th sample and $V_{MC}(\hat{\rho}_{xyI})$ is a simulation-based approximation of the true variance, obtained from an independent run of 10,000 simulations.

Finally, we computed confidence intervals by means of the percentile

method. That is, for each sample, the B bootstrap versions of the correlation coefficient, $\tilde{\rho}_{xyI}^{*(b)}$, $b = 1, \dots, B$. An $(1 - 2\alpha)$ confidence interval is then given by $[\tilde{\rho}_{xyI}^{*(L)}, \tilde{\rho}_{xyI}^{*(U)}]$ with $L = \alpha B$ and $U = (1 - \alpha) B$. Error rates of the confidence intervals (with nominal error rates of 5% and 10% in each tail) were compared.

Table 3 shows the Monte Carlo percent relative bias (RB) of the Bootstrap variance estimator and the error rates. It is clear from Table 3 that the proposed estimator performed well in all the scenarios with an absolute relative bias less than 5%. Also, the error rates were close to the nominal rates in all the cases.

Table 3: Monte Carlo percent RB (in %) and error rates of the Bootstrap variance estimator

RB	Coverage rate					
	5 %			10 %		
	L	U	L+U	L	U	L+U
<i>Population 1</i>						
-4.5	4.7	4.8	9.5	9.6	10.0	19.6
<i>Population 2</i>						
-1.8	6.6	4.7	11.3	12.0	9.9	21.9
<i>Population 3</i>						
2.6	6.4	5.4	11.8	11.9	11.0	22.9

6 Concluding Remarks

In this paper, we proposed a fully efficient version of the Shao and Wang (2002) imputation procedure. Results from a limited simulation study confirm the good performance of the proposed method both in terms of relative bias and relative efficiency. Furthermore, motivated by the reverse framework of Fay (1991) and Shao and Steel (1999), we proposed a bootstrap variance estimator, which performed well both in terms of relative bias and coverage rate of confidence intervals.

References

- Chauvet, G., Deville, J.C. and Haziza, D. (2010). On balanced random imputation in surveys. *To appear in Biometrika*.
- Deville, J.C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25, 193-203.
- Deville, J-C. and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.
- Fay, R.E. (1991). A design-based perspective on missing data Variance. *Proceedings of the 1991 Annual Research Conference, US Bureau of the Census*, 429-440.
- Haziza, D. (2009). Imputation and inference in the presence of missing data. In *Handbook of Statistics, Volume 29, Sample Surveys: Theory Methods and Inference*, Editors: C.R. Rao and D. Pfeiffermann, 215-246.

- Kim, J.K. and Fuller, W.A. (2004). Fractional hot-deck imputation. *Biometrika*, 91, 559-578.
- Rao, J. N. K. and Shao, J. (1992). On variance estimation under imputation for missing data. *Biometrika*, 79, 811-822.
- Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.
- Rubin, D. B. (1976). Inference and missing Data. *Biometrika*, 63, 581-590.
- Santos, R. (1981). Effects of imputation on regression-coefficients. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 140-145.
- Shao, J. and Sitter, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 93, 819-831.
- 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, 254-265.
- Shao, J. and Wang, H. (2002). Sample correlation coefficients based on survey data under regression imputation. *Journal of the American Statistical Association*, 97, 544-552.
- Skinner, C. J. and Rao, J. N. K. (2002). Jackknife variance for multivariate statistics under hot deck imputation from common donors. *Journal of Statistical Planning and Inference*, 102, 149-167.