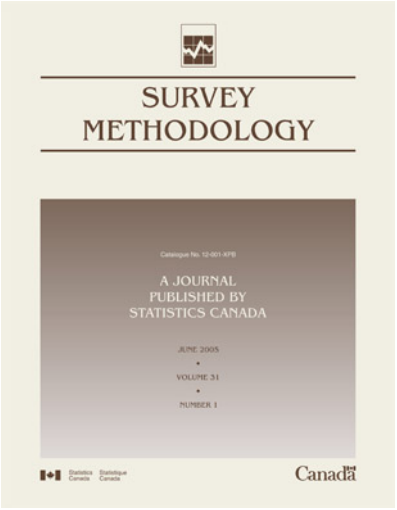
Catalogue no. 12-001-XIE

Survey Methodology

June 2006



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

June 2006

Published by authority of the Minister responsible for Statistics Canada

July 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 32, No. 1, pp. 53-64
Statistics Canada, Catalogue No. 12-001

A Nonresponse Model Approach to Inference Under Imputation for Missing Survey Data

David Haziza and Jon N.K. Rao [1]

Abstract

In the presence of item nonreponse, two approaches have been traditionally used to make inference on parameters of interest. The first approach assumes uniform response within imputation cells whereas the second approach assumes ignorable response but make use of a model on the variable of interest as the basis for inference. In this paper, we propose a third appoach that assumes a specified ignorable response mechanism without having to specify a model on the variable of interest. In this case, we show how to obtain imputed values which lead to estimators of a total that are approximately unbiased under the proposed approach as well as the second approach. Variance estimators of the imputed estimators that are approximately unbiased are also obtained using an approach of Fay (1991) in which the order of sampling and response is reversed. Finally, simulation studies are conducted to investigate the finite sample performance of the methods in terms of bias and mean square error.

Key Words: Bias-adjusted estimator; Deterministic regression imputation; Imputation model approach; Item nonresponse; Nonresponse model approach; Random regression imputation; Variance estimation.

1. Introduction

Item nonresponse occurs in a survey when a sampled element participates in the survey but fails to provide responses on one or more of the survey items (Brick and Kalton 1996). It is usually handled by some form of imputation which involves "filling in" missing values for each item. Imputation may achieve an effective bias reduction, provided suitable auxiliary information is available for all the sampled elements and appropriately incorporated in the imputation model and/or the non-response model.

Imputation offers the following desirable features, among others: (i) it leads to the creation of a complete data file, and (ii) it permits the use of the same survey weights for all items which ensures that the results obtained from different analyses of the completed data set are consistent with one another, unlike the results of analyses from an incomplete data set. However, imputation also presents the following difficulties, among others: (a) marginal imputation for each item distorts the relationship between items, and (b) treating the imputed values as if they were true values may lead to serious underestimation of the variance of imputed estimators, especially when the nonresponse rate is appreciable. Methods that address (a) and (b) have been proposed in the literature.

In this paper, we focus on marginal imputation that is commonly used in many surveys. We first consider deterministic linear regression imputation that includes mean and ratio imputation as special cases. In this method a missing value is replaced by the predicted value obtained by fitting a linear regression model using respondent values and auxiliary variables collected on all the sampled elements. We also consider the case of random linear regression imputation that may be viewed as a deterministic regression imputation plus an added random residual. It includes random hot-deck imputation as a special case.

Let U be a finite population of possibly unknown size N. The objective is to estimate the population total $Y = \sum_U y_i$ of an item y when imputation has been used to compensate for nonresponse on the item values y_i. For brevity, $\sum_A$ will be used for $\sum_{i \in A}$, where $A \subseteq U$. Suppose a probability sample, s, of size n is selected according to a specified design $p(s)$ from U. Under complete response to item y, a design-unbiased estimator of Y is given by the well-known Horvitz-Thompson estimator

$$\hat{Y} = \sum_s w_i y_i, \qquad (1)$$

with sampling (or design) weights $w_i = 1/\pi_i$, where π_i denotes the inclusion probability of population unit i in the sample s, $i = 1, \ldots, N$. Rao (2005) suggested that (1) should be called the Narain-Horvitz-Thompson (NHT) estimator in recognition of the fact that Narain (1951) also discovered (1) independently of Horvitz and Thompson (1952).

In the presence of nonresponse to item y, we use imputation and define an imputed estimator $\hat{Y}_I$ as

$$\hat{Y}_I = \sum_s w_i a_i y_i + \sum_s w_i (1 - a_i) y_i^* = \sum_s w_i \tilde{y}_i, \qquad (2)$$

where y_i^* denotes the value imputed for missing y_i, a_i denotes the response indicator equal to 1 if unit i responds to item y and 0 otherwise and $\tilde{y}_i = a_i y_i + (1 - a_i) y_i^*$. The

1. David Haziza, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada. K1S 5B6.

imputed estimator (2) can be implemented from the imputed data file containing the survey weights w_i and the $\tilde{y}_i$ only, without the knowledge of response indicators a_i. However, the response indicators will be required for variance estimation. Let $p_i = P(a_i = 1)$ be the item y response probability for unit i. In this paper, we assume that the units respond independently of one another, *i.e.*, $p_{ij} = P(a_i = 1, a_j = 1) = p_i p_j$ if $i \neq j$.

As for any method of compensating for missing data, imputation requires some assumptions about the response mechanism and/or the imputation model. In the presence of imputed data, two different approaches are generally used for making inference on totals, means and other parameters of interest: (i) Imputation model (IM) approach; (ii) Nonresponse Model (NM) approach. Approach (i) is also called model-assisted approach (Särndal 1992) and approach (ii) design-based approach (Shao and Steel 1999). NM approach is based on partitioning the population U into J imputation cells and then imputing nonrespondents y–values within each cell using respondent y–values within the same cell as donor values, independently across the J cells. The following assumption is made:

Assumption NM: Response probability for a given item of interest is constant within imputation cells. That is, $p_i = p_v$, say, where the subscript v denotes the imputation cell.

In the NM approach, explicit assumptions about the response mechanism are made. It follows that inference under assumption NM is with respect to repeated sampling and uniform response mechanism within cells. Approach NM has been studied by Rao (1990, 1996), Rao and Shao (1992), Rao and Sitter (1995) and Shao and Steel (1999), among others. For simplicity, we assume a single imputation cell so that $p_i = p$ under assumption NM.

IM approach is based on the following assumption:

Assumption IM: Item values are missing at random (MAR) in the sense that the response probability does not depend on the item value being imputed but may depend on auxiliary variables used for imputation. Further, a model that generates the item values y_i is assumed.

In the IM approach, explicit assumptions about the distribution of item values y_i is made through a model called the "imputation model". It follows that inference under assumption IM is with respect to repeated sampling and the assumed model that generates the finite population of y–values and nonrespondents to item y. Underlying response mechanism is not specified, except for the MAR assumption, unlike in the NM approach. The assumed response mechanism under assumption IM is much weaker than the uniform response within cells under assumption NM, but inferences under assumption IM depends on the assumed population model. IM approach has been studied by Särndal (1992), Deville and Särndal (1994) and Shao and Steel (1999), among others.

Under linear regression imputation, IM approach assumes the following linear regression imputation model:

$$E_m(y_i) = \mathbf{z}_i'\boldsymbol{\gamma},\ V_m(y_i) = \sigma_i^2 = \sigma^2(\boldsymbol{\lambda}'\mathbf{z}_i),$$
$$\mathrm{Cov}_m(y_i, y_j) = 0 \text{ if } i \neq j, \qquad (3)$$

where $\boldsymbol{\gamma}$ is k–vector of unknown parameters, $\mathbf{z}_i$ is a k–vector of auxiliary variables available for all $i \in s$, $\boldsymbol{\lambda}$ is a k–vector of specified constants, σ^2 is an unknown parameter and E_m, V_m, and Cov_m denote respectively the expectation, the variance and the covariance operators with respect to the imputation model. The restriction $\sigma_i^2 = \sigma^2(\boldsymbol{\lambda}'\mathbf{z}_i)$ does not severely restrict the range of imputation models.

In this paper, we propose a third approach, called the Generalized Nonresponse Model (GNM) approach. GNM approach is based on the following assumption:

Assumption GNM: Item values are missing at random (MAR) and response probability is specified as a function of auxiliary variables, $\mathbf{u}_i$, observed on all the sample elements, and unknown parameters $\boldsymbol{\eta}$.

