Doubly robust inference for the distribution function in the presence of missing survey data

Hélène Boistard, Guillaume Chauvet and David Haziza*

May 28, 2015

Abstract

Item nonresponse in surveys occurs when some, but not all, variables are missing. Unadjusted estimators tend to exhibit some bias, called the nonresponse bias, if the respondents differ from the nonrespondents with respect to the study variables. In this paper, we focus on item nonresponse, which is usually treated by some form of single imputation. We examine the properties of doubly robust imputation procedures, which are those that lead to an estimator that remains consistent if either the outcome variable or the nonresponse mechanism are adequately modeled. We establish the double robustness property of the imputed estimator of the finite population distribution function under random hot-deck imputation within classes. We also discuss the links between our approach and that of Chambers and Dunstan (1986). The results of a simulation study support our findings.

Key words: Distribution function, doubly robust inference, imputation model approach, nonresponse model approach.

*Hélène Boistard (helene@boistard.fr), Toulouse School of Economics (GREMAQ, Université Toulouse 1). Guillaume Chauvet (chauvet@ensai.fr), ENSAI(CREST), Laboratoire de Statistique d'Enquête, Campus de Ker Lann, 35170 Bruz, France. David Haziza (David.Haziza@umontreal.ca), Département de mathématiques et de statistique, Université de Montréal, Québec, H3C 3J7, Canada.

1 Introduction

No matter how carefully survey staff try to maximize response, it is virtually certain that some degree of nonresponse will occur in large scale surveys. Survey statisticians distinguish unit nonresponse from item nonresponse. Unit nonresponse occurs because some of the sampled units refuse to respond or because of the inability to contact them. When some, but not all, variables are missing, we are in presence of item nonresponse. The latter occurs, for example, because some sample units refuse to respond to sensitive items, do not know the answer to some items, or because of edit failures. Unadjusted estimators tend to exhibit some bias, called the nonresponse bias, if the respondents differ from the nonrespondents with respect to the study variables. To reduce the nonresponse bias, weight adjustment procedures are used in the context of unit nonresponse, whereas item nonresponse is usually treated through single imputation, whereby a missing value is replaced by a single value.

In this paper, we restrict our attention to weighted random hot-deck imputation, which consists of selecting a respondent (donor) at random from the set of respondents with probability proportional to some weight (to be defined later), and then using donor's item value to "fill in" for the missing value of a nonrespondent (recipient). Random hot-deck imputation is widely used in practice, especially in social and household surveys, because it tends to preserve the distribution of the variable being imputed, which is desirable when estimating non-smooth functions such as quantiles. Also, it leads to actual (observed) imputed values, which is especially important when the variable to be imputed is categorical. Finally, with random hot-deck imputation, several missing variables may be imputed using a single donor, while satisfying post-imputation edit constraints specified by subject-matter specialists. This helps in preserving the relationship between variables of interest.

In order to study the properties of estimators in the presence of missing data,

two approaches are customarily used: (i) the nonresponse model (NM) approach that requires the specification of a nonresponse model describing the unknown nonresponse mechanism and (ii) the imputation model (IM) (also called the outcome regression model) approach that requires the specification of a model describing the distribution of the study variable. We consider doubly robust imputation/estimation procedures that have attracted some attention in recent years. An estimator is said to be doubly robust if it remains asymptotically unbiased and consistent if either model is true. Thus, doubly robust procedures offer some protection against misspecification of one model or the other. For infinite populations, the reader is referred to Robins et al. (1994), Scharfstein et al. (1999), Bang and Robins (2005), Tan (2006), Kang and Schafer (2008), Rubin and van der Laan (2008) and Cao et al. (2009), among others. In the context of finite population sampling, doubly robust procedures have been studied in Kott (1994), Kim and Park (2006), Haziza and Rao (2006), Chauvet and Haziza (2012), Kim and Haziza (2014) and Haziza et al. (2014), among others. So far, the literature on doubly robust inference has focussed on estimating simple parameters such as means and totals. Surprisingly, little attention has been paid to the problem of distribution functions in the presence of missing data. Notable exceptions include Cheng and Chu (1996), Liu et al. (2011) and Zhao et al. (2013). In the last two papers, the authors consider augmented inverse probability weighted imputation procedure to estimate the distribution function of a response variable. In this paper, we adopt a different approach that consist of using a doubly robust version of the customary weighted random hot-deck imputation procedure (see Haziza and Rao, 2006) to "fill in" for the missing values. Our objective is to produce a complete rectangular data file, which allows the secondary analysts to obtain point estimates using complete data estimation procedures.

The paper is organized as follows. After defining some notation, the underlying models and the random imputation procedure are described in Section 2. The main results are established in Section 3 under the following assumptions: (i)

the sampling design is non-informative (Pfeffermann, 1993); (ii) the data are MAR (Rubin, 1976) and (iii) the units respond independently of one another. The assumption (iii) can be relaxed to consider the case of a correlated response behaviour, see the discussion in Section 7. Since this assumption is fairly usual in the literature and so as to simplify the presentation, we keep the assumption of independent response behaviour in the body of the paper. In Section 4, we discuss the link between the proposed method and an estimation procedure proposed by Chambers and Dunstan (1986) in the context of model-based estimation of finite population distribution functions. A simulation study is conducted in Section 5. In Section 6, we illustrate the proposed methodology using data modeled from one industry in the Monthly Retail Trade Survey conducted by the U.S. Census Bureau. We make some final remarks in Section 7.

2 Theoretical set-up

Consider a finite population U of size N. We are interested in estimating the finite population distribution function, $F_{N,y}(t) = N^{-1}\sum_{i\in U} 1(y_i \leq t)$, defined for $t \in \mathbb{R}$, where y denotes a study variable and $1(.)$ the usual indicator function. A sample s of size n is selected according to a given sampling design $p(\cdot)$. Let $d_i = 1/\pi_i$ be the sampling weight attached to unit i, with π_i denoting its first-order inclusion probability in the sample. The inclusion probabilities π_i are assumed to be known for all $i \in U$. A complete data estimator of $F_{N,y}(t)$ is

$$\hat{F}_{N,y}(t) = \sum_{i\in s} \tilde{d}_i 1(y_i \leq t), \tag{1}$$

where $\tilde{d}_i = \left(\sum_{j\in s} d_j\right)^{-1} d_i$. When some y-values are missing, an estimator of $F_{N,y}(t)$ is the imputed estimator

$$\hat{F}_{I,y}(t) = \sum_{i\in s} \tilde{d}_i r_i 1(y_i \leq t) + \sum_{i\in s} \tilde{d}_i (1-r_i) 1(y_i^* \leq t), \tag{2}$$

where y_i^* denotes the imputed value used to replace the missing y_i and r_i is a response indicator attached to unit i such that $r_i = 1$ if y_i is observed and $r_i = 0$ if y_i is missing.

Assume that the finite population U is divided into G mutually disjoint imputation cells, $U_1, \ldots, U_G$. Let n_g be the size of $s_g = s \bigcap U_g$, $g = 1, \ldots, G$ and s_{rg} be the set of respondents in cell g, of size n_{rg}. The elements in cell U_g are assumed to be a realization of independently and identically distributed random variables with mean μ_g and variance σ_g^2; that is,

$$m : y_i \sim (\mu_g, \sigma_g^2), \quad i \in U_g. \tag{3}$$

The imputation cells are formed on the basis of auxiliary information, $\mathbf{x}$, recorded for both respondents and nonrespondents. Model (3) is the common mean model within imputation cells. It is often called the imputation model (IM) or the outcome regression model.

