VARIANCE ESTIMATION FOR THE GENERALIZED REGRESSION ESTIMATOR UNDER TWO-PHASE SAMPLING - A MODIFIED APPROACH

Martin Axelson, University of Örebro
Department of Statistics, University of Örebro, SE-701 82 Örebro, Sweden

KEY WORDS: Regression estimator; Two-phase sampling; Finite population; Linearization; Variance estimation;

1 Introduction

Generalized regression estimation under two-phase sampling is a cost-effective and efficient technique for estimation of a finite population total. Important references are Särndal and Swensson (1987) and Särndal, Swensson, and Wretman (1992, section 9.7), where a thorough summary of generalized regression estimation under two-phase sampling is given, and general results, covering all possible combinations of available auxiliary information, are given for the generalized regression estimator (*GREG*).

Even though all of the available auxiliary information is used for estimation of the total, the two variance estimators suggested by Särndal et al. (1992, p. 362), referred to as the reference variance estimators in the sequel, are primarily based on second-phase sample information. Thus, it seems reasonable to ask whether the auxiliary information available for the elements not included in the second-phase sample could, and if so, should, be used more extensively for variance estimation as well. This is not a new question. An early reference is Cochran (1953, section 12.7), while more recent references are Dorfman (1994), Rao and Sitter (1995), Axelson, Breidt, and Carriquiry (1996), and Sitter (1997).

In this paper a new approach to variance estimation using linearization techniques is proposed. The new approach is general, in that it allows for arbitrary sampling designs in each of the two phases, in combination with any possible set-up of available auxiliary information. Compared to the reference estimators, this new approach makes more extensive use of the available auxiliary information. If the available auxiliary information can be utilized to obtain reasonably accurate predictions of the study variable, the new approach is expected to result in less variable variance estimators than the reference alternatives. The paper is concluded with a few examples, that relate the new approach to variance estimation to results given by Rao and Sitter (1995), Axelson et al. (1996), and Sitter (1997).

2 Two-phase sampling

Let $U = \{1, \ldots, k, \ldots, N\}$ denote a finite population. Associated with each element $k \in U$ is a value y_k, where y denotes the study variable. The parameter of interest is the population total of y, i.e. $t_y = \sum_{k \in U} y_k = \sum_U y_k$. To estimate t_y, we will use information collected through a two-phase sampling procedure.

A first-phase sample s_a, of size n_{s_a}, is drawn from U according to a design denoted $p_a(\cdot)$. The first-phase first- and second-order inclusion probabilities are denoted π_{ak} and π_{akl} respectively. Given s_a, a second-phase sample s, of size n_s, is drawn from s_a according to a design denoted $p(\cdot|s_a)$. The second-phase first- and second-order inclusion probabilities are denoted $\pi_{k|s_a}$ and $\pi_{kl|s_a}$ respectively. By assumption, $\pi_{akl} > 0$ for all $k\&l \in U$, and $\pi_{kl|s_a} > 0$ for all $k\&l \in s_a$. For the elements included in the second-phase sample, the value of the study variable is obtained, i.e., ultimately, y_k is known for all $k \in s$.

To simplify expressions derived in subsequent sections, set $\Delta_{akl} = \pi_{akl} - \pi_{ak}\pi_{al}$ and $\Delta_{kl|s_a} = \pi_{kl|s_a} - \pi_{k|s_a}\pi_{l|s_a}$. Moreover, let the symbol $\check{}$ symbolize division by π_{ak}, and, in analogous manner, let the symbol $\check{}$ symbolize division by $\pi_{ak}\pi_{k|s_a}$. Hence, for example, $\check{y}_k = y_k/\pi_{ak}$, which is defined for all $k \in U$, and $\check{\check{y}}_k = \check{y}_k/\pi_{k|s_a} = y_k/(\pi_{ak}\pi_{k|s_a})$, which is defined for all $k \in s_a$.

3 The generalized regression estimator under two-phase sampling

Assume that in addition to y_k, there is also a vector $\mathbf{x}_k$ of J auxiliary values associated with each element $k \in U$. In the *GREG*, the auxiliary information available at the element level ultimately serves to get predicted values of the study variable. Following Särndal et al. (1992, section 9.7), we partition $\mathbf{x}_k$ as $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$, where $\mathbf{x}_{1k}$ denotes a $J_1 \times 1$-vector comprised of auxiliary values known for all $k \in U$ beforehand, while $\mathbf{x}_{2k}$ denotes a $(J - J_1) \times 1$-vector comprised of auxiliary values unknown at the onset of the study. In the sequel, it is assumed

that $\mathbf{x}_k$ and $\mathbf{x}_{1k}$ differ, and that the very purpose of the first-phase sample is to obtain $\mathbf{x}_{2k}$ for all $k \in s_a$. Typically, $\mathbf{x}_2$ is chosen because it is assumed to be a strong, yet relatively inexpensive, predictor for y, whereas $\mathbf{x}_1$ often primarily is of administrative character and therefore can be assumed to be weaker than $\mathbf{x}_2$ as a predictor for y.

