Communications for Statistical Applications and Methods
2015, Vol. 22, No. 2, 137–146

DOI: http://dx.doi.org/10.5351/CSAM.2015.22.2.137
Print ISSN 2287-7843 / Online ISSN 2383-4757

A Dual Problem of Calibration of Design Weights Based on Multi-Auxiliary Variables

J. Al-Jurarha[1,a]

[a]Department of Statistics, Yarmouk University, Jordan

Abstract

Singh (2013) considered the dual problem to the calibration of design weights to obtain a new generalized linear regression estimator (GREG) for the finite population total. In this work, we have made an attempt to suggest a way to use the dual calibration of the design weights in case of multi-auxiliary variables; in other words, we have made an attempt to give an answer to the concern in Remark 2 of Singh (2013) work. The same idea is also used to generalize the GREG estimator proposed by Deville and Särndal (1992). It is not an easy task to find the optimum values of the parameters appear in our approach; therefore, few suggestions are mentioned to select values for such parameters based on a random sample. Based on real data set and under simple random sampling without replacement design, our approach is compared with other approaches mentioned in this paper and for different sample sizes. Simulation results show that all estimators have negligible relative bias, and the multivariate case of Singh (2013) estimator is more efficient than other estimators.

Keywords: Dual problem, GREG, multi-auxiliary variables, mean squared error, bias.

1. Introduction

In survey sampling we are usually interested in estimating the finite population total, ratio, variance, $\ldots, etc$. The availability of auxiliary variables are usually used to increase the precision of such estimators. Consider the finite population U of N units indexed by the set $\{1, 2, \ldots, N\}$. For the i^{th} unit, let y_i be the value of the interest variable Y, and $x_{1i}, x_{2i}, \ldots, x_{pi}$ be the values of the auxiliary variables $X_1, X_2, \ldots, X_p$, respectively. Based on the probability sampling design $p(\cdot)$, draw a random sample s from U. The first order inclusion probability π_i is defined by $\pi_i = \sum_{s \ni i} p(s)$, and the second inclusion probability π_{ij} is defined by $\pi_{ij} = \sum_{s \ni i,\, j} p(s)$, for $i \neq j$, and $\pi_{ij} = \pi_i$ when $i = j$. The probability sampling design $p(\cdot)$ is assumed to be a measurable design. Define $t_{x_1} = \sum_{i \in U} x_{1_i}$, $t_{x_2} = \sum_{i \in U} x_{2_i}, \ldots, t_{x_p} = \sum_{i \in U} x_{p_i}$ be the population totals for the auxiliary variables $X_1, X_2, \ldots, X_p$, respectively.

Horvitz and Thompson (1952) proposed the following estimator

$$\hat{t}_{y\pi} = \sum_{i \in U} \frac{y_i}{\pi_i} I_{\{i \in s\}} = \sum_{i \in U} d_i y_i I_{\{i \in s\}} = \sum_{i \in s} d_i y_i \tag{1.1}$$

to estimate the population total t_y, where $I_{\{i \in s\}}$ is one if $i \in s$ and zero otherwise, and $d_i = 1/\pi_i$ are the sampling design weights. However, $\hat{t}_{y\pi}$ does not use the availability of the auxiliary variables.

[1] Department of Statistics, Yarmouk University, Irbid, Jordan. E-mail: jjararha@gmail.com

Published 31 March 2015 / journal homepage: http://csam.or.kr

The availability of the auxiliary variables and the calibration on the auxiliary variables can be used to increase the precision of our estimators. Stearns and Singh (2008) summarized the developments by several researchers on the GREG estimators and used the calibration idea to propose three new estimators of the variance of the GREG estimators.

The methods of estimation the population ratio $\theta_{xy} = t_y/t_x$ are heavily discussed in the literature. To get more advantages from the availability of the auxiliary variables, Al-Jararha and Bataineh (2015) extended the idea of estimating θ_{xy} to use the availability of such auxiliary variables. Their approach, is more efficient than the regular estimators proposed in the literature.

