Efficient Calibration for Some Variants of Double Sampling

Takis Merkouris*

Abstract
Recent methodology on combination of multiple-survey data through generalized regression is adapted to certain variants of double sampling arising in practice. A computationally simple calibration scheme that gives rise to efficient generalized regression estimators for characteristics surveyed in the "second-phase" sample is investigated within a framework of optimal regression estimation. This, one-step, calibration scheme makes efficient use of all available auxiliary information in the first-phase and second-phase samples and greatly facilitates variance estimation.

Key Words: Two-phase sampling, nested double sampling, non-nested double sampling, generalized regression estimator, calibration, composite estimator.

1. Introduction

Double sampling, also known as two-phase sampling, is commonly used in large scale surveys as a cost-effective survey method. In the first phase a large sample drawn from the target population provides auxiliary information that is inexpensive to collect, and in the second phase a subsample is used to collect information on the variables of interest. The population totals for some of the auxiliary variables may be known.

In a common estimation procedure in two-phase sampling, regression techniques are used to incorporate all the available auxiliary information into the survey estimation and thus improve the precision of estimates. Generalized regression estimators, also viewed as calibration estimators, were studied in the context of two-phase sampling by Särndal et al. (1992, ch. 9), Hidiroglou and Särndal (1998), Hidiroglou (2001), and Estevao and Särndal (2003). In particular, Hidiroglou (2001) included in his discussion a sampling design, termed non-nested double sampling, in which one of the samples is not nested in the other nor is it necessarily selected from the same frame.

In this paper a computationally simple calibration procedure that gives rise to efficient generalized regression estimators is proposed for certain variants of nested double sampling arising in practice. These variants are distinguished from the standard type of nested double sampling by the following features. (i) The separate use of the first-phase sample for a large scale survey. (ii) The second-phase sample is statistically independent from its complement — the two component samples have the same sampling design. (iii) The second-phase sample is not necessarily smaller than its complement, but usually it is. (iv) The type and amount of auxiliary information collected in the first-phase sample is not determined by the objectives of the survey based on the second-phase sample. A unified presentation of the proposed procedure includes also the non-nested double sampling. Three special cases of interest are described below.

Nested double sampling, Case I.
In this sampling scheme, a large sample s is used for the objectives of one (the main) survey, and a sub-sample s_2 of it — the "second-phase" sample — is used for another survey with different target variables. The characteristic feature of this sub-sample is that it is made up of one or more of the independent parts comprising the entire sample s, so that the sub-sample s_2 and its complement s/s_2 are independent samples. Some of the auxiliary variables (common to the two phases) are known at population level, and can be used for calibration, and some are known at sample level only. Examples of this convenient sampling scheme include household surveys that use a second-phase sample made up of some of the sub-samples (panels) comprising a Labour Force Survey (LFS). In these "supplement" surveys (e.g., travel survey, health survey) the target variables are different from the LFS target variables.

The setting of this nested double-sampling is depicted in Figure 1, where $\mathbf{x}_1$ denotes the vector of auxiliary variables for which the associated vector of population totals is known, $\mathbf{x}_2$ denotes the vector of auxiliary variables for which the vector of population totals is not known, $\mathbf{y}_1$ denotes the vector of target variables surveyed in the entire sample s, and $\mathbf{y_2}$ denotes the vector of target variables surveyed in the second-phase sample s_2. In the context of this variant of nested double sampling, the target variable of interest is $\mathbf{y}_2$.

Nested double sampling. Case II.
This variant of case I is about a single repeated survey, in which a sub-sample may be used periodically to collect information on a set of additional target variables. Collection of all data is done simultaneously from the entire

*Athens University of Economics and Business, Patision 76, 10434 Athens, Greece

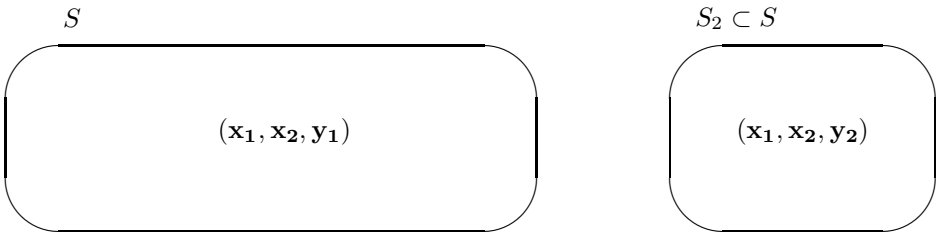

Figure 1: Nested double sampling. Case I.

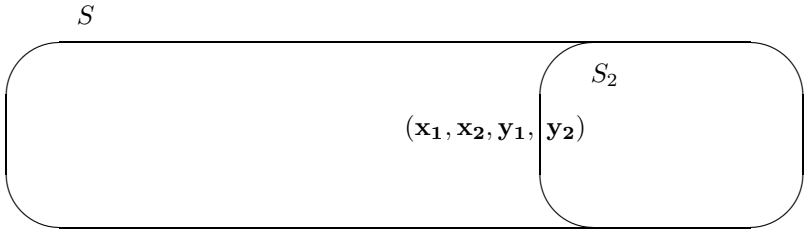

Figure 2: Nested double sampling. Case II.

sample using the same questionnaire, but the module with the additional variables is administered only to the sub-sample in order to reduce response burden and cost. The motivating example of such setting is the multiple-panel Labour Force Surveys in the European Union that use a sub-sample made up of one or more panels to collect information on additional (*structural*) variables.

In Figure 2 depicting this variant of nested double sampling, $\mathbf{y}_2$ denotes the vector of the additional variables surveyed in the second-phase sample s_2.

