Statistica Sinica **24** (2014), 1001-1015
doi:http://dx.doi.org/10.5705/ss.2013.038

INSTRUMENTAL-VARIABLE CALIBRATION ESTIMATION IN SURVEY SAMPLING

Seunghwan Park and Jae Kwang Kim

Seoul National Univeristy and Iowa State University

Abstract: The prediction model, which makes effective use of auxiliary information available throughout the population, is often used to derive efficient estimation in survey sampling. To protect against failure of the assumed model, asymptotic design unbiasedness is often imposed in the prediction estimator. An instrumental-variable calibration estimator can be considered to achieve the model optimality among the class of calibration estimators that is asymptotically design unbiased.

In this paper, we propose a new calibration estimator that is asymptotically equivalent to the optimal instrumental-variable calibration estimator. The resulting weights are no smaller than one and can be constructed to achieve the range restrictions. The proposed method can be extended to calibration estimation under two-phase sampling. Some numerical results are presented using the data from the 1997 National Resource Inventory of the United States.

Key words and phrases: Asymptotic design unbiasedness, exponential tilting, regression estimation, weighting.

1. Introduction

Consider a finite population of size N. Let y be the variable of interest with value y_i for unit i in the population. Suppose we are interested in estimating the population total $Y = \sum_{i=1}^{N} y_i$. Assume that a probability sampling design is used to select a sample from the finite population. Let A be the set of sample indices realized from the sampling design. Let $\pi_i = P(i \in A)$ be the first-order inclusion probability of unit i. The Horvitz-Thompson estimator

$$\hat{Y}_{HT} = \sum_{i \in A} d_i y_i$$

is unbiased for $Y = \sum_{i=1}^{N} y_i$, where $d_i = \pi_i^{-1}$ is the design weight for unit i.

Now, suppose that a p-dimensional auxiliary vector $\mathbf{x}_i$ is available from the entire population. In this case, one can postulate a superpopulation model that describes the structural relationship between y_i and $\mathbf{x}_i$ in the population. For example, the linear regression model

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + e_i, \tag{1.1}$$

with $e_i \sim (0, v_i\sigma^2)$, can be imposed where $\boldsymbol{\beta}$ and σ^2 are unknown model parameters and $v_i = v(\mathbf{x}_i)$ is a known function of $\mathbf{x}_i$. We assume that model (1.1) holds for the sampled part and also for the non-sampled part of the population. The model can be used to build a prediction estimator for Y:

$$\hat{Y}_p = \sum_{i \in A} y_i + \sum_{i \in A^c} \hat{y}_i \tag{1.2}$$

where $\hat{y}_i$ satisfies $E(\hat{y}_i - y_i \mid I_i = 0) \cong 0$, where $I_i = 1$ if $i \in A$ and $I_i = 0$ otherwise. Under (1.1), we can use

$$\hat{y}_i = \mathbf{x}_i' \left(\sum_{i \in A} q_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i \in A} q_i \mathbf{x}_i y_i \tag{1.3}$$

for some q_i. Brewer (1963), Royall (1970, 1976), and others have adapted the linear model prediction theory to the finite population situation and have derived the best linear unbiased predictor (BLUP). When the sampling design is noninformative (Pfeffermann (1993)), the BLUP of Y under (1.1) can be obtained by the choice of $q_i = 1/v_i$ in (1.3). However, the prediction estimator is not necessarily justified if the regression model does not hold. To guarantee asymptotic design unbiasedness (ADU) of the prediction estimator, Brewer (1979) suggested using $q_i = (d_i - 1)$ in (1.3). Isaki and Fuller (1982) suggested using $q_i = d_i^2$ in (1.3) and discussed conditions for the ADU. Wright (1983) also showed that the BLUP of Y can satisfy the ADU property if and only if $v_i(d_i - 1) = \mathbf{a}'\mathbf{x}_i$ for some p-dimensional constant vector $\mathbf{a}$.

The prediction estimator (1.2) with the predictor in (1.3) can be written as $\hat{Y}_p = \sum_{i \in A} w_i y_i$ where w_i satisfies

$$\sum_{i \in A} w_i \mathbf{x}_i = \sum_{i=1}^{N} \mathbf{x}_i. \tag{1.4}$$

The equation (1.4) is often called the calibration equation. The weights satisfying (1.4) are often called calibration weights and the estimator using the calibration weights is called the calibration estimator. Deville and Särndal (1992), Fuller (2002), and Kim and Park (2010) provided comprehensive overviews for calibration estimation.

In this paper, we investigate the ADU property of the prediction estimator in a more general class of prediction estimators under the regression superpopulation model in (1.1). The proposed estimator can be directly written as a calibration estimator and some optimal choice of instrumental variable is discussed. Furthermore, the proposed estimator is extended to construct a two-step calibration estimator for two-phase sampling. Some numerical results are presented.

2. Prediction Estimation

In this section, we briefly review the ADU theory of the prediction estimator and discuss some choices of q_i in (1.3). Here, we assume that the only available information in the finite population is $\mathbf{X}$. The prediction estimator in (1.2) can be written as

$$\hat{Y}_p = \sum_{i=1}^{N} \hat{y}_i + \sum_{i \in A} d_i \left(y_i - \hat{y}_i\right), \tag{2.1}$$

if the predicted values are constructed to satisfy

$$\sum_{i \in A} \left(d_i - 1\right)\left(y_i - \hat{y}_i\right) = 0. \tag{2.2}$$

Since the estimator for form (2.1) satisfies ADU conditions in general, we have only to impose (2.2) in computing the predicted values. Condition (2.2) is referred to as *Internally Bias Calibrated* (IBC) condition by Firth and Bennett (1998) and IBC is a sufficient condition for the ADU property in the prediction estimator.

