A COMPARISON OF TWO REGRESSION ESTIMATORS FOR TWO-PHASE SAMPLING

X. Li and J. D. Opsomer

Department of Statistics, Iowa State University, Ames, IA 50011, USA

KEY WORDS: Calibration, Auxiliary Variables

1. Introduction

Two-phase sampling designs are commonly used for large-scale surveys in which a complete population sampling frame is not available or the variables of interest is expensive or difficult to collect. One or both of these situations often occur in natural resource surveys, for instance. In those types of surveys, it is also common to have access to auxiliary information, which can be used to improve the accuracy of estimation. In two-phase sampling designs, the auxiliary information may exist at two levels. Estevao and Särndal (2002) used the term *complete auxiliary information* to refer to situation in which auxiliary information is available at both the population and the phase one level. They showed that there are nine different ways to incorporate complete auxiliary information into two-phase calibration estimation. In this article, we consider two of those ways in the context of two-phase regression estimation.

204 Snedecor Hall, Ames, IA 50011
lixiaoxi@iastate.edu

Regression techniques are a common way to incorporate the auxiliary information into the survey estimation for two-phase sampling, using either linear, ratio or categorical (post-stratification) models. Different two-phase regression estimators were studied by Särndal and Swensson (1987) and Särndal et al. (1992, ch. 9). Armstrong and St. Jean (1994) applied regression estimators for a two-phase design in a survey of Statistics Canada. Deville and Särndal (1992) discussed a general method of calibration estimation as an alternative to regression estimation, and noted the close connections between both approaches. Dupont (1995) also discussed the relationship between regression and calibration. Estevao and Särndal (2002) expressed their two-phase estimators as calibration estimators, even though they can also be viewed as regression estimators. Sitter (1997) and Fuller (1998) examined replication variance estimation for the two-phase regression estimator.

The outline of this article is as follows. In Section 2 we introduce the two regression estimators that are considered in this article. One was discussed by Särndal et al. (1992). The other was described

by Fuller (1998). For brevity we will call them *Särndal's estimator* and *Fuller's estimator*, respectively, in future discussions. The calibration properties of these two estimators are examined in Section 3. Section 4 reports a simulation study that compares the finite sample properties of these two regression estimators. Some conclusions are drawn in Section 5.

2. Two Regression Estimators

The objective of regression estimators is to use available auxiliary information to improve the precision of estimates. In the context of two-phase sampling, the study variables $\mathbf{y}$ are known only for the second-phase sample, whereas the auxiliary variables $\mathbf{x}_k$ are known either at the population level or phase one level. As stated in the previous section, we study the complete auxiliary information case, where subsets of the variables in $\mathbf{x}_k$ are known for phase one and for the population.

First, we define some notation. We consider sampling from a finite population U. Let s_a and s denote the first-phase and second-phase samples, respectively, and write $\pi_{ak} = \Pr(k \in s_a)$ for the phase one inclusion probability for $k \in U$ and $\pi_{k|s_a} = \Pr(k \in s | k \in s_a)$ for its phase two conditional inclusion probabilities, and let $\pi_k^* = \pi_{k|s_a}\pi_{ak}$. Let $\mathbf{x}_{1k}$ be the vector of J_1 auxiliary variables known for all $k \in U$, where $J_1 \geq 2$ since $\mathbf{x}_{1k}$ will be assumed to contain the intercept. Let $\mathbf{x}_{2k}$ be the vector of J_2 auxiliary variables known for all $k \in s_a$. For an element $k \in s_a$, the complete auxiliary information is thus the vector

$$\mathbf{x}_k' = (\mathbf{x}_{1k}', \mathbf{x}_{2k}').$$

Finally, let $\mathbf{y}$ denote the vector of J_3 study variables.

