Two-Step Calibration of Design Weights in Survey Sampling

Sarjinder Singh and Stephen Andrew Sedory
Department of Mathematics and Statistics
Texas A&M University-Kingsville
Kingsville, TX 78363, USA
E-mail: sarjinder@yahoo.com

ABSTRACT

In this paper, a new two-step technique for the calibration of design weights is proposed. In the first step, the calibration weights are set proportional to the design weights in a given sample. In the second step, the constants of proportionality are determined based on different objectives of the investigator such as bias reduction or minimum mean squared error. Many estimators available in the literature can be shown to be special cases of the proposed two-step calibrated estimator. A simulation study, based on a real data set, is included at the end. A few technical issues are raised with respect to the use of the proposed calibration technique, both limitations and benefits are discussed.

Keywords: Calibration, estimation of total, bias and variance.

1. INTRODUCTION

A new method to calibrate the design weights in the Horvitz-Thompson (1952) estimator was considered by Deville and Särndal (1992) by making use of auxiliary information. Their proposed calibration methods provide a class of estimators. Some well-known estimators such as the classical ratio-estimator belong to this class. Several authors, including Singh (2003, 2004, 2006, 2012), Farrell and Singh (2002, 2005), Wu and Sitter (2001), Estevao and Särndal (2003), Kott (2003), Montanari and Ranalli (2005), Rueda et al. (2006, 2007) among others, considered the Deville and Särndal (1992) method and derived important calibrated estimators. But, so far, an estimator, from the class of calibrated estimators derived by the Deville and Särndal (1992) method, that is always more efficient than the traditional linear regression estimator in the presence of a linear trend, has not been found in the literature. In the present paper, we consider a subclass of the class of calibrated estimators provided by Deville and Särndal (1992) such that at the first step the calibrated weights are set proportional to the design weights in a given sample and at the second step the constants of proportionality are determined based on different requirements of the investigator.

For simplicity, we consider a single auxiliary variable. For a sample s and for $i \in s$, let $\left(y_i,\ x_i\right)$ be the values observed. The population total of the auxiliary variable x, $X = \sum_{i \in \Omega} x_i$, is assumed to be known. The objective is to estimate the population total $Y = \sum_{i \in \Omega} y_i$. The well known Horvitz and Thompson (1952) estimator of the population total Y is given by:

$$\hat{Y}_{\text{HT}} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} d_i y_i \tag{1.1}$$

where $d_i = 1/\pi_i$ are called design weights and π_i are called first order inclusion probabilities. Deville and Särndal (1992) proposed the calibrated estimator of the population total Y as:

$$\hat{Y}_{\text{ds}} = \sum_{i \in s} w_i y_i \tag{1.2}$$

where the calibrated weights w_i, $i \in s$ are obtained by minimizing the chi-squared distance defined as:

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

subject to the calibration constraint:

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

Here q_i, $i \in s$ are suitably chosen weights. In many situations the value of q_i is 1. The form of the estimator (1.2) depends upon the choice of q_i. Minimization of (1.3) subject to calibration equation (1.4), leads to the calibrated weight:

$$w_i = d_i + \frac{d_i q_i x_i}{\sum_{i \in s} d_i q_i x_i^2} \left(X - \sum_{i \in s} d_i x_i \right) \tag{1.5}$$

Substitution of the value of w_i from (1.5) into (1.2) leads to the generalized regression (GREG) estimator of the population total Y as:

$$\hat{Y}_{\text{G}} = \hat{Y}_{HT} + \hat{\beta}_{ds} \left(X - \hat{X}_{HT} \right) \tag{1.6}$$

where

$$\hat{\beta}_{\text{ds}} = \left(\sum_{i \in s} d_i q_i x_i y_i \right) \Big/ \left(\sum_{i \in s} d_i q_i x_i^2 \right) \tag{1.7}$$

Wu and Sitter (2001) suggest a second calibration constraint, in addition to the calibration constraint (1.4), given by:

$$\sum_{i \in s} w_i = N \tag{1.8}$$

Wu and Sitter (2001) estimator takes the form:

$$\hat{Y}_{WS} = \hat{Y}_{HT} + \hat{\beta}_1\left(X - \hat{X}_{HT}\right) + \hat{\beta}_0\left(N - \sum_{i \in s} d_i\right) = \hat{Y}_{LR} + \hat{\beta}_0(N - \hat{N}) \tag{1.9}$$

where

$$\hat{\beta}_0 = \frac{\left(\sum_{i \in s} d_i q_i x_i^2\right)\left(\sum_{i \in s} d_i q_i y_i\right) - \left(\sum_{i \in s} d_i q_i x_i y_i\right)\left(\sum_{i \in s} d_i q_i x_i\right)}{\left(\sum_{i \in s} d_i q_i\right)\left(\sum_{i \in s} d_i q_i x_i^2\right) - \left(\sum_{i \in s} d_i q_i x_i\right)^2} \tag{1.10}$$

and

$$\hat{\beta}_1 = \frac{\left(\sum_{i \in s} d_i q_i\right)\left(\sum_{i \in s} d_i q_i x_i y_i\right) - \left(\sum_{i \in s} d_i q_i y_i\right)\left(\sum_{i \in s} d_i q_i x_i\right)}{\left(\sum_{i \in s} d_i q_i\right)\left(\sum_{i \in s} d_i q_i x_i^2\right) - \left(\sum_{i \in s} d_i q_i x_i\right)^2} \tag{1.11}$$

have their usual meanings. Singh (2003, 2004, 2006, 2011, 2012), Stearns and Singh (2008) and Singh and Arnab (2011) suggest replacing of the constraint (1.8) with a new constraint:

$$\sum_{i \in s} w_i = \sum_{i \in s} d_i \tag{1.12}$$

to get the traditional linear regression estimator due to Hansen, Hurwitz and Madow (1953) given by:

$$\hat{Y}_{LR} = \hat{Y}_{HT} + \hat{\beta}_{ols}\left(X - \hat{X}_{HT}\right) \tag{1.13}$$

where

$$\hat{\beta}_{ols} = \hat{\beta}_1 \tag{1.14}$$

Singh and Arnab (2011) have shown that it is not likely that the Wu and Sitter (2001) estimator could perform better than the linear regression estimator in the presence of a linear trend. Kim (2010) suggests using the constraint (1.8) if the population size is know and (1.12) if the population size is unknown. (for details about Kim's (2010) contribution please refer to Singh (2011) "Dual problem of calibration of design weights")

There seems to be a debate over the choice of the calibration constraint set by Wu and Sitter (2011) and that set by Singh (2003). We provide here an alternative way of looking at the issue which we hope will add some clarity.