Let $p_i = P(r_i = 1)$ be the response probability to item y for unit i. We assume that units respond independently of one another; that is, $p_{ij} = P(r_i = 1, r_j = 1) = p_i p_j$, $i \neq j$. Further, we assume that the p_i's can be modeled trough a parametric model

$$p_i = p(\mathbf{x}_i, \boldsymbol{\alpha}) \tag{4}$$

for some vector of unknown parameters $\boldsymbol{\alpha}$. Model (4) is called the nonresponse model (NM). The estimated response probability $\hat{p}_i$ attached to unit i is $\hat{p}_i = p\left(\mathbf{x}_i, \hat{\boldsymbol{\alpha}}\right)$, where $\hat{\boldsymbol{\alpha}}$ is a consistent estimator of $\boldsymbol{\alpha}$.

We assume that the data are Missing At Random (Rubon, 1976):

$$E_m(y_i \mid \mathbf{x}_i, r_i = 1) = E_m(y_i \mid \mathbf{x}_i, r_i = 0), \tag{5}$$

where $E_m(.)$ denotes the expectation with respect to the imputation model (3).

In order to study the theoretical properties of point estimators we consider two inferential approaches: the NM approach, whereby, inference is made with respect to the joint distribution induced by the sampling design and the assumed nonresponse model given by (4); the IM approach, whereby inference is made with respect to the joint distribution induced by the imputation model (3), the sampling design and the nonresponse model. In the latter approach, explicit assumptions about the non-response mechanism are not required but the data are assumed to be MAR.

Based on (3), it may be tempting to estimate $F_{N,y}(t)$ by

$$\hat{F}_{I,y}(t) = \sum_{g=1}^{G} \left\{ \sum_{i \in s_g} \tilde{d}_i r_i 1(y_i \leq t) + \sum_{i \in s_g} \tilde{d}_i (1 - r_i) 1(\bar{y}_{rg} \leq t) \right\}, \tag{6}$$

for $t \in \mathbb{R}$, where $\bar{y}_{rg} = \left(\sum_{i \in s_g} d_i r_i \right)^{-1} \sum_{i \in s_g} d_i r_i y_i$ denotes the weighted mean of respondents in cell g. This corresponds to mean imputation within cells. However, the estimator (6) is generally biased as mean imputation tends to distort the distribution of the variable being imputed. Indeed, the variability of the study variable y after imputation within each imputation cell is smaller than the natural variability that would have been observed in the complete data case. The relative distortion within each imputation cell increases as the expected response rate within cells decreases.

We consider an alternative approach, whereby a missing value is treated through random hot-deck imputation within classes. More specifically, missing y_i in cell U_g is replaced with

$$y_i^* = y_j \text{ for } j \in s_{rg}, \tag{7}$$

with probability

$$pr(y_i^* = y_j) = \tilde{\omega}_j = \frac{d_j \ \frac{1-\hat{p}_j}{\hat{p}_j}}{\sum_{l \in s_g} r_l d_l \ \frac{1-\hat{p}_l}{\hat{p}_l}}. \tag{8}$$

The use of the imputed values (7) leads to a doubly robust estimator of a population total (or mean). That is, the resulting estimator remains consistent if either the imputation model (3) or the nonresponse model (4) is correctly specified; see Haziza and Rao (2006). In the next section, we establish the uniform consistency property of $\hat{F}_I(t)$ with respect to both the NM approach and the IM approach. We adopt the following notation: let $E_p(\cdot)$, $E_q(\cdot)$ and $E_I(\cdot)$ denote the expectation with respect to the sampling design, the nonresponse model and the imputation mechanism, respectively.

3 Main results

We study the asymptotic properties of the estimated distribution function under the random imputation procedure described in Section 2. We assume that there exists a sequence of sampling designs and finite populations, indexed by ν, such that the population size N_ν, the sample size n_ν and the number of respondents $n_{r\nu}$ tend to infinity when $\nu \to \infty$. Though we suppress the index ν to simplify the notation, the limits are understood as when $\nu \to \infty$.

Under mild regularity conditions, Theorem 1 in Chauvet et al. (2011) implies that, for any $t \in \mathbb{R}$, $\hat{F}_{I,y}(t) - F_{N,y}(t)$ converges in probability to 0 under the IM approach. It is thus sufficient to prove consistency under the NM approach. We make the following regularity assumptions:

C1a: For any $i \neq j \in U$, $\pi_{ij} - \pi_i \, \pi_j \leq 0$;

C1b: $\max_{i \neq j \in U} |\pi_{ij} - \pi_i \, \pi_j| = O(n^{-1})$;

C2: There exists some constant $0 < f < 1$ such that $n/N \to f$;

C3: There exists some constants $C_1, C_2 > 0$ such that for any $i \in U$:

$$C_1 \frac{N}{n} \leq d_i \leq C_2 \frac{N}{n};$$

C4: There exists some constant $0 < \kappa < 1$ such that $\kappa < p_i$ for any $i \in s$;

C5: $E_{pq}\left(\sum_{g=1}^{G}\left|\frac{N}{\sum_{k\in s_g} d_k \frac{1-p_k}{p_k} r_k} - \frac{N}{\sum_{k\in s_g} d_k(1-p_k)}\right|\right)^2 = O(n^{-1})$.

C6: For any cell U_g, $E_{pq}\left(\max_{j\in s_{rg}}\left|\frac{\tilde{\omega}_j - \breve{\omega}_j}{\breve{\omega}_j}\right|\right) \to 0$ and $E_{pq}\left(\max_{j\in s_{rg}}\left|\frac{\tilde{\omega}_j - \breve{\omega}_j}{1-\breve{\omega}_j}\right|\right) \to 0$, where

$$\breve{\omega}_j \quad = \quad \frac{d_j \ \frac{1-p_j}{p_j}}{\sum_{l\in s_{rg}} d_l \ \frac{1-p_l}{p_l}}$$

is obtained from $\tilde{\omega}_j$ by replacing the estimated $\hat{p}_j$ with the true probability p_j.

Assumptions C1a, C1b, C2 and C3 are standard regularity conditions, see for example Breidt and Opsomer (2000). In particular, Assumption C3 guarantees that no extreme weight dominates the others.

Theorem 1 *Suppose that assumption C1a or C1b holds. Suppose that assumptions C2–C6 hold. Then*

$$E\left|\hat{F}_{I,y}(t) - F_{N,y}(t)\right| \underset{\nu\to\infty}{\longrightarrow} 0. \tag{9}$$

In particular, $\hat{F}_{I,y}(t) - F_{N,y}(t)$ converges in probability to 0 for any $t \in \mathbb{R}$, and the imputed values (7) lead to a consistent imputed pointwise estimator of the distribution function with respect to the NM approach.

