



Catalogue no. 11-522-XIE

**Statistics Canada International Symposium
Series - Proceedings**

**Symposium 2003: Challenges
in Survey Taking for the Next
Decade**

2003



Proceedings of Statistics Canada Symposium 2003
Challenges in Survey Taking for the Next Decade

CALIBRATED IMPUTATION IN SURVEYS UNDER A QUASI-MODEL-ASSISTED APPROACH

Jean-François Beaumont¹

ABSTRACT

In this paper, we propose to use calibrated imputation in the context of the quasi-model-assisted approach. This technique consists of finding final imputed values as close as possible to preliminary imputed values and that are calibrated to satisfy constraints. Using appropriate constraints, we show that the resulting imputed estimator is approximately unbiased for estimation of linear population parameters such as domain totals. We use the Taylor linearization technique of Binder (1983) to obtain a variance estimator under a general nonresponse model. We show that the nonresponse variance can be partitioned into two components: the first component is obtained by assuming that the nonresponse model parameters are known while the second component corresponds to the effect of estimating the nonresponse model parameters. We also discuss the use of calibrated imputation when facing issues such as editing and missing values in the auxiliary variables used to obtain the preliminary imputed values.

KEYWORDS: Estimating Function; Imputation Model; Nonresponse Model; Quasi-model-assisted Approach; Taylor Linearization; Two-phase Sampling.

1. INTRODUCTION

Estimating functions are often used to justify the form of imputed estimators in surveys. For example, linear regression imputation, including its special cases such as ratio imputation, makes use of estimating functions to obtain imputed values. Linear regression imputation is widely used and has indeed been implemented in many imputation systems such as the Generalized Edit and Imputation System (GEIS) developed at Statistics Canada (Bissonnette and Girard, 1998). It is quite natural to justify estimating function imputation by means of a model for the variable of interest, often called an imputation model. It is therefore natural to use the selected imputation model to evaluate the properties (bias, variance, etc.) of the imputed estimator and to make inferences. This approach was proposed by Särndal (1992) in the context of ratio imputation and was extended to more general regression imputation methods and sampling designs by Deville and Särndal (1994). For the Generalized Regression (GREG) estimator, it was studied by Gagnon, Lee, Rancourt and Särndal (1996). In a Bayesian context, this approach was proposed by Rubin (1978) and Rubin (1987, chapter 3).

Even if an imputation model is used for evaluation and/or inference purposes, it is not possible to avoid making some assumptions about the unknown nonresponse mechanism. In other words, it is necessary to postulate a nonresponse model. As in many papers (for example, Särndal, 1992, Deville and Särndal, 1994, and Shao and Steel, 1999), it is common to only assume that the nonresponse mechanism is independent of (or unconfounded with) the random term in the imputation model. This is a sufficient condition to make the nonresponse ignorable with respect to the imputation model. Sometimes, more explicit nonresponse models are used and response probabilities are estimated. This is usually required when it is assumed that the nonresponse mechanism depends directly on the variable of interest.

In this paper, a different approach is considered. The approach can be termed *quasi-model-assisted* in the sense that the properties of the imputed estimator are evaluated with respect to the sampling design and a nonresponse model, and that inferences do not depend on the validity of the imputation model. The role of the imputation model is simply to justify the form of the imputed estimator and to help reduce its nonresponse variance. As in Oh and Scheuren (1983), we use the term “quasi” to reflect the fact that the nonresponse mechanism is not known and that a nonresponse model is required. This quasi-model-assisted approach was also used by Rao and Shao (1992), Rao and Sitter (1995), Fay (1996), Rao (1996), Shao and Sitter (1996) and Shao and Steel (1999), among others, who all assumed a uniform (within imputation

¹ Jean-François Beaumont, Statistics Canada, Household Survey Methods Division, Ottawa, Ontario, Canada, K1A 0T6 (e-mail: Jean-Francois.Beaumont@statcan.ca)

classes or not) nonresponse model. Here, we do not restrict ourselves to uniform nonresponse and consider more general nonresponse models. In practice, this quasi-model-assisted approach is useful when it is difficult to find a satisfactory imputation model while a sufficiently good nonresponse model can be found and validated. In such a case, it is more reasonable to evaluate the properties of the imputed estimator only with respect to the sampling design and the nonresponse model and not to rely on the imputation model to make valid inferences. Of course, this does not mean that no effort should be made to finding the best imputation model possible since the nonresponse variance of the imputed estimator is affected by the choice of an imputation model.

In this paper, we propose to use calibrated imputation to compensate for missing values. This technique consists of finding final imputed values as close as possible to preliminary imputed values, according to some distance function, and that are calibrated to satisfy constraints. Using appropriate constraints, we show that the resulting imputed estimator is approximately unbiased for estimation of linear population parameters and is valid under a quasi-model-assisted approach. Estimating functions, potentially justified by some imputation model, are used to obtain the preliminary imputed values. This idea of modifying imputed values to satisfy constraints was suggested by Mantel, Singh and Yu (1995) in the context of two-phase sampling, although they did not consider any specific distance function. In their paper, they used small area estimation techniques to determine appropriate constraints. Beaumont (2000) also considered calibrated imputation but used different constraints, which restricted the application to the linear imputation model. For categorical variables of interest, calibrated imputation has recently been studied by Favre, Matei and Tillé (2003) and by Liu and Rancourt (2001). It is closely related to reverse calibration proposed in the context of outliers and robust estimation by Ren (2002) and by Beaumont and Alavi (2003). Reverse calibration consists of finding modified values as close as possible to the original values of outliers under constraints motivated by robust estimation techniques. It is, however, slightly different than calibrated imputation since the original values are unknown in the context of nonresponse. The idea of finding imputed values that satisfy constraints can also be found in Deville (2002) although a different approach is considered.