In this paper, we assume that the probability of response, p_i, for unit i, is linked to an l–vector of auxiliary variables $\mathbf{u}_i$ according to a logistic model so that

$$p_i = f(\mathbf{u}_i'\boldsymbol{\eta}) = \exp(\mathbf{u}_i'\boldsymbol{\eta}) / \exp(1 + \mathbf{u}_i'\boldsymbol{\eta}), \qquad (4)$$

where $\boldsymbol{\eta}$ is the l–vector of model parameters. Model (4) is the assumed nonresponse model. It can be validated from the values a_i and $\mathbf{u}_i$ for $i \in s$. Note that a_i and $\mathbf{u}_i$ are item specific. Also, note that assumption NM is a special case of assumption GNM. As in NM approach, explicit assumptions about the response mechanism are made and inference under assumption GNM is with respect to repeated sampling and the assumed response mechanism.

Recall that imputation is designed to reduce the nonresponse bias, assuming that the available auxiliary variables can explain the item to be imputed and/or the item response probability. Hence, in practice, the choice of the approach (IM or GNM) should be dictated by the quality of the imputation model and the nonresponse model. The choice between modeling the item response probability and modeling the item of interest will depend on how much reliance one is ready to place on the two models. Although it may seem intuitively more appealing to model the item of interest, there are some cases encountered in practice for which it may be easier to model the response probability (GNM approach). For example, the Capital Expenditures Survey at Statistics Canada produces data on investment made in Canada, in all types of Canadian industries. For this survey, two important variables of interest are capital

expenditures on new construction (CC) and capital expenditures on new machinery and new equipment (CM). In a given year, a large number of businesses have not invested any amount of money on new construction or new machinery. As a result, the sample data file contains a large number of zeros for the two variables CC and CM. In this case, modeling the variables of intrest (CC or CM) may prove to be difficult.

Survey design weights are generally used in linear regression imputation. The resulting imputed estimator of a population total is "robust" in the sense that it is approximately unbiased under either assumption NM or assumption IM. However, the imputed estimator is generally biased under assumption GNM. In this paper, we propose a new method of linear regression imputation that is robust in the sense of leading to approximately unbiased estimators under either assumption GNM or assumption IM.

Section 2 develops a new method of deterministic linear regression imputation as well as random linear regression imputation, and demonstrates the robustness property in estimating a population total Y. Results of a simulation study on the finite-sample performance of the imputed estimator under the new method of imputation are reported in section 3. Variance estimators are derived in section 4, using the 'reverse' approach of Fay (1991) in which the order of sampling and response is reversed:

Population → census with nonrespondents → sample with nonrespondents.

Simulation results on variance estimators are also given. Finally, the case of domain means is investigated in section 5.

2. Estimation of a Total

In this section, we study the bias of the imputed estimator $\hat{Y}_I$. The total error, $\hat{Y}_I - Y$, may be decomposed as

$$\hat{Y}_I - Y = (\hat{Y} - Y) + (\hat{Y}_I - \hat{Y}). \tag{5}$$

The term $\hat{Y} - Y$ in (5) is called the sampling error, whereas the term $\hat{Y}_I - \hat{Y}$ is called the nonresponse/imputation error. Note that there is no imputation error under deterministic imputation. Since the sampling error does not depend on nonresponse and imputation method, we focus on the nonresponse/imputation error $\hat{Y}_I - \hat{Y}$ and evaluate its properties conditionally on the sample s. Under the NM or GNM approach, the conditional nonresponse bias is defined as $E_r(\hat{Y}_I - \hat{Y} \mid s)$, where $E_r(.)$ denotes the expectation with respect to the response mechanism. Under the IM approach, the conditional nonresponse bias is defined as $E_r E_m(\hat{Y}_I - \hat{Y} \mid s)$ under MAR assumption.

2.1 Deterministic Regression Imputation

Deterministic regression imputation uses the imputed values

$$y_i^* = \mathbf{z}_i' \hat{\boldsymbol{\gamma}}_r \tag{6}$$

for missing y_i, where

$$\hat{\boldsymbol{\gamma}}_r = \left(\sum_s w_i a_i \mathbf{z}_i \mathbf{z}_i' / (\boldsymbol{\lambda}' \mathbf{z}_i) \right)^{-1} \sum_s w_i a_i \mathbf{z}_i y_i / (\boldsymbol{\lambda}' \mathbf{z}_i) \tag{7}$$

is the weighted least squares estimator of $\boldsymbol{\gamma}$ in the model (3), based on the sample elements responding to item y. Using (6), the imputed estimator (2) can be written as

$$\hat{Y}_I = \hat{Y}_r + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_\mathbf{r})' \hat{\boldsymbol{\gamma}}_r, \tag{8}$$

where $\hat{Y}_r = \sum_s w_i a_i y_i$, $\hat{\mathbf{Z}} = \sum_s w_i \mathbf{z}_i$ and $\hat{\mathbf{Z}}_\mathbf{r} = \sum_s w_i a_i \mathbf{z}_i$. Note that the imputed estimator (8) is similar to a regression estimator in the case of two-phase sampling.

Under assumption NM, $E_r(a_i \mid s) = p$ and the conditional nonresponse bias, $E_r(\hat{Y}_I - \hat{Y} \mid s)$, is approximately equal to 0. Furthermore, under assumption IM and regression model (3), the conditional nonresponse bias $E_r E_m(\hat{Y}_I - \hat{Y} \mid s)$, is equal to 0. However, under assumption GNM, the conditional nonresponse bias is given by

$$E_r(\hat{Y}_I - \hat{Y} \mid s) \approx -\sum_s w_i (1 - p_i)(y_i - \mathbf{z}_i' \hat{\boldsymbol{\gamma}}_p) \equiv \mathrm{B}(\hat{Y}_I \mid s), \tag{9}$$

where

$$\hat{\boldsymbol{\gamma}}_p = \left(\sum_s w_i p_i \mathbf{z}_i \mathbf{z}_i' / (\boldsymbol{\lambda}' \mathbf{z}_i) \right)^{-1} \sum_s w_i p_i \mathbf{z}_i y_i / (\boldsymbol{\lambda}' \mathbf{z}_i). \tag{10}$$

This result follows from the fact that under assumption GNM, $E_r(a_i \mid s) = p_i$. Hence, the choice of imputed values (6) is, in general, not suitable under assumption GNM. For the special case of assumption NM with $p_i = p$, the last term in (9) vanishes, noting that $(\sum_s w_i \mathbf{z}_i')\hat{\boldsymbol{\gamma}}_p = \boldsymbol{\lambda}'(\sum_s w_i \mathbf{z}_i \mathbf{z}_i' / (\boldsymbol{\lambda}' \mathbf{z}_i))\hat{\boldsymbol{\gamma}}_p = \boldsymbol{\lambda}'(\sum_s w_i \mathbf{z}_i y_i / (\boldsymbol{\lambda}' \mathbf{z}_i)) = \sum_s w_i y_i$.

2.2 A Bias-Adjusted Estimator

We assume for now that the response probabilities p_i are known. A natural approach for eliminating the bias of $\hat{Y}_I$ under assumption GNM is to consider a bias-adjusted estimator of the form

$$\hat{Y}_I^a = \hat{Y}_I - \hat{B}(\hat{Y}_I \mid s), \tag{11}$$

where $\hat{B}(\hat{Y}_I \mid s)$ is an estimator of $\mathrm{B}(\hat{Y}_I \mid s)$:

$$\hat{B}(\hat{Y}_I \mid s) = -\sum_s w_i a_i \frac{(1 - p_i)}{p_i} (y_i - \mathbf{z}_i' \hat{\boldsymbol{\gamma}}_r). \tag{12}$$

Note that $E_r[\hat{B}(\hat{Y}_I \mid s) \mid s] \approx \mathrm{B}(\hat{Y}_I \mid s)$ under assumption GNM. Substituting (12) in (11), we get a bias-adjusted estimator as

$$\hat{Y}_I^a = \sum_s \frac{w_i}{p_i} a_i y_i + \left(\sum_s w_i \mathbf{z}_i' - \sum_s \frac{w_i}{p_i} a_i \mathbf{z}_i' \right) \hat{\boldsymbol{\gamma}}_r. \tag{13}$$

Note that (13) is also in the form of a two phase regression estimator.

