CS-BIGS 1(2): 102-118
© 2007 CS-BIGS

http://www.bentley.edu/csbigs/vol1-2/haziza.pdf

Handling Item Non Response in Surveys

David Haziza
Université de Montréal, Québec, Canada

Gordon Kuromi
Statistics Canada, Ottawa, Canada

Item nonresponse occurs inevitably in almost all surveys conducted by statistical agencies because some sampled units refuse to respond to sensitive items or may not know the answer to some items. Item nonresponse is usually treated by using single imputation, which consists of creating a single value to replace a missing value. The main effect of item nonresponse is that when the respondents and the nonrespondents are different with respect to the survey variables, then nonresponse bias is introduced. Also, since the observed sample size is smaller than the sample size initially planned, nonresponse has the effect of leading to estimators with larger variance than that which would have been obtained if complete response had been achieved. This increase in variance is called the nonresponse variance. Finally, imputation has the effect of distorting the relationships between variables. In this paper, using data observed in the context of the Canadian Community Health Survey, we propose to investigate empirically the properties of estimators of population means, domain means and finite population coefficients of correlation in terms of bias and mean square error when regression imputation has been used to fill in the missing values. The exposition is easily accessible to readers with some background in linear regression.

Keywords: *Canadian Community Health Survey; Coefficient of correlation; Domain mean; Nonresponse bias; Nonresponse variance; Imputation model; Regression imputation; Random hot deck imputation.*

1. Introduction

Despite the best efforts made by survey staff to maximize response, it is almost certain that some degree of nonresponse will occur in large scale surveys. Essentially, survey statisticians distinguish between two types of nonresponse: total or unit nonresponse (when no information is collected on a sampled unit) and partial or item nonresponse (when the absence of information is limited to some variables only). Unit nonresponse occurs, for example, when the sampled unit is not-at-home or refuses to participate in the survey. Item nonresponse may occur if the sampled unit refuses to respond to sensitive items, or if it does not know the answer to some items. Also, missing values occur when a sampled unit fails at least one edit rule. Generally, weighting adjustment methods are used to compensate for unit nonresponse whereas imputation is used to compensate for item nonresponse. The main idea behind a weighting adjustment is to increase the sampling weights of the respondents in order to compensate for the nonrespondents, while imputation is a process where one or more 'plausible values' are produced to replace a missing value. The main effects of (unit or item) nonresponse include: (i) bias of point estimators; (ii) increase in the variance of point estimators and (iii) bias of complete data variance estimators.

In this paper, we focus on single imputation, which consists in creating a single imputed value to replace a missing value, resulting in a single data file. The fact that imputation leads to a complete data file is seen by many data users and analysts as a desirable feature.

If respondents and nonrespondents are different with respect to the survey variables then nonresponse bias is introduced. Also, since the observed sample size is smaller than the sample size initially planned, nonresponse usually has the effect of leading to estimators with larger variance than the variance of those that would have been obtained if complete response had been achieved. This increase in variance is called the nonresponse variance. The role of imputation methods is thus to reduce the nonresponse bias and to control the nonresponse variance as much as possible. The key to achieving these goals is to use auxiliary variables available for both respondents and nonrespondents. Consequently, imputation is essentially a modeling exercise. The quality of the estimates will thus depend on the availability (at the imputation stage) of good auxiliary information and its judicious use in the imputation strategy.

It is nonetheless important to note that imputation carries certain risks. The most significant include: (1) Even though imputation leads to the creation of a complete data file, inferences are valid only if the underlying assumptions about the response mechanism and/or the imputation model are satisfied. (2) Some imputation methods tend to distort the distribution of the variables of interest (i.e., the variables being imputed). (3) Treating the imputed values as if they were observed may lead to a substantial underestimation of the variance of the estimator, especially if the item nonresponse rate is appreciable. (4) Marginal imputation for each item separately has the effect of distorting relationships between variables.

The outline of the paper is as follows: in section 2, we outline definitions and assumptions used in this paper, and define the concept of an imputed estimator. We also describe two families of imputation methods frequently used in practice: deterministic regression imputation and random regression imputation. Section 3 introduces the important concepts of nonresponse bias and nonresponse variance. The concept of ignorability of the nonresponse mechanism is also discussed. In section 4, we describe the data set used in this paper, and perform simulation studies to investigate the magnitude of the nonresponse bias and nonresponse variance under several scenarios. The problem of estimating totals (or means) for subgroups called domains is treated in section 5. We also perform simulation studies to investigate the bias of the resulting estimators. In section 6, we consider the problem of estimating the finite population coefficient of correlation between two variables after they have been marginally imputed. Finally, we conclude in section 7.

2. Framework and Assumptions

Consider a finite population U of N individuals. Our goal is to estimate the population mean of a variable of interest y (for example, the weight of an individual), $\bar{Y} = \frac{1}{N}\sum_{i \in U} y_i$. We select a random sample, s, of size n, according to a given sampling design $p(s)$. If we had complete response to the variable y, we could, for example, use the well known Horvitz-Thompson estimator (e.g., Lohr, 1999, page 96) given by

$$\bar{y}_{HT} = \frac{1}{N}\sum_{i \in s} w_i y_i, \qquad (2.1)$$

where $w_i = 1/\pi_i$ denotes the sampling weight attached to unit i and π_i denotes its inclusion probability in the sample. The estimator $\bar{y}_{HT}$ is unbiased for $\bar{Y}$ with respect to the sampling design and we write $E_p(\bar{y}_{HT}) = \bar{Y}$, where $E_p(.)$ denotes the expectation with respect to the sampling design. In the presence of nonresponse to item y, it is not possible to compute the estimator $\bar{y}_{HT}$ since some y-values are missing. We define an imputed estimator of $\bar{Y}$ given by

$$\bar{y}_I = \frac{1}{N}\left[\sum_{i \in s} w_i r_i y_i + \sum_{i \in s} w_i (1 - r_i) y_i^*\right], \qquad (2.2)$$

where r_i is a response indicator attached to unit i such that $r_i = 1$ if unit i has responded to item i and $r_i = 0$ otherwise, and y_i^* denotes the imputed value used to replace the missing value y_i. Note that the imputed estimator (2.2) is simply the weighted mean of the observed and the imputed values. Also, note that the imputed value y_i^* depends on the imputation method used.

In this paper, we consider deterministic and random regression imputation. We assume that a vector of q auxiliary variables, $\mathbf{z}$, is available for each sampled unit.

Regression imputation is motivated by the following linear regression model:

$$m\colon y_i = \mathbf{z}_i'\boldsymbol{\beta} + \varepsilon_i,$$

$$E_m\left(\varepsilon_i\right) = 0, \quad E_m\left(\varepsilon_i\varepsilon_j\right) = 0 \text{ if } i \neq j, \quad E_m\left(\varepsilon_i^2\right) = \sigma^2, \tag{2.3}$$

where $E_m\left(.\right)$ denotes the expectation with respect to the model (2.3). The model (2.3) is called the imputation model.

Deterministic regression imputation consists of replacing the missing value y_i by the predicted value, $\hat{y}_i$, obtained by fitting the model (2.3) using the respondents *y*-values. That is,

$$y_i^* = \hat{y}_i = \mathbf{z}_i'\hat{\mathbf{B}}_\mathbf{r}, \tag{2.4}$$

where $\hat{\mathbf{B}}_\mathbf{r} = \left(\sum_{i\in s} w_i r_i \mathbf{z}_i \mathbf{z}_i'\right)^{-1}\left(\sum_{i\in s} w_i r_i \mathbf{z}_i y_i\right)$ is the weighted least square estimator of $\boldsymbol{\beta}$. A special case of (2.4) is mean imputation which is obtained by setting $\mathbf{z}_i = 1$. In this case, the estimated regression coefficient $\hat{\mathbf{B}}_\mathbf{r}$ reduces to the weighted mean of the respondents,

$$\bar{y}_r = \sum_{i\in s} w_i r_i y_i \Big/ \sum_{i\in s} w_i r_i .$$

Random regression imputation can be viewed as deterministic regression imputation with an added random component. Let s_r denote the set of respondents to item *y*. The imputed value used for missing y_i is given by

$$y_i^* = \hat{y}_i + e_i^*, \tag{2.5}$$

where $\hat{y}_i$ is given by (2.4), $e_i^* = e_j,\ j \in s_r$ such that $P\left(e_i^* = e_j\right) = w_j \Big/ \sum_{k\in s} w_k r_k$ and $e_j = y_j - \mathbf{z}_j'\hat{\mathbf{B}}_r$. A special case of random regression imputation is random hot deck imputation which is obtained by setting $\mathbf{z}_i = 1$ in (2.5). It can be viewed as mean imputation with an added random residual. In other words, random hot deck imputation consists of selecting at random (with probability proportional to the sampling weight) a respondent *y*-value to replace a missing value.