In section 2, we begin by introducing some notation. We restrict our attention to estimation of linear population parameters, such as domain totals. We also assume that the GREG estimator (or any calibration estimator) would be used if there were no nonresponse. Then, in section 3, we describe calibrated imputation in the context of the quasi-model-assisted approach. In section 4, we use the Taylor linearization technique of Binder (1983) to obtain a variance estimator under a general nonresponse model. We show that the nonresponse variance can be partitioned into two components: the first component is obtained by assuming that the nonresponse model parameters are known while the second component corresponds to the effect of estimating the nonresponse model parameters. In section 5, we discuss the use of calibrated imputation when facing issues such as editing and missing values in the auxiliary variables used to obtain the preliminary imputed values. Finally, we conclude with a brief summary and discussion in the last section.

2. BACKGROUND AND NOTATION

Let us assume that it is desired to estimate a linear population parameter $t_{dy} = \sum_{k \in P} d_k y_k$ for a given population P , where y_k is the value of a variable of interest y for population unit k and d_k is the value of some variable d for population unit k . In practice, d is often a domain indicator variable; that is, $d_k = 1$ if unit k belongs to a domain of interest and $d_k = 0$ otherwise. Since it is usually not possible to survey all population units, a random sample s is selected according to some sampling design $p(s)$. The selection probability for unit k is denoted by π_k , the joint selection probability of units k and l is denoted by π_{kl} and expectations under the sampling design are denoted by $E_p(\cdot)$. We also assume that a vector of auxiliary variables \mathbf{x}_1 is available for all sample units and that population totals, $\mathbf{t}_{\mathbf{x}_1} = \sum_{k \in P} \mathbf{x}_{1k}$, are known for these variables. In the absence of nonresponse, it would be possible to use a calibration estimator $\hat{t}_{dy} = \sum_{k \in s} \tilde{w}_k d_k y_k$ of t_{dy} , where \tilde{w}_k , for $k \in s$, are estimation weights (see Deville and Särndal, 1992). The Generalized Regression (GREG) estimator is an important special case of calibration estimators with the estimation weights given by $\tilde{w}_k = w_k g_{1k}$, where $w_k = 1/\pi_k$ and

$$g_{1k} = 1 + \frac{\mathbf{x}'_{1k}}{v_{1k}} \left(\sum_{k \in s} \frac{w_k}{v_{1k}} \mathbf{x}_{1k} \mathbf{x}'_{1k} \right)^{-1} \left(\mathbf{t}_{x_1} - \sum_{k \in s} w_k \mathbf{x}_{1k} \right). \quad (2.1)$$

The constant v_{1k} corresponds to the variance structure of the estimation model underlying the GREG estimator. Using a first-order Taylor approximation (see, for example, Särndal, Swensson and Wretman, 1992, chapter 6), it can be shown that the GREG estimator is approximately p -unbiased; that is, $E_p(\hat{t}_{dy}) \approx t_{dy}$, and that its variance can be estimated by

$$\hat{V}_p(\hat{t}_{dy}) = \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl} \pi_k \pi_l} e_{1k} e_{1l}, \quad (2.2)$$

where $e_{1k} = d_k y_k - \mathbf{x}'_{1k} \hat{\mathbf{B}}_1$ and

$$\hat{\mathbf{B}}_1 = \left(\sum_{k \in s} \frac{w_k}{v_{1k}} \mathbf{x}_{1k} \mathbf{x}'_{1k} \right)^{-1} \left(\sum_{k \in s} \frac{w_k}{v_{1k}} \mathbf{x}_{1k} d_k y_k \right).$$

In almost all surveys, there are missing values and the variable y is only observed for part of the sample s . A random set of respondents s_r is observed according to a nonresponse mechanism $q(s_r | s)$. Nonresponse can be viewed as a second phase of selection with the difference that the nonresponse mechanism is not controlled by the survey sampler and is unknown. Therefore, it is necessary to postulate a nonresponse model. We assume that the unknown nonresponse mechanism depends on a vector of auxiliary variables \mathbf{z} , observed for every sample unit, and on a vector of unknown parameters $\boldsymbol{\alpha}$. In other words, the nonresponse mechanism can be expressed as $q(s_r | s, \{\mathbf{z}_k; k \in s\}; \boldsymbol{\alpha})$. The response probability of sample unit k and the joint response probability of two different sample units k and l are denoted by $p_k(\boldsymbol{\alpha})$ and $p_{kl}(\boldsymbol{\alpha})$ respectively. Expectations under the nonresponse model are denoted by $E_q(\cdot | s)$ and expectations under the sampling design and the nonresponse model are denoted by $E_{pq}(\cdot)$. We also assume that an estimator $\boldsymbol{\alpha}^*$ of $\boldsymbol{\alpha}$ is obtained by using a q -unbiased estimating function $\mathbf{U}_1(\cdot)$; that is, $\mathbf{E}_q(\mathbf{U}_1(\boldsymbol{\alpha}) | s) = \mathbf{0}$. The estimator $\boldsymbol{\alpha}^*$ is thus implicitly defined by the equation $\mathbf{U}_1(\boldsymbol{\alpha}^*) = \mathbf{0}$. The response probability $p_k(\boldsymbol{\alpha})$ and the joint response probability $p_{kl}(\boldsymbol{\alpha})$ are estimated by $p_k(\boldsymbol{\alpha}^*)$ and $p_{kl}(\boldsymbol{\alpha}^*)$ respectively.

In the presence of nonresponse, \hat{t}_{dy} cannot be computed since the variable y is not observed for the random set of nonrespondents s_m . However, we assume that the variable d is not subject to nonresponse and is observed for every sample unit. A common remedy to the problem of nonresponse consists of imputing the missing y -values. This leads to the imputed estimator $\hat{t}_{dy}^I = \sum_{k \in s} \tilde{w}_k d_k y_{\bullet k}$, where

$$y_{\bullet k} = \begin{cases} y_k & \text{if } k \in s_r, \\ y_k^* & \text{otherwise,} \end{cases} \quad (2.3)$$

and y_k^* is the final imputed value for unit k .