In practice, response probabilities p_i are unknown. Suppose we can obtain estimators $\hat{p}_i$ of p_i by modelling p_i according to the nonresponse model (4). Then, a bias-adjusted estimator is obtained by replacing p_i in (13) with $\hat{p}_i$. This estimator is also approximately conditionally unbiased under assumption IM. Hence, the bias-adjusted estimator (13) is robust in the sense of validity under either assumption IM or assumption GNM. However, unlike the imputed estimator $\hat{Y}_I$ given by (2), the bias-adjusted estimator $\hat{Y}_I^a$ cannot be computed without the knowledge of the response identifiers, a_i, and the estimated response probabilities, $\hat{p}_i$. Hence, both the response indicators and the estimated response probabilities must be provided with the imputed data file to implement $\hat{Y}_I^a$, which may not be the case in practice. This drawback of $\hat{Y}_I^a$ can be eliminated by using the new imputation method, given in section 2.3, that leads to an approximately unbiased estimator under either assumption GNM or assumption IM without the knowledge of a_i and $\hat{p}_i$ on the imputed data file. However, for variance estimation, access to a_i and $\hat{p}_i$ is needed.

2.3 Modified Deterministic Regression Imputation

We assume for now that the response probabilites p_i are known. We then use the imputed values

$$y_i^* = \mathbf{z}_\mathbf{i}'\tilde{\boldsymbol{\gamma}}_s \tag{14}$$

for missing y_i and obtain the form of $\tilde{\boldsymbol{\gamma}}_s$ that leads to an approximately unbiased estimator under assumption GNM.

2.3.1 Approximately Unbiased Estimator

The following lemma gives the form of $\tilde{\boldsymbol{\gamma}}_s$ that leads to an approximately unbiased estimator under assumption GNM.

Lemma 1: *Under assumption GNM, the choice of* $\tilde{\boldsymbol{\gamma}}_s$ *that leads to* $E_r(\hat{Y}_I - \hat{Y} \mid s) = 0$ *is given by*

$$\tilde{\boldsymbol{\gamma}}_{s,N} = \left[\sum_s w_i(1-p_i)\mathbf{z}_\mathbf{i}\mathbf{z}_\mathbf{i}'/(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i})\right]^{-1} \sum_s w_i(1-p_i)\mathbf{z}_\mathbf{i} y_i/(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i}). \tag{15}$$

Proof: The conditional nonresponse bias of $\hat{Y}_I$ with $y_i^* = \mathbf{z}_\mathbf{i}'\tilde{\boldsymbol{\gamma}}_s$ under assumption GNM is given by

$$E_r(\hat{Y}_I - \hat{Y} \mid s) = -\sum_s w_i(1-p_i)(y_i - \mathbf{z}_\mathbf{i}'\tilde{\boldsymbol{\gamma}}_s).$$

Noting that $(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i})/(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i}) = 1$, it follows that $E_r(\hat{Y}_I - \hat{Y} \mid s) = 0$ if $\tilde{\boldsymbol{\gamma}}_s$ satisfies

$$\boldsymbol{\lambda}'\left[\sum_s w_i(1-p_i)\mathbf{z}_\mathbf{i}(y_i - \mathbf{z}_\mathbf{i}'\tilde{\boldsymbol{\gamma}}_s)/(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i})\right] = 0. \tag{16}$$

The choice $\tilde{\boldsymbol{\gamma}}_s = \tilde{\boldsymbol{\gamma}}_{s,N}$ satisfies (16).

Note that $\tilde{\boldsymbol{\gamma}}_{s,N}$ is unknown since the y–values are only observed for $i \in s_r$ and the response probabilities p_i are unknown. An estimator of $\tilde{\boldsymbol{\gamma}}_{s,N}$, based on the responding units and estimated response probabilites $\hat{p}_i$, is given by

$$\tilde{\boldsymbol{\gamma}}_r = \left[\sum_s w_i a_i \frac{(1-\hat{p}_i)}{\hat{p}_i}\mathbf{z}_\mathbf{i}\mathbf{z}_\mathbf{i}'/(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i})\right]^{-1} \sum_s w_i a_i \frac{(1-\hat{p}_i)}{\hat{p}_i}\mathbf{z}_\mathbf{i} y_i/(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i}). \tag{17}$$

We have $E_r(\tilde{\boldsymbol{\gamma}}_r \mid s) \approx \tilde{\boldsymbol{\gamma}}_{s,N}$ so that $\tilde{\boldsymbol{\gamma}}_r$ is conditionally approximately unbiased for $\tilde{\boldsymbol{\gamma}}_{s,N}$ under assumption GNM. Hence, using the imputed values

$$y_i^* = \mathbf{z}_\mathbf{i}'\hat{\boldsymbol{\gamma}}_r \tag{18}$$

in (2) with $\tilde{\boldsymbol{\gamma}}_r$ given by (17), leads to an approximately unbiased estimator of the total Y under assumption GNM. Note that $\tilde{\boldsymbol{\gamma}}_r$ is a weighted least square estimator of $\boldsymbol{\gamma}$ with respect to a new set of weights, $\tilde{w}_i/(\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i})$, where $\tilde{w}_i = w_i((1-\hat{p}_i)/\hat{p}_i)$. Hence, the procedure increases the weights w_i for those units with $\hat{p}_i < 1/2$ and decreases the weights for those units with $\hat{p}_i > 1/2$. The imputed estimator can be implemented from the imputed data file containing the sampling weights w_i and the $\tilde{y}_i$ only; response identifiers a_i and estimated response probabilities, $\hat{p}_i$, are not required. However, a_i and $\hat{p}_i$ are needed for variance estimation. Note that the producer of the imputed data file uses the information on a_i and $\mathbf{u}_\mathbf{i}$ to fit the response model (4) and generate the imputed values y_i^* given by (18).

The use of imputed values (18) also leads to an approximately unbiased estimator of Y under assumption IM. First, under the regression model (3), noting that $E_m(y_i \mid s) = \mathbf{z}_\mathbf{i}'\boldsymbol{\gamma}$ and $E_m(\tilde{\boldsymbol{\gamma}}_r \mid s) = \boldsymbol{\gamma}$, we have $E_m(\hat{Y}_I - \hat{Y} \mid s) = 0$ and $E_r E_m(\hat{Y}_I - \hat{Y} \mid s) = 0$ without specifying the underlying MAR response mechanism. Hence, the use of imputed values (18) leads to a robust imputed estimator in the sense of validity under both approaches. Finally, it is interesting to note that the imputed values (18) can also be obtained using the method of calibration imputation (Beaumont 2005). Calibration imputation consists of finding final imputed values as close as possible to original imputed values according to some distance function, subject to the calibration constraint.

Two particular cases of modified regression imputation (18) are of interest: (i) modified ratio imputation with $\mathbf{z}_\mathbf{i} = z_i$ and $\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i} = z_i$; (ii) modified mean imputation with $\mathbf{z}_\mathbf{i} = 1$ and $\boldsymbol{\lambda}'\mathbf{z}_\mathbf{i} = 1$. In case (i), the imputed values (18) reduce to

$$y_i^* = \frac{\sum_s \tilde{w}_i a_i y_i}{\sum_s \tilde{w}_i a_i z_i} z_i. \tag{19}$$

In case (ii), the imputed values (18) reduce to

$$y_i^* = \frac{\sum_s \tilde{w}_i a_i y_i}{\sum_s \tilde{w}_i a_i}. \qquad (20)$$

Under uniform response $p_i = p$, the imputed values (19) and (20) reduce to $(\sum_s w_i a_i y_i / \sum_s w_i a_i z_i) z_i$ and $\bar{y}_r = \sum_s w_i a_i y_i / \sum_s w_i a_i$ respectively, which are the usual values that survey practioners use for ratio and mean imputation (Rao and Sitter 1995).

2.3.2 Optimal Choice of $\tilde{\gamma}_s$

We now turn to the "optimal" choice of $\tilde{\gamma}_s$ by minimizing the conditional mean square error of the imputed estimator $\hat{Y}_I$ with $y_i^* = \mathbf{z}_i' \tilde{\gamma}_s$. The conditional mean square error of the imputed estimator $\hat{Y}_I$ is given by

$$\begin{aligned} \mathrm{MSE}_r(\hat{Y}_I \mid s) &= V_r(\hat{Y}_I \mid s) + [\mathrm{Bias}(\hat{Y}_I \mid s)]^2 \\ &= \sum_s w_i^2 p_i (1-p_i)(y_i - \mathrm{z}_i' \tilde{\gamma}_s)^2 \\ &\quad + \left[\sum_s w_i (1-p_i)(y_i - \mathrm{z}_i' \tilde{\gamma}_s) \right]^2, \qquad (21) \end{aligned}$$

where $V_r(.\mid s)$ denotes the conditional nonresponse variance with respect to the response mechanism, given the sample s. We search for $\tilde{\gamma}_s$ that minimizes $\mathrm{MSE}_r(\hat{Y}_I \mid s)$.

