Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Linearization variance estimators for model parameters from complex survey data

by Abdellatif Demnati and J.N.K. Rao

December 2010

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2010
Vol. 36, No. 2, pp. 193-201
Statistics Canada, Catalogue No. 12-001-X

Linearization variance estimators for model parameters from complex survey data

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

Abstract

Taylor linearization methods are 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 unified approach to deriving Taylor linearization variance estimators that leads directly to a unique variance estimator that satisfies the above considerations for general designs. When analyzing survey data, finite populations are often assumed to be generated from super-population 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 the combined process of generating the finite population from the assumed super-population model and the selection of the sample according to the specified sampling design. In this paper, we obtain an estimator of the total variance, using the Demnati-Rao approach, when the characteristics of interest are assumed to be random variables generated from a super-population model. We illustrate the method using ratio estimators and estimators defined as solutions to calibration weighted estimating equations. Simulation results on the performance of the proposed variance estimator for model parameters are also presented.

Key Words: Calibration; Ratio estimators; Total variance; Logistic regression; Weighted estimating equations.

1. Introduction

In survey sampling, estimation of a finite population total $Y = \sum_{k=1}^{N} y_k \equiv Y(y)$ is often of interest, where N is the size of the finite population. For a general sampling design with positive inclusion probabilities π_k, a custumary design unbiased estimator of the total Y is given by $\hat{Y} = \sum_{i \in s} y_i/\pi_i \equiv \sum_{k=1}^{N} d_k(s) y_k$, where s is a sample, $d_k(s) = a_k(s)/\pi_k$ are the design weights with $a_k(s) = 1$ if $k \in s$ and $a_k(s) = 0$ otherwise. We use operator notation and write $\hat{Y}(z) = \sum_{k=1}^{N} d_k(s) z_k$ so that $\hat{Y} = \hat{Y}(y)$. Henceforth, all the sums are considered on the whole population and hence write $\sum_{k=1}^{N} y_k = \sum y_k$ and $\hat{Y}(z) = \sum d_k(s) z_k$, to simplify the notation. Again, using the operator notation, we denote an unbiased estimator of the variance of $\hat{Y}(z)$ as a quadratic function, $\vartheta(z)$, in the z_k's.

More complex estimators of a total Y based on known population auxiliary information, such as ratio and regression estimators, and estimators of more complex parameters obtained as solutions to sample weighted estimating equations, such as estimators of "census" logistic regression coefficients, are also often used in practice. Estimators that can be expressed as a general functional $T(\hat{M})$ have also been studied, where $\hat{M}$ denotes a measure that allocates the weight $d_k(s)$ to y_k; for example, $T(\hat{M}) = \int x d\hat{M}(x) = \sum d_k(s) y_k$ if the population parameter is the total $T(M) = \int x dM(x) = Y$, where the measure M allocates a unit mass to each y_k (Deville 1999). Large-sample estimation of the variance of such complex estimators, $\hat{\theta}$ say, has received considerable attention in the literature. In particular, Taylor linearization methods of estimating the variance of $\hat{\theta}$ are generally applicable to any sampling design that permits an unbiased variance estimator $\vartheta(z)$ of $\hat{Y}(z)$. Binder (1983) studied estimators $\hat{\theta}$ that are solutions to weighted estimating equations and applied Taylor linearization to obtain a variance estimator that can be expressed as $\vartheta(\tilde{z})$, where the linearized variable $\tilde{z}_k$ depends on unknown parameters, and $\tilde{z}_k$ is replaced by an estimator z_k that may be based on the substitution method. Deville (1999) derived a Taylor linearization variance estimator of the functional $T(\hat{M})$ as $\vartheta(\tilde{z})$, where $\tilde{z}_k = I_T(M; y_k)$ denotes the influence function of T at y_k, and then replaced $\tilde{z}_k$ by the sample estimator $z_{k1} = I_T(\hat{M}; y_k)$. For example, when $\hat{\theta}$ is the ratio estimator $(\hat{Y}/\hat{X})X = \hat{R}X$ of the total Y, where $\hat{X} = \hat{Y}(x)$ and $X = Y(x)$ is the known total of an auxiliary variable x, we get $\tilde{z}_k = y_k - Rx_k$ and $z_{k1} = y_k - \hat{R}x_k$. However, $z_k = (X/\hat{X})(y_k - \hat{R}x_k)$ is also a candidate to estimate $\tilde{z}_k$ and the resulting $\vartheta(z)$ is often preferred over $\vartheta(z_1)$; see Demnati and Rao (2004). Thus the choice of an

1. Abdellatif Demnati, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6. E-mail: Abdellatif.Demnati@statcan.gc.ca; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6. E-mail: JRao@math.carleton.ca.

estimator of $\tilde{z}_k$ is somewhat arbitrary under Deville's approach.

Demnati and Rao (2004) studied general estimators that can be expressed as smooth functions of the weights $\boldsymbol{d}(s) = \{d_1(s), ..., d_N(s)\}^T$, say $\hat{\theta} = f(\boldsymbol{d}(s))$, and obtained a Taylor linearization variance estimator directly as $\vartheta(z)$ with known linearized variables $z_k = \partial f(b)/\partial b_k |_{\boldsymbol{b}=\boldsymbol{d}(s)}$ without estimating $\tilde{z}_k$ first and then replacing it by an estimator. For example, in the case of the ratio estimator their method automatically leads to z_k given above. This method can be applied to a variety of estimators including estimators of "census" logistic regression parameters based on calibration weights (Demnati and Rao 2004). Previous work on direct variance estimation includes Binder (1996).

When analyzing survey data, the population values y_k, $k = 1, ..., N$, are often assumed to be generated from a super-population model, and the user is often interested in making inferences on the model parameters. Let θ_N be a "census" parameter, *i.e.*, an estimator of a model parameter θ when the population y_k -values are all known, and let $\hat{\theta}$ be a design-unbiased estimator of θ_N, the "census" parameter. Suppose that $\hat{\theta}$ is design-model unbiased for θ, *i.e.*, $E_m E_p(\hat{\theta}) = \theta$, where E_m and E_p respectively denote the expectations with respect to the design and the model. Then the total variance of $\hat{\theta}$ is $V(\hat{\theta}) = E_m E_p(\hat{\theta} - \theta)^2$ which can be decomposed as

$$V(\hat{\theta}) = E_m V_p(\hat{\theta}) + V_m(\theta_N), \qquad (1.1)$$

where $V_p(\hat{\theta}) = E_p(\hat{\theta} - \theta_N)^2$ is the design variance of $\hat{\theta}$ and $V_m(\theta_N)$ is the model variance of θ_N. It follows from (1.1) that the total variance may be estimated using a design-based estimator of $V_p(\hat{\theta})$ if the last term $V_m(\theta_N)$ is negligible relative to $E_m V_p(\hat{\theta})$. In that case, the distinction between θ_N and θ can be ignored (Skinner, Holt and Smith 1989, page 14). On the other hand, it is necessary to estimate the total variance $V(\hat{\theta})$ when the model variance $V_m(\theta_N)$ is not negligable relative to $E_m V_p(\hat{\theta})$. This requires consideration of the joint design and model random processes. Molina, Smith and Sugden (2001) argued that the combined process of generation of the finite population and selection of the sample should be the basis for analytical inferences on model parameters. Rubin-Bleuer and Şchiopu-Kratina (2005) have provided a mathematical framework for joint model and design-based inference. However, a broadly applicable method is needed for the estimation of total variance. The main purpose of this paper is to provide such a method, by extending the Demnati-Rao approach for finite population parameters.