3. CALIBRATED IMPUTATION

Before finding final imputed values using calibrated imputation, preliminary imputed values $\hat{\mu}_k^*$ must first be determined. In this paper, we use deterministic single imputation to find preliminary imputed values. In particular, we focus on

estimating function imputation. With this imputation method, preliminary imputed values $\hat{\mu}_k^*$ depend on a vector of auxiliary variables \mathbf{x} , observed for every sample unit, and on a vector of estimated parameters $\hat{\mathbf{B}}^*$; that is, $\hat{\mu}_k^* = h(\mathbf{x}_k; \hat{\mathbf{B}}^*)$, for some known function $h(\cdot; \cdot)$. The vector \mathbf{x} may contain some auxiliary variables included in the estimation and nonresponse models and/or other variables. The requirement that \mathbf{x} be observed for every sample unit is relaxed in section 5.1. The vector $\hat{\mathbf{B}}^*$ is implicitly defined by the equation $\mathbf{U}_2(\mathbf{a}^*, \hat{\mathbf{B}}^*) = \mathbf{0}$, where $\mathbf{U}_2(\cdot, \cdot)$, is a q -unbiased estimating function for $\hat{\mathbf{B}}$ in the sense that there exists a vector $\hat{\mathbf{B}}$ such that $\mathbf{E}_q(\mathbf{U}_2(\mathbf{a}, \hat{\mathbf{B}}) | s) = \mathbf{0}$. It is usually reasonable to justify the choice of the auxiliary variables \mathbf{x} , the estimating function $\mathbf{U}_2(\cdot, \cdot)$ and the function $h(\cdot; \cdot)$ by an imputation model m such that $E_m(y_k | \mathbf{x}_k) = h(\mathbf{x}_k; \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a vector of unknown model parameters. However, this is not a requirement of our method to obtain an approximately q -unbiased imputed estimator, and \mathbf{x} , $\mathbf{U}_2(\cdot, \cdot)$ and $h(\cdot; \cdot)$ can be chosen quite arbitrarily. Nevertheless, these choices should not be left at random since they have an impact on the nonresponse variance of the imputed estimator, as shown in section 4.

In practice, we are often interested in estimating not only one population parameter but a vector of population parameters $\mathbf{t}_{dy} = \sum_{k \in P} \ddot{\mathbf{d}}_k y_k$, where $\ddot{\mathbf{d}}_k$ is a vector of values for population unit k . The vector $\ddot{\mathbf{d}}_k$ is not subject to missing values and, d_k and t_{dy} are elements of $\ddot{\mathbf{d}}_k$ and \mathbf{t}_{dy} respectively. Once the vectors \mathbf{a}^* and $\hat{\mathbf{B}}^*$ are obtained, missing values can then be imputed and \mathbf{t}_{dy} can be estimated. It is quite common and natural to define the final imputed values by $y_k^* = \hat{\mu}_k^*$. However, this does not necessarily lead to an approximately pq -unbiased imputed estimator for the vector of population parameters \mathbf{t}_{dy} . To make the nonresponse bias negligible, we propose to use calibrated imputation. Under the quasi-model-assisted approach, this technique consists of finding final imputed values y_k^* , for $k \in s_m$, as close as possible to the preliminary imputed values $\hat{\mu}_k^*$ and that are calibrated to satisfy the constraint $\hat{\mathbf{t}}_{dy}^I = \hat{\mathbf{t}}_{dy}^*$, where $\hat{\mathbf{t}}_{dy}^I = \sum_{k \in s} \tilde{w}_k \ddot{\mathbf{d}}_k y_k^*$ and

$$\hat{\mathbf{t}}_{dy}^* = \sum_{k \in s} \tilde{w}_k \ddot{\mathbf{d}}_k \hat{\mu}_k^* + \sum_{k \in s_r} \frac{\tilde{w}_k}{p_k(\mathbf{a}^*)} \ddot{\mathbf{d}}_k (y_k - \hat{\mu}_k^*) . \tag{3.1}$$

The goal of the constraint is to obtain an approximately q -unbiased imputed estimator. This can be seen by noting that the estimator (3.1) has the form of a (nonlinear) GREG estimator. It is also easy to see from (3.1) that the imputed estimator $\hat{\mathbf{t}}_{dy}^I$ is implicitly defined by the equation

$$\mathbf{U}_3(\mathbf{a}^*, \hat{\mathbf{B}}^*, \hat{\mathbf{t}}_{dy}^I) = \hat{\mathbf{t}}_{dy}^I - \sum_{k \in s} \tilde{w}_k d_k \hat{\mu}_k^* - \sum_{k \in s_r} \frac{\tilde{w}_k}{p_k(\mathbf{a}^*)} d_k (y_k - \hat{\mu}_k^*) = 0 , \tag{3.2}$$

where $\mathbf{U}_3(\cdot, \cdot, \cdot)$ is a q -unbiased estimating function for $\hat{\mathbf{t}}_{dy}$; that is, $\mathbf{E}_q(\mathbf{U}_3(\mathbf{a}, \hat{\mathbf{B}}, \hat{\mathbf{t}}_{dy}) | s) = \mathbf{0}$. This last equality is true provided that the following assumption holds:

- A1) The nonresponse mechanism is independent of \tilde{w}_k , $\hat{\mu}_k = h(\mathbf{x}_k; \hat{\mathbf{B}})$, y_k and d_k , for $k \in s$, after conditioning on s and on \mathbf{z}_k , for $k \in s$. As a result, all quantities in $\mathbf{U}_3(\mathbf{a}, \hat{\mathbf{B}}, \hat{\mathbf{t}}_{dy})$ can be treated as fixed, except the random set s_r .

Therefore, a careful choice of the auxiliary variables \mathbf{z} to be included in the nonresponse model is critical to make this assumption reasonably satisfied.