In order to introduce the notation for the different estimators, let z denote a generic variable (either y or x), and

$\bar{\mathbf{z}}_N$ = the population mean of $\mathbf{z}$;

$\bar{\mathbf{z}}_{\pi_a}$ = the π-weighted estimator of the population mean of $\mathbf{z}$, based on phase one;

$\bar{\mathbf{z}}_\pi$ = the π-weighted estimator of the population mean of $\mathbf{z}$, based on phase two;

$\hat{\bar{\mathbf{z}}}_{reg_a}$ = the regression estimator for the population mean, based on phase one (and hence using $\mathbf{x}_1$ only as auxiliary variable);

$\hat{\bar{\mathbf{z}}}_{reg}$ = the regression estimator for the population mean, based on phase two (using both $\mathbf{x}_1$ and $\mathbf{x}_2$). We will consider two types:

$$\hat{\bar{\mathbf{z}}}_{reg}^F = \text{Fuller's estimator},$$
$$\hat{\bar{\mathbf{z}}}_{reg}^S = \text{Särndal's estimator},$$

further defined below.

$\hat{\boldsymbol{\beta}}_{z_a}$ = the weighted regression coefficients for $\mathbf{z}_k$ on phase 1 (using $\mathbf{x}_1$ only);

$\hat{\boldsymbol{\beta}}_z$ = the weighted regression coefficients for $\mathbf{z}_k$ on phase 2, using both $\mathbf{x}_1$ and $\mathbf{x}_2$;

$\hat{\boldsymbol{\beta}}_{1z}$ = the weighted regression coefficients vector for $\mathbf{z}_k$ on phase 2, but only using $\mathbf{x}_1$;

For simplicity, we consider a single y ($J_3 = 1$) in the remainder of the discussion. The first regression estimator considered is given by Särndal et al. (1992), and is defined as

$$\widehat{\bar{y}}^{S}_{reg} = \frac{1}{N}\left(\sum_{U}\widehat{y}_{1k} + \sum_{s_a}\frac{\widehat{y}_k - \widehat{y}_{1k}}{\pi_{ak}} + \sum_{s}\frac{y_k - \widehat{y}_k}{\pi^*_k}\right). \quad (1)$$

where

$$\widehat{y}_{1k} = \mathbf{x}'_{1k}\widehat{\boldsymbol{\beta}}_{1y} \quad \text{for } k \in U;$$
$$\widehat{\boldsymbol{\beta}}_{1y} = \left(\sum_{s}\frac{\mathbf{x}_{1k}\mathbf{x}'_{1k}}{\pi^*_k}\right)^{-1}\sum_{s}\frac{\mathbf{x}_{1k}y_k}{\pi^*_k}. \quad (2)$$

$$\widehat{y}_k = \mathbf{x}'_k\widehat{\boldsymbol{\beta}}_{y} \quad \text{for } k \in U;$$
$$\widehat{\boldsymbol{\beta}}_{y} = \left(\sum_{s}\frac{\mathbf{x}_{k}\mathbf{x}'_{k}}{\pi^*_k}\right)^{-1}\sum_{s}\frac{\mathbf{x}_{k}y_k}{\pi^*_k}. \quad (3)$$

The other is given by Fuller (1998), as

$$\widehat{\bar{y}}^{F}_{reg} = \bar{y}_{\pi} + \begin{pmatrix} \bar{\mathbf{x}}_{1N} - \bar{\mathbf{x}}_{1\pi} \\ \hat{\bar{\mathbf{x}}}_{2reg_a} - \bar{\mathbf{x}}_{2\pi} \end{pmatrix}' \widehat{\boldsymbol{\beta}}_{y}. \quad (4)$$

where $\hat{\bar{\mathbf{x}}}_{2reg_a}$ is the estimator for $\bar{\mathbf{x}}_{2N}$, based on the regression of $\mathbf{x}_{2k}$ on $\mathbf{x}_{1k}$, given by

$$\hat{\bar{\mathbf{x}}}'_{2reg_a} = \bar{\mathbf{x}}'_{2\pi_a} + (\bar{\mathbf{x}}_{1N} - \bar{\mathbf{x}}_{1\pi_a})'\,\widehat{\boldsymbol{\beta}}_{x_{2a}}.$$

where $\widehat{\boldsymbol{\beta}}_{x_{2a}}$ contains the regression coefficients constructed from s_a,

$$\widehat{\boldsymbol{\beta}}_{x_{2a}} = \left(\sum_{s_a}\frac{\mathbf{x}_{1k}\mathbf{x}'_{1k}}{\pi_{ak}}\right)^{-1}\sum_{s_a}\frac{\mathbf{x}_{1k}\mathbf{x}'_{2k}}{\pi_{ak}}.$$