By construction, the prediction estimator with $\hat{y}_i$ computed by (1.3) satisfies

$$\sum_{i \in A} \left(y_i - \hat{y}_i\right) \mathbf{x}_i q_i = 0.$$

Thus, the prediction estimator using $\hat{y}_i$ in (1.3) satisfies ADU if $\mathbf{x}_i$ contains $q_i^{-1}(d_i - 1)$, which is consistent with the result of Wright (1983) for the particular choice of $q_i = v_i^{-1}$.

We consider a more general class of prediction estimators of form (1.2) with

$$\hat{y}_i = \mathbf{x}_i' \left(\sum_{i \in A} \mathbf{z}_i \mathbf{x}_i' \right)^{-1} \sum_{i \in A} \mathbf{z}_i y_i \tag{2.3}$$

for some $\mathbf{z}_i = \mathbf{z}(\mathbf{x}_i, d_i, v_i)$ such that $\sum_{i \in A} \mathbf{z}_i \mathbf{x}_i'$ is nonsingular. By construction, the predicted values in (2.3) satisfy

$$\sum_{i \in A} \mathbf{z}_i (y_i - \hat{y}_i) = \mathbf{0}.$$

Thus, the prediction estimator using the predicted values in (2.3) satisfies ADU if $\mathbf{z}_i$ contains $(d_i - 1)$. The prediction estimator using (2.3) also satisfies the calibration condition in (1.4) as it can be written as $\hat{Y}_p = \sum_{i \in A} w_i y_i$ with

$$w_i = 1 + \left(\sum_{i \in A^c} \mathbf{x}_i' \right) \left(\sum_{i \in A} \mathbf{z}_i \mathbf{x}_i' \right)^{-1} \mathbf{z}_i. \tag{2.4}$$

Calibration estimators using the weights of the form in (2.4) are sometimes called instrumental-variable (IV) calibration estimators (Estavo and Särndal (2000);

Kott (2006); Kim (2010)). Variable $\mathbf{z}_i$ is the instrumental variable and can be chosen to improve statistical efficiency.

To discuss the optimal choice of $\mathbf{z}_i = \mathbf{z}(\mathbf{x}_i, d_i, v_i)$ in the IV calibration estimator, let B_z be the probability limit of $\hat{\boldsymbol{\beta}}_z = \left(\sum_{i\in A} \mathbf{z}_i\mathbf{x}_i'\right)^{-1}\sum_{i\in A}\mathbf{z}_i y_i$. If $\mathbf{z}_i$ contains $(d_i - 1)$, then we can write $(d_i - 1) = \mathbf{a}'\mathbf{z}_i$ for some $\mathbf{a}$ and

$$\begin{aligned} N^{-1}\sum_{i\in A^c}\left(y_i - \mathbf{x}_i'B_z\right) &= p\lim\left\{N^{-1}\sum_{i\in A}(d_i-1)\left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_z\right)\right\} \\ &= p\lim\left\{N^{-1}\sum_{i\in A}\mathbf{a}'\mathbf{z}_i\left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_z\right)\right\} = 0. \end{aligned}$$

Thus, we can write

$$\begin{aligned} \hat{Y}_p - Y &= \sum_{i\in A^c}\mathbf{x}_i'\left(\hat{\boldsymbol{\beta}}_z - B_z\right) \\ &= \left(\sum_{i\in A^c}\mathbf{x}_i'\right)\left(\sum_{i\in A}\mathbf{z}_i\mathbf{x}_i'\right)^{-1}\left\{\sum_{i\in A}\mathbf{z}_i\left(y_i - \mathbf{x}_i'B_z\right)\right\}. \end{aligned} \tag{2.5}$$

The anticipated variance (AV) of $\hat{Y}_p$, defined by $AV(\hat{Y}_p) = E\{(\hat{Y}_p - Y)^2\}$ where the expectation is taken with respect to the joint distribution of the superpopulation model (1.1) and the sampling mechanism, is

$$AV(\hat{Y}_p) \cong E\left\{\left(\sum_{i\in A^c}\mathbf{x}_i'\right)\left(\sum_{i\in A}\mathbf{z}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i\in A}\mathbf{z}_i\mathbf{z}_i'v_i\right)\left(\sum_{i\in A}\mathbf{x}_i\mathbf{z}_i'\right)^{-1}\left(\sum_{i\in A^c}\mathbf{x}_i\right)\right\}. \tag{2.6}$$

Because $E\left\{\left(y_i - \mathbf{x}_i'B_z\right)^2\right\} \cong E\left\{\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2\right\} = v_i$, the anticipated variance in (2.6) is minimized with respect to $\mathbf{z}_i$, $i \in A$, at $\mathbf{z}_i^* = \mathbf{x}_i/v_i$, $i \in A$; see Section S1 of the Supplementary Note. Therefore, if $\mathbf{x}_i/v_i$ contains $(d_i - 1)$, then the IV calibration estimator using the weight in (2.4) satisfies ADU and achieves the minimum AV.

