Journal of Official Statistics
Vol. 10. No. 4, 1994, pp. 381–394

Variance Estimation for the Regression Imputed Horvitz-Thompson Estimator

Jean-Claude Deville[1] *and Carl-Erik Särndal*[2]

Abstract: Imputation has found widespread use in surveys with missing data but can lead to incorrect inferences, e.g., invalid confidence intervals, unless care is exercised. This paper develops a procedure for valid variance estimation in surveys where regression imputation is used for the missing values. The imputed values are derived from the fit of a multiple regression model, with a multivariate auxiliary variable as predictor. Features of this new procedure are: (i) it is based on single value imputation (as opposed to the computationally more demanding multiple imputation); (ii) the variance estimation is valid for an arbitrary probability sampling design and for an arbitrary response mechanism of the unconfounded type; and (iii) the calculation of the variance estimate can be carried out with the standard formulas programmed in the existing computer packages for survey data.

Key words: Nonresponse; imputation; auxiliary information; unconfounded response mechanism; model assisted inference.

1. Introduction

Rising nonresponse rates have made the treatment of survey data increasingly dependent on imputation and consequently on a proper treatment of the effects that imputation has on the survey results, especially on their accuracy. This paper discusses single value regression imputation. That is, if y_k is a missing observation, impute $\hat{y}_k = \mathbf{x}_k'\hat{\boldsymbol{\beta}}$, where $\mathbf{x}_k$ is a known vector of predictors for unit k and $\hat{\boldsymbol{\beta}}$ is derived from the fit of a multiple regression model to data on the respondents. We propose a practical method for variance estimation when such imputation is used.

[1] I.N.S.E.E., 18 Boulevard Adolphe-Pinard, 75675 Paris Cedex 14, France.
[2] Université de Montréal, Département de mathematiques et de statistique, C.P. 6128, Succursale A, Montréal H3C 3J7, Canada.

A commonly held view of imputation is that it can successfully serve to patch up the sample. One seeks to reestablish the situation that would have prevailed if all units selected for the sample had responded. This outlook has both strengths and weaknesses. Working with a complete data set has computational advantages even if some data are artificial. It is an unavoidable weakness that the inferences are only as sound as the models that are invoked, critically or uncritically, in the fabrication of imputed values.

Point estimation is reestablished in the sense that if t is the population total to be estimated, one typically uses the same estimator formula $\hat{t}$ as in the case of 100% response, and $\hat{t}$ is computed on the data set after imputation, that is, the data set consisting in part of actually observed

data (for the respondents), in part of imputations (for the nonrespondents). This is current practice in many survey organizations.

In a similar spirit one may try to reestablish variance estimation. Here the problem is more intricate. A usually incorrect approach is to take $\hat{V}$, the standard variance estimator formula appropriate for 100% response, simply compute it on the data set after imputation and use it as an indicator of the precision attained in the survey. A naive user may be tempted to do this, especially since standard formulas are already programmed in existing software for survey data. That this can be completely misleading has been recognized for some time. The standard formulas, computed on data containing both actual observations and imputations, will often lead to considerable underestimation of the variance.

Multiple imputation methods, promoted by Rubin (1987), are perhaps the best known remedy at this point in time. Although theoretically sound, multiple imputation requires considerable data handling and storage and is not ideal in repeated surveys carried out in national statistical agencies, where rapid, standardized computation is required to meet deadlines in official statistics production.

We work in this paper with single value imputation. Given that statistical agencies employ a variety of sampling designs in their surveys, our aim is to present a method with enough generality that it can be applied to any probability sampling design.

Although naive use of a standard variance estimator formula is incorrect, the question arises: With a more sophisticated use of the standard formula, can we still profit from existing computer software and arrive at a valid variance estimate in the case of imputation? We show that with single value regression imputation it is possible to embed the standard formula in a series of computations leading to a valid variance estimate. Thus with little extra programming, existing software can be used, a considerable step forward. Computationally our method entails modifying standard software by going into an "imputation mode" for variance estimation.

There is a considerable literature on uses of regression methods and auxiliary information for handling nonresponse. One approach is modeling of the response mechanism. Ekholm and Laaksonen (1991) use a logistic regression model to estimate unknown response probabilities. Binder (1991) concentrates on qualitative variables of interest and uses log-linear models to explain the nonresponse mechanism. Another use of auxiliary information is to form regression estimators; these have been found effective for reducing nonresponse bias, see Bethlehem (1988) and Särndal and Swensson (1987). Finally, imputation is often carried out with the aid of a regression fit; Hinde and Chambers (1991) recently studied regression imputation strategies and the use of the iterative EM-algorithm.

The approach in this paper is in the spirit of Särndal, Swensson and Wretman (1992), in particular the nonresponse chapter in that book, but goes further. Ideally, the inference should be model free, but with nonresponse this goal cannot be realized. Instead the inference becomes model assisted, which implies a desire to stay close to the classical randomization theory approach. Nonresponse is treated as the second phase of selection incurred after the sample selection. If there is no nonresponse, the formulas should reduce to those that everyone recognizes as

"standard" design based formulas for the case of complete response.

The work reported in this paper was inspired in part by efforts at Statistics Canada to improve the variance estimation procedure for cases of single value imputation. It extends the method in Särndal (1990, 1992), who gave detailed results for ratio imputation under simple random sampling, which is a special case of the more general method developed in this paper. The ratio imputation case is revisited in the following (Sections 3, 5–7) to illustrate how our more general method unfolds. Also at Statistics Canada work is underway to study jackknife variance estimation in the presence of imputation, see Rao (1992) and Kovar and Chen (1992).

2. Complete Response and Standard Software for Variance Estimation

We consider a finite population $U = \{1, \ldots, k, \ldots, N\}$. Denote by y_k the true value for the unit k of a variable of interest y. The objective is to estimate the y-total $t = \sum_U y_k$. From U, a probability sample s is drawn with known probability $p(s)$. We say that the survey has *complete response* if the entire sample y-data set $y_s = \{y_k: k \in s\}$ is observed and available for estimation purposes. An entirely design based inference is possible in this case using the inclusion probabilities $\pi_k = \sum_{s \ni k} p(s)$, $k = 1, \ldots, N$, which are assumed to be known and positive. The Horvitz-Thompson estimator $\hat{t} = \sum_s w_k y_k$ with weights $w_k = 1/\pi_k$ estimates $t = \sum_U y_k$ without bias. (If M is a set of units, $M \subseteq U$, let us write $\sum_M$ to denote $\sum_{k \in M}$, for example, $\sum_U y_k = \sum_{k \in U} y_k$.) The design based variance of $\hat{t}$ can be expressed as a quadratic form, $V_p(y_U) = \sum\sum_U \Delta_{k\ell} y_k y_\ell$, where $y_U = \{y_k: k \in U\}$ denotes the y-data for the entire population and $\Delta_{k\ell} = \pi_{k\ell}/(\pi_k \pi_\ell) - 1$. Here, $\pi_{k\ell}$ is the probability that both k and ℓ are included the sample, and $\pi_{kk} = \pi_k$. (If M is a set of units, $\sum\sum_M$ will be our shorthand for the double sum $\sum\sum_{k \in M, \ell \in M}$.) This variance is estimated without bias from the sample data $y_s = \{y_k: k \in s\}$ by $\hat{V}_p(y_s) = \sum\sum_s A_{k\ell} y_k y_\ell$, where $A_{k\ell} = \Delta_{k\ell}/\pi_{k\ell} = 1/(\pi_k \pi_\ell) - 1/\pi_{k\ell}$.