The optimal choice, $\tilde{\gamma}_{\mathrm{opt}}$, of $\tilde{\gamma}_s$ is complex, but in the special case of ratio imputation, $\tilde{\gamma}_{\mathrm{opt}}$ reduces to

$$\tilde{\gamma}_{\mathrm{opt}} = \frac{\sum_s w_i (1-p_i) y_i \sum_s w_i (1-p_i) z_i + \sum_s w_i^2 p_i (1-p_i) y_i z_i}{\left[\sum_s w_i (1-p_i) z_i \right]^2 + \sum_s w_i^2 p_i (1-p_i) z_i^2}. \qquad (22)$$

Assume that the sampling weights w_i satisfy $\max(n / N w_i) = O(1)$ and that a positive constant C exists such that $C < p_i$. Then,

$$\begin{aligned} \tilde{\gamma}_{\mathrm{opt}} &= \frac{\sum_s w_i (1-p_i) y_i}{\sum_s w_i (1-p_i) z_i} + O\left(\frac{1}{n}\right) \\ &= \tilde{\gamma}_{s,N} + O\left(\frac{1}{n}\right). \end{aligned}$$

Hence, for large sample sizes, the choice $\tilde{\gamma}_{s,N}$ is nearly optimal for ratio imputation. Similarly, $\tilde{\gamma}_{s,N}$ is nearly optimal for mean imputation which is a special case of ratio imputation.

2.4 Random Regression Imputation

Random imputation can be viewed as deterministic imputation plus a random noise. Let s_r and s_m denote the sets of sample respondents and nonrespondents respectively, and let $e_j = (y_j - \mathbf{z}_j' \hat{\gamma}_r) / (\lambda' \mathbf{z}_j)^{1/2}$ be the standardized residuals for the respondents $j \in s_r$ under deterministic regression imputation. Further, $e_i^* = e_j$ with $P(e_i^* = e_j) = w_j / \sum_s w_i a_i$ independently for each $i \in s_m$. Then, random regression imputation uses the imputed values $y_i^* = \mathbf{z}_i' \hat{\gamma}_r + \epsilon_i^*, i \in s_m$, where $\epsilon_i^* = (\lambda' \mathbf{z}_i)^{1/2} (e_i^* - \bar{e}_r)$ with $\bar{e}_r = \sum_s w_j a_j e_j / \sum_s w_j a_j$. Let $E_*(.)$ denote the expectation with respect to the random imputation process. We have $E_*(\epsilon_i^*) = 0$ and $E_*(\hat{Y}_I)$ equals (8). Hence, the imputed estimator $\hat{Y}_I$ is approximately unbiased under either assumption NM or assumption IM. It may be noted that random regression imputation covers random (weighted) hot-deck imputation as a special case. To see this, consider the mean imputation model $E_m(y_i) = \gamma, V_m(y_i) = \sigma^2$ and $\mathrm{Cov}_m(y_i, y_j) = 0, i \neq j$. We have $\hat{\gamma}_r = \sum_s w_i a_i y_i / \sum_s w_i a_i = \bar{y}_r$, the weighted mean of the respondent y–values, and $e_j = y_j - \bar{y}_r$. Therefore, $y_i^* = \bar{y}_r + \epsilon_i^* = y_j$ corresponds to the respondent value y_j drawn at random with probability $w_j / \sum_s w_i a_i$.

The imputed estimator based on random regression imputation is asymptotically biased under assumption GNM. To obtain an approximately unbiased estimator for Y, we propose modified random regression imputation. Let $\tilde{e}_j = (y_j - \mathbf{z}_j' \tilde{\gamma}_r) / (\lambda' \mathbf{z}_j)^{1/2}$ and $\tilde{e}_i^* = \tilde{e}_j$ with $P(\tilde{e}_i^* = \tilde{e}_j) = \tilde{w}_j / \sum_s \tilde{w}_i a_i$ independently for each $i \in s_m$, where $\tilde{\gamma}_r$ is given by (17) and $\tilde{w}_i = w_i (1 - \hat{p}_i) / \hat{p}_i$. Then, modified random regression imputation uses the imputed values $y_i^* = \mathbf{z}_i' \tilde{\gamma}_r + \tilde{\epsilon}_i^*$, where $\tilde{\epsilon}_i^* = (\lambda' \mathbf{z}_i)^{1/2} (\tilde{e}_i^* - \tilde{e}_r)$ with $\tilde{e}_r = \sum_s \tilde{w}_j a_j \tilde{e}_j / \sum_s \tilde{w}_j a_j$. We have $E_*(\epsilon_i^*) = 0$ and $E_*(\hat{Y}_I)$ equals the imputed estimator under modified deterministic regression imputation. Hence, the imputed estimator $\hat{Y}_I$ is approximately unbised under either assumption GNM or assumption IM. For the special case of mean imputation model, we have $\tilde{\gamma}_r = \sum_s \tilde{w}_i a_i y_i / \sum_s \tilde{w}_i a_i$ and $y_i^* = y_j$ corresponds to the respondent value y_j drawn at random with probability $\tilde{w}_j / \sum_s \tilde{w}_i a_i$.

3. Simulation Studies

We performed two simulation studies to investigate the finite sample performance of the proposed deterministic modified regression and modified random regression imputation methods in terms of relative bias and relative root mean square error. The first simulation study compares the performance of the traditional deterministic regression imputation and the proposed modified deterministic regression imputation when the imputation model and/or the nonresponse model are not correctly specified. The second simulation study compares the performance of the imputed estimator obtained by using imputation classes based on the estimated response probabilities and weighted mean imputation (traditional) with the imputed estimator obtained by using the proposed modified deterministic regression imputation method.

3.1 Simulation Study 1

We generated a finite population of size $N = 1{,}000$ containing 3 variables: a variable of interest y and two auxiliary variables z_1 and z_2. To do so, we first generated z_1 and z_2 independently from an exponential distribution with mean 4 and 30 respectively. Then the y–values were generated according to the regression model

$$y_i = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \epsilon_i,$$

where the ϵ_i's are generated from a normal distribution with mean 0 and variance σ^2. The values of the parameters γ_0, γ_1 and γ_2 were respectively set to 20, 2 and 0.1 and the variance σ^2 was chosen to lead to a model R^2–value approximately equal to 0.75. The objective is to estimate the population total $Y = \sum_U y_i$.

We generated $R = 5{,}000$ simple random samples without replacement of size $n = 100$ from the finite population. In each sample, nonresponse to item y was generated according to the following response mechanisms:

Mechanism 1: Response probability p_{1i} for unit i is given by the logistic regression model

$$\log \frac{p_{1i}}{1 - p_{1i}} = \lambda_0 + \lambda_1 z_{1i}.$$

Mechanism 2: Response probability p_{2i} for unit i is given by the logistic regression model

$$\log \frac{p_{2i}}{1 - p_{2i}} = \lambda_0 + \lambda_1 y_i.$$

The values of λ_0 and λ_1 were chosen to give an overall response rate approximately equal to 70%. The response indicators a_{1i} and a_{2i} were generated independently from a Bernoulli distribution with parameters p_{1i} and p_{2i}, respectively. Note that in the case of the nonresponse mechanism 2, the response mechanism is nonignorable in the sense that the probability of response depends on the variable of interest y.

To compensate for the nonresponse to item y, we used the traditional deterministic regression imputation for which the imputed values are given by (6) and the modified deterministic regression imputation for which the imputed values are given by (18). Imputations were based on the models for y and for p listed in Table 1 as $y_{(1)}, y_{(2)}, y_{(3)}, y_{(4)}$ and $p_{(1)}, p_{(2)}, p_{(3)}$. Note that $p_{(1)}$ corresponds to response mechanism 1 and $y_{(1)}$ to the model generating the population.