Deterministic regression imputation tends to distort the distribution of the variables of interest (i.e., the variables being imputed). The magnitude of the distortion depends on the response rate as well as the adequacy of the model. If the response rate is high and/or the model explains the variable being imputed well, we can expect the distortion to be small to moderate. Random regression imputation tends to preserve the distribution of the variable being imputed but it suffers from an additional component of variance due to the use of a random imputation mechanism.

3. Nonresponse Bias and Nonresponse Variance

To study the properties (for example, bias, variance and mean square error) of the imputed estimator (2.2), we use the standard decomposition of the total error of $\bar{y}_I$ as a starting point:

$$\bar{y}_I - \bar{Y} = \left(\bar{y}_{HT} - \bar{Y}\right) + \left(\bar{y}_I - \bar{y}_{HT}\right) \tag{3.1}$$

The first term $\bar{y}_{HT} - \bar{Y}$ on the right-hand side of (3.1) is called the sampling error of $\bar{y}_I$ whereas the second term $\bar{y}_I - \bar{y}_{HT}$ is called the nonresponse error of $\bar{y}_I$. The concepts of nonresponse bias and nonresponse variance are defined in sections 3.1 and 3.2, respectively. But first, we define the concept of nonresponse mechanism.

Let $p_i = P\left(r_i = 1\right)$ be the response probability to item *y* for individual *i*. We assume that the individuals respond independently of one another; that is, $p_{ij} = P\left(r_i = 1, r_j = 1\right) = p_i p_j$. The unknown distribution of the response indicators, $p\left(r_i \mid s\right)$, is called the nonresponse mechanism. Since it is unknown, we must make some assumptions about the nonresponse mechanism (see section 3.1).

3.1 Nonresponse bias

The bias of the imputed estimator is defined as

$$\begin{aligned}\text{Bias}\left(\bar{y}_I\right) &= E\left(\bar{y}_I - \bar{Y}\right)\\ &= E_p E_q\left(\bar{y}_I - \bar{Y} \mid s\right)\\ &= E_p\left(\bar{y}_{HT} - \bar{Y}\right) + E_p E_q\left(\bar{y}_I - \bar{y}_{HT} \mid s\right)\\ &= E_p\left(B_q\right),\end{aligned}$$

where $B_q = E_q\left(\bar{y}_I - \bar{y}_{HT} \mid s\right)$ denotes the conditional nonresponse bias and $E_q\left(.\right)$ denotes the expectation with respect to the nonresponse mechanism. Hence, the imputed estimator $\bar{y}_I$ is unbiased for $\bar{Y}$ if $B_q = 0$ for any sample *s*. Therefore, the nonresponse bias is the average difference between the imputed estimator obtained after imputation and the estimator we would have obtained had no nonresponse been observed. The question that comes to mind is: when is the nonresponse bias, B_q, (approximately) equal to zero?

The nonresponse bias will be negligible if the vector of auxiliary variables $\mathbf{z}$ is correctly specified and the nonresponse mechanism is ignorable with respect to the selected imputation model, which occurs when the probability of response, p_i, is independent of the error term, ε_i, in the imputation model. In other words, it is ignorable if, after accounting for $\mathbf{z}$ in the imputation procedure, the response probability does not depend on the error term. Otherwise, the nonresponse mechanism is said to be nonignorable. A special case of an ignorable nonresponse mechanism occurs when all the individuals have the same response probability; that is $p_i = p$. In this case, the nonresponse mechanism is said to be uniform. When the nonresponse mechanism is ignorable, the data are said to be *Missing At Random* (MAR). When it is nonignorable, the data are said to be *Not Missing At Random* (NMAR).

Consider the case of a scalar *z* and suppose that the probability of response depends on the variable *z*. If the variable of interest *y* is related to *z* (so the error term of the imputation model depends on *z*), then the nonresponse mechanism is ignorable if *z* is used in the imputation procedure and we can expect the nonresponse bias to be negligible; otherwise the nonresponse mechanism is nonignorable and the resulting estimators are potentially considerably biased. If the variable *z* is not related to *y*, then there is no need to include *z* in the imputation model in order to reduce the nonresponse bias. If the probability of response depends directly on the variable of interest, then the nonresponse mechanism is automatically nonignorable. In this case, the main issue is the nonresponse bias that will remain even after accounting for auxiliary variables in the imputation model. However, note that if the imputation model is rich and has good predictive power, we can expect to achieve a good bias reduction. In practice, we do not know if the nonresponse mechanism is ignorable or not but we can expect that, as the imputation model becomes 'richer', the nonresponse bias will typically decrease.

In practice, it is generally not possible to know if there is a presence of nonresponse bias, and if there is, we do not know anything about its magnitude. What we know is that the nonresponse bias tends to be large if the respondents and the nonrespondents have significantly different characteristics and it increases as the response rate decreases. Consequently, it is important to perform a complete modeling exercise which includes model building as well as model validation. Model validation is particularly important because it gives us some confidence that the model at hand is reasonable. It includes the detection of outliers and the examination of plots such as a plot of residuals vs. the predicted values, a plot of residuals vs. the auxiliary variables selected in the model and a plot of residuals vs. variables not selected in the model.

3.2 Nonresponse variance

In practice, we should make every effort to reduce the non-response bias by selecting an appropriate imputation method. Assuming that the imputed estimator $\overline{y}_I$ is unbiased for $\overline{Y}$, its variance can be expressed as

$$\begin{aligned}
V(\overline{y}_I) &= E(\overline{y}_I - \overline{Y})^2 \\
&= E_pE_q(\overline{y}_{HT} - \overline{Y} \mid s)^2 + E_pE_q(\overline{y}_I - \overline{y}_{HT} \mid s)^2 \\
&\qquad + 2E_pE_q\left[(\overline{y}_{HT} - \overline{Y})(\overline{y}_I - \overline{y}_{HT}) \mid s\right] \\
&= E_p(\overline{y}_{HT} - \overline{Y})^2 + E_pE_q(\overline{y}_I - \overline{y}_{HT} \mid s)^2 \\
&= V_{SAM} + V_{NR},
\end{aligned}$$

where $V_{SAM} = V_p(\overline{y}_{HT})$ denotes the sampling variance and $V_{NR} = E_pV_q(\overline{y}_I - \overline{y}_{HT} | s)$ denotes the nonresponse variance. In the case of random regression imputation, there is an extra term in the total variance which represents the variance associated with the imputation mechanism that consists in randomly selecting the residuals. The magnitude of the sampling variance depends on the sampling procedure being used, the selected sample size and the type of population being sampled. The magnitude of the nonresponse variance depends on the response rate as well as the quality of the imputation model used to construct the imputed values. The nonresponse variance tends to increase as the response rate decreases because the number of respondents decreases. Also, for a given response rate, the nonresponse variance tends to be small if the imputation model has a good predictive power. Indeed, consider the extreme situation for which the relationship between the variable of interest and the auxiliary variables is perfect (i.e., all the points lie on the regression line). In this case, the nonresponse variance is zero regardless of the response rate because we are able to determine exactly the missing values.