Non-nested double sampling.
This sampling scheme involves two separate samples, s_1 and s_2, drawn from the same frame, or from different frames that represent the same target population. The sampling designs for the two samples may be different. The two samples have common auxiliary variables, some of which may be known at the population level and some at the sample level. Information on the target variables is collected only in one (the main) sample. The other sample serves as additional source of auxiliary information, and is usually drawn from an administrative file and is much larger. An example of a survey with non-nested double sampling is the Canadian Survey of Employment, Payroll and Hours (see Hidiroglou 2001).

The setting of the nested double-sampling is depicted in Figure 3, where $\mathbf{x}_1$ denotes the vector of auxiliary variables for which the associated vector of population totals is known, $\mathbf{x}_2$ denotes the vector of auxiliary variables for which the vector of population totals is not known, and $\mathbf{y}$ is the target variable surveyed in the main sample s_2.

Information on variables that are common between s_2 and its complement s/s_2 in nested double sampling, or between s_2 and s_1 in non-nested double sampling, may be combined to enhance the efficiency of estimates for the target variables $\mathbf{y}_2$ ($\mathbf{y}$ for non-nested double sampling) surveyed only in the second-phase sample s_2. This can be achieved by a special regression setup functioning as an extended calibration procedure whereby estimates for the common vector $\mathbf{x}_2$, or for both $\mathbf{x}_2$ and $\mathbf{y}_1$ in case II of nested double sampling, based on the two samples are calibrated to each other.

The proposed calibration is performed on the entire sample s in nested double sampling, or on the combined sample $s_1 \cup s_2$ in non-nested double sampling, and involves also the auxiliary variable $\mathbf{x}_1$. This, one-step, calibration scheme makes efficient use of all available auxiliary information in the first-phase and second-phase samples. The resulting composite calibration estimator for $\mathbf{y}_2$ ($\mathbf{y}$ for non-nested double sampling) is a special form of generalized regression estimator, and for certain sampling designs it is the optimal regression estimator. This estimation

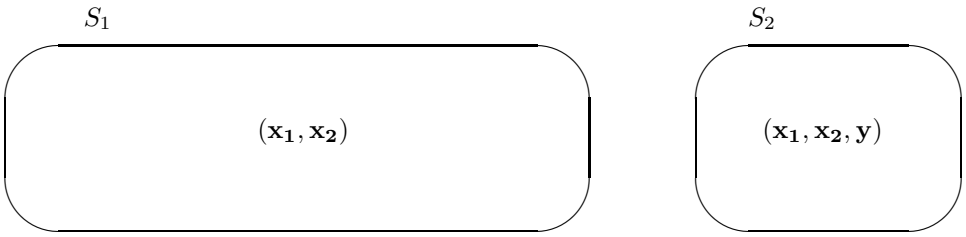

Figure 3: Non-nested double sampling.

procedure is very practical and greatly facilitates variance estimation.

Non-nested double sampling and case I of nested double sampling fall within the domain of application of the methods of Renssen and Nieuwenbroek (1997) and Merkouris (2004) for integrating independent surveys with common variables through regression for more efficient estimation. The proposed method is an adaptation of the method of Merkouris (2004). Non-nested double sampling has also been dealt with by Hidiroglou (2001), but in a different manner. No other method of using data on $\mathbf{x}_2$ and $\mathbf{y}_1$ from the entire sample s in the estimation for $\mathbf{y}_2$ has been proposed thus far for case II of nested double sampling.

The paper is organized as follows. A preliminary discussion of generalized regression estimation as calibration estimation is presented in Section 2. Calibration estimation for the two variants of nested double sampling is presented in Section 3. Composite calibration estimators for all variables of interest are discussed within a framework of optimal regression. Calibration estimation for non-nested double sampling is presented in Section 4. Some remarks on the merits of the proposed estimation method relative to certain alternative methods are made in Section 5.

2. Calibration and Generalized Regression: A Review

Consider a finite population labeled by $U = \{1, \ldots, k, \ldots N\}$, from which a probability sample s of size n is drawn according to a sampling design with known first - and second - order inclusion probabilities π_k and π_{kl} $(k, l \in U)$. Consider the sampling weight vector $\mathbf{w}$ with kth entry defined as $w_k = (1/\pi_k)I(k \in s)$, where I denotes the indicator variable, and let $\mathbf{Y} \in \mathbb{R}^{N\times r}$ denote the population matrix of an r-dimensional survey variable of interest $\mathbf{y}$. The Horvitz - Thompson (HT) estimator of the total $\mathbf{t_y} = \mathbf{Y'1}$, where $\mathbf{1}$ is the unit N-vector, is given by $\hat{\mathbf{Y}} = \mathbf{Y'w}$ $(=\sum_U w_k y_k)$. For the population matrix $\mathbf{X} \in \mathbb{R}^{N\times p}$ of a p-dimensional auxiliary variable $\mathbf{x}$, assume that the total $\mathbf{t_x} = \mathbf{X'1}$ is known. Let also $\mathbf{\Lambda}$ be the diagonal "weighting" matrix that has w_k/q_k as kkth entry, where q_k is a positive constant — the typical default value $q_k = 1$ for all k will be assumed, unless indicated otherwise. The subvectors and submatrices corresponding to the sample are designated by s. A vector of "calibrated" weights, $\mathbf{c}_s \in \mathbb{R}^n$, can be constructed to satisfy the constraints $\mathbf{X}_s'\mathbf{c}_s = \mathbf{t_x}$ while minimizing the generalized least squares distance $(\mathbf{c}_s - \mathbf{w}_s)'\mathbf{\Lambda}_s^{-1}(\mathbf{c}_s - \mathbf{w}_s)$. Assuming that $\mathbf{X}_s$ is of full rank p, this calibration procedure generates the vector

$$\mathbf{c}_s = \mathbf{w}_s + \mathbf{\Lambda}_s\mathbf{X}_s(\mathbf{X}_s'\mathbf{\Lambda}_s\mathbf{X}_s)^{-1}(\mathbf{t_x} - \mathbf{X}_s'\mathbf{w}_s). \tag{1}$$