From each simulated sample, we calculated the imputed estimator $\hat{Y}_I$ given by (2) with the imputed values (6) and (18), based on selected combinations of the models $y_{(a)}$ and $p_{(b)}; a = 1, \ldots, 4; b = 1, 2, 3$. As a measure of the bias of an imputed estimator $\hat{Y}_I$, we used the percent simulated relative bias (RB) given by

$$\mathrm{RB}(\hat{Y}_I) = \frac{\mathrm{Bias}(\hat{Y}_I)}{Y} \times 100, \tag{23}$$

where

$$\mathrm{Bias}(\hat{Y}_I) = \frac{1}{R} \sum_{r=1}^{R} \hat{Y}_I^{(r)} - Y \tag{24}$$

and $\hat{Y}_I^{(r)}$ denotes the value of $\hat{Y}_I$ for the r–th simulated sample. As a measure of variability of an imputed estimator $\hat{Y}_I$, we used the percent simulated relative root mean square error (RRMSE) given by

$$\mathrm{RRMSE}(\hat{Y}_I) = \frac{\sqrt{\mathrm{MSE}(\hat{Y}_I)}}{Y} \times 100, \tag{25}$$

where

$$\mathrm{MSE}(\hat{Y}_I) = \frac{1}{R} \sum_{r=1}^{R} (\hat{Y}_I^{(r)} - Y)^2. \tag{26}$$

Table 1
Models Used for Imputation

Models for y	Intercept	z_1	z_2
$y_{(1)}$	Yes	Yes	Yes
$y_{(2)}$	Yes	No	Yes
$y_{(3)}$	Yes	Yes	No
$y_{(4)}$	No	Yes	Yes
Models for p_i	Intercept	z_1	z_2
$p_{(1)}$	Yes	Yes	No
$p_{(2)}$	Yes	No	Yes
$p_{(3)}$	No	Yes	No

Results on relative bias and RRMSE are shown in Table 2 for the the samples generated by reponse mechanism 1 and in Table 3 for the samples generated by the response mechanism 2. From Table 2, it is clear that, when the imputation is performed according to the correct model (*i.e.*, $y_{(1)}$), traditional deterministic regression imputation leads to an approximately unbiased estimator and it is more efficient than the modified deterministic regression imputation in terms of RRMSE. As noted by a referee, modified deterministic regression imputation can lead to more efficient estimators than traditional deterministic regression. That is, there are scenarios (not considered here) for which the proposed modified deterministic regression imputation method may be more efficient than the traditional deterministic regression imputation method.

When the imputation model is incorrectly specified (*e.g.*, $y_{(2)}$ and $y_{(4)}$), deterministic imputation leads to biased estimators whereas the bias of the modified determinisic imputation is small to negligible, provided the nonresponse model is correctly specified (*i.e.*, $p_{(1)}$). As a result, RRMSE for the deterministic imputation is larger than that for the

modified deterministic regression imputation. When both imputation and nonresponse models are not correctly specified (*e.g.*, $y_{(4)} - p_{(2)}$), all the estimators are biased.

From Table 3, it is clear that, for the case of mechanism 2, the imputed estimator obtained under modified regression imputation performs equally or better than the imputed estimator obtained under traditional regression imputation in all the scenarios. This result is not suprising since achieving an effective bias reduction in the case of nonignorable nonresponse requires the use of all the appropriate auxiliary information available. The auxiliary information used in the case of the proposed modified regression imputation is richer than the one used in the case of regression imputation since it uses the auxiliary variables that are related to both the variable of interest y and the response probability whereas regression imputation uses only the auxiliary variables related to the variable of interest y.

Table 2
Relative Bias (%) and RRMSE (%) of Imputed Estimators Under Response Mechanism 1

Scenario	Bias (traditional)	Bias (proposed)	RRMSE (traditional)	RRMSE (proposed)
$y_{(1)} - p_{(1)}$	0.19	−0.01	1.85	2.33
$y_{(2)} - p_{(1)}$	5.20	0.16	5.60	2.66
$y_{(3)} - p_{(1)}$	0.17	−0.04	1.87	2.37
$y_{(4)} - p_{(1)}$	−14.80	−3.50	15.00	6.70
$y_{(1)} - p_{(2)}$	0.19	0.12	1.85	1.86
$y_{(4)} - p_{(2)}$	−14.80	−14.80	15.00	14.60
$y_{(1)} - p_{(3)}$	0.19	0.05	1.85	1.88

Table 3
Relative Bias (%) and RRMSE (%) of Imputed Estimators Under Response Mechanism 2

Scenario	Bias (traditional)	Bias (proposed)	RRMSE (traditional)	RRMSE (proposed)
$y_{(1)} - p_{(1)}$	1.84	1.83	2.55	2.54
$y_{(2)} - p_{(1)}$	4.46	1.84	4.89	2.65
$y_{(3)} - p_{(1)}$	2.03	2.02	2.70	2.70
$y_{(4)} - p_{(1)}$	−4.58	−3.04	5.07	3.81
$y_{(1)} - p_{(2)}$	1.84	1.84	2.55	2.55
$y_{(4)} - p_{(2)}$	−4.58	−1.70	5.07	2.88
$y_{(1)} - p_{(3)}$	1.84	1.84	2.55	2.55

3.2 Simulation Study 2

We generated a finite population of size $N = 1{,}000$ containing 3 variables: a variable of interest y and three auxiliary variables z_1, z_2 and z_3, by first generating z_1, z_2 and z_3 independently from an exponential distribution with mean 100 and then generating the y–values according to the regression model

$$y_i = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 z_{1i}^2 + \epsilon_i,$$

where the ϵ_i's are generated from a normal distribution with mean 0 and variance σ^2. The values of the parameters $\gamma_0, \gamma_1, \gamma_2$ and γ_3 were respectively fixed to 20, 10, 0.5 and 10. The variance σ^2 was chosen to lead to a model R^2 approximately equal to 0.66. The objective is to estimate the population mean $\bar{Y} = \sum_U y_i / N$. In order to focus on the nonresponse/imputation error, we considered the case of a census, *i.e.*, $n = N = 1{,}000$. From the simulated population, nonresponse to item y was generated according to the following response mechanisms:

Mechanism 1: Response probability p_{1i} for unit i is given by the logistic model

$$\log \frac{p_{1i}}{1 - p_{1i}} = \lambda_0 + \lambda_1 z_{1i} + \lambda_2 z_{3i}.$$

Mechanism 2: Response probability p_{2i} for unit i is given by the logistic model

$$\log \frac{p_{2i}}{1 - p_{2i}} = \lambda_0 + \lambda_1 y_i + \lambda_2 z_{3i}.$$

The values of λ_0, λ_1 and λ_2 were chosen to give an overall response rate approximately equal to 70%. Response indicators a_{1i} and a_{2i} were then genrated independently $R = 1{,}000$ times from a Bernoulli distribution with parameters p_{1i} and p_{2i}, respectively.

To compensate for nonresponse, two strategies were used: The first strategy consisted in dividing the sample, s, into imputation classes $s_1, s_2, \ldots, s_C$ based on the auxiliary variables z_1, z_2 and z_3. To form the classes, we used the score method which may be described as follows: Using the auxiliary information, we first estimated the response probabilites, p_i, to obtain $\hat{p}_i$ for both the respondents and the nonrespondents using logistic regression on z_1, z_2 and z_3. Using the $\hat{p}_i$'s, we then partitioned the population into C classes using the procedure FASTCLUS of SAS (that uses the k–means classification algorithm). The score method leads to a partition of the population in such a way that, within classes, units (respondents and nonrespondents) are homogeneous with respect to $\hat{p}_i$–values. The second strategy used the proposed modified regression imputation method based on the auxiliary variables z_1, z_2 and z_3. The goal of the simulation study is to compare the performances of two imputed estimators of the population mean $\bar{Y}$: (a) Imputed estimator based on the C imputation classes:

$$\bar{y}_I^C = \sum_{c=1}^{C} \frac{\hat{N}_c}{\hat{N}} \bar{y}_{Ic}, \qquad (27)$$

where

$$\bar{y}_{Ic} = \frac{1}{\hat{N}_c}\left[\sum_{s_c} w_i a_i y_i + \sum_{s_c} w_i (1 - a_i) y_i^*\right],$$

and $\hat{N}_c = \sum_{s_c} w_i$. We used weighted mean imputation within classes; *i.e.*, $y_i^* = \sum_{s_c} w_i a_i y_i / \sum_{s_c} w_i a_i$.

(b) Imputed estimator based on the proposed modified regression imputation, denoted $\bar{y}_I$:

$$\bar{y}_I = \frac{1}{\hat{N}}\left[\sum_s w_i a_i y_i + \sum_s w_i (1-a_i) y_i^*\right], \qquad (28)$$

where the imputed values y_i^* are given by (18) using $\mathbf{z}_i' = (z_{1i}, z_{2i})'$ and $\hat{N} = \sum_s w_i$. For mechanism 1, the response probabilities p_i were correctly estimated using the variable z_1 and z_3 whereas the variables z_1, z_2 and z_3 were used to estimate p_i for mechanism 2.

Note that $w_i = 1$ in this simulation study for all $i \in U$ because no sampling is involved. Finally, Table 4 compares these estimators in terms of relative bias, given by (23) and RRMSE, given by (25). From Table 4, it is clear that the proposed imputed estimator (28) performs considerably better than the estimator (27) based on imputation classes in terms of RRMSE for both mechanism 1 and mechanism 2.

