

Linearization Variance Estimators for Survey Data

Abdellatif 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, (ii) validity under a conditional repeated sampling framework. In this paper, a new approach to deriving Taylor linearization variance estimators is proposed. It leads directly to a variance estimator which satisfies the above considerations at least in a number of important cases. The method is applied to a variety of problems, covering estimators of a total as well as other estimators defined either explicitly or implicitly as solutions of estimating equations. In particular, estimators of logistic regression parameters with calibration weights are studied. It leads to a new variance estimator for a general class of calibration estimators that includes generalized raking ratio and generalized regression estimators. The proposed method is extended to two-phase sampling to obtain a variance estimator that makes fuller use of the first phase sample data compared to traditional linearization variance estimators.

Key Words: Calibration; Design weights; Estimating equations; Raking ratio estimator; Regression estimators; Two-phase sampling.

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 $z_k = y_k - (\bar{y}/\bar{x})x_k$ and (n, N) denote the sample and population sizes. By linearizing the jackknife variance estimator, v_J , a different linearization variance estimator, $v_{JL} = (\bar{X}/\bar{x})^2 v_L$, is obtained. This variance estimator 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 ratio-adjusted 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 approximately model unbiased in the sense of $E_m(v_{JL}) \approx V_m(\hat{Y}_R)$, where E_m denotes model expectation and $V_m(\hat{Y}_R)$ is the model variance of \hat{Y}_R under a "ratio model": $E_m(y_k) = \beta x_k; k = 1, \dots, N$ and the y_k 's are independent with model variance $V_m(y_k) = \sigma^2 x_k, \sigma^2 > 0$. Thus, v_{JL} is a good choice from either the design-based or the model-based perspective.

Binder (1996) presented an elegant "cookbook" approach to Taylor linearization that leads directly to v_{JL} -type linearization variance estimators. He applied the method to smooth functions of estimated totals, $g(\hat{Y}_1, \dots, \hat{Y}_m)$, generalized regression estimators and the Wilcoxon rank sum statistic. To illustrate Binder's method, consider a ratio estimator

$$\hat{Y}_R = (\hat{Y}/\hat{X})X = \hat{R}X,$$

1. Abdellatif Demnati, Social Survey Methods Division, Statistics Canada, R.H. Coats Bldg, 15th Floor, Ottawa, Ontario, Canada, K1A 0T6; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6.

where $\hat{Y} = \sum_{k=1}^N d_k(s) y_k = \hat{Y}(y)$, $\hat{X} = \sum_{k=1}^N d_k(s) x_k = \hat{Y}(x)$ and the $d_k(s)$ are the design weights with $d_k(s) = 0$ if the population element k is not in the sample s , e.g., $d_k(s) = (1/\pi_k) a_k(s)$ where π_k is the probability of including the element k in the sample s , $a_k(s) = 1$ if $k \in s$, $a_k(s) = 0$ otherwise, and \sum denotes summation over the population elements. The weights are assumed to provide a design unbiased estimator \hat{Y} of Y , i.e., $E(d_k(s)) = 1$ for $k = 1, \dots, N$. Now take the total differential of \hat{Y}_R to get

$$(d\hat{Y}_R) = (d\hat{R})X = \frac{X}{\hat{X}} [(d\hat{Y}) - \hat{R}(d\hat{X})], \quad (1.1)$$

and replace all the total differentials in (1.1) by deviations of estimators from their respective population parameters, e.g., $d\hat{Y}_R$ is changed to $\hat{Y}_R - Y$. Then (1.1) yields

$$\hat{Y}_R - Y = \sum d_k(s) z_k - \frac{X}{\hat{X}} (Y - \hat{R}X), \quad (1.2)$$

where

$$z_k = \frac{X}{\hat{X}} (y_k - \hat{R}x_k). \quad (1.3)$$

The term $\sum d_k(s) z_k$ in (1.2) reduces to zero, but it is retained for variance estimation. On the other hand, the last term of (1.2) is ignored for variance estimation. Thus, $\hat{Y}_R - Y$ is represented as $\sum d_k(s) z_k = \hat{Y}(z)$ for the purpose of variance estimation. Denoting an unbiased variance estimator of $\hat{Y} = \hat{Y}(y)$ as $v(y)$, Binder's variance estimator of \hat{Y}_R is given by $v(z)$. The linearization variance estimator $v(z)$, obtained from (1.3), agrees with v_{JL} for simple random sampling and stratified multistage sampling if the sample is treated as if the primary sampling units are sampled with replacement. Note that the jackknife method is not applicable generally for any sampling design.

For the estimator $\hat{\theta} = g(\hat{Y}_1, \dots, \hat{Y}_m)$ of a smooth function to totals, $\theta = g(Y_1, \dots, Y_m)$, Binder's (1996) method leads to

$$\hat{\theta} - \theta = \sum d_k(s) z_k + \dots$$

with

$$z_k = \sum_{i=1}^m (\partial g(\mathbf{a}) / \partial a_i |_{\mathbf{a}=\hat{\mathbf{Y}}}) y_{ki}, \quad (1.4)$$

where $\hat{\mathbf{Y}} = (\hat{Y}_1, \dots, \hat{Y}_m)^T$ and $\mathbf{a} = (a_1, \dots, a_m)^T$. It follows from (1.4) that the partial derivatives, $\partial g(\mathbf{a}) / \partial a_i$, are evaluated at $\hat{\mathbf{Y}}$ to obtain z_k 's, whereas in the standard method (see e.g., Andersson and Nordberg 1994) they are evaluated at $\mathbf{Y} = (Y_1, \dots, Y_m)^T$ before getting z_k and then substituting estimates for the unknown components. For example, for the ratio estimator \hat{Y}_R the term X/\hat{X} disappears from z_k in the standard procedure because X/\hat{X} becomes 1 when \hat{X} is replaced by X .

Although Binder's (1996) approach is simple and attractive, a more rigorous and broadly applicable method is needed. In section 2, we propose an alternative approach that is theoretically justifiable and at the same time leads directly to a v_{JL} -type variance estimator for general designs. We apply the method, in section 3, 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, e.g., estimators of logistic regression parameters with design weights calibrated to known auxiliary population totals. We also obtain a new variance estimator for a general class of calibration estimators that includes generalized raking ratio and generalized regression estimators. Section 4 extends the proposed method to two-phase sampling to obtain a variance estimator that makes fuller use of the first phase sample data compared to traditional linearization variance estimators.

For the case of independent and identically (iid) random variables y_1, \dots, y_n with distribution function $F(y)$, estimation of general parameters $\theta = T(F)$ has been studied extensively in the literature (see e.g., Huber 1981). A natural estimator of $\theta = T(F)$ is $\hat{\theta} = T(\hat{F})$, where $\hat{F}(y)$ is the empirical distribution function given by $\hat{F}(y) = n^{-1} \sum_{k=1}^n I(y_k \leq y)$ with $I(y_k \leq y) = 1$ if $y_k \leq y$ and $I(y_k \leq y) = 0$ if $y_k > y$. For example, if $T(F)$ is the population mean $\int y dF(y)$, then $T(\hat{F}) = \int y d\hat{F}(y) = n^{-1} \sum_{k=1}^n y_k = \bar{y}$, the sample mean. Note that \hat{F} assigns equal mass, $1/n$ to each of the sample values y_1, \dots, y_n . If T is "sufficiently regular", then $T(\hat{F})$ may be linearized near F in terms of the influence curve (or function) of $T(\cdot)$ given by

$$IC(y, F, T) = \lim_{a \rightarrow 0} [T((1-a)F + a\delta_y) - T(F)] / a, \quad (1.5)$$

where δ_y denotes the point mass 1 at y . We have

$$\begin{aligned} \sqrt{n}[T(\hat{F}) - T(F)] &= \sqrt{n} \int IC(y, F, T) d\hat{F}(y) + \sqrt{n}R_n \\ &= \frac{1}{\sqrt{n}} \sum_{k=1}^n \tilde{z}_k + \sqrt{n}R_n \end{aligned} \quad (1.6)$$

where $\tilde{z}_k = IC(y_k, F, T)$ and $\sqrt{n}R_n$ is a remainder term. If $\sqrt{n}R_n$ is asymptotically negligible in the sense that $\sqrt{n}R_n$ converges in probability to zero as $n \rightarrow \infty$ (denoted $\sqrt{n}R_n \rightarrow_p 0$) then it follows from (1.6) that $\sqrt{n}[T(\hat{F}) - T(F)]$ is asymptotically normal with mean 0 and variance

$$A(F, T) = \int [IC(y, F, T)]^2 dF(y), \quad (1.7)$$

noting that the terms \tilde{z}_k in (1.6) are iid random variables. As noted by Huber (1981, page 13), $\sqrt{n}R_n$ is "often" asymptotically negligible, but the proof of this property may not be easy for general functionals $T(F)$. Serfling (1980, section 6.2) gave the following two conditions for

$\sqrt{n}R_n \rightarrow_p 0$, applicable for general random variables y_1, \dots, y_n (not necessarily iid): (i) $T(\cdot)$ is “stochastically differentiable” at F ; (ii) $\sqrt{n} \sup | \hat{F}(y) - F(y) |$ is bounded in probability, where \sup is over y . Condition (ii) is satisfied in the iid case, but it may not be easy to prove (ii) for complex sampling designs. Condition (i) means that there exists a functional $T(F; F_n - F)$ such that $T(F_n) - T(F) = n^{-1} \sum_{k=1}^n T(F; \delta_{y_k} - F) + R_n$, where R_n is of lower order in probability than $\sup | F_n(y) - F(y) |$ as the latter tends to zero. This condition may not be easy to verify for general $T(\cdot)$. Serfling (1980) suggested that in practice it is more effective to analyse R_n directly using “the method of differential inequalities”.

A natural estimator of the asymptotic variance $A(F, T)$ is

$$A(\hat{F}, T) = \frac{1}{n} \sum_{k=1}^n [\text{IC}(y_k, \hat{F}, T)]^2, \quad (1.8)$$

where $\text{IC}(y, \hat{F}, T)$ is the influence curve evaluated at $F = \hat{F}$. It follows that a linearization variance estimator of $T(\hat{F})$ is

$$V_L [T(\hat{F})] = A(\hat{F}, T) / n. \quad (1.9)$$

Practical implementation of $v_L [T(\hat{F})]$ involves the computation of $\text{IC}(y_k, \hat{F}, T)$ for each specified T . The latter can be avoided by using the jackknife method. Substituting \hat{F} for F and $-1/(n-1)$ for a in (1.5), we obtain a jackknife estimator of $\text{IC}(y_k, F, T)$ as $z_{kj} = (n-1) [T(\hat{F}) - T(\hat{F}_{-k})]$, where $\hat{F}_{-k}(y)$ is the empirical distribution function obtained when y_k is omitted. The resulting jackknife variance estimator $T(\hat{F})$ is

$$\begin{aligned} v_J [T(\hat{F})] &= \frac{1}{n(n-1)} \sum_{k=1}^n z_{kj}^2 \\ &= \frac{n-1}{n} \sum_{k=1}^n [T(\hat{F}_{-k}) - T(\hat{F})]^2; \end{aligned} \quad (1.10)$$

see e.g., Hampel, Ronchetti, Rousseeuw and Stahel (1986, page 95). If $\text{IC}(y, F, T)$ does not depend smoothly on F , then the jackknife variance estimator may not be consistent for the variance of $T(\hat{F})$; for example, when $T(\hat{F})$ is the sample median.

Campbell (1980) attempted to extend the above results for the iid case to general sampling designs, using the design weights $d_k(s)$. The population (or census) parameter θ is now given by $\theta = T(F_N)$, where $F_N(y)$ is the population distribution function that assigns equal mass, $1/N$, to each of the N population values y_1, \dots, y_N . An empirical distribution function is given by $\hat{F}(y) = \sum_{k \in s} \tilde{d}_k(s) I(y_k \leq y)$, where $\tilde{d}_k(s) = d_k(s) / \sum_{l \in s} d_l(s)$ are the normalized design weights. Note that $\hat{F}(y)$ assigns the mass $\tilde{d}_k(s)$ to the element $k \in s$. An estimator of

$\theta = T(F_N)$ is given by $\hat{\theta} = T(\hat{F})$. For example, if $T(F_N)$ is the population mean $\int y dF_N(y)$, then $T(\hat{F}) = \int y d\hat{F}(y) = \sum_{k \in s} d_k(s) y_k / \sum_{k \in s} d_k(s)$, the design-weighted sample mean. Campbell (1980) followed the linearization (1.6) for the iid case and concluded that $\sqrt{n} [T(\hat{F}) - T(F_N)]$ is asymptotically normal with mean 0 and variance

$$\begin{aligned} A(F_N, T) &= n \text{Var} \left[\sum_{k \in s} d_k(s) \tilde{z}_k / \sum_{k \in s} d_k(s) \right] \\ &\approx n \text{Var} \left[\sum_{k \in s} d_k(s) \{ (\tilde{z}_k - R) / N \} \right], \end{aligned} \quad (1.11)$$

using the approximate variance of a ratio, where $R = \sum_{k \in s} \tilde{z}_k / N$ is the population mean of \tilde{z}_k 's and $\tilde{z}_k = \text{IC}(y_k, F_N, T)$. Denoting the unbiased variance estimator of $\hat{Y} = \hat{Y}(y) = \sum_{k \in s} d_k(s) y_k$ as $v(y)$, it follows from (1.11) that a linearization variance estimator of $T(\hat{F})$ is given by

$$v_L [T(\hat{F})] = v[(z - \hat{R}) / \hat{N}], \quad (1.12)$$

where

$$z_k = \text{IC}(y_k, \hat{F}, T), \quad (1.13)$$

and

$$\hat{R} = \sum_{k \in s} d_k(s) z_k / \sum_{k \in s} d_k(s). \quad (1.14)$$

To avoid the computation of z_k 's, Campbell (1980) proposed a jackknife estimator of \tilde{z}_k for each $k \in s$. It is given by

$$z_{kj} = \frac{1 - \tilde{d}_k(s)}{\tilde{d}_k(s)} [T(\hat{F}) - T(\hat{F}_{-k})], \quad (1.15)$$

where

$$d\hat{F}_{-k}(y) = \begin{cases} \frac{d\hat{F}(y) - \tilde{d}_k(s)}{1 - \tilde{d}_k(s)} & \text{if } y = y_k \\ \frac{d\hat{F}(y)}{1 - \tilde{d}_k(s)} & \text{if } y \neq y_k. \end{cases} \quad (1.16)$$

The resulting linearization variance estimator is given by $v[(z_j - \hat{R}_j) / \hat{N}]$. Note that the proposed jackknife method is different from the customary jackknife for survey sampling. For example, for stratified multistage sampling, the customary jackknife deletes sample clusters in turn whereas the Campbell method deletes elements in turn. Also, the customary jackknife is not always applicable (e.g., unequal probability sampling without replacement) unlike the Campbell method which uses the unbiased variance estimator $v(y)$ of the total \hat{Y} for the given design and then replaces y by $(z_j - \hat{R}_j) / \hat{N}$. However, the computations involved in the Campbell method can be very heavy because it requires the computation of $T(\hat{F}_{-k})$ for each element $k \in s$; in large-scale surveys the number of sample

elements can be very large, as in the Canadian Labour Force Survey.