Deville and Särndal (1992) proposed the following estimator

$$\hat{t}_{y.ds} = \sum_{i \in U} w_i y_i I_{\{i \in s\}} = \sum_{i \in s} w_i y_i, \tag{1.2}$$

for estimating t_y, where w_i $i \in s$ are the new sampling design weights that calibrated the sampling design weights d_i defined by Equation (1.1) based on the calibration on the known population total for the auxiliary variable X and the chi-square distance. The calibrated weights w_i are obtained by minimizing the chi-square distance

$$D = \frac{1}{2} \sum_{i \in s} \frac{(w_i - d_i)^2}{d_i q_i} \tag{1.3}$$

subject to the constraint condition

$$t_x = \sum_{i \in s} w_i x_i. \tag{1.4}$$

As a result of this, the calibrated weights w_i are given by

$$w_i = d_i + \frac{t_x - \hat{t}_{x\pi}}{\sum_{i \in s} d_i q_i x_i^2} d_i q_i x_i, \tag{1.5}$$

Therefore, Equation (1.2) reduced to

$$\hat{t}_{y.ds} = \hat{t}_{y\pi} + \hat{\beta}_{ds} \left(t_x - \hat{t}_{x\pi} \right) \tag{1.6}$$

which is a GREG type estimator. Where $\hat{\beta}_{ds} = \sum_{i \in s} d_i q_i x_i y_i / \sum_{i \in s} d_i q_i x_i^2$, $\hat{t}_{x\pi}$ is the Horvitz and Thompson (1952) estimator of t_x. In most cases $q_i = 1$ and other weights cases may be considered.

Singh (2013) adopted the dual approach of calibration to estimate the population total t_y. The proposed estimator for t_y is given by

$$\hat{t}_{y.sin} = \sum_{i \in s} w_i^* y_i \tag{1.7}$$

The calibrated sampling design weights w_i^* are obtained by minimizing

$$\hat{t}_{x.sin} = \sum_{i \in s} w_i^* x_i \tag{1.8}$$

subject to

$$\sum_{i \in s} w_i^* = \sum_{i \in s} d_i \tag{1.9}$$

and

$$\alpha = \frac{1}{2}\sum_{i\in s}\frac{\left(w_i^* - d_i\right)^2}{d_i q_i}. \tag{1.10}$$

The calibrated weights w_i^* are given by

$$w_i^* = d_i + \frac{d_i q_i \left(x_i - \frac{\sum_{i\in s} d_i q_i x_i}{\sum_{i\in s} d_i q_i}\right)}{\sqrt{\sum_{i\in s} d_i q_i \left(x_i - \frac{\sum_{i\in s} d_i q_i x_i}{\sum_{i\in s} d_i q_i}\right)^2}}\sqrt{\alpha}. \tag{1.11}$$

Define $\sqrt{\alpha}$ by

$$\sqrt{\alpha} = \frac{t_x - \hat{t}_{x\pi}}{\sqrt{\sum_{i\in s} d_i q_i \left(x_i - \frac{\sum_{i\in s} d_i q_i x_i}{\sum_{i\in s} d_i q_i}\right)^2}}. \tag{1.12}$$

Based on the central limit theorem,

$$\sqrt{\alpha} \sim N(0,\ 1).$$

From Equation (1.7), the estimator $\hat{t}_{y.sin}$ reduced to

$$\hat{t}_{y.sin} = \hat{t}_{y\pi} + \hat{\beta}_{sin}\left(t_x - \hat{t}_{x\pi}\right) \tag{1.13}$$

which is a GREG type estimator. Where

$$\hat{\beta}_{sin} = \frac{\sum_{i\in s} d_i q_i \left(y_i - \frac{\sum_{i\in s} d_i q_i y_i}{\sum_{i\in s} d_i q_i}\right)\left(x_i - \frac{\sum_{i\in s} d_i q_i x_i}{\sum_{i\in s} d_i q_i}\right)}{\sum_{i\in s} d_i q_i \left(x_i - \frac{\sum_{i\in s} d_i q_i x_i}{\sum_{i\in s} d_i q_i}\right)^2}. \tag{1.14}$$

In Remark 2, Singh (2013) wrote “It is not apparently clear: how the use of the dual to the calibration of design weights can be developed for the case of multi-auxiliary information?” In this paper, we will answer this question. Furthermore, the estimator proposed by Deville and Särndal (1992), defined by Equation (1.6), will be generalized to use multi-auxiliary variables.

2. Calibration and Multi-Auxiliary Variables

In this section, we will generalize the estimators proposed by Deville and Särndal (1992) and Singh (2013) to use multi-auxiliary variables. In other words, to generalize the estimators defined by Equations (1.6) and (1.13) to the case of the availability of p auxiliary variables $X_1, X_2, \ldots, X_p$.