2. TWO-STEP CALIBRATION TECHNIQUE

In the first step, we suggest setting the calibration weight w_i proportional to the design weight d_i, that is:

$$w_i \propto d_i \tag{2.1}$$

or equivalently,

$$w_i = k_i d_i \tag{2.2}$$

where k_i are constants of proportionality to be determined based on different options that might be considered by an investigator. Summing both sides of (2.2), we have:

$$\sum_{i \in s} w_i = \sum_{i \in s} k_i d_i \tag{2.3}$$

Now we consider the Lagrange function given by:

$$L = \frac{1}{2} \sum_{i \in s} \frac{(w_i - d_i)^2}{d_i q_i} - \lambda_0 \left(\sum_{i \in s} w_i - \sum_{i \in s} k_i d_i \right) - \lambda_1 \left(\sum_{i \in s} w_i x_i - X \right) \tag{2.4}$$

where λ_0 and λ_1 are the Lagrange multipliers.

On setting $\frac{\partial L}{\partial w_i} = 0$, we get

$$w_i = d_i + \lambda_0 d_i q_i + \lambda_1 d_i q_i x_i \tag{2.5}$$

On substituting (2.5) into (2.3) and (1.4), we have

$$\lambda_0 \sum_{i \in s} d_i q_i + \lambda_1 \sum_{i \in s} d_i q_i x_i = \sum_{i \in s} (k_i - 1) d_i \tag{2.6}$$

and

$$\lambda_0 \sum_{i \in s} d_i q_i x_i + \lambda_1 \sum_{i \in s} d_i q_i x_i^2 = (X - \hat{X}_{\text{HT}}) \tag{2.7}$$

On solving for λ_0 and λ_1 and substituting into (2.5), the calibrated weights are given by:

$$w_i = d_i + \frac{d_i q_i \left[\left(\sum_{i \in s} d_i q_i x_i^2 \right) \left(\sum_{i \in s} (k_i - 1) d_i \right) - \left(\sum_{i \in s} d_i q_i x_i \right) \left(X - \hat{X}_{\text{HT}} \right) \right]}{\left(\sum_{i \in s} d_i q_i \right) \left(\sum_{i \in s} d_i q_i x_i^2 \right) - \left(\sum_{i \in s} d_i q_i x_i \right)^2}$$

$$+\frac{d_i q_i x_i\left[\left(X-\hat{X}_{\mathrm{HT}}\right)\left(\sum_{i\in s} d_i q_i\right)-\left(\sum_{i\in s} d_i q_i x_i\right)\left(\sum_{i\in s}(k_i-1)d_i\right)\right]}{\left(\sum_{i\in s} d_i q_i\right)\left(\sum_{i\in s} d_i q_i x_i^2\right)-\left(\sum_{i\in s} d_i q_i x_i\right)^2} \qquad (2.8)$$

On substituting the calibrated weights (2.8) into (1.2), we obtain a new class of estimators of the population total given by:

$$\hat{Y}_{\text{new}}=\hat{Y}_{HT}+\hat{\beta}_1\left(X-\hat{X}_{HT}\right)+\hat{\beta}_0\sum_{i\in s}(k_i-1)d_i \qquad (2.9)$$

Obvious Special Cases:

(a) If $k_i=1$ for all $i\in s$, the proposed estimator (2.9) reduces to the traditional regression estimator derived by Singh (2003).

(b) If $k_i=\frac{N}{n}\pi_i$ for $i\in s$, the proposed estimator (2.9) reduces to the estimator due to Wu and Sitter (2001). In the next section, we show that such a choice of k_i for all $i\in\Omega$ makes the Wu and Sitter (2001) estimator equivalent to the traditional linear regression estimator.

3. PROPERTIES OF THE PROPOSED ESTIMATOR

In order to study the properties of the proposed estimator, we define:

$$\varepsilon_0=\frac{\hat{Y}_{HT}}{Y}-1,\ \varepsilon_1=\frac{\hat{X}_{HT}}{X}-1,\ \delta_0=\frac{\hat{\beta}_0}{\beta_0}-1,\ \delta_1=\frac{\hat{\beta}_1}{\beta_1}-1 \text{ and } \eta=\frac{\sum_{i\in s}(k_i-1)d_i}{\sum_{i\in\Omega}(k_i-1)}-1$$

so that

$$E(\varepsilon_0)=E(\varepsilon_1)=E(\delta_0)=E(\delta_1)=E(\eta)=0$$

Now the new proposed estimator $\hat{Y}_{new}$ in terms of ε_0, ε_1, δ_0, δ_1 and η can be easily written as:

$$\hat{Y}_{\text{new}}=Y(1+\varepsilon_0)+\beta_1(1+\delta_1)[X-X(1+\varepsilon_1)]+\beta_0(1+\delta_0)[\sum_{i\in\Omega}(k_i-1)](1+\eta)$$

or

$$\hat{Y}_{\text{new}}=Y+Y\,\varepsilon_0-\beta_1X\varepsilon_1-\beta_1X\varepsilon_1\delta_1+\beta_0\left[\sum_{i\in\Omega}k_i-N\right](1+\delta_0+\eta+\delta_0\eta) \qquad (3.1)$$