In Section 2, we consider the case of a scalar parameter θ and present linearization variance estimators by expanding the Demnati and Rao (2004) approach. The method is illustrated for the special case of a ratio estimator of a super-population mean θ. Results of Section 2 are extended in Section 3 to estimators of a vector parameter $\boldsymbol{\theta}$ obtained as solutions to weighted estimating equations, and the method is illustrated for the special case of parameters of a logistic regression model. Simulation results are also presented.

2. Scalar model parameter

2.1 Point estimators

Consider a finite population U of N elements, and let $d_k(s) = a_k(s)/\pi_k$ be the design weights attached to the population element k, where $a_k(s) = 1$ if element k is in the sample s and $a_k(s) = 0$ otherwise, and π_k is the inclusion probability associated with k. We consider estimators $\hat{\theta}$ of a scalar parameter θ that can be expressed as functions of random variables under the design and the assumed model. In particular, $\hat{\theta} = f(\boldsymbol{A}_d)$, where $\boldsymbol{A}_d$ is a $(p+1) \times N$ matrix with columns $\boldsymbol{d}_k = (d_k h_{1k}, d_k h_{2k}, ..., d_k h_{(p+1)k})^T \equiv (d_{1k}, ..., d_{(p+1)k})^T$ where $d_k = d_k(s)$ is random under the design, $h_{1k} = 1$, and h_{ik} $(i = 2, ..., p+1)$ are random under the model.

For example, consider the ratio model with fixed covariates x_k:

$$E_m(y_k) = \beta x_k,\ V_m(y_k) = \sigma^2 x_k,\ \mathrm{Cov}_m(y_k,\ y_t) = 0,\quad k \neq t,\ k, t = 1, ..., N, \qquad (2.1)$$

where E_m, V_m, and Cov_m denote model expectation, model variance, and model covariance respectively and $\sigma^2 > 0$. Suppose that we are interested in estimating the super-population mean $\theta = E_m(\bar{Y}) = N^{-1} \sum E_m(y_k) = \beta \bar{X}$ where $\bar{Y}$ is the finite population mean of y. In this case, a ratio estimator of θ is given by

$$\hat{\theta} = \bar{X}(\hat{Y}/\hat{X}) \equiv \bar{X}\hat{R}, \qquad (2.2)$$

where $\hat{Y} = \sum d_k(s) y_k$ and $\hat{X} = \sum d_k(s) x_k$ are the design-unbised estimators of the totals Y and X, and $\bar{X}$ is the know population mean of x. We can write the ratio estimator (2.2) in the form $\hat{\theta} = \bar{X}(\sum d_{2k})/\sum d_{1k} x_k$, where $d_{1k} = d_k(s)$ and $d_{2k} = d_k(s) y_k$. This is a special case of $f(\boldsymbol{A}_d)$ with $p = 1$ and $h_{2k} = y_k$.

Let E_p be the design expectation and $E = E_m E_p$ be the total expectation. Then, we have $E(d_{1k}) = E_m(1) = 1 \equiv \mu_{1k}$ and $E(d_{ik}) = E_m(g_{ik}) \equiv \mu_{ik}$, $i = 2, ..., p+1$, noting that $E_p(d_k(s)) = 1$. We assume that $f(\boldsymbol{A}_\mu) = \theta$, where $\boldsymbol{A}_\mu$ is a $(p+1) \times N$ matrix with columns $\boldsymbol{\mu}_k = (\mu_{1k}, \mu_{2k}, ..., \mu_{(p+1)k})^T$. Hence, $\hat{\theta}$ is asymptotically pm -unbised for θ. In the special case of the ratio estimator, we have $f(\boldsymbol{A}_\mu) = \beta \bar{X} = \theta$, noting that $\mu_{1k} = 1$ and $\mu_{2k} = \beta x_k$.

2.2 Linearization variance estimator

We first derive an estimator of the total variance of a linear estimator $\hat{U} = \sum \boldsymbol{u}_k^T \boldsymbol{d}_k$, where $\boldsymbol{u}_k$ is a vector of constants. The total variance of $\hat{U}$ may be decomposed as

$$V(\hat{U}) = E_m V_p(\hat{U}) + V_m E_p(\hat{U}) \equiv I + II, \quad (2.3)$$

where V_p and V_m denote design variance and model variance respectively. A design-unbiased estimator of the component I of the total variance (2.3) is obtained by estimating the design variance $V_p(\hat{U})$ for fixed $\boldsymbol{h}_k = (h_{1k}, ..., h_{(p+1)k})^T$. Now, noting that $\hat{U} = \sum b_k d_k(s)$ is the standard Narain-Horvitz-Thompson (NHT) estimator of the total $U = \sum b_k$ when $b_k = \boldsymbol{u}_k^T \boldsymbol{h}_k$ are fixed conditionally, we can use either the Sen-Yates-Grandy (SYG) variance estimator for fixed sample size designs or the Horvitz-Thompson (HT) variance estimator for arbitrary designs. The SYG estimator is given by

$$\begin{aligned} \text{est}(I) &= \vartheta_{\text{SYG}}(\hat{U}) \\ &= \sum\sum_{k<t} d_{kt}(s) \frac{(\pi_k \pi_t - \pi_{kt})}{\pi_k \pi_t} (b_k - b_t)^2, \end{aligned} \quad (2.4)$$

where $d_{kt}(s) = \{a_k(s)\, a_t(s)\}/\pi_{kt}$ and π_{kt} is the inclusion probability for units k and t $(k \neq t)$. The HT variance estimator is given by

$$\text{est}(I) = \vartheta_{\text{HT}}(\hat{U}) = \sum\sum_{} d_{kt}(s) \frac{(\pi_{kt} - \pi_k \pi_t)}{\pi_k \pi_t} b_k b_t, \quad (2.5)$$

where $d_{kk}(s) = d_k(s)$. For the special case of stratified random sampling (2.4) and (2.5) are identical.

Turning to the component II of the total variance (2.3), we have $V_m E_p(\hat{U}) = V_m(\sum \boldsymbol{u}_k^T \boldsymbol{h}_k) = \sum\sum \boldsymbol{u}_k^T \text{Cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t)\, \boldsymbol{u}_t$ and a pm-unbiased estimator is therefore given by

$$\text{est}(II) = \sum\sum d_{kt}(s)\, \boldsymbol{u}_k^T \text{cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t)\, \boldsymbol{u}_t, \quad (2.6)$$

after replacing $\text{Cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t)$ by an estimator $\text{cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t)$. The estimator of total variance (2.3) is now given by $\text{est}(I) + \text{est}(II)$. We denote it, in operator notation, as $\vartheta(\boldsymbol{u})$.