The point-wise consistency of the imputed distribution function may be strengthened to uniform consistency, making use of the following additional assumption:

C8: For all $\epsilon > 0$, there exists ν_0, $M \in \mathbb{N}$ and $t_1 < \ldots < t_M$ such that $\forall N \geq \nu_0$,

$$0 \leq F_{N,y}(t_i-) - F_{N,y}(t_{i-1}) \leq \epsilon. \tag{10}$$

Theorem 2 *Assume that the conditions in Theorem 1 hold. Assume that condition C8 holds. Then,*

$$\sup_{t\in\mathbb{R}}\left|\hat{F}_{I,y}(t)-F_{N,y}(t)\right|\xrightarrow{Pr}0 \quad as \quad N\to\infty.$$

4 Link with Chambers and Dunstan (1986)

In the context of model-based inference, Chambers and Dunstan (1986) proposed an estimator of the distribution function, where the values of the non-sampled units are predicted through a linear regression model; see also Valliant et al. (2000). In this section, we derive in the somehow different context of missing data, a Dunstan-Chambers type estimator and establish the link with our approach. First, the complete data estimator $\hat{F}_{N,y}(t)$ in (1) can be written as

$$\hat{F}_{N,y}(t)=\sum_{g=1}^{G}\left\{\sum_{i\in s_g}\tilde{d}_i r_i 1(y_i\le t)+\sum_{i\in s_g}\tilde{d}_i(1-r_i)1(y_i\le t)\right\}. \tag{11}$$

While the first term on the right hand-side of (11) can be computed from the responding units, the second term needs to be estimated. The expectation of the latter with respect to model (3) is

$$\begin{aligned}\sum_{g=1}^{G}\sum_{i\in s_g}\tilde{d}_i(1-r_i)P\left(y_i\le t\right) &= \sum_{g=1}^{G}\sum_{i\in s_g}\tilde{d}_i(1-r_i)P\left(y_i-\mu_g\le t-\mu_g\right)\\ &= \sum_{g=1}^{G}\sum_{i\in s_g}\tilde{d}_i(1-r_i)P\left(\epsilon_i\le t-\mu_g\right)\\ &= \sum_{g=1}^{G}\sum_{i\in s_g}\tilde{d}_i(1-r_i)G_g\left(t-\mu_g\right), \end{aligned}\tag{12}$$

where $G_g(x)=P(\epsilon_i\le x)$ denotes the distribution function of the errors $\epsilon_i=y_i-\mu_g$ in cell g. Following Chambers and Dunstan (1986), a natural estimator of (12) consists of replacing G_g by an estimator based on the residuals observed

on the responding units, given by

$$\hat{G}_g(x) = \sum_{j \in s_g} \tilde{\omega}_i r_i 1(\hat{e}_j \leq x),$$

where $\hat{e}_j = y_j - \bar{y}^*_{rg}$ with $\bar{y}^*_{rg} = \left(\sum_{i \in s_g} d_i r_i \tilde{\omega}_i\right)^{-1} \sum_{i \in s_g} d_i r_i \tilde{\omega}_i y_i$. This leads to the following Dunstan-Chambers type estimator of $F_{N,y}(t)$:

$$\begin{aligned} \tilde{F}_{I,y}(t) &= \sum_{g=1}^{G} \left\{ \sum_{i \in s_g} \tilde{d}_i r_i 1(y_i \leq t) + \sum_{i \in s_g} \tilde{d}_i (1 - r_i) \sum_{j \in s_g} \tilde{\omega}_i r_i 1(\hat{e}_j \leq t - \bar{y}^*_{rg}) \right\} \\ &= \sum_{g=1}^{G} \left\{ \sum_{i \in s_g} \tilde{d}_i r_i 1(y_i \leq t) + \sum_{i \in s_g} \tilde{d}_i (1 - r_i) \sum_{j \in s_g} \tilde{\omega}_i r_i 1(y_j \leq t) \right\}. \quad (13) \end{aligned}$$

Now, the expectation of (2) under the imputation procedure (7) with respect to the imputation mechanism is given by

$$\begin{aligned} E_I\{\hat{F}_{I,y}(t)\} &= \sum_{g=1}^{G} \left\{ \sum_{i \in s_g} \tilde{d}_i r_i 1(y_i \leq t) + \sum_{i \in s_g} \tilde{d}_i (1 - r_i) \sum_{j \in s_g} \tilde{\omega}_i r_i 1(y_j \leq t) \right\} \\ &= \tilde{F}_{I,y}(t), \end{aligned}$$

which is identical to (13). That is, the Dunstan-Chambers type estimator can be viewed as the integrated imputed estimator with respect to the imputation mechanism. It is worth noting that the Dunstan-Chambers type estimator is also doubly robust for the distribution function. That is,

$$E\left|\tilde{F}_{I,y}(t) - F_{N,y}(t)\right| \underset{\nu \to \infty}{\longrightarrow} 0 \quad (14)$$

under the conditions of Theorem 1, for both the IM approach and the NM approach. The proof is briefly sketched in Appendix.

5 Simulation study

We performed a limited simulation study to evaluate the proposed method, in terms of relative bias and relative efficiency. We first generated a finite population of size $N = 10,000$, with one variable of interest y and two auxiliary variables x_1 and x_2. The auxiliary variables were generated according to a Gamma distribution with shape and scale parameters 5 and 2, respectively. Given the x_1-values and the x_2-values, the y-values were generated according to the model

$$y_i = \beta_0 + \beta_1 \, x_{1i} + \beta_2 \, x_{2i} + \eta_i. \tag{15}$$

The parameters β_0, β_1 and β_2 were set to 10, 1 and 1, respectively. The η_i were generated according to a Normal distribution with mean 0 and variance σ^2, whose value was set so as to obtain a coefficient of determination (R^2) approximately equal to 0.7. Results obtained with a coefficient of determination approximately equal to 0.2 or 0.3 were similar, and are thus not presented.

We were interested in estimating the distribution function $F_{N,y}(t)$ for $t = t_\alpha$, with t_α the α-th quantile. We considered $\alpha = 0.05, 0.25, 0.50, 0.75$ and 0.95 in the simulation.

From the population, we selected $1,000$ samples of size $n = 500$ by simple random sampling without replacement. In each sample, nonresponse was generated according to the nonresponse mechanism

$$Pr(r_i = 1 | x_{1i}, x_{2i}) \quad = \quad \frac{\exp\left(-1 + 1.6 \; x_{1i} + 1.6 \; x_{2i}\right)}{1 + \exp\left(-1 + 1.6 \; x_{1i} + 1.6 \; x_{2i}\right)}. \tag{16}$$

The coefficients in (16) were chosen to lead to an average response rate approximately equal to 0.6. Results obtained with average response rates of 0.4 and 0.5 led to similar results and are thus not presented.

To replace the missing values, we used random hot-deck imputation within classes based on different working models. In each case, we first formed 10 imputation classes using the so-called score method (Haziza and Beaumont, 2007). First, predicted values $\hat{y}_i$ were obtained for both respondents and nonrespondents using a linear regression model based on an $\mathbf{x}$-vector of auxiliary variables. That is, $\hat{y}_i = \mathbf{x}_i^\top \hat{\mathbf{B}}_r$, $i = 1, \cdots, n$, where $\hat{\mathbf{B}}_r$ denotes the weighted least square estimator based on the responding units. Then, imputation classes, based on the $\hat{y}$-values, were formed using the equal quantile method. That is, the $\hat{y}$-values were ordered from the smallest value to the largest value and the sample was partitioned into 10 equal size imputation classes. Within each class, the missing values were imputed according to (7).

