LINEARIZATION VARIANCE ESTIMATORS FOR SURVEY DATA WITH MISSING RESPONSES

A. Demnati and J. N. K. Rao
A. Demnati, Statistics Canada, 15-G, R.H. Coats Bldg., Ottawa, ON, K1A 0T6

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

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": $E_m(y_i) = \beta x_i;\ i = 1,\ldots,N$ 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 $g(\hat{\underset{\sim}{Y}})$ of estimated totals $\hat{\underset{\sim}{Y}} = (\hat{Y}_1,\ldots,\hat{Y}_m)^T$, where $\hat{Y}_j = \sum_{i\in U} d_i(s)\, y_{ij}$ is an estimator of the population Y_j, $j = 1,\ldots,m$, where $d_i(s) = 0$ if the unit i is not in the sample s, U is the set of population units, and $\theta = g(\underset{\sim}{Y})$

with $\utilde{Y} = (Y_1,...,Y_m)^T$. We may write $\hat\theta$ as $\hat\theta = f(\utilde{d}(s), \utilde{A}_y)$ and $\theta = f(\utilde{1}, \utilde{A}_y)$, where $\utilde{A}_y$ is an $m \times N$ matrix with j^{th} column $\utilde{y}_j = (y_{1j},...,y_{mj})^T$, $j = 1,...,N$, $\utilde{d}(s) = (d_1(s),...,d_N(s))^T$ and $\utilde{1}$ is the N-vector of 1's. For example, if $\hat\theta$ denotes the ratio estimator $\hat{Y}_R = \left[(\sum_{i \in U} d_i(s)\, y_i)/(\sum_{i \in U} d_i(s)\, x_i)\right] X$, then $m=2$, $y_{1i} = y_i$, $y_{2i} = x_i$ and $f(\utilde{1}, \utilde{A}_y)$ reduces to the total Y, noting that $(Y/X)X = Y$. Note that $\hat{Y}_R$ is a function of $\utilde{d}(s)$, $\utilde{y}$ and $\utilde{x}$ and the known total X, but we dropped X for simplicity and write $\hat{Y}_R = f(\utilde{d}(s), \utilde{y}, \utilde{x})$. 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 $\breve{\utilde{Y}} = \sum b_i \utilde{y}_i$ for arbitrary real numbers $\utilde{b} = (b_1,...,b_N)^T$, and $f(\utilde{b}, \utilde{A}_y) = f(\utilde{b})$. Demnati and Rao (2001) showed that the Taylor linearization of $\hat\theta - \theta$, namely

$$\hat\theta - \theta = g(\hat{\utilde{Y}}) - g(\utilde{Y}) \approx \left(\partial g(\utilde{a})/\partial \utilde{a}\right)^T\Big|_{\utilde{a}=\utilde{Y}} \left(\hat{\utilde{Y}} - \utilde{Y}\right),$$

is equivalent to

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

where $\partial g(\utilde{a})/\partial \utilde{a} = (\partial g(\utilde{a})/\partial a_1,...,\partial g(\utilde{a})/\partial a_m)^T$ and $\tilde{\utilde{z}} = (\tilde{z}_1,...,\tilde{z}_N)^T$ with $\tilde{z}_k = \partial f(\utilde{b})/\partial b_k|_{\utilde{b}=\utilde{1}}$. 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, $var(\hat\theta) \approx v(\tilde{z})$, where $v(y)$ denotes the variance estimator of $\hat{Y} = \hat{Y}(y)$ in operator notation. Now we replace $\tilde{z}_k$ by $z_k = \partial f(\utilde{b})/\partial b_k|_{\utilde{b}=\utilde{d}(s)}$, since $\tilde{z}_k$'s are unknown, to get a linearization variance estimator

$$v_L(\hat\theta) = v(z). \qquad (2.2)$$

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 $\partial f(\utilde{b})/\partial b_k$ at $\utilde{b} = \utilde{1}$ to get $\tilde{\utilde{z}}$ and then substitute estimates for the unknown components of $\tilde{\utilde{z}}$. 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 $\sum$ denotes summation over $i \in U$. Then $f(\utilde{b}) = X\left[(\sum b_i y_i)/(\sum b_i x_i)\right] = X\,\hat{Y}(b)/\hat{X}(b)$ and