Note that $\widehat{\boldsymbol{\beta}}_{x_{2a}}$ can be a matrix if $J_2 > 1$. The estimator $\widehat{\bar{y}}^{S}_{reg}$ in (1) is constructed based on two assumed superpopulation models. Model ξ_1 is the model for the relationship used to construct the $\widehat{y}_{1k}$ in (2):

$$\begin{aligned} \mathrm{E}_{\xi_1}(y_k) &= \boldsymbol{x}'_{1k}\boldsymbol{\beta}_1, \\ \mathrm{Var}_{\xi_1}(y_k) &= \sigma^2_1, \end{aligned}$$

and model ξ_2 is used for $\widehat{y}_k$ in (3):

$$\begin{aligned} \mathrm{E}_{\xi_2}(y_k) &= \boldsymbol{x}'_{k}\boldsymbol{\beta}_2, \\ \mathrm{Var}_{\xi_2}(y_k) &= \sigma^2_2. \end{aligned}$$

In these models, it is reasonable to assume that $\sigma^2_2 < \sigma^2_1$, since ξ_2 contains additional predictor variables, but this is not used in the estimation. The estimator $\widehat{\bar{y}}^{S}_{reg}$ in (1) is composed of three components. The first one represents a population-level model prediction based on ξ_1, the second component is a first-phase sample "correction" component based on model ξ_2, and the third one is an additional second-phase sample correction component.

Similarly, the Fuller estimator in (4) is based on two assumed superpopulation models. The first one is the same model ξ_2 as above, and the other is ξ_3, given by

$$\begin{aligned} \mathrm{E}_{\xi_3}(\boldsymbol{x}_{2k}) &= \boldsymbol{x}'_{1k}\boldsymbol{\beta}_3, \\ \mathrm{Var}_{\xi_3}(\boldsymbol{x}_{2k}) &= \sigma^2_3. \end{aligned}$$

The estimator $\widehat{\bar{y}}^{F}_{reg}$ in (4) can be rewritten as

$$\widehat{\bar{y}}^{F}_{reg} = \frac{1}{N}\left[\sum_{U}\left(\mathbf{x}'_{1k}\,,\ \mathbf{x}'_{1k}\widehat{\boldsymbol{\beta}}_{x_{2a}}\right)\widehat{\boldsymbol{\beta}}_{y} + \sum_{s_a}\frac{\left(\mathbf{0}\,,\ \mathbf{x}'_{2k} - \mathbf{x}'_{1k}\widehat{\boldsymbol{\beta}}_{x_{2a}}\right)}{\pi_{ak}}\widehat{\boldsymbol{\beta}}_{y} + \sum_{s}\frac{y_k - \mathbf{x}'_k\widehat{\boldsymbol{\beta}}_{y}}{\pi^*_k}\right]. \quad (5)$$

In this form, the Fuller's estimator is seen to be similar to Särndal's estimator in (1) and to be composed of a model-based population prediction and two correction components corresponding for both phases.

We would like to compare the statistical properties of both estimators. It

is well-known that, under assumptions not further explored here, both estimators are design consistent and asymptotically design unbiased, so that both will be "well-behaved" for sufficiently large sample sizes. However, it is of interest to know whether one of these estimators will generally be more efficient than the other, in either an asymptotic or finite sample sense. Because they involve different models as explained above, there is no easy way to answer this question in general. Hence, in this article we will compare both estimators in two ways. First, we will evaluate their calibration properties. Second, we will conduct a simulation experiment to compare the design mean squared errors of both estimators.

3. Calibration Properties of Estimators

Calibration is an approach that "modifies" a design-based estimator $\bar{y}_\pi$ to incorporate auxiliary information. For a general discussion of calibration and its properties, see Deville and Särndal (1992).