We considered four distinct scenarios:

Scenario 1: Both the imputation and the nonresponse model were correctly specified. The predicted values $\hat{y}_i$ were obtained using $\mathbf{x} = (1, x_1, x_2)^\top$ as the vector of auxiliary variables, whereas the $\hat{p}_i$'s in (7) where obtained through a logistic regression model using $\mathbf{x} = (1, x_1, x_2)^\top$ as the vector of auxiliary variables.

Scenario 2: Only the nonresponse model was correctly specified. The predicted values $\hat{y}_i$ were obtained using $\mathbf{x} = (1, x_1)^\top$ as the vector of auxiliary variables, whereas the $\hat{p}_i$'s in (7) where obtained trough a logistic regression model using $\mathbf{x} = (1, x_1, x_2)^\top$ as the vector of auxiliary variables.

Scenario 3: Only the imputation model was correctly specified. The predicted values $\hat{y}_i$ were obtained using $\mathbf{x} = (1, x_1, x_2)^\top$ as the vector of auxiliary variables. We used two misspecified nonresponse models, which led to Scenario 3a and Scenario 3b. In Scenario 3a, we set $\omega_i = 1$ for all i in (7), which led to unweighted random hot-deck imputation within classes. Note that this corresponds to a nonresponse model containing only the intercept, leading to the overall response rate as the estimated response probability for all i. In Scenario 3b, the $\hat{p}_i$'s in (7) were obtained through

a logistic regression model using $\mathbf{x} = (1, x_1)^\top$ as the vector of auxiliary variables.

Scenario 4: Both the nonresponse model and the imputation model were misspecified. The predicted values $\hat{y}_i$ were obtained using $\mathbf{x} = (1, x_1)^\top$ as the vector of auxiliary variables. We used two misspecified nonresponse models, which led to Scenario 4a and Scenario 4b. In Scenario 4a, we set $\omega_i = 1$ for all i in (7) as in Scenario 3a. In Scenario 4b, the $\hat{p}_i$'s in (7) where obtained through a logistic regression model using $\mathbf{x} = (1, x_1)^\top$ as in Scenario 3b.

Then, in each sample, we computed the imputed estimator of $F_{N,y}(t)$, denoted by $\hat{F}_{I,y}(t)$, given by (2). To measure the bias of $\hat{F}_{I,y}(t)$, we used the percent Monte Carlo relative bias

$$\text{RB}\{\hat{F}_{I,y}(t)\} = \frac{E_{MC}\{\hat{F}_{I,y}(t)\} - F_{N,y}(t)}{F_{N,y}(t)} \times 100, \tag{17}$$

where $E_{MC}\{\hat{F}_{I,y}(t)\} = \sum_{r=1}^{1000} \hat{F}_{I,y}^{(r)}(t)/1000$, with $\hat{F}_{I,y}^{(r)}(t)$ denoting the estimator $\hat{F}_{I,y}(t)$ for the r-th sample, $r = 1, \ldots, 1000$. To measure the variability of $\hat{F}_{I,y}(t)$, we used the percent Monte Carlo relative root mean square error

$$\text{RRMSE}\{\hat{F}_{I,y}(t)\} = \frac{\sqrt{\text{MSE}\{\hat{F}_{I,y}(t)\}}}{F_{N,y}(t)} \times 100,$$

where

$$\text{MSE}\{\hat{F}_{I,y}(t)\} = \frac{1}{1000} \sum_{r=1}^{1000} \{\hat{F}_{I,y}^{(r)}(t) - F_{N,y}(t)\}^2.$$

The results are shown in Table 1. The imputed estimator $\hat{F}_{I,y}(t)$ showed a small bias in Scenarios 1-3 for all values of α. We note a slight bias in Scenarios 2a and 2b for $\alpha = 0.05$ with values of absolute RB equal to 2.0% and 1.7%, respectively. These results suggest that the imputation procedure (7) lead to a doubly robust estimator of the distribution function. As expected, when both models were misspecified (Scenario 4a et 4b), the imputed estimator $\hat{F}_{I,y}(t)$

		α				
		0.05	0.25	0.50	0.75	0.95
Scenario 1	RB	−0.2	0.1	0.3	0.1	0.0
	RRMSE	32.4	10.8	5.6	2.9	1.1
Scenario 2	RB	−0.7	−0.1	−0.0	0.0	0.0
	RRMSE	33.3	10.7	5.7	2.9	1.1
Scenario 3a	RB	−2.0	−0.6	−0.1	−0.0	−0.1
	RRMSE	31.1	10.8	5.5	2.9	1.1
Scenario 3b	RB	−1.7	−0.1	0.0	0.0	−0.0
	RRMSE	32.2	10.7	5.5	2.9	1.1
Scenario 4a	RB	−19.2	−13.7	−9.3	−5.0	−1.2
	RRMSE	34.0	17.1	11.1	6.1	1.8
Scenario 4b	RB	−19.3	−13.5	−9.0	−4.9	−1.0
	RRMSE	33.4	17.0	10.8	6.0	1.7

Table 1: Monte Carlo percent relative bias and percent relative root mean square error of the imputed estimator of the distribution function for several values of α

exhibited a significant bias for all the scenarios, except for $\alpha = 0.95$. Turning to the efficiency of $\hat{F}_{I,y}(t)$, we note that for a given value of α, the first three scenarios led to similar values of RRMSE.

6 Application to the Monthly Retail Trade Survey

For the purpose of illustration of the proposed methodology, we used data modeled from one industry in the Monthly Retail Trade Survey (MRTS) conducted by the U.S. Census Bureau (Mulry et al., 2014). For confidentiality reasons, the real data could not be used but the simulated data are realistic and designed to match closely the original survey data in the first moments and in the correlation structure. As variables of interest, we considered the sales (y_1) and the inventories (y_2). Both y_1 and y_2 were either observed jointly or missing jointly. As a size measure, we used the variable receipts (x) that was available on the sampling frame. The stratum identifiers and the sampling weights d_i

were available on the data file.

The data set came from a stratified simple random sampling design with six strata U_h, including one take-all stratum. For simplicity, we focussed on the five take-some strata. The take-all stratum contained very large units that may require a particular nonresponse treatment, which is why it was not considered further. Table 2 shows the number of sampled units n_h and the number of respondents n_{rh} in each stratum. The estimated response probability inside the stratum U_h is $\hat{p}_h = \frac{n_{rh}}{n_h}$.