In its most general form, the generalized regression estimator can be written as

$$\begin{aligned}\hat{t}_r = \hat{t}_{y\pi^*} &+ \left(\mathbf{t}_{x_1} - \hat{\mathbf{t}}_{x_1\pi}\right)' \widehat{\mathbf{B}}_{1s} \\ &+ \left(\hat{\mathbf{t}}_{x\pi} - \hat{\mathbf{t}}_{x\pi^*}\right)' \widehat{\mathbf{B}}_s,\end{aligned} \tag{1}$$

where $\hat{t}_{y\pi^*} = \sum_s \breve{y}_k$ is unbiased for t_y, $\mathbf{t}_{x_1} = \sum_U \mathbf{x}_{1k}$, $\hat{\mathbf{t}}_{x_1\pi} = \sum_{s_a} \breve{\mathbf{x}}_{1k}$ is unbiased for $\mathbf{t}_{x_1}$, $\hat{\mathbf{t}}_{x\pi}$ is analogous to $\hat{\mathbf{t}}_{x_1\pi}$, $\hat{\mathbf{t}}_{x\pi^*}$ is analogous to $\hat{t}_{y\pi^*}$, $\widehat{\mathbf{B}}_{1s} = \widehat{\mathbf{M}}^{-1}_{\mathbf{x}_1\mathbf{x}_1 s}\widehat{\mathbf{M}}_{\mathbf{x}_1 y s}$, and $\widehat{\mathbf{B}}_s = \widehat{\mathbf{M}}^{-1}_{\mathbf{x}\mathbf{x}s}\widehat{\mathbf{M}}_{\mathbf{x}ys}$, with $\widehat{\mathbf{M}}_{\mathbf{x}_1\mathbf{x}_1 s} = \sum_s \breve{\mathbf{x}}_{1k}\mathbf{x}'_{1k}/w_{1k}$, $\widehat{\mathbf{M}}_{\mathbf{x}_1 y s} = \sum_s \breve{\mathbf{x}}_{1k} y_k/w_{1k}$, $\widehat{\mathbf{M}}_{\mathbf{x}\mathbf{x}s} = \sum_s \breve{\mathbf{x}}_k\mathbf{x}'_k/w_k$, and $\widehat{\mathbf{M}}_{\mathbf{x}ys} = \sum_s \breve{\mathbf{x}}_k y_k/w_k$. In $\widehat{\mathbf{B}}_{1s}$ and $\widehat{\mathbf{B}}_s$, w_{1k} and w_k, respectively, are weights, pre-specified by the statistician, that reflect the relative importance, based on a priori knowledge, assigned to element k by the survey statistician.

It is not uncommon that no auxiliary information is available at the onset of the study, i.e. $\mathbf{x} = \mathbf{x}_2$. Under these circumstances, all terms involving $\mathbf{x}_1$ drop out of (1).

4 The variance of $\hat{t}_r$

Clearly, since $\widehat{\mathbf{B}}_{1s}$ and $\widehat{\mathbf{B}}_s$ are stochastic, $\hat{t}_r$ is not unbiased for t_y. However, using a linearized version of the *GREG*, it is possible to conclude (see, e.g., Särndal and Swensson, 1987) that $\hat{t}_r$ is approximately unbiased for t_y, provided that the sample size is large in each of the two phases.

Let $E_{1kU} = y_k - \mathbf{x}'_{1k}\mathbf{B}_{1U}$ and $E_{ks_a} = y_k - \mathbf{x}'_k\widehat{\mathbf{B}}_{s_a}$, where $\mathbf{B}_{1U} = \mathbf{M}^{-1}_{\mathbf{x}_1\mathbf{x}_1 U}\mathbf{M}_{\mathbf{x}_1 y U}$ with $\mathbf{M}_{\mathbf{x}_1\mathbf{x}_1 U} = \sum_U \mathbf{x}_{1k}\mathbf{x}'_{1k}/w_{1k}$ and $\mathbf{M}_{\mathbf{x}_1 y U} = \sum_U \mathbf{x}_{1k} y_k/w_{1k}$, and $\widehat{\mathbf{B}}_{s_a} = \widehat{\mathbf{M}}^{-1}_{\mathbf{x}\mathbf{x}s_a}\widehat{\mathbf{M}}_{\mathbf{x}ys_a}$ with $\widehat{\mathbf{M}}_{\mathbf{x}\mathbf{x}s_a} = \sum_{s_a} \breve{\mathbf{x}}_k\mathbf{x}'_k/w_k$ and $\widehat{\mathbf{M}}_{\mathbf{x}ys_a} = \sum_{s_a} \breve{\mathbf{x}}_k y_k/w_k$. Hence, E_{1kU} is defined for $k \in U$, while E_{ks_a} is defined for all $k \in s_a$. Särndal and Swensson (1987) show that the variance of $\hat{t}_r$, which can be written as

$$V(\hat{t}_r) = V_1 + V_2, \tag{2}$$

where $V_1 = V_{p_a}[E(\hat{t}_r|s_a)]$ and $V_2 = E_{p_a}[V(\hat{t}_r|s_a)]$, is well approximated by

$$AV(\hat{t}_r) = AV_1 + AV_2, \tag{3}$$

where $AV_1 = \sum\sum_U \Delta_{akl}\breve{E}_{1kU}\breve{E}_{1lU}$ and $AV_2 = E_{p_a}\left(\sum\sum_{s_a}\Delta_{kl|s_a}\breve{\breve{E}}_{ks_a}\breve{\breve{E}}_{ls_a}\right)$, where $\sum\sum_U$ is short for $\sum_{k\in U}\sum_{l\in U}$ and $\sum\sum_{s_a}$ is analogous to $\sum\sum_U$.