More explicitly, we want to find the final imputed values y_k^* , for $k \in s_m$, that minimize the distance function

$$\sum_{k \in s} u_k (y_k - \hat{\mu}_k^*)^2 \tag{3.3}$$

subject to the constraint $\hat{\mathbf{t}}_{dy}^I = \hat{\mathbf{t}}_{dy}^*$. The quantity u_k is a weight that can be calculated for every sample unit (or at least for the nonrespondents). If an imputation model is used to justify the form of the preliminary imputed values then a natural choice would be to take $u_k = \tilde{w}_k / \hat{\sigma}_k^2$, where $\hat{\sigma}_k^2$ is an estimate of the model variance $V_m(y_k | \mathbf{x}_k)$. Other distance functions than (3.3) could be considered to determine how close final imputed values are to preliminary imputed values. We chose the distance function (3.3) since it is natural in the context of generalized least-squares estimation and it makes the derivation of imputed values easier. Using the method of Lagrange multipliers and some straightforward algebra, it can be shown that the imputed values that minimize (3.3) subject to $\hat{\mathbf{t}}_{dy}^I = \hat{\mathbf{t}}_{dy}^*$ are given by

$$y_k^* = \hat{\mu}_k^* + \frac{\tilde{w}_k}{u_k} (\mathbf{d}_k)' \left(\sum_{k \in s_m} \frac{\tilde{w}_k^2}{u_k} (\mathbf{d}_k)(\mathbf{d}_k)' \right)^{-1} \sum_{k \in s_r} \tilde{w}_k \frac{(1 - p_k(\mathbf{a}^*))}{p_k(\mathbf{a}^*)} (y_k - \hat{\mu}_k^*) \mathbf{d}_k \tag{3.4}$$

Therefore, the final imputed values are obtained by adding a nonresponse bias adjustment term to the preliminary imputed values $\hat{\mu}_k^*$. This adjustment term vanishes and $y_k^* = \hat{\mu}_k^*$ when the second sum in (3.4) is equal to 0. This is the case if, for example, the following two conditions are satisfied:

- i) $\mathbf{U}_2(\mathbf{a}^*, \hat{\mathbf{B}}^*) = \sum_{k \in s_r} a_k (y_k - \hat{\mu}_k^*) \frac{\mathbf{x}_k}{v_k} = \mathbf{0}$, where $a_k = \tilde{w}_k (1 - p_k(\mathbf{a}^*)) / p_k(\mathbf{a}^*)$ is a regression weight and v_k is a known constant potentially justified by an assumption about the model variance $V_m(y_k | \mathbf{x}_k)$;
- ii) $\ddot{\mathbf{d}}_k v_k = \Lambda \mathbf{x}_k$, where Λ is a matrix of known constants.

Condition (ii) is satisfied by including $\ddot{\mathbf{d}}_k v_k$ in the vector \mathbf{x}_k . However, this might not always be appropriate to do when the number of population parameters to be estimated is very large or, in other words, when the dimension of \mathbf{d}_k is very large. Typically, to avoid instability of the estimator $\hat{\mathbf{B}}^*$, the dimension of \mathbf{x}_k is not too large and only the variables that are sufficiently correlated with y are included in the vector \mathbf{x} . As a result, the dimension of \mathbf{d}_k is likely to be larger than the dimension of \mathbf{x}_k in practice. It is therefore not always appropriate to ignore the nonresponse bias adjustment term.

The estimator $\hat{\boldsymbol{\theta}}^* = (\mathbf{a}^*)', (\hat{\mathbf{B}}^*)', \hat{t}_{dy}^I$ is implicitly defined by the equation

$$\mathbf{U}(\hat{\boldsymbol{\theta}}^*) = \begin{pmatrix} \mathbf{U}_1(\mathbf{a}^*) \\ \mathbf{U}_2(\mathbf{a}^*, \hat{\mathbf{B}}^*) \\ \mathbf{U}_3(\mathbf{a}^*, \hat{\mathbf{B}}^*, \hat{t}_{dy}^I) \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ 0 \end{pmatrix}, \tag{3.5}$$

where the estimating function $\mathbf{U}(\cdot)$ is q -unbiased for $\hat{\boldsymbol{\theta}} = (\mathbf{a}', \hat{\mathbf{B}}', \hat{t}_{dy}^I)'$; that is $\mathbf{E}_q(\mathbf{U}(\hat{\boldsymbol{\theta}}) | s) = \mathbf{0}$. Using a first-order Taylor approximation (see Binder, 1983), it can be easily shown that the estimator $\hat{\boldsymbol{\theta}}^*$ is approximately q -unbiased for $\hat{\boldsymbol{\theta}}$ and that, in particular, the imputed estimator \hat{t}_{dy}^I is approximately q -unbiased for the full response calibration estimator \hat{t}_{dy} ; that is, $\mathbf{E}_q(\hat{t}_{dy}^I | s) \approx \hat{t}_{dy}$. The main difference with Binder (1983) is that, here, expectations are evaluated with respect to

the nonresponse model while, in the Binder paper, they are evaluated with respect to the sampling design. Since $E_q(\hat{t}_{dy}^I | s) \approx \hat{t}_{dy}$, \hat{t}_{dy}^I is also approximately pq -unbiased for the population parameter t_{dy} ; that is, $E_{pq}(\hat{t}_{dy}^I) \approx t_{dy}$. This property of approximate q -unbiasedness of the imputed estimator \hat{t}_{dy}^I holds no matter the validity of the imputation model (if specified). Consequently, the role of the imputation model is simply to provide a way to improve the efficiency of the imputed estimator, without relying on its adequacy to achieve approximate q -unbiasedness. However, the nonresponse model must hold for this quasi-model-assisted approach to be valid.