In particular, for simple random sampling without replacement (SRSWOR) with the sampling fraction $f = n/N$, then $\Delta_{kk} = f^{-1} - 1$ and $A_{kk} = f^{-1}(f^{-1} - 1)$ for all k; $\Delta_{k\ell} = -(f^{-1} - 1)/(N - 1)$ and $A_{k\ell} = -f^{-1}(f^{-1} - 1)/(n - 1)$ for all $k \neq \ell$.

The variance estimate $\hat{V}_p(y_s) = \sum\sum_s A_{k\ell} y_k y_\ell$ is computed routinely for various sampling designs with software available in most survey organizations, for example, SUPERCARP (Hidiroglou, Fuller, and Hickman 1980), SESUDAAN (Shah 1981) and STRATOR (Lasarre 1989). Such software is of course an asset but uncritical use can lead the analyst astray in surveys with nonresponse, unless he or she is aware of the pitfalls with imputation. The software empowers the analyst to let the formula $\hat{V}_p(\cdot)$ operate mechanically on any data set $z_s = \{z_k: k \in s\}$ for which a value z_k has been specified for every $k \in s$. This operation yields the number $\hat{V}_p(z_s)$. When the data set z_s is composed in part of actual observations and in part by imputations $\hat{V}_p(z_s)$ can be a completely misleading indicator of precision. We call $\hat{V}_p(\cdot)$ the *standard formula* and show in the following how it can be used to calculate a proper variance estimate when the void caused by missing values is filled by single value regression imputation.

3. Regression Imputation

Now consider nonresponse and imputation. We then call s the desired sample. The

sampling design, $p(s)$, expresses the known probability of drawing the desired sample s from U; π_k and $\pi_{k\ell}$ denote the known inclusion probabilities under this design. The set of respondents is denoted by r, the set of nonrespondents by $o = s - r$. The *response mechanism*, denoted $q(r|s)$, expresses the unknown conditional probability that the subset r responds, given s. We observe y_k for $k \in r$ only. That is, $y_r = \{y_k\colon k \in r\}$ are observed data and $y_o = \{y_k\colon k \in o\}$ are missing data requiring imputation.

Denote by $\hat{y}_k$ the imputed value for a unit $k \in o$. In this paper, $\hat{y}_k$ is a regression prediction derived from a model fit with the aid of an auxiliary value $\mathbf{x}_k$, a J-dimensional vector available for all $k \in s$. The *model*, denoted ξ, states that for $k \in U$

$$\begin{aligned} y_k &= \mathbf{x}_k'\boldsymbol{\beta} + \epsilon_k \\ E_\xi(\epsilon_k) &= 0, \quad E_\xi(\epsilon_k^2) = \sigma_k^2 = \sigma^2\mathbf{x}_k'\boldsymbol{\lambda}, \\ E_\xi(\epsilon_k\epsilon_\ell) &= 0 \ (k \neq \ell). \end{aligned} \tag{3.1}$$

Here, $\boldsymbol{\beta}$ is an unknown regression coefficient vector and $\boldsymbol{\lambda}$ a vector of specified constants, both of dimension J. To impose the variance structure $\sigma_k^2 = \sigma^2\mathbf{x}_k'\boldsymbol{\lambda}$ does not severely restrict the range of possible imputations. Two simple models of this type are $y_k = \beta + \epsilon_k$ with $E_\xi(\epsilon_k^2) = \sigma^2$, which leads to imputation by the respondent mean, and $y_k = x_k\beta + \epsilon_k$ with $E_\xi(\epsilon_k^2) = \sigma^2 x_k$, which leads to ratio imputation, as pointed out at the end of this section. In the following we return repeatedly to these two models to show the progression of our argument. The example with mean imputation serves mainly to confirm that our approach leads to the variance estimator that most would agree is the "standard one" for this simple case; in practice, mean imputation is too elementary to be of great interest. However, it is clear that the model formulation (3.1) covers a wide range of possibilities for imputation with regression. All these can be treated according to the general method that we develop. Compared to jackknifing and other resampling methods the computation is not extensive even for multiple regression models. The computations are carried out by closed form expressions according to an algorithm summarized in Section 9.

The unknown $\boldsymbol{\beta}$ is estimated from respondent data by weighted least squares

$$\hat{\mathbf{B}}_a = \mathbf{T}_{r,a}^{-1}\sum_r a_k\mathbf{x}_k y_k/\sigma_k^2 \tag{3.2}$$

where $\mathbf{T}_{r,a} = \sum_r a_k\mathbf{x}_k\mathbf{x}_k'/\sigma_k^2$, a $J \times J$ matrix whose inverse is assumed to exist and the a_k, $k \in r$, are known weights. Because $\sigma_k^2 = \sigma^2\mathbf{x}_k'\boldsymbol{\lambda}$, and $\boldsymbol{\lambda}$ is known, we can calculate $\hat{\mathbf{B}}_a$ without knowing σ^2. (Later we need to estimate σ^2; see Section 7.) The subscript a refers to the weighting system, for which more than one viable alternative exists. Two immediate options at the statistical analyst's disposal are:

i. $a_k = w_k$. This choice can be justified by the argument that when the response is complete, this weighting yields a $\hat{\mathbf{B}}_a$ which is design consistent for $\mathbf{B} = (\sum_U \mathbf{x}_k\mathbf{x}_k'/\sigma_k^2)^{-1}\sum_U \mathbf{x}_k y_k/\sigma_k^2$, which in turn is the best linear unbiased estimator of $\boldsymbol{\beta}$ under the model if the entire population U was observed;

ii. $a_k = 1$ for every k. This choice can be justified by the fact that it produces the best linear unbiased estimator of $\boldsymbol{\beta}$ under the model. That is, the weighting $a_k = 1$ is entirely model inspired.

The weighting $a_k = 1$ minimizes the variance under the model of $\hat{\mathbf{B}}_a$ viewed as an estimator of $\boldsymbol{\beta}$. Thus $a_k = 1$ is better from this limited perspective than $a_k = w_k$. But our main interest lies in estimating the

variance (conceived in the more intricate "ξpq sense" explained in Section 4) of $\hat{t}_\bullet$ given by (3.3) viewed as an estimator of the total $t = \sum_U y_k$. With this objective, the choice between $a_k = 1$ and $a_k = w_k$ is far from obvious. Attempting to use analytical comparisons to settle the issue is likely to fail because of the complexity of the expressions. However, we do not foresee that our method for estimating the variance of $\hat{t}_\bullet$ will be so sensitive to the choice of the a_k. To opt for $a_k = 1$ is not unreasonable, in view of the model's dominating importance for the imputation stage. Also the choice $a_k = 1$ happens to yield mathematically more tractable formulas, and it dominates in the following. Note that for a self-weighting design, all $w_k = 1/\pi_k$ are equal, and (i) and (ii) give the same result.