The calibration estimator of the total $\mathbf{t_y}$ is obtained as

$$\mathbf{Y}_s'\mathbf{c}_s = \mathbf{Y}_s'\mathbf{w}_s + \mathbf{Y}_s'\mathbf{\Lambda}_s\mathbf{X}_s(\mathbf{X}_s'\mathbf{\Lambda}_s\mathbf{X}_s)^{-1}(\mathbf{t_x} - \mathbf{X}_s'\mathbf{w}_s), \tag{2}$$

which can take the form of a generalized regression (GREG) estimator

$$\hat{\mathbf{Y}}^R = \hat{\mathbf{Y}} + \hat{\boldsymbol{\beta}}'(\mathbf{t_x} - \hat{\mathbf{X}}), \tag{3}$$

where $\hat{\mathbf{X}} = \mathbf{X}_s'\mathbf{w}_s$ is the HT estimator of $\mathbf{t_x}$, and $\hat{\boldsymbol{\beta}} = (\mathbf{X}_s'\mathbf{\Lambda}_s\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{\Lambda}_s\mathbf{Y}_s$ is the matrix of sample regression coefficients. By construction the GREG estimator (3) has the calibration property that $\hat{\mathbf{X}}^R = \mathbf{t_x}$, that is, the GREG estimator of the total for $\mathbf{x}$ is equal to the known associated population total ("control" total). A formulation of the GREG estimator as a calibration estimator is given in Deville and Särndal (1992), and an extensive discussion of it is given in Särndal et al. (1992).

3. Calibration for Nested Double Sampling

Let s_1 denote the complement s/s_2 of s_2. For the survey settings described above, s_1 and s_2 are assumed to be independent sub-samples of a large sample s and have sizes n_1 and n_2, respectively. The inverses of the sub-sample

proportions $\phi = n_1/(n_1+n_2)$ and $(1-\phi) = n_2/(n_1+n_2)$ can be viewed as sub-sampling weights, which multiplied by the sampling weight vectors $\mathbf{w}_{s_1}$ and $\mathbf{w}_{s_2}$ of s_1 and s_2, respectively, may give rise to estimates based on only one of these sub-samples. In particular, $1/(1-\phi)\mathbf{w}_{s_2}$ can be viewed as the second-phase sampling weight of s_2. The vector $\mathbf{x}_2$ for case I and the vectors $\mathbf{x}_2$ and $\mathbf{y}_1$ for case II will be used in the proposed extended calibration procedure, which includes also the vector $\mathbf{x}_1$, to "borrow strength" from s_1 when estimation is based on s_2. To distinguish the different types of the common auxiliary information in this calibration procedure, we re-denote $\mathbf{x}_1$ by $\mathbf{x}$, and $(\mathbf{x}_2,\, \mathbf{y}_1)$ by $\mathbf{z}$. For the p-dimensional vector $\mathbf{x}$, the vector of population totals $\mathbf{t}_\mathbf{x}$ is known, whereas the vector of population totals $\mathbf{t}_\mathbf{z}$ is unknown. The same regression setup will be used for cases I and II, when $\mathbf{z} = \mathbf{x}_2$ and $\mathbf{z} = (\mathbf{x}_2, \mathbf{y}_1)$, respectively.

A simultaneous regression for the two samples using the setup $\mathbf{X}_s = diag(\mathbf{X}_{s_i})$, $\mathbf{\Lambda}_s = diag(\mathbf{\Lambda}_{s_i})$, $\mathbf{w}_s = (\mathbf{w}'_{s_1}, \mathbf{w}'_{s_2})'$, $\mathbf{t} = (\mathbf{t}'_\mathbf{x}, \mathbf{t}'_\mathbf{x})'$, generates a vector of calibrated weights, $\mathbf{c}_{xs}$, given by

$$\mathbf{c}_{xs} = \begin{pmatrix} \mathbf{w}_{s_1} \\ \mathbf{w}_{s_2} \end{pmatrix} + \begin{pmatrix} \mathbf{\Lambda}_{s_1}\mathbf{X}_{s_1}(\mathbf{X}'_{s_1}\mathbf{\Lambda}_{s_1}\mathbf{X}_{s_1})^{-1}[\mathbf{t}_\mathbf{x} - \mathbf{X}'_{s_1}\mathbf{w}_{s_1}] \\ \mathbf{\Lambda}_{s_2}\mathbf{X}_{s_2}(\mathbf{X}'_{s_2}\mathbf{\Lambda}_{s_2}\mathbf{X}_{s_2})^{-1}[\mathbf{t}_\mathbf{x} - \mathbf{X}'_{s_2}\mathbf{w}_{s_2}] \end{pmatrix}. \tag{4}$$

The two components of $\mathbf{c}_{xs}$ give rise to the independent GREG estimators $\hat{\mathbf{Z}}_1^R = (1/\phi)\mathbf{Z}'_{s_1}\mathbf{c}_{xs_1}$ and $\hat{\mathbf{Z}}_2^R = [1/(1-\phi)]\mathbf{Z}'_{s_2}\mathbf{c}_{xs_2}$ of the total $\mathbf{t}_\mathbf{z}$ of the type shown in (3). Both these estimators incorporate the auxiliary information in $\mathbf{x}$. Obviously, the GREG estimator based on the entire sample s is given by the combination $\phi\hat{\mathbf{Z}}_1^R + (1-\phi)\hat{\mathbf{Z}}_2^R$. Combining information on $\mathbf{z}$ is accomplished by adding to the regression procedure the calibration constraint that the two estimators of $\mathbf{t}_\mathbf{z}$ are calibrated to each other, that is, they are aligned. This involves the extended regression matrix and the corresponding vector of control totals

$$\boldsymbol{\mathcal{X}}_s = \begin{pmatrix} \mathbf{X}_{s_1} & \mathbf{0} & (1-\phi)\mathbf{Z}_{s_1} \\ \mathbf{0} & \mathbf{X}_{s_2} & -\phi\mathbf{Z}_{s_2} \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} \phi\mathbf{t}_\mathbf{x} \\ (1-\phi)\mathbf{t}_\mathbf{x} \\ \mathbf{0} \end{pmatrix}. \tag{5}$$