Deville (1999) and Berger (2002) obtained results very similar to those of Campbell (1980). Instead of using the natural probability measure \hat{F} , they considered functionals of the form $T(\hat{M})$, where \hat{M} denotes a measure that allocates the design weight $d_k(s)$ to any point y_k for k in s and zero to units k not in s . 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 of the N points y_k in the finite population U . Suppose that $T(\cdot)$ is of degree α in the sense that $N^{-\alpha} T(\cdot)$ tends to a limit for some $\alpha \geq 0$. Typically, $\alpha = 0$ or 1 ; for example, $\alpha = 1$ if $T(M)$ is the total Y and $\alpha = 0$ if $T(M)$ is the ratio $R = Y/X$. Deville (1999) used the following asymptotic approximation:

$$\sqrt{n} N^{-\alpha} [T(\hat{M}) - T(M)] \approx \frac{\sqrt{n}}{N} \sum (d_k(s) - 1) \tilde{z}_k, \quad (1.17)$$

where $d_k(s) = 0$ if k is not in the sample s . Further $\tilde{z}_k = IT(M; y_k)$ with IT denoting the influence function of $T(M)$ defined by

$$IT(M; y) = \lim_{t \rightarrow 0} \frac{1}{t} [T(M + t\delta_y) - T(M)]. \quad (1.18)$$

As noted earlier, it is not easy to justify the approximation (1.17) for general functionals $T(\cdot)$. Deville (1999) developed rules for evaluating $IT(M; y)$ for selected functionals $T(\hat{M})$. Berger (2002) used the jackknife method to estimate $\tilde{z}_k = IT(M, y_k)$, similar to Campbell (1980).

Noting that $\sum d_k(s) \tilde{z}_k = \hat{Y}(\tilde{z})$ it follows from (1.17) that a linearization variance estimator of $N^{-\alpha} T(\hat{M})$ is given by $N^{-2} v(\tilde{z})$. But \tilde{z}_k depends on unknown parameters and the corresponding estimator, z_k , may not be unique. For example, suppose $T(\hat{M}) = \hat{Y}_R = (\hat{Y}/\hat{X})X$, then $\alpha = 1$ and $\tilde{z}_k = y_k - R x_k$, where $R = Y/X$. In this case, two possible candidates for z_k are $z_k = y_k - \hat{R} x_k$ and $z_k = (X/\hat{X})(y_k - \hat{R} x_k)$. Thus, the choice of z_k in the presence of auxiliary information, such as a known total X , is not unique under Deville's approach. Unlike Deville's approach, our method leads to a unique choice z_k and it avoids the calculation of \tilde{z}_k to determine z_k . Our z_k satisfies desirable properties mentioned section 1, at least in a number of important cases.

2. The Method

To motivate the method, we start with a simple general case where the estimator $\hat{\theta}$ of a parameter θ can be expressed as a smooth function $g(\hat{Y})$ of estimated totals $\hat{Y} = (\hat{Y}_1, \dots, \hat{Y}_i, \dots, \hat{Y}_m)^T$, where $Y_i = \sum_{k \in U} d_k(s) y_{ik}$, $i = 1, \dots, m$, is an estimator of the total $Y_i = \sum_{k \in U} y_{ik}$, and

$\theta = g(\mathbf{Y})$ with $\mathbf{Y} = (Y_1, \dots, Y_i, \dots, Y_m)^T$. We may write $\hat{\theta}$ as $\hat{\theta} = f(\mathbf{d}(s), \mathbf{A}_y)$ and $\theta = f(\mathbf{1}, \mathbf{A}_y)$, where \mathbf{A}_y is an $m \times N$ matrix with k^{th} column $\mathbf{y}_k = (y_{k1}, \dots, y_{ki}, \dots, y_{km})^T$, $k = 1, \dots, N$, $\mathbf{d}(s) = (d_1(s), \dots, d_N(s))^T$ and $\mathbf{1}$ is the N -vector of 1's. For example, if $\hat{\theta}$ denotes the ratio estimator $\hat{Y}_R = [(\sum d_k(s) y_k) / (\sum d_k(s) x_k)] X$, then $m = 2$, $y_{1k} = y_k$, $y_{2k} = x_k$ and $f(\mathbf{1}, \mathbf{A}_y)$ reduces to the total Y , noting that $(Y/X)X = Y$. Note that \hat{Y}_R is a function of $\mathbf{d}(s)$, \mathbf{y} and \mathbf{x} and the known total X , but we dropped X for simplicity and write $\hat{Y}_R = f(\mathbf{d}(s), \mathbf{y}, \mathbf{x})$.

Taylor linearization of $\hat{\theta}$ around \mathbf{Y} gives the approximation

$$\sqrt{n} N^{-\alpha} (\hat{\theta} - \theta) \approx \frac{\sqrt{n}}{N} (\partial g(\mathbf{a}) / \partial \mathbf{a})^T \Big|_{\mathbf{a}=\mathbf{Y}} (\hat{\mathbf{Y}} - \mathbf{Y}) \quad (2.1)$$

where $\partial g(\mathbf{a}) / \partial \mathbf{a} = (\partial g(\mathbf{a}) / \partial a_1, \dots, \partial g(\mathbf{a}) / \partial a_m)^T$ and $N^{-\alpha} g(\cdot)$ tends to a limit for some $\alpha \geq 0$. Asymptotic normality of $\sqrt{n} N^{-\alpha} (\hat{\theta} - \theta)$ follows from (2.1), provided a central limit theorem for $\sqrt{n} N^{-1} (\hat{\mathbf{Y}} - \mathbf{Y})$ holds and $g(\cdot)$ has continuous first derivatives in a neighbourhood of the mean \bar{Y} . Krewski and Rao (1981) justified (2.1) for stratified sampling.

Let $\check{\mathbf{Y}} = \sum b_k \mathbf{y}_k$ for arbitrary real numbers $\mathbf{b} = (b_1, \dots, b_N)^T$, and $g(\check{\mathbf{Y}}) = f(\mathbf{b}, \mathbf{A}_y) = f(\mathbf{b})$. Noting that $\check{\mathbf{Y}} = \mathbf{A}_y \mathbf{d}(s)$ and $\mathbf{Y} = \mathbf{A}_y \mathbf{1}$, we can express (2.1) as

$$\begin{aligned} & \sqrt{n} N^{-\alpha} (\hat{\theta} - \theta) \\ & \approx \frac{\sqrt{n}}{N} (\partial g(\check{\mathbf{Y}}) / \partial \check{\mathbf{Y}})^T \Big|_{\check{\mathbf{Y}}=\mathbf{Y}} \mathbf{A}_y (\mathbf{d}(s) - \mathbf{1}) \\ & = \frac{\sqrt{n}}{N} \sum_{k=1}^N (\partial f(\mathbf{b}) / \partial \check{\mathbf{Y}})^T \Big|_{b=1} \mathbf{y}_k (d_k(s) - 1), \quad (2.2) \end{aligned}$$

noting that $\check{\mathbf{Y}} = \mathbf{Y}$ is equivalent to $\mathbf{b} = \mathbf{1}$. Now we substitute $\mathbf{y}_k = \partial \check{\mathbf{Y}} / \partial b_k \Big|_{b=1}$ in (2.2) to get

$$\begin{aligned} & \sqrt{n} N^{-\alpha} (\hat{\theta} - \theta) \\ & \approx \frac{\sqrt{n}}{N} \sum_{k=1}^N (\partial f(\mathbf{b}) / \partial b_k) \Big|_{b=1} (d_k(s) - 1) \\ & = \frac{\sqrt{n}}{N} \check{\mathbf{z}}^T (\mathbf{d}(s) - \mathbf{1}), \quad (2.3) \end{aligned}$$

where $\check{\mathbf{z}} = (\check{z}_1, \dots, \check{z}_N)^T$ with $\check{z}_k = \partial f(\mathbf{b}) / \partial b_k \Big|_{b=1}$.

A variance estimator of the right hand side of (2.3) is given by $(n/N^2) v(\check{\mathbf{z}})$, where $v(\check{\mathbf{z}})$ is the variance estimator of the estimated total $\sum d_k(s) \check{z}_k = \hat{Y}(\check{\mathbf{z}})$. Since \check{z}_k 's are unknown, we replace \check{z}_k by $z_k = \partial f(\mathbf{b}) / \partial b_k \Big|_{b=d(s)}$, to get $(n/N^2) v(\mathbf{z})$. Thus, a linearization variance estimator of $\hat{\theta}$ is given by

$$v_L(\hat{\theta}) = (N^{2\alpha} / N^2) v(\mathbf{z}), \quad (2.4)$$

which reduces to $v(\mathbf{z})$ if $\alpha = 1$. Note that $v_L(\hat{\theta})$ given by (2.4) is simply obtained from the formula $v(y)$ for \hat{Y} by replacing y_k by z_k for $k \in s$. Note that we do not first

evaluate the partial derivatives $\partial f(\mathbf{b})/\partial b_k$ at $\mathbf{b} = \mathbf{1}$ to get \tilde{z} and then substitute estimates for the unknown components of \tilde{z} . Our method, therefore, is similar in spirit to Binder's approach. The variance estimator $v_L(\hat{\theta})$ is valid because z_k is a consistent estimator of \tilde{z}_k .

Example 2.1 Suppose $\hat{\theta}$ is the ratio estimator $\hat{Y}_R = X[(\sum d_k(s)y_k)/(\sum d_k(s)x_k)]$ of the total Y . Then $f(\mathbf{b}) = X[(\sum b_k y_k)/(\sum b_k x_k)]$ and

$$\partial f(\mathbf{b})/\partial b_k = X \frac{y_k \sum b_k x_k - x_k \sum b_k y_k}{(\sum b_k x_k)^2}.$$

Therefore,

$$z_k = \partial f(\mathbf{b})/\partial b_k |_{\mathbf{b}=\mathbf{d}(s)} = \frac{X}{\hat{X}}(y_k - \hat{R}x_k)$$

which agrees with (1.3). Thus, our variance estimator $v_L(\hat{Y}_R)$ is identical to Binder's (1996) variance estimator, $v(z)$, noting that $\alpha = 1$.

Our derivation is simple and natural. On the other hand, in the standard linearization method, $\hat{\theta}$ is first expressed in terms of elementary components $\hat{Y}_1, \dots, \hat{Y}_m$ as $g(\hat{\mathbf{Y}})$ and the partial derivatives $\partial g(\mathbf{a})/\partial a_j$ are then evaluated at $\mathbf{a} = \mathbf{Y}$. It is interesting to note that all the components of $\hat{\mathbf{Y}}$ use the same weights $d_k(s)$ and our approach always takes first derivatives of $f(\mathbf{b})$ with respect to b_k at $\mathbf{b} = \mathbf{d}(s)$. It is not necessary to first express $\hat{\theta}$ in terms of elementary components.

3. Calibration Estimators

The ratio estimator can be viewed as a calibration estimator, $\hat{Y}_R = \sum w_k(s)y_k$, with explicit weights $w_k(s) = (X/\hat{X})d_k(s)$ and satisfying the calibration constraint $\sum w_k(s)x_k = X$. Calibration estimators of a total Y of the form $\hat{Y}_w = \sum w_k(s)y_k$ with explicit weights $w_k(s)$ and satisfying the calibration constraints $\sum w_k(s)\mathbf{x}_k = \mathbf{X}$ are widely used, where $\mathbf{x}_k = (x_{1k}, \dots, x_{qk})^T$ and $\mathbf{X} = (X_1, \dots, X_q)^T$ is the vector of known totals of auxiliary variables $x_j, j = 1, \dots, q$. In subsection 3.1 we consider the generalized regression (GREG) estimator and then study a general class of regression calibration estimators in subsection 3.2. Extension to estimators, $\hat{\theta}$, obtained as solutions of estimating equations is presented in subsection 3.3. The case of general calibration estimators is investigated in subsection 3.4.