Note carefully that if $k_i=\frac{N}{n}\pi_i$ for all $i\in\Omega$, then $(\sum_{i\in\Omega}k_i-N)=0$ for any unequal probability sampling design. Thus the choice $k_i=\frac{N}{n}\pi_i$ for all $i\in\Omega$ makes the Wu and Sitter (2001) estimator the same as the linear regression estimator. In other words, for a choice of $k_i = \frac{N}{n}\pi_i$ for all $i \in \Omega$, the traditional linear regression estimator is the optimal estimator among the Wu and Sitter (2001) class of estimators. See Singh and Arnab (2011) where they have shown through a simulation study that the Wu and Sitter (2001) estimator is likely to remain less efficient than the traditional linear regression estimator.

Now a natural question arises? Is there any choice of the proportionality constants k_i such that the proposed estimator can perform more efficiently than the linear regression estimator? As is often the case these does not seem to be an obvious answer, but we made an attempt in this direction as follows.

Remark: It is not easy to analytically find individual values of k_i which results either in reduction in bias or minimum mean squared error of the newly estimator.

Assume $k_i = k$ for all $i \in s$, then we set $\sum_{i\in s} w_i \propto \sum_{i\in s} d_i$ or equivalently $\sum_{i\in s} w_i = k \sum_{i\in s} d_i$. Taking expected value on both sides of (3.1), and setting the bias in the estimator equal to zero, we get:

$$\sum_{i\in\Omega} k_i = Nk = N + \frac{\beta_1 X E(\varepsilon_1\delta_1)}{\beta_0\{1 + E(\varepsilon_0\eta)\}} \tag{3.2}$$

Thus to get an unbiased estimator of the population total, one obvious choice of the proportionality constants k is given by:

$$k = 1 + \frac{Cov\left(\hat{\beta}_1, \hat{X}_{HT}\right)}{N\beta_0 + Cov\left(\hat{\beta}_0, \sum_{i\in s} d_i\right)} \quad \text{for all } i \in \Omega \tag{3.3}$$

So to obtain an exactly unbiased estimator of the population total, the value of the proportionality constants k_i remain constant, but differ from unity.

Let us look at another choice for the values of the proportionality constants (k_i) based on minimizing the mean squared error of the resultant estimator $\hat{Y}_{\text{new}}$. For simplicity, let us assume $k_i = k$ for all $i \in \Omega$. Under such an assumption, we get

$$\eta = \frac{\sum_{i\in s}(k_i - 1)d_i}{\sum_{i\in\Omega}(k_i - 1)} - 1 = \frac{\sum_{i\in s}(k-1)d_i}{\sum_{i\in\Omega}(k-1)} - 1 = \frac{\sum_{i\in s} d_i}{N} - 1$$

and the new estimator can be written as:

$$\hat{Y}_{\text{new}} = Y + Y\ \varepsilon_0 - \beta_1 X\varepsilon_1 - \beta_1 X\varepsilon_1\delta_1 + \beta_0 N\left(K - 1\right)\left(1 + \delta_0 + \eta + \delta_0\eta\right) \tag{3.4}$$

By the definition of the mean squared error, we have:

$$\begin{aligned}
MSE(\hat{Y}_{new}) &= E\left[\hat{Y}_{new} - Y\right]^2 \\
&= E\left[Y\varepsilon_0 - \beta_1 X\varepsilon_1 - \beta_1 X\varepsilon_1\delta_1 + \beta_0 N(K-1)(1+\delta_0+\eta+\delta_0\eta)\right]^2 \\
&= E\Big[Y^2\varepsilon_0^2 + \beta_1^2 X^2\varepsilon_1^2 + \beta_1^2 X^2\varepsilon_1^2\delta_1^2 + \beta_0^2 N^2(K-1)^2(1+\delta_0+\eta+\delta_0\eta)^2 \\
&\quad - 2YX\beta_1\varepsilon_0\varepsilon_1 - 2Y\beta_1 X\varepsilon_0\varepsilon_1\delta_1 + 2\beta_0 YN(K-1)(\varepsilon_0+\varepsilon_0\delta_0+\varepsilon_0\eta+\varepsilon_0\delta_0\eta) \\
&\quad + 2\beta_1^2 X^2\varepsilon_1^2\delta_1 - 2\beta_0\beta_1 XN(K-1)(\varepsilon_1+\varepsilon_1\delta_0+\varepsilon_1\eta+\varepsilon_1\delta_0\eta) \\
&\quad - 2\beta_0\beta_1 XN(K-1)(\varepsilon_1\delta_1+\varepsilon_1\delta_0\delta_1+\varepsilon_1\delta_1\eta+\varepsilon_1\delta_0\delta_1\eta)\Big]
\end{aligned}$$

By taking the expected values and by neglecting the higher order terms, we have

$$\begin{aligned}
MSE(\hat{Y}_{new}) &\approx V(\hat{Y}_{HT}) + \beta_1^2 V(\hat{X}_{HT}) - 2\beta_1 Cov(\hat{Y}_{HT}, \hat{X}_{HT}) \\
&\quad + (K-1)^2\left[N^2\beta_0^2 + N^2 V(\hat{\beta}_0) + \beta_0^2 V(\hat{N}) + 4N\beta_0 ECov(\hat{\beta}_0, \hat{N})\right] \\
&\quad + 2(K-1)\Big[N\beta_0 Cov(\hat{X}_{HT}, \hat{\beta}_1) + \beta_1 NCov(\hat{X}_{HT}, \hat{\beta}_0) + \beta_0\beta_1 Cov(\hat{X}_{HT}, \hat{N}) \\
&\quad - NCov(\hat{Y}_{HT}, \hat{\beta}_0) - \beta_0 Cov(\hat{Y}_{HT}, \hat{N})\Big]
\end{aligned}$$