Stratum h	1	2	3	4	5
Number of units N_h	$9,493$	$6,244$	$3,259$	$1,397$	463
Sample size n_h	145	123	72	57	61
Number of respondents n_{rh}	73	75	55	44	54

Table 2: Total number of units, number of sampled units and number of respondents inside strata

6.1 Imputation for sales

For the sales, we performed random hot-deck imputation within imputation cells that were defined as follows: each stratum was divided into 2 or 3 imputation cells based of the x-variable. That is, in each stratum, we ordered the units with respect to their x-values and imputation cells were formed so that each imputation cell had approximately the same number of sampled units. We note G_h^1 the number of imputation cells U_{hg} within the stratum U_h. We used $G_1^1 = G_2^1 = 3$ cells for the strata $h = 1$ and $h = 2$, and we used $G_3^1 = G_4^1 = G_5^1 = 2$ cells for the three remaining strata. In this case, the imputed estimator of the distribution function in (2) may be rewritten as

$$\hat{F}_{I,y_1}(t) = \frac{1}{N} \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{g=1}^{G_h^1} \left\{ \sum_{i \in s_{hg}} r_i 1(y_{1i} \leq t) + \sum_{i \in s_{hg}} (1 - r_i) 1(y_{1i}^* \leq t) \right\} \quad (18)$$

with $s_{hg} = s \cap U_{hg}$, of size n_{hg}. Within each imputation cell, random hot-deck imputation based on the imputation weights

$$\tilde{\omega}_j = \frac{d_j}{\sum_{l \in s_g} r_l d_l} = \frac{1}{n_{rhg}} \quad \text{for} \quad j \in s_{rhg} \tag{19}$$

was used, where $s_{rhg} = s_r \cap U_{hg}$ denotes the subset of respondents in U_{hg}, of size n_{rhg}. The imputation weights in (**??**) are a special case of the imputation weights in (8). This is essentially equivalent to assuming equal response probabilities within imputation cells. Point estimates for different values of t are shown in Table 3.

In addition, we estimated the variance of $\hat{F}_{I,y_1}(t)$ under the NM approach. For simplicity, we neglected the sampling rates within strata. Our variance estimator uses the fact that conditionally on n_{hg} and n_{rhg}, we can treat the set of respondents in each imputation cell as a simple random sample without replacement. In this case, an approximately unbiased estimator of the total variance of $\hat{F}_{I,y_1}(t)$ is

$$\begin{aligned} v_1\{\hat{F}_{I,y_1}(t)\} &= \frac{1}{N^2} \sum_{h=1}^{H} \left(\frac{N_h}{n_h}\right)^2 \sum_{g=1}^{G_h^1} \frac{\{n_{hg}\}^2}{n_{rhg}-1} \hat{F}_{rhg,y1}(t)\{1-\hat{F}_{rhg,y1}(t)\} \\ &+ \frac{1}{N^2} \sum_{h=1}^{H} \left(\frac{N_h}{n_h}\right)^2 \sum_{g=1}^{G_h^1} \{n_{mhg}\} \hat{F}_{rhg,y1}(t)\{1-\hat{F}_{rhg,y1}(t)\}, \end{aligned} \tag{20}$$

where

$$\hat{F}_{rhg,y1}(t) = \frac{1}{n_{rhg}} \sum_{i \in s_{rhg}} 1(y_i \leq t),$$

and where n_{mhg} denotes the size of the set of non-respondents s_{mhg} in U_{hg}. The first term on the right-hand side of (20) is an estimator of the variance due to both the sampling design and the non-response mechanism, while the second term is an estimator of the imputation variance due to the random selection

of donors within each imputation cell. We also considered a with-replacement bootstrap variance estimator, which is obtained as follows:

1. Draw a simple random sample with replacement s^*_{rhg} of size n_{rhg} from s_{rhg}, and a simple random sample with replacement s^*_{mhg} of size n_{mhg} from s_{mhg} for any cell U_{hg}.

2. For any $i \in s^*_{mhg}$, the value y_i is replaced with $y^{**}_i = y_j$ for $j \in s^*_{rhg}$ with probability $\tilde{\omega}_j$. We compute

$$\hat{F}^*_{I,y_1}(t) = \frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{g=1}^{G^1_h}\left\{\sum_{i\in s^*_{rhg}} 1(y_{1i}\leq t) + \sum_{i\in s^*_{mhg}}(1-r_i)1(y^{**}_{1i}\leq t)\right\}.$$

3. Repeat Steps 1 and 2 a large number of times, B, to get $\hat{F}^{*(1)}_{I,y_1}(t), \ldots, \hat{F}^{*(B)}_{I,y_1}(t)$. The variance of $\hat{F}_{I,y_1}(t)$ is estimated by

$$v_{boot}\{\hat{F}_{I,y_1}(t)\} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{F}^{*(b)}_{I,y_1}(t) - \frac{1}{B}\sum_{c=1}^{B}\hat{F}^{*(c)}_{I,y_1}(t)\right)^2. \quad (21)$$

The estimated coefficients of variation obtained from (20) and (21) with $B = 1,000$, are given in Table 3. We note that both variance estimation procedures led to very similar results.

t ($\times 1,000$)	300	700	1,000	2,000	5,000	8,000	10,000
$\hat{F}_{I,y_1}(t)$	0.20	0.33	0.44	0.63	0.91	0.97	0.99
$cv_1\{\hat{F}_{I,y_1}(t)\}$ (%)	7.41	2.51	1.72	1.46	0.26	0.26	0.09
$cv_{boot}\{\hat{F}_{I,y_1}(t)\}$ (%)	7.29	2.58	1.72	1.44	0.26	0.26	0.09

Table 3: Imputed distribution function and estimated coefficients of variation for the variable y_1

6.2 Imputation for inventories

For the inventories, the imputation cells were formed as follows: all the sampled units were ordered according to their x-values. Then, we divided the ordered

sample in $G^2 = 10$ imputation cells of approximately equal size. Unlike for the sales, note that the cells cut across the sampling strata. In this context, the imputed estimator of the distribution function in (2) may be rewritten as

$$\hat{F}_{I,y_2}(t) = \frac{1}{N} \sum_{g=1}^{G^2} \sum_{h=1}^{H} \frac{N_h}{n_h} \left\{ \sum_{i \in s_{hg}} r_i 1(y_{2i} \leq t) + \sum_{i \in s_{hg}} (1 - r_i) 1(y_{2i}^* \leq t) \right\}. \quad (22)$$

Once again, we used random hot-deck imputation within each cell, whereby the imputation weights were given by (8). Note that these imputation weights were not constant inside the imputation cells (unlike for the sales) since the cells cut across strata. Point estimates for several values of t and estimated coefficients of variation based on the bootstrap estimator (21) with $B = 1,000$ are shown in Table 4.

t ($\times 1,000$)	400	800	1,300	3,700	5,500	8,500
$\hat{F}_{I,y_2}(t)$	0.13	0.22	0.28	0.59	0.75	0.85
$cv_{boot}\{\hat{F}_{I,y_2}(t)\}$ (%)	10.52	7.40	6.26	3.30	2.17	1.33

Table 4: Imputed distribution function and estimated coefficients of variation for the variable y_2

7 Discussion

We established the double robustness property of the imputed estimator of the distribution function under random hot-deck within classes. Our results were based on the assumption that units respond independently of one another. In practice, a correlated response behaviour may occur. To cover such cases, we briefly describe an extension of the proposed hot-deck method. Denote by $\hat{p}_{ij}$ the estimation of the joint response probability, and by

$$\hat{p}_{j|i} = \frac{\hat{p}_{ij}}{\hat{p}_i}$$

the estimation of the conditional response probability $p_{j|i} = \frac{p_{ij}}{p_i}$. The extended random hot-deck imputation is as follows: missing y_i in cell U_g is replaced with

$$y_i^* = y_j \text{ for } j \in s_{rg}, \tag{23}$$

with probability

$$pr(y_i^* = y_j) \quad = \quad \tilde{\omega}_{j|i} = \frac{d_j \ \frac{1-\hat{p}_{j|i}}{\hat{p}_{j|i}}}{\sum_{l \in s_g} r_l d_l \ \frac{1-\hat{p}_{l|i}}{\hat{p}_{l|i}}}. \tag{24}$$

Mimicking the proof of Theorem 1, it can be shown that equation (9) still holds under this hot-deck procedure; the proof is available from the authors. In the particular case when the units respond independently, we obtain $p_{j|i} = p_j$ and $\hat{p}_{j|i} = \hat{p}_j$, which leads to the random hot-deck imputation method described in Section 2.