Let t_{x_i} be the population total for the i^{th} auxiliary variable, $i = 1, \ldots, p$. Assume that $t_{x_1}, t_{x_2}, \ldots, t_{x_p}$ are known. Based on a probability sampling design $p(\cdot)$, draw a random sample s from U. Define

$$\mathbf{t_x} = \tau_1 t_{x_1} + \tau_2 t_{x_2} + \cdots + \tau_p t_{x_p}, \tag{2.1}$$

$$\hat{\mathbf{t}}_{\mathbf{x}\pi} = \tau_1 \hat{t}_{x_{1\pi}} + \tau_2 \hat{t}_{x_{2\pi}} + \cdots + \tau_p \hat{t}_{x_{p\pi}}, \tag{2.2}$$

and

$$\check{\mathbf{t}}_{\mathbf{x}} = \tau_1 \check{t}_{x_1} + \tau_2 \check{t}_{x_2} + \cdots + \tau_p \check{t}_{x_p}, \tag{2.3}$$

where $\hat{t}_{x_{i\pi}} = \sum_{j\in s} d_j x_{ij}$, $\check{t}_{x_i} = \sum_{j\in s} w_j x_{ij}$, $0 \le \tau_i \le 1$ for $i = 1, \ldots, p$, and $\sum_{i=1}^{p} \tau_i = 1$.

To generalize the estimator proposed by Deville and Särndal (1992), minimize

$$D = \frac{1}{2}\sum_{i\in s}\frac{(w_i - d_i)^2}{d_i q_i} \tag{2.4}$$

subject to the calibration equation

$$\mathbf{t_x} = \check{\mathbf{t}}_\mathbf{x}, \tag{2.5}$$

where $\mathbf{t_x}$ and $\check{\mathbf{t}}_\mathbf{x}$ are defined by Equations (2.1) and (2.3), respectively. The Lagrange function L is defined by

$$L = \frac{1}{2}\sum_{i\in s}\frac{(w_i - d_i)^2}{d_i q_i} + \lambda\left(\mathbf{t_x} - \check{\mathbf{t}}_\mathbf{x}\right). \tag{2.6}$$

Differentiate L with respect to w_i and equate to zero. As a result of this

$$w_i - d_i = \lambda d_i q_i \left(\tau_1 x_{1i} + \tau_2 x_{2i} + \cdots + \tau_p x_{pi}\right). \tag{2.7}$$

Multiply both sides of Equation (2.7) by $\tau_1 x_{1i} + \tau_2 x_{2i} + \cdots + \tau_p x_{pi}$, sum over $i \in s$, and use the calibration Equation (2.5), we have

$$\begin{aligned}\lambda &= \frac{\tau_1\left(t_{x_1} - \hat{t}_{x_{1\pi}}\right) + \cdots + \tau_p\left(t_{x_p} - \hat{t}_{x_{p\pi}}\right)}{\sum_{i\in s} d_i q_i\left(\tau_1 x_{1i} + \tau_2 x_{2i} + \cdots + \tau_p x_{pi}\right)^2}\\ &= \frac{\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi}}{\sum_{i\in s} d_i q_i\left(\tau_1 x_{1i} + \tau_2 x_{2i} + \cdots + \tau_p x_{pi}\right)^2}.\end{aligned} \tag{2.8}$$

From Equation (2.7), the calibrated weights w_i are

$$w_i = d_i + \frac{d_i q_i\left(\tau_1 x_{1i} + \tau_2 x_{2i} + \cdots + \tau_p x_{pi}\right)}{\sum_{i\in s} d_i q_i\left(\tau_1 x_{1i} + \tau_2 x_{2i} + \cdots + \tau_p x_{pi}\right)^2}\left(\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi}\right). \tag{2.9}$$

Multiply both sides of Equation (2.9) by y_i and sum over $i \in s$, we have

$$\begin{aligned}\hat{t}_{y.gds} &= \hat{t}_{y\pi} + \frac{\sum_{i\in s} d_i q_i\left(\tau_1 x_{1i} + \tau_2 x_{2i} + \cdots + \tau_p x_{pi}\right) y_i}{\sum_{i\in s} d_i q_i\left(\tau_1 x_{1i} + \tau_2 x_{2i} + \cdots + \tau_p x_{pi}\right)^2}\left(\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi}\right)\\ &= \hat{t}_{y\pi} + \hat{\beta}_{gds}\left(\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi}\right),\end{aligned} \tag{2.10}$$

where

$$\begin{aligned}\hat{\beta}_{gds} &= \frac{\sum_{i\in s} d_i q_i\left(\tau_1 x_{1i} + \tau_2 x_{2i} + \cdots + \tau_p x_{pi}\right) y_i}{\sum_{i\in s} d_i q_i\left(\tau_1 x_{1i} + \tau_2 x_{2i} + \cdots + \tau_p x_{pi}\right)^2}\\ &= \frac{\tau_1 \sum_{i\in s} d_i q_i x_{1i} y_i + \tau_2 \sum_{i\in s} d_i q_i x_{2i} y_i + \cdots + \tau_p \sum_{i\in s} d_i q_i x_{pi} y_i}{\sum_{i\in s} d_i q_i\left(\tau_1 x_{1i} + \tau_2 x_{2i} + \cdots + \tau_p x_{pi}\right)^2}.\end{aligned} \tag{2.11}$$

The estimator given by Equation (2.10) is a GREG estimator and generalizes the estimator proposed by Deville and Särndal (1992).