5 Estimation of the variance of $\hat{t}_r$

5.1 The standard approach

As an estimator for $V(\hat{t}_r)$ in (2), Särndal et al. (1992, section 9.7) suggest

$$\hat{V}_o^{g_1;g} = \hat{V}_{1o}^{g_1} + \hat{V}_{2o}^{g}, \tag{4}$$

where

$$\hat{V}_{1o}^{g_1} = \sum\sum_s \frac{\Delta_{akl}}{\pi_{akl}\pi_{kl|s_a}} g_{1ks_a}\check{e}_{1k}g_{1ls_a}\check{e}_{1l} \tag{5}$$

and

$$\hat{V}_{2o}^{g} = \sum\sum_s \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}} g_{ks}\check{e}_k g_{ls}\check{e}_l, \tag{6}$$

where e_{1k} and e_k are the sample-based counterparts to E_{1kU} and E_{ks_a} respectively,

$$g_{ks} = 1 + \left(\hat{\mathbf{t}}_{x\pi} - \hat{\mathbf{t}}_{x\pi^*}\right)' \widehat{\mathbf{M}}^{-1}_{\mathbf{x}\mathbf{x}s}\mathbf{x}_k/w_k^2,$$

and

$$g_{1ks_a} = 1 + \left(\mathbf{t}_{x_1} - \hat{\mathbf{t}}_{x_1\pi}\right)' \widehat{\mathbf{M}}^{-1}_{\mathbf{x}_1\mathbf{x}_1 s_a}\mathbf{x}_{1k}/w_{1k}^2,$$

where $\widehat{\mathbf{M}}_{\mathbf{x}_1\mathbf{x}_1 s_a}$ is analogous to $\widehat{\mathbf{M}}_{\mathbf{x}\mathbf{x}s_a}$. This estimator follows from an extension of the g-weighted residual technique for one-phase sampling, which was proposed by Särndal (1982) and further elaborated in Särndal, Swensson, and Wretman (1989). A simplified estimator for $V(\hat{t}_r)$ is given by

$$\hat{V}_o^{1;1} = \hat{V}_{1o}^{1} + \hat{V}_{2o}^{1}, \tag{7}$$

where $\hat{V}_{1o}^{1}$ and $\hat{V}_{2o}^{1}$ follow from (5) and (6) respectively, by setting all g-weights equal to unity.

The estimators $\hat{V}_o^{g_1;g}$ and $\hat{V}_o^{1;1}$ are both known to have acceptable large sample properties, in that they both are approximately design unbiased and yield satisfactory results when used for construction of confidence intervals. However, considering conditional inference, as advocated by, for example, Holt and Smith (1979) and Rao (1985), we expect $\hat{V}_o^{g_1;g}$ to be the better estimator, in analogy with the findings regarding one-phase sampling in Särndal et al. (1989).

Despite the fact that $\hat{V}_o^{g_1;g}$ and $\hat{V}_o^{1;1}$ have acceptable large sample properties, the variability of the estimators may be unduly large. In Section 5.2, we propose a new approach to estimation of V_1 which makes more complete use of the observed sample information. If the available auxiliary information has strong predictive power, we expect the new approach, when combined with either $\hat{V}_{2o}^{g}$ or $\hat{V}_{2o}^{1}$, to yield estimators for $V(\hat{t}_r)$ that are more efficient than either of $\hat{V}_o^{g_1;g}$ and $\hat{V}_o^{1;1}$.

5.2 The new approach

In the search for a new estimator for V_1, we start by considering a variance estimator, implicitly suggested by the term AV_1 in (3), that could be used if y_k was recorded for $k \in s_a$. That is, using the g-weighted technique with $E_{1ks_a} = y_k - \mathbf{x}'_{1k}\widehat{\mathbf{B}}_{1s_a}$, where $\widehat{\mathbf{B}}_{1s_a}$ is analogous to $\widehat{\mathbf{B}}_{s_a}$,

$$\hat{V}^{g_1}_{1,hyp;s_a} = \sum\sum_{s_a} \frac{\Delta_{akl}}{\pi_{akl}} g_{1ks_a}\check{E}_{1ks_a} g_{1ls_a}\check{E}_{1ls_a} \quad (8)$$

will serve as the starting point in the search for an alternative estimator for V_1. Since $E_{1ks_a} = D_{ks_a} + E_{ks_a}$, where $D_{ks_a} = \mathbf{x}'_k\widehat{\mathbf{B}}_{s_a} - \mathbf{x}'_{1k}\widehat{\mathbf{B}}_{1s_a}$, (8) can be rewritten as

$$\hat{V}^{g_1}_{1,hyp;s_a} = \hat{C}^{g_1}_{1DDs_a} + 2\hat{C}^{g_1}_{1DEs_a} + \hat{C}^{g_1}_{1EEs_a}, \quad (9)$$

where

$$\hat{C}^{g_1}_{1DDs_a} = \sum\sum_{s_a} \frac{\Delta_{akl}}{\pi_{akl}} g_{1ks_a}\check{D}_{ks_a} g_{1ls_a}\check{D}_{ls_a}, \quad (10)$$

and $\hat{C}^{g_1}_{1DEs_a}$ and $\hat{C}^{g_1}_{1EEs_a}$ are obtained from (10) by substitution of, respectively, $\check{E}_{ls_a}$ for $\check{D}_{ls_a}$ and $\check{E}_{ks_a}$ for $\check{D}_{ks_a}$ *and* $\check{E}_{ls_a}$ for $\check{D}_{ls_a}$. By focusing on estimation of the conditional parameters $\hat{C}^{g_1}_{1DDs_a}$, $\hat{C}^{g_1}_{1DEs_a}$ and $\hat{C}^{g_1}_{1EEs_a}$, an alternative approach to estimation of V_1 is suggested.