Among the many components of the vectors $\mathbf{x}_2$ and $\mathbf{y}_1$ we include in the calibration procedure those most correlated with the variable $\mathbf{y_2}$. Let the dimension of $\mathbf{z}$ be q. Assume then that $(\mathbf{X}_{s_i}\ \mathbf{Z}_{s_i})$ is of full rank $p+q$ and write $\boldsymbol{\mathcal{X}}_s$ in partition form as $\boldsymbol{\mathcal{X}}_s = (\mathbf{X}_s\ \boldsymbol{\mathcal{Z}}_s)$, where $\mathbf{X}_s$ and $\boldsymbol{\mathcal{Z}}_s$ are of dimension $(n_1+n_2)\times 2p$ and $(n_1+n_2)\times q$, respectively. Reset the default value $q_{ik} = 1$ in the entries of $\mathbf{\Lambda}_{s_i}$ to $q_{1k} = 1/\phi$ for every unit k of s_1 and to $q_{2k} = 1/(1-\phi)$ for every unit k of s_2. Next let $\mathbf{L}_s = \mathbf{\Lambda}_s(\mathbf{I} - \mathbf{P}_{\mathbf{X}_S})$, with $\mathbf{P}_{\mathbf{X}_S} = \mathbf{X}_s(\mathbf{X}'_s\mathbf{\Lambda}_s\mathbf{X}_s)^{-1}\mathbf{X}'_s\mathbf{\Lambda}_s$, and notice that $\mathbf{X}_S = diag(\mathbf{X}_{s_i})$ implies $\mathbf{L}_S = diag(\mathbf{L}_{s_i})$, where $\mathbf{L}_{s_i} = \mathbf{\Lambda}_{s_i}(\mathbf{I} - \mathbf{P}_{\mathbf{X}_{s_i}})$, in obvious notation for $\mathbf{\Lambda}_{s_i}$ and $\mathbf{P}_{\mathbf{X}_{s_i}}$. Then, in an adaptation of Merkouris (2004), for weight vector $\mathbf{w}_S = (\mathbf{w}'_{s_1}, \mathbf{w}'_{s_2})'$ and weighting matrix $\mathbf{\Lambda}_s = diag(\mathbf{\Lambda}_{s_i})$, the regression procedure based on the partitioned matrix $\boldsymbol{\mathcal{X}}_s$ generates the vector of calibrated weights

$$\begin{aligned} \mathbf{c}_s &= \mathbf{c}_{xs} + \mathbf{L}_s\boldsymbol{\mathcal{Z}}_s(\boldsymbol{\mathcal{Z}}'_s\mathbf{L}_s\boldsymbol{\mathcal{Z}}_s)^{-1}(\mathbf{0} - \boldsymbol{\mathcal{Z}}'_s\mathbf{c}_{xs}) \\ &= \begin{pmatrix} \mathbf{c}_{xs_1} \\ \mathbf{c}_{xs_2} \end{pmatrix} + \begin{pmatrix} (1-\phi)\mathbf{L}_{s_1}\mathbf{Z}_{s_1} \\ -\phi\mathbf{L}_{s_2}\mathbf{Z}_{s_2} \end{pmatrix} \left[(1-\phi)^2\mathbf{Z}'_{s_1}\mathbf{L}_{s_1}\mathbf{Z}_{s_1} + \phi^2\mathbf{Z}'_{s_2}\mathbf{L}_{s_2}\mathbf{Z}_{s_2}\right]^{-1}\left[\phi\mathbf{Z}'_{s_2}\mathbf{c}_{xs_2} - (1-\phi)\mathbf{Z}'_{s_1}\mathbf{c}_{xs_1})\right]. \end{aligned}$$

It is easy to verify that the vector $\mathbf{c}_s$ satisfies all the calibration constraints, namely, $(1/\phi)\mathbf{X}'_{s_1}\mathbf{c}_{s_1} = [1/(1-\phi)]\mathbf{X}'_{s_2}\mathbf{c}_{s_2} = \mathbf{t}_\mathbf{x}$ and $\boldsymbol{\mathcal{Z}}'_s\mathbf{c}_s = (1-\phi)\mathbf{Z}'_s\mathbf{c}_{s_1} - \phi\mathbf{Z}'_s\mathbf{c}_{s_2} = \mathbf{0}$. This explains the use of the coefficients ϕ and $1-\phi$ in 5.

For any single variable y_2 surveyed only in sample s_2, we can obtain a composite GREG estimator $\hat{Y}_2^{CR} = [1/(1-\phi)]\mathbf{Y}'_{s_2}\mathbf{c}_{s_2}$ of its total $\mathbf{t}_{y_2}$ that has the form

$$\hat{Y}_2^{CR} = \hat{Y}_2^R - \hat{\mathbf{B}}_{y_2}[\hat{\mathbf{Z}}_2^R - \hat{\mathbf{Z}}_1^R], \tag{6}$$

where $\hat{\mathbf{B}}_{y_2} = \phi^2\mathbf{Y}'_{s_2}\mathbf{L}_{s_2}\mathbf{Z}_{s_2}[(1-\phi)^2\mathbf{Z}'_{s_1}\mathbf{L}_{s_1}\mathbf{Z}_{s_1} + \phi^2\mathbf{Z}'_{s_2}\mathbf{L}_{s_2}\mathbf{Z}_{s_2}]^{-1}$. The expression of the composite GREG estimator $\hat{Y}_2^{CR}$ in (6) allows a direct comparison with its single-sample counterpart $\hat{Y}_2^R$.