On setting $\frac{\partial MSE(\hat{Y}_{new})}{\partial(K-1)} = 0$ and keeping the same constants β_0 and β_1 we have,

$$K = 1 + \frac{N\beta_0 Cov(\hat{X}_{HT}, \hat{\beta}_1) + \beta_1 NCov(\hat{X}_{HT}, \hat{\beta}_0) + \beta_0\beta_1 Cov(\hat{X}_{HT}, \hat{N}) - NCov(\hat{Y}_{HT}, \hat{\beta}_0) - \beta_0 Cov(\hat{Y}_{HT}, \hat{N})}{N^2\beta_0^2 + N^2 V(\hat{\beta}_0) + \beta_0^2 V(\hat{N}) + 4\beta_0 NCov(\hat{\beta}_0, \hat{N})}$$

The minimum mean squared error of the proposed estimator is given by:

$$\begin{aligned}
\text{Min.MSE}(\hat{Y}_{\text{new}}) &= V(\hat{Y}_{HT})\left(1-\rho_{xy}^2\right) \\
&- \frac{\left\{N\beta_0 Cov(\hat{X}_{HT}, \hat{\beta}_1) + \beta_1 NCov(\hat{X}_{HT}, \hat{\beta}_0) + \beta_0\beta_1 Cov(\hat{X}_{HT}, \hat{N}) - NCov(\hat{Y}_{HT}, \hat{\beta}_0) - \beta_0 Cov(\hat{Y}_{HT}, \hat{N})\right\}^2}{N^2\beta_0^2 + N^2 V(\hat{\beta}_0) + \beta_0^2 V(\hat{N}) + 4\beta_0 NCov(\hat{\beta}_0, \hat{N})}
\end{aligned} \qquad (3.5)$$

where $\rho_{xy} = \frac{Cov(\hat{Y}_{HT}, \hat{X}_{HT})}{\sqrt{V(\hat{Y}_{HT})V(\hat{X}_{HT})}}$ is the correlation coefficient.

Note that for simple random sampling (SRS), we have:

$V(\hat{N}) = 0$, $Cov(\hat{Y}_{HT}, \hat{N}) = 0$, $Cov(\hat{X}_{HT}, \hat{N}) = 0$, and $Cov(\hat{\beta}_0, \hat{N}) = 0$

Thus for SRS design, we have:

$$Min.MSE(\hat{Y}_{new}) = \frac{(1-f)}{n} S_y^2 (1-\rho_{xy}^2) - \frac{\left\{N\beta_0 Cov(\hat{X}_{HT}, \hat{\beta}_1) + \beta_1 N Cov(\hat{X}_{HT}, \hat{\beta}_0) - N Cov(\hat{Y}_{HT}, \hat{\beta}_0)\right\}^2}{N^2 \beta_0^2 + N^2 V(\hat{\beta}_0)} \qquad (3.6)$$

From (3.5) and (3.6), it can be seen that there must a choice of K such that the proposed estimator remains more efficient than the linear regression estimator. As pointed out by Singh (2012), there is no estimator better than the linear regression estimator in the presence of a linear trend. It seems that the optimum value of K is likely to be very close to one, and of course if it is one then we get the traditional linear regression estimator.

In the next section we simulated numerical values of the optimizing constant K such that the proposed new estimator can perform better than the linear regression estimator.

4. SIMULATION STUDY

In the next section, we demonstrate performance of the proposed two-step estimator through simulation studies. In the simulation study, we consider the following five estimators and compare them with each other:

(a) Ratio Estimator:

$$\hat{Y}_{Ratio} = \hat{Y}_{HT}\left(\frac{X}{\hat{X}_{HT}}\right) = \hat{\theta}_1 \text{ (say)} \qquad (4.1)$$

(b) GREG Estimator:

$$\hat{Y}_{GREG} = \hat{Y}_{HT} + \hat{\beta}_{ds}\left(X - \hat{X}_{HT}\right) = \hat{\theta}_2 \text{ (say)} \qquad (4.2)$$

$$\text{where } \hat{\beta}_{ds} = \frac{\sum_{i \in s} d_i x_i y_i}{\sum_{i \in s} d_i x_i^2}$$

(c) Wu and Sitter Estimator:

$$\hat{Y}_{WS} = \hat{Y}_{HT} + \hat{\beta}_1\left(X - \hat{X}_{HT}\right) + \hat{\beta}_0\left(N - \sum_{i \in s} d_i\right) = \hat{\theta}_3 \text{ (say)} \qquad (4.3)$$

where

$$\hat{\beta}_0 = \frac{\left(\sum_{i \in s} d_i x_i^2\right)\left(\sum_{i \in s} d_i y_i\right) - \left(\sum_{i \in s} d_i x_i y_i\right)\left(\sum_{i \in s} d_i x_i\right)}{\left(\sum_{i \in s} d_i\right)\left(\sum_{i \in s} d_i x_i^2\right) - \left(\sum_{i \in s} d_i x_i\right)^2} \qquad (4.4)$$

and

$$\hat{\beta}_1 = \frac{\left(\sum_{i \in s} d_i\right)\left(\sum_{i \in s} d_i x_i y_i\right) - \left(\sum_{i \in s} d_i y_i\right)\left(\sum_{i \in s} d_i x_i\right)}{\left(\sum_{i \in s} d_i\right)\left(\sum_{i \in s} d_i x_i^2\right) - \left(\sum_{i \in s} d_i x_i\right)^2} \qquad (4.5)$$