If $\mathbf{x}_i/v_i$ does not contain $(d_i - 1)$, we no longer have the ADU property. To obtain a design-consistent IV calibration estimator that is close to optimal in the sense of minimizing the anticipated variance in (2.6), we can impose the ADU condition into the instrumental variable $\mathbf{z}_i$ by choosing $\mathbf{z}_i' = (z_{0i}, \mathbf{z}'_{1i})$ where $z_{0i} = d_i - 1$ and $\mathbf{z}_{1i} = \mathbf{x}_{1i}/v_i$, with $\mathbf{x}_i' = (1, \mathbf{x}_{1i}')$. In this case, it is equivalent to the prediction estimator for form (1.2) with

$$\hat{y}_i = \left(1, \mathbf{x}_{1i}'\right)\left\{\sum_{i\in A}\begin{pmatrix} d_i - 1 \\ \mathbf{x}_{1i}/v_i\end{pmatrix}'\begin{pmatrix}1 \\ \mathbf{x}_{1i}\end{pmatrix}\right\}^{-1}\sum_{i\in A}\begin{pmatrix} d_i - 1 \\ \mathbf{x}_{1i}/v_i\end{pmatrix} y_i. \tag{2.7}$$

The prediction estimator using (2.7) was originally proposed by Brewer, Muhammad, and Tam (1988). The optimal regression estimator considered by

Isaki and Fuller (1982), which uses $\hat{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_2$ where $\hat{\boldsymbol{\beta}}_2 = \left(\sum_{i\in A}\pi_i^{-2}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}$ $\sum_{i\in A}\pi_i^{-2}\mathbf{x}_i y_i$, under a sampling design with $\pi_i \propto v_i^{1/2}$ also achieves the minimum AV. Note that the prediction estimator using $\hat{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_2$ also belongs to the class of the IV calibration estimator for form (2.4). As pointed out by Isaki and Fuller (1982), the ADU property holds for the choice of $\mathbf{x}_i = (\pi_i, \pi_i^2, \mathbf{x}_{1i})$ when $\hat{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_2$ is used in (1.2).

The prediction estimator using $\hat{y}_i$ in (2.7), which can also be written as an IV calibration estimator using (2.4), has such nice statistical properties as ADU and some optimality under (1.1). However, such optimality is not tenable for multipurpose sampling because there are many y-variables in the survey. The IV calibration estimator using (2.4) can be quite inefficient for some y_i if the working model (1.1) is far from the true model. Furthermore, the weights in (2.4) can take extreme values and some modification is needed to guarantee that they satisfy some range restriction.

3. Proposed Method

We consider a calibration estimator that can be viewed as a prediction estimator under (1.1) and with some range restriction in the weights. The calibration weights in (2.4) can take negative values, and avoiding this has been an important practical problem in survey sampling (Huang and Fuller (1978)).

Given the instrumental variable in (2.5), the proposed calibration estimator is

$$\hat{Y}_{cal,p} = \sum_{j\in A}\left\{1 + (d_j - 1)\exp\left(\hat{\lambda}_0 + \hat{\boldsymbol{\lambda}}_1'\mathbf{z}_{1j}^*\right)\right\} y_j, \tag{3.1}$$

where $\mathbf{z}_{1j}^* = \mathbf{z}_{1j}/(d_j - 1)$, $\exp(-\hat{\lambda}_0) = (N-n)^{-1}\sum_{j\in A}(d_j - 1)\exp(\hat{\boldsymbol{\lambda}}_1'\mathbf{z}_{1j}^*)$, and $\hat{\boldsymbol{\lambda}}_1$ satisfies (1.4). This estimator is a modified version of the exponential tilting calibration estimator

$$\hat{Y}_{ET} = \sum_{j\in A} d_j\exp(\hat{\lambda}_0 + \hat{\boldsymbol{\lambda}}_1'\mathbf{z}_{1j}^*)y_j$$

considered in Kim (2010). The proposed calibration weights satisfy $\tilde{w}_i \geq 1$, which makes sense because a unit in the sample represents at least one unit in the population. Some computational details for finding $(\hat{\lambda}_0, \hat{\boldsymbol{\lambda}}_1)$ are discussed in Section S2 of Supplementary Note.

Using Taylor linearization, we can show that $\hat{Y}_{cal,p}$ satisfies

$$N^{-1}(\hat{Y}_{cal,p} - \hat{Y}_p) = o_p(n^{-1/2}). \tag{3.2}$$

where

$$\hat{Y}_p = \sum_{i\in A} y_i + \sum_{i\in A^c}\hat{y}_i$$

with $\hat{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_z$ and

$$\hat{\boldsymbol{\beta}}_z = \left\{\sum_{i\in A}\begin{pmatrix} d_i - 1 \\ \mathbf{z}_{1i}\end{pmatrix}\begin{pmatrix} 1 \\ \mathbf{x}_{1i}\end{pmatrix}'\right\}^{-1}\sum_{i\in A}\begin{pmatrix} d_i - 1 \\ \mathbf{z}_{1i}\end{pmatrix} y_i.$$

A proof of (3.2) is a straightforward application of Theorem 1 of Kim and Park (2010) and is sketched in Section S3 of Supplementary Note. By (3.2), the proposed calibration estimator is asymptotically equivalent to a prediction estimator using an instrumental variable $\mathbf{z}_i = (d_i - 1, \mathbf{z}_{1i})$. The first component, $d_i - 1$, is needed for the ADU property and the other component can be chosen to improve efficiency.