We now turn to the estimation of total variance of $\hat{\theta}$. Following Demnati and Rao (2004), a Taylor expansion of $\hat{\theta} - \theta$ may be written as

$$\hat{\theta} - \theta \approx \sum \tilde{\boldsymbol{z}}_k^T (\boldsymbol{d}_k - \boldsymbol{\mu}_k) \quad (2.7)$$

where $\tilde{\boldsymbol{z}}_k = \partial f(\boldsymbol{A}_b)/\partial \boldsymbol{b}_k \,|_{\boldsymbol{A}_b = \boldsymbol{A}_\mu}$ and $\boldsymbol{A}_b$ is a $(p+1) \times N$ matrix with k^{th} column $\boldsymbol{b}_k$, a vector of arbitrary real numbers. The approximation (2.7) is valid for any $\hat{\theta}$ that can be expressed as a smooth function of estimated totals. Following Demnati and Rao (2004), a linearization estimator of the total variance is now given by

$$\vartheta_{\text{DR}}(\hat{\theta}) = \vartheta(\boldsymbol{z}), \quad (2.8)$$

which is obtained from $\vartheta(\boldsymbol{u})$ by replacing $\boldsymbol{u}_k$ by the "linearized variable" $\boldsymbol{z}_k = \partial f(\boldsymbol{A}_b)/\partial \boldsymbol{b}_k \,|_{\boldsymbol{A}_b = \boldsymbol{A}_d}$. A rigorous theoretical justification of (2.8) follows along the lines of Deville (1999).

2.3 Special case of ratio estimator

For the ratio estimator $\hat{\theta} = \bar{X}\hat{R}$ of the model parameter $\theta = \beta \bar{X}$, $\boldsymbol{z}_k$ reduces to

$$\boldsymbol{z}_k = (\bar{X}/\hat{X})(-\hat{R}\, x_k, 1)^T = (z_{1k}, z_{2k})^T. \quad (2.9)$$

Further, b_k in (2.4) or (2.5) is replaced by

$$\begin{aligned} \boldsymbol{z}_k^T \boldsymbol{h}_k &= z_{1k} + z_{2k} y_k \\ &= (\bar{X}/\hat{X})(y_k - \hat{R}\, x_k) \equiv (\bar{X}/\hat{X})\, e_k, \end{aligned}$$

using (2.9). Also, replacing $\boldsymbol{u}_k$ by $\boldsymbol{z}_k$ in (2.6) we get

$$\boldsymbol{z}_k^T \text{cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t)\, \boldsymbol{z}_t = z_{2k} z_{2t} \text{cov}_m(y_k, y_t).$$

Under the ratio model (2.1) with unspecified model variance $V_m(y_k) = \sigma_k^2$, $k = 1, ..., N$, we can estimate $\sigma_k^2 = E_m(y_k - \beta x_k)^2$ by $(y_k - \hat{R} x_k)^2$ and letting $\text{cov}_m(y_k, y_t) = 0$, for $k \neq t$.

We now study the special case of simple random sampling without replacement. In this case, both (2.4) and (2.5) reduce to

$$\text{est}(I) = \left(\frac{\bar{X}}{\bar{x}}\right)^2 \frac{1}{n}\left(1 - \frac{n}{N}\right) s_e^2, \quad (2.10)$$

where $s_e^2 = \sum a_k(s)\, e_k^2/(n-1)$, and (2.6) reduces to

$$\text{est}(II) = \left(\frac{\bar{X}}{\bar{x}}\right)^2 \frac{(n-1)}{nN} s_e^2. \quad (2.11)$$

Hence, using (2.10) and (2.11), the variance estimator (2.8) reduces to

$$\begin{aligned} \vartheta_{\text{DR}}(\hat{\theta}) &= \text{est}(I) + \text{est}(II) \\ &= \left(\frac{\bar{X}}{\bar{x}}\right)^2 \frac{1}{n} \frac{N-1}{N} s_e^2. \end{aligned} \quad (2.12)$$

It is interesting to note that the "g-weight" $\bar{X}/\bar{x}$ appears automatically in $\vartheta_{\text{DR}}(\hat{\theta})$, given by (2.12), and that the finite population correction $1 - n/N$ is absent in $\vartheta_{\text{DR}}(\hat{\theta})$ unlike in $est(I)$ given by (2.10).

In the customary approach to the estimation of total variance (see *e.g.*, Korn and Graubard 1998) $V(\hat{\theta})$ is first written as

$$V(\hat{\theta}) = E_m V_p(\hat{\theta}) + V_m E_p(\hat{\theta})$$

$$\approx E_m V_p(\hat{\theta}) + V_m(\bar{Y})$$

$$= E_m V_p(\hat{\theta}) + N^{-2} \sum E_m (y_k - \beta x_k)^2, \quad (2.13)$$

under the ratio model with unspecified σ_k^2, $k = 1, ..., N$. The first term $E_m V_p(\hat{\theta})$ in (2.13) is then estimated by a design-consistent estimator of $V_p(\hat{\theta})$, typically by (2.10) without the g-factor $(\bar{X}/\bar{x})^2$. The second term is estimated by $N^{-2} \sum d_k(s)(y_k - \hat{R} x_k)^2 = (nN)^{-1}(n-1)s_e^2$. The sum of the two estimated terms then equals (2.12) without the g-factor. We denote this customary variance estimator by $\vartheta_{\text{cus}}(\hat{\theta})$. On the other hand, if (2.10) with the g-factor is used to estimate $V_p(\hat{\theta})$, the sum of this estimated term and the previous estimator of the second term leads to a "hybrid" variance estimator

$$\vartheta_{\text{mix}}(\hat{\theta}) = \text{est}(I) + (nN)^{-1}(n-1)s_e^2,$$

where the g-term is absent in the last term. It is clear from the above results that the choice of estimator of total variance under the customary approach is not unique, unlike under the proposed approach.

If the parameter of interest is $\beta = \theta/\bar{X}$ instead of θ, then $\hat{\beta} = \hat{\theta}/\bar{X} = \hat{R}$ and $\vartheta_{\text{DR}}(\hat{\beta})$ under simple random sampling is give by

$$\vartheta_{\text{DR}}(\hat{\beta}) = \bar{X}^{-2} \vartheta_{\text{DR}}(\hat{\theta}) = \bar{x}^{-2} \frac{1}{n} \frac{N-1}{N} s_e^2. \quad (2.14)$$

The customary approach leads to the same variance estimator, (2.14).