In our study, calibration implies that we require the phase two weighted averages of the auxiliary variables to be exactly equal to the population mean in the case of $\mathbf{x}_{1k}$, and to the estimated population mean based on the first-phase sample for the $\mathbf{x}_{2k}$. When we use calibrated weights to estimate the population means for other variables, the properties of the resulting estimators will depend on the strength of the association between these variables and the calibration variables. For sufficiently strong relationship between the study variable and calibration variables, the calibrated estimators should be more efficient than uncalibrated estimators.

Now we calibrate (1). Replace y_k by $\mathbf{x}'_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})$ and define

$$\widehat{\boldsymbol{\beta}}_{x_2} = \left(\sum_s \frac{\mathbf{x}_{1k}\mathbf{x}'_{1k}}{\pi^*_k} \right)^{-1} \sum_s \frac{\mathbf{x}_{1k}\mathbf{x}'_{2k}}{\pi^*_k},$$

so that

$$\widehat{\bar{\mathbf{x}}}^S_{reg} = \left(\bar{\mathbf{x}}_{1N} \ , \ \bar{\mathbf{x}}'_{2\pi_a} + (\bar{\mathbf{x}}_{1N} - \bar{\mathbf{x}}_{1\pi_a})' \widehat{\boldsymbol{\beta}}_{x_2} \right), \quad (6)$$

the estimated mean for $\mathbf{x}_k$. Hence, we see that $\widehat{\bar{y}}^S_{reg}$ is calibrated exactly for $\bar{\mathbf{x}}_{1N}$, and is calibrated for a regression estimator of $\bar{\mathbf{x}}_{2N}$.

Secondly, we calibrate (4). Replace y_k by $\mathbf{x}'_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})$ again and we obtain

$$\widehat{\bar{\mathbf{x}}}^F_{reg} = \left(\bar{\mathbf{x}}_{1N} \ , \ \bar{\mathbf{x}}'_{2\pi_a} + (\bar{\mathbf{x}}_{1N} - \bar{\mathbf{x}}_{1\pi_a})' \widehat{\boldsymbol{\beta}}_{x_{2a}} \right). \quad (7)$$

Hence, $\widehat{\bar{y}}^S_{reg}$ is also calibrated exactly for $\bar{\mathbf{x}}_{1N}$ and for a phase one regression estimator of $\bar{\mathbf{x}}_{2N}$.

We can see that the only difference between (6) and (7) is that (6) uses $\widehat{\boldsymbol{\beta}}_{x_2}$ as the regression coefficient vector while (7) uses $\widehat{\boldsymbol{\beta}}_{x_{2a}}$. Since $\widehat{\boldsymbol{\beta}}_{x_{2a}}$ is computed on the first phase sample, which has typically much larger sample size than the second phase sample, we conclude that Fuller's estimator has better calibrated weights, because it exploits all the observations available in s_a in the computation of $\widehat{\boldsymbol{\beta}}_{x_{2a}}$. In contrast, Särndal's estimator only uses observations in s to compute the regression coefficient, and hence appears to not make efficient use of the data. It therefore seems reasonable to assume that Fuller's

estimator might have better finite sample properties than Särndal's estimator, even if asymptotically this difference is negligible.

4. Simulation

We conducted a simulation to study the performance of the two regression estimators. For this purpose, finite populations of size $N = 10,000$ were created using the model

$$y_k = 10 + 5x_{1k} - 5x_{2k} + \varepsilon_k, \qquad (8)$$

where $\varepsilon_k \sim N(0, \sigma^2_\varepsilon)$, and we will investigate two levels for σ^2_ε. We also consider three cases for the relationship between x_{2k} and x_{1k}. In the first two cases, x_{2k} and x_{1k} are linearly related through the model

$$x_{2k} = 1 - x_{1k} + \eta_k, \qquad (9)$$

with $x_{1k} \sim U(0,1)$ and $\eta_k \sim N(0, \sigma^2_\eta)$, for two levels of σ^2_η. In the third case, x_{1k} and x_{2k} are linearly independent and are both generated as $U(0,1)$.