3.1 Generalized Regression Estimator

The GREG estimator of total Y is given by \hat{Y}_w with calibration weights $w_k(s) = d_k(s)g_k(\mathbf{d}(s))$, where

$$g_k(\mathbf{d}(s)) = 1 + (\mathbf{X} - \hat{\mathbf{X}})^T \left(\sum d_k(s)c_k \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} c_k \mathbf{x}_k \quad (3.1)$$

with specified constants c_k and $\hat{\mathbf{X}} = \sum d_k(s)\mathbf{x}_k$ (cf., Särndal *et al.* 1989). The ratio estimator, \hat{Y}_R , is a special case with $q=1$ (i.e., scalar x_k) and $c_k = x_k^{-1}$, and $g_k(\mathbf{d}(s))$, is given by (3.1), reduces to X/\hat{X} .

The GREG estimator may be expressed as a differentiable function of estimated totals. Hence, the general theory of section 2 is applicable and it remains to evaluate $z_k = \partial f(\mathbf{b})/\partial b_k |_{\mathbf{b}=\mathbf{d}(s)}$, where $f(\mathbf{b}) = \sum (b_k g_k(\mathbf{b}))y_k$ is obtained by replacing $\mathbf{d}(s)$ by \mathbf{b} in the formula for \hat{Y}_w . Noting that $\partial \mathbf{A}(\mathbf{b})^{-1}/\partial b_k = -\mathbf{A}(\mathbf{b})^{-1}(\partial \mathbf{A}(\mathbf{b})/\partial b_k)\mathbf{A}(\mathbf{b})^{-1}$, where $\mathbf{A}(\mathbf{b}) = \sum b_k c_k \mathbf{x}_k \mathbf{x}_k^T$, we get

$$\begin{aligned} \partial (b_k g_k(\mathbf{b}))/\partial b_k &= g_k(\mathbf{b}) - \mathbf{x}_k^T \mathbf{A}(\mathbf{b})^{-1} b_k c_k \mathbf{x}_k \\ &\quad - (\mathbf{X} - \hat{\mathbf{X}}(\mathbf{b}))^T \mathbf{A}(\mathbf{b})^{-1} (c_k \mathbf{x}_k \mathbf{x}_k^T) \mathbf{A}(\mathbf{b})^{-1} (b_k c_k \mathbf{x}_k) \end{aligned} \quad (3.2)$$

and for $l \neq k$

$$\begin{aligned} \partial (b_l g_l(\mathbf{b}))/\partial b_k &= -\mathbf{x}_k^T \mathbf{A}(\mathbf{b})^{-1} (b_l c_l \mathbf{x}_l) \\ &\quad - (\mathbf{X} - \hat{\mathbf{X}}(\mathbf{b}))^T \mathbf{A}(\mathbf{b})^{-1} (c_k \mathbf{x}_k \mathbf{x}_k^T) \mathbf{A}(\mathbf{b})^{-1} (b_l c_l \mathbf{x}_l). \end{aligned} \quad (3.3)$$

It now follows from (3.2) and (3.3), that

$$\partial f(\mathbf{b})/\partial b_k = g_k(\mathbf{b})e_k(\mathbf{b}), \quad (3.4)$$

where

$$e_k(\mathbf{b}) = y_k - \mathbf{x}_k^T \mathbf{B}(\mathbf{b}) \quad (3.5)$$

with $\mathbf{B}(\mathbf{b}) = \mathbf{A}^{-1}(\mathbf{b})(\sum_k b_k c_k \mathbf{x}_k y_k)$. Therefore, $z_k = \partial f(\mathbf{b})/\partial b_k |_{\mathbf{b}=\mathbf{d}(s)}$ reduces to

$$z_k = g_k(\mathbf{d}(s))e_k, \quad (3.6)$$

where $e_k = y_k - \mathbf{x}_k^T \hat{\mathbf{B}}$ with $\hat{\mathbf{B}} = \mathbf{B}(\mathbf{d}(s))$.

The variance estimator of \hat{Y}_w , resulting from (3.6), namely $v(z)$, takes account of the g -weights, $g_k(\mathbf{d}(s))$, unlike the standard linearization variance estimator (see e.g., Särndal *et al.* 1991, page 237). It agrees with the model-assisted variance estimator of Särndal *et al.* (1989). It also agrees with the jackknife linearization variance estimator when the latter is applicable (Yung and Rao 1996).

3.2 A General Class of Regression Calibration Weights

We now turn to a general class of regression calibration weights of the form $w_k(s) = d_k(s)h_k(\mathbf{d}(s))$ with

$$h_k(\mathbf{d}(s)) = 1$$

$$+ (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{Q}}^{-1} (c_k \mathbf{x}_k + \sum_{l \neq k} d_l(s)c_{kl} \mathbf{x}_l), \quad (3.7)$$

where the ab th element of $\hat{\mathbf{Q}}$ is given by

$$\begin{aligned} \hat{q}_{ab} &= \sum_{k=1}^N d_k(s)c_k x_{ak} x_{bk} \\ &\quad + \sum_{k=1}^N \sum_{l \neq k}^N d_k(s)d_l(s)c_{kl} x_{ak} x_{bl} \end{aligned}$$

for specified constants c_k and $c_{kl} (= c_{lk})$. The class (3.7) covers the GREG estimator as well as the “optimal” linear regression estimator with $d_k(s) = (1/\pi_k) a_k(s)$. In the former case $c_{kl} = 0$ while the optimal linear regression estimator uses $c_k = (1 - \pi_k)/\pi_k$ and $c_{kl} = (\pi_{kl} - \pi_k \pi_l)/\pi_{kl}$, $k \neq l$, where π_{kl} is the probability of including both elements k and l in the sample s (Montanari 1998).

The calibration weights $w_k(s)$ may be written as

$$w_k(s) = d_k(s) + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{Q}}^{-1} (d_k(s) c_k \mathbf{x}_k + \sum_{l \neq k} d_{kl}(s) c_{kl}^* \mathbf{x}_l), \quad (3.8)$$

where $d_{kl}(s) = d_k(s) d_l(s) / E[d_k(s) d_l(s)]$, $c_{kl}^* = c_{kl} E[d_k(s) d_l(s)]$ and

$$\hat{q}_{ab} = \sum_{k=1}^N d_k(s) c_k x_{ak} x_{bk} + \sum_{k=1}^N \sum_{l \neq k}^N d_{kl}(s) c_{kl}^* x_{ak} x_{bl}.$$

Note that $Ed_k(s) = 1$ and $Ed_{kl}(s) = 1$. If $d_k(s) = (1/\pi_k) a_k(s)$ then $d_{kl}(s)$ reduces to $d_{kl}(s) = a_k(s) a_l(s) / \pi_{kl}$ and $c_{kl}^* = (\pi_{kl} - \pi_k \pi_l) / (\pi_k \pi_l)$. We can regard the calibration estimator \hat{Y}_w resulting from (3.8) as a function of totals, by expressing a quadratic form as a total of synthetic variables (Sitter and Wu 2002). Therefore, we can use the method of section 2 and write $\hat{Y}_w = f(\mathbf{d}^{(1)}(s), \mathbf{d}^{(2)}(s), \mathbf{y}) = \sum d_k(s) h(\mathbf{d}^{(1)}(s), \mathbf{d}^{(2)}(s)) y_k$ where $\mathbf{d}^{(1)}(s) = \mathbf{d}(s)$ and $\mathbf{d}^{(2)}(s)$ is the vector of elements $d_{kl}(s)$, $k < l$, arranged in a sequence. Now, following the derivation of (2.3), we get

$$\hat{Y}_w - Y \approx \sum_k \tilde{z}_k (d_k(s) - 1) + 2 \sum_{k < l} \tilde{z}_{kl} (d_{kl}(s) - 1) \quad (3.9)$$

where

$$\tilde{z}_k = \partial f(\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \mathbf{y}) / \partial b_k |_{b^{(1)}=1, b^{(2)}=1},$$

$$\tilde{z}_{kl} = \partial f(\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \mathbf{y}) / \partial b_{kl} |_{b^{(1)}=1, b^{(2)}=1},$$

$\mathbf{b}^{(1)} = \mathbf{b} = (b_1, \dots, b_N)^T$ and $\mathbf{b}^{(2)}$ is the vector of arbitrary real numbers b_{kl} , $k < l$, arranged in the same order as the elements $d_{kl}(s)$ in $\mathbf{d}^{(2)}(s)$. Using (3.9), a variance estimator of \hat{Y}_w is approximately given by the variance estimator of $\sum_k \tilde{z}_k d_k(s) + 2 \sum_{k < l} \tilde{z}_{kl} d_{kl}(s)$, denoted by $v(\tilde{\mathbf{z}}^{(1)}, \tilde{\mathbf{z}}^{(2)})$.

Since $v(\tilde{\mathbf{z}}^{(1)}, \tilde{\mathbf{z}}^{(2)})$ involves the unknown values \tilde{z}_k and \tilde{z}_{kl} , we replace \tilde{z}_k by $z_k = \partial f(\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \mathbf{y}) / \partial b_k |_{b^{(1)}=d^{(1)}(s), b^{(2)}=d^{(2)}(s)}$ and \tilde{z}_{kl} by $z_{kl} = \partial f(\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \mathbf{y}) / \partial b_{kl} |_{b^{(1)}=d^{(1)}(s), b^{(2)}=d^{(2)}(s)}$ to get $v(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$. Unfortunately the variance estimator $v(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$ involves third order and fourth order moments $E[d_k(s) d_l(s) d_q(s)]$ and $E[d_k(s) d_l(s) d_q(s) d_r(s)]$ in addition to the second moments $E[d_k(s) d_l(s)]$, whereas the variance estimator for the generalized regression estimator requires only the second moments. In particular, if $d_k(s) = (1/\pi_k) a_k(s)$ we required third and fourth order inclusion probabilities π_{klq} and π_{klqr} as well as the second order inclusion probabilities π_{kl} .

The calculation of z_k and z_{kl} involves the derivatives $\partial [b_l h(\mathbf{b}^{(1)}, \mathbf{b}^{(2)})] / \partial b_k$ for $l = k$ and $l \neq k$ and the derivatives $\partial [b_l h(\mathbf{b}^{(1)}, \mathbf{b}^{(2)})] / \partial b_{kl}$ for $l = k$ and $l \neq k$. After simplification, we get

$$z_k = [1 + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{Q}}^{-1} c_k \mathbf{x}_k] e_k^*$$

and

$$z_{kl} = (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{Q}}^{-1} c_{kl}^* \mathbf{x}_l e_k^*,$$

where

$$e_k^* = y_k - \mathbf{x}_k^T \hat{\mathbf{B}}^*$$

with

$$\hat{\mathbf{B}}^* = \hat{\mathbf{Q}}^{-1} \left(\sum_k d_k(s) c_k \mathbf{x}_k y_k + \sum_{k \neq l} d_{kl}(s) c_{kl}^* \mathbf{x}_l y_k \right).$$

Note that the customary Taylor linearization variance estimation uses $v(e^*)$, while $v(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$ would involve the residuals e_k^* as well as the g -weights $1 + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{Q}}^{-1} c_k \mathbf{x}_k$ and $(\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{Q}}^{-1} c_{kl}^* \mathbf{x}_l$. If $c_{kl} = 0$ for all $k \neq l$, then $z_{kl} = 0$ and $v(\mathbf{z}^{(1)}, \mathbf{z}^{(2)})$ reduces to $v(\mathbf{z})$ with z_k given by (3.6). Thus the GREG result of subsection 3.1 is a special case.

3.3 Estimating Equations

We now turn to a vector parameter $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$ defined either explicitly or implicitly as the solution to “census” estimating equations $\mathbf{S}(\boldsymbol{\theta}) = \sum_{k=1}^N \mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{0}$. A calibration estimator $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_p)^T$ with GREG calibration weights $w_k(s) = d_k(s) g_k(\mathbf{d}(s))$ is obtained as the solution to sample estimating equations:

$$\hat{\mathbf{S}}(\hat{\boldsymbol{\theta}}) = \sum w_k(s) \mathbf{u}_k(\hat{\boldsymbol{\theta}}) = \mathbf{0}, \quad (3.10)$$

where $\mathbf{u}_k(\hat{\boldsymbol{\theta}})$ and $\hat{\mathbf{S}}(\hat{\boldsymbol{\theta}})$ are $(p \times 1)$ vectors (Binder 1983). For example for logistic regression with scalar θ , we have $u_k(\boldsymbol{\theta}) = (y_k - p_k(\boldsymbol{\theta})) a_k$, where $p_k(\boldsymbol{\theta}) = P(y_k = 1 | a_k) = \exp(\theta a_k) / (1 + \exp(\theta a_k))$ and a_k is the predictor variable. Note that $\hat{\boldsymbol{\theta}}$, in this case, is the implicit solution to (3.10) and obtained iteratively using Newton-Raphson or Fisher scoring method.