4. Simulation Study

4.1 Description of the data set

The data set used for the simulation studies is a subset of a sample collected between January 2005 and December 2005 for the Canadian Community Health Survey (CCHS), which is a cross-sectional survey that collects information related to health status, health care utilization and health determinants for the Canadian population. This data represents our 'population' that will be used as a starting point for the simulation studies.

The original data file was a Public Use Microdata File (PUMF) obtained for the CCHS Cycle 3.1 (2005). This file contained more than 1000 variables. We selected a relatively small subset of 16 variables, which are described in Appendix A. Information on the file was collected for more than 100,000 records. Records with missing or incomplete responses were dropped, which reduced the number of records to the neighborhood of 80,000. Then a simple random sample of 10,000 records was taken for the case study population data set.

In this paper, the main variable of interest is HWTEGWTK (self reported weight in kg of an individual). All the other variables on the file will be treated as auxiliary variables. To determine the set of variables related to the variable HWTEGWTK, we performed a regression analysis with the variable HWTEGWTK as the dependent variable and the other variables as independent variables. The selected independent variables are shown in Appendix B. From this point on, we refer to this model as the full model. Note that the 16 age categories for the variable DHHEAGE of the CCHS PUMF were collapsed into 7 categories in the derived variable AGE_GROUP, and therefore this derived variable rather than the original variable from the CCHS PUMF file is described in Appendix A. Also, two interaction variables were defined and added to the population file for the purpose of modeling: (i) the interaction between AGE_GROUP and DHHE_SEX (called AGE_SEX_NUMERIC) and (ii) the interaction between AGE_GROUP and CCCE_071 (called AGE_BP_NUMERIC).

4.2 Implementation

In this section, we perform simulation studies in order to illustrate the concepts on nonresponse bias and nonresponse variance. From the CCHS population, we selected $R = 1000$ random samples according to simple random sampling without replacement with sample sizes $n = 1000$ and $n = 10,000$. Note that the latter case is the census case. In each selected sample, we generated nonresponse according to four distinct nonresponse mechanisms, as follows: first, we assigned a response probability, p_i, to each unit in the sample according to a logistic function; that is, we have

$$p_i = \frac{\exp\{\mathbf{x}_i'\boldsymbol{\gamma}\}}{1+\exp\{\mathbf{x}_i'\boldsymbol{\gamma}\}},$$

where $\mathbf{x}_i$ is a vector of variables and $\boldsymbol{\gamma}$ denotes a vector of parameters. Table 1 presents the vector $\mathbf{x}_i$ for each nonresponse mechanism. Then, for each sampled unit, a response indicator r_i was generated according to a Bernoulli distribution with parameter p_i. The response rates were set to either 0.5 or 0.9. That is, we generated the p_i's so that their mean was equal to either 0.5 or 0.9.

Table 1. Nonresponse Mechanisms Used to Generate Nonresponse

Non-response mechanism	1	2	3	4
$\mathbf{x}_i$	1	1, DHHE_SEX, AGE_GROUP, HWTEGHTM	1, HWTEG-WTK	1, CCCE_011, CCCE_031, INCEGHH

In each simulated sample, we used three imputation methods to compensate for nonresponse to item *y*: mean imputation, deterministic regression imputation and random regression imputation. Both deterministic and random regression imputation were based on the full model described in Appendix B.

Then, in each sample, we calculated the imputed estimator, $\bar{y}_I$, given by (2.2). The Monte Carlo average of an estimator $\hat{\theta}$ is defined by

$$E_{MC}\left(\hat{\theta}\right) = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}^{(r)}, \tag{4.1}$$

where $\hat{\theta}^{(r)}$ denotes the estimator $\hat{\theta}$ in the *r*-th simulated sample, $r = 1,\ldots,R.$ As a measure of bias of $\bar{y}_I$, we used the Monte Carlo percent relative bias given by

$$RB_{MC}\left(\bar{y}_I\right) = 100\times\frac{E_{MC}\left(\bar{y}_I\right)-\bar{Y}}{\bar{Y}}, \tag{4.2}$$

where $E_{MC}\left(\bar{y}_I\right)$ is obtained from (4.1) by replacing $\hat{\theta}$ with $\bar{y}_I$. As a measure of variability of the imputed estimator $\overline{y}_I$, we used the Monte Carlo mean square error given by

$$MSE_{MC}(\overline{y}_I) = E_{MC}\left(\overline{y}_I - \overline{Y}\right)^2,$$

where $E_{MC}\left(\overline{y}_I - \overline{Y}\right)^2$ is obtained from (4.1) by replacing $\hat{\theta}$ with $\left(\overline{y}_I - \overline{Y}\right)^2$.

To investigate the relative magnitude of the variance components (variance due to sampling and due to nonresponse and imputation), we calculated the following Monte Carlo measures: $V_{SAM}^{MC} = E_{MC}\left[E_{MC}\left(\overline{y}_{HT}\right) - \overline{Y}\right]^2$, which is obtained from (4.1) by replacing $\hat{\theta}$ with $\left[E_{MC}\left(\overline{y}_{HT}\right) - \overline{Y}\right]^2$ and $E_{MC}\left(\overline{y}_{HT}\right)$ is obtained from (4.1) by replacing $\hat{\theta}$ with $\overline{y}_{HT}$; $V_{NR}^{MC} = E_{MC}\left[\overline{y}_I - E_{MC}\left(\overline{y}_{HT}\right)\right]^2$, which is obtained from (4.1) by replacing $\hat{\theta}$ with $\left[\overline{y}_I - E_{MC}\left(\overline{y}_{HT}\right)\right]^2$. Finally, to get an idea of the relative increase in variance when random regression imputation is used as opposed to deterministic regression imputation, we computed the following measure:

$$\lambda = \frac{MSE_{MC}^{\text{Random}}(\overline{y}_I)}{MSE_{MC}^{\text{Deterministic}}(\overline{y}_I)}, \qquad (4.3)$$

where $MSE_{MC}^{\text{Random}}(\overline{y}_I)$ denotes the mean square error of the imputed estimator $\overline{y}_I$ when random regression imputation has been used, whereas $MSE_{MC}^{\text{Deterministic}}(\overline{y}_I)$ denotes the mean square error of the imputed estimator $\overline{y}_I$ under deterministic regression imputation.

4.2 Discussion of the results

Note that the nonresponse mechanism 1 corresponds to uniform response under which all the individuals have the same response probability. The nonresponse mechanism 2 depends on auxiliary variables that are strongly related to the variable of interest HWTEGWTK (see Appendix B). Hence, if the variables DHHE_SEX, AGE_GROUP and HWTEGHTM are used in the imputation procedure, the nonresponse mechanism will be ignorable; otherwise, it will be nonignorable and the resulting estimators will be biased. The nonresponse mechanism 3 depends directly on the variable of interest. Therefore, it is automatically nonignorable. Finally, the nonresponse mechanism 4 depends on variables that are related poorly or not at all to HWTEGWTK. These variables were not selected in the full model. Therefore, the nonresponse mechanism is automatically ignorable.

Figure 1 shows the relative error of the Horvitz-Thompson estimator (2.1) that we would have obtained in the complete data case and that of the imputed estimator (2.2) for each replicate. Note that the relative error (in %) of an estimator $\hat{\theta}$ of a parameter θ is defined as $100 \times \left(\frac{\hat{\theta} - \theta}{\theta}\right)$. As expected, the relative error of the Horvitz-Thompson estimator 'centers around' zero, showing that the Horvitz-Thompson estimator is unbiased for the population mean. Also, Figure 1 shows that the imputed estimator under mean imputation and nonresponse mechanism 2 (with $n/N = 0.1$ and $p = 0.5$) is clearly biased since its relative error centers around 15%. This result can be easily explained by the fact that the probability of response depends on some auxiliary variables that are also related to the variable of interest HWTEGWTK but mean imputation fails to account for these variables. Figure 2 shows that both the Horvitz-Thompson estimator (2.1) and the imputed estimator (2.2) are unbiased under mean imputation and the nonresponse mechanism 4 (with $n/N = 0.1$ and $p = 0.5$). This result is not surprising since, in the case of the nonresponse mechanism 4, the probability of response depends on variables that are not related to the variable of interest HWTEGWTK, so there is no need to include them in the imputation procedure.