Also, note that the assumption of independent response behaviour is generally not tenable for multi-stage surveys (e.g., household surveys) as units within clusters tend to be correlated with respect to the variable being imputed as well with respect to the response behaviour. In this context, a more appropriate imputation model would be the linear mixed model within imputation cells

$$y_{ki} = \mu_g + \nu_k + \epsilon_{ki} \tag{25}$$

if element (ij) belongs to class g, where y_{ki} denotes the y-value attached to unit i in cluster k, ν_k is i-th cluster random effect and ϵ_{ki} is the residual error; e.g., Haziza and Rao (2010) and Lago and Clark (2015) for more details. Also, estimation of response probabilities based upon conditional logistic regression in the context of correlated responses has been studied by Skinner and D'Arrigo (2011). A doubly robust random hot-deck imputation procedure may be obtained by using a random best linear unbiased prediction procedure (Lago and Clark, 2015) based on (25). That is, the imputed value $\hat{y}_{ki}$ for missing y_{ki} is

given by

$$\hat{y}_{ki} = \hat{\mu}_g + \hat{\nu}_k + e^*_{ki},$$

where $\hat{\mu}_g$ is a suitable estimator of μ_g, $\hat{\nu}_k$ is a suitable predictor of ν_k and the e^*_{ki}'s are residuals selected at random with appropriate probabilities. The choice of appropriate probabilities is a topic of future research.

Estimating the variance of $\hat{F}_{I,y}(t)$ in the case of nonnegligible sampling fractions is a challenging problem. Mashreghi et al. (2015) proposed a doubly robust bootstrap procedure and showed empirically that the proposed procedure performed well in terms of bias and coverage of confidence intervals for distribution functions and quantiles. As for a doubly robust point estimator, a doubly robust variance estimator remains consistent for the true variance (which can be expressed as the sum of the sampling, nonresponse and imputation variances) if either the nonresponse or the imputation model is correctly specified. The bootstrap procedure of Mashreghi et al. (2015) belongs to the class of pseudo-populations bootstrap methods, whereby a pseudo-population is first constructed from the set of respondents before selecting the samples and generating nonresponse within each selected sample.

Under random hot-deck imputation, the imputed estimator $\hat{F}_{I,y}(t)$ suffers from the imputation variance that arises from the selection of donors within classes, leading to a potentially inefficient estimator. Reducing/eliminating the imputation variance may be achieved by extending the results of Chauvet et al. (2011) to the case of a distribution function. The idea is to select donors at random while respecting appropriate constraints, which can be achieved through a balanced selection of donors. An alternative to balanced imputation is fractional imputation, whereby several imputed values are used to replaced a missing value, each being assigned a fractional weight; see Kim and Fuller (2004). This topic is currently under investigation.

In practice, it is often required to estimate more complex parameters such as quantiles or complex indicators of poverty and inequality such as Gini coefficients. These parameters depend on the distribution function. Thus, it would be useful to establish the theoretical properties of imputed estimators such as double robustness for this type of complex parameters.

Appendix A: Proofs of results

A preliminary lemma

For any missing y_i, we note y_i^{**} the value that would have been imputed if the true response probabilities were known. That is, for missing y_i in cell U_g, we take:

$$y_i^{**} = y_j \quad \text{for } j \in s_{rg} \quad \text{with probability } \breve{\omega}_j = \frac{d_j \ \frac{1-p_j}{p_j}}{\sum_{l \in s_{rg}} d_l \ \frac{1-p_l}{p_l}}. \tag{26}$$

We note

$$T_1 \quad = \quad \sum_{i \in s} \tilde{d}_i (1 - r_i) \left\{ 1(y_i^* \leq t) - 1(y_i^{**} \leq t) \right\} \tag{27}$$

Lemma 1 *Suppose that assumption C1a or C1b holds. Suppose that assumptions C2–C6 hold. Then* $E|T_1| \underset{\nu \to \infty}{\longrightarrow} 0$.

Proof. We first note that the imputed value y_i^* (using the estimated probabilities $\hat{p}_l,\ l \in s_{rg}$) and the virtual imputed value y_i^{**} (using the true probabilities $p_l,\ l \in s_{rg}$) may be obtained as follows. Following Algorithm 6.2 in [**?**], we consider α_g the largest value in $[0, 1]$ such that

$$0 \leq \omega_j^0 = \frac{\tilde{\omega}_j - \alpha_g \ \breve{\omega}_j}{1 - \alpha_g} \leq 1 \qquad \text{for any } j \in s_{rg},$$

which may be alternatively defined as $\alpha_g = \min\left(\min_{j\in s_{rg}} \frac{\tilde{\omega}_j}{\check{\omega}_j}, \min_{j\in s_{rg}} \frac{1-\tilde{\omega}_j}{1-\check{\omega}_j}\right)$. Then, let

$$\begin{aligned} y_i^{**} &= y_j \quad \text{for } j \in s_{rg} \quad \text{with probability } \check{\omega}_j, \\ y_i^{0*} &= y_j \quad \text{for } j \in s_{rg} \quad \text{with probability } \omega_j^0, \end{aligned}$$

and let ϵ_g denote independent random variables such that $\epsilon_g = 1$ with probability α_g, and $\epsilon_g = 0$ with probability $1 - \alpha_g$. Then any missing y_i in cell U_g is replaced with $y_i^* = \epsilon_g\ y_i^{**} + (1-\epsilon_g)\ y_i^{0*}$. It is straightforward to show that this procedure leads to $y_i^* = y_j$ for $j \in s_{rg}$ with probability $\tilde{\omega}_j$. This joint imputation procedure enables to generate an imputed value y_i^* which is close to the imputed value y_i^{**} that we would obtain with the true probabilities $p_l,\ l \in s_{rg}$. In fact, y_i^* and y_i^{**} are identical with a probability α_g.