The estimator of a ratio of totals Y and $A = \sum a_k$ is obtained as the explicit solution of (3.10) with $u_k(\boldsymbol{\theta}) = y_k - \theta a_k$: $\hat{\boldsymbol{\theta}} = \sum w_k(s) y_k / \sum w_k(s) a_k = \hat{Y} / \hat{A}$. In this case, $\hat{\boldsymbol{\theta}}$ is a function of estimated totals and hence our method for functions of totals is applicable. It remains to evaluate $\partial f(\mathbf{b}) / \partial b_k$, where $f(\mathbf{b}) = \sum b_k g_k(\mathbf{b}) y_k / \sum b_k g_k(\mathbf{b}) a_k$. We have

$$\partial f(\mathbf{b}) / \partial b_k = \sum_{l=1}^N [\partial (b_l g_l(\mathbf{b})) / \partial b_k] \hat{\mathbf{A}}(\mathbf{b})^{-1} (y_l - f(\mathbf{b}) a_l),$$

where $\hat{\mathbf{A}}(\mathbf{b}) = \sum b_l g_l(\mathbf{b}) a_l$. Now using (3.4) and (3.5), it is easy to verify that z_k reduces to

$$z_k = g_k(\mathbf{d}(s)) \hat{\mathbf{A}}^{-1} e_k^*$$

where

$$e_k^* = u_k(\hat{\boldsymbol{\theta}}) - \mathbf{x}_k^T \hat{\mathbf{B}}_{\boldsymbol{\theta}}$$

with $\hat{\mathbf{B}}_u$ obtained from $\hat{\mathbf{B}}$ by changing y_k to $u_k(\hat{\boldsymbol{\theta}})$. Note that the residuals e_k^* has the same form as the GREG residuals e_k with y_k changed with $u_k(\hat{\boldsymbol{\theta}})$.

In general, the solution $\hat{\boldsymbol{\theta}}$ to the estimating equations (3.10) may not be expressible as a function of estimated totals. We therefore follow Binder's (1983) approach and write the linearization estimator of the covariance matrix of $\hat{\boldsymbol{\theta}}$ as

$$\mathbf{v}_L(\hat{\boldsymbol{\theta}}) = [\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})]^{-1} \hat{\boldsymbol{\Sigma}}_S(\hat{\boldsymbol{\theta}}) [\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})]^{-1}, \quad (3.11)$$

where $\hat{\mathbf{J}}(\boldsymbol{\theta}) = -\partial \hat{\mathbf{S}}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ and $\hat{\boldsymbol{\Sigma}}_S(\hat{\boldsymbol{\theta}})$ is the estimated covariance matrix $\mathbf{v}_L(\hat{\mathbf{S}}(\boldsymbol{\theta})) = \hat{\boldsymbol{\Sigma}}_S(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Binder (1983) gave regularity conditions for the validity of (3.11). Noting that $\hat{\mathbf{S}}(\boldsymbol{\theta})$ is a vector of estimated totals with GREG weights $d_k(s)g_k(\mathbf{d}(s))$, it follows from (3.6) and (3.11) that

$$\mathbf{v}_L(\hat{\boldsymbol{\theta}}) = \mathbf{v}(z) \quad (3.12)$$

where

$$z_k = [\hat{\mathbf{J}}(\hat{\boldsymbol{\theta}})]^{-1} g_k(\mathbf{d}(s)) e_k^* \quad (3.13)$$

with $e_k^* = (e_{k1}^*, \dots, e_{kp}^*)^T$ and

$$e_{kj}^* = u_{jk}(\hat{\boldsymbol{\theta}}) - \mathbf{x}_k \hat{\mathbf{B}}_{ju}; j = 1, \dots, p.$$

Further, $\hat{\mathbf{B}}_{ju}$ is obtained from $\hat{\mathbf{B}}_j$ by changing y_k to $u_{jk}(\hat{\boldsymbol{\theta}})$ and $\mathbf{v}(z)$ is the estimated covariance matrix of the vector of estimated totals $\hat{\mathbf{Z}} = \sum d_k(s)z_k$, where $u_{jk}(\hat{\boldsymbol{\theta}})$ is the j^{th} element of $u_k(\hat{\boldsymbol{\theta}})$. The result (3.12) agrees with the jackknife linearization variance estimator, v_{JL} , for stratified multistage sampling obtained by Rao, Yung and Hidiroglou (2002).

The result (3.12)–(3.13) may also be obtained directly by writing $\hat{\boldsymbol{\theta}}$ as $f(\mathbf{d}(s))$ and evaluating $z_k = \partial f(\mathbf{b}) / \partial b_k |_{b=d(s)}$. We denote $\hat{\boldsymbol{\theta}}(\mathbf{b}) = f(\mathbf{b})$ as the solution of $\sum (b_k g_k(\mathbf{b})) \mathbf{u}_k(\hat{\boldsymbol{\theta}}(\mathbf{b})) = \mathbf{0}$, i.e.,

$$\sum (b_k g_k(\mathbf{b})) \mathbf{u}_k(\hat{\boldsymbol{\theta}}(\mathbf{b})) = \mathbf{0}, \quad (3.14)$$

We now take the derivative of (3.14) with respect to b_k to get

$$\sum_{l=1}^N [\partial (b_l g_l(\mathbf{b})) / \partial b_k] \mathbf{u}_l(\hat{\boldsymbol{\theta}}(\mathbf{b})) + \sum_{l=1}^N (b_l g_l(\mathbf{b})) [\partial \mathbf{u}_l(\hat{\boldsymbol{\theta}}(\mathbf{b})) / \partial (\hat{\boldsymbol{\theta}}(\mathbf{b}))] \partial (\hat{\boldsymbol{\theta}}(\mathbf{b})) / \partial b_k. \quad (3.15)$$

Substituting (3.2) and (3.3) for $\partial (b_l g_l(\mathbf{b})) / \partial b_k$ in (3.15), we obtain (3.13) after simplification. This result shows that our method is also directly applicable to general estimators $\hat{\boldsymbol{\theta}}$ under Binder's (1983) regularity conditions.

3.4 A General Class of Calibration Estimators

The calibration weights, $w_k(s)$, associated with the GREG estimator \hat{Y}_w may not be always nonnegative. To get around this difficulty, generalized raking ratio weights are

often used. These weights are always nonnegative, but the method can lead to some extreme weights (Deville and Särndal 1992).

The generalized raking weights belong to the class

$$w_k(s) = d_k(s) F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}) \quad (3.16)$$

with $F(a) = e^a$, where the LaGrange multiplier $\hat{\boldsymbol{\lambda}}$ is determined by solving the calibration equations

$$\sum w_k(s) \mathbf{x}_k = \sum d_k(s) F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}) \mathbf{x}_k = \mathbf{X}. \quad (3.17)$$

The GREG weights correspond to $F(a) = 1 + a$ in which case $\hat{\boldsymbol{\lambda}} = (\sum d_k(s) \mathbf{x}_k \mathbf{x}_k^T)^{-1} (\mathbf{X} - \hat{\mathbf{X}})$.

In general, the calibration estimator $\hat{Y}_w = \sum w_k(s) y_k$ with weights $w_k(s)$ given by (3.16) may not be expressible as a function of estimated totals. We therefore follow Binder's (1983) approach and expand $F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}})$ around $\boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ denotes the probability limit of $\hat{\boldsymbol{\lambda}}$. We get

$$F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}) \approx F(\mathbf{x}_k^T \boldsymbol{\lambda}) + f(\mathbf{x}_k^T \boldsymbol{\lambda}) \mathbf{x}_k^T (\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}), \quad (3.18)$$

where $f(a) = \partial F(a) / \partial a$. Further, by expanding the calibration equations (3.17) around $\boldsymbol{\lambda}$, we obtain after simplification,

$$\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda} \approx -\hat{\mathbf{Q}}_{\boldsymbol{\lambda}}^{-1} (\hat{\mathbf{S}}_{\boldsymbol{\lambda}} - \mathbf{X}) \quad (3.19)$$

where $\hat{\mathbf{Q}}_{\boldsymbol{\lambda}} = \sum d_k(s) f(\mathbf{x}_k^T \boldsymbol{\lambda}) \mathbf{x}_k \mathbf{x}_k^T$ and $\hat{\mathbf{S}}_{\boldsymbol{\lambda}} = \sum d_k(s) F(\mathbf{x}_k^T \boldsymbol{\lambda}) \mathbf{x}_k$. Note that both $\hat{\mathbf{Q}}_{\boldsymbol{\lambda}}$ and $\hat{\mathbf{S}}_{\boldsymbol{\lambda}}$ are of the form of estimated totals. Substituting (3.19) into (3.18) gives

$$F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}) \approx F(\mathbf{x}_k^T \boldsymbol{\lambda}) - f(\mathbf{x}_k^T \boldsymbol{\lambda}) \mathbf{x}_k^T \hat{\mathbf{Q}}_{\boldsymbol{\lambda}}^{-1} (\hat{\mathbf{S}}_{\boldsymbol{\lambda}} - \mathbf{X}). \quad (3.20)$$

Using the approximation (3.20) in (3.16), it follows that \hat{Y}_w is approximated by a differentiable function of estimated totals. Hence, the general theory of section 2 is applicable and it remains to evaluate $z_k = \partial h(\mathbf{b}) / \partial b_k |_{b=d(s)}$, where $h(\mathbf{b}) = \sum b_k g_k^*(\mathbf{b}) y_k$ with

$$g_k^*(\mathbf{b}) = F(\mathbf{x}_k^T \boldsymbol{\lambda}) - f(\mathbf{x}_k^T \boldsymbol{\lambda}) \mathbf{x}_k^T \mathbf{Q}_{\boldsymbol{\lambda}}(\mathbf{b})^{-1} (\mathbf{S}_{\boldsymbol{\lambda}}(\mathbf{b}) - \mathbf{X})$$

where $\mathbf{Q}_{\boldsymbol{\lambda}}(\mathbf{b}) = \sum b_k f(\mathbf{x}_k^T \boldsymbol{\lambda}) \mathbf{x}_k \mathbf{x}_k^T$ and $\mathbf{S}_{\boldsymbol{\lambda}}(\mathbf{b}) = \sum b_k F(\mathbf{x}_k^T \boldsymbol{\lambda}) \mathbf{x}_k$. After simplifications, we get

$$z_k = F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}) (y_k - \mathbf{x}_k^T \hat{\mathbf{B}}_{\boldsymbol{\lambda}}) = F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}) e_{k\boldsymbol{\lambda}}, \quad (3.21)$$

where

$$\hat{\mathbf{B}}_{\boldsymbol{\lambda}} = \left(\sum d_k(s) f(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}) \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} \sum d_k(s) f(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}) \mathbf{x}_k y_k.$$

Singh and Folsom (2000) obtained a similar result, using a somewhat different approach.

The result (3.21) may also be obtained directly along the lines of (3.2) and (3.3) by writing \hat{Y}_w as $f(\mathbf{d}(s))$ and evaluating $z_k = \partial f(\mathbf{b}) / \partial b_k |_{b=d(s)}$, where $f(\mathbf{b}) = \sum b_k g_k(\mathbf{b}) y_k$ with $g_k(\mathbf{b}) = F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}(\mathbf{b}))$. We have

$$\begin{aligned} \partial(b_k g_k(\mathbf{b})) / \partial b_k &= g_k(\mathbf{b}) \\ &+ b_k f(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}(\mathbf{b})) \mathbf{x}_k^T (\partial \hat{\boldsymbol{\lambda}}(\mathbf{b}) / \partial b_k), \end{aligned} \quad (3.22)$$

and for $l \neq k$

$$\partial(b_l g_l(\mathbf{b})) / \partial b_k = b_l f(\mathbf{x}_l^T \hat{\boldsymbol{\lambda}}(\mathbf{b})) \mathbf{x}_l^T (\partial \hat{\boldsymbol{\lambda}}(\mathbf{b}) / \partial b_k). \quad (3.23)$$

To evaluate $\partial \hat{\boldsymbol{\lambda}}(\mathbf{b}) / \partial b_k$, we take the derivatives of the calibration equations (3.17) with $\mathbf{d}(s)$ replaced by $\mathbf{b} : \sum b_k F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}(\mathbf{b})) \mathbf{x}_k - \mathbf{X} = \mathbf{0}$. This gives

$$\begin{aligned} \mathbf{0} &= F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}(\mathbf{b})) \mathbf{x}_k \\ &+ \sum_l b_l f(\mathbf{x}_l^T \hat{\boldsymbol{\lambda}}(\mathbf{b})) \mathbf{x}_l \mathbf{x}_l^T (\partial \hat{\boldsymbol{\lambda}}(\mathbf{b}) / \partial b_k) \end{aligned}$$

or

$$\begin{aligned} \partial \hat{\boldsymbol{\lambda}}(\mathbf{b}) / \partial b_k &= \\ &- \left(\sum b_k f(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}(\mathbf{b})) \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}}(\mathbf{b})) \mathbf{x}_k. \end{aligned} \quad (3.24)$$

Substituting (3.24) into (3.22) and (3.23), we get (3.21) after simplification.

Deville and Särndal (1992) showed that the asymptotic variance of \hat{Y}_w for general $F(\cdot)$ is equivalent to the asymptotic variance of the GREG estimator which involves the ‘‘census’’ regression coefficient \mathbf{B} . Using this result they obtained a variance estimator of \hat{Y}_w for general $F(\cdot)$, by replacing \mathbf{B} by $\hat{\mathbf{B}} = (\sum w_k(s) \mathbf{x}_k \mathbf{x}_k^T)^{-1} \sum w_k(s) \mathbf{x}_k y_k$, where $w_k(s) = d_k(s) F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}})$. The resulting z_k agrees with our z_k given by (3.21) if $f(a) = F(a)$, i.e., in the case of generalized raking weights. In the case of GREG estimator, we have $F(x) = 1 + x$, $f(x) = 1$ and $\hat{\boldsymbol{\lambda}} = (\sum d_k(s) \mathbf{x}_k \mathbf{x}_k^T)^{-1} (\mathbf{X} - \hat{\mathbf{X}})$. It readily follows that $F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}})$ reduces to the customary g -weight $g_k(\mathbf{d}(s)) = 1 + (\mathbf{X} - \hat{\mathbf{X}})^T (\sum d_k(s) \mathbf{x}_k \mathbf{x}_k^T)^{-1} \mathbf{x}_k$, and $e_{k\lambda} = y_k - \mathbf{x}_k^T \hat{\boldsymbol{\lambda}}$ reduces to $e_k = y_k - \mathbf{x}_k^T \hat{\mathbf{B}}$ with $\hat{\mathbf{B}} = (\sum d_k(s) \mathbf{x}_k \mathbf{x}_k^T)^{-1} \sum d_k(s) \mathbf{x}_k y_k$. Note that our z_k in this case is different from the z_k of Deville and Särndal (1992), but agrees with a commonly used z_k (Särndal, Swensson and Wretman 1989).

