Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada **Statistique Canada**

Canada

Proceedings of Statistics Canada Symposium 2005
Methodological Challenges for Future Information Needs

LINEARIZATION VARIANCE ESTIMATORS FOR MODEL PARAMETERS FROM COMPLEX SURVEY DATA

A. Demnati and J. N. K. Rao[1]

ABSTRACT

In survey sampling, Taylor linearization is often used to obtain variance estimators for calibration estimators of totals and nonlinear finite population (or census) 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, and (ii) validity under a conditional repeated sampling framework. Demnati and Rao (2004) proposed a new approach to deriving Taylor linearization variance estimators that leads directly to a unique variance estimator that satisfies the above considerations for general designs. Afterwards, Demnati and Rao (2002) considered the case of missing responses when adjustment for complete nonresponse and imputation for item nonresponse based on smooth functions of observed values, in particular ratio imputation, are used. When analyzing survey data, finite populations are often assumed to be generated from superpopulation models, and analytical inferences on model parameters are of interest. If the sampling fractions are small, then the sampling variance captures almost the entire variation generated by the design and model random processes. However, when the sampling fractions are not negligible, the model variance should be taken into account in order to construct valid inferences on model parameters under both randomization processes. In this paper, we focus on total variance estimation using the Demnati-Rao approach when the characteristics of interest are assumed to be random variables generated from a superpopulation model. We illustrate the method using ratio estimators and estimators defined as solutions to calibration weighted estimating equations. Application to a zero-inflated Poisson model is also given.

KEY WORDS: Calibration; Weighted estimating equations; Ratio estimators; Total variance; Zero-inflated Poisson.

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 unlike a resampling method such as the jackknife, and it is computationally simpler than the latter method. 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, and (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 $\vartheta_L = N^2(n^{-1} - N^{-1})s_z^2$ does not track the conditional variance of $\hat{Y}_R$ given $\bar{x}$, unlike the jackknife variance estimator ϑ_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 $z_k = y_k - (\bar{y}/\bar{x})x_k$ and (n, N) denote the sample and population sizes. By linearizing the jackknife variance estimator, ϑ_J, we obtain a different linearization variance estimator, $\vartheta_{JL} = (\bar{X}/\bar{x})^2 \vartheta_L$, 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, ϑ_{JL} or ϑ_J may be preferred over ϑ_L. Valliant (1993) obtained ϑ_{JL} for the post-stratified estimator and conducted a simulation study to demonstrate that both ϑ_J and ϑ_{JL} possess good conditional properties given the estimated post-

[1] *A. Demnati, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6 (Abdellatif.Demnati@statcan.ca); J. N. K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6 (JRao@math.carleton.ca)*

strata counts. Särndal, Swensson and Wretman (1989) showed that ϑ_{JL} is both asymptotically design unbiased and asymptotically model unbiased in the sense of $E_m(\vartheta_{JL}) = V_m(\hat{Y}_R)$, where E_m denotes model expectation and $V_m(\hat{Y}_R)$ is the model variance of $\hat{Y}_R$ under a "ratio model": $E_m(y_k) = \beta x_k$; $k = 1,...,N$ and the y_k's are independent with model variance $V_m(y_k) = \sigma^2 x_k$, $\sigma^2 > 0$. Thus, ϑ_{JL} is a good choice from either the design-based or the model-based perspective. Demnati and Rao (2004) proposed a new approach to variance estimation that is theoretically justifiable and at the same time leads directly to a ϑ_{JL}-type variance estimator for general designs. Demnati and Rao (2004) applied the method under the design based approach 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 sampling variance estimator that makes fuller use of the first phase sample data compared to traditional linearization variance estimators. Demnati and Rao (2002) extended their method to the case of missing responses when adjustments for complete nonresponse and imputation for item nonresponse based on smooth functions of observed values, in particular ratio imputation, are used.