If $k \in o$, the value $\hat{y}_k = \mathbf{x}_k'\hat{\mathbf{B}}_a$ is imputed. The *y-data after imputation*, denoted $y_{\bullet s} = \{y_{\bullet k}: k \in s\}$, are such that

$$y_{\bullet k} = \begin{cases} y_k & \text{if } k \in r \\ \hat{y}_k = \mathbf{x}_k'\hat{\mathbf{B}}_a & \text{if } k \in o \end{cases}.$$

Imputed values must be flagged for identification in the data file.

It is current practice to estimate $t = \sum_U y_k$ by the standard formula $\hat{t} = \sum_s w_k y_k$, computed on data after imputation. That is, the estimator after imputation is

$$\hat{t}_\bullet = \sum_s w_k y_{\bullet k} = \sum_r w_k y_k + \sum_o w_k \hat{y}_k. \tag{3.3}$$

This is the *regression imputed Horvitz-Thompson estimator*. It has the important property of reducing to the ordinary Horvitz-Thompson estimator, $\hat{t} = \sum_s w_k y_k$, in two cases:

i. if there is no nonresponse, that is, $r = s$;
ii. if the imputations are perfect substitutes, that is, $\hat{y}_k = y_k$ for every $k \in o$.

In the formula (3.3), an imputed value is thus treated as a real observation in that its weight remains unchanged at $w_k = 1/\pi_k$. This expresses the survey statistician's belief that no bias is introduced by creating artificial values from the model ξ. This belief seems to be the prime statistical reason why estimates calculated in part from made-up, artificial values should be accepted at all in producing important national statistics. There is faith in the imputation method; statistically speaking this translates as an assumption that no bias is introduced.

Clearly alternatives to (3.3) could, and should perhaps, be entertained. To fall back on the formula for complete response is convenient; existing software can be used directly. But instead one could try to construct "the best possible estimator" (in some well defined sense) of t, given the reduced data set remaining after nonresponse. This would not necessarily be the regression imputed Horvitz-Thompson estimator given by (3.3). However, (3.3) represents such an important current practice that it is necessary to closely examine the question of how variance estimation is to be carried out with this estimator. To answer this question is our goal in this paper.

An alternative view of (3.3) is that it embodies a reweighting of the respondent y-values. We can write $\hat{t}_\bullet = \sum_r w_k^* y_k$, where $w_k^* = w_k + (\sum_{\ell \in o} w_\ell \mathbf{x}_\ell)' \mathbf{T}_{r,a}^{-1} a_k \mathbf{x}_k / \sigma_k^2$ is a transformation of the original weight w_k.

Let us illustrate by means of the two simple examples.

Mean imputation. Assume SRSWOR with $w_k = N/n$ for all k. Suppose no auxiliary values are available to impute by, so pose the extremely simple model ξ given by $y_k = \beta + \epsilon_k$ with $E_\xi(\epsilon_k^2) = \sigma^2$. That is, $\mathbf{x}_k = 1$ for all k and $\boldsymbol{\lambda} = 1$ in the general model statement (3.1). From (3.2), with

$a_k = 1$, we get $\hat{B}_a = \bar{y}_r = \sum_r y_k/m$, where the number of respondents, m, is random. (Since w_k is constant in this case, the option $a_k = w_k$ would lead to the same result.) For every $k \in o = s - r$, impute the respondent mean, $\hat{y}_k = \bar{y}_r$. From (3.3), the estimator of t becomes $\hat{t}_\bullet = N\bar{y}_r$.

Ratio imputation. Assume SRSWOR with $w_k = N/n$ for all k. Suppose x_k is a positive scalar known for every $k \in s$. Let the model ξ be $y_k = x_k\beta + \epsilon_k$ with $E_\xi(\epsilon_k^2) = \sigma^2 x_k$. Let us take $\lambda = 1$ in (3.1) and $a_k = 1$ for all k in (3.2), yielding $\hat{B}_a = \bar{y}_r/\bar{x}_r$. ($a_k = w_k$ leads to the same result.) The imputed values are $\hat{y}_k = x_k\bar{y}_r/\bar{x}_r$ for $k \in o$, hence the name ratio imputation. The estimator of t resulting from (3.3) is $\hat{t}_\bullet = N\bar{x}_s\bar{y}_r/\bar{x}_r$. (The index r or s on $\bar{x}$ or $\bar{y}$ specifies the set on which the arithmetic mean is calculated, so $\bar{x}_r = \sum_r x_k/m$, $\bar{x}_s = \sum_s x_k/n$, and so on.)

To treat imputed values as real observations may work for point estimation, if the model is realistic, but will usually be misleading for variance estimation, even when the model holds. The standard formula for variance estimation, computed on the complete response data, $y_s = \{y_k: k \in s\}$, would have given $\hat{V}_p(y_s) = \sum\sum_s A_{k\ell} y_k y_\ell$, where $A_{k\ell} = 1/(\pi_k\pi_\ell) - 1/\pi_{k\ell}$. Computation on the y-data after imputation, $y_{\bullet s} = \{y_{\bullet k}: k \in s\}$, gives instead

$$\hat{V}_p(y_{\bullet s}) = \sum\sum_s A_{k\ell} y_{\bullet k} y_{\bullet \ell}. \tag{3.4}$$

To use $\hat{V}_p(y_{\bullet s})$ by itself as indicator of the precision of $\hat{t}_\bullet$ is again to act as if imputed values are as good as observed values. But (3.4) underestimates the true variance, often dramatically, especially with high nonresponse. Thus a confidence interval computed with the aid of (3.4) will often be much too short for the confidence level aimed at. This is a well known fact, and multiple imputation methods will, at least for simpler sampling designs, improve the situation. However, single (rather than multiple) imputation is attractive for routine statistics production in national statistical agencies. In this case terms must be added to $\hat{V}_p(y_{\bullet s})$. We now derive these terms and show how they can be computed.

4. Variance and Variance Estimation

The desired sample s is selected with the known probability $p(s)$. The response mechanism, $q(r|s)$, represents the usually unknown probability that the set r responds, given s. A general form of the response mechanism is $q(\cdot|s) = q(\cdot|s, y_s, \mathbf{x}_s)$. That is, it depends on s, on the x-data $\mathbf{x}_s = \{\mathbf{x}_k: k \in s\}$ and on the y-data $y_s = \{y_k: k \in s\}$. In particular, if it depends on $\mathbf{x}_s$ but not on y_s, so that $q(\cdot|s) = q(\cdot|s, \mathbf{x}_s)$, then we call the mechanism unconfounded, resembling terminology in Rubin (1983). Put differently, to be unconfounded, $q(r|s)$ must not depend on the residual set $\epsilon_s = \{\epsilon_k: k \in s\}$ in the model ξ in (3.1) used to create imputed values. Let E_ξ, E_p and E_q be the expectation operators with respect to ξ, $p(s)$ and $q(r|s)$, respectively. We use the well known anticipated variance, $E_\xi E_p E_q(\hat{t}_\bullet - t)^2$, to assess the precision of $\hat{t}_\bullet$.