Now, we can write

$$T_1 \quad = \quad \sum_{g=1}^{G}(1-\epsilon_g)\sum_{i\in s_g}\tilde{d}_i(1-r_i)\left\{1(y_i^{0*}\le t) - 1(y_i^{**}\le t)\right\}$$

so that

$$\begin{aligned} |T_1| \quad &\le \quad \sum_{g=1}^{G}(1-\epsilon_g)\sum_{i\in s_g}\tilde{d}_i(1-r_i) \\ &\le \quad \sum_{g=1}^{G}(1-\epsilon_g) \end{aligned}$$

and

$$\begin{aligned} E_{pq}|T_1| \quad &\le \quad E_{pq}\left\{\sum_{g=1}^{G}(1-\alpha_g)\right\} \\ &\le \quad E_{pq}\left\{\sum_{g=1}^{G}\max\left(\max_{j\in s_{rg}}\left|\frac{\tilde{\omega}_j-\check{\omega}_j}{\check{\omega}_j}\right|, \max_{j\in s_{rg}}\left|\frac{\tilde{\omega}_j-\check{\omega}_j}{1-\check{\omega}_j}\right|\right)\right\}. \end{aligned}$$

The result thus follows from assumption (C6). ■

Proof of Theorem 1

Let $t \in \mathbb{R}$. First, the total error of $\hat{F}_{I,y}(t)$ may be written as

$$\hat{F}_{I,y}(t) - F_{N,y}(t) \quad = \quad \left\{\hat{F}_{N,y}(t) - F_{N,y}(t)\right\} + \left\{\hat{F}_{I,y}(t) - \hat{F}_{N,y}(t)\right\}.$$

The first term on the right-hand side of the previous expression is the sampling error, whereas the second term is the nonresponse error. Under the assumption C1a or C1b, and under the assumptions C2 and C3, we easily prove that

$$E\left|\hat{F}_{N,y}(t) - F_{N,y}(t)\right| \quad \underset{\nu\to\infty}{\longrightarrow} \quad 0, \tag{28}$$

see for example [**?**] and [**?**]. It is thus sufficient to prove that

$$E\left|\hat{F}_{I,y}(t) - \hat{F}_{N,y}(t)\right| \quad \underset{\nu\to\infty}{\longrightarrow} \quad 0. \tag{29}$$

The nonresponse error term may be written as

$$\begin{aligned} \hat{F}_{I,y}(t) - \hat{F}_{N,y}(t) &= \sum_{i\in s} \tilde{d}_i(1-r_i)\left\{1(y_i^* \le t) - 1(y_i \le t)\right\} \qquad (30)\\ &= T_1 + T_2, \end{aligned}$$

where

$$T_2 \quad = \quad \sum_{i\in s} \tilde{d}_i(1-r_i)\left\{1(y_i^{**} \le t) - 1(y_i \le t)\right\}. \tag{31}$$

and T_1 is given in (27). From Lemma 1, it is sufficient to prove that $E|T_2| \underset{\nu\to\infty}{\longrightarrow} 0$. This is equivalent to prove that $E|\tilde{T}_2| \underset{\nu\to\infty}{\longrightarrow} 0$ where

$$\tilde{T}_2 \quad = \quad N^{-1}\sum_{i\in s} d_i(1-r_i)\left\{1(y_i^{**} \le t) - 1(y_i \le t)\right\}. \tag{32}$$

We proceed by showing that

$$E_{pqI}\left(\tilde{T}_2\right) \underset{\nu\to\infty}{\longrightarrow} 0, \tag{33}$$

$$V_{pqI}\left(\tilde{T}_2\right) \underset{\nu\to\infty}{\longrightarrow} 0. \tag{34}$$

We begin with (33). We have

$$\begin{aligned} E_I\left(\tilde{T}_2\right) &= N^{-1}\sum_{g=1}^{G}\sum_{i\in s_g} d_i(1-r_i)\sum_{j\in s_g}\breve{\omega}_j r_j\left\{1(y_j\le t)-1(y_i\le t)\right\} \\ &= U_1+U_2, \end{aligned} \tag{35}$$

where

$$\begin{aligned} U_1 &= N^{-1}\sum_{g=1}^{G}\sum_{i\in s_g} d_i(1-r_i)\sum_{j\in s_g}\breve{\omega}_j r_j\left\{1(y_j\le t)-1(y_i\le t)\right\}, \\ U_2 &= N^{-2}\sum_{g=1}^{G} X_g\sum_{i\in s_g} d_i(1-r_i)\sum_{j\in s_g} d_j\frac{1-p_j}{p_j} r_j\left\{1(y_j\le t)-1(y_i\le t)\right\}, \end{aligned}$$

with

$$\begin{aligned} \breve{\omega}_j &= \frac{d_j\ \frac{1-p_j}{p_j}}{\sum_{l\in s_g} d_l\ (1-p_l)} \text{ for any } j\in s_g, \\ X_g &= \left(\frac{N}{\sum_{k\in s_g} d_k\ \frac{1-p_k}{p_k} r_k}-\frac{N}{\sum_{k\in s_g} d_k(1-p_k)}\right). \end{aligned}$$

We have

$$E_q(U_1) = 0 \tag{36}$$

and

$$\begin{aligned} |U_2| &\leq N^{-2} \sum_{g=1}^{G} |X_g| \sum_{i \in s_g} d_i \sum_{j \in s_g} d_j \frac{1-\kappa}{\kappa} \\ &\leq \frac{1-\kappa}{\kappa} \left(\sum_{g=1}^{G} |X_g| \right) \times \left(N^{-1} \sum_{i \in s} d_i \right)^2 . \end{aligned} \tag{37}$$

From (C3) and (C5), equation (37) leads to

$$E_{pq}(|U_2|) \underset{\nu \to \infty}{\longrightarrow} 0. \tag{38}$$

From (36) and (38), we obtain (33).

Now, we consider (34). We have $V_{pqI}(\tilde{T}_2) = E_{pq} V_I(\tilde{T}_2) + V_{pq} E_I(\tilde{T}_2)$. Also,

$$\begin{aligned} V_I(\tilde{T}_2) &= N^{-2} \sum_{g=1}^{G} \sum_{i \in s_g} d_i^2 (1-r_i) \sum_{j \in s_g} \breve{\omega}_j r_j \left\{ 1(y_j \leq t) - \sum_{k \in s_g} \breve{\omega}_k r_k 1(y_k \leq t) \right\}^2 \\ &\leq N^{-2} \sum_{g=1}^{G} \sum_{i \in s_g} d_i^2 (1-r_i) \\ &\leq N^{-2} \sum_{i \in s} d_i^2 . \end{aligned}$$

The assumptions $C2$ and $C3$ imply that $V_I(\tilde{T}_2) = O(n^{-1})$, so that

$$E_{pq} V_I(\tilde{T}_2) \underset{\nu \to \infty}{\longrightarrow} 0. \tag{39}$$

We now consider the term

$$\begin{aligned} V_{pq} E_I(\tilde{T}_2) &= V_{pq}(U_1 + U_2) \\ &= V_{pq}(U_1) + V_{pq}(U_2) + Cov_{pq}(U_1, U_2). \end{aligned}$$

Since $E_q(U_1) = 0$, we have $V_{pq}(U_1) = E_p V_q(U_1)$, and after some algebra we have $V_q(U_1) = O(n^{-1})$, so that $V_{pq}(U_1) \underset{\nu \to \infty}{\longrightarrow} 0$. On the other hand, $V_{pq}(U_2) \leq$

$E_{pq}(U_2^2)$, and using equation (37), we have from assumptions (C3) and (C5) that $E_{pq}(U_2^2) \underset{\nu\to\infty}{\longrightarrow} 0$. Hence $V_{pq}(U_2) \underset{\nu\to\infty}{\longrightarrow} 0$. Using the Cauchy-Schwarz inequality, we obtain $Cov_{pq}(U_1, U_2) \underset{\nu\to\infty}{\longrightarrow} 0$. Consequently,

$$V_{pq}E_I(\tilde{T}_2) \quad \to \quad 0. \tag{40}$$

From (39) and (40), we obtain (34). This completes the proof.

