

11-613
no. 2002-007
c. 1



Statistics
Canada Statistique
Canada

NOT FOR LOAN
NE S'EMPRUNTE PAS

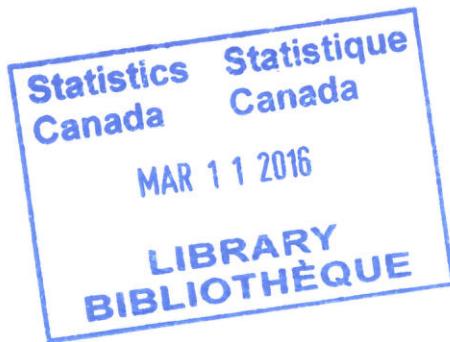


Methodology Branch

Social Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes sociales



Canadä



WORKING PAPER
METHODOLOGY BRANCH

**Linearization Variance Estimators
for Survey Data
with Missing Responses**

SSMD-2002-007E/F

A. Demnati and J. N. K. Rao

Statistics Canada

November 2002

The work presented in this paper is the responsibility of the author and does not necessarily represent the views or policies of Statistics Canada.

November 2002

Linearization Variance Estimators for Survey Data with Missing Responses

A. Demnati and J. N. K. Rao

SUMMARY

In survey sampling, Taylor linearization is often used to obtain variance estimators for nonlinear finite population parameters such as ratios, regression and correlation coefficients which can be expressed as smooth functions of totals. Taylor linearization is generally applicable to any sampling design, but it can lead to multiple variance estimators that are asymptotically design unbiased under repeated sampling. The choice among the variance estimators requires other considerations such as (i) approximate unbiasedness for the model variance of the estimator under an assumed model, (ii) validity under a conditional repeated sampling framework. Demnati and Rao (2001) proposed a new approach to deriving Taylor linearization variance estimators that leads directly to a unique variance estimator that satisfies the above considerations. In this paper, we extended the work of Demnati and Rao (2001) to deal with missing data problem. We derived valid variance estimators under weighting adjustment, which is often used to compensate for complete nonresponse, as well as under imputation based on smooth functions of observed values, in particular ratio imputation, which is often used to produce a complete data set.

Key Words: Item nonresponse; Ratio imputation; Taylor linearization; Unit nonresponse; Weighting adjustment.

A. Demnati, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6
E-mail: demnabd@statcan.ca

J. N. K. Rao, Carleton University, School of Mathematics and Statistics, Ottawa, Ontario, Canada, K1S 5B6
E-mail: jrao@math.carleton.ca.

INDEX

1. INTRODUCTION.....	7
2. FULL RESPONSE	9
3. ITEM NONRESPONSE	11
4. NEW METHOD: MISSING RESPONSES.....	13
CONCLUDING REMARKS	17
REFERENCES.....	17

1. INTRODUCTION

Taylor linearization is a popular method of variance estimation for complex statistics such as ratio and regression estimators and logistic regression coefficient estimators. It is generally applicable to any sampling design that permits unbiased variance estimation for linear estimators, and it is computationally simpler than a resampling method such as the jackknife. However, it can lead to multiple variance estimators that are asymptotically design unbiased under repeated sampling. The choice among the variance estimators, therefore, requires other considerations such as (i) approximate unbiasedness for the model variance of the estimator under an assumed model, (ii) validity under a conditional repeated sampling framework. For example, in the context of simple random sampling and the ratio estimator, $\hat{Y}_R = (\bar{y}/\bar{x})X$, of the population total Y , Royall and Cumberland (1981) showed that a commonly used linearization variance estimator $v_L = N^2(n^{-1} - N^{-1})s_z^2$ does not track the conditional variance of \hat{Y}_R given \bar{x} , unlike the jackknife variance estimator v_J . Here \bar{y} and \bar{x} are the sample means, X is the known population total of an auxiliary variable x , s_z^2 is the sample variance of the residuals $z_i = y_i - (\bar{y}/\bar{x})x_i$ and (n, N) denote the sample and population sizes. By linearizing the jackknife variance estimator, v_J , we obtain a different linearization variance estimator, $v_{JL} = (\bar{X}/\bar{x})^2 v_L$, which also tracks the conditional variance as well as the unconditional variance, where $\bar{X} = X/N$ is the mean of x . As a result, v_{JL} or v_J may be preferred over v_L . Yung and Rao (1996) considered generalized regression and ratio-adjusted post-stratified estimators under stratified multistage sampling and obtained a jackknife linearization variance estimator, v_{JL} , by linearizing v_J . Valliant (1993) also obtained v_{JL} for the post-stratified estimator and conducted a simulation study to demonstrate that both v_J and v_{JL} possess good conditional properties given the estimated post-strata counts. Särndal, Swensson and Wretman (1989) showed that v_{JL} is both asymptotically design unbiased and asymptotically model unbiased in the sense of $E_m(v_{JL}) = Var_m(\hat{Y}_R)$, where E_m denotes model expectation and $Var_m(\hat{Y}_R)$ is the model variance of \hat{Y}_R under a “ratio model”: $E_m(y_i) = \beta x_i$; $i = 1, \dots, N$ and the y_i ’s are independent with model variance $Var_m(y_i) = \sigma^2 x_i$, $\sigma^2 > 0$. Thus, v_{JL} is a good choice from either the design-based or the model-based perspective. Demnati and Rao (2001) proposed a new approach to variance estimation that is theoretically justifiable and at the same time leads directly to a v_{JL} -type variance estimator for general designs. This method is presented in section 2.

In the presence of missing responses, weighting adjustment is often used to compensate for complete nonresponse while imputation is commonly used with the goal of making the data complete and obtaining estimates from the complete data. However, treating the adjusted weights as the design weights, the imputed values as true values and applying standard variance estimation formulas can lead to serious underestimation if the nonresponse rate is appreciable. In recent years, several methods that correctly estimate the variance of an estimator under imputation have been proposed. Rao (1996), Shao and Steel (1999) and others studied variance estimation under ratio imputation, while variance estimation under both weighting adjustment and imputation remains unexplored.

The main purpose of this paper is to extend the Demnati and Rao (2001) method to the case of missing responses when adjustment for complete nonresponse and imputation based on smooth functions of observed values, in particular ratio imputation, are used. Section 2 gives a brief account of the method for the case of full response. The method of Shao and Steel (1999) is described in section 3, while section 4 presents the extension of Demnati-Rao method.

2. FULL RESPONSE

To motivate the Demnati-Rao (2001) method for full response, suppose an estimator $\hat{\theta}$ of a parameter θ can be expressed as a differentiable function $g(\hat{Y})$ of estimated totals $\hat{Y}=(\hat{Y}_1, \dots, \hat{Y}_m)^T$, where $\hat{Y}_j = \sum_{i \in U} d_i(s) y_{ij}$ is an estimator of the population Y_j , $j=1, \dots, m$, where $d_i(s)=0$ if the unit i is not in the sample s , U is the set of population units, and $\theta=g(Y)$ with $Y=(Y_1, \dots, Y_m)^T$. We may write $\hat{\theta}$ as $\hat{\theta}=f(d(s), A_y)$ and $\theta=f(1, A_y)$, where A_y is an $m \times N$ matrix with j^{th} column $y_j=(y_{1j}, \dots, y_{mj})^T$, $j=1, \dots, N$, $d(s)=(d_1(s), \dots, d_N(s))^T$ and 1 is the N -vector of 1's. For example, if $\hat{\theta}$ denotes the ratio estimator $\hat{Y}_R = [\sum_{i \in U} d_i(s) y_i] / [\sum_{i \in U} d_i(s) x_i] X$, then $m=2$, $y_{1i}=y_i$, $y_{2i}=x_i$ and $f(1, A_y)$ reduces to the total Y , noting that $(Y/X)X=Y$. Note that \hat{Y}_R is a function of $d(s)$, y and x and the known total X , but we dropped X for simplicity and write $\hat{Y}_R=f(d(s), y, x)$. If the Horvitz-Thompson weights are used, then $d_i(s)=1/\pi_i$ for $i \in s$, where π_i is the probability of selecting unit i in the sample s .