2.4 Simulation study

We conducted a small simulation study to examine the performances of different variance estimators, both unconditionally and conditionally on $\hat{X}$. We first generated $R = 2{,}000$ finite populations $\{y_1, ..., y_N\}$ each of size $N = 393$, from the ratio model

$$y_k = 2x_k + x_k^{1/2} \varepsilon_k, \quad (2.15)$$

with independent values ε_k generated from $N(0, 1)$, where the fixed x_k are the "number of beds" for the Hospitals population studied in Valliant, Dorfman and Royall (2000, page 424-427). One simple random sample of specified size n is drawn from each generated population. Our parameter of interest is $\theta = \beta \bar{X}$, where $\beta = 2$.

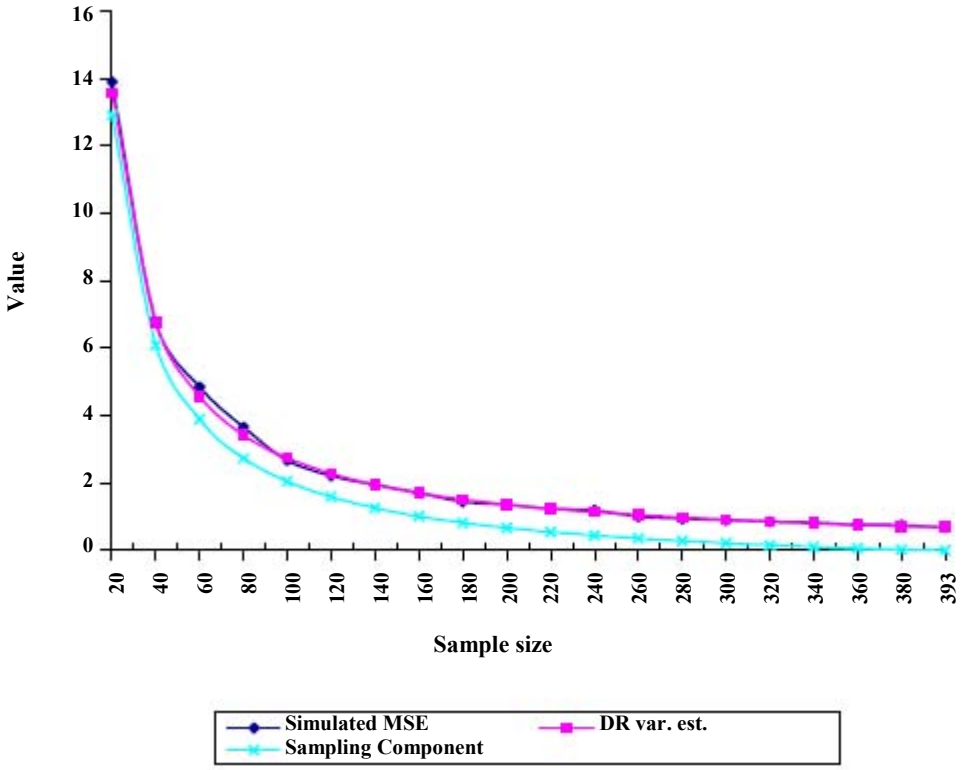

Figure 1 **Averages of variance estimates for selected sample sizes compared to estimated MSE of the ratio estimator. $\bar{\vartheta}_{\text{DR}}$ = DR var. est., $\bar{\vartheta}_s$ = Sampling component: ratio model**

Simulated total MSE of the ratio estimator $\hat{\theta} = \bar{X}(\bar{y}/\bar{x})$ is calculated as $M(\hat{\theta}) = R^{-1}\sum_{r=1}^{2,000}(\hat{\theta}_r - \theta)^2$, where $\hat{\theta}_r$ is the value of $\hat{\theta}$ for the r^{th} simulated sample and $(\bar{y}, \bar{x})$ are the sample means. We calculated the total variance estimate $\vartheta_{\text{DR}}(\hat{\theta})$, and its components $\vartheta_s = \text{est}(I)$ and $\vartheta_m = \text{est}(II)$ from each simulated sample r and their averages $\bar{\vartheta}_{\text{DR}}$, $\bar{\vartheta}_s$, and $\bar{\vartheta}_m$ over r. Figure 1 gives a plot of the average of variance estimates, $\bar{\vartheta}_{\text{DR}}$ and $\bar{\vartheta}_s$, and the simulated total MSE for $n = 20, 40, ..., 380, 393$. In the case of $n = N$, $\bar{\vartheta}_s = 0$. It is seen from Figure 1, that ϑ_{DR} is approxiamatly unbiased, whereas ϑ_s leads to severe underestimation as the sample size, n, increases.

We also examined the conditional performance of the variance estimators under simple random sampling given $\bar{x}$, by conducting another simulation study for inference on θ, using model (2.15). The study is similar to the study of Royall and Cumberland (1981) for inference on the finite population mean $\theta_N = \bar{Y}$ from a fixed population $\{y_1, ..., y_N\}$. We generated $R = 20{,}000$ finite populations $\{y_1, ..., y_N\}$, each of size $N = 393$ from (2.15) using the number of beds as x_k, and from each population we then selected one simple random sample of size $n = 100$. We arranged the 20,000 samples in ascending order of $\bar{x}$-values and then grouped them into 20 groups each of size 1,000 such that the first group, G_1, contained 1,000 samples with the smallest $\bar{x}$-values, the next group, G_2, contained the next 1,000 smallest $\bar{x}$-values, and so on to get $G_1, ..., G_{20}$. For each of the 20 groups so formed, we calculated the average values of the ratio estimates $\hat{\theta} = \bar{X}(\bar{y}/\bar{x})$ and the mean estimates $\bar{y}$, and the resulting conditional relative bias (CRB) in estimating $\theta = 2\bar{X}$; see Figure 2. It is clear from Figure 2 that $\bar{y}$ is conditionally biased unlike $\hat{\theta}$: negative CRB (-14%) for G_1 increasing to positive CRB (+14%) for G_{20}. Note that both $\bar{y}$ and $\hat{\theta}$ are unconditionally unbiased for θ. The conditional bias of $\hat{\theta}$ and $\bar{y}$ in estimating the model parameter θ is similar to the conditional bias in estimating the "census" parameter $\theta_N = \bar{Y}$, as observed by Royall and Cumberland (1981).

We also calculated the conditional MSE of $\hat{\theta}$ and the associated CRB of the variance estimators ϑ_{DR}, ϑ_{cus} and ϑ_{mix} based on the average values of ϑ_{DR}, ϑ_{cus} and ϑ_{mix} in each group; see Figure 3. It is evident from Figure 3 that CRB of ϑ_{cus} ranges from -28% to 20% across the groups whereas ϑ_{DR} exhibits no such trend and its CRB is less than 5% in absolute value except for G_6 and G_{20}. Also, the CRB of ϑ_{mix} is largely negative and below that of ϑ_{DR} for the first half of the groups and above for the second half, but ϑ_{mix} exhibits no visible trends unlike ϑ_{cus}.