By crossing the cases for the model for y_k with those for the model for x_{2k}, we obtain six different scenarios for the overall population model, which we will identify by the coefficients of determination of both models, R^2_y and $R^2_{x_2}$. Specifically, we varied the model for x_{2k} and the model variances so that the six cases correspond to the combinations $(R^2_y, R^2_{x_2})$ with $R^2_y = 0.25$ or 0.75, and $R^2_{x_2} = 0, 0.25$ or 0.75.

Two-phase samples were drawn from each of the populations, with simple random sampling without replacement in both phases. The sample sizes were $n_a = 2,000$ for phase one, and $n = 20$ or 200 for phase two. For each sample, we calculate $\widehat{\bar{y}}^F_{reg}$, $\widehat{\bar{y}}^S_{reg}$ and $\widehat{\bar{y}}_\pi$ for all six population. For comparison, we also compute $\widehat{\bar{y}}_\pi$, the phase two expansion estimator of $\bar{y}_N$. Each simulation setting is repeated $B = 10,000$ times.

Table 1 reports the simulated bias of $\widehat{\bar{y}}^F_{reg}$, $\widehat{\bar{y}}^S_{reg}$ and $\widehat{\bar{y}}_\pi$ for all populations and phase two sample sizes, as a percentage of the finite population mean $\bar{y}_N$. This table shows that all three estimators are essentially unbiased, even at the smaller sample size, since the largest simulated bias was approximately 0.3% at $n = 20$.

Now we compare the design variances (or, equivalently, mean squared error) of Särndal's and Fuller's estimators. Table 2 reports the values of $V(\widehat{\bar{y}}^F_{reg})$ and $V(\widehat{\bar{y}}^S_{reg})$ for different populations and two phase two sample sizes, as a percentage of $V(\widehat{\bar{y}}_\pi)$. The relative MSEs of both estimators appear to be primarily determined by the quality of the regression model for the y_k as a function of the $\boldsymbol{x}_k$, with the models with high R^2_y achieving relative MSE less than 0.3 in all cases, while the models with low R^2_y achieve relative MSEs between 0.75 and 0.9.

A striking feature of the results in Table 2 is that there is almost no difference between $V(\widehat{\bar{y}}^F_{reg})$ and $V(\widehat{\bar{y}}^S_{reg})$. While the effect of the second regression model is indeed visible in the Table 2, it is very small. This appears to contradict the results from Section 3, since equations (6) and (7) showed a difference in the calibration properties of both estimators through the regression coefficients $\widehat{\boldsymbol{\beta}}_{x2_a}$ and $\widehat{\boldsymbol{\beta}}_{x2}$. Hence, in order to study the effect of these

	$(\bar{\hat{y}}^F_{reg} - \bar{y}_N)/\bar{y}_N(\%)$		$(\bar{\hat{y}}^S_{reg} - \bar{y}_N)/\bar{y}_N(\%)$		$(\bar{\hat{y}}_\pi - \bar{y}_N)/\bar{y}_N(\%)$	
Population	$n = 200$	$n = 20$	$n = 200$	$n = 20$	$n = 200$	$n = 20$
1: $R^2_y = 0.75, R^2_{x_2} = 0.75$	0.011	-0.035	0.011	-0.035	-0.008	-0.030
2: $R^2_y = 0.75, R^2_{x_2} = 0.25$	-0.010	0.022	-0.009	0.022	-0.023	0.028
3: $R^2_y = 0.25, R^2_{x_2} = 0.75$	0.017	0.115	0.017	0.115	-0.029	0.053
4: $R^2_y = 0.25, R^2_{x_2} = 0.25$	0.035	0.305	0.035	0.305	0.047	0.168
5: $R^2_y = 0.75, R^2_{x_2} = 0$	0.002	-0.008	0.002	-0.007	0.010	0.069
6: $R^2_y = 0.25, R^2_{x_2} = 0$	-0.018	0.135	-0.018	0.135	-0.037	0.168

Table 1: Simulated relative bias for $\bar{\hat{y}}^F_{reg}$, $\bar{\hat{y}}^S_{reg}$ and $\bar{\hat{y}}_\pi$ for six populations and two phase two sample sizes (in percent).