Table 4
Relative Bias (%) and RRMSE (%) of Imputed Estimators

Imputed estimator*	Number of classes	RB	RRMSE
$\bar{y}_I^C$ (mechanism 1)	1	14.4	14.5
	5	−0.02	4.26
	10	−0.85	7.33
	20	−0.20	8.61
	30	−0.03	8.61
	40	0.03	9.09
	50	0.06	9.44
$\bar{y}_I$ (mechanism 1)	–	1.11	1.90
$\bar{y}_I^C$ (mechanism 2)	1	29.0	29.1
	5	21.4	21.4
	10	21.0	21.1
	20	20.9	21.0
	30	20.9	21.0
	40	21.0	21.0
	50	21.0	21.0
$\bar{y}_I$ (mechanism 2)	–	10.9	10.9

* $\bar{y}_I^C$ given by (27) and $\bar{y}_I$ given by (28).

4. Variance Estimation

In this section, we derive a variance estimator of the imputed estimator $\hat{Y}_I$, using the reverse approach of Fay (1991). The total variance of $\hat{Y}_I$ under a particular deterministic imputation method, is given by

$$V(\hat{Y}_I - Y) = E_r V_p(\hat{Y}_I - Y \mid \mathbf{a}) + V_r E_p(\hat{Y}_I - Y \mid \mathbf{a}), \qquad (29)$$

where $\mathbf{a} = (a_1, \ldots, a_N)'$ is the vector of response indicators, (Shao and Steel 1999). An estimator of the overall variance $V(\hat{Y}_I - Y)$ in (29) is given by $v_t = v_1 + v_2$, where v_1 is an estimator of $V_p(\hat{Y}_I - Y \mid \mathbf{a})$ conditional on the response indicators a_i, and v_2 is an estimator of $V_r[E_p(\hat{Y}_I - Y \mid \mathbf{a})]$. The estimator v_1 does not depend on the response mechanism or the imputation model, and hence v_1 is valid under either assumption GNM or assumption IM.

Under the corresponding random imputation, the variance of the imputed estimator $\hat{Y}_I$ is given by

$$V(\hat{Y}_I - Y) = E_r V_p E_*(\hat{Y}_I - Y \mid \mathbf{a}) + E_r E_p V_*(\hat{Y}_I - Y \mid \mathbf{a}) + V_r E_p E_*(\hat{Y}_I - Y \mid \mathbf{a}), \qquad (30)$$

where $V_*(.)$ denotes the variance operator with respect to random imputation. We assume that $E_*(\hat{Y}_I \mid \mathbf{a})$ agrees with the imputed estimator for the deterministic case. Hence, $E_r V_p E_*(\hat{Y}_I - Y \mid \mathbf{a})$ is estimated by v_1 for the deterministic case. Similarly, $V_r E_p E_*(\hat{Y}_I - Y \mid \mathbf{a})$ is estimated by v_2 for the deterministic case. The additional contribution to variance due to random imputation comes from the component $E_r E_p V_*(\hat{Y}_I - Y \mid \mathbf{a})$, which is estimated by $v_* = V_*(\hat{Y}_I - Y \mid \mathbf{a})$. Hence, it follows from (30) that the overall variance $V(\hat{Y}_I - Y)$ is estimated by $v_t = v_1 + v_* + v_2$. The term v_* is absent for deterministic imputation.

4.1 Known p_i

In this section, we assume that the reponse probabilites p_i are known. We first consider the case of modified deterministic regression imputation in section 4.1.1. The case of modified random regression imputation is studied in section 4.1.2.

4.1.1 Modified Deterministic Regression Imputation

Under modified deterministic regression imputation, the imputed estimator with known p_i may be written as

$$\hat{Y}_{Ip} = \sum_s w_i a_i y_i + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_\mathbf{r})' \tilde{\boldsymbol{\gamma}}_{\boldsymbol{rp}}, \qquad (31)$$

where

$$\tilde{\boldsymbol{\gamma}}_{\boldsymbol{rp}} = \left[\sum_s w_i a_i \frac{(1-p_i)}{p_i} \mathbf{z}_\mathbf{i} \mathbf{z}_\mathbf{i}' / (\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i})\right]^{-1} \left[\sum_s w_i a_i \frac{(1-p_i)}{p_i} \mathbf{z}_\mathbf{i} y_i / (\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i})\right]. \qquad (32)$$

To obtain v_1, we use standard Taylor linearization which leads to

$$\hat{Y}_{Ip} - Y \approx \sum_s w_i \tilde{\xi}_{ip}, \qquad (33)$$

where

$$\tilde{\xi}_{ip} = a_i y_i + (1-a_i) \mathbf{z}_\mathbf{i}' \tilde{\boldsymbol{\gamma}}_{\boldsymbol{rp}} + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_\mathbf{r})' \tilde{\mathbf{T}}_\mathbf{p}^{-1} a_i \frac{(1-p_i)}{p_i} \frac{1}{(\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i})} \mathbf{z}_\mathbf{i} (y_i - \mathbf{z}_\mathbf{i}' \tilde{\boldsymbol{\gamma}}_{\boldsymbol{rp}})$$

with $\tilde{\mathbf{T}}_\mathbf{p} = \sum_s w_i a_i ((1-p_i)/p_i) \mathbf{z}_\mathbf{i} \mathbf{z}_\mathbf{i}' / (\boldsymbol{\lambda}' \mathbf{z}_\mathbf{i})$. Denoting the variance estimator of the full sample estimator as

$\hat{Y} = \sum_s w_i y_i$ as $v(y)$, it follows from (33) that an estimator of $V_p(\hat{Y}_I - Y \mid \mathbf{a})$ is given by

$$v_1 = v(\tilde{\xi}_p), \tag{34}$$

which is obtained by replacing y_i by $\tilde{\xi}_{ip}$ in the formula for $v(y)$.

To obtain the second component v_2, first note that

$$E_p(\hat{Y}_{Ip} - Y \mid \mathbf{a}) \approx \sum_s a_i y_i + \sum_U (1 - a_i)\boldsymbol{\gamma}_p - Y,$$

where

$$\boldsymbol{\gamma}_p = \left[\sum_U a_i \frac{(1-p_i)}{p_i}\mathbf{z_i}\mathbf{z_i}'/(\boldsymbol{\lambda}'\mathbf{z_i})\right]^{-1} \sum_U a_i \frac{(1-p_i)}{p_i}\mathbf{z_i} y_i/(\boldsymbol{\lambda}'\mathbf{z_i}).$$

Using Taylor linearization, it can be shown that

$$V_r[E_p(\hat{Y}_{Ip} - Y \mid \mathbf{a})] \approx \sum_U p_i(1-p_i)\zeta_i^2, \tag{35}$$

where

$$\zeta_i = \left[1 + \frac{(1-p_i)}{p_i}\frac{1}{(\boldsymbol{\lambda}'\mathbf{z_i})}(\mathbf{Z} - \mathbf{Z_r})'\mathbf{T_p^{-1}}\mathbf{z_i}\right](y_i - \mathbf{z_i}'\boldsymbol{\gamma}_p)$$

with $\mathbf{Z} = \sum_U \mathbf{z_i}$, $\mathbf{Z_r} = \sum_U a_i\mathbf{z_i}$ and $\mathbf{T_p} = \sum_U a_i((1-p_i)/p_i)\mathbf{z_i}\mathbf{z_i}'/(\boldsymbol{\lambda}'\mathbf{z_i})$. The component v_2 is then obtained by estimating the unknown quantities in (35), which leads to

$$v_2 = \sum_s w_i a_i (1-p_i)\hat{\zeta}_i^2, \tag{36}$$

where

$$\hat{\zeta}_i = \left[1 + \frac{(1-p_i)}{p_i}\frac{1}{(\boldsymbol{\lambda}'\mathbf{z_i})}(\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_\mathbf{r})'\tilde{\mathbf{T}}_\mathbf{p}^{-1}\mathbf{z_i}\right](y_i - \mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_{rp}).$$

An estimator of the total variance v_t is obtained as the sum of (34) and (36): $v_t = v_1 + v_2$. In practice, the response probabilities are unknown. As a result, it is not possible to calculate the variance estimator v_t. A simple solution consists in replacing p_i by the estimated response probabilities $\hat{p}_i$ in (34) and (36) and use the resulting v_t as the variance estimator of $\hat{Y}_I$. As we show in a simulation sudy in section 4.3, this simple method gives acceptable results.