(d) Hansen, Hurwitz and Madow’s estimator:

$$\hat{Y}_{LR} = \hat{Y}_{HT} + \hat{\beta}_{ols}\left(X - \hat{X}_{HT}\right) = \hat{\theta}_4 \text{(say)} \tag{4.6}$$

where

$$\hat{\beta}_{ols} = \hat{\beta}_1$$

(e) Two-Step Calibration Method Estimator:

$$\hat{Y}_{\text{new}} = \hat{Y}_{HT} + \hat{\beta}_1\left(X - \hat{X}_{HT}\right) + \hat{\beta}_0 \sum_{i \in s}(k_i - 1)d_i = \hat{\theta}_5 \text{(say)} \tag{4.7}$$

We use a dataset, FEV.DAT, available on the CD that accompanies the text by Rosner (2006), that contains data on 654 children from the Childhood Repertory Disease Study done in Boston. Among the variables are height, age, and FEV (forced expiratory volume). We consider the problem of estimating total (or average) height of the population using age at the estimation stage and FEV at the selection stage using the Midzuno (1952) and Sen (1953) sampling scheme. A pictorial representation of the datasets used is shown in the following scatter plots:

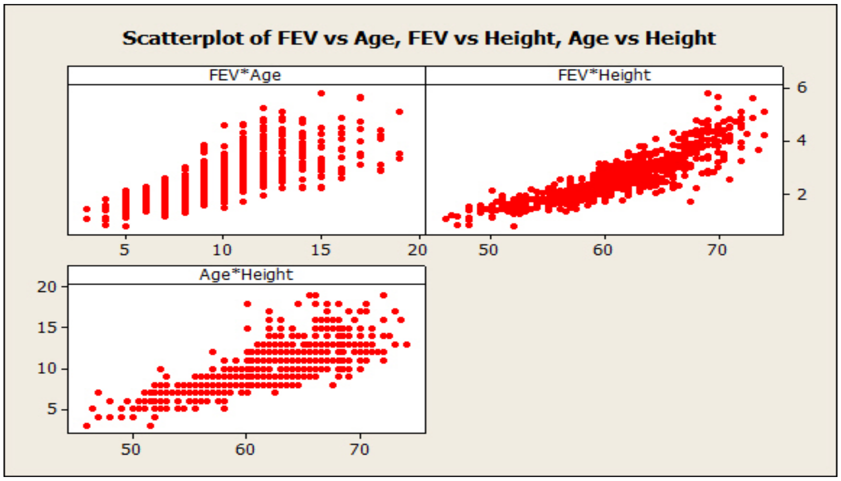

Fig. 4.1. Scatter plots of the three variables considered in the study.

To investigate various situations, we apply Box-Cox type transformations on all the three variables in the population as:

$$Y_i = \frac{(Height)^T - 1}{T}, \quad X_i = \frac{(Age)^T - 1}{T} \text{ and } Z_i = \frac{(FEV)^T - 1}{T} \tag{4.8}$$

for different choice of values of T. For a given value of $T = 2.0$ (say), from the population of size $N = 654$, we select a sample of size n (say). Now we compute four different population correlation coefficients:

$$\rho_{xy} = \frac{\sum_{i \in \Omega}\left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)}{\sqrt{\sum_{i \in \Omega}\left(X_i - \bar{X}\right)^2 \sum_{i \in \Omega}\left(Y_i - \bar{Y}\right)^2}} \tag{4.9}$$

$$\rho_{xz} = \frac{\sum_{i\in\Omega}\left(Z_i - \bar{Z}\right)\left(X_i - \bar{X}\right)}{\sqrt{\sum_{i\in\Omega}\left(Z_i - \bar{Z}\right)^2 \sum_{i\in\Omega}\left(X_i - \bar{X}\right)^2}} \tag{4.10}$$

$$\rho_{yz} = \frac{\sum_{i\in\Omega}\left(Y_i - \bar{Y}\right)\left(Z_i - \bar{Z}\right)}{\sqrt{\sum_{i\in\Omega}\left(Y_i - \bar{Y}\right)^2 \sum_{i\in\Omega}\left(Z_i - \bar{Z}\right)^2}} \tag{4.11}$$

and

$$\rho^*_{xy} = \frac{(\sum_{i\in\Omega} d_i)(\sum_{i\in\Omega} d_i x_i y_i) - (\sum_{i\in\Omega} d_i x_i)(\sum_{i\in\Omega} d_i y_i)}{\sqrt{(\sum_{i\in\Omega} d_i)(\sum_{i\in\Omega} d_i x_i^2) - (\sum_{i\in\Omega} d_i x_i)^2}\sqrt{(\sum_{i\in\Omega} d_i)(\sum_{i\in\Omega} d_i y_i^2) - (\sum_{i\in\Omega} d_i y_i)^2}} \tag{4.12}$$

For the Midzuno-Sen sampling scheme, the value of the first order inclusion probability π_i is given by:

$$\pi_i = \frac{(N-n)}{N-1}P_i + \frac{(n-1)}{(N-1)}, \text{ where } P_i = Z_i \Big/ \sum_{i\in\Omega} Z_i \,. \tag{4.13}$$

Note that the population correlation coefficients ρ_{xy}, ρ_{xz} and ρ_{yz} are simple Pearson's correlation coefficients. Their values play more of a role when one applies the simple random sampling design. But for any unequal probability sampling design, such as Midzuno-Sen sampling design, the weighted Pearson's correlation coefficient ρ^*_{xy} plays the key role. Thus, for each value of T and n we also computed the value of ρ^*_{xy}. The first unit has been selected with probability-proportional-to-size sampling by using the cumulative total method, and the remaining $(n-1)$ units are selected by SRSWOR sampling by using the standard IMSL subroutine: CALL RNSRI (NEWNS,NKK,IR). For $B = 5{,}000$ samples, we computed the empirical relative bias in the kth estimator $(k = 1,2,3,4,5)$ as:

$$\text{RB}\left(\hat{\theta}_k\right) = \frac{\frac{1}{B}\sum_{j=1}^{B}(\hat{\theta}_k)_j - \bar{Y}}{\bar{Y}} \times 100\% = \text{RB(k)} \qquad \text{(say)} \tag{4.14}$$

and the mean squared error is computed as:

$$\text{MSE}\left(\hat{\theta}_k\right) = \frac{1}{B}\sum_{j=1}^{B}\left[(\hat{\theta}_k)_j - \bar{Y}\right]^2 \tag{4.15}$$