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

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, $\hat{Y}^{\bullet}$, with imputed item nonresponse values. They assumed that the estimated total $\hat{Y}^{\bullet}$ can be expressed as a smooth function of totals, $\hat{Y}^{\bullet} = \psi(\hat{\utilde{T}}_o)$, where $\hat{\utilde{T}}_o = \sum d_i(s)\, diag(\utilde{o}_i) \utilde{t}_i$, t_{ki} is the value of y_i or the value of some other variable used to impute y_i, and $\utilde{o}_i = (o_{i1},...,o_{ip})^T$ 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 $\hat{y}_i^{\bullet} = \hat{R}_o x_i$, 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 $\hat{Y}^{\bullet}$ is given by

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

where $o_{i1} = o_i$ and $o_{i2} = 1 - o_i$. It follows from (3.1) that $\hat{Y}^{\bullet}$ is of the form $\psi(\hat{\utilde{T}}_o)$ with $\utilde{o}_i = (o_{i1}, o_{i1}, o_{i2})^T$ and $\utilde{t}_i = (y_i, x_i, x_i)^T$.

We assume deterministic imputation. We have $Var(\hat{Y}^{\bullet} - Y) = V_1 + V_2$, where $V_1 = E_o[Var_s(\hat{Y}^{\bullet} - Y)]$, $V_2 = Var_o[E_s(\hat{Y}^{\bullet} - Y)]$, 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 $\psi(\hat{\utilde{T}}_o)$ for given $\utilde{o}_i$'s. They also obtained an estimator, v_2, of $V_2 \approx [\nabla\phi(E_o \utilde{T}_o)]^T \utilde{C} [\nabla\phi(E_o \utilde{T}_o)]$, where $\phi(\utilde{T}_o) = \psi(E_s \hat{\utilde{T}}_o) - Y$, by deriving $\utilde{C}$ with kl^{th} element $c_{kl} = cov_o(\sum o_{ki} t_{ki}, \sum o_{li} t_{li})$ 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 $\hat{Y}^{\bullet}$.

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 Y is estimated by a weighted sample total

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

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

$$\hat{Y}^{\bullet} = \sum \tilde{w}_i(s)\, \hat{y}_i = \underset{\sim}{\hat{y}}^T \underset{\sim}{\tilde{w}}(s), \quad (4.2)$$

where $\underset{\sim}{\hat{y}} = (\hat{y}_1, \ldots, \hat{y}_N)^T$ and $\hat{y}_i = o_i y_i + (1 - o_i) \hat{y}_i^{\bullet}$. In subsection 4.1, we study the case of item nonresponse only (i.e., $\tilde{w}_i(s) = d_i(s)$) assuming $\hat{Y}^{\bullet}$ can be expressed as a smooth function of totals $\sum d_i(s)\, diag(\underset{\sim}{o}_i) \underset{\sim}{t}_i$, where t_{ki} the value of y_i or the value of some other variable used to impute y_i. 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 $\hat{Y}^{\bullet}$ is assumed to be a smooth function of totals $\sum d_i(s)\, diag(\underset{\sim}{o}_i) \underset{\sim}{t}_i$, as in Shao and Steel (1999). In this case, $\hat{Y}^{\bullet}$ may be expressed as $f(\underset{\sim}{A}_w, \underset{\sim}{A}_y)$, where $\underset{\sim}{A}_w = diag(\underset{\sim}{d}(s))\, \underset{\sim}{A}_o$ is an $m \times N$ matrix with j^{th} column $\underset{\sim}{w}_j(s) = (w_{1j}(s), \ldots, w_{mj}(s))^T$, $j = 1, \ldots, N$. The vector $\underset{\sim}{w}_j(s)$ is defined as

$$\begin{aligned} \underset{\sim}{w}_j(s) &= (w_{1j}(s), \ldots, w_{mj}(s))^T \\ &= (o_{1j} d_j(s), \ldots, o_{mj} d_j(s))^T = \underset{\sim}{o}_j\, d_j(s), \end{aligned}$$