When analyzing survey data, the finite population values $\boldsymbol{y} = (y_1,...,y_N)^T$ are often assumed to be generated from a superpopulation model, and analytical inferences on model parameters are of interest. If the sampling fractions are negligible, the sampling variance captures almost the entire variation generated by the design and model random processes. However, when the sampling fractions are not small, the model variance is not negligible in comparison to the total variance. In this case, the model variance should also be accounted for in order to construct valid inferences on model parameters under both random processes.

Molina, Smith and Sugden (2001) obtained general expressions for the mean and covariance function of the sample data $diag(\boldsymbol{a}(s))\boldsymbol{y}$ and for the sample totals under the joint processes, where $\boldsymbol{a}(s) = (a_1(s),...,a_N(s))^T$, $a_k(s) = 1$ if element k belongs to the sample s and $a_k(s) = 0$ otherwise. There is no doubt that the combined process of selection of the sample and generation of the finite population should be the basis for analytic inferences, as argued by Molina, Smith and Sugden (2001). However, a broadly applicable method is needed for total variance estimation. In section 2, we extend the Demnati-Rao approach to total variance estimation when the characteristics of interest are assumed to be random variables generated from a superpopulation model. The method is theoretical justifiable and at the same time leads directly to a unique estimator of total variance with desirable properties. We apply the method to ratio estimators of model parameters in section 2. The method is then extended in section 3 to estimators that are defined as solutions to weighted estimating equations, using generalized regression (GREG) calibration weights.

2. RATIO ESTIMATOR WHEN y_k IS RANDOM

Suppose that $\hat{\theta}$ is the ratio estimator $\hat{\theta} = X(\sum y_k d_k(s))/(\sum x_k d_k(s)) \equiv X\hat{R}$ and the model parameter is $\theta = E_m(Y) = \sum E_m(y_k)$, where the sum is over the N elements in the population, $d_k(s) = 0$ if element k is not in the sample s and X is the known total of x. Let $\boldsymbol{d}_k = (d_{1k}, d_{2k})^T$, where $d_{1k} = d_k(s)$, $d_{2k} = d_k(s)y_k$, and s is suppressed in $\boldsymbol{d}_k(s)$ for simplicity. We may then express $\hat{\theta}$ as $\hat{\theta} = f(\boldsymbol{A}_d) = X(\sum d_{2k})/(\sum x_k d_{1k})$, where $\boldsymbol{A}_d$ is a $2 \times N$ matrix with k^{th} column $\boldsymbol{d}_k$. We use the Horvitz-Thompson (HT) weights $d_k(s) = a_k(s)/\pi_k$, where $a_k(s)$ is the sample membership indicator variable and π_k is the inclusion probability. In this case, letting $E = E_m E_p$ denote the total expectation, we have $E(d_{1k}) = E_m E_p(d_k(s)) = E_m(1) = 1 \equiv \mu_{1k}$ and $E(d_{2k}) = E_m(y_k E_p(d_k(s))) = E_m(y_k) \equiv \mu_{2k}$, where E_p denotes expectation with respect to the design. Hence, $E\hat{\theta} \approx f(\boldsymbol{A}_\mu) = \theta$, where $\boldsymbol{A}_\mu$ is a $2 \times N$ matrix with k^{th} column $\boldsymbol{\mu}_k = (\mu_{1k}, \mu_{2k})^T$.

Demnati and Rao (2004) have shown that the Taylor expansion of $\hat{\theta}-\theta$ maybe written as

$$\hat{\theta}-\theta \approx \sum \tilde{z}_k^T (\boldsymbol{d}_k - \boldsymbol{\mu}_k), \tag{2.1}$$

where $\tilde{z}_k = \partial f(\boldsymbol{A}_b)/\partial \boldsymbol{b}_k \mid_{\boldsymbol{A}_b=\boldsymbol{A}_\mu}$ and $\boldsymbol{A}_b$ is a $2\times N$ matrix of arbitrary real numbers with k^{th} column $\boldsymbol{b}_k$. This result is true for any $\hat{\theta}$ that can be expressed as a smooth function of estimated totals. Using operator notation, let $\vartheta(\boldsymbol{u})$ denote the estimator of total variance of a linear estimator $\hat{U}=\sum \boldsymbol{u}_k^T \boldsymbol{d}_k$. Then the Demnati-Rao (DR) linearization variance estimator of $\hat{\theta}$ is simply given by