To generalize the estimator proposed by Singh (2013), minimize

$$\hat{t}_{x.sin} = \sum_{i \in s} w_i^* \left[\tau_1 x_{1i} + \tau_2 x_{2i} + \cdots + \tau_p x_{pi}\right] \tag{2.12}$$

subject to

$$\sum_{i \in s} w_i^* = \sum_{i \in s} d_i \tag{2.13}$$

and

$$\alpha = \frac{1}{2} \sum_{i \in s} \frac{\left(w_i^* - d_i\right)^2}{d_i q_i}. \tag{2.14}$$

The Lagrange function L is defined by

$$L = \hat{t}_{x.sin} - \lambda_1 \left(\sum_{i \in s} w_i^* - \sum_{i \in s} d_i \right) - \lambda_2 \left(\frac{1}{2} \sum_{i \in s} \frac{\left(w_i^* - d_i\right)^2}{d_i q_i} - \alpha \right). \tag{2.15}$$

Differentiate L with respect to w_i^*, and equate to zero, we have

$$\tau_1 x_{1i} + \tau_2 x_{2i} + \cdots + \tau_p x_{pi} - \lambda_1 - \lambda_2 \frac{w_i^* - d_i}{d_i q_i} = 0. \tag{2.16}$$

Therefore,

$$w_i^* = d_i + \frac{1}{\lambda_2} \left[\left(\tau_1 x_{1i} + \tau_2 x_{2i} + \cdots + \tau_p x_{pi}\right) d_i q_i - d_i q_i \lambda_1\right]. \tag{2.17}$$

Multiply Equation (2.16) by $d_i q_i$, sum over $i \in s$, and solve for λ_1, we have

$$\lambda_1 = \frac{1}{\sum_{i \in s} d_i q_i} \left[\tau_1 \sum_{i \in s} d_i q_i x_{1i} + \tau_2 \sum_{i \in s} d_i q_i x_{2i} + \cdots + \tau_p \sum_{i \in s} d_i q_i x_{pi}\right]. \tag{2.18}$$

Insert the w_i^* into Equation (2.14), and solve for λ_2, we have

$$\lambda_2^2 = \frac{1}{2\alpha} \sum_{i \in s} \frac{1}{d_i q_i} \left[\left(\tau_1 x_{1i} + \tau_2 x_{2i} + \cdots + \tau_p x_{pi}\right) d_i q_i - d_i q_i \lambda_1\right]^2. \tag{2.19}$$

Therefore,

$$\lambda_2 = \pm \frac{1}{\sqrt{2\alpha}} \sqrt{\sum_{i \in s} d_i q_i \left[\left(\tau_1 x_{1i} + \tau_2 x_{2i} + \cdots + \tau_p x_{pi}\right) - \lambda_1\right]^2}. \tag{2.20}$$

Using the same choice of Singh (2013), and from Equations (2.17), (2.18), and (2.20) the calibrated weights w_i^* are

$$w_i^* = d_i + \frac{\psi_{1i}}{\sqrt{\psi_2}}\sqrt{2\alpha}, \tag{2.21}$$

where

$$\psi_{1i} = \tau_1\left(x_{1i} - \frac{\sum_{i\in s} d_i q_i x_{1i}}{\sum_{i\in s} d_i q_i}\right) d_i q_i + \cdots + \tau_p\left(x_{pi} - \frac{\sum_{i\in s} d_i q_i x_{pi}}{\sum_{i\in s} d_i q_i}\right) d_i q_i \tag{2.22}$$

and

$$\psi_2 = \sum_{i\in s} \frac{1}{d_i q_i}\psi_{1i}^2. \tag{2.23}$$

Multiply both sides of Equation (2.21) by y_i, sum over $i \in s$, we have

$$\hat{t}_{y.gsin} = \hat{t}_{y\pi} + \frac{\psi_1}{\sqrt{\psi_2}}\sqrt{2\alpha}, \tag{2.24}$$