For the q-dimensional common variable $\mathbf{z}$, we have the composite GREG estimator of $\mathbf{t}_\mathbf{z}$ based on s_2

$$\hat{\mathbf{Z}}_2^{CR} = \hat{\mathbf{Z}}_2^R - \hat{\mathbf{B}}[\hat{\mathbf{Z}}_2^R - \hat{\mathbf{Z}}_1^R], \tag{7}$$

where, $\hat{\mathbf{B}} = \phi^2\mathbf{Z}'_{s_2}\mathbf{L}_{s_2}\mathbf{Z}_{s_2}[(1-\phi)^2\mathbf{Z}'_{s_1}\mathbf{L}_{s_1}\mathbf{Z}_{s_1} + \phi^2\mathbf{Z}'_{s_2}\mathbf{L}_{s_2}\mathbf{Z}_{s_2}]^{-1}$. This estimator can be written in the form

$$\hat{\mathbf{Z}}_2^{CR} = \hat{\mathbf{B}}\hat{\mathbf{Z}}_1^R + (\mathbf{I} - \hat{\mathbf{B}})\hat{\mathbf{Z}}_2^R, \tag{8}$$

and is identical, by construction, to the composite estimator $\hat{\mathbf{Z}}_1^{CR}$ based on s_1.

The estimator (8) is of interest in case II, in which the composite vector $\mathbf{z}$ includes some of the components of the common target vector $\mathbf{y}_1$. The expected improvement of precision is greater for these variables than it is for any y_2, for which the additional information from s_1 stems indirectly from its correlation with $\mathbf{z}$. For a component y of $\mathbf{y}_1$ not included in $\mathbf{z}$, we may obtain one composite GREG estimator of the type shown in (6) based on s_2 and one, its counterpart, based on s_1. Each of these estimators incorporates information from the entire sample s through the common vector $\mathbf{z}$, and thus is more efficient than the simple GREG estimator based on the same sub-sample. However, a more efficient estimator making use of all available information in s on the particular variable involves the entire set of calibrated weights $\mathbf{c}_s$ and is given by

$$\mathbf{Y}_s'\mathbf{c}_s = \hat{Y}^{CR} = \phi\hat{Y}_1^R + (1-\phi)\hat{Y}_2^R + [\phi\hat{\mathbf{B}}_{y_1} - (1-\phi)\hat{\mathbf{B}}_{y_2}][\hat{\mathbf{Z}}_2^R - \hat{\mathbf{Z}}_1^R]. \tag{9}$$

The estimator (9) is a weighted average of the estimator (6) and its counterpart based on s_1, and if y is a component of $\mathbf{z}$ the estimator (9) reverts to (8).

The approximate (large sample) design variance of $\hat{Y}_2^{CR}$, denoted by $AV(\hat{Y}_2^{CR})$ is given by

$$AV(\hat{Y}_2^{CR}) = AV(\hat{Y}_2^R) + \mathbf{B}_{y_2}[AV(\hat{\mathbf{Z}}_1^R) + AV(\hat{\mathbf{Z}}_2^R)]\mathbf{B}_{y_2}' - 2\mathbf{B}_{y_2}[AC(\hat{Y}_2^R, \hat{\mathbf{Z}}_2^R)]', \tag{10}$$

where $\mathbf{B}_{y_2} = \phi^2\mathbf{Y}_2'\mathbf{L}_2\mathbf{Z}[(1-\phi)^2\mathbf{Z}'\mathbf{L}_1\mathbf{Z} + \phi^2\mathbf{Z}'\mathbf{L}_2\mathbf{Z}]^{-1}$ is the population counterpart of $\hat{\mathbf{B}}_{y_2}$ and AC denotes approximate covariance. Here $\mathbf{L}_1 = \phi(\mathbf{I} - \mathbf{P}_\mathbf{X})$ and $\mathbf{L}_2 = (1-\phi)(\mathbf{I} - \mathbf{P}_\mathbf{X})$, with $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

The approximate variance of $\hat{\mathbf{Z}}_2^{CR}$ is given by

$$AV(\hat{\mathbf{Z}}_2^{CR}) = \mathbf{B}AV(\hat{\mathbf{Z}}_1^R)\mathbf{B}' + (\mathbf{I} - \mathbf{B})AV(\hat{\mathbf{Z}}_2^R)(\mathbf{I} - \mathbf{B})', \tag{11}$$

where $\mathbf{B} = \phi^2\mathbf{Z}'\mathbf{L}_2\mathbf{Z}[(1-\phi)^2\mathbf{Z}'\mathbf{L}_1\mathbf{Z} + \phi^2\mathbf{Z}'\mathbf{L}_2\mathbf{Z}]^{-1}$ is the population counterpart of $\hat{\mathbf{B}}$.