The percent relative efficiency of the 5th estimator $\hat{\theta}_5$ with respect to the other three estimators $\hat{\theta}_k$ for $k = 1,2,3,4$ is computed as:

$$\mathrm{RE}\left(\hat{\theta}_k, \hat{\theta}_5\right) = \frac{MSE\left(\hat{\theta}_k\right)}{MSE\left(\hat{\theta}_5\right)} \times 100\% = \mathrm{RE(k,5)} \qquad \text{(say)} \qquad (4.16)$$

The value of T is changed from 0.5 to 4.5 with a step of 0.5, and the sample size n changed from 20 to 60. The proposed two-step calibration estimator remains more efficient than the traditional regression estimator for transformations chosen with $T \geq 2$. A pictorial representation of such data sets which have been used in the simulation study is given in Fig.4.2.

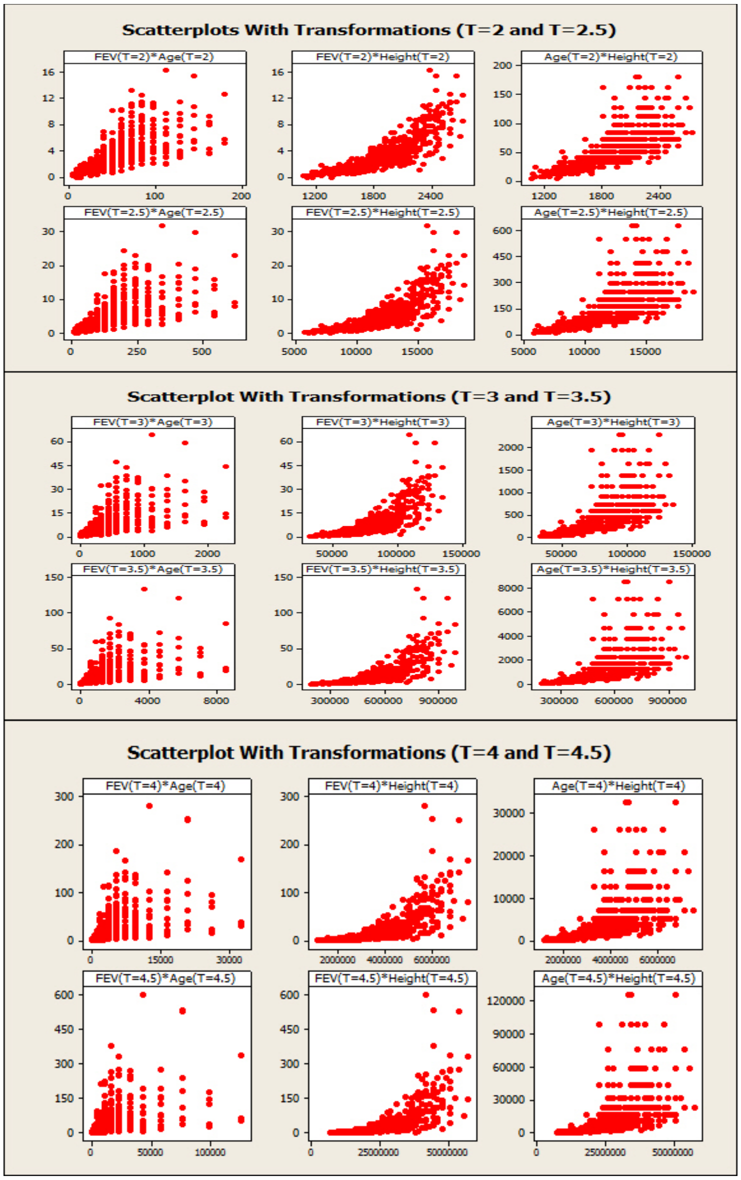

Fig. 4.2. A pictorial representation of data sets used.

Fig. 4.2 shows that as the value of T increases from 2.0 to 4.5 with a step of 0.5, the scatterness in the datasets becomes wider and wider, and breaks the assumption of linearity. Figure 4.3 shows dot plots of the values of the transformation T and the sample size n for which the proposed two-step estimator remains more efficient than the linear regression estimator and the absolute value of the relative bias remains less than 10%. There are more situations when the value of T is close to 4.5 and sample size n is close to 20 where the proposed estimator performs better than the linear regression estimator. Singh and Arnab (2011) reported that the estimator of Wu and Sitter (2001) is likely to remain less efficient if sample size is large, and the same conclusion seems to be true in the case of the proposed two-step calibration estimator.