$$\vartheta_{DR}(\hat{\theta}) = \vartheta(\boldsymbol{z}), \tag{2.2}$$

which is obtained from $\vartheta(\boldsymbol{u})$ by replacing $\boldsymbol{u}_k$ by $\boldsymbol{z}_k = \partial f(\boldsymbol{A}_b)/\partial \boldsymbol{b}_k \mid_{\boldsymbol{A}_b=\boldsymbol{A}_d}$. Note that $\boldsymbol{z}_k$ is a consistent estimator of $\tilde{z}_k$. For the ratio estimator $\hat{\theta}$, we have

$$\boldsymbol{z}_k = (z_{1k}, z_{2k})^T = (X/\hat{X})(-\hat{R}x_k, 1)^T. \tag{2.3}$$

It remains to evaluate $\vartheta(\boldsymbol{u})$. We have

$$\vartheta(\boldsymbol{u}) = \sum\sum \boldsymbol{u}_k^T \operatorname{cov}(\boldsymbol{d}_k, \boldsymbol{d}_t)\boldsymbol{u}_t, \tag{2.4}$$

where

$$\operatorname{cov}(\boldsymbol{d}_k, \boldsymbol{d}_t) = d_{kt}(s)\begin{bmatrix} 0 & 0 \\ 0 & \operatorname{cov}_m(y_k, y_t) \end{bmatrix} + d_{kt}(s)\frac{(1-\omega_{kt})}{\omega_{kt}}\boldsymbol{v}_k \boldsymbol{v}_t^T. \tag{2.5}$$

In (2.5), $\boldsymbol{v}_k = (1, y_k)^T$,

$d_{kt}(s) = a_k(s)a_t(s)/\pi_{kt}$, $d_{kk}(s) = d_k(s)$,

$\omega_{kt} = \pi_k\pi_t/\pi_{kt}$, $(1-\omega_{kt})/\omega_{kt} = (\pi_{kt}-\pi_k\pi_t)/(\pi_k\pi_t)$,

and $\operatorname{cov}_m(y_k, y_t)$ is an estimator of the covariance of y_k and y_t under the assumed model, where π_{kt} is the joint inclusion probability for $k \neq t$, and $\pi_{kk} = \pi_k$. When the model covariance of y_k and y_t is zero, $\operatorname{cov}_m(y_k, y_t)$ is taken as zero.

Substituting $\boldsymbol{z}_k$ in (2.3) for $\boldsymbol{u}_k$ in (2.4), we get

$$\begin{aligned} \vartheta_{DR}(\hat{\theta}) &= \sum\sum d_{kt}(s) z_{k;m} z_{t;m} \operatorname{cov}_m(y_k, y_t) + \sum\sum d_{kt}(s) z_{k;s} z_{t;s} (1-\omega_{kt})/\omega_{kt} \\ &\equiv \vartheta_m + \vartheta_s \end{aligned}. \tag{2.6}$$

where $z_{k;m} = z_{2k} = X/\hat{X}$ and $z_{k;s} = \boldsymbol{z}_k^T \boldsymbol{v}_k = z_{1k} + z_{2k}y_k = (X/\hat{X})(y_k - \hat{R}x_k)$. Note that the first term, ϑ_m, on the right side of (2.6) corresponds to the model and the second term, ϑ_s, corresponds to the sampling design.