The composite GREG estimators $\hat{Y}_2^{CR}$ and $\hat{\mathbf{Z}}_2^{CR}$, based on the second-phase sample s_2, incorporate information from s_1, and thus should be more efficient than the simple GREG estimators $\mathbf{Y}_2^R$ and $\hat{\mathbf{Z}}_2^R$. The efficiency gain will be smaller for $\hat{Y}_2^{CR}$, which borrows strength from s_1 indirectly through the correlation of y_2 with $\mathbf{z}$ — it is clear from (10) that the efficiency of $\hat{Y}_2^{CR}$ depends on the strength of correlation between y_2 and $\mathbf{z}$. It is important to note here that the particular specification of the values of q_{ik} in the entries of $\mathbf{\Lambda}_{s_i}$ entails that the coefficients $\hat{\mathbf{B}}_{y_2}$ and $\hat{\mathbf{B}}$, generated implicitly by the regression procedure, account for the differential in sample size between s_1 and s_2. Without this specification of the q_{ik} it would not necessarily follow that the composite estimators $\hat{Y}_2^{CR}$ and $\hat{\mathbf{Z}}_2^{CR}$ would be more efficient than the estimators $\mathbf{Y}_2^R$ and $\hat{\mathbf{Z}}_2^R$. Values of q_{ik} that yield the most efficient composite estimators can be specified for certain sampling designs; see Merkouris (2004). It can be proved in these situations that $\mathbf{B}_{y_2} = AC(\hat{Y}_2^R, \hat{\mathbf{Z}}_2^R)[AV(\hat{\mathbf{Z}}_1^R) + AV(\hat{\mathbf{Z}}_2^R)]^{-1}$ and $\mathbf{B} = AV(\hat{\mathbf{Z}}_2^R)[AV(\hat{\mathbf{Z}}_1^R) + AV(\hat{\mathbf{Z}}_2^R)]^{-1}$, so that these are the optimal (variance minimizing) coefficients in (10) and (11). For instance, (10) becomes $AV(\hat{Y}_2^{CR}) = AV(\hat{Y}_2^R) - AC(\hat{Y}_2^R, \hat{\mathbf{Z}}_2^R)[AV(\hat{\mathbf{Z}}_1^R) + AV(\hat{\mathbf{Z}}_2^R)]^{-1}[AC(\hat{Y}_2^R, \hat{\mathbf{Z}}_2^R)]'$. Clearly then, the stronger the correlation of y_2 with $\mathbf{z}$, and the larger n_1 is relative to n_2, the more efficient $\hat{Y}_2^{CR}$ becomes relative to $\hat{Y}_2^R$. In more general settings, the efficiency gain will be somewhat smaller, as the coefficients $\mathbf{B}_{y_2}$ and $\mathbf{B}$ (incorporating the generic specification $q_{ik} = (n_1 + n_2)/n_i$) will only be approximations of the optimal ones. This is because for general sampling designs, $\mathbf{B}_{y_2}$ and $\mathbf{B}$ may not precisely reflect the relative interaction of design and regression effects between the two samples.

In the present setting, where we assume that the same auxiliary vector $\mathbf{x}$ is used in the simultaneous calibration of s_1 and s_2, we notice that $\mathbf{B}_{y_2} = \phi\mathbf{Y}_2'(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Z}[\mathbf{Z}'(\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Z}]^{-1}$ and $\mathbf{B} = \phi\mathbf{I}$. It follows then from (8) and (9) that for large sample sizes $\hat{\mathbf{Z}}_2^{CR} = \phi\hat{\mathbf{Z}}_1^R + (1-\phi)\hat{\mathbf{Z}}_2^R$ and $\hat{Y}^{CR} = \phi\hat{Y}_1^R + (1-\phi)\hat{Y}_2^{CR}$. The same estimators $\hat{\mathbf{Z}}_2^{CR}$ and $\hat{Y}^{CR}$ would be obtained using only the part $\mathbf{X}_s = diag(\mathbf{X}_{s_i})$ of the design matrix $\boldsymbol{\mathcal{X}}_s$ in (5) for the simultaneous calibration of s_1 and s_2, regardless of the size of these two sub-samples. This is reasonable, because composite estimation through the extended calibration setup (5) does not use any information for the common variables beyond what is available in the first-phase sample s.

4. Calibration for Non-Nested Double Sampling

In this context, where the two separate (assumed independent) samples s_1 and s_2 are drawn from the same frame or from different frames representing the same target population, the sampling weights of both s_1 and s_2 aggregate to the same population total. The sampling designs for s_1 and s_2 may differ. Furthermore, the vectors of auxiliary variables with known population totals may be different in the two samples, with those in sample s_2 being a subset of those in sample s_1.

As in case I of nested double sampling, the vector $\mathbf{x}_2$ with unknown population totals will be used in an extended calibration procedure to "borrow strength" from s_1 in the estimation for the target variables $\mathbf{y}$ surveyed in s_2. Thus,

now $\mathbf{z} = \mathbf{x}_2$. The extended calibration setup is similar to that used in the previous section. Here we use

$$\boldsymbol{\mathcal{X}}_s = \begin{pmatrix} \mathbf{X}_{s_1} & \mathbf{0} & \mathbf{Z}_{s_1} \\ \mathbf{0} & \mathbf{X}_{s_2} & -\mathbf{Z}_{s_2} \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} \mathbf{t}_{\mathbf{x}_1} \\ \mathbf{t}_{\mathbf{x}_2} \\ \mathbf{0} \end{pmatrix}, \tag{12}$$

where the subscript i in $\mathbf{t}_{\mathbf{x}_i}$ indicates possibly different sets of auxiliary variables with known population totals being used in the two samples. We assume then that $(\mathbf{X}_{s_i}\ \mathbf{Z}_{s_i})$ is of full rank $p_i + q$ and write $\boldsymbol{\mathcal{X}}_s$ in partition form as $\boldsymbol{\mathcal{X}}_s = (\mathbf{X}_s\ \boldsymbol{\mathcal{Z}}_s)$, where $\mathbf{X}_s$ and $\boldsymbol{\mathcal{Z}}_s$ are of dimension $(n_1 + n_2) \times (p_1 + p_2)$ and $(n_1 + n_2) \times q$, respectively. The value q_{ik} in the entries of $\boldsymbol{\Lambda}_{s_i}$ is set now to $q_{1k} = \phi$ for every unit k of s_1 and to $q_{2k} = 1 - \phi$ for every unit k of s_2. Here $\phi = \tilde{n}_i/(\tilde{n}_1 + \tilde{n}_2)$, where $\tilde{n}_i = n_i/d_i$ is the effective sample size for s_i — d_i denoting design effect. Then, with weight vector $\mathbf{w}_s = (\mathbf{w}'_{s_1}, \mathbf{w}'_{s_2})'$ and weighting matrix $\boldsymbol{\Lambda}_s = diag(\boldsymbol{\Lambda}_{s_i})$, the regression procedure based on the partitioned matrix $\boldsymbol{\mathcal{X}}_s$ generates the vector of calibrated weights