4. VARIANCE ESTIMATION

The analogy with two-phase sampling (see, for example, Hidiroglou and Särndal, 1998) can be used to estimate the variance of the imputed estimator \hat{t}_{dy}^I . Using the fact that $E_q(\hat{t}_{dy}^I | s) \approx \hat{t}_{dy}$, the pq -variance of \hat{t}_{dy}^I can be approximated by:

$$V_{pq}(\hat{t}_{dy}^I) \approx V_p(\hat{t}_{dy}) + E_p V_q(\hat{t}_{dy}^I | s). \tag{4.1}$$

The first term of the right-hand side of (4.1), $V_p(\hat{t}_{dy})$, is often called the sampling variance while the second term, $E_p V_q(\hat{t}_{dy}^I | s)$, is called the nonresponse variance.

In the case of complete response, the sampling variance can be estimated by (2.2). However, (2.2) cannot be computed in the presence of nonresponse since y_k is not observed for the nonresponding units $k \in s_m$. Using the analogy with two-phase sampling, the sampling variance can nevertheless be estimated by

$$\hat{V}_p^*(\hat{t}_{dy}) = \sum_{k \in s_r} \sum_{l \in s_r} \frac{1}{p_{kl}(\mathbf{a}^*)} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl} \pi_k \pi_l} e_{1k}^* e_{1l}^*, \tag{4.2}$$

where $e_{1k}^* = d_k y_k - \mathbf{x}'_{1k} \hat{\mathbf{B}}_1^*$ and

$$\hat{\mathbf{B}}_1^* = \left(\sum_{k \in s_r} \frac{w_k}{p_k(\mathbf{a}^*)} \frac{1}{v_{1k}} \mathbf{x}_{1k} \mathbf{x}'_{1k} \right)^{-1} \left(\sum_{k \in s_r} \frac{w_k}{p_k(\mathbf{a}^*)} \frac{1}{v_{1k}} \mathbf{x}_{1k} d_k y_k \right).$$

An approximately pq -unbiased estimator of the nonresponse variance $E_p V_q(\hat{t}_{dy}^I | s)$ can be obtained by finding an approximately q -unbiased estimator for the conditional nonresponse variance $V_q(\hat{t}_{dy}^I | s)$. Using the Taylor linearization technique of Binder (1983), and replacing the expectations with respect to the sampling design by expectations with respect to the nonresponse model, we have that $V_q(\hat{\theta}^* | s)$ can be approximated by $\dot{V}_q(\hat{\theta}^* | s)$, where

$$\dot{V}_q(\hat{\theta}^* | s) = [\mathbf{H}(\hat{\theta})]^{-1} \Sigma(\hat{\theta}) [\mathbf{H}'(\hat{\theta})]^{-1}, \tag{4.3}$$

and where

$$\mathbf{H}(\tilde{\theta}) = \mathbf{E}_q \left(\frac{\partial \mathbf{U}(\tilde{\theta})}{\partial \tilde{\theta}'} \mid s \right), \text{ for some vector } \tilde{\theta} = (\tilde{\mathbf{a}}', \tilde{\mathbf{B}}', \tilde{t}_{dy})', \text{ and } \Sigma(\hat{\theta}) = \mathbf{E}_q (\mathbf{U}(\hat{\theta}) \mathbf{U}'(\hat{\theta}) \mid s).$$

Therefore, the conditional nonresponse variance $V_q(\hat{t}_{dy}^I | s)$ is approximated by the value in the last row and in the last column of the matrix defined in the right-hand side of (4.3).

To obtain a more explicit expression for $V_q(\hat{t}_{dy}^I | s)$, let us first denote the imputed estimator \hat{t}_{dy}^I by $\hat{t}_{dy}^I(\mathbf{a}^*)$ to emphasize

the fact that the imputed estimator depends on a vector of estimated nonresponse model parameters \mathbf{a}^* . The imputed estimator that would be obtained if the nonresponse model parameters were known is therefore denoted by $\hat{t}_{dy}^I(\mathbf{a})$. Then, after straightforward but somewhat tedious algebra, it can be shown (see Beaumont, 2004) that the approximate variance $\dot{V}_q(\hat{t}_{dy}^I(\mathbf{a}^*)|s)$ for $V_q(\hat{t}_{dy}^I(\mathbf{a}^*)|s)$ is given by

$$\dot{V}_q(\hat{t}_{dy}^I(\mathbf{a}^*)|s) = \dot{V}_q(\hat{t}_{dy}^I(\mathbf{a})|s) + \left\{ -\mathbf{H}_{31}(\hat{\theta})\dot{V}_q(\mathbf{a}^*|s)\mathbf{H}'_{31}(\hat{\theta}) - 2\mathbf{H}_{31}(\hat{\theta})\dot{\mathbf{C}}_q(\mathbf{a}^*, \hat{t}_{dy}^I|s) \right\}, \quad (4.4)$$