Under simple random sampling,

$$\vartheta_s = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\left(\frac{X}{\hat{X}}\right)^2 s_e^2, \tag{2.7}$$

where $s_e^2 = \sum a_k(s)(y_k - \hat{R}x_k)^2/(n-1)$ and $\hat{R} = \bar{y}/\bar{x}$. Further, under the ratio model

$$E_m(y_k) = \beta x_k,\ Cov_m(y_k, y_t) = 0,\ k \neq t, \tag{2.8}$$

$\theta = \beta X$, the model variance of y_k, $V_m(y_k) = E_m(y_k - \beta x_k)^2$ is estimated robustly by $(y_k - \hat{R}x_k)^2$, and

$$\vartheta_m = \frac{N}{n}\left(\frac{X}{\hat{X}}\right)^2 (n-1)s_e^2. \tag{2.9}$$

Note that ϑ_m remains valid under misspecification of the model variance of y_k.

Now combining (2.7) with (2.9), we get an estimator of the total variance of $\hat{\theta}$ as

$$\vartheta_{DR}(\hat{\theta}) = \frac{N^2}{n}\left(\frac{X}{\hat{X}}\right)^2 \frac{N-1}{N} s_e^2 . \tag{2.10}$$

It is interesting to note that the "g-weight" $X/\hat{X}$ appears automatically in $\vartheta_{DR}(\hat{\theta})$, and that the finite population correction $1-n/N$ is absent in $\vartheta_{DR}(\hat{\theta})$. The Demnati-Rao approach leads to a unique choice of variance estimator that preserves the g-factors automatically.

In the traditional approach to estimation of total variance, $V(\hat{\theta})$ is written as $E_m V_p(\hat{\theta}) + V_m E_p(\hat{\theta}) \approx E_m V_p(\hat{\theta}) + V_m(Y) = E_m V_p(\hat{\theta}) + \sum E_m(y_k - \beta x_k)^2$ under the ratio model. The first term is estimated by an estimator of $V_p(\hat{\theta})$, typically using ϑ_s without the g-factor. The second term is estimated robustly by $\sum d_k(s)(y_k - \hat{R}x_k)^2 = (N/n)(n-1)s_e^2$. The sum of the two estimated terms equals (2.10) without the g-factor. The choice of variance estimator under the traditional approach is not unique.

If the parameter of interest is $\beta = \theta / X$, then $\hat{\beta} = \hat{\theta}/X = \hat{R}$ and

$$\vartheta_{DR}(\hat{\beta}) = \vartheta_{DR}(\hat{\theta}/X) = \frac{N^2}{n}\frac{1}{\hat{X}^2}\frac{N-1}{N}s_e^2 , \tag{2.11}$$

under the ratio model. The traditional approach typically uses the same variance estimator of $\hat{\beta}$.

3. SURVEY WEIGHTED ESTIMATING EQUATIONS

Suppose the supperpopulation model on the responses y_k is specified by a generalized linear model with mean $E_m(y_k) = \mu_k(\boldsymbol{\theta}) = h(\boldsymbol{u}_k^T\boldsymbol{\theta})$, where $\boldsymbol{u}_k$ is a $p\times 1$ vector of explanatory variables and $h(.)$ is a "link" function. The model parameter of interest is $\boldsymbol{\theta}$. For example, the choice $h(a) = a$ gives linear regression model and $h(a) = e^a/(1+e^a)$ gives logistic regression model for binary responses y_k.

We define census estimating equations (CEE) as $\boldsymbol{l}(\boldsymbol{\theta}) = \sum \boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{0}$ with $E_m \boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{0}$, and the solution to CEE gives the census parameter $\hat{\boldsymbol{\theta}}_N$. For example, $\boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{u}_k(y_k - \boldsymbol{\mu}_k(\boldsymbol{\theta}))$ for linear and logistic regression models. We use generalized regression (GREG) calibration weights $w_k(s) = d_k(s)g_k(\boldsymbol{d}(s))$, where the "g-weights" are given by

$$g_k(\boldsymbol{d}(s)) = 1 + (\boldsymbol{X} - \hat{\boldsymbol{X}})^T[\sum d_k(s)c_k\boldsymbol{x}_k\boldsymbol{x}_k^T]^{-1}c_k\boldsymbol{x}_k , \tag{3.1}$$