where

$$\begin{aligned}
\psi_1 &= \sum_{i\in s} \psi_{1i} y_i \\
&= \sum_{i\in s}\left[\tau_1\left(x_{1i} - \frac{\sum_{i\in s} d_i q_i x_{1i}}{\sum_{i\in s} d_i q_i}\right) d_i q_i + \cdots + \tau_p\left(x_{pi} - \frac{\sum_{i\in s} d_i q_i x_{pi}}{\sum_{i\in s} d_i q_i}\right) d_i q_i\right] y_i \\
&= \sum_{i\in s}\left[\tau_1\left(x_{1i} - \frac{\sum_{i\in s} d_i q_i x_{1i}}{\sum_{i\in s} d_i q_i}\right)\left(y_i - \frac{\sum_{i\in s} d_i q_i y_i}{\sum_{i\in s} d_i q_i}\right) d_i q_i \right.\\
&\quad \left. + \cdots + \tau_p\left(x_{pi} - \frac{\sum_{i\in s} d_i q_i x_{pi}}{\sum_{i\in s} d_i q_i}\right)\left(y_i - \frac{\sum_{i\in s} d_i q_i y_i}{\sum_{i\in s} d_i q_i}\right) d_i q_i\right] \\
&= \tau_1 \sum_{i\in s}\left(x_{1i} - \frac{\sum_{i\in s} d_i q_i x_{1i}}{\sum_{i\in s} d_i q_i}\right)\left(y_i - \frac{\sum_{i\in s} d_i q_i y_i}{\sum_{i\in s} d_i q_i}\right) d_i q_i \\
&\quad + \cdots + \tau_p \sum_{i\in s}\left(x_{pi} - \frac{\sum_{i\in s} d_i q_i x_{pi}}{\sum_{i\in s} d_i q_i}\right)\left(y_i - \frac{\sum_{i\in s} d_i q_i y_i}{\sum_{i\in s} d_i q_i}\right) d_i q_i.
\end{aligned} \tag{2.25}$$

The best choice of $\sqrt{2\alpha}$ is

$$\sqrt{2\alpha} = \frac{\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi}}{\sqrt{\psi_2}} \tag{2.26}$$

and by the central limit theorem,

$$\sqrt{2\alpha} \sim N(0,1). \tag{2.27}$$

From Equation (2.24), $\hat{t}_{y.gsin}$ reduces to

$$\hat{t}_{y.gsin} = \hat{t}_{y\pi} + \frac{\psi_1}{\psi_2}\left(\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi}\right), \tag{2.28}$$

which is a GREG type estimator and generalizes the estimator proposed by Singh (2013).

Remark 1. Al-Yaseen (2014) showed that the estimator given by Equation (1.13) can be obtained theoretically and without using the approach given by Equation (1.12), which clarifies the concern mentioned in Singh (2013), Remark 1.

Remark 2. The estimator $\hat{t}_{y.gsin}$ which is defined by Equation (2.28) depends on $\tau_1, \ldots, \tau_p$. To find the MSE for $\hat{t}_{y.gsin}$ and differentiate with respect to $\tau_1, \ldots, \tau_p$, equate to zero and solve for $\tau_1, \ldots, \tau_p$ is difficult even for $p = 2$.

To apply our approach to real data set, we have to assign values for $\tau_1, \ldots, \tau_p$. Consider the following suggestions:

1. You can assign the weights by your own choice.
2. Give the auxiliary variables equal weights *i.e.* $\tau_1 = \cdots = \tau_p = 1/p$.
3. The weight of i^{th} auxiliary variable depends on the sample correlation between the i^{th} auxiliary variable and the interest variable Y. At this case, define τ_i by

$$\tau_i = \frac{|\mathrm{corr}\,(x_i, y)|}{\sum_{j=1}^{p} |\mathrm{corr}(x_j, y)|}, \quad \text{for } i = 1, \ldots, p,$$

where

$$\mathrm{corr}\,(x_i, y) = \frac{\sum_{j=1}^{n} d_j x_{i_j} y_j - \frac{\sum_{j=1}^{n} d_j x_{i_j} \sum_{j=1}^{n} d_j y_j}{\sum_{j=1}^{n} d_j}}{\sqrt{\left(\sum_{j=1}^{n} d_j x_{i_j}^2 - \frac{\left(\sum_{j=1}^{n} d_j x_{i_j}\right)^2}{\sum_{j=1}^{n} d_j}\right)\left(\sum_{j=1}^{n} d_j y_j^2 - \frac{\left(\sum_{j=1}^{n} d_j y_j\right)^2}{\sum_{j=1}^{n} d_j}\right)}}$$

is the sample correlation between X_i and Y.