Figure 3 shows that the asymptotic distribution of the imputed estimator is normal under the nonresponse

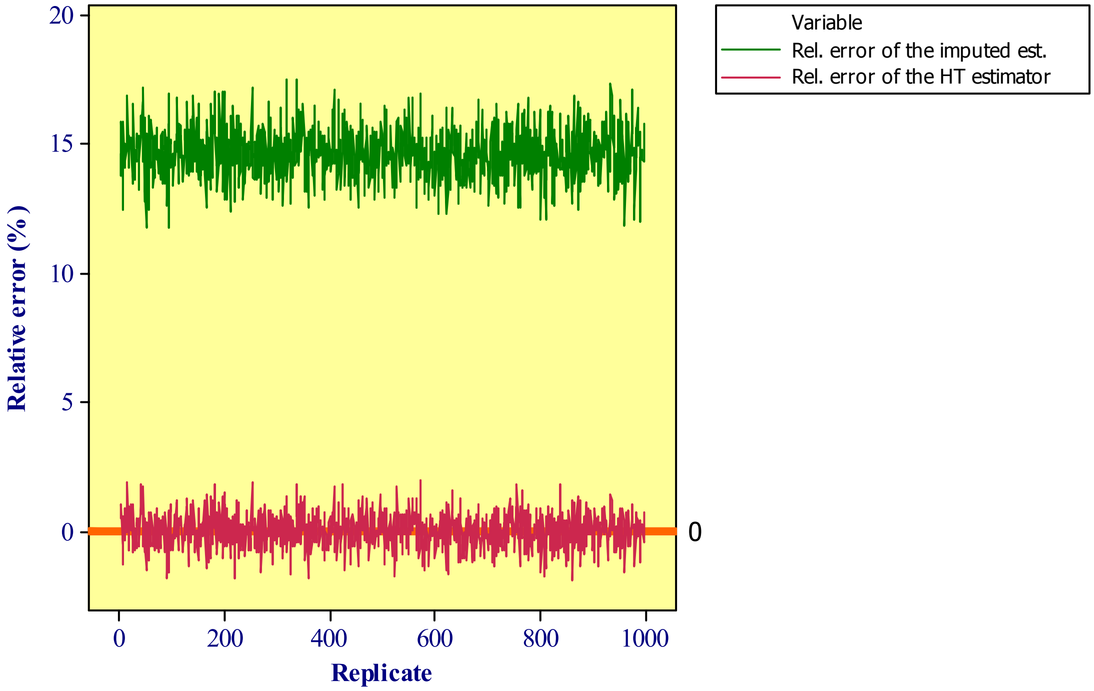

Figure 1: Relative Error of the Imputed estimator vs. the Horvitz-Thompson (complete data) estimator for mean imputation with $n/N = 0.1$, $p = 0.5$ and nonresponse mechanism 2

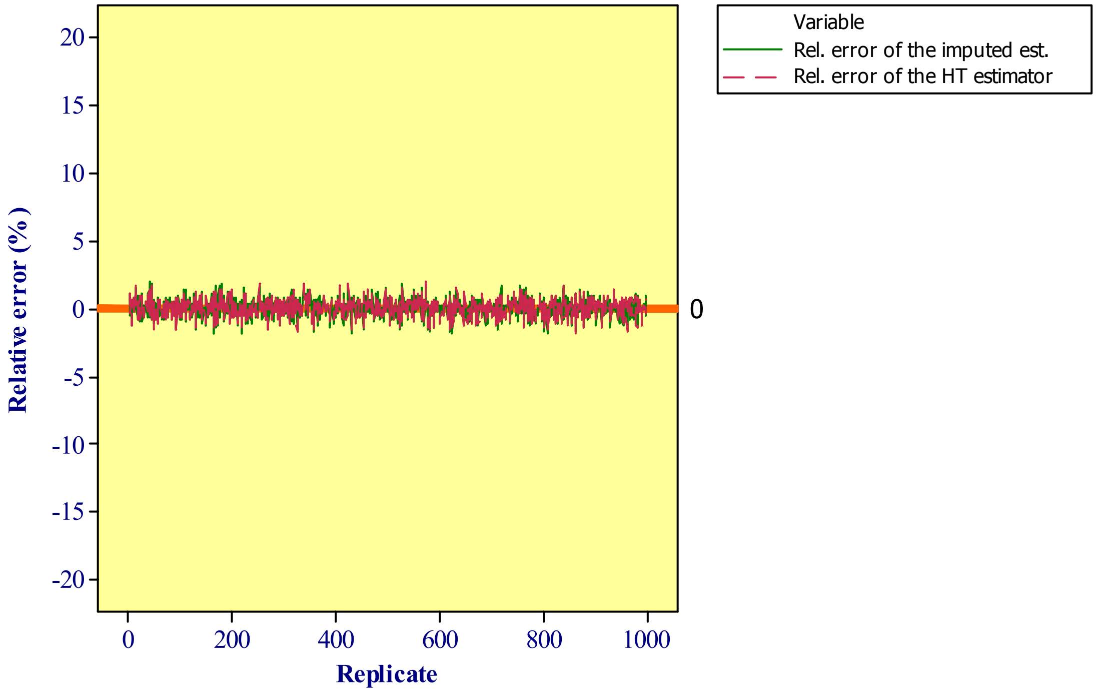

Figure 2: Relative Error of the Imputed estimator vs. the Horvitz-Thompson (complete data) estimator for mean imputation with $n/N = 0.1$, $p = 0.5$ and nonresponse mechanism 4

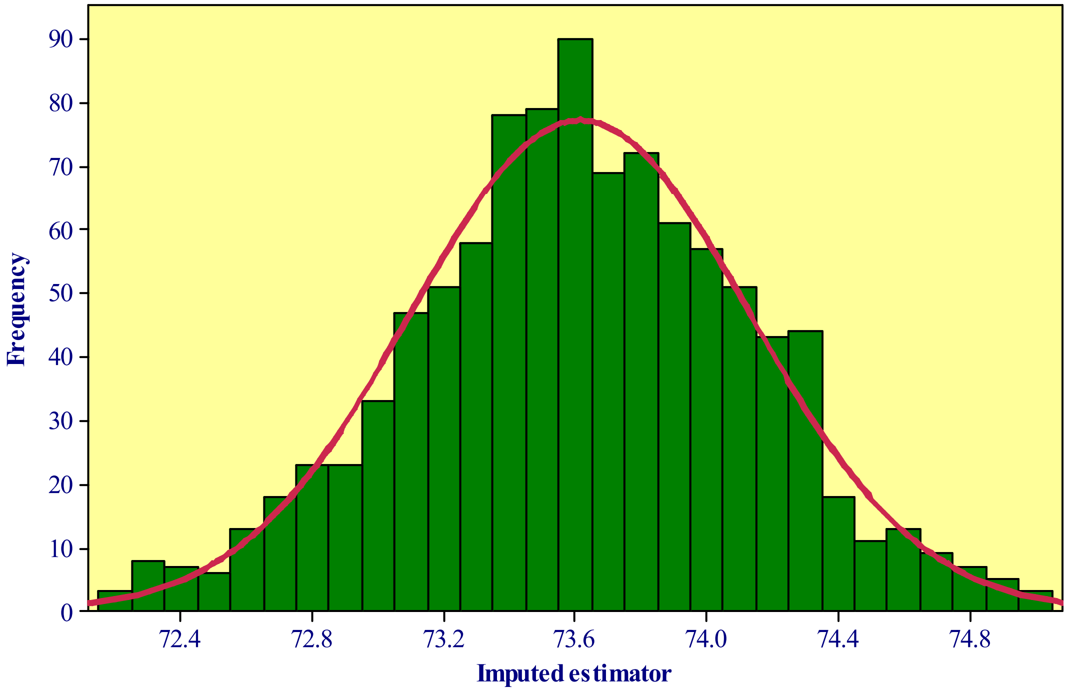

Figure 3: Distribution of the imputed estimator

mechanism 1 with $n/N = 0.1$ and $p = 0.5$. In the context of arbitrary sampling design and imputation, the central limit theorem is not easy to prove but holds for many cases encountered in practice. A similar shape is obtained under all the scenarios considered in the simulation study.