for specified c_k, $\hat{\boldsymbol{X}} = \sum d_k(s)\boldsymbol{x}_k$ is the HT estimator of the known total $\boldsymbol{X}$ of a $q\times 1$ vector of calibration variables $\boldsymbol{x}_k$ and $\boldsymbol{d}(s)$ is the $N\times 1$ vector of HT weights $d_k(s)$. The resulting GREG estimator of the total Y, namely $\hat{Y} = \sum w_k(s)y_k$ has the calibration property $\sum w_k(s)\boldsymbol{x}_k = \boldsymbol{X}$ (Särndal *et al*., 1989).

We use the calibration weights to estimate the CEE. The calibration weighted estimating equations are given by

$$\hat{\boldsymbol{l}}(\boldsymbol{\theta}) = \sum d_k(s)g_k(\boldsymbol{d}(s))\boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{0} . \tag{3.2}$$

The solution to (3.2) gives the calibration weighted estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, and $\hat{\boldsymbol{\theta}}$ is approximately design-model unbiased for $\boldsymbol{\theta}$, i.e., $E_m E_p(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$. It follows from (3.2) that $\hat{\boldsymbol{\theta}}$ is of the form $\boldsymbol{f}(\boldsymbol{A}_d)$ with $\boldsymbol{d}_k = (d_k(s), d_k(s)\boldsymbol{l}_k^T(\boldsymbol{\theta}))^T$, where $\boldsymbol{f}(\boldsymbol{A}_d)$ is a $p\times 1$ vector and $\boldsymbol{A}_d$ is a $(p+1)\times N$ matrix with k^{th} column $\boldsymbol{d}_k$.

Following the implicit differentiation method of Demnati and Rao (2004), $\boldsymbol{Z}_k = \partial \boldsymbol{f}(\boldsymbol{A}_b)/\partial \boldsymbol{b}_k \,|_{\boldsymbol{A}_b = \boldsymbol{A}_d}$ is evaluated as

$$\boldsymbol{Z}_k^T = [\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})]^{-1} g_k(\boldsymbol{d}(s))(-\hat{\boldsymbol{B}}_l^T \boldsymbol{x}_k, \boldsymbol{I}_p), \tag{3.3}$$

with

$$\hat{\boldsymbol{B}}_l = [\sum d_k(s) c_k \boldsymbol{x}_k \boldsymbol{x}_k^T]^{-1} \sum d_k(s) c_k \boldsymbol{x}_k \boldsymbol{l}_k^T(\hat{\boldsymbol{\theta}}), \tag{3.4}$$

$$\hat{\boldsymbol{J}}(\boldsymbol{\theta}) = \sum d_k(s) g_k(\boldsymbol{d}(s))(\partial \boldsymbol{l}_k(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^T), \tag{3.5}$$

and $\boldsymbol{I}_p$ is the identity matrix. The DR linearization variance estimator of $\hat{\boldsymbol{\theta}}$ is obtained from (2.4) and (2.5) by replacing $\boldsymbol{u}_k^T$ by the $p \times (p+1)$ matrix $\boldsymbol{Z}_k^T$, $\boldsymbol{v}_k$ by $(1, \boldsymbol{l}_k^T(\boldsymbol{\theta}))^T$ and $\text{cov}_m(y_k, y_t)$ by an estimator of the $p \times p$ covariance matrix of $\boldsymbol{l}_k(\boldsymbol{\theta})$ under the assumed model. After simplification, we get

$$\vartheta_{DR}(\hat{\boldsymbol{\theta}}) = \vartheta_m + \vartheta_s, \tag{3.6}$$

where ϑ_s is the sampling estimated covariance matrix given by

$$\vartheta_s = [\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})]^{-1} \sum\sum d_{kt}(s) g_k(\boldsymbol{d}(s)) g_t(\boldsymbol{d}(s))(1-\omega_{kt}) / \omega_{kt} \boldsymbol{e}_k^*(\hat{\boldsymbol{\theta}}) \boldsymbol{e}_t^*(\hat{\boldsymbol{\theta}}) [\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})]^{-1\text{T}}, \tag{3.7}$$

with

$$\boldsymbol{e}_k^*(\hat{\boldsymbol{\theta}}) = \boldsymbol{l}_k(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{B}}_l^T \boldsymbol{x}_k. \tag{3.8}$$