Now consider two-phase sampling, with $\mathbf{x}_{1k}$ known for $k \in U$, $\mathbf{x}_k$ observed for $k \in s_a$, and y_k observed for $k \in s$. If $\widehat{\mathbf{B}}_{1s_a}$ and $\widehat{\mathbf{B}}_{s_a}$ were known, $\hat{C}^{g_1}_{1DDs_a}$ could easily be computed and conditionally unbiased estimators for $\hat{C}^{g_1}_{1DEs_a}$ and $\hat{C}^{g_1}_{1EEs_a}$ would be given by

$$\hat{C}^{g_1}_{1DEs} = \sum\sum_{s} \frac{\Delta_{akl}}{\pi_{akl}\pi_{kl|s_a}} g_{1ks_a}\check{D}_{ks_a} g_{1ls_a}\check{E}_{ls_a} \quad (11)$$

and $\hat{C}^{g_1}_{1EEs}$ respectively, where $\hat{C}^{g_1}_{1EEs}$ is obtained from $\hat{C}^{g_1}_{1DEs}$ by substituting $\check{E}_{ks_a}$ for $\check{D}_{ks_a}$. Hence,

$$\hat{V}^{g_1}_{1,hyp} = \hat{C}^{g_1}_{1DDs_a} + 2\hat{C}^{g_1}_{1DEs} + \hat{C}^{g_1}_{1EEs} \quad (12)$$

would be conditionally unbiased for $\hat{V}^{g_1}_{1,hyp;s_a}$.

Despite the fact that $\hat{V}^{g_1}_{1,hyp}$ is purely hypothetical, it nevertheless suggests an estimator of practical interest. Replacing the unobservable D_{ks_a} and E_{ks_a} in (12) with $d_k = \mathbf{x}'_k\widehat{\mathbf{B}}_s - \mathbf{x}'_{1k}\widehat{\mathbf{B}}_{1s}$, available for $k \in s_a$, and e_k, available for $k \in s$, we get, in obvious notation,

$$\hat{V}^{g_1}_{1a} = \hat{C}^{g_1}_{1dds_a} + 2\hat{C}^{g_1}_{1des} + \hat{C}^{g_1}_{1ees}. \quad (13)$$

Rather than replacing E_{ks_a} with e_k, one might consider using $g_{ks}e_k$, in analogy with using $\hat{V}^g_{2o}$ as an estimator for $V_2 \doteq AV_2$. Thus, an alternative estimator for V_1 is given by

$$\hat{V}^{g_1g}_{1a} = \hat{C}^{g_1}_{1dds_a} + 2\hat{C}^{g_1g}_{1des} + \hat{C}^{g_1g}_{1ees}, \quad (14)$$

where $\hat{C}^{g_1g}_{1des}$ and $\hat{C}^{g_1g}_{1ees}$ are obtained from, respectively, $\hat{C}^{g_1}_{1des}$ and $\hat{C}^{g_1}_{1ees}$ by substituting $g_{ks}e_k$ for e_k.

In deriving $\hat{V}^{g_1}_{1a}$ and $\hat{V}^{g_1g}_{1a}$, $\hat{V}^{g_1}_{1,hyp;s_a}$ in (9) served as the starting point. Thus, using

$$\hat{V}^{1}_{1hyp;s_a} = \sum\sum_{s_a} \frac{\Delta_{akl}}{\pi_{akl}} \check{E}_{1ks_a}\check{E}_{1ls_a}$$

as the starting point, rather than (9), the above approach results in yet two more estimators for V_1. Setting $g_{1ks_a} = 1$ for all $k \in s_a$ in (13) and (14), we get, respectively,

$$\hat{V}^{1}_{1a} = \hat{C}^{1}_{1dds_a} + 2\hat{C}^{1}_{1des} + \hat{C}^{1}_{1ees} \quad (15)$$

and

$$\hat{V}^{g}_{1a} = \hat{C}^{1}_{1dds_a} + 2\hat{C}^{g}_{1des} + \hat{C}^{g}_{1ees}. \quad (16)$$

Theoretically, in order to construct an estimator for $V(\hat{t}_r)$, any one of (13), (14), (15), and (16) can be combined with either $\hat{V}^g_{2o}$ in (4) or $\hat{V}^1_{2o}$ in (7). Thus, the new approach to estimation of V_1 leads to no less than eight possible estimators for $V(\hat{t}_r)$, all of which are given in Table 1 below.