We now discuss the results presented in Tables 2 and 3. Note that we only presented the results corresponding to mean and regression imputation because random regression imputation is, on average, asymptotically equivalent to deterministic imputation. This is explained by the fact that the average of the added residual with respect to the imputation procedure is equal to zero. It is clear that under the nonresponse mechanism 1 (uniform nonresponse) the imputed estimator is unbiased in all scenarios. This is not surprising because, under uniform nonresponse, the set of respondents can be viewed essentially as a simple random sample without replacement. Therefore, the set of respondents (and therefore the set of nonrespondents) will include all kind of individuals: the people with a small weight, the people with a medium weight and the people with a large weight. In other words the distribution of the variable HWTEGWTK in the set of respondents is identical to that of HWTEGWTK in the set of nonrespondents. As a result, we can expect the mean of respondents for the variable HWTEGWTK to be close to its population mean. Turning to the MSE of the imputed estimator (which corresponds to the variance of the imputed estimator) under the nonresponse mechanism 1, we see that regression imputation leads to a smaller MSE than mean imputation under all the scenarios. This is due to the fact that the model underlying the regression imputation procedure (i.e., the full model) has better predictive power than the model underlying mean imputation (which includes only the intercept). In the case of a census (i.e., $n/N = 1$), the MSE of the imputed estimator virtually reduces to the nonresponse variance since the bias is negligible and the sampling variance is identically equal to zero. The results show that a good imputation model can reduce both the nonresponse bias as well as the nonresponse variance. Finally, note that the MSE of the estimators decreases as the sampling fraction increases, as expected.

For the nonresponse mechanism 2, the imputed estimator is biased under mean imputation, which can be explained by the fact that the response probability depends on the variables DHHE_SEX, AGE_GROUP and HWTEGHTM, and that these variables explain also the variable of interest HWTEGWTK. However, mean imputation fails to account for these three variables, which in turns leads to biased estimators. Tables 4-6 show the response rates (across the R samples) for the three variables. It is clear that the response rates to the variable

HWTEGWTK for men and women are very different (71% for men and 32% for women). The response rates by age group are also very different from one category to another (see Table 4). The response rate for the first age group (individuals whose age is between 12 and 14) is especially low (around 6%). Finally, the response rates for the four quartiles of the continuous variable HWTEGHTM are shown in Table 6. Once again, it is clear that as the height increases, so does the response probability to the variable HWTEGWTK. These results are not surprising given the way we defined the nonresponse mechanism 2. In light of these results, it is crucial to include these variables in the imputation model, which is satisfied when we perform deterministic regression imputation based on the full model. Finally, note that the relative bias decreases as the response rate increases for a given sampling fraction. On the other hand, it is clear that relative bias does not change as the sampling fraction increases for a given response rate. This result is important because it shows that the nonresponse bias is not a function of the sample size but rather a function of the response rate.

For the nonresponse mechanism 3 (which is nonignorable), we note that the imputed estimator is biased in all scenarios, as expected. However, we see that, although we cannot eliminate the nonresponse bias completely, we can reduce it significantly by using an imputation model with good predictive power. For example, with *n*/*N* = 0.1 and *p* = 0.5, we obtained a relative bias equal to 14.7% under mean imputation, whereas it is equal to 10.4% under deterministic regression imputation.

For the nonresponse mechanism 4, it is clear from Tables 2 and 4 that the imputed estimator has a negligible bias in all the scenarios. This is due to the fact that, under the nonresponse mechanism 4, the probability of response to HWTEGWTK depends on variables that are not related to this variable (CCCE_011, CCCE_031, INCEGHH). As a result, the response mechanism is ignorable since the response probability is independent of the error in the imputation model.

Table 4: Response rate by age group with $n/N = 0.1$, $p = 0.5$ and the nonresponse mechanism 2

Age group	1	2	3	4	5	6	7
Response rate	0.06	0.28	0.47	0.57	0.58	0.55	0.37

Table 5: Response rate by sex with $n/N = 0.1$, $p = 0.5$ and the nonresponse mechanism 2

Sex	1	2
Response rate	0.71	0.32

Table 6: Response rate by height quartile with $n/N = 0.1$, $p = 0.5$ and the nonresponse mechanism 2

Height quartile	1	2	3	4
Response rate	0.19	0.39	0.62	0.80

Table 2: Monte Carlo percent relative bias and mean square error with *n*/*N* = 0.1

	$p = 0.5$				$p = 0.9$			
	Mean imputation		Deterministic regression imputation		Mean imputation		Deterministic regression imputation	
	Relative bias (%)	MSE	Relative bias (%)	MSE	Relative bias (%)	MSE	Relative bias (%)	MSE
Nonresponse mechanism 1	0.0	0.6	0.0	0.4	0.0	0.4	0.02	0.3
Nonresponse mechanism 2	7.5	30.72	-0.5	0.8	1.6	1.7	-0.0	0.3
Nonresponse mechanism 3	14.7	117.0	10.4	59.0	2.9	4.7	1.5	1.4
Nonresponse mechanism 4	0.2	0.56	0.0	0.4	0.0	0.3	0.0	0.3

Table 3: Monte Carlo percent relative bias and mean square error with *n*/*N* = 1

	$p = 0.5$				$p = 0.9$			
	Mean imputation		Deterministic regression imputation		Mean imputation		Deterministic regression imputation	
	Relative bias (%)	MSE	Relative bias (%)	MSE	Relative bias (%)	MSE	Relative bias (%)	MSE
Nonresponse mechanism 1	0.0	0.03	0.0	0.02	0.0	0.003	0.0	0.0021
Nonresponse mechanism 2	7.5	30.2	-0.5	0.2	1.6	1.71	-0.1	0.0048
Nonresponse mechanism 3	14.7	117	10.4	58.7	2.9	4.41	1.4	1.15
Nonresponse mechanism 4	0.2	0.044	0.0	0.018	0.0	0.0037	0.0	0.0021

Table 7: Contribution (in %) of V_{SAM} and V_{NR} to the total variance under mean and deterministic regression imputation with $n/N = 0.1$ and nonresponse mechanism

$p=0.5$				$p=0.9$			
Mean imputation		Deterministic regression imputation		Mean imputation		Deterministic regression imputation	
V_{SAM}^{MC}	V_{NR}^{MC}	V_{SAM}^{MC}	V_{NR}^{MC}	V_{SAM}^{MC}	V_{NR}^{MC}	V_{SAM}^{MC}	V_{NR}^{MC}
45.0	55.0	58.0	42.0	88.6	11.4	92.5	7.5

Turning to the variance components of the imputed estimator, Table 7 shows the contribution (in %) of the sampling variance, V_{SAM}^{MC}, and that of the nonresponse variance, V_{NR}^{MC}. Note that under the nonresponse mechanism 1 (uniform nonresponse), the imputed estimator is unbiased (see Table 2 and Table 3). It is clear from Table 7 that the contribution of the nonresponse variance decreases as the response rate increases. For example, under mean imputation the nonresponse variance contributes for 55% of the total variance when the response rate is set to 50%, whereas it contributes only for 11.4% of the total variance when the response rate is set to 90%. This result is not surprising since as the response increases, we expect the non-response variance to decrease. Also, it is clear that for a given response rate, the nonresponse variance deceases as the imputation model has more predictive power. For example, when the response rate is set to 50%, the contribution of the nonresponse variance is equal to 55% under mean imputation and only 42% under deterministic regression imputation.

Table 8 shows the increase in mean square error when random regression imputation is used as opposed to deterministic regression imputation under the nonresponse mechanism 1. For a response rate of 50%, the total variance of the imputed estimator under random regression imputation is 20% larger than that of the imputed estimator under deterministic regression imputation. For a response rate of 90%, the relative increase is only equal to approximately 5%.