The model estimated covariance matrix depends on the assumed model covariance structure. If $Cov_m(\boldsymbol{l}_k(\boldsymbol{\theta}), \boldsymbol{l}_t^T(\boldsymbol{\theta})) = \boldsymbol{0}$ for $k \neq t$, and $\boldsymbol{V}_m(\boldsymbol{l}_k(\boldsymbol{\theta})) = E_m(\boldsymbol{l}_k(\boldsymbol{\theta}) \boldsymbol{l}_k^T(\boldsymbol{\theta}))$ is estimated robustly by $\boldsymbol{l}_k(\hat{\boldsymbol{\theta}}) \boldsymbol{l}_k^T(\hat{\boldsymbol{\theta}})$, then the model estimated covariance matrix, ϑ_m, reduces to

$$\vartheta_m = [\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})]^{-1} \sum d_k(s) g_k^2(\boldsymbol{d}(s)) \boldsymbol{l}_k(\hat{\boldsymbol{\theta}}) \boldsymbol{l}_k^T(\hat{\boldsymbol{\theta}}) [\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})]^{-1\text{T}}. \tag{3.9}$$

Note that for the linear regression and logistic regression models, $Cov_m(\boldsymbol{l}_k(\boldsymbol{\theta}), \boldsymbol{l}_t^T(\boldsymbol{\theta})) = \boldsymbol{0}$ for $k \neq t$ if the y_k's are uncorrelated under the model, noting that $\boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{u}_k(y_k - \boldsymbol{\mu}_k(\boldsymbol{\theta}))$.

Example: Zero-Inflated Poisson Model

We now report the results of a simulation study on the finite sample bias of the new variance estimator $\vartheta_{DR}(\hat{\boldsymbol{\theta}})$ given by (3.6). We consider a zero-inflated Poisson regression model that is often used for count data with excess zeros. The model assumes that with probability $1 - p_k$ the value of k^{th} element, y_k, is always 0 and with probability p_k it is $a_k\ (\geq 0)$ that is drawn from a Poisson (λ_k) distribution (Lambert, 1992). We assume that $p_k = p = e^{\alpha}/(1+e^{\alpha}) \approx 0.62$ and $\lambda_k = \lambda = e^{\beta} \approx 2.7$ with $\alpha = 0.5$ and $\beta = 1$ so that the model parameters are α and β. We generated 300 populations, each of size $N = 1{,}000$, from the above zero-inflated Poisson model, and from each generated population we selected 300 samples using Bernoulli sampling with probability $\pi = 0.1$. To implement calibration, we generated constants x_k $(k = 1, ..., N)$ from Bernoulli(0.6) and fixed them over the simulation runs. Using these x_k and the design weights $d_k(s) = a_k(s)/\pi$, we calculated the GREG weights with $c_k = 1$.

The estimating (or score) function $\boldsymbol{l}_k(\boldsymbol{\theta})$ under the above model has two components, where $\boldsymbol{\theta} = (\alpha, \beta)^T$:

$$l_{1k}(\boldsymbol{\theta}) = \frac{I_k(1-p_k) + p_k f(y_k)}{1 - p_k + p_k f(y_k)} \partial \log f(y_k) / \partial \beta \tag{3.10}$$

$$l_{2k}(\boldsymbol{\theta}) = \frac{I_k - p_k + p_k f(y_k)}{1 - p_k + p_k f(y_k)} \partial \log p_k / \partial \alpha, \tag{3.11}$$