Let $\check{Y}=\sum b_i y_i$ for arbitrary real numbers $b=(b_1, \dots, b_N)^T$, and $f(b, A_y)=f(b)$. Demnati and Rao (2001) showed that the Taylor linearization of $\hat{\theta}-\theta$, namely

$$\hat{\theta}-\theta=g(\hat{Y})-g(Y) \approx (\partial g(a)/\partial a)|_{a=Y} (\hat{Y}-Y),$$

is equivalent to

$$\begin{aligned} \hat{\theta}-\theta &\approx \sum_{k=1}^N \left(\frac{\partial f(b)}{\partial b_k} \right) \Big|_{b=1} (d_k(s) - 1) \\ &= \tilde{z}^T (d(s) - 1), \end{aligned} \tag{2.1}$$

where $\partial g(a)/\partial a=(\partial g(a)/\partial a_1, \dots, \partial g(a)/\partial a_m)^T$ and $\tilde{z}=(\tilde{z}_1, \dots, \tilde{z}_N)^T$ with $\tilde{z}_k=\partial f(b)/\partial b_k|_{b=1}$. It follows from (2.1) that a variance estimator of $\hat{\theta}$ is approximately given by the variance estimator of the estimated total $\sum d_i(s) \tilde{z}_i = \hat{Y}(\tilde{z})$; that is, $\text{var}(\hat{\theta}) \approx v(\tilde{z})$, where $v(y)$ denotes the variance estimator of $\hat{Y}=\hat{Y}(y)$ in operator notation. Now we replace \tilde{z}_k by $z_k=\partial f(b)/\partial b_k|_{b=d(s)}$, since \tilde{z}_k 's are unknown, to get a linearization variance estimator

$$v_L(\hat{\theta})=v(z). \tag{2.2}$$

Note that $v_L(\hat{\theta})$ given by (2.2) is simply obtained from the formula $v(y)$ for $\hat{Y}=\hat{Y}(y)$ by replacing y_i by z_i for $i \in s$. Note that we do not first evaluate the partial derivatives $\partial f(b)/\partial b_k$ at $b=1$ to get \tilde{z} and then substitute estimates for the unknown components of \tilde{z} . Our method, therefore, is similar in spirit to Binder(1996)'s approach. The variance estimator $v_L(\hat{\theta})$ is valid because z_i is a consistent estimator of \tilde{z}_i .

Suppose $\hat{\theta}$ is the ratio estimator $\hat{Y}_R = X \left[(\sum d_i(s) y_i) / (\sum d_i(s) x_i) \right]$, where Σ denotes summation over $i \in U$. Then $f(b) = X \left[(\sum b_i y_i) / (\sum b_i x_i) \right] = X \hat{Y}(b) / \hat{X}(b)$ and

$$z_k = \frac{\partial f(b)}{\partial b_k} \Big|_{b=d(s)} = \frac{X}{\hat{X}} \left(y_k - \hat{R} x_k \right).$$

For simple random sampling, $v_L(\hat{Y}_R) = v(z)$ agrees with $v_{JL} = (\bar{X}/\bar{x})^2 v_L$.

Demnati and Rao (2001) applied the method to a variety of problems, covering regression calibration estimators of a total Y and other estimators defined either explicitly or implicitly as solutions of estimating equations. They obtained a new variance estimator for a general class of calibration estimators that includes generalized raking ratio and generalized regression estimators. They also extended the method to two-phase sampling and obtained a variance estimator that makes fuller use of the first phase sample data compared to traditional linearization variance estimators.

3. ITEM NONRESPONSE

Following Fay (1991), Shao and Steel (1999) proposed a method of deriving variance estimators for the Horvitz-Thompson-type estimated total, \hat{Y}^* , with imputed item nonresponse values. They assumed that the estimated total \hat{Y}^* can be expressed as a smooth function of totals, $\hat{Y}^* = \psi(\hat{T}_o)$, where $\hat{T}_o = \sum d_i(s) \text{diag}(\varrho_i) t_i$, t_{ki} is the value of y_i or the value of some other variable used to impute y_i , and $\varrho_i = (o_{i1}, \dots, o_{ip})^T$ is a vector of response indicator variables. For example, consider ratio imputation when the auxiliary variable x_i is available for all $i \in s$. A missing y_i is then imputed by $\hat{y}_i = \hat{R}_o x_i$, where $\hat{R}_o = (\sum d_i(s) o_i y_i) / (\sum d_i(s) o_i x_i)$ and o_i is the response indicator for y_i , i.e. $o_i = 1$ if y_i is observed and $o_i = 0$ if y_i is missing. The imputed estimator \hat{Y}^* is given by

$$\begin{aligned}\hat{Y}^* &= \sum d_i(s) o_i y_i + \sum d_i(s) (1 - o_i) \hat{R}_o x_i \\ &= \sum d_i(s) o_i y_i \left(1 + \sum d_i(s) o_{2i} x_i / \sum d_i(s) o_{1i} x_i \right)\end{aligned}\quad (3.1)$$

where $o_{i1} = o_i$ and $o_{i2} = 1 - o_i$. It follows from (3.1) that \hat{Y}^* is of the form $\psi(\hat{T}_o)$ with $\varrho_i = (o_{i1}, o_{i2}, o_{i3})^T$ and $t_i = (y_i, x_i, x_i)^T$.

We assume deterministic imputation. We have $\text{Var}(\hat{Y}^* - Y) = V_1 + V_2$, where $V_1 = E_o[\text{Var}_s(\hat{Y}^* - Y)]$, $V_2 = \text{Var}_o[E_s(\hat{Y}^* - Y)]$, E_o and Var_o stand for the expectation and variance with respect to the response mechanism, and E_s and Var_s stand for the expectation and variance with respect to sampling under a given design. Shao and Steel (1999) obtained a variance estimator, v_1 , of V_1 using a standard linearization variance estimator of $\psi(\hat{T}_o)$ for given ϱ_i 's. They also obtained an estimator, v_2 , of $V_2 \approx [\nabla \phi(E_o \hat{T}_o)]^T C [\nabla \phi(E_o \hat{T}_o)]$, where $\phi(\hat{T}_o) = \psi(E_o \hat{T}_o) - Y$, by deriving C with kl^{th} element $c_{kl} = \text{cov}_o(\sum o_{ki} t_{ki}, \sum o_{li} t_{li})$ and by substituting estimators for the unknown quantities. For simple random sampling and ratio imputation, Shao and Steel (1999) obtained v_1 as

$$v_1 = N^2 \frac{(1 - n/N)}{n(n-1)} \left[\left(\frac{\bar{x}}{\bar{x}_o} \right)^2 \frac{s_d^2}{n_o} + 2 \frac{\bar{x}}{\bar{x}_o} \frac{\hat{R}_o s_{dx}}{n} + \frac{\hat{R}_o^2 s_x^2}{n} \right], \quad (3.2)$$

where \bar{x} and s_x^2 are the sample mean and sample variance of the x_i 's, $\bar{x}_o = \sum_{i \in s} o_i x_i / n_o$ is the mean of x_i 's for the respondents, n_o is the number of respondents, $s_d^2 = \sum_{i \in s} o_i (y_i - \hat{R}_o x_i)^2 / (n_o - 1)$, and $s_{dx} = \sum_{i \in s} o_i x_i (y_i - \hat{R}_o x_i) / (n_o - 1)$.

Further, under the assumption of uniform response (i.e., that the o_i 's are independent and identically distributed with mean p_y and variance $p_y(1 - p_y)$), Shao and Steel (1999) obtained v_2 as

$$v_2 = [X/X_o]^2 \hat{p}_y (1 - \hat{p}_y) N s_d^2, \quad (3.3)$$

where $\hat{p}_y = \sum_i o_i d_i(s) / \sum_i d_i(s)$. The sum of (3.2) and (3.3) gives the variance estimator of \hat{Y}^* .

Shao and Steel's (1999) method is based on the classical linearization approach which consists of (i) expressing the estimator in terms of elementary components, (ii) evaluating the partial derivatives at the population level and (iii) then estimating the unknown parameters in the formula. As a result, the corresponding variance estimator may not be unique. Our method avoids expressing the estimator in terms of elementary components and thus leads directly to a unique variance estimator with desirable properties. We present our method for ratio imputation in subsection 4.1, while the case of variance estimation under weighting adjustment for complete nonresponse and ratio imputation for item nonresponse is investigated in subsection 4.2.