Figure 4 reports the conditional coverage rates (CCR) of normal theory confidence intervals based on ϑ_{DR}, ϑ_{cus}, ϑ_{mix} and ϑ_s (ignoring the component ϑ_m) for nominal level of 95%. As expected, the use of ϑ_s leads to severe undercoverage because the sampling fraction, 100/393, is significant. On the other hand, CCR associated with ϑ_{DR} is closer to nominal level across groups, while ϑ_{cus} exhibits a trend across groups with CCR ranging from 91% to 97%. Further, CCR associated with ϑ_{mix} is slightly below that of ϑ_{DR} for the first half of the groups but ϑ_{mix} and ϑ_{DR} perform similarly.

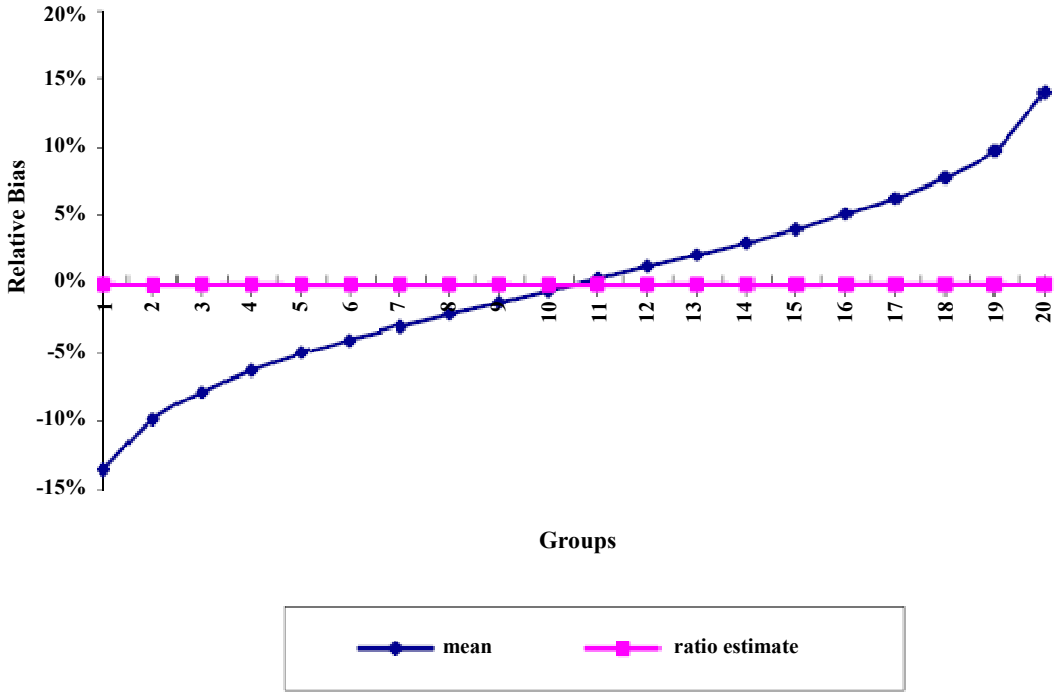

Figure 2 Conditional relative bias of the expansion and ratio estimators: ratio model

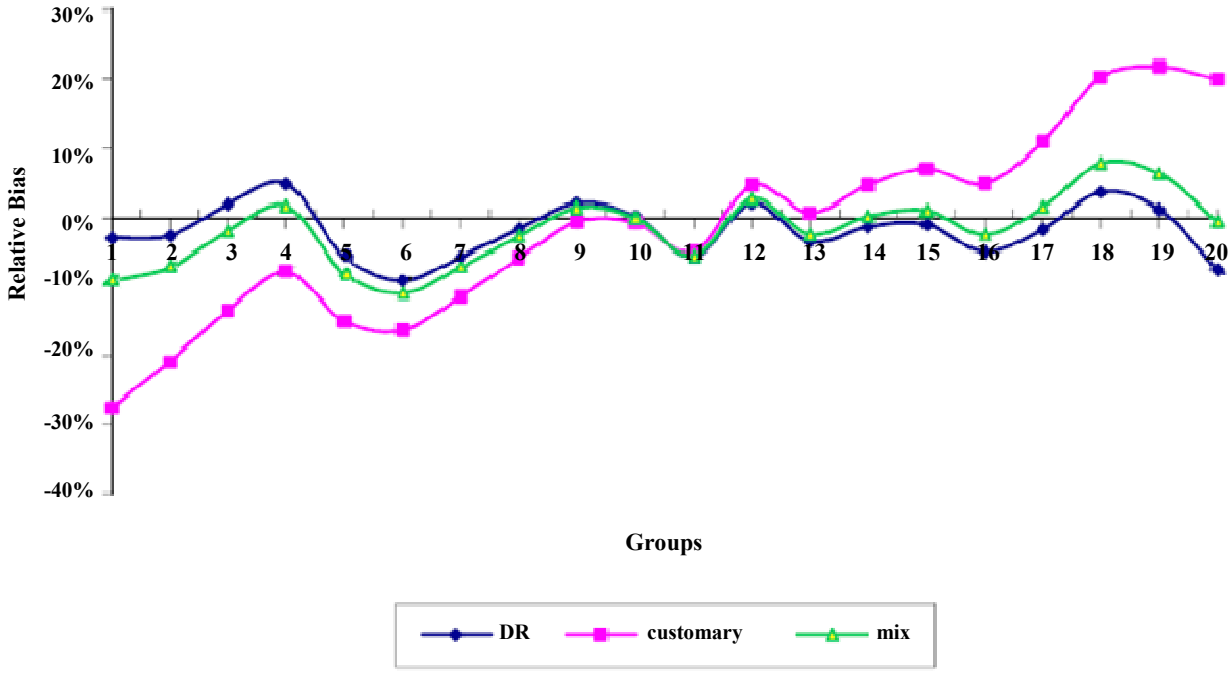

Figure 3 Conditional relative bias of variance estimators ϑ_{DR}, ϑ_{cus} and ϑ_{mix}: ratio model