What unbiasedness, if any, can be claimed for the estimator $\hat{t}_\bullet$ endorsed by current practice? It is in fact ξpq-unbiased, that is, $E_\xi E_p E_q(\hat{t}_\bullet - t) = 0$, under two conditions specified below. To see this, decompose the total error, $\hat{t}_\bullet - t$, into sampling error, $\hat{t} - t$, and error due to imputation $\hat{t}_\bullet - \hat{t}$, where $\hat{t} = \sum_s w_k y_k$ is the Horvitz-Thompson estimator for complete response. We have

$$\begin{aligned}\xi pq\text{-bias } (\hat{t}_\bullet) &= E_\xi E_p E_q(\hat{t}_\bullet - t) \\ &= E_\xi E_p(\hat{t} - t) + E_\xi E_p E_q(\hat{t}_\bullet - \hat{t}).\end{aligned}$$

The first term, $E_\xi E_p(\hat{t} - t)$, is zero, because $E_p(\hat{t} - t) = 0$ by the fact that the complete response Horvitz-Thompson estimator $\hat{t}$ is

design unbiased for t. To claim that the second term, $E_\xi E_p E_q(\hat{t}_\bullet - \hat{t})$, is zero takes two conditions:

A. the mechanism $q(r|s)$ is unconfounded (but otherwise unknown), and
B. the imputation model ξ given by (3.1) holds, so that the imputation error $\hat{t}_\bullet - \hat{t} = -\sum_s w_k(y_k - \hat{y}_k)$ has zero model expectation, given any s and r.

If (A) holds, changing the order of the operators, $E_\xi E_p E_q$ into $E_p E_q E_\xi$, is allowed without affecting the value of the expectation. If (B) also holds, $E_\xi\{(\hat{t}_\bullet - \hat{t})|s, r\} = 0$. In words, given any s and r, the difference between the imputed estimator $\hat{t}_\bullet$ and the complete response estimator $\hat{t}$ differs from zero only by random error having zero expectation under the model used to impute. Thus, under (A) and (B), $E_\xi E_p E_q(\hat{t}_\bullet - \hat{t}) = 0$; consequently, $\hat{t}_\bullet$ is ξpq-unbiased for t.

Note that condition (A) does not imply that the response mechanism is of the simple kind called uniform. A uniform mechanism is a naive assumption implying that all units respond with the same probability, that is, the nonresponse is not selective in any way. But numerous studies have shown that the problem with nonresponse is precisely that it is nonuniform, often related to observable characteristics of the units, such as age and sex in the case of individuals. Such nonuniform nonresponse is permitted under an unconfounded mechanism; we can have very different response probabilities for the different units, as long as these probabilities depend on the x_k-values but not on the y_k-values. In the following, we assume that (A) and (B) apply.

Confounded response causes the problem that $\hat{\mathbf{B}}_a$ given in (3.2) is a model biased estimator of the regression coefficient β. The imputed values $\hat{y}_k = \mathbf{x}_k'\hat{\mathbf{B}}_a$ for $k \in o$ are thereby distorted, and both the point estimator $\hat{t}_\bullet$ and the corresponding variance estimator suffer from bias.

Now consider the ξpq-variance (or the anticipated variance under ξ) of $\hat{t}_\bullet$, that is, $E_\xi E_p E_q(\hat{t}_\bullet - t)^2$. Using the decomposition $\hat{t}_\bullet - t = (\hat{t} - t) + (\hat{t}_\bullet - \hat{t})$ and permuting $E_\xi E_p E_q$ into $E_p E_q E_\xi$ (allowed under assumption A), we obtain

$$E_\xi E_p E_q(\hat{t}_\bullet - t)^2 = E_\xi V_p + E_p E_q V_{imp\xi}$$

where $V_p = V_p(y_U) = \sum\sum_U \Delta_{k\ell} y_k y_\ell$ is the design based variance of $\hat{t}$ and

$$V_{imp\xi} = E_\xi\{(\hat{t}_\bullet - \hat{t})^2 + 2(\hat{t}_\bullet - \hat{t})(\hat{t} - t)|s, r\}. \tag{4.1}$$

The decomposition of the total anticipated variance, $V_{tot} = E_\xi E_p E_q(\hat{t}_\bullet - t)^2$, now reads

$$V_{tot} = V_{sam} + V_{imp}$$

where $V_{sam} = E_\xi V_p$, called the *sampling variance*, is the anticipated design based variance for the case of complete response, and the second component, $V_{imp} = E_p E_q V_{imp\xi}$, called *imputation variance*, is the variance added through imputation.

We now construct estimators, $\hat{V}_{sam}$ and $\hat{V}_{imp}$, of the components V_{sam} and V_{imp}. For complete response we require, as is reasonable, that they reduce to $\hat{V}_{imp} = 0$ and $\hat{V}_{sam} = \hat{V}_p(y_s) = \sum\sum_s A_{k\ell} y_k y_\ell$, the standard formula computed on complete response y-data.

a. Estimation of the sampling variance, V_{sam}. Mechanical calculation by the standard formula gives the result $\hat{V}_p(y_{\bullet s})$ given by (3.4). This understates the sampling variance which is correctly estimated by $\hat{V}_p(y_s) = \sum\sum_s A_{k\ell} y_k y_\ell$, but this latter quantity cannot be calculated, so we must estimate the difference. Find $C = E_\xi\{\hat{V}_p(y_s) - \hat{V}_p(y_{\bullet s})\}$ and estimate it by $\hat{C}$ satisfying $E_\xi(\hat{C}) = C$. It then follows that

$\hat{V}_p(y_{\bullet s}) + \hat{C}$ will estimate V_{sam} without ξpq-bias: We have $E_\xi E_p E_q\{\hat{V}_p(y_{\bullet s}) + \hat{C}\} = E_\xi V_p = V_{sam}$, because

$$E_\xi E_p E_q\{\hat{V}_p(y_{\bullet s}) + \hat{C}\}$$
$$= E_p E_q\{E_\xi(\hat{V}_p(y_{\bullet s})) + C\}$$
$$= E_p E_q\{E_\xi(\hat{V}_p(y_s)\}$$
$$= E_\xi E_p \hat{V}_p(y_s) = E_\xi V_p.$$

b. Estimation of the imputation variance, V_{imp}. This is conceptually simple. Find an estimator $\hat{V}_{imp\xi}$ that is model unbiased for the conditional imputation variance $V_{imp\xi}$ given in (4.1). This is a standard problem in inferential statistics. Then $\hat{V}_{imp\xi}$ will estimate V_{imp} without ξpq-bias, because $E_\xi E_p E_q \hat{V}_{imp\xi} = E_p E_q E_\xi \hat{V}_{imp\xi} = E_p E_q V_{imp\xi} = V_{imp}$.

Details for finding $\hat{C}$ and $\hat{V}_{imp\xi}$ such that $E_\xi \hat{C} = C$ and $E_\xi \hat{V}_{imp\xi} = V_{imp\xi}$ will be given in the next section. At this point, we summarize in the form of a general principle.

Procedure for variance estimation: If conditions (A) and (B) hold, a ξpq-unbiased estimator of the variance of $\hat{t}_\bullet = \sum_s w_k y_{\bullet k}$ is given by

$$\hat{V}_{tot} = \hat{V}_{sam} + \hat{V}_{imp}.$$

The sampling variance is estimated by

$$\hat{V}_{sam} = \hat{V}_p(y_{\bullet s}) + \hat{C}$$

where $\hat{V}_p(y_{\bullet s})$ is calculated by the standard formula applied to the data after imputation, and $\hat{C}$ satisfying $E_\xi(\hat{C}) = C = E_\xi\{\hat{V}_p(y_s) - \hat{V}_p(y_{\bullet s})\}$ is added to avoid underestimation of the sampling variance. The imputation variance is estimated by

$$\hat{V}_{imp} = \hat{V}_{imp\xi}$$

where $\hat{V}_{imp\xi}$ satisfies $E_\xi \hat{V}_{imp\xi} = V_{imp\xi}$ given by (4.1).