Table 8: Relative increase in variance, λ, given by (4.3) with n/N = 0.1 under the non-response mechanism 1 and random regression imputation

$p=0.5$	$p=0.9$
1.20	1.05

5. Estimation of Domain Means

In practice, estimates for various domains (subpopulations) are needed for most surveys. For example, in the context of CCHS, estimates of the average weight may be required by age-sex group or by province. Let $U_d \subseteq U$ be a domain of interest of size N_d. The domain mean, $\bar{Y}_d$, can be expressed as

$$\bar{Y}_d = \sum_{i\in U} d_i y_i \Big/ \sum_{i\in U} d_i , \tag{5.1}$$

where d_i is a domain indicator such that $d_i = 1$ if unit *i* belongs to U_d and $d_i = 0$, otherwise. In the absence of nonresponse, an asymptotically unbiased estimator of $\bar{Y}_d$ is given by

$$\bar{y}_d = \frac{\sum_{i\in s} w_i d_i y_i}{\sum_{i\in s} w_i d_i}.$$

That is, $E_p\left(\bar{y}_d\right) \approx \bar{Y}_d$. In the presence of nonresponse to item *y*, we define an imputed estimator by

$$\bar{y}_{dI} = \frac{1}{\sum_{i\in s} w_i d_i}\left[\sum_{i\in s} w_i d_i r_i y_i + \sum_{i\in s} w_i d_i \left(1-r_i\right) y_i^*\right]. \tag{5.2}$$

The imputed estimator (5.2) is simply the weighted mean of the observed and imputed values within the domain. An important question arises in the context of domain estimation: should we account for the domain of interest at the imputation stage? The answer depends on whether or not the domain of interest is related to the variable being imputed. If the domain is highly related to the variable being imputed, then not accounting for the domain in the imputation procedure may lead to imputed estimators with considerable bias. In this case, the nonresponse mechanism is nonignorable because the response probability depends on the error term of the imputation model. As we illustrate next with the CCHS data, the magnitude of the bias increases as the response rate decreases, or as the distance between the domain mean, $\bar{Y}_d$, and the overall mean $\bar{Y}$, increases. In other words, if the behaviour of individuals within the domain is very different than that of the individuals in the rest of the population, then it is important to include the domain variables in the imputation model. On the other hand, if the domain is not related to the variable being imputed, then there is no need to include it because in this case, we expect the behaviour of the individuals within the domain to be similar to that of the individuals in the rest of the population.

To illustrate this problem, we performed several simulation studies. We are interested in estimating the mean for two domains: the first domain is the age-sex group (consisting of 14 categories obtained by cross-classifying the variables AGE_GROUP and DHHE_SEX), whereas the second domain is the province (which consists of 10 categories).

From the population, we selected $R = 500$ samples of size $n = 10,000$ (census case). In each selected sample, nonresponse was generated according to a uniform nonresponse mechanism with probability $p = 0.5$. From each simulated sample, we used two imputation methods: the overall mean imputation and deterministic regression imputation. For the latter method, we used two distinct imputation models: the first corresponds to the full model (see Appendix B), whereas the second includes all the variables of the full model, except the variables AGE_GROUP, DHHE_SEX and the interaction terms AGE_GROUP * DHHE_SEX and AGE_GROUP * CCCE_071 (interaction term for age category with presence of high blood pressure). Thus, in the second model (which we call the incomplete regression model), we purposely omitted the domain variables. Note that that the overall mean of the respondents is probably very close to the population mean since the mechanism used to generate nonresponse to the variable HWTEGWTK gives equal response probability to all the population units.

Table 9 clearly shows that the weight of an individual varies greatly from one age-sex group to another. The range goes from 51.0 kg for females whose age is between 12 and 14 years old to 85 kg for males whose age is between 50 and 64 years old. Thus, a quick look to Table 9 shows that the weight of an individual and its age-sex are strongly related. On the other hand, the mean of the variable HWTEGWTK does not vary much from one province to another. The range goes from 70.3 kg for people living in Quebec to 77.4 for people living in Newfoundland and Labrador as shown in Table 10. Hence, the variables HWTEGWTK and PROVINCE do not seem to be strongly related.

We calculated the Monte Carlo average of the imputed estimator $\overline{y}_{dI}$, given by (4.1) with $\hat{\theta}$ replaced by $\overline{y}_{dI}$. Figures 4-7 show the Monte Carlo average of the imputed estimator for each age-sex group along with its true mean. Under overall mean imputation (see Figure 4), it is clear that the imputed estimator is biased for most of the domains. The bias is especially large for domains 1 and 2 (respectively men and women whose age is between 12 and 14 years old). In these two domains, the average weight is particularly low in comparison with the average weight in the population. Since the nonresponse mechanism is uniform, we expect the mean of the respondents to be close to the population mean. Hence, performing mean imputation in domains 1 and 2 is clearly inadequate since a nonrespondent in domain 1 or 2 (whose weight is probably in the 50 kg range) is imputed by the mean of the respondents, which is much too high. As a result, the imputed estimator has a positive bias as shown in Figure 4. Under regression imputation using the incomplete model, the results in terms of bias are better than those obtained under mean imputation (see Figure 5) because even though the imputation model does not include the variables involving age and sex, the model has more predictive power than the one containing only the intercept (which corresponds to mean imputation). Finally, it is clear from Figure 6, that when the imputation model contains all the appropriate variables (especially the domain variables), the bias of the imputed estimator is negligible for all domains.

Turning to the domain *PROVINCE*, Figure 7 shows that mean imputation leads to imputed estimators with a small bias for all provinces. This result is not surprising since the province is not strongly related to the weight of an individual. In other words, the weight does not vary much from one province to another. Therefore, imputing with the overall mean of the respondents is reasonable in most domains.

In conclusion, in order to ensure approximately unbiased estimators for domains, it is important to include in the imputation those domain variables that are related to the variable being imputed. Failure to do so may result in considerably biased estimators, especially for domains that exhibit an atypical behavior.

Table 9. Population mean by age-sex group

Age-sex group	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Mean	54.5	51.0	71.1	58.8	80.5	65.6	85.4	68.3	85.0	70.0	82.0	69.8	77.5	63.9

Table 10. Population Mean by Province

Province	1	2	3	4	5	6	7	8	9	10	11
Mean	77.4	72.6	74.9	73.8	70.3	73.9	77	76.3	74.8	73.1	76.3

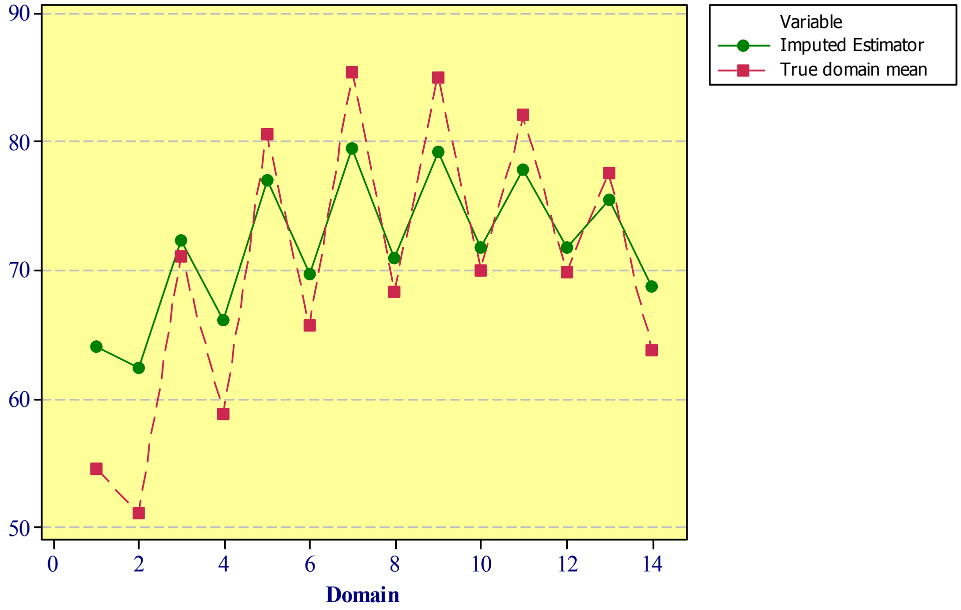

Figure 4: Plot of the imputed estimator vs. true population mean by age-sex under mean imputation