Proof of Theorem 2

Let $\epsilon > 0$ and $\eta > 0$. According to Condition C8, let ν_0 and $t_1, \ldots, t_M$ such that (10) is satisfied for all $N \geq \nu_0$. Let $t \in \mathbb{R}$ and i such that $t_{i-1} \leq t < t_i$ and $N \geq \nu_0$. By monotonicity of $\hat{F}_{I,y}$, we have

$$\hat{F}_{I,y}(t) \leq \hat{F}_{I,y}(t_i-)$$

and by (10),

$$F_{N,y}(t) \geq F_{N,y}(t_i-) - \epsilon.$$

Hence,

$$\hat{F}_{I,y}(t) - F_{N,y}(t) \leq \hat{F}_{I,y}(t_i-) - F_{N,y}(t_i-) + \epsilon.$$

Similarly,

$$\hat{F}_{I,y}(t) - F_{N,y}(t) \geq \hat{F}_{I,y}(t_{i-1}) - F_{N,y}(t_{i-1}) - \epsilon.$$

Taking the supremum over all t yields

$$\sup_{t\in\mathbb{R}} \left| \hat{F}_{I,y}(t) - F_{N,y}(t) \right| \leq \epsilon + X_{N,\epsilon},$$

where we have defined

$$X_{N,\epsilon} = \max_{i=1,\ldots,M} \max \left\{ \hat{F}_{I,y}(t_{i-1}) - F_{N,y}(t_{i-1}), \hat{F}_{I,y}(t_i-) - F_{N,y}(t_i-) \right\}.$$

Because $\hat{F}_{I,y} - F_{N,y}$ converges to 0 in probability as $N \to \infty$, there exists ν_1 such that for any $N \geq \nu_1$, $Pr(|X_{N,\epsilon}| > \epsilon) \leq \eta$. Let now $N \geq \max\{\nu_0, \nu_1\}$. For such N,

$$Pr\left(\sup_{t\in\mathbb{R}} \left|\hat{F}_{I,y}(t) - F_{N,y}(t)\right| > 2\epsilon\right) < \eta.$$

This concludes the proof.

Proof of the double robustness of the Dunstan-Chambers type estimator

In view of Theorem 1, it is sufficient to prove that

$$E\left|\hat{F}_{I,y}(t) - \tilde{F}_{I,y}(t)\right| \underset{\nu\to\infty}{\longrightarrow} 0 \tag{41}$$

both under the IM approach and the NM approach. The result will follow from

$$V\{\hat{F}_{I,y}(t) - \tilde{F}_{I,y}(t)\} \to 0, \tag{42}$$

and since $E_I(\hat{F}_{I,y}(t) - \tilde{F}_{I,y}(t)) = 0$, it is sufficient to prove that

$$E_p E_q V_I\{\hat{F}_{I,y}(t)\} \to 0. \tag{43}$$

We have

$$\begin{aligned} V_I\{\hat{F}_{I,y}(t)\} &= \sum_{g=1}^{G}\sum_{i\in s_g} \tilde{d}_i^2(1-r_i) \sum_{j\in s_g} \tilde{\omega}_j r_j \left\{1(y_j \leq t) - \sum_{k\in s_g} \tilde{\omega}_k r_k 1(y_k \leq t)\right\}^2 \\ &\leq \sum_{g=1}^{G}\sum_{i\in s_g} \tilde{d}_i^2(1-r_i) \\ &\leq \sum_{i\in s} \tilde{d}_i^2. \end{aligned}$$

From Assumption (C3), $V_I\{\hat{F}_{I,y}(t)\}$ is $O(n^{-1})$, which completes the proof.

References

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

Breidt, F.J. and Opsomer, J.D. (2000). Local polynomial regresssion estimators in survey sampling. *Ann. Statist.* **28** 1026–1053.

Cao, W., Tsiatis, A.A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika.* **96** 723–734.

Cardot, H., Chaouch, M., Goga, C; and Labruère, C. (2010). Properties of design-based functional principal components analysis. *J. Statist. Plann. Inference.* **140** 75–91.

Chambers, R. L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika.* **73** 597–604.

Chauvet, G. (2014). A note on the consistency of the Narain-Horvitz-Thompson estimator. *ArXiv e-print.*

Chauvet, G., Deville, J.-C. and Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika.* **98** 459–471.

Chauvet, G. and Haziza, D. (2012). Fully efficient estimation of coefficients of correlation in the presence of imputed data. *Canad. J. Statist.*, **40**, 124–149.

Cheng, P.E. and Chu, C.K. (1996). Kernel estimation of distribution functions and quantiles with missing data. *Statist. Sinica.* **6** 63–78.

Haziza, D. and Beaumont, J-F. (2007). On the construction of imputation classes in surveys. *Internat. Statist. Rev.* **75** 25–43.

Haziza, D., Nambeu, C-O. and Chauvet, G. (2014). Doubly robust imputation procedures for populations containing a large amount of zeroes in surveys. *Canad. J. Statist.*, **42**, 650–669.

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

Haziza, D. and Rao, J.N.K. (2010). Variance estimation in two-stage cluster sampling under imputation for missing data. *J. Stat. Theory Pract.* **4** 827–844.

Kang, J.D.Y. and Schafer, J.L. (2008). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539.

Kim, J.K. and Fuller, W.A. (2004). Fractional hot-deck imputation. *Biometrika.* **91** 559–578.

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

Kim, J.K. and Park, H. (2006). Imputation using response probability. *Canad. J. Statist.* **34** 171–182.

Kott, P.S. (1994). A note on handling nonresponse in sample surveys. *J. Amer. Statist. Assoc.* **89** 693–696.

Lago, L.P. and Clark, R.G. (2015). Imputation of Household Survey Data Using Linear Mixed Models. *To appear in Aust. N. Z. J. Stat.*

Liu, X., Liu, P. and Zhou, Y. (2011). Distribution estimation with auxiliary information for missing data. *J. Statist. Plann. Inference.* **141** 711–724.

Mashreghi, Z., Haziza, D. and Léger, C. (2015). Pseudo-population bootstrap methods for imputed survey data. *Submitted.*

Mulry, M.H., Oliver, B.E. and Kaputa, S.J. (2014). Detecting and Treating Verified Influential Values in a Monthly Retail Trade Survey. *J. Official Statist.* **30** 721–747.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Internat. Statist. Rev.* **61** 317–337.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficient when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866.

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

Rubin, D.B. and van der Laan, M.J. (2008). Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *Int. J. Biostat.* **4**.

Scharfstein, D.O., Rotnitzky, A. and Robins, J.M. (1999). Adjusting for non-ignorable drop-out using semiparametric nonresponse models (with discussion and rejoinder). *J. Amer. Statist. Assoc.* **94** 1096–1146.

Skinner, C.J. and D'Arrigo, J. (2011). Inverse probability weighting for clustered nonresponse. *Biometrika.* **98** 953–966.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *J. Amer. Statist. Assoc.* **101** 1619–1637.

Tillé, Y. (1995). *Sampling algorithms*, Springer, New York.

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). *Finite population sampling and inference: a prediction approach*, Wiley, New York.

Zhao, P.-Y., Tang, M.-L. and Tang, N.-S. (2013). Robust estimation of distribution functions and quantiles with non-ignorable missing data. *Canad. J. Statist.* **41** 575–595.