Calibration weight is an exponential function of $\mathbf{z}_{1i}^* = \mathbf{z}_{1i}/(d_i - 1)$,

$$w_i = 1 + (d_i - 1)\exp\left(\hat{\lambda}_0 + \hat{\boldsymbol{\lambda}}_1'\frac{\mathbf{z}_{1i}}{d_i - 1}\right) \tag{3.3}$$

and extreme values of $\mathbf{z}_{1i}/(d_i - 1)$ can lead to extreme weights. As a remedy, instead of using $\mathbf{z}_{1i} = \mathbf{x}_{1i}/v_i$ for optimal estimation, one can take c_i in $\mathbf{z}_{1i} = \mathbf{x}_{1i}/(v_i c_i)$ so that

$$\frac{w_i - 1}{d_i - 1} = \exp\left(\hat{\lambda}_0 + \hat{\boldsymbol{\lambda}}_1'\frac{\mathbf{z}_{1i}}{d_i - 1}\right) < K, \tag{3.4}$$

for a predetermined upper bound K. The choice of $c_i = 1$ gives us back the best prediction estimator not necessarily satisfying the range restriction in the final weights. Roughly, K can be in the range of three to five. The proposed method, in line with the range restriction $w_i \in [1, K)$, uses $\mathbf{z}_{1i} = \mathbf{x}_{1i}/(v_i c_i^*)$, where $c_i^* = 1$ if it satisfies (3.4). Otherwise c_i^* is the value that makes the ratio $(w_i - 1)/(d_i - 1)$ equal to K. Use of instrumental variable for range restricted calibration estimation was also considered by Kott (2011).

For variance estimation, first assume that there is a consistent estimator for the variance of $\hat{Y}_d$ given by

$$\hat{V}(\hat{Y}_d) = \sum_{i\in A}\sum_{j\in A}\Omega_{ij}y_i y_j \tag{3.5}$$

and that $\hat{V}(\hat{Y}_d)$ satisfies

$$\frac{\hat{V}(\hat{Y}_d)}{V(\hat{Y}_d)} = 1 + o_p(1),$$

for any y with bounded fourth moment. Using (3.2), the asymptotic variance of $\hat{Y}_{cal,p}$ is asymptotically equivalent to the variance of $\hat{Y}_p$. Moreover, $\hat{Y}_p$ can be expressed as $\hat{Y}_p = \hat{Y}_d + (\mathbf{X} - \hat{\mathbf{X}}_d)'\hat{\boldsymbol{\beta}}_z$, since $\hat{y}_i$ satisfies (2.2). The asymptotic variance of the $\hat{Y}_p$ has the form

Table 1. Data structure for two-phase sampling.

Set	Size	Observation
Population (U)	N	$\mathbf{x}_{1i}$
First-phase sample (A_1)	n_1	$\mathbf{x}_i = (\mathbf{x}_{1i}, \mathbf{x}_{2i})$
Second-phase sample (A_2)	n_2	$\mathbf{x}_i, y_i$

$$V(\hat{Y}_p) \cong V\Big\{ \sum_{i \in A} d_i \left(y_i - \mathbf{x}_i' B_z \right) \Big\}. \tag{3.6}$$

Thus the estimator for the variance of $\hat{Y}_{cal,p}$ has the form

$$\hat{V}(\hat{Y}_{cal,p}) = \sum_{i \in A} \sum_{j \in A} \Omega_{ij} g_i g_j \left(y_i - \mathbf{x}_i' \hat{\beta}_z \right) \left(y_j - \mathbf{x}_j' \hat{\beta}_z \right), \tag{3.7}$$

where $g_i = w_i / d_i$, with w_i provided by (3.3).

4. Calibration for Two-Phase Sampling

We discuss an extension of the proposed method to two-phase sampling. Two-phase sampling, or double sampling, is a cost-effective technique widely used in survey sampling (Hidiroglou (2001)); Rao (1973); Kim, Navarro, and Fuller (2006); Fuller (2003)). In two-phase sampling, a first-phase sample A_1 with size n_1 is drawn from the population U under a sampling design with the first-order inclusion probabilities π_{1k}. Given the first-phase sampling A_1, a second-phase sample A_2 with size n_2 is drawn from A_1 under a sampling scheme with the first-order conditional probabilities $\pi_{2k|A_1} = \pi_{2k|1k}$. We denote the first-phase design weight of unit k as $d_{1k} = \pi_{1k}^{-1}$, the second-phase conditional design weight of unit k as $d_{2k|1k} = \pi_{2k|1k}^{-1}$, and the final design weight as $d_{2k} = d_{1k} d_{2k|1k}$. Assume that there is a vector of auxiliary variables that can be partitioned as $\mathbf{x}_k' = (\mathbf{x}_{1k}, \mathbf{x}_{2k})'$, with $\mathbf{x}_{1k}$ observed for the entire population and $\mathbf{x}_{2k}$ observed up to the first-phase sample. The study variable y is observed only in the second-phase sample A_2. Table 1 presents the data structure of two-phase sampling.

In this two-phase sampling setup, we can consider a prediction estimator for $Y = \sum_{i=1}^{N} y_i$ of the form

$$\hat{Y}_{tp,p} = \sum_{j \in A_2} y_j + \sum_{j \in A_1 / A_2} \hat{y}_{2j} + \sum_{j \in U / A_1} \hat{y}_{1j}, \tag{4.1}$$

where $\hat{y}_{1j} = \mathbf{x}_{1j}' \hat{\mathbf{B}}_1$, $\hat{y}_{2j} = \mathbf{x}_j' \hat{\boldsymbol{\beta}}$ with

$$\hat{\mathbf{B}}_1 = \left(\sum_{j \in A_1} \frac{\mathbf{x}_{1j} \mathbf{x}_{1j}'}{v_j} \right)^{-1} \left\{ \sum_{j \in A_2} \frac{\mathbf{x}_{1j} y_j}{v_j} + \sum_{j \in A_1 / A_2} \frac{\mathbf{x}_{1j} \hat{y}_{2j}}{v_j} \right\},$$

$$\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2')' = \left(\sum_{j \in A_2} \frac{\mathbf{x}_j \mathbf{x}_j'}{v_j} \right)^{-1} \sum_{j \in A_2} \frac{\mathbf{x}_j y_j}{v_j}.$$