4. NEW METHOD: MISSING RESPONSES

After weight adjustment for complete nonresponse and imputation for item nonresponse, the population total \hat{Y} is estimated by a weighted sample total

$$\hat{Y}^* = \sum \tilde{w}_i(s) o_i y_i + \sum \tilde{w}_i(s)(1 - o_i) \hat{y}_i^*, \quad (4.1)$$

where $\tilde{w}_i(s)$ is the adjusted weight and \hat{y}_i^* denote the imputed value for unit i . The estimator (4.1) can be rewritten as

$$\hat{Y}^* = \sum \tilde{w}_i(s) \hat{y}_i = \hat{\mathcal{L}}^T \tilde{\mathcal{W}}(s), \quad (4.2)$$

where $\hat{\mathcal{L}} = (\hat{y}_1, \dots, \hat{y}_N)^T$ and $\hat{y}_i = o_i y_i + (1 - o_i) \hat{y}_i^*$. In subsection 4.1, we study the case of item nonresponse only (i.e., $\tilde{w}_i(s) = d_i(s)$) assuming \hat{Y}^* can be expressed as a smooth function of totals $\sum d_i(s) \text{diag}(\varrho_j) t_i$, where t_{ki} the value of y_{ij} or the value of some other variable used to impute y_{ij} . Subsection 4.2 deals with the more general case of weight adjustment for complete nonresponse and imputation for item nonresponse.

4.1. Imputation for item nonresponse

The imputed estimator \hat{Y}^* is assumed to be a smooth function of totals $\sum d_i(s) \text{diag}(\varrho_j) t_i$, as in Shao and Steel (1999). In this case, \hat{Y}^* may be expressed as $f(\mathcal{A}_w, \mathcal{A}_y)$, where $\mathcal{A}_w = \text{diag}(d(s)) \mathcal{A}_o$ is an $m \times N$ matrix with j^{th} column $\underline{w}_j(s) = (w_{1j}(s), \dots, w_{mj}(s))^T$, $j=1, \dots, N$. The vector $\underline{w}_j(s)$ is defined as

$$\begin{aligned} \underline{w}_j(s) &= (w_{1j}(s), \dots, w_{mj}(s))^T \\ &= (o_{1j} d_j(s), \dots, o_{mj} d_j(s))^T = \varrho_j d_j(s), \end{aligned}$$

where $\varrho_j = (o_{1j}, \dots, o_{mj})^T$ is the vector of indicator variables corresponding to the vector $\underline{y}_j = (y_{1j}, \dots, y_{mj})^T$. For simplicity we drop \mathcal{A}_y and denote $\hat{Y}^* = f(\mathcal{A}_w)$. Under ratio imputation, we have $m=2$, $y_{1j}=x_j$, $y_{2j}=y_j$, $o_{1j}=1$, $o_{2j}=o_j$, $\underline{w}_j(s) = (w_{1j}(s), w_{2j}(s))^T = (d_j(s), o_j d_j(s))^T$ and

$$\hat{Y}^* = \sum w_{2i}(s) (y_i - \hat{R}_o x_i) + \sum w_{1i}(s) \hat{R}_o x_i,$$

with $\hat{R}_o = (\sum w_{2i}(s) y_i) / (\sum w_{2i}(s) x_i)$.

Because the estimator $\hat{Y}^* = f(\mathcal{A}_w)$ is a function of totals, we can use Demnati and Rao (2001) approach to approximate its variance by the variance of a linear function

$$\text{Var}(\hat{Y}^*) \approx \text{Var}(\hat{Y}_L^*)$$

with

$$\hat{Y}_L^* = \sum (\varrho_i d_i(s))^T \tilde{\mathcal{Z}}_i = \sum \underline{w}_i^T(s) \tilde{\mathcal{Z}}_i,$$

where \tilde{z}_i is the vector of derivatives of $f(\mathcal{A}_b)$ with respects to b_k evaluated at $\mathcal{A}_b = E(\mathcal{A}_w)$, where \mathcal{A}_b is a $N \times m$ matrix of arbitrary real numbers, $f(\mathcal{A}_b)$ is obtained by replacing \mathcal{A}_w by \mathcal{A}_b in the formula for \hat{Y}^* and b_k is a column vector of \mathcal{A}_b . The total variance of \hat{Y}_L^* can then be estimated by

$$v(\hat{Y}_L^*) = v_s(\mathcal{Q}^T \mathcal{Z}) + v_o(z), \quad (4.4)$$

where \mathcal{Z}_k is the vector of derivatives of the estimator $f(\mathcal{A}_b)$ with respects to b_k evaluated at $\mathcal{A}_b = \mathcal{A}_w$, and $v_o(z)$ is an estimator of $V(\sum \text{diag}(\mathcal{Q}_i) z_i)$. Under independent response mechanism,

$$v_o(z) = \sum \mathcal{Z}_i^T \text{cov}_o(\mathcal{Q}_i) \mathcal{Z}_i, \quad (4.5)$$

where $\text{cov}_o(\mathcal{Q}_i)$ is an (approximately) unbiased estimator of $E(\mathcal{Q}_i \mathcal{Q}_i^T) - E(\mathcal{Q}_i) E(\mathcal{Q}_i^T)$.

Under ratio imputation, we have

$$\begin{aligned} \mathcal{Z}_k &= \frac{\partial}{\partial b_k} \left(\sum b_{2i} (y_i - \hat{R}_o(\mathcal{A}_b) x_i) + \sum b_{1i} \hat{R}_o(\mathcal{A}_b) x_i \right) \Big|_{\mathcal{A}_b = \mathcal{A}_w} \\ &= \left(\hat{R}_o x_k, (\hat{X}/\hat{X}_o)(y_k - \hat{R}_o x_k) \right)^T. \end{aligned} \quad (4.6)$$

It follows from (4.6) that

$$z_k = o_k(\hat{X}/\hat{X}_o)(y_k - \hat{R}_o x_k) + \hat{R}_o x_k. \quad (4.7)$$

Therefore, $v_s(\mathcal{Q}^T \mathcal{Z})$ equals $v_1 = v(z)$. Under simple random sampling, $v(z)$ with z_k given by (4.7) agrees with (3.2) of Shao and Steel (1999). Further,

$$\text{cov}_o(\mathcal{Q}_i) = d_i(s) \begin{pmatrix} 0 & 0 \\ 0 & o_i(1 - \hat{\xi}_{io}) \end{pmatrix},$$

where $\hat{\xi}_{io}$ is an estimator of probability of response for unit i . Therefore, $v_o(z)$, given by (4.5), reduces to

$$v_o(z) = (\hat{X}/\hat{X}_o)^2 \sum d_i(s) o_i (1 - \hat{\xi}_{io}) (y_i - \hat{R}_o x_i)^2. \quad (4.8)$$

Under simple random sample and uniform response mechanism (4.8) reduces to

$$v_o(z) = \frac{N}{n} (\hat{X}/\hat{X}_o)^2 (1 - n_o/n) \sum o_i (y_i - \hat{R}_o x_i)^2 \quad (4.9)$$

which is the Shao and Steel (1999) estimator v_2 given by (3.3).

4.2. Weight adjustment and imputation for item nonresponse

Let r_i be the partial response indicator variable for the i^{th} unit, i.e. $r_i = 0$ if there is complete nonresponse and $r_i = 1$ if there is partial response. The partial response indicator variable r_i is related to the item response variable indicators o_p , $p=1\dots m$ by

$$r_i = 1 - \prod_{p=1}^m (1 - o_{ip}). \quad (4.10)$$

We have

$$\text{Cov}(r_i o_{ip}) = E(r_i o_{ip}) - E(r_i)E(o_{ip}),$$

for any response variable indicator o_{ip} . Noting that $r_i o_{ip} = o_{ip}$ for any o_{iq} ,

$$r_i o_{ip} = [1 - (1 - o_{ip}) \prod_{q \neq p} (1 - o_{iq})] o_{ip} = o_{ip}.$$

Hence,

$$\text{Cov}(r_i o_{ip}) = E(o_{ip}) - E(r_i)E(o_{ip}) = E(o_{ip})(1 - E(r_i)).$$

An estimator of $\text{Cov}(r_i o_{ip})$ maybe taken as

$$\text{cov}(r_i o_{ip}) = o_{ip}(1 - \hat{\xi}_{ir})$$

with $\hat{\xi}_{ir} = \hat{E}(r_i)$ and $\hat{E}(\cdot)$ denotes an estimator for $E(\cdot)$.