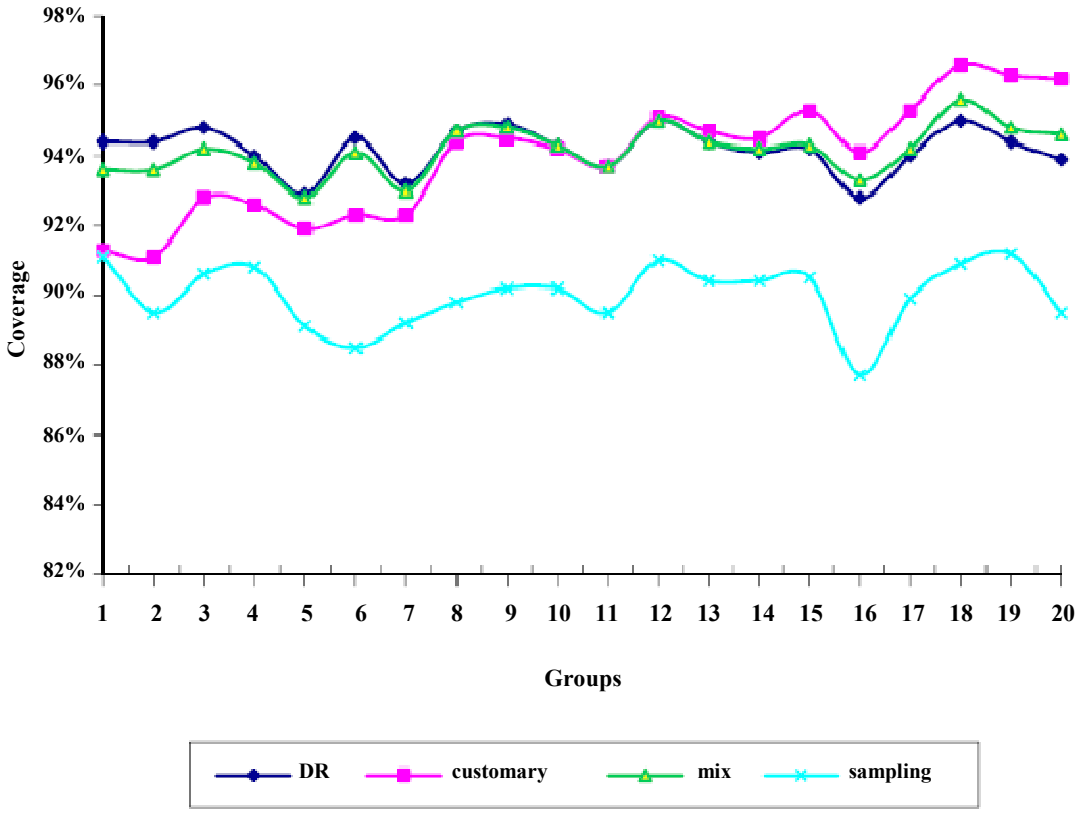

Figure 4 Conditional coverage rates of normal theory confidence intervals based on ϑ_{DR}, ϑ_{cus}, ϑ_{mix} and ϑ_{s} for nominal level of 95%: ratio model

3. Calibration weighted estimating equations

3.1 Estimators of model parameters

Suppose that the super-population model on the responses y_k is specified by a generalized linear model (McCullagh and Nelder 1989) with mean $E_m(y_k) = \mu_k(\boldsymbol{\theta}) = h(\boldsymbol{x}_k^T\boldsymbol{\theta})$, where $\boldsymbol{x}_k$ is a $p\times 1$ vector of explanatory variables, $\boldsymbol{\theta}$ is the p-vector of model parameters and $h(.)$ is a "link" function. For example, $h(a) = a$ gives a linear regression model and $h(a) = e^a/(1+e^a)$ gives a logistic regression model for binary responses y_k.

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

$$g_k(\boldsymbol{d}(s)) = 1 + (\boldsymbol{T} - \hat{\boldsymbol{T}})^T\left[\sum d_k(s)c_k\boldsymbol{t}_k\boldsymbol{t}_k^T\right]^{-1}c_k\boldsymbol{t}_k,$$

for specified c_k, where $\hat{\boldsymbol{T}} = \sum d_k(s)\boldsymbol{t}_k$ is the HT estimator of the known total $\boldsymbol{T}$ of a $q\times 1$ vector of calibration variables $\boldsymbol{t}_k$ and $\boldsymbol{d}(s)$ is the $N\times 1$ vector of the weights $d_k(s)$. The GREG weights, $w_k(s)$, have the calibration property $\sum w_k(s)\boldsymbol{t}_k = \boldsymbol{T}$ and lead to efficient estimators $\tilde{Y} = \sum w_k(s)y_k$ of totals $Y = \sum y_k$, when y_k and t_k are linearly related (Särndal, Swensson and Wretman 1989, chapter 6).

We use the calibration weights, $w_k(s)$, to estimate the CEE. The calibration weighted estimating equations are given by

$$\tilde{\boldsymbol{l}}(\boldsymbol{\theta}) = \sum w_k(s)\boldsymbol{l}_k(\boldsymbol{\theta}) = \sum d_k(s)g_k(\boldsymbol{d}(s))\boldsymbol{l}_k(\boldsymbol{\theta}) = \mathbf{0}. \quad (3.1)$$

The solution to (3.1), obtained by the Newton-Raphson-type iterative method, gives the calibration-weighted estimator $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, and $\tilde{\boldsymbol{\theta}}$ is approximately design-model unbiased for $\boldsymbol{\theta}$, *i.e.*, $E(\tilde{\boldsymbol{\theta}}) \approx \boldsymbol{\theta}$. It follows from (3.1) that $\tilde{\boldsymbol{\theta}}$ 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 $\boldsymbol{d}_k$. Here we have $h_{1k} = 1$ and $(h_{2k}, ..., h_{(p+1)k}) = \boldsymbol{l}_k(\boldsymbol{\theta})$.

3.2 Linearized variance estimators

We first extend the result on variance estimation for the scalar case $\hat{U} = \sum \boldsymbol{b}_k^T\boldsymbol{d}_k$ (Section 2.2) to the vector case $\hat{\boldsymbol{U}} = \sum \boldsymbol{U}_k\boldsymbol{d}_k = \sum \boldsymbol{b}_k^T d_k(s)$, where $\boldsymbol{b}_k = \boldsymbol{U}_k\boldsymbol{h}_k$ is a p-vector and $\boldsymbol{U}_k$ is a $p\times(p+1)$ matrix with rows $\boldsymbol{u}_{jk}^T$, $j = 1, ..., p$. In this case, the SYG variance estimator (2.4) is changed to

$$\mathbf{est}(I) = \vartheta_{\text{SYG}}(\hat{U}) = \sum\sum_{k<t} d_{kt}(s)\frac{(\pi_k\pi_t - \pi_{kt})}{\pi_k\pi_t}(\boldsymbol{b}_k - \boldsymbol{b}_t)(\boldsymbol{b}_k - \boldsymbol{b}_t)^T. \quad (3.2)$$

Similarly, the H-T variance estimator (2.5) is changed to

$$\mathbf{est}(I) = \vartheta_{\text{HT}}(\hat{U}) = \sum\sum d_{kt}(s)\frac{(\pi_{kt} - \pi_k\pi_t)}{\pi_k\pi_t}\boldsymbol{b}_k\boldsymbol{b}_t^T. \quad (3.3)$$

Turning to the component II of the total variance of $\hat{\boldsymbol{U}}$, (2.6) is changed to

$$\mathbf{est}(II) = \sum\sum d_{kt}(s)\boldsymbol{U}_k\,\text{cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t)\boldsymbol{U}_t^T. \quad (3.4)$$

The total variance of $\hat{\boldsymbol{U}}$ is estimated by the sum of (3.2) and (3.4) for fixed sample size designs or by the sum of (3.3) and (3.4) for arbitrary designs.