Here, the implicit model is

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + e_i \tag{4.2}$$

with $e_i \sim (0, v_i)$. By the definition of $\hat{y}_{2j}$, we can write $\hat{\mathbf{B}}_1 = \hat{\boldsymbol{\beta}}_1 + \hat{\boldsymbol{\beta}}_{1,x} \hat{\boldsymbol{\beta}}_2$, where

$$\hat{\boldsymbol{\beta}}_{1,x} = \left(\sum_{j \in A_1} \frac{\mathbf{x}_{1j} \mathbf{x}_{1j}'}{v_j} \right)^{-1} \sum_{j \in A_1} \frac{\mathbf{x}_{1j} \mathbf{x}_{2j}'}{v_j}.$$

After some algebra, $\hat{Y}_{tp,p}$ in (4.1) can be expressed as

$$\begin{aligned} \hat{Y}_{tp,p} &= \sum_{j \in U} \mathbf{x}_{1j}' \hat{\mathbf{B}}_1 + \sum_{j \in A_1} (\mathbf{x}_{2j}' - \mathbf{x}_{1j}' \hat{\boldsymbol{\beta}}_{1,x}) \hat{\boldsymbol{\beta}}_2 + \sum_{j \in A_2} (y_j - \mathbf{x}_j' \hat{\boldsymbol{\beta}}) \\ &= \sum_{j \in U} \hat{y}_{1j} + \sum_{j \in A_1} (\hat{y}_{2j} - \hat{y}_{1j}) + \sum_{j \in A_2} (y_j - \hat{y}_{2j}). \end{aligned} \tag{4.3}$$

However, the prediction estimator (4.3) does not necessarily satisfy the ADU property.

To satisfy ADU, we consider more general predicted values for y using the instrumental vector variable $\mathbf{z}_i = \mathbf{z}(\mathbf{x}_i, d_{1i}, d_{2i}, v_i)$, with $\mathbf{z}_i = (\mathbf{z}_{1i}', \mathbf{z}_{2i}')'$ where $\mathbf{z}_{1i}$ is available for the population and $\mathbf{z}_{2i}$ is available up to A_1. Let $\hat{y}_{1i,z} = \mathbf{x}_{1i}' \hat{\mathbf{B}}_{1,z}$, $\hat{y}_{2i,z} = \mathbf{x}_i' \hat{\boldsymbol{\beta}}_z$, where $\hat{\mathbf{B}}_{1,z} = \left(\hat{\boldsymbol{\beta}}_{1,z} + \hat{\boldsymbol{\beta}}_{x,z} \hat{\boldsymbol{\beta}}_{2,z} \right)$, $\hat{\boldsymbol{\beta}}_{x,z} = \left(\sum_{j \in A_1} \mathbf{z}_{1j} \mathbf{x}_{1j}' \right)^{-1} \sum_{j \in A_1} \mathbf{z}_{1j} \mathbf{x}_{2j}'$, and $\hat{\boldsymbol{\beta}}_z = (\hat{\boldsymbol{\beta}}_{1,z}', \hat{\boldsymbol{\beta}}_{2,z}')' = \left(\sum_{j \in A_2} \mathbf{z}_j \mathbf{x}_j' \right)^{-1} \sum_{j \in A_2} \mathbf{z}_j y_j$. Then a prediction estimator under two-phase sampling using the instrumental variable $\mathbf{z}_i$ has the form of

$$\begin{aligned} \hat{Y}_{tp,z} &= \sum_{j \in A_2} y_j + \sum_{j \in A_1 / A_2} \hat{y}_{2j,z} + \sum_{j \in U / A_1} \hat{y}_{1j,z} \\ &= \sum_{j \in U} \hat{y}_{1j,z} + \sum_{j \in A_1} (\hat{y}_{2j,z} - \hat{y}_{1j,z}) + \sum_{j \in A_2} (y_j - \hat{y}_{2j,z}). \end{aligned} \tag{4.4}$$

If $(d_{1i} - 1)$ is included in $\mathbf{z}_{1i}$ and $(d_{2i} - 1)$ is included in $\mathbf{z}_i$, then $\hat{Y}_{tp,z}$ can be expressed as

$$\hat{Y}_{tp,r} = \sum_{j \in U} \hat{y}_{1j,z} + \sum_{j \in A_1} d_{1j} (\hat{y}_{2j,z} - \hat{y}_{1j,z}) + \sum_{j \in A_2} d_{2j} (y_j - \hat{y}_{2j,z}). \tag{4.5}$$

Expression (4.5) suggests that $\hat{Y}_{tp,z}$ satisfies the ADU property. Thus, we assume that $(d_{1i} - 1)$ and $(d_{2i} - 1)$ are included in the column space of $\mathbf{z}_{1i}$ and $\mathbf{z}_i$, respectively.