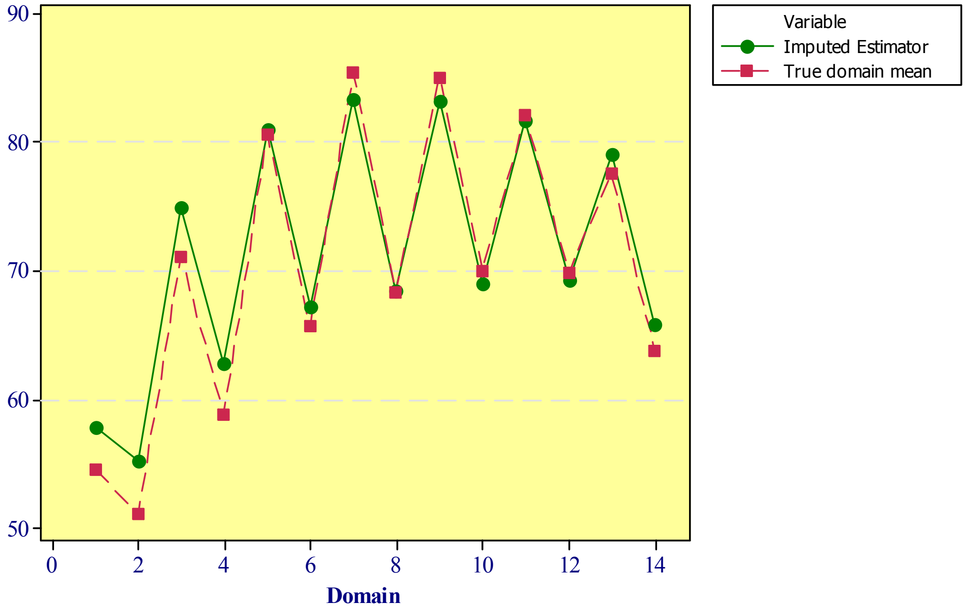

Figure 5: Plot of the imputed estimator vs. true population mean by age-sex under regression imputation (incomplete model

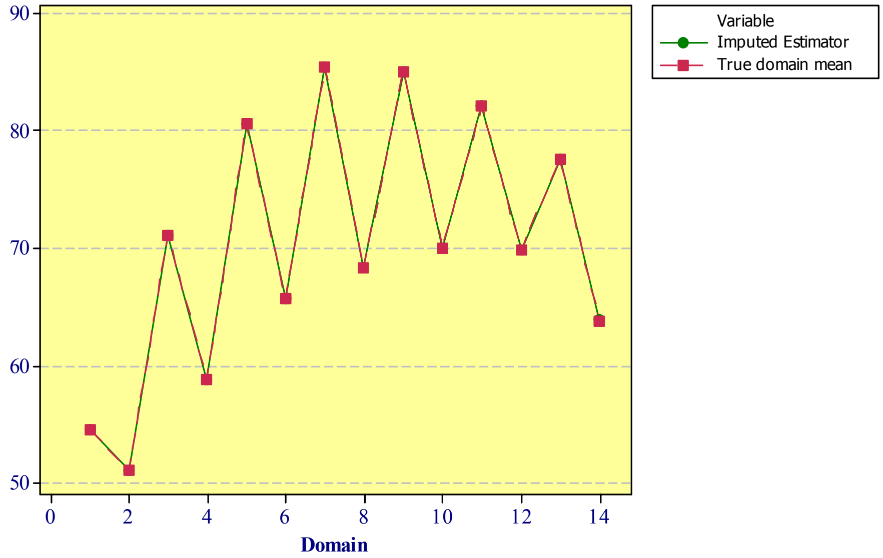

Figure 6: Plot of the imputed estimator vs. true population mean by age-sex under regression imputation (full model)

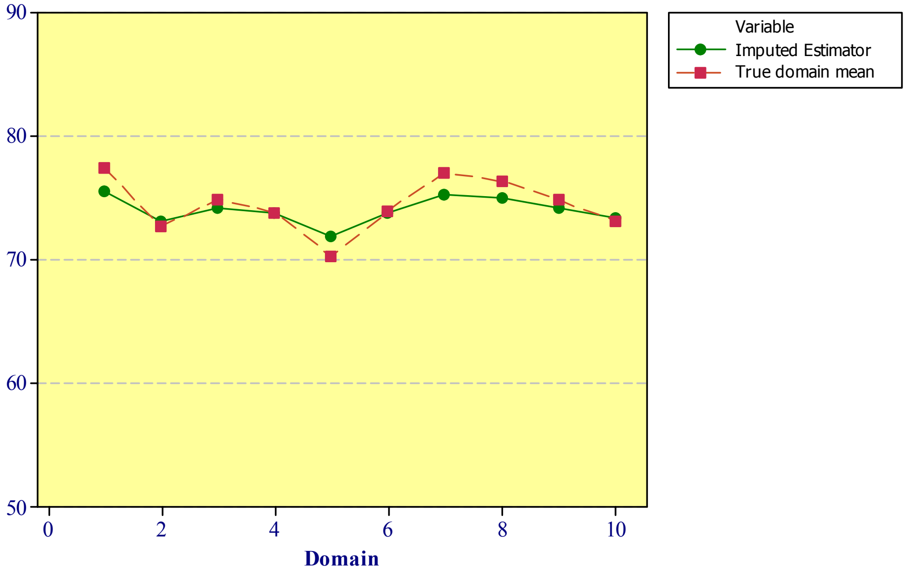

Figure 7: Plot of the imputed estimator vs. true population mean by province under mean imputation

6. Estimation of Coefficients Of Correlation

In this section, we are interested in estimating the finite population coefficient of correlation between two variables *x* and *y*, given by

$$\rho_{xy} = \frac{1}{N-1}\left[\frac{\sum_{i\in U} x_i y_i - N\bar{X}\bar{Y}}{S_x S_y}\right], \qquad (6.1)$$

where $S_x = \left[\frac{1}{N-1}\sum_{i\in U}\left(x_i - \bar{X}\right)^2\right]^{1/2}$ and S_y is similarly defined. Here, both variables are potentially missing and will be imputed. Obtaining an approximately unbiased estimator of ρ_{xy} requires obtaining an approximately unbiased estimator for each term separately, $\sum_{i\in U} x_i y_i$, $\bar{X}$, $\bar{Y}$, S_x and S_y. As we have seen in section 3, obtaining a good estimator of $\bar{X}$ and $\bar{Y}$ requires a good modeling exercise to ensure that the underlying imputation model is at least reasonable. Obtaining an approximately unbiased estimator of the population variability S_x and S_y is possible if a random imputation method is used as opposed to a deterministic imputation method. The use of a random imputation method is necessary because deterministic regression imputation distorts the distribution of the variable being imputed. In particular, the variability of the *y*-values (or *x*-values) after imputation does not provide an unbiased estimator of S_y^2 (or S_x^2). Random regression imputation tends to preserve the distribution of the variables being imputed, particularly their variability. The estimation of the term, $\sum_{i\in U} x_i y_i$, proves to be problematic since this term is a measure of the relationship between the two variables *x* and *y*. Marginal imputation that imputes independently both variables tends to attenuate the relationship between the variables being imputed. This phenomenon can be easily explained by the fact that, after imputation, we are in presence of pairs (x, y) that would not have been observed had there been complete response to both variables. As a result, the estimator of the term $\sum_{i\in U} x_i y_i$, obtained after imputation is generally negatively biased. The magnitude of the bias depends on the response rates to item *x* and *y* as well as the coefficient of correlation between the two variables. If the variables are strongly related and the response rate is low, we can expect the bias of the coefficient of correlation after imputation to be considerable. If the variables *x* and *y* are not related, then the coefficient of correlation computed after imputation would have a negligible bias because, in this case, there is no relationship to preserve.