where

$$\begin{aligned} \dot{V}_q(\hat{t}_{dy}^I(\mathbf{a})|s) &= V_q(\mathbf{U}_3(\mathbf{a}, \hat{\mathbf{B}}, \hat{t}_{dy})|s) = \sum_{k \in s} \sum_{l \in s} \frac{p_{kl}(\mathbf{a}) - p_k(\mathbf{a})p_l(\mathbf{a})}{p_k(\mathbf{a})p_l(\mathbf{a})} [\tilde{w}_k(y_k - \hat{\mu}_k)d_k] [\tilde{w}_l(y_l - \hat{\mu}_l)d_l], \\ \dot{V}_q(\mathbf{a}^*|s) &= [\mathbf{H}_{11}(\hat{\theta})]^{-1} \mathbf{E}_q(\mathbf{U}_1(\mathbf{a})\mathbf{U}'_1(\mathbf{a})|s) [\mathbf{H}_{11}(\hat{\theta})]^{-1}, \\ \dot{\mathbf{C}}_q(\mathbf{a}^*, \hat{t}_{dy}^I|s) &= -\dot{V}_q(\mathbf{a}^*|s)\mathbf{H}'_{31}(\hat{\theta}) + \mathbf{H}_{11}^{-1}(\hat{\theta})\mathbf{E}_q(\mathbf{U}_1(\mathbf{a})\mathbf{U}_3(\mathbf{a}, \hat{\mathbf{B}}, \hat{t}_{dy})|s), \\ \mathbf{H}_{31}(\hat{\theta}) &= \mathbf{E}_q \left(\left. \frac{\partial \mathbf{U}_3(\tilde{\mathbf{a}}, \tilde{\mathbf{B}}, \tilde{t}_{dy})}{\partial \tilde{\mathbf{a}}'} \right| s \right) \Bigg|_{\tilde{\theta} = \hat{\theta}} = \sum_{k \in s} \frac{\tilde{w}_k}{p_k(\mathbf{a})} (y_k - \hat{\mu}_k) d_k \left(\left. \frac{\partial p_k(\tilde{\mathbf{a}})}{\partial \tilde{\mathbf{a}}'} \right|_{\tilde{\mathbf{a}} = \mathbf{a}} \right), \\ \mathbf{H}_{11}(\tilde{\theta}) &= \mathbf{E}_q \left(\left. \frac{\partial \mathbf{U}_1(\tilde{\mathbf{a}})}{\partial \tilde{\mathbf{a}}'} \right| s \right). \end{aligned}$$

As a result, the approximate conditional nonresponse variance $\dot{V}_q(\hat{t}_{dy}^I(\mathbf{a}^*)|s)$ can be partitioned into two components. The first component, $\dot{V}_q(\hat{t}_{dy}^I(\mathbf{a})|s)$, corresponds to the approximate conditional nonresponse variance that would be obtained if the nonresponse model parameters were known while the second component, the part within brackets in (4.4), corresponds to the effect of estimating the nonresponse model parameters. This effect can be either positive or negative. It can be shown that it is negative for the logistic nonresponse model (for details, see Beaumont, 2004).

In practice, it is quite tempting to assume that \mathbf{a}^* is sufficiently close to \mathbf{a} to approximate the conditional nonresponse variance $V_q(\hat{t}_{dy}^I(\mathbf{a}^*)|s)$ by the first component of (4.4), $\dot{V}_q(\hat{t}_{dy}^I(\mathbf{a})|s)$. This approximation greatly simplifies variance estimation. It was indeed used by Beaumont and Mitchell (2002) in the development of the System for Estimation of Variance due to Nonresponse and Imputation (SEVANI). In the context of bootstrap variance estimation, Mantel, Nadon and Yeo (2000) found empirically that treating \mathbf{a}^* as being fixed often leads to slightly larger variance estimates than those obtained without the approximation. From (4.4), we can see that this approximation is appropriate when $\mathbf{H}_{31}(\hat{\theta}) = \mathbf{0}$. This last equation is verified if all sample units respond with the same probability and if conditions (i) and (ii) of section 3 are satisfied. However, the second component of (4.4) may not always be negligible, especially if the imputation model m is not properly specified. In such a case, not only is $\mathbf{H}_{31}(\hat{\theta})$ not likely to be negligible but also $\mathbf{E}_m \left\{ \mathbf{H}_{31}(\hat{\theta}) | s \right\}$. In a simulation study, Beaumont (2004) shows that neglecting the second component of (4.5) leads to reasonably conservative inferences.

It is interesting to note that equation (4.4) does not depend on the estimating function $\mathbf{U}_2(\dots)$. Consequently, the function $\mathbf{U}_2(\dots)$ can be quite complex without adding any complexity in the expression for $\dot{V}_q(\hat{t}_{dy}^I(\mathbf{a}^*)|s)$. Also, we can see that the approximate variance $\dot{V}_q(\hat{t}_{dy}^I(\mathbf{a}^*)|s)$ depends on the residuals $e_k = y_k - \hat{\mu}_k$. It is therefore important to carefully choose the vector of auxiliary variables \mathbf{x} , the estimating function $\mathbf{U}_2(\dots)$ and the function $h(\dots)$ in such a way that the residuals e_k are kept as small as possible. This can be done by finding the most appropriate imputation model possible.

Note, however, that the validity of variance expression (4.4) does not depend on the validity of the imputation model.

A variance estimator for $V_q(\hat{t}_{dy}^I(\mathbf{a}^*) | s)$ can be obtained by estimating the unknown quantities in (4.4) and is given by

$$\hat{V}_q(\hat{t}_{dy}^I(\mathbf{a}^*) | s) = \hat{V}_q(\hat{t}_{dy}^I(\mathbf{a}) | s) + \left\{ -\hat{\mathbf{H}}_{31}(\hat{\boldsymbol{\theta}}^*) \hat{V}_q(\mathbf{a}^* | s) \hat{\mathbf{H}}'_{31}(\hat{\boldsymbol{\theta}}^*) - 2 \hat{\mathbf{H}}_{31}(\hat{\boldsymbol{\theta}}^*) \hat{\mathbf{C}}_q(\mathbf{a}^*, \hat{t}_{dy}^I | s) \right\}, \quad (4.5)$$