Note that we can express

$$\hat{Y}_{tp,z} = \sum_{i \in A_2} w_{1i} y_i + \sum_{i \in A_1/A_2} w_{1i} \hat{y}_{2i,z} = \sum_{i \in A_2} w_{2i} y_i,$$

where

$$w_{1i} = 1 + \sum_{j \in A_1^c} \mathbf{x}_{1j}' \left(\sum_{j \in A_1} \mathbf{z}_{1j} \mathbf{x}_{1j}' \right)^{-1} \mathbf{z}_{1i}, \tag{4.6}$$

$$\begin{aligned} w_{2i} &= w_{1i} + \sum_{i \in A_1/A_2} w_{1i} \mathbf{x}_i' \left(\sum_{j \in A_2} \mathbf{z}_j \mathbf{x}_j' \right)^{-1} \mathbf{z}_i \\ &= 1 + \left(\sum_{i \in A_1} w_{1i} \mathbf{x}_i - \sum_{i \in A_2} \mathbf{x}_i \right)' \left(\sum_{j \in A_2} \mathbf{z}_j \mathbf{x}_j' \right)^{-1} \mathbf{z}_i. \end{aligned} \tag{4.7}$$

The weights w_{1i} and w_{2i} satisfy

$$\sum_{i \in A_1} w_{1i} \mathbf{x}_{1i} = \sum_{i \in U} \mathbf{x}_{1i} \tag{4.8}$$

$$\sum_{i \in A_2} w_{2i} \mathbf{x}_i = \sum_{i \in A_1} w_{1i} \mathbf{x}_i. \tag{4.9}$$

Thus, both w_{1i} and w_{2i} are well calibrated for the population total of $\mathbf{x}_{1i}$ and also provide consistency for $\mathbf{x}_{2i}$. Such calibration can be called two-step calibration, as discussed by Dupont (1995). In step one, w_{1i} are constructed to satisfy (4.8). In step two, w_{2i} are constructed to satisfy (4.9) using the calibration weights w_{1i} computed from step one. The resulting calibration estimator is efficient if the linear model (4.2) holds. Modified exponential-tilting calibration weights can be constructed similarly.

The two-step calibration method requires that we observe the individual information of $\mathbf{x}_i$ in A_1 when computing w_{1i} from (4.8). Instead of using (4.8) in step one and (4.9) in step two, the following two-step calibration method can also be considered.

Step 1: For $i \in A_2$, compute the initial calibration weights $w_{2i}^{(1)}$ from A_2,

$$w_{2i}^{(1)} = 1 + (d_{2i} - 1) \exp\left(\hat{\lambda}_0^{(1)} + \frac{\mathbf{z}_{1i}'}{d_{2i} - 1} \hat{\boldsymbol{\lambda}}_1^{(1)} + \frac{\mathbf{z}_{2i}'}{d_{2i} - 1} \hat{\boldsymbol{\lambda}}_2^{(1)} \right),$$

where $(\hat{\lambda}_0^{(1)}, \hat{\boldsymbol{\lambda}}_1^{(1)}, \hat{\boldsymbol{\lambda}}_2^{(1)})$ satisfies

$$\sum_{i \in A_2} w_{2i}^{(1)} \mathbf{x}_i = \sum_{i \in A_1} d_{1i} \mathbf{x}_i. \tag{4.10}$$

Step 2: Use $w_{2i}^{(1)}$ in [Step 1] to compute

$$w_{2i}^{(2)} = 1 + \left(w_{2i}^{(1)} - 1\right) \exp\left\{\hat{\lambda}_0^{(2)} + \left(\frac{d_{2i}}{d_{1i}}\right) \frac{\mathbf{z}_{1i}'}{w_{2i}^{(1)} - 1} \hat{\boldsymbol{\lambda}}_1^{(2)}\right\},$$

where $(\hat{\lambda}_0^{(2)}, \hat{\boldsymbol{\lambda}}_1^{(2)})$ satisfies

$$\sum_{i \in A_2} w_{2i}^{(2)} \mathbf{x}_{1i} = \sum_{i \in U} \mathbf{x}_{1i}. \tag{4.11}$$

Such a two-step calibration does not need to compute the calibration weights w_{1i} for the first-phase sample, which is quite convenient in practice. In Section S4 of Supplementary Note, we briefly show that the proposed two-step calibration method for two-phase sampling is asymptotically equivalent to the classical two-step calibration method in (4.8)−(4.9).

5. Simulation Study

To compare the estimators, we performed a limited simulation study. We considered two study variables, y_1 and y_2, and generated a population of size $N = 10,000$ with

$$\begin{aligned} x_i &\sim N(4,1), \\ e_i &\sim N(0, 0.25x_i^2), \\ z_i &\sim 0.5x_i + \chi^2(0.5) + 5, \\ y_{1i} &= 1 + x_i + e_i, \\ y_{2i} &= (x_i - 1)^2 + e_i, \end{aligned}$$

with x_i and e_i independent. From the population, we generated $B = 2,000$ Monte Carlo samples, of sizes $n = 500$ and $n = 1,000$, under simple random sampling and probability proportional to size z_i sampling with replacement. The parameters of interest are population means for y_1 and y_2. We considered five estimators: direct estimator without calibration, denoted by *HT*; generalized regression estimator of Deville and Särndal (1992), denoted by *GREG*; best linear prediction estimator, denoted by *Prediction*; bias-corrected prediction estimator, denoted by *B − C prediction*; proposed calibration estimator, denoted by *New*. Under (1.1) with $v_i = x_i^2$, the five estimators are

$$\begin{aligned} HT \quad &: \sum_{i \in A} d_i y_i, \\ GREG \quad &: \sum_{i \in A} d_i y_i + \Big(\sum_{i=1}^{N} \mathbf{x}_i' - \sum_{i \in A} d_i \mathbf{x}_i'\Big) \Big(\sum_{i \in A} d_i \frac{\mathbf{x}_i \mathbf{x}_i'}{v_i}\Big)^{-1} \sum_{i \in A} d_i \frac{\mathbf{x}_i y_i}{v_i}, \\ Prediction \quad &: \sum_{i \in A} y_i + \sum_{i \in A^c} \mathbf{x}_i' \Big(\sum_{i \in A} \frac{\mathbf{x}_i \mathbf{x}_i'}{v_i}\Big)^{-1} \sum_{i \in A} \frac{\mathbf{x}_i y_i}{v_i}, \end{aligned}$$