A widely-used approach to adjust for complete nonresponse is to employ a new set of weights, $\tilde{w}_i(s)$, with i^{th} element equals to

$$\tilde{w}_i(s) = d_i(s) r_i g_i(d(s), \mathcal{L}, \mathcal{A}_\chi), \quad (4.11)$$

where $g_i(d(s), \mathcal{L}, \mathcal{A}_\chi)$ is known as the g-weights in the context of regression estimator and \mathcal{A}_χ a matrix of auxiliary variables known for all units in the sample. The ratio estimator is a special case of (4.11) for which the g-weight reduces to

$$g_i(d(s), \mathcal{L}, \mathcal{A}_\chi) = \frac{\sum d_i(s) \chi_i}{\sum d_i(s) r_i \chi_i} = \frac{\hat{\chi}}{\hat{\chi}_r}, \quad (4.12)$$

where $\hat{\chi} = \sum d_i(s) \chi_i$ and $\hat{\chi}_r = \sum d_i(s) r_i \chi_i$. The weight adjustment using the ratio (4.12) is a special case of the class of calibration weights obtained through the regression estimator. Generalized raking weights are also used to compensate for complete nonresponse. Another way to adjust for complete non-response is to weight each observation by the inverse probability of responding in which case

$$g_i(d(s), \mathcal{L}, \mathcal{A}_\chi) = \hat{\xi}_{ir}^{-1},$$

where

$$\hat{\xi}_{ir} = \xi_{ir}(\mathcal{L}, d(s)) = \Pr(r_i = 1 | d(s), \mathcal{A}_\chi),$$

is the estimator of probability of response defined as solution to an estimating equation of the form

$$\hat{U}(\hat{\xi}_r) = \sum d_i(s) u_i(r_i, \chi_i, \hat{\xi}_{ir}) = 0.$$

In the logistic case, we have

$$\hat{U}(\hat{\xi}_r) = \sum d_i(s) (r_i - \hat{\xi}_{ir}) \chi_i = 0,$$

where

$$u_i(r_i, \chi_i, \hat{\xi}_{ir}) = (r_i - \hat{\xi}_{ir}) \chi_i,$$

$$\hat{\xi}_{ir} = \exp(\chi_i^T \beta) / (1 + \exp(\chi_i^T \beta)) = Pr(r_i = 1 | \chi_i, \beta),$$

and χ_i is the vector of predictor variables.

Under the above weight adjustment methods, the variance can be obtained along the line of Demnati and Rao (2001) method by expressing \hat{Y}^* as $f(\mathcal{A}_w)$ and then by differentiating $f(\mathcal{A}_b)$ with respect to b_k . Details are omitted for simplicity but we illustrate the calculation for the estimator (4.1) under the ratio weight adjustment (4.12) and ratio imputation, i.e.,

$$\hat{Y}^* = \sum_i o_i \tilde{w}_i(s) y_i + \sum_j (1 - o_j) \tilde{w}_j(s) \hat{R}_o x_j,$$

with

$$\hat{R}_o = \frac{\sum \tilde{w}_i(s) o_i y_i}{\sum \tilde{w}_i(s) o_i x_i},$$

and

$$\tilde{w}_i(s) = d_i(s) r_i \hat{\chi} / \hat{\chi}_r.$$

We have

$$w_i(s) = (1, o_i, r_i)^T d_i(s),$$

$$z_k = \left(x_k \hat{R}_o (\hat{X}_r / \hat{\chi}_r), (\hat{\chi} / \hat{\chi}_r) (\hat{X}_r / \hat{R}_o) (y_k - \hat{R}_o x_k), (\hat{\chi} / \hat{\chi}_r) \hat{R}_o (x_k - (\hat{X}_r / \hat{\chi}_r) \chi_k) \right)^T,$$

and

$$cov_o(Q_i^T) = d_i(s) \begin{pmatrix} 0 & 0 & 0 \\ 0 & o_i(1 - \hat{\xi}_{io}) & o_i(1 - \hat{\xi}_{ir}) \\ 0 & o_i(1 - \hat{\xi}_{ir}) & r_i(1 - \hat{\xi}_{ir}) \end{pmatrix}.$$

CONCLUDING REMARKS

We have presented a new approach to variance estimation under missing responses. A valid variance estimator is given under a variety of weighting adjustment methods often used for unit nonresponse as well as under imputation based on smooth functions of observed values, in particular ratio imputation, which often used for item nonresponse. Extensions to nearest neighbor imputation and panel surveys are under investigation.

REFERENCES

Binder, D. (1996), "Linearization Methods for Single Phase and Two-Phase Samples: A Cookbook Approach", *Survey Methodology*, 22, 17-22.

Demnati, A. and Rao, J. N. K. (2001), Linearization Variance Estimators for Survey Data, Methodology Branch Working Paper, SSMD-2001-010E. Statistics Canada.

Fay, R. E. (1991), "A Design-Based Perspective on Missing Data Variance", in *Proceeding of the 1991 Annual Research Conference, US Bureau of the census*, 429-440.

Rao, J. N. K. (1996), "On Variance Estimation With Imputed Survey Data (with discussion)", *Journal of the American Statistical Association*, 91, 499-520.

Royall, R. M., and Cumberland, W. G. (1981), "An Empirical Study of the Ratio Estimator and Estimators of its Variance", *Journal of the American Statistical Association*, 76, 66-77.

Särndal, C.-E., Swensson, B., and Wretman, J.H. (1989), "The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total", *Biometrika*, 76, 527-537.

Shao, J. and Steel, P. (1999), Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions, *Journal of the American Statistical Association*, 94, 254-265.

Valliant, R. (1993), "Poststratification and Conditional Variance Estimation", *Journal of the American Statistical Association*, 88, 89-96.

Yung, W. and Rao, J. N. K. (1996), "Jackknife Linearization Variance Estimators under Stratified Multi-Stage Sampling", *Survey Methodology*, 22, 23-31.

- Nous avons présenté une nouvelle méthode d'estimation de la variance en cas de réponses manquantes. Nous donnons un estimateur valide de la variance pour diverses méthodes d'ajustement de pondération utilisées fréquemment pour tenir compte de la non-réponse partielle. L'estimation à l'imputation par fonction losses des valeurs observées, en particulier l'imputation par quotient, qui est utilisée fréquemment pour tenir compte de la non-réponse partielle. L'estimation à l'imputation basée sur des fonctions losses des valeurs observées, en particulier l'imputation par quotient basee sur des fonctions losses des valeurs observées, en particulier l'imputation par quotient, qui est utilisée fréquemment pour tenir compte de la non-réponse partielle. L'estimation à l'imputation par quotient basee sur des fonctions losses des valeurs observées, en particulier l'imputation par quotient, qui est utilisée fréquemment pour tenir compte de la non-réponse partielle. L'estimation à l'imputation par quotient basee sur des fonctions losses des valeurs observées, en particulier l'imputation par quotient, qui est utilisée fréquemment pour tenir compte de la non-réponse partielle. L'estimation à l'imputation par quotient basee sur des fonctions losses des valeurs observées, en particulier l'imputation par quotient, qui est utilisée fréquemment pour tenir compte de la non-réponse partielle.
- Binder, D. (1996), « Linearization methods for single phase and two-phase samples: a cookbook approach», *Survey Methodology*, 22, 17-22.
- Demnati, A. and Rao, J. N. K. (2001), Linearization variance estimators for survey data, *Methodology Branch Working Paper*, SSM-D-2001-010E, Statistics Canada.
- Fay, R. E. (1991), « A design-based perspective on missing data variance», in *Proceeding of the 1991 Annual Research Conference, US Bureau of the census*, 429-440.
- Rao, J. N. K. (1996), « On variance estimation with imputed survey data (with discussion)», *Journal of the American Statistical Association*, 91, 499-520.
- Royall, R. M., and Cumberland, W. G. (1981), « An empirical study of the ratio estimator and estimators of its variance», *Journal of the American Statistical Association*, 76, 66-77.
- Sarndal, C.-E., Swensson, B., and Wretman, J.H. (1989), « The Weighted residual technique for estimating the variance of the general regression estimator of the finite population total», *Biometrika*, 76, 527-537.
- Shao, J. and Steel, P. (1999), « Variance estimation for survey data with composite imputation and nonnegligible sampling fractions», *Journal of the American Statistical Association*, 94, 254-265.
- Valliant, R. (1993), « Poststratification and conditional variance estimation», *Journal of the American Statistical Association*, 88, 89-96.
- Yung, W. and Rao, J. N. K. (1996), « Jackknife linearization variance estimators under stratified multi-Stage sampling», *Survey Methodology*, 22, 23-31.