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

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

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

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

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

with

$$\hat{Y}_L^{\bullet} = \sum (\underset{\sim}{o}_i d_i(s))^T\ \underset{\sim}{\tilde{z}}_i = \sum \underset{\sim}{w}_i^T(s)\, \underset{\sim}{\tilde{z}}_i,$$

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

$$v(\hat{Y}_L^{\bullet}) = v_s(\underset{\sim}{o}^T \underset{\sim}{z}) + v_o(\underset{\sim}{z}), \quad (4.4)$$

where $\underset{\sim}{z}_k$ is the vector of derivatives of the estimator $f(\underset{\sim}{A}_b)$ with respects to $\underset{\sim}{b}_k$ evaluated at $\underset{\sim}{A}_b = \underset{\sim}{A}_w$, and $v_o(\underset{\sim}{z})$ is an estimator of $V(\sum diag(\underset{\sim}{o}_i) \underset{\sim}{z}_i)$. Under independent response mechanism,

$$v_o(\underset{\sim}{z}) = \sum \underset{\sim}{z}_i^T cov_o(\underset{\sim}{o}_i) \underset{\sim}{z}_i, \quad (4.5)$$

where $cov_o(\underset{\sim}{o}_i)$ is an (approximately) unbiased estimator of $E(\underset{\sim}{o}_i \underset{\sim}{o}_i^T) - E(\underset{\sim}{o}_i) E(\underset{\sim}{o}_i^T)$.

Under ratio imputation, we have

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

It follows from (4.6) that

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

Therefore, $v_s(\underset{\sim}{o}^T\underset{\sim}{z})$ 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,

$$cov_o(\underset{\sim}{o}_i) = d_i(s)\begin{pmatrix} 0 & 0 \\ 0 & o_i(1-\hat{\xi}_{io}) \end{pmatrix},$$

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

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

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

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

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, $p = 1...m$ by

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

We have

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

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,

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

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

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

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

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

$$\tilde{w}_i(s) = d_i(s)\, r_i\, g_i(\underset{\sim}{d}(s), \underset{\sim}{r}, \underset{\sim}{A}_\chi), \qquad (4.11)$$

where $g_i(\underset{\sim}{d}(s), \underset{\sim}{r}, \underset{\sim}{A}_\chi)$ is known as the g-weights in the context of regression estimator and $\underset{\sim}{A}_\chi$ 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

$$g_i(\underset{\sim}{d}(s), \underset{\sim}{r}, \underset{\sim}{A}_\chi) = \frac{\sum d_i(s)\chi_i}{\sum d_i(s) r_i \chi_i} = \frac{\hat{\chi}}{\hat{\chi}_r}, \qquad (4.12)$$

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

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

where

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

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

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

In the logistic case, we have

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

where

$$u_i(r_i, \underset{\sim}{\chi}_i, \hat{\xi}_{ir}) = (r_i - \hat{\xi}_{ir})\underset{\sim}{\chi}_i ,$$

$$\xi_{ir} = \exp(\underset{\sim}{\chi}_i^T \underset{\sim}{\beta})/(1 + \exp(\underset{\sim}{\chi}_i^T \underset{\sim}{\beta})) = Pr(r_i = 1 | \underset{\sim}{\chi}_i, \underset{\sim}{\beta}),$$

and $\underset{\sim}{\chi}_i$ 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 $\hat{Y}^{\bullet}$ as $f(\underset{\sim}{A}_w)$ and then by differentiating $f(\underset{\sim}{A}_b)$ with respect to $\underset{\sim}{b}_k$. 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.,

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

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

$$\underset{\sim}{w}_i(s) = (1, o_i, r_i)^T d_i(s),$$

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

,

and

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

5. 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.

6. 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.

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),“Postsratification 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.