where

$$\begin{aligned} \hat{V}_q(\hat{t}_{dy}^I(\mathbf{a}) | s) &= \sum_{k \in s_r} \sum_{l \in s_r} \frac{p_{kl}(\mathbf{a}^*) - p_k(\mathbf{a}^*) p_l(\mathbf{a}^*)}{p_{kl}(\mathbf{a}^*) p_k(\mathbf{a}^*) p_l(\mathbf{a}^*)} [\tilde{w}_k (y_k - \hat{\mu}_k^*) d_k] [\tilde{w}_l (y_l - \hat{\mu}_l^*) d_l], \\ \hat{V}_q(\mathbf{a}^* | s) &= [\hat{\mathbf{H}}_{11}(\hat{\boldsymbol{\theta}}^*)]^{-1} \hat{\boldsymbol{\Sigma}}_{11}(\hat{\boldsymbol{\theta}}^*) [\hat{\mathbf{H}}'_{11}(\hat{\boldsymbol{\theta}}^*)]^{-1}, \\ \hat{\mathbf{C}}_q(\mathbf{a}^*, \hat{t}_{dy}^I | s) &= -\hat{V}_q(\mathbf{a}^* | s) \hat{\mathbf{H}}'_{31}(\hat{\boldsymbol{\theta}}^*) + \hat{\mathbf{H}}_{11}^{-1}(\hat{\boldsymbol{\theta}}^*) \hat{\boldsymbol{\Sigma}}_{13}(\hat{\boldsymbol{\theta}}^*), \\ \hat{\mathbf{H}}_{31}(\hat{\boldsymbol{\theta}}^*) &= \sum_{k \in s_r} \frac{\tilde{w}_k}{p_k^2(\mathbf{a}^*)} (y_k - \hat{\mu}_k^*) d_k \left(\frac{\partial p_k(\tilde{\mathbf{a}})}{\partial \tilde{\mathbf{a}}'} \bigg|_{\tilde{\mathbf{a}} = \mathbf{a}^*} \right), \\ \hat{\mathbf{H}}_{11}(\hat{\boldsymbol{\theta}}^*) &= \frac{\partial \mathbf{U}_1(\tilde{\mathbf{a}})}{\partial \tilde{\mathbf{a}}'} \bigg|_{\tilde{\mathbf{a}} = \mathbf{a}^*}. \end{aligned}$$

and where $\hat{\boldsymbol{\Sigma}}_{11}(\hat{\boldsymbol{\theta}}^*)$ and $\hat{\boldsymbol{\Sigma}}_{13}(\hat{\boldsymbol{\theta}}^*)$ are estimators for $\mathbf{E}_q(\mathbf{U}_1(\mathbf{a}) \mathbf{U}'_1(\mathbf{a}) | s)$ and $\mathbf{E}_q(\mathbf{U}_1(\mathbf{a}) \mathbf{U}_3(\mathbf{a}, \hat{\mathbf{B}}, \hat{t}_{dy}) | s)$ respectively. The total variance $V_{pq}(\hat{t}_{dy}^I(\mathbf{a}^*))$ can then be estimated by adding the sampling variance estimator (4.2) to the nonresponse variance estimator (4.5).

5. SOME ISSUES

5.1 Missing values in the x-variables

So far, we have assumed that the auxiliary variables used to obtain the preliminary imputed values were not subject to missing values. However, it is common practice to use some variables of interest to impute others. In such a case, we are in a situation where the vector of auxiliary variables \mathbf{x}_k is potentially not fully observed for some sample units and more than one imputation method might be used. For example, if we have two variables of interest y and x , we might want to impute the missing y -value for a given unit k using a linear regression model between y and x , if x_k is known, and using mean imputation, otherwise. Therefore, the preliminary imputed value for a given unit k , $\hat{\mu}_k^*$, depends on the nonresponse pattern $\mathbf{r}_k^{(x)}$, where the j^{th} component of $\mathbf{r}_k^{(x)}$ is equal to 0 or 1 depending on whether the j^{th} component of \mathbf{x}_k is missing or not respectively; that is, $\hat{\mu}_k^* = h(\mathbf{x}_k, \mathbf{r}_k^{(x)}; \hat{\mathbf{B}}^*)$. Of course, $\hat{\mu}_k^*$ cannot depend on the unobserved values of the vector \mathbf{x}_k . The estimating function $\mathbf{U}_2(\cdot, \cdot)$ can become quite complex when many different imputation methods are used. Fortunately, this does not add any complexity to variance estimation since the variance estimator (4.5) does not depend on this estimating function. As a result, the theory presented in sections 3 and 4 can be applied directly when we face the problem of missing values in the \mathbf{x} -variables. Note that assumption (A1) in section 3 must still hold, with $\hat{\mu}_k = h(\mathbf{x}_k; \hat{\mathbf{B}})$ replaced by $\hat{\mu}_k = h(\mathbf{x}_k, \mathbf{r}_k^{(x)}; \hat{\mathbf{B}})$. Therefore, it is important to consider including $\mathbf{r}^{(x)}$ in the vector of auxiliary variables \mathbf{z} to make this assumption reasonably satisfied since $\mathbf{r}_k^{(x)}$, for $k \in s$, might be related to the y -variable nonresponse pattern.

5.2 Editing

Edit rules are often used to restrict the possible values to be imputed. Calibrated imputation can deal with this issue by considering the set of all edit rules as an additional constraint in the minimization process. This does not affect the properties of the vector of imputed estimators $\hat{\mathbf{t}}_{dy}^I$ since the final imputed values are still calibrated to satisfy the constraint $\hat{\mathbf{t}}_{dy}^I = \hat{\mathbf{t}}_{dy}^*$. However, it has an effect on the final imputed values. Moreover, it is generally not possible to obtain a closed-form solution using the method of Lagrange multipliers and a numerical algorithm is usually required to obtain the final imputed values. This is because the edit rules are often more complex than simple equations. Also, since the edit rules often involve more than one variable of interest at a time, a multivariate distance function is needed. This area requires further investigation, especially in the context of categorical variables of interest.

6. CONCLUSION

In this paper, we have proposed to use calibrated imputation in the context of the quasi-model-assisted approach. This technique consists of finding final imputed values as close as possible to preliminary imputed values and that are calibrated to satisfy constraints. The constraints are chosen in such a way that the resulting imputed estimator is approximately unbiased under a nonresponse model, not necessarily restricted to the uniform nonresponse model. It is not necessary to explicitly specify an imputation model although it is natural to use an imputation model to justify the form of the imputed estimator. Even if an imputation model is used, it does not need to be correctly specified to obtain valid inferences.