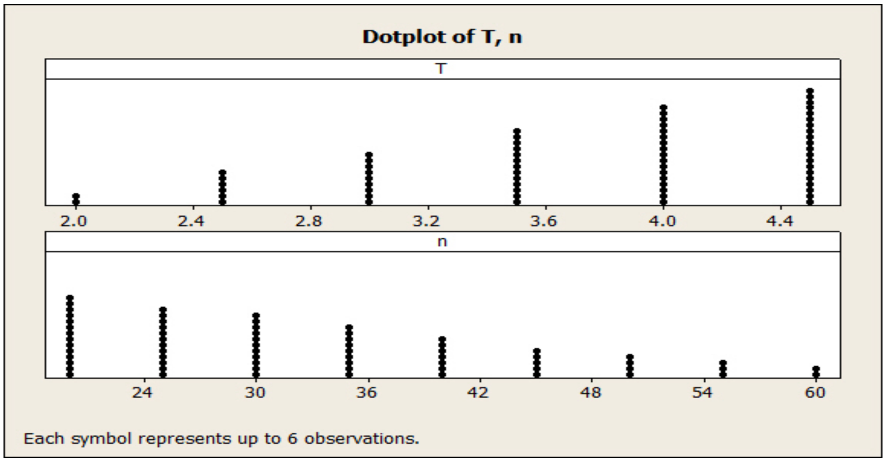

Fig. 4.3. Practicable values of T and n where proposed estimator performs well.

Figure 4.4 is devoted to displaying the values of K and the values of the percent relative efficiency (RE) of the proposed two-step estimator with respect to the traditional linear regression estimator. The presentation shows that the choice of K ranges between 0.980 and 0.996. The percent relative efficiency (RE) value changes between 103% and 108.8%.

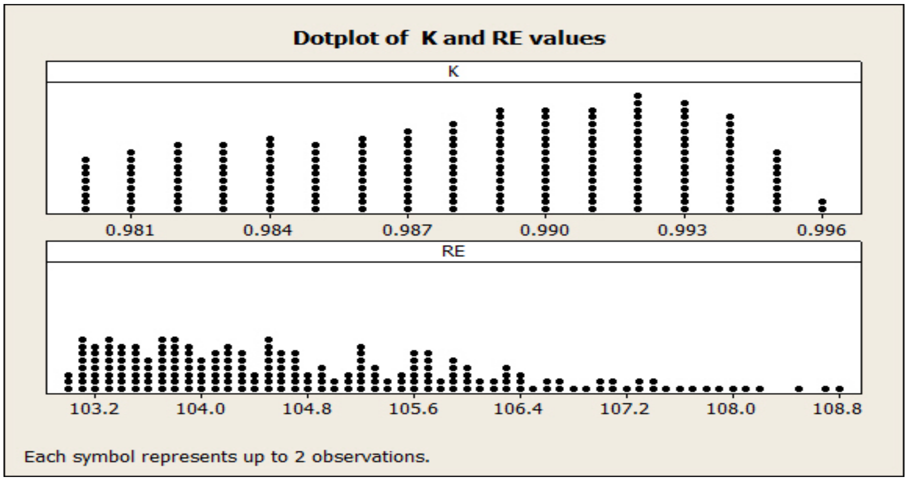

Fig. 4.4. Practicable values of K and percent relative efficiency.

Table 4.1 provides descriptive statistics of the values of the unknown parameter K which are used in the simulation study. Out of 386 cases, where the two-step estimator performs better, the minimum value of K is found to be 0.9800, maximum value to be 0.996 and with a median value to be 0.9890. The average value of K is 0.9883 with a standard deviation of 0.00439.

Table 4.1. Choice of value of K .

Variables	freq	Mean	StDev	Min	Med	Max
K	386	0.9883	0.00439	0.9800	0.9890	0.996

Table 4.2 provides descriptive statistics of the values of the population correlation coefficients ρ_{xy} , ρ_{xz} , ρ_{yz} and ρ_{xy}^* which are used in the simulation study. Recall that as pointed out by Singh and Arnab (2011), only the value of the weighted correlation coefficient ρ_{xy}^* is important where comparing estimators with unequal probability sampling schemes. Out of 386 cases, where the two-step estimator performs better, the minimum value of ρ_{xy}^* is found to be 0.52965, maximum value to be 0.70000 and median value to be 0.53800. The average value of ρ_{xy}^* is 0.52965 with a standard deviation of 0.08504. Thus the proposed two-step estimator is unlikely to perform better than the linear regression estimator if the value of the correlation coefficient is more than 0.70 in a population similar to the one being considered here.

Table 4.2. Choice of values of ρ_{xy} , ρ_{xz} , ρ_{yz} and ρ_{xy}^* .

Variables	freq	Mean	StDev	Min	Med	Max
ρ_{xy}	386	0.60003	0.04913	0.54800	0.58200	0.72800
ρ_{xz}	386	0.56397	0.04987	0.51100	0.54600	0.69400
ρ_{yz}	386	0.71830	0.04171	0.67300	0.70400	0.82300
ρ_{xy}^*	386	0.52965	0.08504	0.33100	0.53800	0.70000

Pictorial presentations of such populations are given in Figure 4.2. It shows that if a dataset has a non-linear pattern, then the proposed two-step calibration method estimator is found to perform better than the linear regression estimator. It reconfirms the statement of Singh (2012) that in the presence of linear trend there is no estimator which is better than the linear regression estimator. Table 4.3 provides descriptive statistics of the values of the percent relative efficiencies RE(1,5), RE(2,5), RE(3,5) and RE(4,5). Out of 386 cases, where the two-step estimator attains the minimum value of RE(4,5) to be 103.50%, maximum value to be 108.80% and with a median value to be 104.50%. The average value of RE(4,5) is 104.77 with a standard deviation of 1.32.

Table 4.3. Values of RE(1,5), RE(2,5), RE(3,5) and RE(4,5).

Variables	freq	Mean	StDev	Min	Med	Max
RE(1,5)	386	1355.7	71.2	1163.9	1353.3	1531.8
RE(2,5)	386	486.40	122.12	326.90	447.05	962.40
RE(3,5)	386	111.87	2.43	107.30	111.80	118.60
RE(4,5)	386	104.77	1.32	103.00	104.50	108.80

Figure 4.5 gives histograms of the percent relative efficiency values for the proposed two-step estimator with respect to the ratio (or RE(1,5)), GREG (or RE(2,5)), Wu and Sitter (2001) (or RE(3,5)) and the linear regression estimator (or RE(4,5)).