At present, little is known about the performance of the suggested estimators, why further work is needed before any general recommendations can be made about which estimator to use in practice. However, by examining the simplified expressions for the two special cases when (i) $\pi_{ak} = 1$ for all $k \in U$ (i.e. $s_a = U$), and (ii) $\pi_{k|s_a} = 1$ for all $k \in s_a$ (i.e. $s{=}s_a$), and comparing them to the reference estimators given in (4) and (7), a group of four estimators emerges as more interesting than the remaining ones. Note that under (i) and (ii), we are simply studying generalized regression estimation under one-phase sampling, the auxiliary information being either (i) $\mathbf{x}$, or (ii) $\mathbf{x}_1$. Thus, it follows from (4) that $\hat{V}^{g_1;g}_o$ simplifies to $\hat{V}^g_{2o}$ under (i) and $\hat{V}^{g_1}_{1o}$ under (ii). Similarly, it follows from (7) that $\hat{V}^{1;1}_o$ simplifies to $\hat{V}^1_{2o}$ under (i) and $\hat{V}^1_{1o}$ under (ii). Thus, no matter if the auxiliary information used is $\mathbf{x}$ or $\mathbf{x}_1$, using $\hat{V}^{g_1;g}_o$ under one-phase sampling amounts to using the g-weighted residual technique for variance estimation, whereas using $\hat{V}^{1;1}_o$ amounts to using the simplified approach that completely ignores the g-weights. This feature of $\hat{V}^{g_1;g}_o$ and $\hat{V}^{1;1}_o$, which we choose to label *phase-compliance*, is attractive. As indicated, only four of the estimators in Table 1

are phase-compliant, namely $\hat{V}_a^{g_1;g}$, $\hat{V}_a^{g_1g;g}$, $\hat{V}_a^{1;1}$, and $\hat{V}_a^{g;1}$. Since phase-compliance is a desirable feature of any estimator for $V(\hat{t}_r)$ under two-phase sampling, we restrict our attention to these estimators in the sequel.

Estimator	Simplified expr.		Phase-compl.
	$\pi_{ak}=1$ for $k \in U$	$\pi_{k\|s_a}=1$ for $k \in s_a$	
$\hat{V}_a^{1;1} = \hat{V}_{1a}^1 + \hat{V}_{2o}^1$	$\hat{V}_{2o}^1$	$\hat{V}_{1o}^1$	yes
$\hat{V}_a^{1;g} = \hat{V}_{1a}^1 + \hat{V}_{2o}^g$	$\hat{V}_{2o}^g$	$\hat{V}_{1o}^1$	no
$\hat{V}_a^{g;1} = \hat{V}_{1a}^g + \hat{V}_{2o}^1$	$\hat{V}_{2o}^1$	$\hat{V}_{1o}^1$	yes
$\hat{V}_a^{g;g} = \hat{V}_{1a}^g + \hat{V}_{2o}^g$	$\hat{V}_{2o}^g$	$\hat{V}_{1o}^1$	no
$\hat{V}_a^{g_1;1} = \hat{V}_{1a}^{g_1} + \hat{V}_{2o}^1$	$\hat{V}_{2o}^1$	$\hat{V}_{1o}^{g_1}$	no
$\hat{V}_a^{g_1;g} = \hat{V}_{1a}^{g_1} + \hat{V}_{2o}^g$	$\hat{V}_{2o}^g$	$\hat{V}_{1o}^{g_1}$	yes
$\hat{V}_a^{g_1g;1} = \hat{V}_{1a}^{g_1g} + \hat{V}_{2o}^1$	$\hat{V}_{2o}^1$	$\hat{V}_{1o}^{g_1}$	no
$\hat{V}_a^{g_1g;g} = \hat{V}_{1a}^{g_1g} + \hat{V}_{2o}^g$	$\hat{V}_{2o}^g$	$\hat{V}_{1o}^{g_1}$	yes

Table 1: Simplified expressions for $\hat{V}_a^{1;1}$, $\hat{V}_a^{1;g}$, $\hat{V}_a^{g;1}$, $\hat{V}_a^{g;g}$ $\hat{V}_a^{g_1;1}$, $\hat{V}_a^{g_1;g}$, $\hat{V}_a^{g_1g;1}$, and $\hat{V}_a^{g_1g;g}$ when (i) all $\pi_{ak}=1$ for all $k \in U$, and (ii) all $\pi_{k|s_a}=1$ for all $k \in s_a$.

6 Discussion

As mentioned in Section 5, the new approach to variance estimation makes more complete use of the observed sample information. Therefore, it should result in estimators for $V(\hat{t}_r)$ that are more efficient than either of $\hat{V}_o^{g_1;g}$ and $\hat{V}_o^{1;1}$, given reasonable circumstances, i.e. circumstances such that two-phase regression estimation is preferred to ordinary double expansion estimation. To exemplify, let us compare $\hat{V}_a^{1;1} = \hat{V}_{1a}^1 + \hat{V}_{2o}^1$ with $\hat{V}_o^{1;1} = \hat{V}_{1o}^1 + \hat{V}_{2o}^1$ in terms of variances.