Bibliographie

Le plus proche voisin et aux enquêtes par panel est à l'étude.

qui est utilisée fréquemment pour tenir compte de la non-réponse partielle. L'estimation à l'imputation par fonction losses des valeurs observées, en particulier l'imputation par quotient, qui est utilisée fréquemment pour tenir compte de la non-réponse partielle. L'estimation à l'imputation par fonction losses des valeurs observées, en particulier l'imputation par quotient, qui est utilisée fréquemment pour tenir compte de la non-réponse partielle. L'estimation à l'imputation par fonction losses des valeurs observées, en particulier l'imputation par quotient, qui est utilisée fréquemment pour tenir compte de la non-réponse partielle. L'estimation à l'imputation par fonction losses des valeurs observées, en particulier l'imputation par quotient, qui est utilisée fréquemment pour tenir compte de la non-réponse partielle. L'estimation à l'imputation par fonction losses des valeurs observées, en particulier l'imputation par quotient, qui est utilisée fréquemment pour tenir compte de la non-réponse partielle.

Nous avons présenté une nouvelle méthode d'estimation de la variance en cas de réponses manquantes. Nous donnons un estimateur valide de la variance pour diverses méthodes d'ajustement de pondération utilisées fréquemment pour tenir compte de la non-réponse partielle. L'estimation à l'imputation par quotient basee sur des fonctions losses des valeurs observées, en particulier l'imputation par quotient, qui est utilisée fréquemment pour tenir compte de la non-réponse partielle.

Conclusion



et

$$\int \left((\chi \chi / X) - x \chi^o Y (\chi \chi) (\chi^o Y - \chi) (\chi / X) (\chi \chi) (\chi / X) \chi^o Y x \right) = \tilde{z}$$

$$(s)^i d_{\underline{J}} (s^i o^i x^i) = (s)^i w^i$$

Nous avons

$$\chi / \chi^i d(s)^i p = (s)^i \underline{w}$$

et

$$\frac{\chi^i o(s) \underline{w}}{\chi^i o(s) \underline{w}} = \underline{Y}$$

avec

$$\tilde{y}_i^o = \sum_j o_j w_j(s) y_j + \sum_j (1 - o_j) \tilde{w}_j(s) \tilde{P}^o x_j,$$

sous ajustement de la pondération par quotient (4.12) et imputation par quotient, c'est-à-dire rapport à \tilde{x}_i . Nous omittons les détails par souci de simplicité, mais nous illustrons le calcul pour l'estimateur (4.1) suivant la méthode de Demnati et Rao (2001) en exprimant \tilde{y}_i^o sous la forme $f(\tilde{A}^o)$, puis en dérivant $f(\tilde{A}^o)$ par rapport aux méthodes d'ajustement de la pondération susmentionnée, nous pouvons obtenir la variance en

et \tilde{x}_i^o est le vecteur des variables explicatives.

$$\tilde{\xi}_{ii}^o = \exp(\tilde{x}_i^T \tilde{\beta}) / (1 + \exp(\tilde{x}_i^T \tilde{\beta})) = \Pr(r_i^o = 1 | \tilde{x}_i^o, \tilde{\beta}),$$

$$u_i(r_i^o, \tilde{x}_i^o, \tilde{\xi}_{ii}^o) = (r_i^o - \tilde{\xi}_{ii}^o) \tilde{x}_i^o,$$

ou

$$U(\tilde{\xi}_{ii}^o) = \sum_i d_i(s) u_i(r_i^o, \tilde{x}_i^o, \tilde{\xi}_{ii}^o) = 0,$$

Dans le cas logistique, nous avons

$$U(\tilde{\xi}_{ii}^o) = \sum_i d_i(s) u_i(r_i^o, \tilde{x}_i^o, \tilde{\xi}_{ii}^o) = 0.$$

est l'estimateur de la probabilité de réponse défini comme étant la solution d'une équation d'estimation de la forme

$$\tilde{\xi}_{ii}^o = \tilde{\xi}_{ii}^o(\tilde{x}, \tilde{d}(s)) = \Pr(r_i^o = 1 | \tilde{d}(s), \tilde{A}^o),$$

ou

$$g_i(\tilde{d}(s), \tilde{x}, \tilde{A}^o) = \tilde{\xi}_{ii}^o,$$

réponse, auquel cas

ou $\tilde{x} = \sum_i d_i(s) x_i$ et $\tilde{x}_i^o = \sum_i d_i(s) r_i^o x_i$. L'ajustement des poids au moyen du quotient (4.12) est un cas spécial de la classe de poids de calage obtenus en utilisant l'estimateur par régression. On utilise également les poids obtenus par ajustement itératif du quotient génératrice pour compenser pour la non-réponse complète. Un autre moyen de tenir compte de la non-réponse complète consiste à pondérer chaque observation par la probabilité inverse de

$$g_i(\tilde{d}(s), \tilde{x}, \tilde{A}^o) = \frac{\sum_i d_i(s) r_i^o x_i}{\sum_i d_i(s) x_i}, \quad (4.12)$$

L'estimateur par quotient est un cas particulier de (4.11) pour lequel les poids y se reduisent à une matrice de variables auxiliaires dont la valeur est connue pour toutes les unités figurant dans l'échantillon.

où les $g_i(\tilde{d}(s), \tilde{x}, \tilde{A}^o)$ sont connus soit le nom de poids g dans le contexte de l'estimateur par régression et \tilde{A}^o est

$$(4.11) \quad \tilde{w}_i(s) = d_i(s) r_i g_i(\tilde{d}(s), x, \tilde{A}^x),$$

ensembles de poids, $\tilde{w}_i(s)$, dont le i^{e} élément est égal à
Une méthode très répandue de correction pour la non-réponse complète consiste à employer un nouvel

où $\zeta_{ip} = E(r_i)$ et $E(\cdot)$ représente un estimateur de $E(\cdot)$.

$$\text{cov}(r_i, o_{ip}) = o_{ip}(1 - \zeta_{ip})$$

Nous pouvons donner un estimateur de $\text{Cov}(r_i, o_{ip})$ sous la forme

$$\text{Cov}(r_i, o_{ip}) = E(o_{ip}) - E(r_i)E(o_{ip}) = E(o_{ip})(1 - E(r_i)) .$$

Donc,

$$r_i o_{ip} = [1 - (1 - o_{ip}) \prod_{j \neq p} (1 - o_{jp})] o_{ip} .$$

pour toute variable indicatrice de réponse o_{ip} . Notant que $r_i o_{ip} = o_{ip}$ pour tout o_{ip} ,

$$\text{Cov}(r_i, o_{ip}) = E(r_i o_{ip}) - E(r_i)E(o_{ip}),$$

Nous savons

$$(4.10) \quad r_i = 1 - \prod_{j=1}^d (1 - o_{ij}) .$$

Soit r_i , la variable indicatrice de réponse partielle pour la i^{e} unité, c.-à-d. $r_i = 1$ si la réponse est complète et $r_i = 0$ si la réponse est partielle. La variable indicatrice de réponse partielle r_i est reliée aux variables indicatrices de réponse à une question o_p , $p = 1 \dots m$ par

partielle

4.2 Ajustement de la pondération et imputation pour tenir compte de la non-réponse

qui est l'estimateur v_2 de Shao et Steel (1999) donné par (3.3).

$$(4.9) \quad v_o(\bar{z}) = \frac{n}{N} \left(\bar{X} / \bar{X}_o \right)^2 (1 - n_o/m) \sum_{i=1}^n (y_i - \bar{Y}_o x_i)^2$$