$$B-C\ prediction: \sum_{i\in A} y_i + \sum_{i\in A^c} \mathbf{x}_i' \left(\sum_{i\in A} \mathbf{z}_i \mathbf{x}_i' \right)^{-1} \sum_{i\in A} \mathbf{z}_i y_i,$$

$$New \qquad : \sum_{i\in A} \left\{ 1 + (d_i - 1)\exp(\hat{\lambda}_0 + \hat{\boldsymbol{\lambda}}_1' z_{1i}^*) \right\} y_i.$$

For deriving the bias-corrected estimators, we used $\mathbf{z}_i = (d_i - 1, x_i/v_i)$, and for deriving the proposed calibration estimators $z_{1i}^* = x_i/\{v_i(d_i - 1)\}$.

The bias and the mean squared errors of the five estimators for the two population means are presented in Table 2. All estimators have smaller mean squared errors than the HT estimator, as expected. The prediction estimator is the most efficient in terms of mean squared error for y_1 because the working linear regression model is true for variable y_1. The mean squared errors of the GREG and B-C prediction estimators are slightly higher than that of the prediction estimator, but the bias of the GREG estimator and the B-C prediction estimator are smaller for y_2. The calibration estimator is similar to the B-C prediction estimator in terms of the two criteria for estimating y_1. For y_2, since the linear regression model is not a good fit, the prediction estimator demonstrates significant bias. The proposed calibration estimator is more robust than the other three estimators in term of both bias and mean squared error.

We considered variance estimation only for bias-corrected prediction and the proposed calibration estimator. The linearization variance estimator in (2.6) was used to compute the variance of each estimator in the simulation. The Monte Carlo relative biases of the linearization variance estimators are all less than 5% in absolute values; they are not presented here.

6. Data Example

To compare the proposed estimator with other estimators we used a small sample of segments from the 1997 National Resource Inventory for the state of Missouri that is presented in Table 2.8 of Fuller (2009). In this sample with all 80 segments chosen from three strata, 79 segments have three sample points and one segment has only two sample points. The sampling design features stratified random sampling for segments and second-stage sampling selected with two or three points within segments. The variable "Weight" in the table is the inverse of the sampling rate and "Segment Size" is the total area of the segment in acres. The variable "Cultivated Cropland" is the fraction of points having cropland in active use multiplied by the segment size. Other variables, "Forest" and "Federal", are defined in the same way.

As for auxiliary variables, the variables "Segment Size" and "Federal" can be considered. We included indicators for three strata in the regression so that the

Table 2. Monte Carlo Biases and Mean squared errors of the five point estimators.

Parameter	Sample Size	Estimator	SRS		PPS	
			Bias	MSE	Bias	MSE
$E(Y_1)$	500	*HT*	-0.0026	0.0096	-0.0040	0.0098
		GREG	-0.0021	0.0082	-0.0023	0.0085
		Prediction	-0.0002	0.0081	0.0065	0.0082
		B − C prediction	-0.0021	0.0082	-0.0022	0.0085
		New	-0.0021	0.0082	-0.0022	0.0085
	1000	*HT*	0.0004	0.0043	-0.0034	0.0047
		GREG	-0.0015	0.0036	-0.0022	0.0042
		Prediction	0.0002	0.0035	0.0061	0.0041
		B − C prediction	-0.0014	0.0036	-0.0021	0.0042
		New	-0.0014	0.0036	-0.0021	0.0042
$E(Y_2)$	500	*HT*	-0.0013	0.0790	-0.0097	0.0694
		GREG	-0.0036	0.0147	-0.0089	0.0173
		Prediction	-0.2111	0.0611	-0.1438	0.0474
		B − C prediction	-0.0028	0.0120	-0.0066	0.0141
		New	0.0006	0.0121	-0.0012	0.0139
	1000	*HT*	0.0092	0.0360	-0.0091	0.0338
		GREG	-0.0015	0.0069	-0.0083	0.0085
		Prediction	-0.1999	0.0477	-0.1513	0.0458
		B − C prediction	-0.0031	0.0057	-0.0067	0.0066
		New	-0.0012	0.0057	-0.0029	0.0065

Table 3. Alternative Estimators and Standard Errors.

Model Variance	Estimator	Cultivated	Forest	Other
$\gamma = 0$	*GREG*	156.88	74.74	178.28
		(16.73)	(13.72)	(17.19)
	B − C prediction	156.89	74.77	178.25
		(16.73)	(13.74)	(17.17)
	Calibration	156.58	74.72	178.6
		(16.58)	(13.74)	(17.12)
$\hat{\gamma} = 1.31$	*GREG*	156.24	75.05	178.61
		(16.50)	(13.77)	(17.15)
	B − C prediction	156.41	74.82	178.67
		(16.53)	(13.74)	(17.11)
	Calibration	156.19	74.98	178.72
		(16.46)	(13.76)	(17.10)

sampling design would be noninformative. To estimate total acres of cultivated cropland, the main variable of interest, we considered the model

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + e_i,$$

where y_i is the acres of cultivated cropland, (x_{1i}, x_{2i}, x_{3i}) is the vector of stratum indicators for segment i, x_{4i} is the total area of segment i, x_{5i} is federal acres,

and $e_i \sim N(0, \sigma_e^2 x_{4i}^{\gamma})$.