Our method, along the lines of section 3.3, can be extended to implicitly defined estimators, $\hat{\theta}_w$, obtained as solutions to estimating equations (3.10) based on the general calibration weights (3.16). Details are omitted for simplicity.

4. Two-Phase Sampling

We extend our method to two-phase sampling, assuming the estimator $\hat{\theta}$ of a parameter θ can be expressed as a differentiable function, $g(\hat{\mathbf{Y}}, \hat{\mathbf{X}}^{(1)})$, of estimated totals, $\hat{\mathbf{Y}} = (\hat{Y}_1, \dots, \hat{Y}_m)^T$, from the second-phase sample and estimated totals, $\hat{\mathbf{X}}^{(1)} = (\hat{X}^{(1)}_1, \dots, \hat{X}^{(1)}_p)^T$, from the first-phase sample only. Here $\hat{Y}_i = \sum_{k=1}^N d_k(s) y_{ik}$, $i = 1, \dots, m$, $\hat{X}^{(1)}_j = \sum_{k=1}^N d_k^{(1)}(s_1) x_{jk}$, $j = 1, \dots, p$, $d_k^{(1)}(s_1)$ denotes the first-phase design weight attached to the k^{th} element with $d_k(s_1) = 0$ if k is not in the first-phase sample s_1 , and $d_k(s)$ is the final design weight attached to the k^{th}

element with $d_k(s) = 0$ if k is not in the second-phase sample s . Further, the parameter $\theta = g(\mathbf{Y}, \mathbf{X})$ with $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_m)^T$ and $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)^T$ denoting the vectors of Y - and X - totals. For example, the two-phase ratio estimator, \hat{Y}_{R2} , is of the form $\hat{\theta} = g(\hat{Y}, \hat{X}, \hat{X}^{(1)})$:

$$\begin{aligned} \hat{Y}_{R2} &= \frac{\hat{Y}}{\hat{X}} \hat{X}^{(1)} = \hat{R} \hat{X}^{(1)} \\ &= \frac{\sum d_k(s) y_k}{\sum d_k(s) x_k} (\sum d_k^{(1)}(s_1) x_k). \end{aligned} \quad (4.1)$$

Note that $\hat{\mathbf{Y}} = (\hat{Y}_1, \hat{Y}_2)^T$ with $\hat{Y}_1 = \hat{Y}$, $\hat{Y}_2 = \hat{X}$, and $\hat{\mathbf{X}}^{(1)} = \hat{X}^{(1)}$. Also, $\theta = g(Y, X, X^{(1)}) = Y$.

For simplicity, consider a $g(\cdot)$ such that $N^{-1} g(\cdot)$ tends to a limit. Taylor linearization of $\hat{\theta} = g(\hat{\mathbf{Y}}, \hat{\mathbf{X}}^{(1)})$ around (\mathbf{Y}, \mathbf{X}) gives

$$\begin{aligned} \hat{\theta} - \theta &= g(\hat{\mathbf{Y}}, \hat{\mathbf{X}}^{(1)}) - g(\mathbf{Y}, \mathbf{X}) \\ &\approx (\partial g(\mathbf{a}, \mathbf{a}^{(1)})) / \partial \mathbf{a} \Big|_{\mathbf{a}=\mathbf{Y}, \mathbf{a}^{(1)}=\mathbf{X}} (\hat{\mathbf{Y}} - \mathbf{Y}) \\ &+ (\partial g(\mathbf{a}, \mathbf{a}^{(1)}) / \partial \mathbf{a}^{(1)})^T \Big|_{\mathbf{a}=\mathbf{Y}, \mathbf{a}^{(1)}=\mathbf{X}} (\hat{\mathbf{X}}^{(1)} - \mathbf{X}). \end{aligned} \quad (4.2)$$

Let $\tilde{\mathbf{Y}} = \sum b_k y_k$ and $\tilde{\mathbf{X}}^{(1)} = \sum b_k^{(1)} x_k$ for arbitrary real numbers $\mathbf{b} = (b_1, \dots, b_N)^T$ and $\mathbf{b}^{(1)} = (b_1^{(1)}, \dots, b_N^{(1)})^T$. Also, let $g(\tilde{\mathbf{Y}}, \tilde{\mathbf{X}}^{(1)}) = f(\mathbf{b}, \mathbf{b}^{(1)}, \mathbf{A}_y, \mathbf{A}_x) = f(\mathbf{b}, \mathbf{b}^{(1)})$, where \mathbf{A}_y is an $m \times N$ matrix with k^{th} column $\mathbf{x}_k = (x_{k1}, \dots, x_{kp})^T$, $k = 1, \dots, N$, and \mathbf{A}_x is an $p \times N$ matrix with k^{th} column $\mathbf{y}_k = (y_{k1}, \dots, y_{km})^T$, $k = 1, \dots, N$. Now following the derivation of (2.3) and noting that $\hat{\mathbf{Y}} = \mathbf{A}_y \mathbf{d}(s)$, $\mathbf{Y} = \mathbf{A}_y \mathbf{1}$, $\hat{\mathbf{X}}^{(1)} = \mathbf{A}_x \mathbf{d}^{(1)}(s_1)$, $\mathbf{X} = \mathbf{A}_x \mathbf{1}$, it can be shown that (4.2) reduces to

$$\hat{\theta} - \theta \approx \tilde{z}^T (\mathbf{d}(s) - \mathbf{1}) + \tilde{z}^{(1)T} (\mathbf{d}^{(1)}(s_1) - \mathbf{1}), \quad (4.3)$$

where $\mathbf{d}(s) = (d_1(s), \dots, d_N(s))^T$ and $\mathbf{d}^{(1)}(s_1) = (d_1^{(1)}(s_1), \dots, d_N^{(1)}(s_1))^T$. Further, $\tilde{z} = (\tilde{z}_1, \dots, \tilde{z}_m)^T$ with $\tilde{z}_k = \partial f(\mathbf{b}, \mathbf{b}^{(1)}) / \partial b_k \Big|_{b=1, b^{(1)}=1}$, and $\tilde{z}^{(1)} = (\tilde{z}_1^{(1)}, \dots, \tilde{z}_p^{(1)})^T$ with $\tilde{z}_k^{(1)} = \partial f(\mathbf{b}, \mathbf{b}^{(1)}) / \partial b_k^{(1)} \Big|_{b=1, b^{(1)}=1}$. It follows from (4.3) that a variance estimator of $\hat{\theta}$ is approximately given by the variance estimator of the estimated total $\sum d_k(s) \tilde{z}_k + \sum d_k^{(1)}(s_1) \tilde{z}_k^{(1)} = \hat{Y}(\tilde{z}) + \hat{X}^{(1)}(\tilde{z}^{(1)})$. We denote the latter variance estimator as $v(\tilde{z}, \tilde{z}^{(1)})$. Now we replace \tilde{z}_k and $\tilde{z}_k^{(1)}$ by $z_k = \partial f(\mathbf{b}, \mathbf{b}^{(1)}) / \partial b_k \Big|_{b=d(s), b^{(1)}=d^{(1)}(s_1)}$ and $z_k^{(1)} = \partial f(\mathbf{b}, \mathbf{b}^{(1)}) / \partial b_k^{(1)} \Big|_{b=d(s), b^{(1)}=d^{(1)}(s_1)}$ respectively, since \tilde{z}_k and $\tilde{z}_k^{(1)}$ are unknown. This lead to a linearization variance estimator

$$v_L(\hat{\theta}) = v(z, z^{(1)}). \quad (4.4)$$

We now consider the special case of a ‘‘double expansion’’ estimator $\hat{Y}(y) = \sum d_k(s) y_k$ with $d_k(s) = \pi_{1k}^{-1} \pi_{2k/1}^{-1}$ for $k \in s$ and the Horvitz-Thompson (H-T) estimator $\hat{X}^{(1)}(x) = \sum d_k^{(1)}(s_1) x_k$ with $d_k^{(1)}(s_1) = \pi_{1k}^{-1}$ for $k \in s_1$, where π_{1k} is probability of including element k in s_1 , and $\pi_{2k/1}$ is the conditional probability of including element k in s

given s_1 . In this case, an unbiased H-T type estimator of $\hat{Y}(y) + \hat{X}^{(1)}(x)$ is given by

$$\begin{aligned}
 v(y, x) &= \sum_{k, l \in s_1} \sum_{k, l \in s} \frac{\pi_{1kl} - \pi_{1k} \pi_{1l}}{\pi_{1kl}} \frac{x_k}{\pi_{1k}} \frac{x_l}{\pi_{1l}} \\
 &+ \sum_{k, l \in s} \sum_{k, l \in s} \frac{\pi_{1kl} - \pi_{1k} \pi_{1l}}{\pi_{kl}^*} \left(\frac{y_k}{\pi_{1k}} \frac{y_l}{\pi_{1l}} + 2 \frac{y_k}{\pi_{1k}} \frac{x_l}{\pi_{1l}} \right) \\
 &+ \sum_{k, l \in s} \sum_{k, l \in s} \frac{\pi_{2kl/1} - \pi_{2k/1} \pi_{2l/1}}{\pi_{2kl/1}} \frac{y_k}{\pi_k^*} \frac{y_l}{\pi_l^*} \quad (4.5)
 \end{aligned}$$

where $\pi_k^* = \pi_{1k} \pi_{2k/1}$, $\pi_{kl}^* = \pi_{1kl} \pi_{2kl/1}$, π_{1kl} is the probability of including both elements k and l in s_1 and $\pi_{2kl/1}$ is the conditional probability of including both elements k and l in s given s_1 . A proof of (4.5) is given in the Appendix. The variance estimator (4.4) is obtained from (4.5) by changing y_k and x_k to z_k and $z_k^{(1)}$ respectively.

Example 4.1 We illustrate the calculation of $v(z, z^{(1)})$ for the two-phase ratio estimator \hat{Y}_{R2} , given by (4.1), for the special case of simple random sampling at both phases: s_1 is a simple random sample of size n and s is a simple random subsample of size m from s_1 . In this case, $\pi_{1k} = n/N$ and $\pi_{2k/1} = m/n$. Further, it follows from (4.1) that for general two-phase design,

$$z_k = \frac{\hat{X}^{(1)}}{\hat{X}}(y_k - \hat{R}x_k) = \frac{\hat{X}^{(1)}}{\hat{X}}e_k \quad (4.6)$$

and

$$z_k^{(1)} = \hat{R}x_k. \quad (4.7)$$

Under simple random sampling at both stages, (4.6) and (4.7) reduce to $z_k = (\bar{x}^{(1)} / \bar{x})e_k$ and $z_k^{(1)} = (\bar{y} / \bar{x})x_k$, where $e_k = y_k - (\bar{y} / \bar{x})x_k$, \bar{y} and \bar{x} are the second-phase sample means of y and x respectively, and $\bar{x}^{(1)}$ is the first-phase sample mean of x . Now substituting z_k and $z_k^{(1)}$ for y and x in (4.5) and noting that $\pi_{1kl} = n(n-1)/[N(N-1)]$, $\pi_{2k/1} = m(m-1)/[n(n-1)]$, $\pi_{1kk} = \pi_{1k}$ and $\pi_{2kk/1} = \pi_{2k/1}$, we get

$$\begin{aligned}
 v_L(\hat{Y}_{R2}) &= N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \hat{R}^2 s_{1x}^2 \\
 &+ N^2 \left(\frac{1}{m} - \frac{1}{n} \right) \left(\frac{\bar{x}^{(1)}}{\bar{x}} \right)^2 s_{2e}^2 \\
 &+ 2N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \hat{R} \frac{\bar{x}^{(1)}}{\bar{x}} s_{ex}, \quad (4.8)
 \end{aligned}$$

where

$$\begin{aligned}
 \hat{R} &= \bar{y} / \bar{x}, s_{1x}^2 = (n-1)^{-1} \sum_{k \in s_1} (x_k - \bar{x}^{(1)})^2, \\
 s_{2e}^2 &= (m-1)^{-1} \sum_{k \in s} (e_k - \bar{e})^2, \\
 s_{2ex} &= (m-1)^{-1} \sum_{k \in s} (e_k - \bar{e})(x_k - \bar{x})
 \end{aligned}$$

and \bar{e} is the second-phase sample mean of e . The formula (4.8) agrees with the formula derived by Rao and Sitter (1995). It is different from the customary formula (Sukhatme and Sukhatme 1970, page 176) which fails to make use of the full x -data $\{x_k, k \in s_1\}$. Rao and Sitter (1995) demonstrated through simulation that $v_L(\hat{Y}_{R2})$ is more efficient than the customary variance estimator. Also, $v_L(\hat{Y}_{R2})$ performed better in tracking the conditional mean squared error of \hat{Y}_{R2} ; see Rao and Sitter (1995, section 3) for details of the simulation study.