where $I_k = 0$ if $y_k = 0$ and $I_k = 1$ if $y_k > 0$, and $f(y_k)$ is the density of Poisson (λ_k). Using (3.10) and (3.11) in (3.2), we calculated $\hat{\boldsymbol{\theta}}$ for each simulated sample. Using $\hat{\boldsymbol{\theta}}$, we then calculated the total variance estimate $\vartheta_{DR}(\hat{\boldsymbol{\theta}})$ and its components ϑ_m and ϑ_s for each sample. Using these values, we evaluated the simulated covariance matrix

of $\hat{\boldsymbol{\theta}}$ and the average values of ϑ_m, ϑ_s and $\vartheta_{DR}(\hat{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{\theta}}$ over simulations. The bias in the estimation of $\boldsymbol{\theta}$ is negligible here: $\alpha = 0.5$ vs. $\bar{\hat{\alpha}} = 0.5169$ and $\beta = 1$ vs. $\bar{\hat{\beta}} = 0.9936$, where $\bar{\hat{\alpha}}$ and $\bar{\hat{\beta}}$ denote the average values of $\hat{\alpha}$ and $\hat{\beta}$. We have the following results on the average values $\bar{\vartheta}_m$, $\bar{\vartheta}_s$ and $\bar{\vartheta}_{DR}$ compared to simulated $\boldsymbol{V}(\hat{\boldsymbol{\theta}})$:

$V(\hat{\alpha}) = 0.561$, $\bar{\vartheta}_{DR}(\hat{\alpha}) = 0.583$, $\bar{\vartheta}_s(\hat{\alpha}) = 0.525$, $\bar{\vartheta}_m(\hat{\alpha}) = 0.0058$,

$V(\hat{\beta}) = 0.0076$, $\bar{\vartheta}_{DR}(\hat{\beta}) = 0.0076$, $\bar{\vartheta}_s(\hat{\beta}) = 0.0069$, $\bar{\vartheta}_m(\hat{\beta}) = 0.0007$,

$Cov(\hat{\alpha}, \hat{\beta}) = -0.0041$, $\bar{\vartheta}_{DR}(\hat{\alpha}, \hat{\beta}) = -0.0042$, $\bar{\vartheta}_s(\hat{\alpha}, \hat{\beta}) = -0.0038$, $\bar{\vartheta}_m(\hat{\alpha}, \hat{\beta}) = 0.0004$,

It is clear that the DR variance and covariance estimators track the corresponding population values, while the use of ϑ_s only leads to underestimation of $V(\hat{\alpha})$ and $V(\hat{\beta})$.

CONCLUDING REMARKS

For estimators of model parameters defined as solutions to GREG calibration weighted estimating equations, we studied total variance estimation, assuming that the characteristics y_k in the finite population are generated from a superpopulation model. We obtained a linearization variance estimator, using the Demnati and Rao (2004) approach. The proposed variance estimator automatically preserves the "g-weights". Also, it remains valid under misspecification of the model variance of y_k, assuming that the model covariance of y_k and y_t is 0 for $k \neq t$. We plan to extend our results to longitudinal survey data, allowing for misspecification of the model covariance over time. Other extensions under study include two-phase sampling and missing responses.

REFERENCES

Demnati, A. and Rao, J. N. K. (2002), "Linearization Variance Estimators for Survey Data With Missing Responses", *Proceeding of the Section Survey Research Methods, American Statistical Association,* pp. 736-740.

Demnati, A. and Rao, J. N. K. (2004), "Linearization Variance Estimators for Survey Data (with discussion)", *Survey Methodology*, 30, pp. 17-34.

Lambert, D. (1992), "Zero-inflated Poisson Regression, With an Application to Defects in Manufacturing", *Technometrics*, 34, pp.1-14.

Molina, E. A., Smith, T. M. F. and Sugden, R. A. (2001), "Modeling Overdispersion for Complex Survey Data", *International Statistical Review, 69, pp. 373-384.*

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, pp. 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, pp. 527-537.

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