To estimate γ, we considered the regression model:

$$\log(\hat{e}_i^2) = \log(\sigma_e^2) + \gamma \log(x_{4i}) + u_i,$$

where $u_i \sim N(0, \sigma_u^2)$. The procedure of estimating β and γ can be iterated with initial values calculated by the ordinary least squares method. Similar estimation procedures are also discussed in Valliant, Dorfman, and Royall (2000).

For deriving the total average of cultivated cropland, we considered three estimators using model variance that depends on γ : generalized regression estimator, *GREG*, bias-corrected prediction estimator, *B − C prediction*, and the proposed calibration estimator, *Calibrtaion*. In addition to cultivated cropland, we also estimated the total acres of forest and nonfederal land in other categories.

Table 3 presents the resulting estimators and their standard errors for $\gamma = 0$ and $\hat{\gamma} = 1.31$. From Table 3, we conclude that it is a more reasonable assumption that the model variance depends on x_{4i}, since the estimators computed when we assume that the model variance depends on x_{4i} have smaller standard errors for total acres of cultivated cropland than the estimators calculated under constant variance assumption. The generalized regression and bias-corrected estimators show similar performance in terms of standard errors under both model variance assumptions. The proposed calibration estimator for the total area of cultivated cropland, when model variance is proportional to x_{4i}^{γ}, is the most efficient among the estimators considered. For estimating total area of forest and other nonfederal land categories, the three estimators do not have significant differences in biases and mean squared errors.

7. Concluding Remarks

Calibration constraint is important in survey estimation. Using a prediction model, instrumental variables can be constructed to achieve optimality in the sense of minimizing the anticipated variance among the class of asymptotic design unbiased estimators satisfying the calibration constraint. The proposed instrumental-variable calibration estimator can be modified to achieve range restrictions on the final weights. The proposed method can be directly applied to two-phase sampling. An alternative two-step calibration method is discussed.

Optimality of the proposed estimator is based on the linear regression model with known variance function. Further investigation of the departure from the model assumptions is a topic of future research.

Acknowledgement

We thank an anonymous referee and an associate editor for their constructive comments. The research of the second author was partially supported by

a Cooperative Agreement between the US Department of Agriculture Natural Resources Conservation Service and Iowa State University.

References

Brewer, K. R. W. (1963). Ratio estimation in finite populations: some results deductible from the assumption of an underlying stochastic process. *Austral. J. Statist.* **5**, 962-973.

Brewer, K. R. W. (1979). A class of robust sampling designs for large-scale surveys. *J. Amer. Statist. Assoc.* **74**, 911-915.

Brewer, K. R. W., Muhammad, H. and Tam, S. M. (1988). How nearly can model-based prediction and design-based estimation be reconciled? *J. Amer. Statist. Assoc.* **83**, 128-132.

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87**, 376-382.

Dupont, F. (1995). Alternative adjustment where there are several levels of auxiliary information. *Surv. Methodol.* **21**, 125-135.

Estavo, V. M., and Särndal, C. E. (2000). A functional form approach to calibration. *J. Official Statist.* **16**, 379-399.

Firth, D. and Bennett, K. E. (1998). Robust models in probability sampling. *J. Roy. Statist. Soc. Ser. B* **60**, 3-21.

Fuller, W. A. (2002). Regression estimation for survey samples. *Surv. Methodol.* **28**, 5-23.

Fuller, W. A. (2003). Estimation for multiple phase samples. In *Analysis of Survey Data,* (Edited by R. L. Chambers and C. J. Skinner), 307-322. Wiley, Chichester.

Fuller, W. A. (2009). *Sampling Statistics.* Wiley, Hoboken, NJ.

Hidiroglou, M. A. (2001). Double sampling. *Surv. Methodol.* **27**, 143-154.

Huang, E. T. and Fuller, W. A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section,* American Statistical Association, 300-305.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.* **77**, 89-96.

Kim, J. K. (2010). Calibration estimation using exponential tilting in sample surveys. *Surv. Methodol.* **36**, 145-155.

Kim, J. K., Navarro, A., and Fuller, W. A. (2006). Replication variance estimation for two-phase stratified sampling. *J. Amer. Statist. Assoc.* **101**, 312-320.

Kim, J. K. and Park, M. (2010). Calibration estimation in survey sampling. *Internat. Statist. Rev.* **78**, 21-39.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Surv. Methodol.* **32**, 133-142.

Kott, P. S. (2011). A nearly pseudo-optimal mehtod for keeping calibration weights from falling below unity in the absence of nonresponse or frame errors. *Pakistan J. Statist.* **27**, 391-396.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Internat. Statist. Rev.* **61**, 317-337.

Rao, J. N. K. (1973). On double sampling for stratification and analytic survey. *Biometrika* **60**, 125-133.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika* **57**, 377-387.

Royall, R. M. (1976). The linear least squares prediction approach to two-stage sampling. *J. Amer. Statist. Assoc.* **71**, 657-664.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach.* Wiley, New York.

Wright, R. L. (1983). Finite population sampling with multivariate auxiliary information. *J. Amer. Statist. Assoc.* **78**, 879-884.

Department of Statistics, Seoul National University, Seoul, 151-747, Korea.

E-mail: kkampsh@gmail.com

Department of Statistics, Iowa State University, Ames, Iowa, 50011, U.S.A.

E-mail: jkim@iastate.edu

(Received February 2013; accepted June 2013)