Dans le cas d'un échantillonnage aléatoire simple et d'un mécanisme de réponse uniforme, (4.8) se réduit à

$$(4.8) \quad v_o(\bar{z}) = (\bar{X} / \bar{X}_o)^2 \sum_{i=1}^n d_i(s) o_i (1 - \zeta_{io}) (y_i - \bar{Y}_o x_i)^2 .$$

à

où ζ_{io} est un estimateur de la probabilité de réponse de l'unité i . Par conséquent, $v_o(\bar{z})$, donné par (4.5), se réduit

$$\text{cov}_o(\tilde{o}_i) = d_i(s) \begin{pmatrix} (\tilde{\beta}_i^T I - \tilde{\beta}_i^T \tilde{\beta}_i) & 0 \\ 0 & 0 \end{pmatrix}$$

donne par (4.7) concorde avec l'expression (3.2) de Shao et Steel (1999). Aussi,

Par conséquent, $v_o(\tilde{z})$ est égal à $v_1 = v(z)$. Dans les conditions d'échantillonnage sélectoire simple, $v(z)$ avec z

$$(4.7) \quad z^k = o^k(X/X^o)(v^k - R^o x^k) + R^o x^k.$$

Il découle de (4.6) que

(4.6)

$$\begin{aligned} & \left(\tilde{R}^o x^k, (X/X^o)(v^k - R^o x^k) \right) = \\ & \tilde{R}^o = \frac{\sum b^k (v^k - R^o x^k)}{\sum b^k} = \tilde{z}^k \end{aligned}$$

Sous imputation par quotient, nous avons

où $\text{cov}_o(\tilde{o}_i)$ est un estimateur (approximativement) non biaisé de $E(\tilde{o}_i \tilde{o}_j^T) - E(\tilde{o}_i)E(\tilde{o}_j^T)$.

$$(4.5) \quad v_o(\tilde{z}) = \sum_{i=1}^n \text{cov}_o(\tilde{o}_i) \tilde{z}_i,$$

où \tilde{z}^k est le vecteur des dérivées de l'estimateur $f(\tilde{A}^k)$ par rapport à \tilde{q}^k évaluées à \tilde{A}^m , et $v_o(\tilde{z})$ est un estimateur de $\text{Var}(\sum_i \text{diag}(\tilde{o}_i) \tilde{z}_i)$. Sous un mécanisme de réponse indépendante,

$$(4.4) \quad v(Y^L_o) = v_s(\tilde{z}) + v_o(\tilde{z}),$$

où \tilde{z} est le vecteur des dérivées de $f(\tilde{A}^k)$ par rapport à \tilde{q}^k évaluées à $\tilde{A}^m = E(\tilde{A}^m)$, où \tilde{A}^k est une matrice $m \times N$ de nombres réels arbitraires, $f(\tilde{A}^k)$ est obtenue par remplacement de \tilde{A}^m par \tilde{A}^k dans la formule de Y^L_o et \tilde{q}^k est un vecteur colonne de \tilde{A}^k . Nous pouvons estimer la variance totale de Y^L_o selon

$$Y^L_o = \sum_i \tilde{o}_i d_i(s) \tilde{z}_i = \sum_i w_i^L(s) \tilde{z}_i,$$

avec

$$\text{Var}(Y^L_o) \approx \text{Var}(Y^L_o)$$

Demandé et Rao (2001) pour approximer sa variance au moyen de la variance d'une fonction linéaire comme l'estimateur $Y^L_o = f(\tilde{A}^m)$ est une fonction des totaux, nous pouvons appliquer la méthode de

$$\text{ou } R^o = (\sum w^{2i}(s)y_i) / (\sum w^{2i}(s)x_i)$$

$$Y^o = \sum w^{2i}(s)(x_i - \bar{x}^o) + \sum w^{1i}(s)\bar{x}^o$$

avons $m=2$, $y_{1j} = x_j$, $y_{2j} = y_j$, $o_{1j} = 1$, $o_{2j} = o_j$, $\bar{w}^f(s) = (w_1^f(s), w_2^f(s))^T$ et

de simplicité, nous laissons tomber A^y , et écrivons $\tilde{Y}^o = f(\tilde{A}^y)$. En cas d'imputation par quotient, nous

ou $\tilde{o}_j = (o_{1j}, \dots, o_{mj})^T$ est le vecteur des variables indicatrices correspondant au vecteur $\tilde{Y}^o = (y_{1j}, \dots, y_{mj})^T$. Par souci

$$\begin{aligned} &= (o_{1j}d^f(s), \dots, o_{mj}d^f(s))^T = \tilde{o}_j d^f(s) \\ &\quad \tilde{w}^f(s) = (w_1^f(s), \dots, w_m^f(s))^T \end{aligned}$$

$j=1, \dots, N$. Le vecteur $\tilde{w}^f(s)$ est défini comme étant

$f(\tilde{A}^y, A^y)$, où $\tilde{A}^y = \text{diag}(\tilde{o}_j)^T A^y$. est une matrice $m \times N$ dont la j^e colonne est $\tilde{w}^f(s) = (w_1^f(s), \dots, w_m^f(s))^T$,

total, $\sum d^f(s)\text{diag}(\tilde{o}_j)^T$, à l'instar de Shao et Steele (1999). Dans ce cas, nous pouvons exprimer \tilde{Y}^o sous la forme
Nous supposons que l'estimateur avec données imputées \tilde{Y}^o est une fonction lisse de

4.1 Imputation pour tenir compte de la non-réponse partielle

de la non-réponse complète et de l'imputation pour tenir compte de la non-réponse partielle.
à \tilde{Y}^o . À la sous-section 4.2, nous traitons le cas plus général de l'ajustement de la pondération pour tenir compte

de total $\sum d^f(s)\text{diag}(\tilde{o}_j)^T$, où t_{ij} est la valeur de y_i ou celle d'une autre variable utilisée pour imputer une valeur unique (c.-à-d. $w^i(s) = d^f(s)$) en supposant que nous pouvons exprimer \tilde{Y}^o sous la forme d'une fonction lisse
ou $\tilde{Y}^o = (\tilde{Y}^o_1, \dots, \tilde{Y}^o_N)^T$ et $\tilde{Y}^o_i = o_i Y^o_i + (1 - o_i) \tilde{Y}^o_i$. À la sous-section 4.1, nous étudions le cas de la non-réponse partielle

$$(4.2) \quad \tilde{Y}^o = \sum w^i(s) \tilde{Y}^o_i = \tilde{Y}^o \tilde{w}(s),$$

l'estimateur (4.1) sous la forme
ou $w^i(s)$ est le poids ajusté et \tilde{Y}^o_i représente la valeur imputée pour l'unité i . Nous pouvons réécrire

$$(4.1) \quad \tilde{Y}^o = \sum w^i(s)o_i Y^o_i + \sum w^i(s)(1 - o_i) \tilde{Y}^o_i,$$

après ajustement de la pondération pour tenir compte de la non-réponse complète et l'imputation pour tenir
compte de la non-réponse partielle, nous estimons le total de la population \tilde{Y} au moyen d'un total pondéré à partir

4. Nouvelle méthode: réponse manquante

La méthode de Shao et Steel (1999) est fondée sur la méthode de linéarisation classique qui consiste à i) exprimer l'estimateur en fonction des paramètres connus dans la formule, ii) calculer les dérivées partielles au niveau de la population et iii) estimer les paramètres nécessaires dans la formule. Par conséquent, l'estimateur correspondant de la variance n'est pas nécessairement unique. Notre méthode permet d'éviter d'exprimer l'estimateur en fonction de composantes élémentaires, donc produit directement un estimateur unique de la variance sans prendre souhaitées. Nous présentons notre méthode pour l'imputation par quotient à la sous-section 4.1, et nous examinons le cas de l'estimation de la variance sous ajustement par quotient pour tenir compte de la non-réponse partielle à la sous-section 4.2.

où $\hat{p}_y = \frac{\sum o_i d_i(s)}{\sum d_i(s)}$. La somme de (3.2) et (3.3) donne l'estimateur de la variance de y^o .