Concluding Remarks

We have presented a unified approach to deriving Taylor linearization variance estimators and applied it to a variety of problems. It leads directly to a variance estimator that has some desirable properties at least in a number of important special cases; in particular, approximate unbiasedness for the model variance of the estimator under an assumed model and validity under a conditional repeated sampling framework. It would be useful to investigate whether such desirable properties also hold for more complex cases such as the general class of calibration estimators (section 3.2), the estimators based on estimating equations (section 3.3) and two-phase sampling (section 4). We are currently investigating various extensions of our method, including variance estimation under imputation for item nonresponse and variance estimation from longitudinal survey data.

Acknowledgments

We thank the Associate Editor and a referee for constructive comments and suggestions. We also thank several colleagues in Statistics Canada for useful suggestions and encouragement, especially Linda Standish, David Binder, Geoff Hole, Richard Burgess and Larry Swain. Demnati's work was made possible by the Small Area and Administrative Data Division of Statistics Canada. J.N.K. Rao's work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

Appendix

Unbiased Variance Estimator of $\hat{Y}(y) + \hat{X}^{(1)}(x)$

The variance of $\hat{Y}(y) + \hat{X}^{(1)}(x)$ is the sum of the variance of $\hat{Y}(y)$, the variance of $\hat{X}^{(1)}(x)$ and twice the covariance of $\hat{Y}(y)$ and $\hat{X}^{(1)}(x)$. An unbiased H-T type estimator of

$V[\hat{Y}(y)]$ is given by Särndal, Swensson and Wretman (1991, chapter 9, page 348):

$$v[\hat{Y}(y)] = \sum_{k,l \in s} \frac{\pi_{1kl} - \pi_{1k}\pi_{1l}}{\pi_{kl}^*} \frac{y_k x_l}{\pi_{1k} \pi_{1l}} + \sum_{k,l \in s} \frac{\pi_{2kl/1} - \pi_{2k/1}\pi_{2l/1}}{\pi_{2kl/1}} \frac{y_k y_l}{\pi_{k}^* \pi_{l}^*}. \quad (\text{A.1})$$

An unbiased H-T type estimator of $V[\hat{X}^{(1)}(x)]$ is given by

$$v[\hat{X}^{(1)}(x)] = \sum_{k,l \in s_1} \frac{\pi_{1kl} - \pi_{1k}\pi_{1l}}{\pi_{1kl}} \frac{x_k x_l}{\pi_{1k} \pi_{1l}}. \quad (\text{A.2})$$

Further,

$$\text{Cov}[\hat{Y}(y), \hat{X}^{(1)}(x)] = E\text{Cov}_2[\hat{Y}(y), \hat{X}^{(1)}(x)] + \text{Cov}[E_2(\hat{Y}(y)), E_2(\hat{X}^{(1)}(x))],$$

where E_2 and Cov_2 denote conditional expectation and conditional covariance given s_1 . Noting that

$$E_2 \hat{Y}(y) = \hat{X}^{(1)}(y), E_2 \hat{X}^{(1)}(x) = \hat{X}^{(1)}(x)$$

and $\text{Cov}_2[\hat{Y}(y), \hat{X}^{(1)}(x)]$ we get

$$\text{Cov}[\hat{Y}(y), \hat{X}^{(1)}(x)] = \text{Cov}[\hat{X}^{(1)}(y), \hat{X}^{(1)}(x)].$$

An unbiased H-T type estimator of $2\text{Cov}[\hat{X}^{(1)}(y), \hat{X}^{(1)}(x)]$ is given by

$$2\text{Cov}[\hat{X}^{(1)}(y), \hat{X}^{(1)}(x)] = 2 \sum_{k,l \in s} \frac{\pi_{1kl} - \pi_{1k}\pi_{1l}}{\pi_{kl}^*} \frac{x_k x_l}{\pi_{1k} \pi_{1l}}. \quad (\text{A.3})$$

The sum of (A.1), (A.2) and (A.3) equals (4.5).

References

- Anderson, C., and Nordberg, L. (1994). A method for variance estimation of non-linear functions of totals in surveys—theory and software implementation. *Journal of Official Statistics*, 10, 395-405
- Berger, Y.G. (2002). A generalized jackknife variance estimator for nonlinear statistics in probability sampling. Technical Report, Department of Social Statistics, University of Southampton.
- 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.
- Campbell, C. (1980). A different view of finite population estimation. *Proceeding of the Section on Survey Research Methods*, American Statistical Association, 319-324.
- Deville, J.C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25, 193-203.
- Deville, J.C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley & Sons, Inc.
- Huber, P.J. (1981). *Robust Statistics*. New York: John Wiley & Sons, Inc.
- Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.
- Montanari, G.E. (1998). On regression estimation of finite population means. *Survey Methodology*, 24, 69-77.
- Rao, J.N.K., and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.
- Rao, J.N.K., Yung, W. and Hidiroglou, M. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā*.
- 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.
- Särndal, C.-E., Swensson, B. and Wretman, J.H. (1991). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley & Sons, Inc.
- Singh, A.C., and Folsom, R.E. (2000). Bias correcting estimating function approach for variance estimation adjusted for poststratification. *Proceeding of the Section on Survey Research Methods*, American Statistical Association, 610-615.
- Sitter, R.R., and Wu, C. (2002). Efficient estimation of quadratic finite population functions in the presence of auxiliary information. *Journal of the American Statistical Association*, 97, 535-544.
- Sukhatme, P.V., and Sukhatme, B.V. (1970). *Sampling Theory of Surveys with Applications*. 2nd ed. London: Asia Publishing House.
- Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.
- Yung, W., and Rao, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, 22, 23-31.

Comment

Phillip S. Kott ¹

The article addresses an impressive number of contexts, many of which have only recently been investigated in the literature, often by Professor Rao himself. I will have little to say here about estimating functions with calibration weights or two-phase sampling, except (mostly) to agree with the solutions advocated in the text. Instead, I will focus on three applications: the ratio estimator under simple random sampling discussed in the Introduction, the general class of regression calibration weights from section 3.2, and the general class of calibration estimators from section 3.4. I will end with a question about the linearization variance estimator in full Horvitz-Thompson form, which has bothered me for some time.

The Ratio Under Simple Random Sampling

Before beginning, let me confess to a certain skepticism about the general method proposed in section 2. I find that techniques of this sort work best when you already know what the answer is. Godambe and Thompson (1986) tried to use estimating functions to settle a controversy then surrounding the best variance estimator for the ratio under simple random sampling. Using the notation in the text, they demonstrated that $(\bar{X}/\bar{x})^2 v_L$ was the proper way to estimate the variance of a ratio estimator, $\hat{Y}_R(\bar{X}/\bar{x})\bar{y}$. Later, Binder (1996) corrected them. He showed that when done properly, $v_{JL}(\bar{X}/\bar{x})^2 v_L$ is produced from estimating-function technology. It helped that he already knew that was the better answer.

As Demanti and Rao state, v_{JL} has both good randomization (design) and model-based properties (here and hereafter I omit the qualifier, “under mild conditions which I assume to hold”). In fact, when n/N is ignorably small, v_{JL} has a relative bias of $O(1/n)$ as an estimator for the model variance of \hat{Y}_R . If the y_k are uncorrelated, then this is not only true when $V_m(y_k) = \sigma^2 x_k$ as stated in the text, but, more generally, when $V_m(y_k) = \sigma_k^2$. Unfortunately, the result is less general when n/N is not ignorably small. In that context, when the y_k are uncorrelated and $V_m(y_k) = \sigma^2 x_k$, a more appropriate estimator for the model variance of \hat{Y}_R is $v_m = [(\bar{X}/\bar{x})^2 - (n/N)(\bar{X}/\bar{x})] [1 - (n/N)]^{-1} v_L$ (Kott and Brewer 2001). As an estimator for the randomization mean squared error of \hat{Y}_R , v_m has a relative bias of $O(1/\sqrt{n})$, just like v_{JL} and v_L .

When simple random sampling is used in practice the sampling fraction is almost always small. Thus, v_{JL} is an

attractive variance/mean-squared-error estimator, and my criticism of Demnati and Rao for advocating it is mild.

A General Class of Regression Calibration Weights

I would generalize the results of section 3.1 in a different manner than the authors do in section 3.2. Following Estavao and Särndal (2002), replace $c_k \mathbf{x}_k$ in equation (3.1) with a vector \mathbf{q}_k having the same dimension as \mathbf{x}_k . The rest of that section follows easily.

One choice for \mathbf{q}_k is

$$\mathbf{q}_{(1)k} = \sum_{j \in U} (\pi_{kj} - \pi_k \pi_j) \mathbf{x}_j / (\pi_k \pi_j),$$

the use of which results in a variant of the randomization-optimal regression estimator proposed by Tillé (1999). Observe that $(\sum_U \mathbf{q}_{(1)k} \mathbf{x}_k^T)^{-1} (\sum_U \mathbf{q}_{(1)k} y_k^T) = [\text{Var}(\hat{X})]^{-1} \text{Cov}(\hat{X}, \hat{Y})$, where Var and Cov denote randomization-based properties.

Another choice, investigated indirectly by Demnati and Rao and likewise resulting into a variant of the randomization-optimal estimator, is

$$\mathbf{q}_{(2)k} = \sum_{j \in s} (\pi_{kj} - \pi_k \pi_j) \mathbf{x}_j / (\pi_{kj} \pi_j).$$

Since $\mathbf{q}_{(2)k}$ is a function of the sample, the authors take us through the complications of section 3.2. This was only necessary for randomization-based inference. I would have gone a different way. Observe that $d_k(s) \mathbf{q}_{(2)k} - d_k(s) \mathbf{q}_{(1)k} = O_p(1/\sqrt{n})$. Replacing one for the other has an asymptotically ignorable effect on $w_k(s)$ (i.e., the relative difference is $O_p(1/n)$).

A General Class of Calibration Estimators

A mild generalization of equation (3.16) allows calibration weights of the form,

$$w_k(s) = d_k(s) F(\mathbf{q}_k^T \hat{\boldsymbol{\lambda}}),$$

where \mathbf{q}_k again has the same dimension as \mathbf{x}_k . For convenience F is assumed positive and twice differentiable around $\mathbf{q}_k^T \boldsymbol{\lambda}$. Without loss of generality, one can assume $\boldsymbol{\lambda}$ (the limit of $\hat{\boldsymbol{\lambda}}$) is $\mathbf{0}$, and $f(0) > 1$. When $\hat{Y}_{GC} = \sum_U w_k(s) y_k$ is a randomization consistent estimator, as I assume it is, $F(0)$ is equal to 1.

Paralleling the development in the text leads ultimately to

$$z_k = F(\mathbf{q}_k^T \hat{\boldsymbol{\lambda}}) (y_k - \mathbf{x}_k^T \hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) = F(\mathbf{q}_k^T \hat{\boldsymbol{\lambda}}) e_{k\boldsymbol{\lambda}},$$

1. Phillip S.Kott, USDA / NASS, 3251 Old Lee Hwy, Fairfax, VA 22030, U.S.A.

where $\hat{\mathbf{B}}_\lambda = [\sum d_k(s) f(\mathbf{q}_k^T \hat{\boldsymbol{\lambda}}) \mathbf{q}_k \mathbf{x}_k^T]^{-1} \sum d_k(s) f(\mathbf{q}_k^T \hat{\boldsymbol{\lambda}}) \mathbf{q}_k y_k$. The presence of the $f(\cdot)$ in the expression of $\hat{\mathbf{B}}_\lambda$ may be a bit of a surprise, but, it turns out, not a meaningful one in this context. For inference under the prediction model, $E_m(y_k | \mathbf{x}_k) = \mathbf{x}_k^T \boldsymbol{\beta}$, the derivative can be replaced by any constant without asymptotic consequence; $\hat{\mathbf{B}}_\lambda$ remains a model unbiased estimator for $\boldsymbol{\beta}$. For randomization-based inference, since $\mathbf{q}_k^T \hat{\boldsymbol{\lambda}} = O_p(1/\sqrt{n})$ and $F(0), f(0) > 0, z_k$ would be unaffected asymptotically if $f(\mathbf{q}_k^T \hat{\boldsymbol{\lambda}})$ were replaced by 1 or by $F(\mathbf{q}_k^T \hat{\boldsymbol{\lambda}})$.

Things change, however, if we push the envelop a bit. Fuller, Loughin and Baker (1994) use calibration to adjust for unit nonresponse by treating sample response as a second phase of sampling. They assume that every element k in the population has a Poisson probability of sample response, π_{2k} , which is independent of whether it is actually chosen for the sample. They further assume $\pi_{2k} = 1/(1 + \mathbf{x}_k^T \boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is unknown and implicitly estimated by calibration. Here we generalize that and assume $\pi_{2k} = 1/F(\mathbf{q}_k^T \boldsymbol{\lambda})$, where F is known, positive, and twice differentiable. In practice, \mathbf{q}_k will likely be identical to \mathbf{x}_k , but it may be reasonable to replace one of more components of \mathbf{x}_k with variables conjectured to be more strongly correlated with response/nonresponse.

Redefining s as the respondent sample and $d_k(s)$ as $(1/\pi_{1k})$ when $k \in s, 0$ otherwise, everything proceeds as before. The difference is that $f(\mathbf{q}_k^T \hat{\boldsymbol{\lambda}})$ in $\hat{\mathbf{B}}_\lambda$ need no longer need be asymptotically identical across the k . Thus, the term can matter even with a large sample.