$$\begin{aligned} \mathbf{c}_s &= \mathbf{c}_{xs} + \mathbf{L}_s \boldsymbol{\mathcal{Z}}_s (\boldsymbol{\mathcal{Z}}'_s \mathbf{L}_s \boldsymbol{\mathcal{Z}}_s)^{-1} (\mathbf{0} - \boldsymbol{\mathcal{Z}}'_s \mathbf{c}_{xs}) \\ &= \begin{pmatrix} \mathbf{c}_{xs_1} \\ \mathbf{c}_{xs_2} \end{pmatrix} + \begin{pmatrix} \mathbf{L}_{s_1} \mathbf{Z}_{s_1} \\ -\mathbf{L}_{s_2} \mathbf{Z}_{s_2} \end{pmatrix} \left[\mathbf{Z}'_{s_1} \mathbf{L}_{s_1} \mathbf{Z}_{s_1} + \mathbf{Z}'_{s_2} \mathbf{L}_{s_2} \mathbf{Z}_{s_2} \right]^{-1} \left[\mathbf{Z}'_{s_2} \mathbf{c}_{xs_2} - \mathbf{Z}'_{s_1} \mathbf{c}_{xs_1}) \right], \end{aligned}$$

where $\mathbf{c}_{xs_i}$ and $\mathbf{L}_{s_i}$ are as in Section 3, except for the respecifications for $\mathbf{t}_{\mathbf{x}_i}$ and q_{ik}. It is easy to verify that the vector $\mathbf{c}_s$ satisfies all the calibration constraints, namely, $\mathbf{X}'_{s_i} \mathbf{c}_{s_i} = \mathbf{t}_{\mathbf{x}_i}$ and $\boldsymbol{\mathcal{Z}}'_s \mathbf{c}_s = \mathbf{Z}'_s \mathbf{c}_{s_1} - \mathbf{Z}'_s \mathbf{c}_{s_2} = \mathbf{0}$.

For any single variable y surveyed in sample s_2, we can obtain a composite GREG estimator $\hat{Y}^{CR} = \mathbf{Y}'_{s_2} \mathbf{c}_{s_2}$ of its total $\mathbf{t}_y$ that has the form

$$\hat{Y}^{CR} = \hat{Y}^R - \hat{\mathbf{B}}_y [\hat{\mathbf{Z}}_2^R - \hat{\mathbf{Z}}_1^R], \tag{13}$$

where $\hat{\mathbf{B}}_y = \mathbf{Y}'_{s_2} \mathbf{L}_{s_2} \mathbf{Z}_{s_2} [\mathbf{Z}'_{s_1} \mathbf{L}_{s_1} \mathbf{Z}_{s_1} + \mathbf{Z}'_{s_2} \mathbf{L}_{s_2} \mathbf{Z}_{s_2}]^{-1}$. The approximate design variance of $\hat{Y}^{CR}$ is given by

$$AV(\hat{Y}^{CR}) = AV(\hat{Y}^R) + \mathbf{B}_y [AV(\hat{\mathbf{Z}}_1^R) + AV(\hat{\mathbf{Z}}_2^R)] \mathbf{B}'_y - 2\mathbf{B}_y [AC(\hat{Y}_2^R, \hat{\mathbf{Z}}_2^R)]', \tag{14}$$

where $\mathbf{B}_y = \mathbf{Y}' \mathbf{L}_2 \mathbf{Z} [\mathbf{Z}' \mathbf{L}_1 \mathbf{Z} + \mathbf{Z}' \mathbf{L}_2 \mathbf{Z}]^{-1}$ is the population counterpart of $\hat{\mathbf{B}}_y$. Here $\mathbf{L}_1 = (1/\phi)(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})$ and $\mathbf{L}_2 = [1/(1-\phi)](\mathbf{I} - \mathbf{P}_{\mathbf{X}_2})$. As in nested double sampling, values of q_{ik} that yield the optimal (variance minimizing) coefficient $\mathbf{B}_{y_2} = AC(\hat{Y}_2^R, \hat{\mathbf{Z}}_2^R)[AV(\hat{\mathbf{Z}}_1^R) + AV(\hat{\mathbf{Z}}_2^R)]^{-1}$ in (14) can be specified for certain sampling designs. These values of q_{ik} are the inverses of the values specified in nested double sampling. As in nested double sampling, the stronger the correlation of y_2 with $\mathbf{z}$, and the larger n_1 is relative to n_2, the more efficient $\hat{Y}^{CR}$ becomes relative to $\hat{Y}^R$.