EXAMPLE 1. We start by noting that

$$V\left(\hat{V}_a^{1;1}\right) - V\left(\hat{V}_o^{1;1}\right) = V\left(\hat{V}_{1a}^1\right) - V\left(\hat{V}_{1o}^1\right) + 2C\left(\hat{V}_{1a}^1 - \hat{V}_{1o}^1, \hat{V}_{2o}^1\right) \quad (17)$$

Now, given the relation $e_{1k} = d_k + e_k$, it follows from (5), setting all g-weights to unity, and (15), that $\hat{V}_{1a}^1$ can be rewritten as

$$\hat{V}_{1a}^1 = \hat{C}_{1ddds_a}^1 - \hat{C}_{1dds}^1 + \hat{V}_{1o}^1, \quad (18)$$

where $\hat{C}_{1dds}^1$ is analogous to $\hat{C}_{1ees}^1$. Under reasonable circumstances, we expect $d_k \doteq e_{1k}$ for a majority of $k \in s_a$, regardless of the particular second-phase sample obtained. This implies that conditionally on s_a, the term $\hat{C}_{1ddds_a}^1 - \hat{C}_{1dds}^1$ in $\hat{V}_{1a}^1$ will serve as an adjustment term, calibrating $\hat{V}_{1o}^1$ towards its conditional expected value and thus reducing its conditional variance. Consequently, since $E\left(\hat{V}_{1a}^1|s_a\right) \doteq E\left(\hat{V}_{1o}^1|s_a\right)$, we conclude from (18) that $V\left(\hat{V}_{1a}^1\right) - V\left(\hat{V}_{1o}^1\right) < 0$ under circumstances deemed reasonable from a practitioner's point of view. Furthermore, since $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')'$ is a stronger predictor than $\mathbf{x}_1$ by assumption, using arguments from standard regression theory, we expect e_{1k} to be larger than e_k, in absolute value, for most $k \in s$. Moreover, we expect a majority of the paired residuals e_{1k} and e_k to coincide in terms of their sign. Indirectly, this indicate that $\hat{V}_{1o}^1$ and $\hat{V}_{2o}^1$ should be positively correlated under circumstances deemed reasonable from a practitioner's point of view, since the observed estimate of $\hat{V}_{1o}^1$ in a sense imposes a restriction on the possible range of outcomes of $\hat{V}_{2o}^1$. Since $\hat{V}_{1a}^1$ to a large extent is based on information observed for $k \in s_a - s$, we expect the correlation between $\hat{V}_{1a}^1$ and $\hat{V}_{2o}^1$ to be negligible. Hence, since $E\left(\hat{V}_{1a}^1|s_a\right) \doteq E\left(\hat{V}_{1o}^1|s_a\right)$, we expect $C\left(\hat{V}_{1a}^1 - \hat{V}_{1o}^1, \hat{V}_{2o}^1\right) \leq 0$ under circumstances deemed reasonable from a practitioner's point of view. From (17) we thus conclude that the above arguments, although rather heuristic by nature, indicate that $\hat{V}_a^{1;1}$ should be more efficient than $\hat{V}_o^{1;1}$ under reasonable circumstances. □

Given a reasonably large sample size in each phase, each one of the four phase-compliant estimators $\hat{V}_a^{1;1}$, $\hat{V}_a^{g;1}$, $\hat{V}_a^{g_1;g}$, and $\hat{V}_a^{g_1g;g}$ is approximately design unbiased for $V(\hat{t}_r)$ and should yield approximately valid confidence intervals. However, using the same arguments as in Section 5, we expect $\hat{V}_a^{g_1;g}$ and $\hat{V}_a^{g_1g;g}$ to exhibit better conditional behavior. Two references regarding conditional vs. unconditional behavior, relevant for this paper, are Rao and Sitter (1995) and Sitter (1997). The references are similar, in that they both concern variance estimation for the *GREG* when simple random sampling without replacement is used in both phases, *SI*, *SI* for short, no auxiliary information is available at the onset of the study, and the auxiliary information observed for $k \in s_a$ is scalar valued.

EXAMPLE 2. In Rao and Sitter (1995), the point estimator of interest is $\hat{\bar{y}}_r = (\bar{y}_s/\bar{x}_s)\,\bar{x}_{s_a} = \hat{B}_s\bar{x}_{s_a}$, where $\bar{y}_s = \sum_s y_k/n_s$, $\bar{x}_s$ is defined analogously, and

$\bar{x}_{s_a}$ is the first-phase analogue to $\bar{x}_s$. Using $w_k \propto x_k$ in (1), it is a matter of simple algebra to show that $\hat{\bar{y}}_r = \hat{t}_r/N$ under the design SI, SI. It is straightforward to show that the reference variance estimator considered by Rao and Sitter,

$$\hat{V}_0\left(\hat{\bar{y}}_r\right) = \frac{c_{s_a}}{n_{s_a}}\left(\hat{B}_s^2 S_{xs}^2 + 2\hat{B}_s S_{xes}\right) + \frac{c_s}{n_s} S_{es}^2, \quad (19)$$

where $c_{s_a} = 1 - n_{s_a}/N$, $c_s = 1 - n_s/N$, $S_{xs}^2 = \sum_s (x_k - \bar{x}_s)^2/(n_s - 1)$, $S_{es}^2 = \sum_s e_k^2/(n_s - 1)$ with $e_k = y_k - \hat{B}_s x_k$, and $S_{xes} = \sum_s x_k e_k/(n_s - 1)$, is identical to the variance estimator for $\hat{\bar{y}}_r$ given by $\hat{V}_o^{1;1}/N^2$. Two variance estimators included in the study are the linearization-type estimators