A linearization variance estimator of the total variance of $\tilde{\boldsymbol{\theta}}$ is obtained from the estimated total variance estimator of $\hat{\boldsymbol{U}}$ by replacing $\boldsymbol{U}_k$ by the linearized variable $\boldsymbol{Z}_k = \partial\boldsymbol{f}(\boldsymbol{A}_b)/\partial\boldsymbol{b}_k\,|_{\boldsymbol{A}_b = \boldsymbol{A}_d}$. Following the implicit differentiation method of Demnati and Rao (2004), $\boldsymbol{Z}_k$ reduces to

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

with

$$\hat{\boldsymbol{B}}_l = \left[\sum d_k(s)c_k\boldsymbol{t}_k\boldsymbol{t}_k^T\right]^{-1}\sum d_k(s)c_k\boldsymbol{t}_k\boldsymbol{l}_k^T(\tilde{\boldsymbol{\theta}}),$$

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

and $\boldsymbol{I}_p$ is the $p\times p$ identity matrix.

After some simplification, the first component $\mathbf{est}(I)$ is given by (3.2) or (3.3) with $\boldsymbol{b}_k$ changed to

$$\boldsymbol{Z}_k\boldsymbol{h}_k = [\tilde{\boldsymbol{J}}(\tilde{\boldsymbol{\theta}})]^{-1}e_k(\tilde{\boldsymbol{\theta}})g_k(\boldsymbol{d}(s)), \quad (3.5)$$

where

$$\boldsymbol{e}_k(\tilde{\boldsymbol{\theta}}) = \boldsymbol{l}_k(\tilde{\boldsymbol{\theta}}) - \hat{\boldsymbol{B}}_l^T\boldsymbol{t}_k.$$

Similarly, the second component $\mathbf{est}(II)$ simplifies to

$$\mathbf{est}(II) = [\tilde{\boldsymbol{J}}(\tilde{\boldsymbol{\theta}})]^{-1}\sum d_k(s)g_k^2(\boldsymbol{d}(s))\boldsymbol{l}_k(\tilde{\boldsymbol{\theta}})\boldsymbol{l}_k^T(\tilde{\boldsymbol{\theta}})[\tilde{\boldsymbol{J}}(\tilde{\boldsymbol{\theta}})]^{-1}, \quad (3.6)$$

if $\text{Cov}_m[\boldsymbol{l}_k(\tilde{\boldsymbol{\theta}})\boldsymbol{l}_t^T(\tilde{\boldsymbol{\theta}})] = \mathbf{0}$ for $k \neq t$.

The total variance estimator of $\tilde{\boldsymbol{\theta}}$ is now estimated by

$$\vartheta_{\text{DR}}(\tilde{\boldsymbol{\theta}}) = \mathbf{est}(I) + \mathbf{est}(II). \quad (3.7)$$

This variance estimator of $\tilde{\boldsymbol{\theta}}$ automatically takes account of the g-weights as in Section 2.

A customary variance estimator of $\tilde{\boldsymbol{\theta}}$, $\vartheta_{\text{cus}}(\tilde{\boldsymbol{\theta}})$, is obtained from (3.7) by ignoring the g-weights in (3.5) and (3.6). Similarly, a hybrid variance estimator, $\vartheta_{\text{mix}}(\tilde{\boldsymbol{\theta}})$, is

obtained from (3.7) by retaining the g-weights in $\mathbf{est}(I)$ and ignoring them in $\mathbf{est}(II)$.

3.3 Simulation study

We conducted a simulation study to compare the relative performances of the three variance estimators ϑ_{DR}, ϑ_{cus}, and ϑ_{mix}, for the special case of a logistic regression model:

$$E_m(y_k) = \mu_k(\boldsymbol{\theta}) = \exp(\boldsymbol{x}_k^T\boldsymbol{\theta})/\{1+\exp(\boldsymbol{x}_k^T\boldsymbol{\theta})\} \qquad (3.8)$$

$$V_m(y_k) = \mu_k(\boldsymbol{\theta})(1-\mu_k(\boldsymbol{\theta})),\ \mathrm{Cov}_m(y_k, y_t) = 0,\ k \neq t.$$

In this case, we have $\boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{x}_k(y_k - \mu_k(\boldsymbol{\theta}))$, and

$$\tilde{\boldsymbol{J}}(\boldsymbol{\theta}) = \sum d_k(s) g_k(\boldsymbol{d}(s)) \boldsymbol{x}_k \boldsymbol{x}_k^T \mu_k(\boldsymbol{\theta})(1-\mu_k(\boldsymbol{\theta})).$$

For the simulation study, we set $\boldsymbol{x}_k = (1,\ x_k)^T$, where the x_k denote the number of beds for the Hospitals population of size $N = 393$ studied in Section 2.2. We implemented post-stratification by dividing the population into two classes with $N_1 = 171$ hospitals k having $x_k < 350$ in class 1 and $N_2 = 122$ hospitals k with $x_k \geq 350$ in class 2. Here, $g_k(\boldsymbol{d}(s)) = N_h/\hat{N}_h$, $h = 1, 2$, if k belongs to class h, where $\hat{N}_h = \sum d_k(s) t_{hk}$ is the design-weight estimator of N_h, and $\boldsymbol{t}_k = (t_{1k}, t_{2k})^T$ is the vector of class indicator variables t_{hk}.

We generated $R = 40{,}000$ finite populations $\{y_1, ..., y_N\}$, each of size $N = 393$, assuming the logistic regression model (3.8) with $\boldsymbol{\theta} = (\theta_0, \theta_1)^T = (-1,\ 0.005)^T$. The parameter of interest is $\theta_1 = 0.005$. From each generated population, we selected one simple random sample of size $n = 150$, and then obtained the calibration-weighted estimated $\tilde{\theta}_1$ and associated variance estimators $\mathrm{est}(I) = \vartheta_s(\tilde{\theta}_1)$, $\vartheta_{DR}(\tilde{\theta}_1)$, $\vartheta_{cus}(\tilde{\theta}_1)$ and $\vartheta_{mix}(\tilde{\theta}_1)$ from each sample r. We obtained the averages of the estimates and the variance estimates as $\mathrm{av}(\hat{\theta}_1) \approx 0.00514$, $\mathrm{av}(\vartheta_{DR}) \approx 0.0989$, $\mathrm{av}(\vartheta_{cus}) \approx 0.0987$, $\mathrm{av}(\vartheta_{mix}) \approx 0.0988$, and $\mathrm{av}(\vartheta_s) \approx 0.0613$. Also, the estimated total MSE of $\hat{\theta}_1$ is equal to 0.0998. Hence, unconditionally the estimator $\tilde{\theta}_1$ is approximately unbiased for θ_1, and the bias of the three variance estimators ϑ_{DR}, ϑ_{cus} and ϑ_{mix} is negligible. On the other hand ignoring the second component and using only the first component, $\mathrm{est}(I) = \vartheta_s(\tilde{\theta}_1)$, leads to severe underestimation, as expected.