In this paper, we have assumed that all population parameters of interest are taken into account at the imputation stage of the survey. Although many population parameters are often known in advance or might be expected to be of interest, this cannot always be the case, especially if a public-use microdata file is produced. The problem with unplanned population parameters is that we cannot be sure that the imputed estimator has a negligible nonresponse bias. Even under an approach relying more on an imputation model, such as in Särndal (1992) or in multiple imputation, it is well known (see, for example, Meng, 1994 and Rubin, 1996) that variables used to define the population parameters that will be estimated need to be considered when selecting the imputation model. Otherwise, there is no assurance that valid imputation-model-based inferences can be obtained. However, if the imputation model has been carefully validated and is rich enough, making imputation-model-based inferences can be an attractive choice for analysts using public-use microdata files. In such a case, resampling methods, such as the bootstrap, or multiple imputation can greatly simplify variance estimation.

ACKNOWLEDGEMENTS

The author would like to thank J.N.K. Rao from Carleton University as well as David Binder, Steve Matthews and Eric Rancourt from Statistics Canada for their useful suggestions that helped improve the quality of this paper.

REFERENCES

- Beaumont, J.-F. (2000), "On Regression Imputation in the Presence of Nonignorable Nonresponse", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 580-585.
- Beaumont, J.-F. (2004), "Calibrated Imputation in Surveys under a Quasi-Model-Assisted Approach", unpublished manuscript, Statistics Canada.
- Beaumont, J.-F., and Alavi, A. (2003), "Robust Generalized Regression Estimation", unpublished manuscript submitted to *Survey Methodology*.

- Beaumont, J.-F., and Mitchell, C. (2002), "The System for Estimation of Variance due to Nonresponse and Estimation (SEVANI)", *Proceedings of Statistics Canada Symposium 2002, Modeling Survey Data for Social and Economic Research*, Statistics Canada.
- Binder, D.A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys", *International Statistical Review*, 51, pp. 279-292.
- Bissonnette, J., and Girard, J. (1998), "Improvements in Imputation Estimators: the ESTIMP Module", unpublished manuscript, Business Survey Methods Division, Statistics Canada.
- Deville, J.-C. (2002), "Imputation par prédiction ou imputation avec aléa?", *Recueil des Journées de méthodologie statistique*, INSEE.
- Deville, J.-C., and Särndal, C.-E. (1992), "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association*, 87, pp. 376-382.
- Deville, J.-C., and Särndal, C.-E. (1994), "Variance Estimation for the Regression Imputed Horvitz-Thompson Estimator", *Journal of Official Statistics*, 10, pp. 381-394.
- Favre, A.-C., Matei, A., and Tillé, Y. (2003), "Calibrated Random Imputation for Qualitative Data", *Journal of Statistical Planning and Inference* (to appear).
- Fay, R.E. (1996), "Alternative Paradigms for the Analysis of Imputed Survey Data", *Journal of the American Statistical Association*, 91, pp. 490-498.
- Gagnon, F., Lee, H., Rancourt, E., and Särndal (1996), "Estimating the Variance of the Generalized Regression Estimator in the Presence of Imputation for the Generalized Estimation System", *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pp. 151-156.
- Hidiroglou, M.A., and Särndal, C.-E. (1998), "Use of Auxiliary Information for Two-Phase Sampling", *Survey Methodology*, 24, pp. 11-20.
- Liu, T.-P., and Rancourt, E. (2001), "Constrained Categorical Imputation for Nonresponse in Surveys", Methodology Branch Working Paper no. HSMD-2001-012E, Statistics Canada.
- Mantel, H., Nadon, S., and Yeo, D. (2000), "Effect of Nonresponse Adjustments on Variance Estimates for the National Population Health Survey", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 221-226.
- Mantel, H.J., Singh, A.C., and Yu, M. (1995), "Mass Imputation for Two Phase Sampling: Use of Small Area Estimation and Calibration Techniques", *Proceedings of the Survey Methods Section, Statistical Society of Canada*, pp. 57- 62.
- Meng, X.L. (1994), "Multiple Imputation with Uncongenial Sources of Input (with Discussion)", *Statistical Science*, 9, pp. 538-574.
- Oh, H.L., and Scheuren, F.J. (1983), "Weighting Adjustment for Unit Nonresponse", In W.G. Madow, I. Olkin, and D.B. Rubin (eds.), *Incomplete Data in Sample Surveys*, Vol. 2, New-York: Academic Press, pp. 143-184.
- Rao, J.N.K. (1996), "On Variance Estimation with Imputed Survey Data" *Journal of the American Statistical Association*, 91, pp. 499-506.

Rao, J.N.K., and Shao, J. (1992), "Jackknife Variance Estimation with Survey Data under Hot-Deck Imputation", *Biometrika*, 79, pp. 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, pp. 453-460.
- Ren, R.. (2002), "Méthodes d'imputation de valeurs aberrantes pour des données d'enquête" *Recueil des Journées de méthodologie statistique*, INSEE.
- Rubin, D.B. (1978), "Multiple Imputation in Sample Surveys - a Phenomenological Bayesian Approach to Nonresponse", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 20-34.
- Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New-York : Wiley.
- Rubin, D.B. (1996), "Multiple Imputation after 18+ Years", *Journal of the American Statistical Association*, 91, pp. 473-489.
- Särndal, C.-E. (1992), "Methods for Estimating the Precision of Survey Estimates when Imputation Has Been Used", *Survey Methodology*, 18, pp. 241-252.
- Särndal, C.-E., Swensson, B., and Wretman, J.H. (1992), *Model Assisted Survey Sampling*, New-York, Springer-Verlag.
- Shao, J., and Sitter, R. (1996), "Bootstrap for Imputed Survey Data", *Journal of the American Statistical Association*, 91, pp. 1278-1288.
- Shao, J., and Steel, P. (1999), "Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions", *Journal of the American Statistical Association*, 94, pp. 254-265.