5. Estimation of the Imputation Variance

We now produce explicit expressions for $\hat{V}_{sam}$ and $\hat{V}_{imp}$. Starting with $\hat{V}_{imp}$, let us evaluate the two terms of (4.1). The first term is

$$E_\xi\{(\hat{t}_\bullet - \hat{t})^2)|s,r\} = \mathbf{t}'_{o,w} V_\xi(\hat{\mathbf{B}}_a|s,r)\mathbf{t}_{o,w} + \sum_o w_k^2 \sigma_k^2 \quad (5.1)$$

where $\mathbf{t}_{o,w} = \sum_o w_k \mathbf{x}_k$ and

$$V_\xi(\hat{\mathbf{B}}_a|s,r) = \mathbf{T}_{r,a}^{-1}\mathbf{T}_{r,aa}\mathbf{T}_{r,a}^{-1}.$$

with $\mathbf{T}_{r,a}$ defined by (3.2) and $\mathbf{T}_{r,aa} = \sum_r a_k^2 \mathbf{x}_k \mathbf{x}'_k/\sigma_k^2$. The second term of (4.1) is

$$2E_\xi\{(\hat{t}_\bullet - \hat{t})(\hat{t} - t)|s,r\} =$$
$$2(-\mathbf{t}'_{o,ww}\mathbf{T}_{r,a}^{-1}\mathbf{t}_{r,a} + \mathbf{t}'_{o,w}\mathbf{T}_{r,a}^{-1}\mathbf{t}_{r,aw}) \quad (5.2)$$

where $\mathbf{t}_{r,a} = \sum_r a_k \mathbf{x}_k$, $\mathbf{t}_{o,ww} = \sum_o w_k^2 \mathbf{x}_k$, $\mathrm{t}_{r,aw} = \sum_r a_k w_k \mathbf{x}_k$. It is easy to see that (5.2) is zero if the sampling design is self-weighting (w_k = constant); otherwise, (5.2) is nonzero but ordinarily small compared to (5.1).

Because σ_k^2 is of the form $\sigma^2 \mathbf{x}'_k \boldsymbol{\lambda}$, we conclude that the sum of (5.1) and (5.2) can be written as $V_{imp\xi} = \sigma^2 G$, where G can always be computed from the sample x-data. Thus σ^2 is the only unknown requiring estimation.

Simplicity is a strong point in favour of the weighting $a_k = 1$ in (3.2). It leads to $\mathbf{T}_{r,a} = \mathbf{T}_{r,aa} = \sigma^{-2}\mathbf{T}$, where $\mathbf{T} = \sum_r \mathbf{x}_k \mathbf{x}'_k/(\mathbf{x}'_k\boldsymbol{\lambda})$, and the sum of (5.1) and (5.2) reduces to

$$V_{imp\xi} = \sigma^2\{\mathbf{t}'_{o,w}\mathbf{T}^{-1}(\mathbf{t}_{s,w} + \mathbf{t}_{r,w}) - \mathbf{t}'_{o,ww}\mathbf{T}^{-1}\mathbf{t}_{r,1}\} \quad (5.3)$$

where $\mathbf{t}_{s,w} = \sum_s w_k \mathbf{x}_k$ and $\mathbf{t}_{r,1} = \sum_r \mathbf{x}_k$. If in addition the design is self-weighting, say, $w_k = W$ for all k, then

$$V_{imp\xi} = \sigma^2 W^2 \left(\sum_o \mathbf{x}_k\right)' \mathbf{T}^{-1} \left(\sum_s \mathbf{x}_k\right). \quad (5.4)$$

In (5.3) and (5.4), σ^2 is the only unknown. What remains in order to have an ξpq-unbiased estimator of the imputation variance is to find a $\hat{\sigma}^2$ based on data for the respondents and satisfying $E_\xi(\hat{\sigma}^2) = \sigma^2$. In Section 7 we provide such an estimator $\hat{\sigma}^2$.

To illustrate, reconsider the two examples started at the end of Section 3.

Mean imputation. We obtain $\mathbf{T} = m$, the number of respondents. Furthermore, $\mathbf{t}_{o,w} = (N/n)(n-m)$ and $\mathbf{t}_{s,w} = N$. From (5.4),

$$V_{imp\xi} = N^2(1/m - 1/n)\sigma^2.$$

The factor $1/m - 1/n$ reminds us of a second phase SRSWOR selection of m respondents from the n desired sample units.

Ratio imputation. In this case, $\mathbf{T} = \sum_r x_k = m\bar{x}_r$, $\mathbf{t}_{o,w} = (N/n)(n-m)\bar{x}_o$ and $\mathbf{t}_{s,w} = N\bar{x}_s$. From (5.4) we get

$$V_{imp\xi} = N^2(1/m - 1/n)(\bar{x}_o\bar{x}_s/\bar{x}_r)\sigma^2. \quad (5.5)$$

The factor $1/m - 1/n$ appears again, which is reassuring. Formula (5.5) also shows that, given s and m, the imputation variance $V_{imp\xi}$ is relatively higher when the response is selective so that small x-value units respond to a greater extent than large x-value units. This is as it should be, because when $\bar{x}_r$ is smaller than $\bar{x}_s$, relatively many large x-value units require imputation, and the total imputation error $\sum_o(\hat{y}_k - y_k)$ tends to have a high variance. To find $\hat{V}_{imp}$, it remains to replace σ^2 in (5.5) by an estimator $\hat{\sigma}^2$; see Section 7.

6. Correction Term for the Sampling Variance

To correct for the underestimation of the sampling variance component, we focus first on the difference between $\hat{V}_p(y_s)$, which is the proper sampling variance estimate but impossible to compute, and $\hat{V}_p(y_{\bullet s})$, which can be computed but is negatively biased

$$\hat{V}_p(y_s) - \hat{V}_p(y_{\bullet s}) = \sum\sum_o A_{k\ell}(y_k y_\ell - \hat{y}_k\hat{y}_\ell) + \sum_{k\in r}\sum_{\ell\in o} A_{k\ell}y_k(y_\ell - \hat{y}_\ell) + \sum_{k\in o}\sum_{\ell\in r} A_{k\ell}y_k(y_\ell - \hat{y}_\ell).$$

Following the procedure in Section 4, let us estimate the model expected value of this difference. That is, find $\hat{C}$ such that $E_\xi(\hat{C}) = C = E_\xi\{\hat{V}_p(y_s) - \hat{V}_p(y_{\bullet s})\}$, then use $\hat{V}_p(y_{\bullet s}) + \hat{C}$ as an unbiased estimator of the sampling variance $V_{sam} = E_\xi V_p$. Simple derivations give $E_\xi\{y_k(y_\ell - \hat{y}_\ell)\} = -a_k\mathbf{x}_k'\mathbf{T}_{r,a}^{-1}\mathbf{x}_\ell$ and $E_\xi(y_k y_\ell - \hat{y}_k\hat{y}_\ell) = \sigma_{k\ell} - \mathbf{x}_k'\mathbf{T}_{r,a}^{-1}\mathbf{T}_{r,aa}\mathbf{T}_{r,a}^{-1}\mathbf{x}_\ell$, where $\sigma_{k\ell} = 0$ if $k \neq \ell$ and $\sigma_{k\ell} = \sigma_k^2$ if $k = \ell$. Since $\sigma_k^2 = \sigma^2\mathbf{x}_k'\boldsymbol{\lambda}$, both expectations equal σ^2 times a quantity that can be calculated from the known x-data. Thus $C = E_\xi\{\hat{V}_p(y_s) - \hat{V}_p(y_{\bullet s})\}$ is of the form $\sigma^2 D$, where D does not contain any unknowns. The only task remaining is to find an estimator $\hat{\sigma}^2$ such that $E_\xi(\hat{\sigma}^2) = \sigma^2 D$; then $\hat{C} = \hat{\sigma}^2 D$ appropriately corrects the mechanically computed term $\hat{V}_p(y_{\bullet s})$.