We also examined the conditional performances of the three variance estimators along the line of Section 2.2. We arranged the 40,000 samples in ascending order of the sample size, n_1, in class 1, and then grouped the samples into twenty groups, each of size 2,000, such that the first group, G_1, contained the 2,000 samples with the smallest n_1-values, the second group, G_2, contained the 2,000 samples with the next smallest n_1-values, and so on to get twenty groups, $G_1, ..., G_{20}$.

We calculated the conditional MSE of $\tilde{\theta}_1$ and the associated conditional relative bias (CRB) of the variance estimators ϑ_{DR}, ϑ_{cus} and ϑ_{mix} based on the average values of ϑ_{DR}, ϑ_{cus} and ϑ_{mix} in each group; see Figure 5. We can see from Figure 5 that CRB of ϑ_{cus} ranges from 20% to -20% across the groups, whereas ϑ_{DR} exibits no such trend and its CRB is less than 5% in absulate value except for two groups. Also, the CRB of ϑ_{mix} exhibits a trend but less prononced than ϑ_{cus}. Figure 6 reports the conditional coverage rates (CCR) of normal theory intervals based on ϑ_{DR}, ϑ_{cus} and ϑ_{mix} for nominal level of 95%. We can see from Figure 6 that ϑ_{cus} exhibits a trend across groups with CCR ranging from 97% to 92%, whereas CCR associated with ϑ_{DR} is close to the nominal level across groups. Further, CCR associated with ϑ_{mix} is slighthy above that of ϑ_{DR} for the first half of the groups and slighty below for the remaing groups.

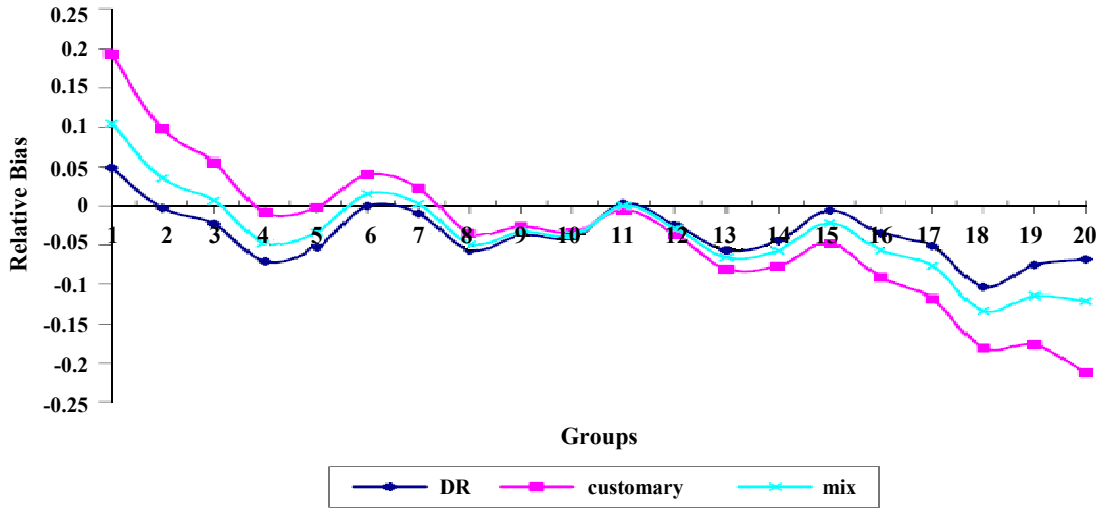

Figure 5 Conditional relative bias of variance estimators: logistic regression

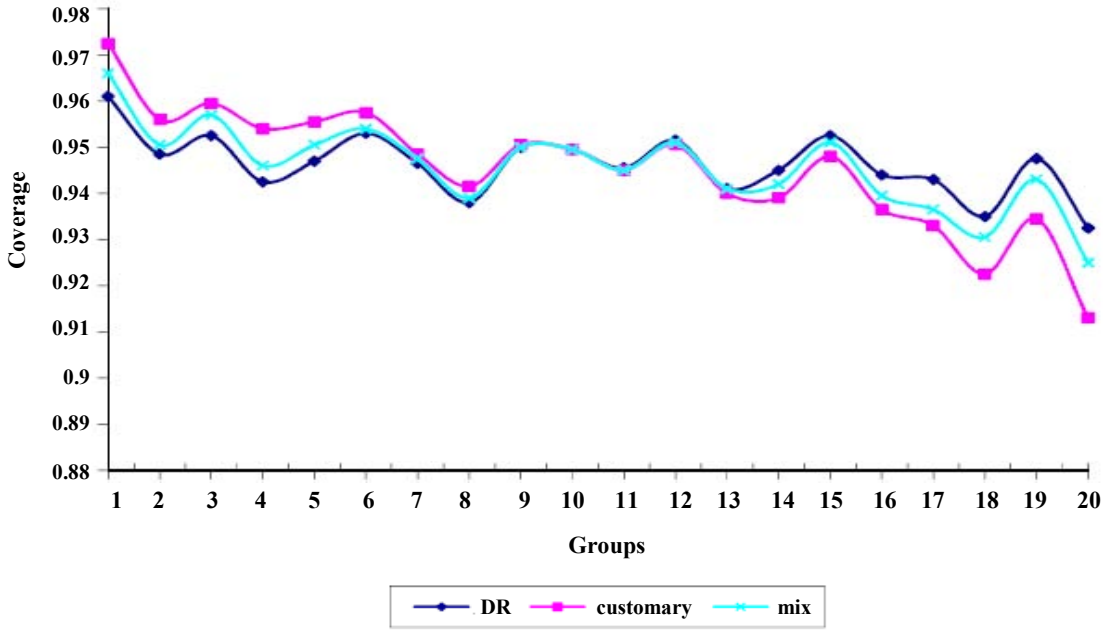

Figure 6 Conditional coverage rates of normal theory confidence intervals for nominal level of 95%: logistic regression

Concluding remarks

We have studied the estimation of total variance of estimators of model parameters under an assumed super-population model. Our approach leads directly to a linearization variance estimator which is shown to perform well under a conditional framework when calibration weights are used for estimation. We are currently investigating extensions of our method to estimation of total variance under imputation for item nonresponse and integration of two independent surveys.

Acknowledgements

We thank two referees for constructive comments and suggestions. J.N.K. Rao's work was partially supported by a grant from Natural Sciences and Engineering Research Council of Canada.

References

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

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. (2004). Linearization variance estimators for survey data (with discussion). *Survey Methodology*, 30, 17-34.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 193-203.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*, New York: John Wiley & Sons, Inc.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

Molina, E.A., Smith, T.M.F. and Sugden, R.A. (2001). Modeling overdispersion for complex survey data. *International Statistical Review*, 69, 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, 66-77.

Rubin-Bleuer, S., and Şchiopu-Kratina, I. (2005). On the two-phase framework for joint model and design-based inference. *Annals of Statistics*, 33, 2789-2810.

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.

Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). *Analysis of Complex Surveys*, New York: John Willey & Sons, Inc.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite population sampling and inference*: *A prediction approach*, New York: John Wiley & Sons, Inc.