4. The weight of i^{th} auxiliary variable depends on the coefficient of variation (cv) of this auxiliary variable.

The variable with less cv has higher weight. At this case, define τ_i by

$$\tau_i = 1 - \frac{\mathrm{cv}\,(x_i)}{\sum_{j=1}^{p} \mathrm{cv}(x_j)}, \quad \text{for } i = 1, \ldots, p.$$

Further, the variable with higher cv has higher weight. At this case, define τ_i by

$$\tau_i = \frac{\mathrm{cv}\,(x_i)}{\sum_{j=1}^{p} \mathrm{cv}(x_j)}, \quad \text{for } i = 1, \ldots, p.$$

Remark 3. Our suggestions are not the optimal choice for the parameters *i.e.* they are not the solution for

$$\frac{\partial\,\mathrm{MSE}\left(\hat{t}_{y.gsin}\right)}{\partial \tau_i} = 0, \quad \text{for } i = 1, \ldots, p.$$

3. Applying Our Approach

In this section, our main goal is to apply our approach to real data set. Further, we will compare the estimators defined by the Equations (1.6), (1.13), (2.10), and (2.28), namely

$$\hat{t}_{y.ds} = \hat{t}_{y\pi} + \hat{\beta}_{ds}\left(t_x - \hat{t}_{x\pi}\right), \quad (3.1)$$

$$\hat{t}_{y.sin} = \hat{t}_{y\pi} + \hat{\beta}_{sin}\left(t_x - \hat{t}_{x\pi}\right), \quad (3.2)$$

$$\hat{t}_{y.gds} = \hat{t}_{y\pi} + \hat{\beta}_{gds}\left(\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi}\right), \quad (3.3)$$

and

$$\hat{t}_{y.gsin} = \hat{t}_{y\pi} + \frac{\psi_1}{\psi_2}\left(\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}\pi}\right) \quad (3.4)$$

based on real data set.

Consider the Forced Expiratory Volume (FEV) data set which was used by Singh (2013). The real data set FEV is an index of pulmonary function that measures the volume of air expelled after one second of constant effort. This data is downloaded from http://www.amstat.org/publications/jse/datasets/fev.dat.txt. The FEV data set was taken from a study conducted in East Boston, Massachusetts, 1980, on 654 children aged from 3 to 19 years who were seen in the childhood respiratory disease (CRD). The variable of interest is $Y :=$ Forced expiratory volume, and the auxiliary variables are $X_1 :=$ Children age, from 3–19 years age, and $X_2 :=$ Children height in inches. For this data set, $t_y = 1724$, $t_{x_1} = 6495$, $t_{x_2} = 39988$, $\text{corr}(X_1, Y) = 0.75646$, $\text{corr}(X_2, Y) = 0.86814$, and $\text{corr}(X_1, X_2) = 0.79194$.

Our main goal is to estimate $t_y = 1724$ (pretend unknown), when $t_{x_1} = 6495$, and $t_{x_2} = 39988$ are assumed known. The estimators $\hat{t}_{y.ds}$ and $\hat{t}_{y.sin}$, defined by Equations (3.1), and (3.2) respectively, are depending on one auxiliary variable; therefore, $\hat{t}_{y.ds}$, and $\hat{t}_{y.sin}$ will be computed based on X_1, and X_2 separately; denoted by $\hat{t}_{y.ds_{x_1}}$, $\hat{t}_{y.ds_{x_2}}$, $\hat{t}_{y.sin_{x_1}}$, and $\hat{t}_{y.sin_{x_2}}$ respectively. However, the estimators $\hat{t}_{y.gds}$ and $\hat{t}_{y.gsin}$, defined by Equations (3.3) and (3.4) respectively, are the generalized form of the estimators $\hat{t}_{y.ds}$ and $\hat{t}_{y.sin}$, respectively. The estimators $\hat{t}_{y.gds}$ and $\hat{t}_{y.gsin}$ depend on the $p = 2$ auxiliary variables X_1 and X_2. Based on suggestion 3 defined earlier, let

$$\tau_1 = \frac{|\text{corr}(x_1, y)|}{|\text{corr}(x_1, y)| + |\text{corr}(x_2, y)|} \quad \text{and} \quad \tau_2 = 1 - \tau_1.$$

The empirical mean (EM) of the estimator $\hat{t}_{y.*}$ of t_y is defined by

$$\text{EM}\left(\hat{t}_{y.*}\right) = \frac{1}{m}\sum_{i=1}^{m} \hat{t}_{y.*_i}, \quad (3.5)$$

where $\hat{t}_{y.*_i}$ is the estimate of t_y based on the i^{th} simulated random sample, $i = 1, \ldots, m$. The empirical relative bias (ERB) of $\hat{t}_{y.*}$ is defined by

$$\text{ERB}\left(\hat{t}_{y.*}\right) = \frac{\text{EM}\left(\hat{t}_{y.*}\right) - t_y}{t_y} \times 100\%. \quad (3.6)$$