For the q-dimensional common variable $\mathbf{z}$, we have the composite GREG estimator of $\mathbf{t}_\mathbf{z}$ based on s_2

$$\hat{\mathbf{Z}}_2^{CR} = \hat{\mathbf{B}} \hat{\mathbf{Z}}_1^R + (\mathbf{I} - \hat{\mathbf{B}}) \hat{\mathbf{Z}}_2^R, \tag{15}$$

where, $\hat{\mathbf{B}} = \mathbf{Z}'_{s_2} \mathbf{L}_{s_2} \mathbf{Z}_{s_2} [\mathbf{Z}'_{s_1} \mathbf{L}_{s_1} \mathbf{Z}_{s_1} + \mathbf{Z}'_{s_2} \mathbf{L}_{s_2} \mathbf{Z}_{s_2}]^{-1}$. In the present context the estimator (15) is not of interest in itself because $\mathbf{z}$ is the auxiliary vector $\mathbf{x}_2$. It is, however, instructive to notice that (13) can be written as

$$\hat{Y}^{CR} = \hat{Y}^R + \mathbf{Y}'_{s_2} \mathbf{L}_{s_2} \mathbf{Z}_{s_2} [\mathbf{Z}'_{s_2} \mathbf{L}_{s_2} \mathbf{Z}_{s_2}]^{-1} [\hat{\mathbf{Z}}_2^{CR} - \hat{\mathbf{Z}}_2^R]. \tag{16}$$

This is the GREG estimator that would be obtained in an extended calibration of sample s_2 only, in which in addition to the control total $\mathbf{t}_{\mathbf{x}_2}$ the composite estimator $\hat{\mathbf{Z}}_2^{CR}$ would be used as control total for the variable $\mathbf{z}$.

5. Concluding Remarks

The proposed one-step calibration procedure, involving the first-phase sample s in nested double sampling or the combined sample $s_1 \cup s_2$ in non-nested double sampling, generates a single set of calibrated weights that can be used to produce a composite estimate for any variable of interest. This is especially convenient in case II of nested double sampling, where incorporating information from s_1 into the estimates for $\mathbf{z}$ is integrated in the same estimation system thereby preserving the internal consistency of all estimates. Obviously, estimation of the variance of the resulting estimators is done using the same technique as when working with only one sample.

An alternative calibration method for incorporating auxiliary information from s_1 into estimators based on the second-phase sample s_2 involves two steps. First, a composite GREG estimator is formed for the total $\mathbf{t}_\mathbf{z}$, using all information in s_1 and s_2, which in turn is used as additional control total in the calibration of s_2. This works as described in the remark following equation (15), in the context of non-nested double sampling. In a similar but more general context of combining information from different surveys through regression, Renssen and Nieuwenbroek (1997) proposed a family of composite GREG estimators for $\mathbf{t}_\mathbf{z}$ that could be used as alternatives to $\hat{\mathbf{Z}}_2^{CR}$ in such a

two-step procedure, resulting in a composite estimator of $\mathbf{t}_y$ of the same form as in (16). In the simplest case when the estimator $\hat{\mathbf{Z}}_2^{CR} = \phi\hat{\mathbf{Z}}_1^R + (1-\phi)\hat{\mathbf{Z}}_2^R$ is used, the two methods give exactly the same composite estimator (16) for large samples and when the vectors of control totals $\mathbf{t}_{\mathbf{x}_1}$ and $\mathbf{t}_{\mathbf{x}_2}$ are identical. The disadvantage of the method of Renssen and Nieuwenbroek is the difficulty in forming a more efficient composite estimator for $\mathbf{t}_\mathbf{z}$ in more general settings and the inconvenience of the two-step calibration procedure. Moreover, variance estimation by resampling techniques (e.g., jackknife), typically used in surveys with complex designs, is very inconvenient or even impossible with this alternative method; see Merkouris (2004) for an extensive discussion of this issue.

An alternative estimation method for the variants of nested double sampling considered in this paper is similar to that used in the standard nested double sampling (where the samples s_1 and s_2 are dependent). In this method, the ordinary GREG estimator based on the entire first-phase sample s may be calculated first, and then used as additional control total in the calibration of s_2 in case I or s in case II. This is similar to the alternative method described above for the non-nested case. Recalling that s_1 and s_2 are independent and representative of the same population, the GREG estimator based on s can be obtained as $\hat{\mathbf{Z}}_2^{CR} = \phi\hat{\mathbf{Z}}_1^R + (1-\phi)\hat{\mathbf{Z}}_2^R$, using the design matrix $\mathbf{X}_s = diag(\mathbf{X}_{s_i})$. As mentioned in the end of Section 3, in large samples this procedure gives the estimators (8) and (9). Without the diagonal structure in the design matrix $\mathbf{X}_s$, this alternative method gives estimators that are different from (8) and (9) in form but only slightly different in efficiency. Practical disadvantages similar to those mentioned in the non-nested double sampling are encountered also in the application of this alternative to the nested double sampling.

As in standard nested double sampling, it has been assumed that in cases I and II the same auxiliary vector with known population total $\mathbf{t}_\mathbf{x}$ is used in the calibration of both s_1 and s_2. However, in the proposed method we may have to use a subset of the auxiliary variables in the calibration of the second-phase sample s_2 when its size is not large enough. In this situation, the efficiency of the proposed method relative to the alternative mentioned above needs further investigation.

REFERENCES

Deville, J. C., and Särndal, C. E. (1992), "Calibratin Estimators in Survey Sampling," *Journal of the American Statistical Association*, 87, 376–382.

Estevao, V. M., and Särndal, C. E. (2004), "The Ten Cases of Auxiliary Information for Calibration in Two-Phase Sampling," *Journal of Official Statistics*, 18, 645–669.

Hidiroglou, M. (2001), "Double Sampling," *Survey Methodology*, 27, 143–154.

Hidiroglou, M., and Särndal, C. E. (1998), "Use of Auxiliary Information for Two-Phase Sampling," *Survey Methodology*, 24, 11–20.

Merkouris, T. (2004), "Combining Independent Regression Estimators from Multiple Surveys," *Journal of the American Statistical Association*, 99, 1131–1139.

Renssen, R. H., and Nieuwenbroek, N. J. (1997), "Aligning Estimates for Common Variables in Two or More Sample Surveys," *Journal of the American Statistical Association*, 92, 368–375.

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