$$\hat{V}_1\left(\hat{\bar{y}}_r\right) = \frac{c_{s_a}}{n_{s_a}}\left(\hat{B}_s^2 S_{xs_a}^2 + 2\hat{B}_s S_{xes}\right) + \frac{c_s}{n_s} S_{es}^2, \quad (20)$$

and

$$\hat{V}_2\left(\hat{\bar{y}}_r\right) = \frac{c_{s_a}}{n_{s_a}}\left(\hat{B}_s^2 S_{xs_a}^2 + 2\frac{\bar{x}_{s_a}}{\bar{x}_s}\hat{B}_s S_{xes}\right) + \frac{c_s}{n_s}\left(\frac{\bar{x}_{s_a}}{\bar{x}_s}\right)^2 S_{es}^2, \quad (21)$$

where $S_{xs_a}^2$ is the first-phase analogue to S_{xs}^2. While $\hat{V}_1\left(\hat{\bar{y}}_r\right)$ is derived using arguments somewhat similar in spirit to those given in Section 5.2, $\hat{V}_2\left(\hat{\bar{y}}_r\right)$ is a linearized version of a jackknife estimator. It is a matter of simple algebra to show that (20) and (21) are identical to $\hat{V}_a^{1;1}/N^2$ and $\hat{V}_a^{g_1 g;g}/N^2$ respectively, under the given model and design. □

In the simulation study conducted by Rao and Sitter (1995), $\hat{V}_1\left(\hat{\bar{y}}_r\right)$ and $\hat{V}_2\left(\hat{\bar{y}}_r\right)$ are both found to be substantially more efficient than $\hat{V}_0\left(\hat{\bar{y}}_r\right)$ unconditionally, under reasonable circumstances. However, from results regarding one-phase sampling (e.g. Royall and Cumberland, 1981a, 1981b; Wu and Deng, 1983; Särndal et al., 1989), one may expect $\hat{V}_2\left(\hat{\bar{y}}_r\right)$ to outperform $\hat{V}_1\left(\hat{\bar{y}}_r\right)$ conditionally, given the approximately ancillary statistic $\bar{x}_{s_a}/\bar{x}_s$. This hypothesis is supported by the results of the simulation study; $\hat{V}_2\left(\hat{\bar{y}}_r\right)$ reveals acceptable conditional behavior, whereas $\hat{V}_1\left(\hat{\bar{y}}_r\right)$ does not.

EXAMPLE 3. In Sitter (1997), the point estimator of interest is $\hat{\bar{y}}_r = \bar{y}_s + \hat{B}_s\left(\bar{x}_{s_a} - \bar{x}_s\right)$, where $\hat{B}_s = S_{ys}^2/S_{xs}^2$ with S_{ys}^2 analogous to S_{xs}^2. Using $\mathbf{x}_k = (1, x_k)'$ and $w_k = 1$ in (1), it follows that $\hat{\bar{y}}_r = \hat{t}_r/N$ under the design SI, SI. Apart from the reference estimator, which is identical to $\hat{V}_o^{1;1}/N^2$, based on $e_k = y_k - \bar{y}_s - \hat{B}_s\left(x_k - \bar{x}_s\right)$, five other variance estimators are studied, including two estimators suggested by Dorfman (1994). Of these, two are linearization-type estimators that are of interest within our context, namely

$$\hat{V}_1\left(\hat{\bar{y}}_r\right) = \frac{c_{s_a}}{n_{s_a}}\hat{B}_s^2 S_{xs_a}^2 + \frac{c_s}{n_s} S_{es}^2 \quad (22)$$

and

$$\hat{V}_2\left(\hat{\bar{y}}_r\right) = \hat{V}_1\left(\hat{\bar{y}}_r\right) + \frac{1}{n_s} R_1^* + \frac{1}{n_{s_a}} R_2^*, \quad (23)$$

where

$$R_1^* = \frac{1}{n_s}\sum_s e_k^2\left(a_{ks}^{*2} + 2a_{ks}^*\right)$$

and

$$R_2^* = \frac{n_{s_a}}{n_s}\frac{2}{n_{s_a} - 1}\hat{B}_s \sum_s \left(x_k - \bar{x}_{s_a}\right) e_k a_{ks}^*,$$

where $a_{ks}^* = a_{ks}/(1 - b_{ks})$, with

$$a_{ks} = n_s\left(x_k - \bar{x}_s\right)\left(\bar{x}_{s_a} - \bar{x}_s\right) / \left[\left(n_s - 1\right) S_{xs}^2\right]$$

and

$$b_{ks} = 1/n_s + \left(x_k - \bar{x}_s\right)^2 / \left[\left(n_s - 1\right) S_{xs}^2\right].$$

The derivation of the estimators (22) and (23) is similar to the derivation of the estimators (20) and (21) respectively, and it is easy to show that, under the given design, (22) is identical to $\hat{V}_a^{1;1}/N^2$. Moreover, from $\hat{V}_a^{g_1 g;g}$ we get

$$\tilde{V}_2\left(\hat{\bar{y}}_r\right) = \frac{\hat{V}_a^{g_1 g;g}}{N^2} = \hat{V}_1\left(\hat{\bar{y}}_r\right) + \frac{c_s}{n_s} R_1 + \frac{c_{s_a}}{n_{s_a}} R_2, \quad (24)$$

where

$$R_1 = \frac{1}{n_s - 1}\sum_s e_k^2\left(a_{ks}^2 + 2a_{ks}\right)$$

and

$$R_2 = \frac{2}{n_s - 1}\hat{B}_s \sum_s \left(x_k - \bar{x}_{s_a}\right) e_k a_{ks}.$$