There is considerable simplification if we settle for the weighting $a_k = 1$ for all k in (3.2). Then $C = \sigma^2 D$ with

$$D = \sum_o A_{kk}\mathbf{x}_k'\boldsymbol{\lambda} - R \quad (6.1)$$

where $R = \sum\sum_s A_{k\ell}\mathbf{x}_k'\mathbf{T}^{-1}\mathbf{x}_\ell - \sum\sum_r A_{k\ell}\mathbf{x}_k'\mathbf{T}^{-1}\mathbf{x}_\ell$ with $T = \sum_r \mathbf{x}_k\mathbf{x}_k'/(\mathbf{x}_k'\boldsymbol{\lambda})$. To compute the simple sum $\sum_o A_{kk}\mathbf{x}_k'\boldsymbol{\lambda}$ is easy, and although R involves two seemingly complex double sums, we now show that R can be computed with any existing software that handles computation of the standard variance estimator formula $\hat{V}_p(\cdot)$ for a given sampling design.

Consider first the case $J = 1$. Then $\mathbf{x}_k = x_k$ is a scalar and we can take $\boldsymbol{\lambda} = 1$

without loss of generality in the model (3.1). In (6.1) we have $\mathbf{T} = m\bar{x}_r$ and

$$R = (m\bar{x}_r)^{-1}\{\hat{V}_p(x_s) - \hat{V}_p(x_{0s})\} \qquad (6.2)$$

where $x_s = \{x_k: k \in s\}$ is the x-data set for the sample, and $x_{0s} = \{x_{0k}: k \in s\}$ is the sample data for the pseudo-variable x_0 which agrees with x for a respondent and is set to be zero for a nonrespondent, that is, $x_{0k} = x_k$ if $k \in r$ and $x_{0k} = 0$ if $k \in o = s - r$. Viewed in this way, the problem of computing R is reduced to that of computing two standard design based variance estimates, $\hat{V}_p(x_s)$ and $\hat{V}_p(x_{0s})$. This is done by letting the standard variance formula $\hat{V}_p(\cdot)$ operate on the data sets x_s and x_{0s}, respectively. Therefore, R is easily computed with any software that calculates $\hat{V}_p(\cdot)$.

Mean imputation. Here, $A_{kk} = N^2(1/n - 1/N)/n$, so $\sum_o A_{kk}x_k = N^2(1/n - 1/N)(n-m)/n$. Further, $\hat{V}_p(x_s) = 0$ because it is the estimated variance for a variable constant at unity, and $\hat{V}_p(x_{0s})$ is the estimated variance for a variable equaling 1 if $k \in r$ and 0 if $k \in o$. We get $T = m$, $R = -N^2(1/n - 1/N)(1 - m/n)/(n-1)$ and finally $C = \sigma^2 N^2(1/n - 1/N)(n-m)/(n-1)$. For example, if the nonresponse rate is $1 - m/n = 30\%$, the understatement portion C accounts for as much as 30% of the entire sampling variance, $V_{sam} = (1/n - 1/N)\sigma^2$, if we approximate $(m-1)/(n-1)$ by m/n.

Ratio imputation. Here we find $\mathbf{T} = m\bar{x}_r$, $\sum_o A_{kk}\mathbf{x}_k = N^2(1/n - 1/N)\{(n-m)/n\}\bar{x}_o$. Moreover, to calculate (6.2) we need $\hat{V}_p(x_s) = (1/n - 1/N)S_{xs}^2$ with $S_{xs}^2 = \sum_s (x_k - \bar{x}_s)^2/(n-1)$ and $\hat{V}_p(x_{0s}) = (1/n - 1/N)S_{x_os}^2$ with $S_{x_os}^2 = \{\sum_r x_k^2 - n^{-1}(\sum_r x_k)^2\}/(n-1)$. Note that $\hat{V}_p(x_s)$ and $\hat{V}_p(x_{0s})$ are standard SRSWOR variance estimates for the sample means $\bar{x}_s$ and $\bar{x}_{0s} = \sum_s x_{0k}/n$, where $x_{0k} = x_k$ if k is a respondent and $x_{0k} = 0$ if k is a nonrespondent. That is, apart from factors in N, n and m, the calculation of D in (6.1) requires two means, $\bar{x}_r$ and $\bar{x}_o$, and two SRSWOR variance estimates, $(1/n - 1/N)S_{xs}^2$ and $(1/n - 1/N)S_{x_os}^2$. The explicit expression for D in (6.1) is relatively simple in this case and is given by

$$D = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1} \times \left(\sum_o x_k - \frac{\sum_o x_k^2}{\sum_r x_k} + \frac{1}{n}\frac{\sum_o x_k \sum_s x_k}{\sum_r x_k}\right). \qquad (6.3)$$

Let us turn to computational aspects for the case $J \geq 2$. Using spectral decomposition, we can write $\mathbf{T} = \sum_{j=1}^{J} v_j \mathbf{U}_j \mathbf{U}_j'$, where the v_j are the eigenvalues of $\mathbf{T}$ arranged in decreasing order (all v_j are positive), and $\mathbf{U}_1, \ldots, \mathbf{U}_J$ form an orthonormal basis of eigenvectors of $\mathbf{T}$. We then have

$$\mathbf{x}_k'\mathbf{T}^{-1}\mathbf{x}_\ell = \sum_{j=1}^{J} v_j^{-1}(\mathbf{x}_k'\mathbf{U}_j)(\mathbf{x}_\ell'\mathbf{U}_j) = \sum_{j=1}^{J} \zeta_{jk}\zeta_{j\ell}$$

where the values $\zeta_{jk} = v_j^{-1/2}\mathbf{x}_k'\mathbf{U}_j$ can be computed for $k \in s$, $j = 1, \ldots, J$. For $j = 1, \ldots, J$, define the data sets $\zeta_{js} = \{\zeta_{jk}: k \in s\}$ and $\zeta_{j0s} = \{\zeta_{j0k}: k \in s\}$, where $\zeta_{j0k} = \zeta_{jk}$ if $k \in r$ and $\zeta_{j0k} = 0$ otherwise. Now R in (6.1) can be written as

$$R = \sum_{j=1}^{J}\{\hat{V}_p(\zeta_{js}) - \hat{V}_p(\zeta_{j0s})\}$$

where $\hat{V}_p(\zeta_{js})$ and $\hat{V}_p(\zeta_{j0s})$ are calculated with the standard variance estimator formula once the sets ζ_{js} and ζ_{j0s} have been obtained from the spectral decomposition. Thus for $J \geq 2$, R is again easily computed with the aid of any software that handles the standard design based formula $\hat{V}_p(\cdot)$.