Now $V(\hat{Y}_{GC}) \approx V(\sum_U d_k(s) z_k)$, where $\sum_U d_k(s) z_k = \sum_U d_k(s) F(\mathbf{q}_k^T \hat{\boldsymbol{\lambda}}) e_{k\lambda}$ is the double expansion estimation. Substituting $1/F(\mathbf{q}_k^T \hat{\boldsymbol{\lambda}})$ for π_{2k} , the variance estimator for \hat{Y}_{GC} becomes (from equation (A.1) with $\pi_{2kj/1} = \pi_{2kj} \pi_{2k} \pi_{2j}$

$$v(\hat{Y}_{GC}) = \sum_{k, j \in s} [(\pi_{1kj} - \pi_{1k} \pi_{1j}) / \pi_{1kj}] \\ d_k(s) F(\mathbf{q}_k^T \hat{\boldsymbol{\lambda}}) e_{k\lambda} d_j(s) F(\mathbf{q}_j^T \hat{\boldsymbol{\lambda}}) e_{j\lambda} \\ + \sum_{k \in s} \pi_{1k} \{ [F(\mathbf{q}_k^T \hat{\boldsymbol{\lambda}})]^2 - [F(\mathbf{q}_k^T \hat{\boldsymbol{\lambda}})] \} [d_k(s) e_{k\lambda}]^2.$$

This differs from the variance estimator in Folsom and Singh (2000) mainly because those authors assume the original sample is chosen using a stratified multistage design employing with-replacement sampling in the first. That, among other things, annihilates the second summation on the right hand side.

Not only does $v(\hat{Y}_{GC})$ estimate the quasi-randomization mean squared error of \hat{Y}_{GC} – “quasi” because a response model is assumed, it also estimates the model variance of \hat{Y}_{GC} . In fact, the relative bias of $v(\hat{Y}_{GC})$ under the

prediction model, $E_m(y_k | \mathbf{x}_k, \mathbf{q}_k) = \mathbf{x}_k^T \boldsymbol{\beta}$, is $O(1/n)$ when the y_k are uncorrelated and $V_m(y_k | \mathbf{x}_k, \mathbf{q}_k) = \mathbf{x}_k^T \boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ (like $\boldsymbol{\beta}$) need not be specified. Surprisingly, the second term in $v(\hat{Y}_{GC})$ provides the model-based correction I recommended for the ratio estimator under simple random sampling *in the absence of nonresponse*.

Does the “Plug-in” Variance Estimator Really Work for the Full Horvitz-Thompson Form?

As I warned parenthetically early on, I have omitted the key phrase, “under mild conditions which I assume to hold,” repeatedly in these comments. Now, I want to turn my attention to what may be one of those conditions. It is standard in variance estimation to replace population (or model) values with sample analogues since their difference is asymptotically ignorable. That is done, for example, by Demnati and Rao in equation (2.4) when they plug in z_k for \tilde{z}_k . The question I want to raise, and for which I do not know the answer, is this. Suppose one is estimating a total with a calibration estimator. The total is $O(N)$, and $O(n) = O(N)$. The estimator’s model variance and randomization mean squared error are also $O(n)$. Is it legitimate to plug in z_k for \tilde{z}_k , where $z_k - \tilde{z}_k = O_p(1/\sqrt{n})$, when there are $n(n-1)/2$ terms in the Horvitz-Thompson – or Yates-Grundy – variance/mean-squared-error estimator? In most practical applications, this is a non-issue, because the variance estimator can be re-expressed with $O(n)$ terms. What if that is not the case?

Let me conclude these remarks by thanking Drs. Demnati and Rao for their stimulating article and *Survey Methodology* for both publishing it and allowing me to provide some comments.

Additional References

- Estevao, V.M., and Särndal, C-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics*, 18, 233-255.
- Fuller, W.A., Loughin, M.M. and Baker, H.D. (1994). Regression weighting for the 1987-88 National Food Consumption Survey. *Survey Methodology*, 20, 75-85.
- Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: their relationship and estimation. *International Statistical Review*, 54, 2, 127-138.
- Kott, P.S., and Brewer K.R.W. (2001). Estimating the model variance of a randomization-consistent regression estimator. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
- Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 811-822.
- Tillé, Y. (1999). Estimation in surveys using conditional inclusion probabilities: complex designs. *Survey Methodology*, 25, 57-66.

Comment

Babubhai V. Shah ¹

This is an excellent paper that removes the mystery underlying Taylor linearization. Most data analysis applications use Horvitz-Thompson weights that are reciprocals of the probabilities of selection. The simplest prescription for deriving the linearization for an estimator $\hat{\theta}$ is as follows:

1. For each observation, create a new variable $z_i = \partial \hat{\theta} / \partial w_i$, where w_i is the reciprocal of the selection probability for the i^{th} observation selected in the sample. In cases where the estimator $\hat{\theta}$ is defined implicitly through estimating equations, the derivative can be computed by differentiating the implicit equations.
2. Define weighted $\hat{T} = \sum w_i z_i$ total.
3. Compute the variance \hat{V} of the total \hat{T} based on the sample design.
4. The variance \hat{V} is the approximate variance of the estimator $\hat{\theta}$.

If the parameter θ is a vector then the variable z_i and the total T are also vectors and \hat{V} is an approximate estimate of the variance covariance matrix of the estimator $\hat{\theta}$.

The steps (1) and (2) specified above produce the correct linearization in the following cases:

- a. Means, proportions, and ratio estimates.
- b. Generalized linear regression models.
- c. Predicted marginal for generalized linear model.
- d. Estimate of the mean from regression imputed data.
- e. Generalized linear regression models with calibrated weights.

- f. Wilcoxon two sample rank sum test.
- g. Estimates of coefficients and the hazard rate in Cox's proportional hazard model.
- h. Estimates of predicted marginal survival in Cox's proportional hazard model.
- i. Two-phase sample survey.

The derivation in the step (1) is uniquely defined and does not contain the true value of the parameter θ , and does not require substitution by the estimator $\hat{\theta}$.

The independence of step (3) for variance computation from the linearization in steps (1) and (2) is aptly demonstrated by the discussion on two-phase sampling in section 4. In most cases, one assumes with replacement sample design to estimate the variance of the total in the step (3). Of course, a better estimate of the variance of the total may be obtained by using all the available information about the sample design. For the case of a two-phase design, step (1) can be performed by using Horvitz Thompson weights for the phase one sampling, and treating the multipliers m_i as data. The multiplier m_i is equal to zero if the observation i is not selected in phase two and is equal to the inverse of the conditional probability $\pi_{2k/1}^{-1}$. The resulting step (2) produces the same total as presented in the paragraph between equations (4.3) and (4.4). The subsequent discussion in section 4, describes the appropriate way to estimate the variance of this total for a two-stage sample design without replacement at each stage, and that calculation is independent of the linearization.

The steps (1) and (2) generate appropriate linearization in all known cases except where the estimator is not a continuous function of the weights w_i , e.g., quantile.

1. Babubhai V. Shah, SAFAL Institute, Inc. E-mail: babushah@earthlink.net.

Comment

Chris Skinner¹

Linearization and replication approaches provide two broad classes of methods for variance estimation in surveys. Both have their relative advantages and it seems important to keep a place for both in the survey statistician's 'toolkit'. This paper deepens our understanding of linearization methods, proposes a general procedure to generate such variance estimators uniquely and provides valuable illustrations of this procedure in some important areas of application.

A linearization method approximates the variance of a statistic of interest by the variance of a linear statistic, for which it is assumed a suitable variance estimator is available. The main issue here is the method used to determine the linear statistic. The standard approach assumes the statistic of interest may be expressed as a differentiable function of a vector of linear statistics (of fixed dimension) and uses Taylor series expansion to determine the approximation. The approach proposed in this paper applies to a more general class of sample-weighted statistics, illustrated by the complex examples in sections 3.2. and 4. The variance estimator is constructed by differentiating the statistic with respect to the sample weights. The approach to linear approximation is closely related to methods based upon the influence function (*e.g.*, equations 1.6 and 1.13) and the paper provides a helpful review of such methods in section 1. The authors note that it is not easy to verify the validity of such methods for statistics which are not smooth functions of (or a fixed number of) linear statistics and it would be interesting to know how far the proposed approach does indeed provide valid variance estimators for statistics, such as quantiles, which are not of this form.

A key feature of the proposed approach, which ensures the unique construction of the variance estimator, is that derivatives are evaluated at values based on the achieved sample, without any initial evaluation of the approximating linear statistic at theoretical population values. Such initial evaluation may lead to non-uniqueness when auxiliary information is available, for example on a population mean, \bar{X} , and it is assumed that this value is equal to the limiting value of a corresponding sample statistic, \bar{x} . For statistics which are smooth functions of linear statistics, it appears that the variance estimator generated by the proposed method may also be constructed by conventional Taylor series methods, provided no initial simplification of the

variance estimator takes place based on such assumptions about auxiliary information. Such construction may, however, be less clear-cut than for the proposed approach.

Assumptions employed by linearization methods differing from the proposed approach, such as that an auxiliary value \bar{X} is the theoretical limiting value of a sample value \bar{x} , are based upon unconditional distributions and so it might be anticipated that the incorporation of such assumptions into a variance estimator might damage the method's conditional properties, especially with respect to statistics such as \bar{x} . The proposed procedure avoids dependence upon such assumptions and, by evaluating derivatives at achieved sample values, may be expected to track conditional properties more closely. (There appear to be parallels with Efron and Hinkley's (1978) arguments in favour of the observed versus the expected information, although the context is rather different.)

The avoidance of dependence upon such assumptions may not only benefit the conditional properties of the proposed approach, but also protect the variance estimator against possible biasing effects of non-sampling errors. The auxiliary population information may differ from the limiting values of the corresponding sample statistics either because of non-response or non-coverage or because of discrepancies in the way the auxiliary variables are measured. In such circumstances, linearization methods differing from the proposed approach might lead to inconsistent variance estimation. For this reason, Fuller (2002, page 10) recommends the use of the g -weights in (3.6), as proposed, especially in the presence of nonresponse (page 15). With regards to the latter case, it seems worth noting that the validity of the proposed procedure does not appear to depend on the requirement that $E(\mathbf{d}(s)) = \mathbf{1}$, provided $\mathbf{1}$ is replaced by $E(\mathbf{d}(s))$ in the development in section 2. In particular, if s denotes unit respondents and non-response may be represented by Poisson sampling with unknown response probabilities then the proposed approach to variance estimation may still be consistent (when based on many standard variance estimators for linear statistics), even if $d(s)$ is based only on sampling inclusion probabilities.

Julia d'Arrigo and I have recently studied the properties of linearization variance estimators under nonresponse in simulation studies as part of the DACSEIS research project (www.dacseis.de) using data from the UK Labour Force

1. Chris Skinner, Southampton Statistical Sciences Research Institute, University of Southampton, Southampton S017 1BJ, United Kingdom. E-mail: cjs@soecsci.soton.ac.uk.

Survey and the German Income and Expenditure Survey. We considered various calibration estimators under Poisson models for unit non-response which were ignorable given the calibrating variables, using standard variance estimators for linear statistics under stratified multi-stage sampling. We indeed found that nonresponse could lead to serious biases in the linearization variance estimators if they failed to take account of the g -weights for GREG estimation (section 3.1.) or ignored the $F(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}})$ term in (3.21). Such biases were absent in the proposed approach.

We also investigated the alternative calibration estimators discussed in section 3.4. Deville and Särndal's (1992) theoretical finding that the asymptotic variance of \hat{Y}_w does not depend on the form of the function $F(\cdot)$ is based on the assumption that $\sum d_k(s) \mathbf{x}_k$ is consistent for \mathbf{X} . This assumption may not hold under various sources of non-sampling error, and is not required for the proposed approach. Hence, the appropriate approximate linear statistic (under departures from this assumption) is defined by (3.21) and the resulting variance estimator may depend on the form of $F(\cdot)$, even asymptotically. The standard linearization variance estimators in which $d_k(s) f(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}})$ in $\hat{\mathbf{B}}_\lambda$ is replaced by $d_k(s)$ or $w_k(s)$ may be inconsistent if these weights differ from $d_k(s) f(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}})$. Despite this theoretical fact, we observed little difference in our simulation study (for each of the functions, $1+u$, $\exp(u)$, and $(1-u)^{-1}$, used for $F(u)$) between the statistical properties of variance estimators based upon these three different choices of weight, $d_k(s) f(\mathbf{x}_k^T \hat{\boldsymbol{\lambda}})$, $d_k(s)$ or $w_k(s)$, in the $\hat{\mathbf{B}}_\lambda$ vector in (3.21). Others studies might produce different findings.

A disadvantage of the linearization methods considered here compared to replication methods is the need for analytic differentiation. It would appear from the examples presented in this paper that the analytic differentiation involved in the proposed method is at least as straightforward as that in standard methods of Taylor series expansion of smooth functions of linear statistics. Nevertheless, in some applications, it may be advantageous to replace the human labour and possible human error arising with analytic differentiation by the use of 'numerical differentiation'. The proposed approach might be described as an *infinitesimal jackknife* method since it perturbs the weight given to each sample observation by an infinitesimal amount to determine the approximating linear statistic. The derivative with respect to a weight in the proposed approach may be approximated numerically by a finite difference approach in which the statistic is recalculated with the weight perturbed by a finite amount for each observation in turn. This approach may be described as a *jackknife* method of linearization. A conventional approach would be to change each weight to zero in turn, perhaps standardizing

for unequal weights as in (1.15). It does not seem essential to replace the original weight by zero and, in principle, each weight might be perturbed in some other way, for example by reducing it by a fixed amount δ , smaller than the minimum value of $d_k(s)$. It seems likely that in many applications the variance estimator arising from such jackknife linearization will have very similar statistical properties to that constructed by the proposed approach. The choice between the estimators is likely to depend more on practical and computational considerations.