	$\mathrm{Var}(\bar{\hat{y}}^F_{reg})/\mathrm{Var}(\bar{\hat{y}}_\pi)$ (%)		$\mathrm{Var}(\bar{\hat{y}}^S_{reg})/\mathrm{Var}(\bar{\hat{y}}_\pi)$ (%)	
Population	$n = 200$	$n = 20$	$n = 200$	$n = 20$
1: $R^2_y = 0.75, R^2_{x_2} = 0.75$	26.41	27.45	26.41	27.45
2: $R^2_y = 0.75, R^2_{x_2} = 0.25$	27.14	27.62	27.16	27.65
3: $R^2_y = 0.25, R^2_{x_2} = 0.75$	76.44	86.27	76.44	86.28
4: $R^2_y = 0.25, R^2_{x_2} = 0.25$	77.16	86.81	77.16	86.82
5: $R^2_y = 0.75, R^2_{x_2} = 0$	28.24	28.53	28.25	28.56
6: $R^2_y = 0.25, R^2_{x_2} = 0$	76.34	86.22	76.35	86.23

Table 2: Values of $\mathrm{Var}(\bar{\hat{y}}^F_{reg})/\mathrm{V}(\bar{\hat{y}}_\pi)$ and $\mathrm{Var}(\bar{\hat{y}}^S_{reg})/\mathrm{V}(\bar{\hat{y}}_\pi)$ for different populations and selected phase 2 sample sizes (in percent).

different regression fits on the properties of both estimators, we computed the values of $\mathrm{Var}(\bar{\mathbf{x}}_{2\pi a})$, $\mathrm{Var}\Big((\bar{\mathbf{x}}_{1N} - \bar{\mathbf{x}}_{1\pi_a})' \widehat{\boldsymbol{\beta}}_{x_{2a}}\Big)$, and $\mathrm{Var}\Big((\bar{\mathbf{x}}_{1N} - \bar{\mathbf{x}}_{1\pi_a})' \widehat{\boldsymbol{\beta}}_{x_2}\Big)$ for different populations and different phase two sample sizes.

Table 3 shows the ratio of variances of the regression components of both calibrated estimators,

$$\frac{\mathrm{Var}\left((\bar{\mathbf{x}}_{1N} - \bar{\mathbf{x}}_{1\pi_a})' \widehat{\boldsymbol{\beta}}_{x_{2a}}\right)}{\mathrm{Var}\left((\bar{\mathbf{x}}_{1N} - \bar{\mathbf{x}}_{1\pi_a})' \widehat{\boldsymbol{\beta}}_{x_2}\right)},$$

and the ratio of the variance of expansion component over the regression component, $\mathrm{Var}(\bar{\mathbf{x}}_{2\pi a})/\mathrm{Var}\left((\bar{\mathbf{x}}_{1N} - \bar{\mathbf{x}}_{1\pi_a})' \widehat{\boldsymbol{\beta}}_{x_2}\right)$. An interesting pattern emerges from these results. For the first four populations, where x_{1k} and x_{2k} are linearly related, the variances of the regression components are quite similar to each other, with the smallest variance ratio equal to 0.85 for a phase two sample size of 20 and the low R^2 model. Hence, for the first four populations, the calibration for both Särndal's and Fuller's estimators is almost identical because the regression components of the calibration are estimated with approximately the same precision.

For populations 5 and 6 where x_{1k} is independent of x_{2k}, the variability of $(\bar{\mathbf{x}}_{1N} - \bar{\mathbf{x}}_{1\pi_a})' \widehat{\boldsymbol{\beta}}_{x_{2a}}$ is much smaller than that of $(\bar{\mathbf{x}}_{1N} - \bar{\mathbf{x}}_{1\pi_a})' \widehat{\boldsymbol{\beta}}_{x_2}$ for both sample sizes, so that there is indeed a big difference in the regression component of the calibration. However, this difference does not translate into a loss of efficiency for $\widehat{\bar{y}}^S_{reg}$ relative to $\widehat{\bar{y}}^F_{reg}$, because when x_{1k} is independent of x_{2k}, the variability of the regression component of the calibration, $\mathrm{Var}\Big((\bar{\mathbf{x}}_{1N} - \bar{\mathbf{x}}_{1\pi_a})' \widehat{\boldsymbol{\beta}}_{x_2}\Big)$, is dwarfed by the variability of the expansion component of the calibration, $\mathrm{Var}(\bar{\mathbf{x}}_{2\pi a})$.