7. The Estimation of σ^2

As noted in Sections 5 and 6, the variance component estimators $\hat{V}_{sam}$ and $\hat{V}_{imp}$ require an estimator of the unknown σ^2 in (3.1). We now present such an estimator, assuming for simplicity the weighting $a_k = 1$ for all k in (3.2). The idea is again to exploit the already programmed formula $\hat{V}_p(\cdot)$. First compute the residuals $e_k = y_k - \hat{y}_k$ for the responding units $k \in r$, then define the "pseudo-residual" set $e_{0s} = \{e_{0k}: \quad k \in s\}$, where $e_{0k} = e_k = y_k - \hat{y}_k$ if $k \in r$ and $e_{0k} = 0$ if $k \in o$. Now, $\hat{V}_p(e_{0s}) = \sum\sum_s A_{k\ell} e_{0k} e_{0\ell} = \sum\sum_r A_{k\ell} e_k e_\ell$, and we obtain

$$E_\xi \hat{V}_p(e_{0s}) = E_\xi \left\{ \left(\sum\sum_r A_{k\ell} e_k e_\ell \right) \middle| s, r \right\} = \sigma^2 Q_r$$

where $Q_r = \sum_r A_{kk} \mathbf{x}_k' \boldsymbol{\lambda} - \sum\sum_r A_{k\ell} \mathbf{x}_k' \mathbf{T}^{-1} \mathbf{x}_\ell$ with $\mathbf{T} = \sum_r \mathbf{x}_k \mathbf{x}_k' / (\mathbf{x}_k' \boldsymbol{\lambda})$. It follows that

$$\hat{\sigma}^2 = \hat{V}_p(e_{0s}) / Q_r \tag{7.1}$$

is a model unbiased estimator of σ^2. Note that Q_r can be calculated with the technique in Section 6 for any dimension $J \geq 1$ of the $\mathbf{x}_k$-vector.

We can now specify the desired variance component estimators $\hat{V}_{sam}$ and $\hat{V}_{imp}$. Defining also $Q_s = \sum_s A_{kk} \mathbf{x}_k' \boldsymbol{\lambda} - \sum\sum_s A_{k\ell} \mathbf{x}_k' \mathbf{T}^{-1} \mathbf{x}_\ell$, we can write (6.1) as $C = \sigma^2 (Q_s - Q_r)$ and take $\hat{C} = \hat{\sigma}^2 (Q_s - Q_r) = (Q_s / Q_r - 1) \hat{V}_p(e_{0s})$ as the desired term to prevent underestimation of the sample variance. That is

$$\hat{V}_{sam} = \hat{V}_p(y_{\bullet s}) + (Q_s / Q_r - 1) \hat{V}_p(e_{0s}) \tag{7.2}$$

is unbiased for the sampling variance component V_{sam}. Moreover, from (5.3)

$$\hat{V}_{imp} = \{ \mathbf{t}_{o,w}' \mathbf{T}^{-1} (\mathbf{t}_{s,w} + \mathbf{t}_{r,w}) - \mathbf{t}_{o,ww}' \mathbf{T}^{-1} \mathbf{t}_{r,1} \} \hat{V}_p(e_{0s}) / Q_r \tag{7.3}$$

is unbiased for the imputation variance component V_{imp}. The total variance estimate is the sum of (7.2) and (7.3). Since Q_s and Q_r can be computed with the technique in Section 6 (with spectral decomposition if $J > 1$), we have a complete procedure for variance estimation of the regression imputed Horvitz-Thompson estimator. It can be carried out with the aid of existing survey sampling software.

Remark. A simplified alternative to (7.1) is to estimate σ^2 by

$$\hat{\sigma}^{*2} = \{ m/(m-J) \} \left(\sum_r e_k^2 \right) \Big/ \left(\sum_r \mathbf{x}_k' \boldsymbol{\lambda} \right). \tag{7.4}$$

Although slightly biased, this simplified estimator can be used in practice with good results. The computational advantage compared to (7.1) is, however, relatively minor.

To illustrate (7.1) to (7.4), consider again the two examples.

Mean imputation. The standard formula computed on the data after imputation gives $\hat{V}_p(y_{\bullet s}) = N^2 (1/n - 1/N) \{ (m-1)/(n-1) \} S_{yr}^2$ with $S_{yr}^2 = \sum_r (y_k - \bar{y}_r)^2 / (m-1)$. From (7.1) and (7.2), $\hat{\sigma}^2 = S_{yr}^2$, $\hat{C} = (Q_s / Q_r - 1) \hat{V}_p(e_{0s}) = N^2 (1/n - 1/N) \quad \{ (n-m)/(n-1) \} \quad S_{yr}^2$. Thus $\hat{V}_{sam} = N^2 (1/n - 1/N) S_{yr}^2$, a natural estimator of the sampling variance. In addition, (7.3) gives $\hat{V}_{imp} = N^2 (1/m - 1/n) S_{yr}^2$. Summing the components, we get

$$\hat{V}_{tot} = \hat{V}_{sam} + \hat{V}_{imp} = N^2 (1/m - 1/N) S_{yr}^2.$$

It is important to note that we get precisely what many survey statisticians would agree is the "natural variance estimator" in this simple case. We have used model assisted reasoning; the perhaps better known route to the same formula is via two-phase sampling argument, assuming the respondent set r to be a SRSWOR subsample from s, as in Oh and Scheuren (1983), and Cochran (1977).

Ratio imputation. The standard formula computed on data after imputation yields

$$\hat{V}_p(y_{\bullet s}) = N^2(1/n - 1/N) \times \sum_s (y_{\bullet k} - \bar{x}_s\hat{B})^2/(n-1)$$

where $y_{\bullet k} = y_k$ if $k \in r$ and $y_{\bullet k} = x_k\hat{B}$ if $k \in o$ where $\hat{B} = \bar{y}_r/\bar{x}_r$. We obtain $Q_r = N^2(1/n - 1/N)\{(m-1)/(n-1)\}\bar{x}_r(1 - m^{-1}cv_{xr}^2)$, where $cv_{xr} = S_{xr}/\bar{x}_r$ is the coefficient of variation of x in r, and $\hat{V}_p(e_{0s}) = N^2(1/n - 1/N)\sum_r e_k^2/(n-1)$, where $e_k = y_k - \hat{B}x_k$. The model unbiased estimator of σ^2 obtained from (7.1) is

$$\hat{\sigma}^2 = [\bar{x}_r\{1 - (1/m)cv_{xr}^2\}]^{-1} \times \left(\sum_r e_k^2\right) \Big/ (m-1). \quad (7.5)$$

With $\hat{\sigma}^2$ defined in this way, the variance component estimators are

$$\hat{V}_{sam} = \hat{V}_p(y_{\bullet s}) + \hat{\sigma}^2 D,$$

$$\hat{V}_{imp} = N^2(1/m - 1/n)(\bar{x}_o\bar{x}_s/\bar{x}_r)\hat{\sigma}^2 \quad (7.6)$$

where D is given by (6.3). These results were given in Särndal (1990, 1992). The simpler, slightly biased σ^2-estimator defined by (7.4) becomes $\sigma^{*2} = \{m/(m-1)\}(\sum_r e_k^2)/(\sum_r x_k)$, which is a good approximation to (7.5).