The empirical mean squares error (EMSE) of $\hat{t}_{y.*}$ is defined by

$$\text{EMSE}\left(\hat{t}_{y.*}\right) = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{t}_{y.*} - t_y\right)^2, \quad (3.7)$$

Table 1: Under srs: Comparisons between different estimators. τ_1 and τ_2 reported in this table are the means of 3000 random values of τ_1, and τ_2 computed from 3000 random samples.

n	τ_1	τ_2		$\hat{t}_{y.ds_{x_1}}$	$\hat{t}_{y.ds_{x_2}}$	$\hat{t}_{y.sin_{x_1}}$	$\hat{t}_{y.sin_{x_2}}$	$\hat{t}_{y.gds}$	$\hat{t}_{y.gsin}$
			EM	1725.692	1724.958	1727.812	1721.621	1726.514	1721.596
30	0.4645	0.5355	ERB	0.0718	0.0292	0.1947	−0.1643	0.1195	−0.1657
			RE	1.7777	2.2253	1.7496	1.0391	3.7901	1.0000
			EM	1726.390	1724.545	1728.090	1722.503	1725.489	1722.892
40	0.4652	0.5348	ERB	0.1123	0.0053	0.2108	−0.1131	0.0600	−0.0906
			RE	1.8328	2.4739	1.8332	1.0393	4.2814	1.0000
			EM	1724.919	1723.686	1725.882	1722.291	1724.290	1722.371
50	0.4656	0.5344	ERB	0.0269	−0.0445	0.0828	−0.1254	−0.0095	−0.1208
			RE	1.8422	2.5459	1.8375	1.0394	4.4303	1.0000
			EM	1725.182	1724.820	1726.113	1723.377	1725.453	1723.35
60	0.4653	0.5347	ERB	0.0422	0.0212	0.0962	−0.0624	0.0579	−0.0640
			RE	1.8420	2.6589	1.8225	1.0434	4.5972	1.0000
			EM	1723.64	1723.305	1724.520	1722.573	1723.643	1722.428
70	0.4651	0.5349	ERB	−0.0472	−0.0666	0.0038	−0.1091	−0.0470	−0.1175
			RE	1.8402	2.4440	1.7906	1.0440	4.2090	1.0000
			EM	1725.074	1725.111	1725.957	1723.969	1725.654	1723.895
80	0.4649	0.5351	ERB	0.0360	0.0381	0.0872	−0.0281	0.0696	−0.0324
			RE	1.9233	2.4416	1.8606	1.0268	4.2271	1.0000

and the empirical relative mean squares error (RE) of the estimator $\hat{t}_{y.*}$ is defined by

$$\mathrm{RE}\left(\hat{t}_{y.*}\right) = \frac{\frac{1}{m}\sum_{i=1}^{m}\left(\hat{t}_{y.*} - t_y\right)^2}{\frac{1}{m}\sum_{i=1}^{m}\left(\hat{t}_{y.gsin} - t_y\right)^2}. \tag{3.8}$$

For the case, $q_i = 1$, for $i \in s$, and under simple random sampling without replacement (srs) design, draw $m = 3000$ random samples of size n from the FEV data set by using procedure `surveyselect` of SAS Institute. For each random sample, compute $\hat{t}_{y.ds_{x_1}}$, $\hat{t}_{y.ds_{x_2}}$, $\hat{t}_{y.sin_{x_1}}$, $\hat{t}_{y.sin_{x_2}}$, $\hat{t}_{y.gds}$, and $\hat{t}_{y.gsin}$. Based on $m = 3000$ random samples, compute EM, ERB, and RE for such estimators. The computations are done by using a macro written in SAS. For $n = 30, 40, 50, 60, 70$, and 80. The results are summarized in Table 1.

Divide the FEV data set into two strata according to the variable sex. From each stratum, under srs, draw a random sample of size n_h and combine the two samples into one sample, and compute $\hat{t}_{y.ds_{x_1}}$, $\hat{t}_{y.ds_{x_2}}$, $\hat{t}_{y.sin_{x_1}}$, $\hat{t}_{y.sin_{x_2}}$, $\hat{t}_{y.gds}$, and $\hat{t}_{y.gsin}$. Based on $m = 3000$ drawn random samples, compute EM, ERB, and RE for each estimator. For $n_h = 15, 20, 25, 30, 35$, and 40 the results are summarized in Table 2.