$$(3.3) \quad V^o = (X/X^o)^2 \hat{p}_y (1 - \hat{p}_y) N s_d^2,$$

Aussi, sous l'hypothèse de réponse uniforme (c.-à-d. si α_i sont indépendamment et identiquement distribuées en ayant une moyenne p_i et une variance $p_i(1-p_i)$), Shao et Steele (1999) ont obtenu V_2 sous la forme

répondants, n_o est le nombre de répondants, $s_d^2 = \sum_{i \in s} (y_i - \bar{R}_o x_i)^2 / (n_o - 1)$, et $s_{dx} = \sum_{i \in s} \alpha_i x_i (y_i - \bar{R}_o x_i)^2 / (n_o - 1)$. où \bar{x} et s_x^2 sont la moyenne et la variance d'échantillon des x_i , $\bar{x}_o = \sum_{i \in s} \alpha_i x_i / n_o$ est la moyenne des x_i pour les

$$V_1 = N^2 \frac{n(n-1)}{(1-n/N)} \left[\frac{\bar{x}}{s_x^2 + 2} \frac{n_o}{\bar{x}} + \frac{\bar{x}_o}{\bar{R}_o^2 s_x^2} \right] \quad (3.2)$$

échantillonnage aléatoire simple et l'imputation par quotient, Shao et Steele (1999) ont obtenu V_1 sous la forme élément $C_{ij} = \text{Cov}_{ij} (\sum_{k \in s} \alpha_k t_k, \sum_{k \in s} \alpha_k t_k)$ et en remplaçant les quantités inconnues par des estimateurs. Pour estimateur, V_2 , de $V_2 = \left[\Delta \phi(E_o T_o) \right] C \left[\Delta \phi(E_o T_o) \right]$, où $\phi(T_o) = \psi(E_o T_o) - Y$, en calculant C dont le

se servant d'un estimateur par linéarisation de la variance de $\psi(T_o)$ étant donnée les α_i . Ils ont obtenu aussi un moyen un plan d'échantillonnage donné. Shao et Steele (1999) ont obtenu un estimateur de la variance, V_1 , de V_1 en mécanisme de réponse, et E_o , et Var_o , représentent l'espérance et la variance sous le mécanisme d'échantillonnage où $V_1 = E_o(Var_o(Y_o - Y))$, $V_2 = Var_o(E_o(Y_o - Y))$, E_o et Var_o représentent l'espérance et la variance sous le Nous supposons que l'imputation est déterministe. Nous avons $Var(Y_o - Y) = V_1 + V_2$,

où $\alpha_1 = o_1$ et $\alpha_2 = 1 - o_1$. Il découle de (3.1) que \hat{Y}_o est de la forme $\psi(T_o)$ avec $\hat{o}_i = (o_{i1}, o_{i2})^T$ et

$$\hat{Y}_o = \sum_{i=1}^n d_i(s) \alpha_i Y_i + \sum_{i=1}^n d_i(s)(1 - o_i) \bar{R}_o x_i \quad (3.1)$$

utilisée pour imputer une valeur à y_i , et $\hat{o}_i = (o_{i1}, \dots, o_{i2})^T$ est un vecteur des variables indicatrices de réponse. Par exemple, considérons l'imputation par quotient lorsqu'on dispose des valeurs de la variable auxiliaire x_i pour tous les $i \in s$. Une valeur manquante de y_i est alors imputée par $\hat{y}_i = \bar{R}_o x_i$, où $\bar{R}_o = (\sum_{i \in s} d_i(s) \alpha_i Y_i) / (\sum_{i \in s} d_i(s) \alpha_i x_i)$ et \bar{x}_o L'estimateur imputé \hat{Y}_o est donné par

est l'indicateur de réponse pour y_i , c.-à-d. $\alpha_i = 1$ si la valeur de y_i est observée et $\alpha_i = 0$ si elle est manquante.

En inspirant de Fay (1991), Shao et Steele (1999) ont proposé une méthode de calcul des estimateurs de la variance de l'estimateur du total de type Horvitz-Thompson, \hat{Y}_o , lorsque des valeurs sont imputées pour tenir compte de la non-réponse partielle. Ils ont supposé que l'on peut exprimer le total estimé \hat{Y}_o sous forme d'une fonction lisse des totaux $\hat{Y}_o = \psi(T_o)$, où $T_o = \sum_{i=1}^n d_i(s) \text{diag}(\hat{o}_i)^T$, est la valeur de y_i ou celle d'une autre variable

3. Non-réponse partielle

Demandé et Rao (2001) ont appliquée la méthode à divers problèmes, qui englobent les estimateurs par régression et par collage d'un total X , et d'autres estimateurs définis explicitement ou implicitement comme étant des solutions des équations d'estimation. Ils ont obtenu un nouveau estimateur de la variance pour une classe générale d'estimateurs par collage qui inclut les estimateurs par la méthode itérative du quotient (ranking ratio) généralisée et les estimateurs par régression qui utilisent deux degrés et ont obtenu un estimateur de la variance à deux degrés qui utilise deux degrés mieux que l'estimation classique de la variance.

Dans le cas de l'échantillonnage sélectif simple, $v_L(Y^L) = v(z)$ concorde avec $v_{LL} = (\underline{X}/\bar{x})^2 v_L$.

$$z^k = \frac{\partial f(\bar{q})}{\partial q^k} \Big|_{\bar{q}=\bar{q}(s)} = \frac{X}{\bar{x}} (Y^k - \bar{F}x^k).$$

$$\text{sur } i \in U. \text{ Alors } f(\bar{q}) = X \left[(\bar{Z} b_i x_i) / (\bar{Z} b_i x_i) \right] = X \left[\bar{Y}(\bar{q}) / \bar{X}(\bar{q}) \right] \text{ et}$$

Supposons que θ est l'estimateur par quotient $\hat{Y}^\theta = X \left[(\bar{Z} d_i(s)x_i) / (\bar{Z} d_i(s)x_i) \right]$, où Z représente la sommation

variance $v_L(\theta)$ est valide, car z^i est un estimateur convergent de \bar{z}^i .

Par conséquent, notre méthode est similaire en esprit à l'approche de Binder (1996). L'estimateur de la variance $v_L(\theta)$ est valide, car $\hat{b}_i = \bar{b}_i$ afin d'obtenir \bar{z}^i pour ensuite substituer les estimations aux composantes inconnues partielles $\partial f(\bar{q}) / \partial b^k$ à $k = i$. Soulignons que nous ne commençons pas par évaluer les dérivées remplaçant y_i par z^i pour $i \in s$. Soulignons que nous ne commençons pas par évaluer les dérivées notées que $v_L(\theta)$ donne par (2.2) obtient simplement à partir de la formule de $v(\theta)$ pour $\hat{Y} = Y(\theta)$ en

$$(2.2) \quad v_L(\theta) = v(z).$$

pour obtenir un estimateur par linéarisation de la variance notation opérationnelle. Maintenant, nous remplaçons \tilde{z}_k par $\tilde{z}_k = \partial f(\tilde{q}) / \partial b_k$, puisque les \tilde{z}_k sont inconnus, estime $\mathbb{E} d_i(s) \tilde{z}_i = \tilde{Y}(z)$; c'est-à-dire, $\text{var}(\theta) \approx \text{v}(z)$, où $v(z)$ représente l'estimateur de la variance de $\tilde{Y} = Y(\theta)$ en estimateur de la variance de θ est donné approximativement par l'estimateur de la variance du total où $\partial g(\tilde{q}) / \partial \tilde{q} = (\partial g(\tilde{q}) / \partial a_1, \dots, \partial g(\tilde{q}) / \partial a_m)^T$ et $\tilde{z} = (z_1, \dots, z_N)^T$ avec $\tilde{z}_k = \partial f(\tilde{q}) / \partial b_k$. Il découle de (2.1) qu'un

$$(2.1) \quad \begin{aligned} \tilde{z} &= (d(s) - \bar{I}) \\ &\approx \theta - \theta \left(\sum_{k=1}^N \partial f(\tilde{q}) / \partial b_k \right) \end{aligned}$$

est équivalente à

$$\tilde{Y} = \tilde{Y}(\tilde{X}, \tilde{\theta}), \quad \theta = g(\tilde{Y}) - g(\tilde{X}) \partial g(\tilde{q}) / \partial \tilde{q}$$

(2001) ont montré que la linéarisation de Taylor de $\theta - \theta$, c'est-à-dire