My final comments are on terminology. There are practical reasons why it may be helpful to give the z_k variable a name. In particular, this may be helpful for the practitioner who, for some complex statistics, has to employ two separate computational steps: (a) construction of the z_k variable, for example using least squares routines when calibration weighting is used, and (b) use of standard variance estimation software for linear statistics. Different names are used for z_k in the literature. Woodruff (1971) is usually acknowledged as the first paper in the survey sampling literature to draw attention to the role of z_k and Andersson and Nordberg (1994) refer to z_k as the *Woodruff transformation*. Woodruff and Causey (1976) refer to the approximating linear statistic as the *linear substitute* and z_k as the *substitute variable*. In the more mainstream statistical literature, Davison and Hinkley (1997, page 46) refer to the z_k as the *empirical influence values*. The term *linearized variable*, as used by Deville (1999), seems to me a simple and natural one. It is consistent with the use of the term *linearized statistic* to denote the approximating linear statistic and the term *linearization* for the method (which is a more suitable general term than Taylor series method for the broad class of approaches considered here).

Additional References

- Davison, A.C., and Hinkley, D.V. (1997). *Bootstrap Methods and their Applications*. Cambridge: Cambridge University Press.
- Efron, B., and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information (with discussion). *Biometrika*, 65, 457-487.
- Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-23.
- Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.
- Woodruff, R.S., and Causey, B.D. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 71, 315-321.

Response from the Authors

1. Introduction

We thank the three discussants, Phillip Kott, Babubhai Shah and Chris Skinner, for their insightful comments. Our rejoinder will attempt to address some of the issues raised by the discussants. The main aim of our paper was to study variance estimation for calibration estimators of population totals and nonlinear parameters, θ , defined as solutions to “census” estimating equations. We proposed a new Taylor linearization approach that provides a unique variance estimator, by avoiding initial evaluation of the linearized statistic at the population values. We have also shown that the variance estimator satisfies some desirable considerations, such as approximate model unbiasedness and validity under a conditional repeated sampling frame work, at least in a number of important cases. We have also shown that in two-phase sampling the variance estimator makes fuller use of the first phase sample data compared to traditional linearization variance estimators.

Kott

Kott’s discussion focused on three applications in our paper: (i) the jackknife linearization variance estimator, v_{JL} , of the ratio estimator $\hat{Y}_R = (\bar{y}/\bar{x})X$ in simple random sampling mentioned in section 1; (ii) the general class of regression calibration weights considered in section 3.2; (iii) the general class of calibration weights studied in section 3.4. Regarding (i), we noted the result that v_{JL} is both asymptotically design unbiased and approximately model unbiased under the ratio model $E_m(y_k) = \beta x_k$ and $V_m(y_k) = \sigma^2 x_k$. Kott is correct in saying that the model bias may not be negligible if the sampling fraction, n/N , is not small. If n/N is “ignorably small”, then model unbiasedness is, in fact, valid under a general variance function $V_m(y_k) = \sigma_k^2$, as noted by Kott and previously by Särndal *et al.* (1989). Under the ratio model, Kott proposes a more appropriate variance estimator, v_m , that is model unbiased even if n/N is not small and also valid under repeated sampling. The leading terms of v_m and v_{JL} are identical, and our new approach captures only the leading term. It should be noted that model-unbiasedness of v_m depends on the validity of the assumption $\sigma_k^2 = \sigma^2 x_k$.

Turning to (ii), we have shown in section 3.2 that if the general class of regression calibration weights, (3.7), are used, our approach leads to a variance estimator that is quite complex, involving third and fourth order moments of the design weights $d_k(s)$ with $d_k(s) = 0$ if the k^{th} population element is not in the sample s . Kott proposes an attractive choice of weights obtained by replacing $c_k x_k$ in the GREG

weight (3.1) with $q_{(1)k} = \sum_{l=1}^N (\pi_{kl} - \pi_k \pi_l) x_l / (\pi_k \pi_l)$. This choice gives a variant of the “optimal” linear regression estimator and also avoids the complexities associated with the variance estimator based on the weights (3.7). This is an interesting and useful proposal, but $q_{(1)k}$ requires the knowledge of the x -vector for all the population elements, unlike (3.7) which depends only on the population total X ; in practice, only X may be available. Moreover, $q_{(1)k}$ depends on all the $N(N-1)/2$ joint inclusion probabilities π_{kl} and hence computation of $q_{(1)k}$ may become cumbersome when the sampling design is based on unequal probability sampling without replacement.

Turning to (iii), Kott proposes a generalization of the calibration weights $w_k(s) = d_k(s)F(x_k^T \hat{\lambda})$ in section 3.4 by replacing x_k with “instrumental” variables q_k having the same dimension as x_k . The corresponding z -variable in the variance estimator $v(z)$ is similar to our (3.21) with $x_k x_k^T$ and $x_k y_k$ in \hat{B}_λ changed to $q_k x_k^T$ and $q_k y_k$ respectively and $F(x_k^T \hat{\lambda})$ changed to $F(q_k^T \hat{\lambda})$. This is an useful extension. Kott notes that \hat{B}_λ remains a model unbiased estimation of B_λ if $f(q_k^T \lambda)$ in \hat{B}_λ is replaced by any constant and the resulting z_k is unaffected asymptotically under repeated sampling. However, Kott also notes that the term $f(q_k^T \lambda)$ can matter even asymptotically if the calibration is used to adjust for unit nonresponse by treating sample response as a second phase of sampling. Using the result for two-phase sampling given in the Appendix, Kott then obtains a corresponding variance estimator, $v(\hat{Y}_{GC})$. This extension for nonresponse setting is also useful. It is indeed surprising that the second term in $v(\hat{Y}_{GC})$ provides the model based correction he recommended for the ratio estimator \hat{Y}_R under simple random sampling in the absence of nonresponse.

Finally, Kott raises a question on the customary “plug-in” or “substitution” method used for variance estimation, as done in (2.4), where we plug in z_k for \tilde{z}_k . He asks if it is legitimate to plug in z_k for \tilde{z}_k , where $z_k - \tilde{z}_k = O_p(1/\sqrt{n})$, when they are $n(n-1)/2$ terms in the variance estimator $v(\tilde{z}_k)$, as in the case of Sen-Yates-Grundy variance estimator. We are not sure if we have understood his point correctly, but as long as $O_p(1/\sqrt{n})$ is uniform in k , say a/\sqrt{n} , then $v(z) = v(\tilde{z}) + \text{lower order terms}$.

Shah

Shah’s prescription (steps 1–4) clearly summarizes our method. Shah also notes that his steps 1 and 2, leading to our z -variable, produces the “correct” linearization in many other important applications not studied in our paper,

including Wilcoxon two sample rank sum test and estimation of regression coefficients and hazard rate in the Cox proportional hazard model. Shah’s unpublished paper (seen by courtesy of the author) spells out the z – variable for those applications, but using design weights. Extension to calibration weights should follow along the lines of section 3.

Shah makes an important point that step 3 for the computation of the variance estimate is independent of the linearization in step 1 and 2 and that it is “aptly demonstrated by the discussion on two-phase sampling in section 4”. He also notes that for two-phase sampling, linearization (step 1) can be performed using only the first-phase H-T weights π_{1k}^{-1} , by treating the second phase weights, $\pi_{2k/1}^{-1}$, if $k \in s$ and 0 if k is not in the second-phase sample s as data, and that the resulting step 2 produces the same approximation as given in our paper. We have verified this equivalence result for the two-phase ratio estimator in Example 4.1, and it is likely to hold generally. Shah’s proposal might simplify the implementation of step 1 to some extent.

Skinner

Skinner gives a clear appraisal of our linearization method and raises a number of important points: (i) terminology, (ii) possible extensions to non-smooth statistics such as quantiles, (iii) modifications of the method to handle unit nonresponse, (iv) possible use of numerical differentiation to calculate the z_k – variables.

With regard to point (i), Skinner notes that it would be useful to give the z_k variable a name since different names have been used in the literature. He suggests that the term *linearized variable*, as used by Deville (1999), is a simple and natural one since it is consistent with the usage of *linearized statistic* to denote the approximating linear statistic and linearization for the method. We are in agreement with Skinner’s suggestion.

Turning to point (ii), a difficulty in extending our proposal to nonsmooth statistics $\hat{\theta} = f(\mathbf{d}(s))$, such as quantiles, is that $f(\cdot)$ is not a differentiable function. A way to get around this difficulty is to approximate $\hat{\theta} - \theta$ by a differentiable function and then apply our method to the approximation. For example, in the case of the p^{th} quantile θ , Francisco and Fuller (1991) and Shao (1991) established the following asymptotic approximation valid for stratified multistage designs:

$$\theta - \theta \approx - \frac{1}{h(\theta)} \{ \hat{F}_w(\theta) - p \},$$

where $\hat{F}_w(\theta) = \sum w_k(s) I(y_k \leq \theta) / \sum w_k(s)$ is the calibration estimator of the distribution function $F(\cdot)$ at θ , $F(\theta) = N^{-1} \sum I(y_k \leq \theta) = p$, and $h(\theta)$ is the value of the density function $h(\cdot)$ at θ . The definition of $h(\cdot)$

requires reference to a sequence of populations (Shao and Rao 1993) or to a superpopulation (Francisco and Fuller 1991). We used $h(\cdot)$ to denote the density rather than the customary $f(\cdot)$ because we used $f(\mathbf{d}(s))$ to denote the estimator $\hat{\theta}$. Now, suppose $w_k(s) = d_k(s) g_k(\mathbf{d}(s))$, where $g_k(\mathbf{d}(s))$ is the GREG weight given by (3.1). We can then use (3.2) and (3.3) to get the linearized variable z_k from the above approximation to $\hat{\theta} - \theta$, by replacing $h(\theta)$ with a suitable estimator $\hat{h}(\hat{\theta})$; for example the kernel-based estimator of $h(\cdot)$ used by Berger and Skinner (2003). Similarly, one can apply the method to general calibration weights, $w_k(s)$, using the results of section 4. Variance estimators of a low income proportion, say $\theta = F(\tau/2)$ where τ is the median income, can also be obtained using the asymptotic approximation for $\hat{\theta} - \theta$ developed by Shao and Rao (1993). Berger and Skinner (2003) studied variance estimation for a low income proportion when generalized raking ratio weights, $w_k(s)$, are used. We can apply the results in section 3.2 to this case, and the resulting linearized variable z_k will account for the calibration. Also, it will be different from the Deville z – variable (10) in Berger and Skinner (2003).

The modification suggested in point (iii) to handle unit nonresponse is very important, and it broadens the applicability of our method. As noted by Skinner, Kott and Fuller (2002), it is important to retain the g – weights in variance estimation whenever the limiting values of the estimators \hat{X} differ from the corresponding control totals X , as in the case of non-response or non-coverage. Our method automatically accounts for the g – weights and may lead to consistent variance estimators in such cases. Empirical results of Skinner with d’Arrigo in this context are very interesting. The case of variance estimators for alternative calibration estimators, studied in section 3.4, relative to customary variance estimators that replace $d_k(s) f(\mathbf{x}_k^T \hat{\lambda})$ in the expression for \hat{B}_λ by $d_k(s)$ or $w_k(s)$ need further study, as noted by Skinner.

It may be noted that unit nonresponse is typically treated as second phase sampling (e.g., Poisson sampling with unknown response probabilities) and Skinner notes that our method may lead to consistent variance estimators even when the estimators are based only on the sampling inclusion probabilities. However, control totals X are needed to get valid estimators of the total Y , under some assumptions on the response probabilities (Fuller 2002, equation (8.4)). We have extended our method to handle weight adjustment for unit nonresponse and imputation for item nonresponse when control totals are not available, assuming uniform response within classes (Demnati and Rao 2002). The resulting variance estimators are naturally more complex compared to Skinner’s modification for unit nonresponse in the presence of control totals.

Turning to point (iv) on the possible use of numerical differentiation to calculate the linearized variables z_k , Woodruff and Causey (1976) used such a method to calculate the derivatives $\partial g(\mathbf{a})/\partial a_i|_{a=\hat{y}}$ given in (1.4) when $\hat{\theta} = g(\hat{Y})$. Skinner proposes perturbing each weight $d_k(s)$ in turn and then recalculating $\hat{\theta}$; for example, by replacing it by a fixed amount δ than the minimum value of $d_k(s), k \in s$. He conjectures that the proposed approach should lead to variance estimators very similar to those obtained through analytical differentiation. It would be useful to study the statistical properties of the proposed approach to analytic differentiation of $f(\mathbf{d}(s))$ with respect to weights $d_k(s)$.

We hope the discussions by Kott, Shah and Skinner will stimulate further work on the approach to variance estimation presented in our paper.

References

- Berger, Y.G., and Skinner, C.J. (2003). Variance estimation for a low income proportion. *Applied Statistics*, 52, 457-468.
- 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, 736-740.
- Francisco, C.A., and Fuller, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.
- Shao, J. (1991). L-statistics in complex problems. Technical Report, University of Ottawa, Ottawa.
- Shao, J., and Rao, J.N.K. (1993). Standard errors for low income proportions estimated from stratified multistage samples. *Sankhyā*, Series B, 55, 393-414.
- Woodruff, R.S., and Causey, B.D. (1976). Computerized method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 71, 315-321.