Hence, there appears to be a "self-balancing" character in Särndal's regression estimator that causes the apparent calibration inefficiency in its formulation to not have an appreciable effect on its overall statistical properties. Note however this inefficiency might still appear in contexts where the regression models are more complicated, or where estimators are computed for small domains, so that the finding in this simulation study should be interpreted cautiously at this point.

5. Conclusion

In this article, we studied the two regression estimators for the complete auxiliary information case. While they both rely on the same auxiliary information, they are not directly comparable because of incompatibilities between the underlying superpopulation models. Hence, calibration was used to compare them. We showed that Fuller's regression estimator has an advantage over Särndal's because it uses more information for the regression coefficients. However, the reported simulation study shows that these two regression estimators are extremely similar in two ways. First, they produce almost identical estimators for population means. Secondly, their design mean squared errors are very similar too, even when the sample sizes are very small.

Note that the calibration inefficiency of Särndal's estimator might still have an effect in contexts where the regression

	$\frac{\text{Var}\left((\bar{\mathbf{x}}_{1N}-\bar{\mathbf{x}}_{1\pi_a})'\widehat{\boldsymbol{\beta}}_{x_{2a}}\right)}{\text{Var}\left((\bar{\mathbf{x}}_{1N}-\bar{\mathbf{x}}_{1\pi_a})'\widehat{\boldsymbol{\beta}}_{x_2}\right)}$		$\frac{\text{Var}(\bar{\mathbf{x}}_{2\pi a})}{\text{Var}\left((\bar{\mathbf{x}}_{1N}-\bar{\mathbf{x}}_{1\pi_a})'\widehat{\boldsymbol{\beta}}_{x_2}\right)}$	
Population	$n=200$	$n=20$	$n=200$	$n=20$
1: $R^2_y=0.75, R^2_{x_2}=0.75$	0.99	0.99	1.35	1.33
2: $R^2_y=0.75, R^2_{x_2}=0.25$	0.98	0.86	3.92	3.45
3: $R^2_y=0.25, R^2_{x_2}=0.75$	0.99	0.97	1.35	1.29
4: $R^2_y=0.25, R^2_{x_2}=0.25$	0.99	0.85	4.09	3.36
5: $R^2_y=0.75, R^2_{x_2}=0$	0.09	0.01	214.97	18.57
6: $R^2_y=0.25, R^2_{x_2}=0$	0.09	0.01	218.91	17.20

Table 3: Comparison of $\text{Var}(\bar{\mathbf{x}}_{2\pi a})$, $\text{Var}\left((\bar{\mathbf{x}}_{1N}-\bar{\mathbf{x}}_{1\pi_a})'\widehat{\boldsymbol{\beta}}_{x_{2a}}\right)$ and $\text{Var}\left((\bar{\mathbf{x}}_{1N}-\bar{\mathbf{x}}_{1\pi_a})'\widehat{\boldsymbol{\beta}}_{x_2}\right)$for different populations and different sample sizes.

models are more complicated, or where estimators are computed for small domains, so that the finding in this simulation study should be interpreted cautiously at this point. Nevertheless, it appears that in many situations, the choice between the two estimators can be driven by application requirements, not by efficiency concerns.

References

Armstrong, J. and H. St. Jean (1994). Generalized regression estimation for a two-phase sample of tax records. *Survey Methodology 20*, 97–105.

Deville, J.-C. and C.-E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*, 376–382.

Dupont, F. (1995). Alternative adjustments where there are several levels of auxiliary information. *Survey Methodology 21*, 125–135.

Estevao, V. M. and C.-E. Särndal (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics 18*, 233–255.

Fuller, W. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica 8*, 1153–1164.

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

Särndal, C.-E., B. Swensson, and J. 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.