When the second-phase sample size, n_s, is large, $a_{ks}^* \doteq a_{ks}$, which implies that the estimators (23) and (24), while not identical, have similar large-sample behavior. □

The properties of the variance estimators considered in Sitter are studied through simulation. Both $\hat{V}_1\left(\hat{\bar{y}}_r\right)$ and $\hat{V}_2\left(\hat{\bar{y}}_r\right)$ are found to be substantially more efficient than $\hat{V}_0\left(\hat{\bar{y}}_r\right)$ unconditionally, with $\hat{V}_1\left(\hat{\bar{y}}_r\right)$ being slightly better. However, when studying the properties conditional on the approximately ancillary statistic $\bar{x}_{s_a} - \bar{x}_s$, a slightly different picture

emerges. Both $\hat{V}_1(\hat{\bar{y}}_r)$ and $\hat{V}_2(\hat{\bar{y}}_r)$ perform well when the population data conform to a homoscedastic regression model, but $\hat{V}_2(\hat{\bar{y}}_r)$ continue to perform well also when the population data deviates slightly from such a model, whereas $\hat{V}_1(\hat{\bar{y}}_r)$ does not. These findings are in line with what might be expected; considering one-phase sampling, some support is given in Royall and Cumberland (1981b) and Särndal et al. (1989). Furthermore, since $\hat{V}_2(\hat{\bar{y}}_r)$ and $\tilde{V}_2(\hat{\bar{y}}_r)$ are approximately equivalent in large samples, the simulation results regarding $\hat{V}_2(\hat{\bar{y}}_r)$ should hold approximately for $\tilde{V}_2(\hat{\bar{y}}_r)$ as well.

The final example concerns the problem of negative variance estimates, which indeed may be a problem of practical concern under two-phase sampling.

EXAMPLE 4. Axelson et al. (1996) consider two-phase regression estimation for the special case when no auxiliary information is available at the onset of the study. For the first phase, they consider two-stage sampling, such that stratified simple random sampling is used in the first stage, whereas an arbitrary design is allowed for in the second stage of the first phase. For the second phase, any design such that $\pi_{kl|s_a} > 0$ for all $k\&l \in s_a$ is allowed for.

As a first choice of variance estimator, they consider $\hat{V}_{HT} = \hat{V}_{1,HT} + \hat{V}_2$, where $\hat{V}_{1,HT}$ is a slightly modified version of $\hat{V}_{1o}^1$, while $\hat{V}_2$ is identical to $\hat{V}_{2o}^1$. Through a simulation study, Axelson et al. show that $\hat{V}_{HT}$, due to unstable performance of $\hat{V}_{1,HT}$, results in negative estimates unacceptably often under circumstances deemed relevant for the particular application considered. In order to avoid this, an alternative estimator for V_1, denoted $\hat{V}_{1,reg}$, is suggested. $\hat{V}_{1,reg}$ utilizes more of the available auxiliary information than $\hat{V}_{1,HT}$, why the authors expect $\hat{V}_{reg} = \hat{V}_{1,reg} + \hat{V}_2$ to perform better than $\hat{V}_{HT}$ under conditions such that the *GREG* outdoes the double-expansion estimator. This expectation is confirmed in the above mentioned simulation study. Although derived using a different line of argument, it is easy to show that $\hat{V}_{1,reg}$ alternatively can be derived using the approach suggested in Section 5.2 with $E\left(\hat{V}_{1,HT}|s_a\right)$ as the starting point. □

References

Axelson, H. M., F. J. Breidt, and A. L. Carriquiry (1996). Two-phase regression estimation for policy analysis using computer simulation experiments. In *Proceedings of the Section on Survey Research Methods*, pp. 320–325. American Statistical Association.

Cochran, W. G. (1953). *Sampling Techniques* (1st ed.). New York: Wiley.

Dorfman, I. H. (1994). A note on variance estimation for the regression estimator in double sampling. *Journal of the American Statistical Association 89*, 137–140.

Holt, D. and T. M. F. Smith (1979). Poststratification. *Journal of the Royal Statistical Society A 142*, 33–46.

Rao, J. N. K. (1985). Conditional inference in survey sampling. *Survey Methodology 11*, 15–31.

Rao, J. N. K. and R. R. Sitter (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika 82*, 453–460.

Royall, R. M. and W. G. Cumberland (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association 76*, 66–77.

Royall, R. M. and W. G. Cumberland (1981b). The finite-population linear regression estimator and estimators of its variance–an empirical study. *Journal of the American Statistical Association 76*, 924–930.

Särndal, C. E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference 7*, 155–170.

Särndal, C. E. and B. Swensson (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and non-response. *International Statistical Review 55*, 279–294.

Särndal, C. E., B. Swensson, and J. H. Wretman (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., B. Swensson, and J. H. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Sitter, R. R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association 92*, 780–787.

Wu, C. F. J. and L. Y. Deng (1983). Estimation of variance of the ratio estimator: an empirical study. In G. E. P. Box et al. (Eds.), *Scientific Inference, Data Analysis and Robustness*, pp. 245–277. New York: Academic Press.