Soit $\tilde{Y} = \tilde{Z}b$, où pour des nombres réels arbitraires $b = (b_1, \dots, b_N)^T$, et $f(\tilde{q}, \tilde{A}) = f(\tilde{q})$. Demnati et Rao

pour tout $i \in s$, où a_i est la probabilité de sélectionner l'unité i dans l'échantillon s .
écrivons $\tilde{Y}^i = f(d(s), \tilde{Y}, \tilde{x})$. Si nous utilisons les coefficients de pondération d'Horvitz-Thompson, alors $d_i(s) = 1/a_i$
de $d(s)$, \tilde{Y}^i et de \tilde{x} , et du total connu X , mais que, par souci de simplicité, nous laissons tomber X et
 $y_{1i} = y_i$, $y_{2i} = x_i$ et $f(\tilde{I}, \tilde{A})$ se réduit au total Y , en remarquant que $(Y/X)X = Y$. Notons que \tilde{Y}^i est une fonction
des 1. Par exemple, si θ représente l'estimateur par quotient $\tilde{Y}^i = [\mathbb{E}_{i \in s} d_i(s) y_i / \mathbb{E}_{i \in s} d_i(s)] X$, alors $m = 2$,
matrice $m \times N$ dont la j^e colonne est $\tilde{Y}_j = (y_{1j}, \dots, y_{mj})^T$, $j = 1, \dots, N$, $d(s) = (d_1(s), \dots, d_N(s))^T$ et I est le vecteur N
avec $\tilde{Y} = (Y_1, \dots, Y_m)^T$. Nous pouvons écrire θ sous la forme $\theta = f(\tilde{d}(s), \tilde{A})$, où \tilde{A} est une
l'unité i ne fait pas partie de l'échantillon s , U est l'ensemble des unités de la population et $\theta = g(\tilde{Y})$,
estimates $\tilde{Y} = (Y_1, \dots, Y_m)^T$, où $\tilde{Y}_j = \mathbb{E}_{i \in s} d_i(s) y_{ij}$ est un estimateur de la population Y_j , $j = 1, \dots, m$ et $d_i(s) = 0$ si
pour modifier l'application de la méthode de Demnati-Rao (2001) en cas de réponse complète, supposeons
qu'un estimateur θ d'un paramètre θ peut être exprimé sous forme d'une fonction dérivable $g(\tilde{Y})$ des totaux

2. Réponse complète

En cas de réponses manquantes, on procéde souvent à l'ajustement des poids pour compenser la non-réponse complète, tandis que l'on a recourt couramment à l'imputation pour compenser la non-réponse partielle afin d'obtenir des données complètes pour calculer des estimations. Cependant, traiter les poids ajustés comme des poids de sondage et les valeurs imputées comme des valeurs réelles, et appliquer des formules standard d'estimation de la variance peut produire une sous-estimation considérable si le taux de non-réponse est important. Ces dernières années, on a proposé plusieurs méthodes pour estimer correctement la variance d'un estimateur en cas d'imputation. Rao (1996), Sha et Steel (1999) et d'autres ont étudié l'estimation de la variance sous imputation par quotient, cependant le problème d'estimation de la variance en cas d'ajustement de la pondération par quotient reste ouvert.

La linéarisation de Taylor est une méthode d'estimation de la variance fréquemment utilisée pour des statistiques complexes, comme les estimateurs par quantile ou par régression, ainsi que les estimateurs des coefficients de régression logistique. Elle est généralement applicable à tout plan jackknife, et les calculs sont plus simples que ceux des méthodes de réechantillonnage, comme le linéarage qui permet d'obtenir une estimation non biaisée de la variance d'estimateurs d'échantillonage qui dépendent de régressions logistiques. Elle est généralement utilisée pour rapporter au total de la population y , Royal et Cumby (1981) ont montré qu'un estimateur par linéarisation utilise couramment, $v_{ij} = N^{-1}(y_{ij} - \bar{y}_i)$, ne permet pas de suivre la variance de la variable x , s_x^2 est la variance de l'échantillon des résidus $z_i = y_i - \bar{y}_i / N$ et (n, N) représente les conditions de l'échantillon et de la population. Si nous linéarisons l'estimateur jackknife de la variance, v_{ij} , nous obtenons un estimateur par linéarisation de la variance v_{ij} . Yung et Rao (1996) ont considéré des estimateurs par régression généralisée et des conditions, où $\bar{x} = \bar{X}/N$ est la moyenne de x . Par conséquent, on pourra préférer v_{ij} ou v_{ij} à $v_{ij} = (X/x)^2 v_{ij}$, qui suit la variance conditionnelle ainsi que la variance non conditionnelle, $s_x^2 = \bar{x}^2 - \bar{x}^2/N$. Par contre, on pourra préférer v_{ij} ou v_{ij} à $v_{ij} = (X/x)^2 v_{ij}$ et v_{ij} possède toujours deux de bonnes propriétés simulatoires pour démontrer que v_{ij} et v_{ij} possèdent toutes deux de bonnes propriétés Valiant (1993) a aussi obtenu v_{ij} pour l'estimateur stratifié à posteriori et a réalisé une étude de simulation pour quatre stratégies à posteriori sous un plan stratifié à plusieurs degrés et ont obtenu un estimateur par linéarisation jackknife de la variance, v_{ij} , par la linéarisation de v_{ij} . Les estimations ajustées par quantile stratifiées à posteriori sont des tailles des strates. Simdal, Svensson et Wretman (1989) ont montré que v_{ij} est à la fois asymptotiquement non biaisé par rapport au plan de conditionnelles, étant donné les estimations des tailles des strates. Simdal, Svensson et Wretman sondage et asymptotiquement non biaisé par rapport au modèle au sens de $E_m(v_{ij}) = E_m(v_{ij})$, où E_m représente l'espérance par rapport au modèle et $V_m(y_j)$ est la variance de y_j par rapport au modèle sous un modèle $Var_m(y_i) = \sigma^2 x_i$, $i = 1, \dots, n$ et les y_i sont indépendants avec comme variance du modèle $Var_m(y_i) = \sigma^2 x_i$, $i > 0$. Donc, v_{ij} est un bon choix tant du point de vue du plan d'échantillonage que du point de vue du modèle. Demnati et Rao (2001) ont proposé une nouvelle méthode d'estimation de la variance qui est justifiable théoriquement et qui donne directement un estimateur de la variance de type v_{ij} pour des plans d'échantillonage généraux. Cette méthode est présente à la deuxième section.

1. Introduction

1. Introduction	7
2. Réponse complète	9
3. Non-réponse partielle	11
4. Nouvelle méthode: réponse manquante	13
Conclusion	19
Bibliographie	19

INDEX

Mots clés : Non-réponse partielle; imputation par quotient; linéarisation de Taylor; non-réponse totale; ajustement de la pondération.

Dans le présent article, nous étendons les travaux de Demnati et Rao (2001) afin de couvrir le problème des données manquantes. Nous élaborons des estimateurs valides de la variance directement à un estimateur unique de la variance qui satisfait les critères susmentionnés. Dans le présent article, nous étendons les travaux de Demnati et Rao (2001) afin de couvrir le problème des données manquantes. Nous élaborons des estimateurs valides de la variance directement à un estimateur unique de la variance qui satisfait les critères susmentionnés. en particulier l'imputation par quotient, souvent utilisée pour produire un ensemble complet complet, ainsi que sous l'imputation basée sur des fonctions lisses des valeurs observées, sous l'ajustement de la pondération, souvent utilisée pour tenir compte de la non-réponse de données. Mots clés : Non-réponse partielle; imputation par quotient; linéarisation de Taylor; non-réponse totale; ajustement de la pondération.

RESUME

B. Demnati et J. N. K. Rao

Estimateurs de la variance par linéarisation pour des données d'enquêtes avec des réponses manquantes

Novembre 2002

Statistique Canada

(Traduction)

A. Demnati et J. N. K. Rao

DMEs-2002-007E/F

Estimateurs de la variance par linéarisation
pour des données d'enquêtes
avec des réponses manquantes

Canada

STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA



1010723651

Methodology Branch Social Survey Methods Division

Dévision des méthodes
d'enquêtes sociales

Direction de la méthodologie