4.1.2 Modified Random Regression Imputation

We first note that

$$V_*(y_i^*) = (\boldsymbol{\lambda}'\mathbf{z_i})\sum_s w_j \frac{(1-p_j)}{p_j} a_i(\tilde{e}_j - \tilde{e}_r)^2 \Big/ \sum_s w_j \frac{(1-p_j)}{p_j} a_j \equiv \tilde{s}_e^2$$

and $\mathrm{Cov}_*(y_i^*, y_j^*) = 0, i \neq j$. Hence, from (2) the component v_*, due to random imputation, is given by

$$v_* = \sum_s w_i^2(1-a_i)V_*(y_i^*) = \sum_s w_i^2(1-a_i)\tilde{s}_e^2. \tag{37}$$

An estimator of the total variance is obtained as the sum of (34), (36) and (37): $v_t = v_1 + v_2 + v_*$. Once again, since the response probabilities p_i are unknown, it is not possible to compute v_* in (37). We propose to replace p_i in (37) by the estimated response probabilities $\hat{p}_i$.

4.2 Unknown p_i

We use Binder's method (Binder 1983) to derive the component v_1 when the response probabilities p_i are estimated. We assume that $p_i = f(\mathbf{u_i}'\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is l–vector of unknown parameters, $\mathbf{u_i}$ is a l–vector of auxiliary variables available for all $i \in s$. For example, in the case of logistic regression, $f(\mathbf{u_i}'\boldsymbol{\eta}) = \exp(\mathbf{u_i}'\boldsymbol{\eta})/\exp(1 + \mathbf{u_i}'\boldsymbol{\eta})$. The estimated response probabilities are given by $\hat{p}_i = f(\mathbf{u_i}'\hat{\boldsymbol{\eta}})$, where $\hat{\boldsymbol{\eta}}$ is a consistent estimator of $\boldsymbol{\eta}$. Let $\boldsymbol{\theta} = (\boldsymbol{\eta}_N', \boldsymbol{\gamma}_N', Y)'$, where $\boldsymbol{\eta}_N$ and $\boldsymbol{\gamma}_N$ are census parameter corresponding to $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$, respectively. An estimator of $\boldsymbol{\theta}$ given by $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\eta}}', \tilde{\boldsymbol{\gamma}}_r', \hat{Y}_I)'$ can be expressed as a solution of the sample estimating equations

$$\hat{\mathbf{S}}(\boldsymbol{\theta}) = 0,$$

where $\hat{\mathbf{S}}(\boldsymbol{\theta}) = \left(\hat{\mathbf{S}}_1(\boldsymbol{\theta}), \hat{\mathbf{S}}_2(\boldsymbol{\theta}), \hat{S}_3(\boldsymbol{\theta})\right)'$ with

$$\hat{\mathbf{S}}_\mathbf{1}(\boldsymbol{\theta}) = \sum_s w_i \mathbf{u_i}[a_i - f(\mathbf{u_i}'\boldsymbol{\eta}_N)] = \mathbf{0},$$

$$\hat{\mathbf{S}}_\mathbf{2}(\boldsymbol{\theta}) = \sum_s w_i a_i \mathbf{z_i}\frac{(1 - f(\mathbf{u_i}'\boldsymbol{\eta}_N))}{f(\mathbf{u_i}'\boldsymbol{\eta}_N)}(y_i - \mathbf{z_i}'\boldsymbol{\gamma}_N)/(\boldsymbol{\lambda}'\mathbf{z_i}) = \mathbf{0}$$

and

$$\hat{S}_3(\boldsymbol{\theta}) = Y - \sum_s w_i \mathbf{z_i}'\boldsymbol{\gamma}_N - \sum_s w_i a_i(y_i - \mathbf{z_i}'\boldsymbol{\gamma}_N) = 0.$$

Let $\hat{\mathbf{J}}(\boldsymbol{\theta}) = (\partial\hat{\mathbf{S}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta})$ be the $(k+l+1)\times(k+l+1)$ matrix of partial derivative. We have

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) = [\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta})]\boldsymbol{\Sigma}(\boldsymbol{\theta})[\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta})]',$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ denotes the $(k+l+1)\times(k+l+1)$ symetric matrix whose ij element is the covariance between $\hat{S}_i(\boldsymbol{\theta})$ and $\hat{S}_j(\boldsymbol{\theta})$ with respect to sampling given the vector of response indicator $\mathbf{a}$. If $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is replaced by a consistent estimator $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$, say, we obtain a consistent variance estimator $\mathbf{v}(\hat{\boldsymbol{\theta}})$ given by

$$\mathbf{v}(\hat{\boldsymbol{\theta}}) = [\hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}})]\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})[\hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}})]'.$$

Since we are interested in the variance estimator, v_1, of $\hat{Y}_I$, we need the final row, $\mathbf{b}$, say, of $\hat{\mathbf{J}}^{-1}(\boldsymbol{\theta})$, evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. It follows that

$$v_1 = \mathbf{b}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})\mathbf{b}'. \tag{38}$$

To obtain the component v_2, we assume that the sampling weights w_i satisfy $\max(n/Nw_i) = O(1)$ and that there exists a positive constant C such that $C < p_i$. Furthermore, we assume that $\hat{\boldsymbol{\eta}} - \boldsymbol{\eta} = O_p(n^{-1/2})$. By Taylor linearization, we have

$$\hat{Y}_I = \hat{Y}_{Ip} + (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})\sum_s p_i^{-1}(y_i - \tilde{\boldsymbol{\gamma}}_a)\partial\frac{f(\mathbf{u_i}'\boldsymbol{\eta})}{\partial\boldsymbol{\eta}} + O_p(N/n),$$

where

$$\tilde{\gamma}_a = \left[\sum_U (1-a_i)\mathbf{z_i}\mathbf{z_i}'/(\boldsymbol{\lambda}'\mathbf{z_i})\right]^{-1}\left[\sum_U (1-a_i)\mathbf{z_i}y_i/(\boldsymbol{\lambda}'\mathbf{z_i})\right].$$

Assuming that $f(\mathbf{u}_i'\boldsymbol{\eta})/\partial\boldsymbol{\eta}$ is uniformly bounded, we have

$$E_p(\hat{Y}_I) = E_p(\hat{Y}_{Ip}) + O_p(N/n^{1/2}).$$

Hence, the component $V_r[E_p(\hat{Y}_{Ip} - Y \mid \mathbf{a})]$ is approximately given by (35) and v_2 is given by (36) with p_i replaced by $\hat{p}_i$. In the case of modified random regression imputation, the component due to random imputation will be estimated by (37) with p_i replaced by $\hat{p}_i$.

4.3 Simulation Study

We performed a limited simulation study to assess the performance of the variance estimators considered in sections 4.1 and 4.2. We generated a population of size $N = 2{,}500$ containing two variables y and z. First, the variable z was generated from a Gamma distribution with scale parameter equal to 4 and shape parameter equal to 10. The y–values were then generated according to the ratio model

$$y_i = \gamma z_i + \epsilon_i,$$

where the ϵ_i's are generated from a normal distribution with mean 0 and variance σ^2. The value of the parameter γ was set to 2 and the variance σ^2 was chosen to lead to a model R^2–value approximately equal to 0.81. The objective is to estimate the population total $Y = \sum_U y_i$.

We generated $R = 10{,}000$ simple random samples without replacement from the finite population using the following sampling fractions n/N: 0.05; 0.1 and 0.25. In each sample, nonresponse to item y was generated according to the following response mechanism: Response probability p_i for unit i is given by the logistic model

$$\log\frac{p_i}{1-p_i} = \lambda_0 + \lambda_1 z_i.$$

The values of λ_0 and λ_1 were chosen to give an overall response rate approximately equal to 70%. The response indicators a_i were then generated independently from a Bernoulli distribution with parameters p_i.

To compensate for the nonresponse to item y, we used the modified deterministic ratio imputation for which the imputed values are given by (19). From each simulated sample, we calculated the imputed estimator $\hat{Y}_I$ given by (2) with the imputed values (19). As a measure of the bias of a variance estimator v, we used the relative bias $[E(v) - \mathrm{MSE}(\hat{Y}_I)]/\mathrm{MSE}(\hat{Y}_I)$. Let v_{naive} denotes the total variance estimator obtained by summing (34) and (36) when the response probabilties p_i are replaced by the estimated response probabilities $\hat{p}_i$ and v_{correct} denotes the total variance estimator obtained by summing (38) and (36) with p_i replaced by $\hat{p}_i$. Table 5 gives the relative bias (in %) of the two variance estimators. It is clear from Table 5 that both variance estimators lead to underestimation, but v_{correct} is slightly better in terms of underestimation. Also, both variance estimators performed well with a relative bias less than –10%. Hence, the simpler variance estimator v_{naive} might be suitable in practice.