8. Empirical Testing of the Method

We have seen that the method outlined in Sections 4 to 7 produces a ξpq-unbiased estimator of the variance of $\hat{t}_\bullet$ under the conditions (A) and (B) in Section 4. That the survey statistician is ready to accept both assumptions is clear, because, as Section 4 points out, both are needed to justify the point estimator $\hat{t}_\bullet$, that is, to establish its unbiasedness. The statistician who trusts these two conditions in order to justify the point estimator must trust them in the variance estimation procedure as well.

Extensive empirical testing to confirm theory has been carried out in the Ratio Imputation example. Lee, Rancourt and Särndal (1994) undertook Monte Carlo simulations to study the performance of the total variance estimator $\hat{V}_{tot} = \hat{V}_{sam} + \hat{V}_{imp}$ derived from (7.6). The main conclusions were:

a. Confirming theory, the simulation showed that $\hat{V}_{tot}$ is a roughly unbiased estimator of the variance when the population scatter agrees well with the assumption that justifies ratio imputation, that is, the linear regression through the origin, $y_k = x_k\beta + \epsilon_k$.

b. According to the theory seen in this paper, $\hat{V}_{tot}$ is essentially unbiased for an arbitrary unconfounded mechanism. Interestingly, the simulation showed that the bias of $\hat{V}_{tot}$ is quite small even for confounded mechanisms such as when the response probability, θ_k, is an explicit function of the y_k-value, for example, $\theta_k = 1 - \exp(-cy_k)$, where c is a constant. In this simulation at least, our method was quite robust to a breakdown of the assumption of unconfoundedness. For the confounded mechanisms examined, the absolute relative bias of $\hat{V}_{tot}$ was only of the order of a few percentage points.

c. If the population scatter displays a relationship other than linear regression through the origin, then some bias was, not unexpectedly, noted in $\hat{V}_{tot}$. The conclusion is that the method is more sensitive to the regression model used to create the imputations than to the assumption of unconfounded response mechanism. However, even in this case, $\hat{V}_{tot}$ improves greatly on the systematic

understatement of the variance that the standard uncorrected formula $\hat{V}_p(y_{\bullet s})$ will produce.

9. Operating Mode

In closing, we reiterate the steps in the variance estimation procedure proposed in this paper. We assume that the weighting $a_k = 1$ for all k is used in (3.2).

1. *Regression fit.* Fit the regression $y_k = \mathbf{x}_k'\boldsymbol{\beta} + \epsilon_k$ using data on the respondents; obtain $\hat{\mathbf{B}} = \mathbf{T}^{-1}\sum_r \mathbf{x}_k y_k/(\mathbf{x}_k'\boldsymbol{\lambda})$ with $\mathbf{T} = \sum_r \mathbf{x}_k\mathbf{x}_k'/(\mathbf{x}_k'\boldsymbol{\lambda})$.
2. *Imputation.* Impute the value $\hat{y}_k = \mathbf{x}_k'\mathbf{B}$ if $k \in o$.
3. *Preparation of data sets and use of standard formula.* If $J = 1$, prepare the data sets $y_{\bullet s}$, x_s, x_{0s} and e_{0s}. Use standard formula $\hat{V}_p(\cdot)$ to compute $\hat{V}_p(y_{\bullet s})$, $\hat{V}_p(x_s)$, $\hat{V}_p(x_{0s})$ and $\hat{V}_p(e_{0s})$, then compute $Q_r = \lambda\sum_r A_{kk}x_k - (\sum_r x_k)^{-1}\hat{V}_p(x_{0s})$ and $Q_s = \lambda\sum_s A_{kk}x_k - (\sum_r x_k)^{-1}\hat{V}_p(x_s)$. If $J \geq 2$, compute instead Q_s and Q_r with the aid of $\hat{V}_p(\zeta_{js})$ and $\hat{V}_p(\zeta_{j0s})$ obtained by spectral decomposition as described in Section 6.
4. *Estimation of model parameter* σ^2. Compute $\hat{\sigma}^2$ from (7.1).
5. *Arriving at variance component estimates and total variance estimate.* Compute $\hat{V}_{sam}$ and $\hat{V}_{imp}$ from (7.2) and (7.3); finally, compute $\hat{V}_{tot} = \hat{V}_{sam} + \hat{V}_{imp}$.

It is in step (3) that existing computer software comes in as a valuable asset for computing the standard design based variance formula $\hat{V}_p(\cdot)$ on several data sets. The other steps are computationally straightforward. Alternatively, σ^2 may be estimated by (7.4).

10. References

Bethlehem, J.G. (1988). Reduction of Nonresponse Bias Through Regression Estimation. Journal of Official Statistics, 4, 251–260.

Binder, D.A. (1991). A Framework for Analyzing Categorical Survey Data with Nonresponse. Journal of Official Statistics, 7, 393–404.

Cochran, W.G. (1977). Sampling Techniques. Third edition. New York: John Wiley.

Ekholm, A. and Laaksonen, S. (1991). Weighting via Response Modeling in the Finnish Household Budget Survey. Journal of Official Statistics, 7, 325–337.

Hidiroglou, M.A., Fuller, W.A., and Hickman, R.D. (1980). SUPERCARP-Sixth Edition. Statistical Laboratory, Survey Section, Iowa State University, Ames, Iowa.

Hinde, R.L. and Chambers, R.L. (1991). Nonresponse Imputation with Multiple Sources of Nonresponse. Journal of Official Statistics, 7, 167–179.

Kovar, J.G. and Chen, E.J. (1994). Jackknife Variance Estimation of Imputed Survey Data. Survey Methodology, 20, 45–52.

Lasarre, S. (1989). STRATOR, version 1–3, présentation technique. Note technique de DEMOSCOPIE-INRETS, Paris.

Lee, H., Rancourt, E., and Särndal, C.E. (1994). Experiments with Variance Estimation from Survey Data with Imputed Values. Journal of Official Statistics, 10, 231–243.

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, 143–184.

Rao, J.N.K. (1992). Jackknife Variance

Estimation under Imputation for Missing Survey Data. Statistics Canada report.

Rubin, D.B. (1983). Conceptual Issues in the Presence of Nonresponse. In Incomplete Data in Sample Surveys, (eds.) W.G. Madow, I. Olkin, and D.B. Rubin, vol. 2. New York: Academic Press, 123–142.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley.

Särndal, C.E. (1990). Methods for Estimating the Precision of Survey Estimates when Imputation Has Been Used. Proceedings of Statistics Canada's Symposium '90: Measurement and Improvement of Data Quality, Ottawa, October 29–31, 1990, 337–347.

Särndal, C.E. (1992). Methods for Estimating the Precision of Survey Estimates when Imputation Has Been Used. Survey Methodology, 18, 241–252.

Särndal, C.E. and Swensson, B. (1987). A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse. International Statistical Review, 55, 275–294.

Särndal, C.E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

Shah, V.V. (1981). SESUDAAN: Standard Errors Program for Computing of Standardized Rates from Sample Survey Data. Research Triangle Institute, Research Triangle Park, North Carolina.

Received February 1993
Revised August 1994