4. Final Remarks

In this paper, the estimators proposed by Deville and Särndal (1992) and Singh (2013) are generalized to use p multi-auxiliary variables that are available in the study. The generalized estimators, $\hat{t}_{y.gds}$ and $\hat{t}_{y.gsin}$, are GREG type estimators.

From Table 1 and Table 2, regardless of the sample size n, the estimators $\hat{t}_{y.ds_{x_1}}$, $\hat{t}_{y.ds_{x_2}}$, $\hat{t}_{y.sin_{x_1}}$, $\hat{t}_{y.sin_{x_2}}$, $\hat{t}_{y.gds}$, and $\hat{t}_{y.gsin}$, have negligible empirical relative bias (ERB). Further, the estimator $\hat{t}_{y.gsin}$ is more efficient than the estimators $\hat{t}_{y.ds_{x_1}}$, $\hat{t}_{y.ds_{x_2}}$, $\hat{t}_{y.sin_{x_1}}$, $\hat{t}_{y.sin_{x_2}}$, $\hat{t}_{y.gds}$, since the other estimators have empirical relative mean squares error (RE) greater than one.

Table 2: Stratified sampling design: Under srs, draw random sample of size n_h from each stratum and combined samples into one sample of size n. τ_1 and τ_2 reported in this table are the means of 3000 random values of τ_1, and τ_2 computed from 3000 random samples.

n_h	τ_1	τ_2		$\hat{t}_{y.ds_{x_1}}$	$\hat{t}_{y.ds_{x_2}}$	$\hat{t}_{y.sin_{x_1}}$	$\hat{t}_{y.sin_{x_2}}$	$\hat{t}_{y.gds}$	$\hat{t}_{y.gsin}$
			EM	1725.640	1723.545	1727.790	1721.359	1724.522	1721.467
15	0.4645	0.5355	ERB	0.0688	−0.0527	0.1935	−0.1795	0.0039	−0.1732
			RE	1.7393	2.3657	1.7785	1.0507	4.0535	1.0000
			EM	1725.586	1724.3	1727.142	1722.638	1725.027	1722.632
20	0.4651	0.5349	ERB	0.0656	−0.0089	0.1559	−0.1053	0.0332	−0.1056
			RE	1.7483	2.3497	1.7454	1.0631	4.0592	1.0000
			EM	1725.721	1724.079	1726.742	1722.956	1724.527	1723.042
25	0.4652	0.5348	ERB	0.0735	−0.0218	0.1327	−0.0869	0.0042	−0.0819
			RE	1.6900	2.5176	1.6865	1.0755	4.3359	1.0000
			EM	1725.368	1724.607	1726.51	1,723.346	1725.147	1723.404
30	0.4653	0.5347	ERB	0.0530	0.0089	0.1192	−0.0643	0.0402	−0.0609
			RE	1.7995	2.4671	1.7629	1.0499	4.2482	1.0000
			EM	1725.316	1724.231	1726.020	1723.451	1724.610	1723.59
35	0.4653	0.5347	ERB	0.0500	−0.0129	0.0908	−0.0581	0.0091	−0.0501
			RE	1.8396	2.3301	1.7979	1.0311	4.0332	1.0000
			EM	1726.023	1725.224	1726.695	1723.992	1725.834	1724.192
40	0.4651	0.5349	ERB	0.0910	0.0446	0.1300	−0.0268	0.0800	−0.0152
			RE	1.8349	2.5009	1.7839	1.0781	4.3225	1.0000

Acknowledgement

The author is grateful to the three referees for their valuable comments and suggestions. Also, my grateful is extended to my friend Dr. Abdul-Razzaq Mugdadi for editing this work.

References

Al-Jthorrarha, J. and Bataineh, A. (2015). Estimation of finite population ratio when other auxiliary variables are available in the study, *Austrian Journal of Statistics*, **44**, 33–46.

Al-Yaseen, A. (2014). Penalized chi-square distance and the dual calibration for estimating the finite population total, *Master Thesis. Statistics Department Yarmouk University*, Jordan.

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

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**, 663–685.

Singh, S. (2013). A dual problem of calibration of design weights, *Statistics: A Journal of Theoretical and Applied Statistics*, **47**, 566–574.

Stearns, M. and Singh, S. (2008). On the estimation of the general parameter, *Computational Statistics and Data Analysis*, **52**, 4253–4271.

Received December 30, 2014; Revised February 9, 2015; Accepted February 9, 2015