To illustrate the problem, we conducted a limited simulation study. Suppose we are interested in estimating the coefficient of correlation between the weight of an individual (HWTEGWTK) and its height (HWTEGHTM). From the CCHS population, we selected $R = 1000$ random samples of size $n = 500$ according to simple random sampling without replacement. In each selected sample, we generated nonresponse so that the probability of responding to the variable *x* but not *y* was set to 20%, the probability of responding to the variable *y* but not *x* was set to 20% and the probability of responding to both variables was set to 40%. To compensate for the nonresponse to variables *x* and *y*, we performed marginal random hot deck (MRHD) imputation. That is, both variables were imputed independently using random hot deck imputation described in section 3. Then the coefficient of correlation between the two variables was computed after imputation. The Monte Carlo percent relative bias of the resulting estimator was found to be equal to -58.5%. That is, the Monte Carlo average of the coefficient of correlation after imputation was found to be equal to 0.22, whereas the true value of the coefficient of correlation between HWTEGHTM and HWTEGWTK is approximately equal to 0.55. This example shows that one needs to be extremely careful when performing statistical analyses after imputation has been performed.

7. Conclusions

The results above clearly show that imputation is essentially a modeling exercise. The choice of the auxiliary variables in the imputation model is very important, especially those which are related to the response probability. Model validation is thus an important step during the imputation process. It includes the detection of outliers and the examination of plots such as the plot of residuals vs. the predicted values, the plot of residuals vs. the auxiliary variables selected in the model and a plot of residuals vs. variables not selected in the model. Also, the imputation method should be chosen with respect to the type of parameter being estimated as well as the nature of the variable being imputed (continuous or categorical). For example, if we are interested in estimating a quantile, deterministic regression imputation should be avoided because it tends to distort the distribution of the variables being imputed. Random imputation methods should be used in this case. Also, if the variable being imputed is categorical, donor imputation that includes random hot deck imputation is preferable to (deterministic or random) regression imputation to avoid the possibility of impossible values in the imputed data file.

The problem of variance estimation was not considered in this paper. It is well known that treating the imputed values as if they were observed could lead to a serious underestimation of the variance of the imputed estimator, especially if the nonresponse rate is appreciable. A variety of variance estimation methods that take the nonresponse and imputation variance into account have been developed in the literature. The reader is referred to the following papers on the topic: Rao and Shao (1992); Särndal (1992); Rao and Sitter (1995); Fay (1996); Shao and Sitter (1996); and Shao and Steel (1999). These methods cover many imputation methods including both deterministic and random regression imputation.

Acknowledgments: We thank Cynthia Bocci from Statistics Canada for constructive comments and suggestions that helped in improving the quality of the paper. David Haziza's work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

Correspondence : david.haziza@umontreal.ca

REFERENCES

Fay, R.E. 1996. Alternative Paradigms for the Analysis of Imputed Survey Data. *Journal of the American Statistical Association*, 91, 490-498.

Rao, J.N.K. 1996. On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.

Rao, J.N.K. and Sitter, R.R. 1995. Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

Rao, J.N.K. and Shao, J. 1992. Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

Särndal, C.-E. 1992. Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used. *Survey Methodology*, 18, 241-252.

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, 254-26.

Statistics Canada 2005. Canadian Community Health Survey, Public Use Microdata, www.statacan.ca.

APPENDIX A: Variables and their definitions

Table A.1: List of variables included in the CCHS case study data set

Variable name	Variable Label (meaning)	Type of variable	Number of values
GEOEGPRV	Province of residence of respondent	nominal categorical	11
AGE_GROUP	Age	ordinal categorical	7
DHHE_SEX	Sex	nominal categorical	2
DHHEGMS	Marital status	nominal categorical	4
CCCE_011	Has food allergies	nominal categorical	2
CCCE_031	Has asthma	nominal categorical	2
CCCE_071	Has high blood pressure	nominal categorical	2
PACEDEE	Daily energy expenditure	continuous	N/A
PACEDPAI	Physical activity index	ordinal categorical	3
SMKE_202	Type of smoker	ordinal categorical	3
ETSE_10	Someone smokes inside home	nominal categorical	2
ALCEDTYP	Type of drinker	nominal categorical	4
ALCEDDLY	Average daily alcohol consumption - (D)	discrete continuous	N/A
INCEGHH	Total hhld inc. from all sources	ordinal categorical	5
HWTEGHTM	Height (metres) / self-reported	ordinal categorical	28
HWTEGWTK	Weight (kgs)/ self-reported	continuous	N/A

Table A.2. List of values taken by variables

Variable name	Variable value	Value Label (meaning)
GEOEGPRV	10	NFLD & LAB.
	11	PEI
	12	NOVA SCOTIA
	13	NEW BRUNSWICK
	24	QUEBEC
	35	ONTARIO
	46	MANITOBA
	47	SASKATCHEWAN
	48	ALBERTA
	59	BRITISH COLUMBIA
	60	YUKON/NWT/NUNA.
AGE_GROUP	1	12 TO 14 YEARS
	2	15 TO 17 YEARS
	3	18 TO 29 YEARS
	4	30 TO 49 YEARS
	5	50 TO 64 YEARS
	6	65 TO 74 YEARS
	7	75 YEARS OR MORE
DHHE_SEX	1	MALE
	2	FEMALE
DHHEGMS	1	MARRIED
	2	COMMON-LAW
	3	WIDOW/SEP/DIV
	4	SINGLE/NEVER MAR
CCCE_011	1	YES
	2	NO
CCCE_031	1	YES
	2	NO
CCCE_071	1	YES
	2	NO
PACEDEE	N/A	N/A
PACEDPAI	1	ACTIVE
	2	MODERATE
	3	INACTIVE
SMKE_202	1	DAILY
	2	OCCASIONALLY
	3	NOT AT ALL
ETSE_10	1	YES
	2	NO
ALCEDTYP	1	REGULAR DRINKER
	2	OCCASIONAL DRINKER
	3	FORMER DRINKER
	4	NEVER DRANK
ALCEDDLY	N/A	N/A
INCEGHH	1	NO OR <$15,000
	2	$15,000-$29,999
	3	$30,000-$49,999
	4	$50,000-$79,999
	5	$80,000 OR MORE
HWTEGHTM	midpoint of range	
HWTEGWTK	N/A	N/A

APPENDIX B:
Regression analysis with HWTEGWTK (weight) as the dependent variable

Class Level Information

Class	Levels	Values
DHHE_SEX	2	1 2
AGE_GROUP	7	1 2 3 4 5 6 7
CCCE_071	2	1 2
ALCEDTYP	4	1 2 3 4
AGE_SEX_NUMERIC	14	1 2 3 4 5 6 7 8 9 10 11 12 13 14
AGE_BP_NUMERIC	12	1 2 3 4 5 6 7 8 9 10 11 12

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	22	1202715.673	54668.894	321.75	<.0001
Error	9977	1695180.999	169.909		
Corrected Total	9999	2897896.673			

R-Square	Coeff Var	Root MSE	HWTEGWTK Mean
0.415031	17.71113	13.03491	73.59728

Source	DF	Type I SS	Mean Square	F Value	Pr > F
HWTEGHTM	1	876302.1629	876302.1629	5157.48	<.0001
DHHE_SEX	1	17454.4488	17454.4488	102.73	<.0001
AGE_GROUP	6	207989.8583	34664.9764	204.02	<.0001
CCCE_071	1	70169.7382	70169.7382	412.98	<.0001
ALCEDTYP	3	18100.9878	6033.6626	35.51	<.0001
AGE_SEX_NUMERIC	6	6313.7893	1052.2982	6.19	<.0001
AGE_BP_NUMERIC	4	6384.6881	1596.1720	9.39	<.0001