Table 5
Relative Bias (%) of the Variance Estimators

f	RB(v_{naive})	RB(v_{correct})
0.05	–6.3	–5.1
0.10	–5.8	–4.1
0.25	–4.3	–3.2

5. Estimation of Domain Means

In practice, estimates for various domains (subpopulations) are often needed. For example, in the Canadian Labour Force Survey, estimates of unemployment are required by age-sex group and by industry at the provincial level. To compensate for item nonresponse, the proposed modified regression imputation may be used. However, the domains must be specified in advance at the imputation stage. In other words, the domain indicators must be part of the imputation model. In practice, domains are generally not specified at the edit and imputation stage and domain estimates are obtained from imputed data based on imputation models without the domain indicators. As a result, the imputed estimators for domains are generally biased. We propose a bias-adjusted estimator, along the lines of section 2.2, to remedy this problem. The bias-adjusted estimator can be obtained at the estimation stage and does not require the specification of the domains at the imputation stage.

A vector of domain means may be expressed as

$$\overline{\mathbf{Y}}_{(\mathbf{d})} = \left(\sum_U \mathbf{x_i}\mathbf{x_i}'\right)^{-1}\sum_U \mathbf{x_i}y_i, \qquad (39)$$

where $\mathbf{x} = (x_{1i}, \ldots, x_{di}, \ldots, x_{Di})'$ is a vector of domain indicators, x_{di}, such that $x_{di} = 1$ if $i \in$ domain d and $x_{di} = 0$, otherwise. We assume that $\mathbf{x}$ is known for all the units $i \in s$. In other words, only item y may be missing. In the absence of nonresponse, an approximately unbiased estimator of $\overline{\mathbf{Y}}_{(d)}$ is given by

$$\hat{\overline{\mathbf{Y}}}_{(\mathbf{d})} = \left(\sum_s w_i\mathbf{x_i}\mathbf{x_i}'\right)^{-1}\sum_s w_i\mathbf{x_i}y_i. \qquad (40)$$

In the presence of nonresponse to item y, an imputed estimator of $\overline{\mathbf{Y}}_{(d)}$ is given by

$$\begin{aligned}\hat{\overline{\mathbf{Y}}}_{\mathbf{I(d)}} &= \hat{\mathbf{T}}^{-1}\left[\sum_s w_i a_i\mathbf{x_i}y_i + \sum_s w_i(1-a_i)\mathbf{x_i}y_i^*\right] \\ &= \hat{\mathbf{T}}^{-1}\sum_s w_i a_i\mathbf{x_i}\tilde{y}_i, \qquad (41)\end{aligned}$$

where $\hat{\mathbf{T}} = \sum_s w_i \mathbf{x_i} \mathbf{x_i}'$. Note that the imputed estimator $\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}}$ in (41) does not require the response identifiers, a_i. Haziza and Rao (2005) showed that the imputed estimator $\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}}$ is biased under assumption NM. They proposed a bias-adjusted estimator which is approximately unbiased under either assumption NM or assumption IM. In this section, we propose an extension of the Haziza-Rao bias-adjusted estimator which is approximately unbiased under either assumption GNM or assumption IM.

It is easily seen that, under assumption GNM, the conditional nonresponse bias of the imputed estimator (41) that uses the modified deterministic regression imputation (18) is given by

$$\text{Bias}\left(\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}} \mid s\right) \approx -\hat{\mathbf{T}}^{-1}\left[\sum_s w_i (1-p_i)\,\mathbf{x_i}(y_i - \mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_{s,N})\right], \quad (42)$$

where $\tilde{\boldsymbol{\gamma}}_{s,N}$ is given by (15). An approximately conditionally unbiased estimator of the bias in (42) is given by

$$\hat{B}\left(\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}} \mid s\right) \approx -\hat{\mathbf{T}}^{-1}\left[\sum_s \tilde{w}_i a_i \mathbf{x_i}(y_i - \mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}})\right], \quad (43)$$

where $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}}$ is given by (17). A bias-adjusted estimator, $\hat{\bar{\mathbf{Y}}}^{\mathbf{a}}_{\mathbf{I(d)}}$, is then obtained as $\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}} - \hat{B}(\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}} \mid s)$, which leads to

$$\hat{\bar{\mathbf{Y}}}^{\mathbf{a}}_{\mathbf{I(d)}} = \hat{\mathbf{T}}^{-1}\left[\sum_s \frac{w_i}{\hat{p}_i} a_i \mathbf{x_i}(y_i - \mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}}) + \sum_s w_i \mathbf{x_i}\mathbf{z_i}'\tilde{\boldsymbol{\gamma}}_{\boldsymbol{r}}\right]. \quad (44)$$

The bias-adjusted estimator (44) is approximately unbiased under either IM or GNM. Hence, it is robust in the sense of validity under both assumption IM or assumption GNM. However, it requires both the response identifiers a_i and the estimated response probabilities $\hat{p}_i$, unlike the imputed estimator $\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}}$ in (41).

It is possible to obtain a bias-adjusted estimator of the form (44) if we use the traditional deterministic regression imputation instead. It is interesting to note that the bias-adjusted estimator is identical to the estimator obtained using calibrated imputation (Beaumont 2005). The latter estimator does not require the knowledge of a_i and $\hat{p}_i$ in the imputed data file but the domains must be specified at the imputation stage, which may not be feasible in practice.

If the nonresponse model (4) contains only the intercept, we have $\hat{p}_i = \hat{p}$, where $\hat{p}$ denotes the overall response rate. In this case, the bias-adjusted estimator (44) reduces to

$$\hat{\bar{\mathbf{Y}}}^{\mathbf{a}}_{\mathbf{I(d)}} = \hat{p}^{-1}\hat{\bar{\mathbf{Y}}}_{\mathbf{I(d)}} + (1-\hat{p}^{-1})\hat{\mathbf{T}}^{-1}\sum_s w_i \mathbf{x_i}\mathbf{z_i}'\hat{\boldsymbol{\gamma}}_{\boldsymbol{I}}, \quad (45)$$

noting that $\hat{\boldsymbol{\gamma}}_{\boldsymbol{r}} = \hat{\boldsymbol{\gamma}}_{\boldsymbol{I}}$, where, under deterministic regression imputation,

$$\hat{\boldsymbol{\gamma}}_{\boldsymbol{I}} = \left(\sum_{i\in s} w_i \mathbf{z_i}\mathbf{z_i}'/(\boldsymbol{\lambda}'\mathbf{z_i})\right)^{-1} \times \left[\sum_{i\in s} w_i a_i \mathbf{z_i} y_i/(\boldsymbol{\lambda}'\mathbf{z_i}) + \sum_{i\in s} w_i (1-a_i)\mathbf{z_i} y_i^*/(\boldsymbol{\lambda}'\mathbf{z_i})\right] = \hat{\boldsymbol{\gamma}}_{\boldsymbol{r}}.$$

Haziza and Rao (2005) obtained the bias-adjusted estimator (45).

Concluding Remarks

For simplicity, we focussed on a single imputation class but our GNM method readily extends to multiple imputation classes by using separate imputations across classes. For example, we could use weighted mean imputation within classes using our modified weights $\tilde{w}_i$. Also, our method can be extended to the case of composite imputation (Sitter and Rao 1997; Shao and Steel 1999) which uses different imputations for missing item values depending on the auxiliary information available. For example, ratio imputation is used when an auxiliary variable *x* is observed and some other imputation when x is not observed. In this case, the IM approach based on the ratio model relating y to x will not be applicable unlike in the case where *x* is observed on all the sampled units.

Acknowledgments

J.N.K. Rao's research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The authors wish to thank the reviewers for useful comments and suggestions.

References

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society*, B, 67, 445-458.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 15, 279-292.

Brick, J.M., and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.

Deville, J.C., and Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference, US Bureau of the Census*, 429-440.

Haziza, D., and Rao, J.N.K. (2005). Inference for domains under imputation for missing survey data. *Canadian Journal of Statistics*, 33, 149-161.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Narain, R.D. (1951). On sampling without replacement with variying probabiltities. *Journal of the Indian Society of Agricultural Statistics*, 2, 169-174.

Rao, J.N.K. (1990). Variance estimation under imputation for missing data. Technical report, Statistics Canada, Ottawa.

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

Rao, J.N.K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology*, 31, 117-138.

Rao, J.N.K., and Shao, J. (1992). On variance estimation under imputation for missing data. *Biometrika*, 79, 811-822.

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.

Särndal, C.-E. (1992). Method for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.

Shao, J., and Steel, P. (1999). Variance Estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Sitter, R., and Rao, J.N.K. (1997). Imputation for missing values and corresponding variance estimation. *Canadian Journal of Statistics*, 25, 61-73.