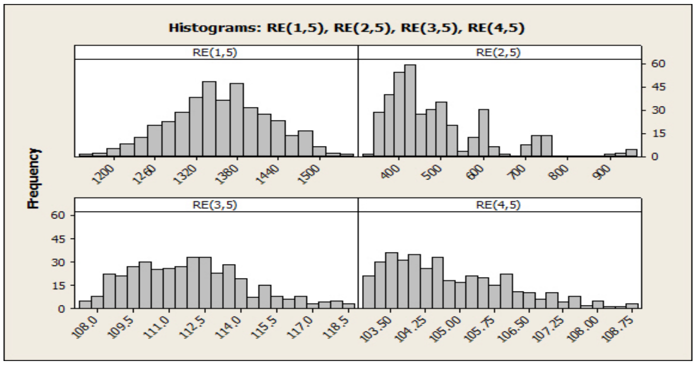

Fig. 4.5. Histograms of percent relative efficiency values.

Figure 4.6 is devoted to study the percent relative bias (RB) values of the five estimators considered in the simulation study.

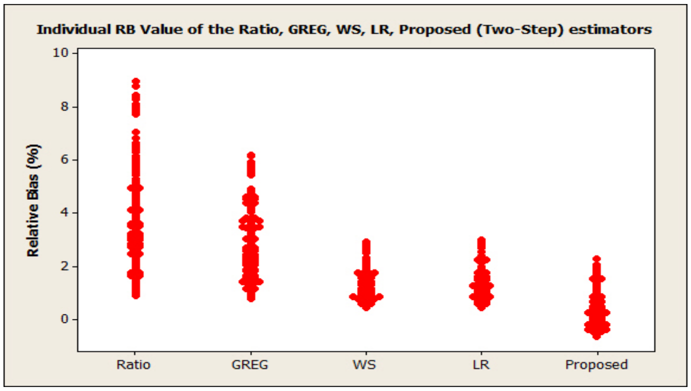

Fig. 4.6. RB values of the five estimators considered.

Table 4.4 shows descriptive statistics values of the percentage relative bias (RB) in the five estimators considered in the simulation study.

Table 4.4. Percent Relative Bias (RB) values.

Variable	*freq*	Mean	StDev	Min	Med	Max
Ratio	386	3.5160	1.5506	0.8860	3.1525	8.9590
GREG	386	2.5744	1.0913	0.7820	2.3030	6.1860
WS	386	1.2241	0.5288	0.4400	1.0845	2.9240
LR	386	1.2483	0.5413	0.4520	1.1015	2.9880
Proposed	386	0.2840	0.5476	-0.6200	0.1845	2.2860

The percent relative bias (RB) value in the ratio estimator varies from 0.8860% to 8.9590% with a median value of 3.1525%; in the GREG estimator varies from 0.7820% to 6.1860% with a median value of 2.3030%; in the Wu and Sitter (2001) estimator varies from 0.4400% to 2.9240% with a median value of 1.0845%; in the linear regression estimator varies from 0.4520% to 2.9880% with a median value of 1.1015%; and in the proposed two-step estimator varies from -0.6200% to 2.2860% with a median value of 0.1845%. Results in Table 4.4 show that the percent relative bias remains negligible in all the estimators considered.

REFERENCES

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

Estevao, V.M. and Sarndal, C.E. (2003). A new perspective on calibration estimators. *Joint Statistical Meeting-Section on Survey Research Methods, 1346-1356.*

Farrell, P.J. and Singh, S. (2002). Penalized chi square distance function in survey sampling. *Proc. of the Joint. Statist. Meet, New York (Available on CD).*

Farrell, P.J. and Singh, S. (2005). Model-assisted higher order calibration of estimators of variance. *Aust. & New Zealand J. Statist.* 47(3), 375-383.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample survey methods and theory*. New York, John Wiley and Sons, 456-464.

Horvitz, D.G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc*., 47, 663-685.

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

Kott, P.S. (2003). An overview of calibration weighting. *Joint Stat. Meet. Sec. of Survey Meth.,* 2241-2252.

Midzuno, H. (1952). On the sampling system with probability proportional to sum of sizes. *Ann. Inst. Statist. Math*., 3, 99--107.

Montanari, G.E. and Ranalli, G. (2005). Nonparametric model calibration estimation in survey sampling. *Jour. Amer. Statist. Assoc.* 100(472), 1429-1442.

Rueda, M., Martinez, S., Martinez, H. and Arcos, A. (2006). Mean estimation with calibration techniques in presence of missing data. *Comp. Stat. and Data Anal.,* 3263-77.

Rueda, M., Martinez, S., Martinez, H. and Arcos, A. (2007). Estimation of the distribution function with calibration methods. *J. Stat. Plann.Inf.,* 137, 435-448.

Singh, S. (2003). *Advanced sampling theory with applications: How Michael 'Selected' Amy*. Kluwer Academic Publisher, pp. 1-1247.

Singh, S. (2004). Golden and Silver Jubilee Year–2003 of the linear regression estimators. *Proc. Of the Joint Stat. Meeting, Toronto, 4382-4389.*

Singh, S. (2006). Survey statisticians celebrate Golden Jubilee Year-2003 of the linear regression estimator. *Metrika*, 1-18.

Singh, S. (2011). A dual problem of calibration of design weights. *Statistics,* 47, 566-574.

Singh, S. (2012).Calibration of design weights using a displacement function. *Metrika*, 85-107.

Singh, S. and Arnab, R. (2011). On calibration of design weights. *Metron,* 185-205.

Stearns, M. and Singh, S. (2008). On the estimation of the general parameter. *Comp. Stat. and Data Anal.*, 52, 4253-4271.

Wu, C. and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *J. Amer. Statist. Assoc.,* 96, 185-193.