Pak. J. Statist.
2016 Vol. 32(3), 443-456

COSMETIC CALIBRATION WITH WEIGHTED OBSERVATIONS

Naila Alam[1] and **Muhammad Hanif**[2]
[1] Kinnaird College for Women, Lahore, Pakistan.
Email: naila.alam1969@gmail.com
[2] National College of Business Administration and Economics
Lahore, Pakistan. Email: drmianhanif@gmail.com

ABSTRACT

In this paper, different cosmetic estimators are obtained, by using different individual weights for auxiliary-variable's observations that are still unexplored, and a simulation study is conducted to assess their performance. The results show that these weights are helpful in reducing heteroskedasticity, resulting in minimum SSE and therefore, may be considered as an alternative of linear transformations. Furthermore, it is shown that by using these weights in linear calibration approach, generated calibration estimators are equivalent to those produced by the instrumental approach.

KEY WORDS

Cosmetic Calibration, Calibration approach, Instrumental Variable.

1. INTRODUCTION

Cosmetic estimation was a result of the attempts to reconcile design based and model based approaches of estimation. Brewer (1979) initiated the idea to use a linear prediction estimator format, and asymptotic design un-biasness to combine model based and design based estimation. Särndal and Wright (1984) extended the idea and used the word "cosmetic". It enables a model-based as well as a design-based interpretation of estimators, making it "attractive". The well-known calibration estimator proposed by Deville and Särndal (1992) is also a cosmetic estimator.

Suppose the objective is to estimate the total of the study variable 'y' in a finite population $U=\{y_1,.....y_N\}$, consisting of N elements. A probability sample s is drawn from the population with a sampling design $p(s)$. y_k is the value of k_{th} observation of the study variable and is known for all $k\in s$. The probability sampling design $p(s)$ generates a known inclusion probability $\pi_k>0$ for each element k in the sample of size n, and the corresponding sampling design weight is $d_k=1/\pi_k$. A vector of p auxiliary variables $\mathrm{x}_k'=\left[x_{k1}.....x_{kj}\ ...\ x_{kp}\right]$ for the element $k\in s$ is associated with y_k, where $j=1,2,.....p$ and the population total to be estimated is $Y=\sum_U y_k$. Also, the vector of total(s) of p auxiliary variables $\mathrm{t_x}=\sum_U x_k$ is known ($\sum_U$ denotes sum of all $k\in U$ and

$\sum_s$ denotes sum on all $k \in s$), and vector of Horvitz Thompson estimators of population total(s) for auxiliary variables is. $\hat{t}_{x\pi} = \sum_s d_k x_k$

The Deville and Särndal (1992) gave the name "calibration estimator" to a weighted estimator of the unknown population total of the form $\hat{t}_{yw} = \sum_s w_k y_k$, where $\{w_k : k \in s\}$ are weights, having minimum distance from the survey design weights $d_k = 1/\pi_k$ and are chosen to satisfy the following calibration constraints:

$$\sum_s w_k x_k = \sum_U x_k \tag{1.1}$$

To minimize the distance between w_k (calibration weights) and d_k (sampling design weights), any distance measure $G_k(w,d)$, which satisfies some basic conditions can be minimized under the constraints given in (1.1).

If $\lambda' = (\lambda_1, ...\lambda_j, ...\lambda_p)$ is the vector of Lagrange multipliers, Lagrangian equation can be written in the following form:

$$\sum_s d_k G(w_k, d_k) - \lambda' \left(\sum_s w_k x_k - \sum_U x_k \right) = 0$$

Hence,

$$\phi_s(\lambda) = \left(t_x - \hat{t}_{x\pi} \right),$$

where, the value of λ can be obtained by using Newton's method of optimization i.e.
$\lambda_{v+1} = \lambda_v + \left\{ \phi'_s(\lambda_v) \right\}^{-1} \left\{ t_x - \hat{t}_{x\pi} - \phi_s(\lambda_v) \right\}$ generated,

$$\lambda = \left(\sum_s d_k q_k x_k x'_k \right)^{-1} \left(t_x - \hat{t}_{x\pi} \right) \tag{1.2}$$

The obtained calibration weights are:

$$w_k = d_k F(q_k x'_k \lambda) = d_k F(u).$$

Different distance functions result in different calibration weights. Deville and Särndal (1992) considered various distance function, and obtained resulting calibration estimators. The chi square distance function $G_k(w_k, d_k) = \dfrac{(w_k - d_k)^2}{2 d_k q_k}$ generates a class of calibration weights that are a linear function of design weights and available auxiliary information:

$$w_k = d_k (1 + q_k x'_k \lambda) \tag{1.3}$$

The obtained calibration estimator, is a generalized linear regression estimator (GREG) proposed by Cassel, Sarndal and Wretman (1976) and therefore, is a cosmetic estimator as it can be interpreted both as model-based and design-based estimator:

$$\hat{t}_{GREG} = \hat{t}_{y\pi} + \mathrm{b'_w}\left(\mathrm{t_x} - \hat{\mathrm{t}}_{\mathrm{x}\pi}\right) \tag{1.4}$$

The distance minimization approach in calibration technique, provides approximately identical estimators for different distance functions. Seeing this, for studying the properties of calibration estimators, Estevao and Särndal (2000) proposed a wider class of functional-form calibration estimators by dropping the condition of distance minimization.

They define a vector $z'_k = \left(z_{1k}, z_{2k}, \ldots\ldots z_{Jk}\right)$ for every $k \in s$, such that

a) $\dim\left(z_k\right) = J = \dim\left(\mathrm{x_k}\right)$

b) and the $J \times J$ matrix $\sum_s q_k \mathrm{z}_k \mathrm{x'}_k$ is nonsingular.

$\mathrm{z'_k} = \left(z_{1k}, z_{2k}, \ldots\ldots z_{pk}\right)$ is an instrumental variable vector whose components are the functions of x_k that is $\mathrm{z'_k} = \left(f(x_{1k}) , f(x_{2k}) , \ldots\ldots, f(x_{pk}) \right)$. They consider

$$\mathrm{z'_k} = \left(x_{1k}^{\,m-1}, x_{2k}^{\,m-1}, \ldots\ldots, x_{pk}^{\,m-1}\right) \text{for } m \geq 1,\ j = 1,2,\ldots.p \text{ and. } x_{jk} > 0$$

Thus, the functional form of calibration weights $w_{k,CALF} = d_k + d_k q_k \lambda'_s z_k$ are obtained by satisfying the calibration constraints (1.1), such that value of λ is determined implicitly by these constraints. The functional-form calibration weights can be written as:

$$w_{k,CALF} = d_k + \hat{R}'_k\left(t_x - \hat{t}_{x\pi}\right), \tag{1.5}$$

where $\hat{\mathrm{R}}_\mathrm{k} = \left(\sum_s d_k q_k \mathrm{z_k x'_k} \right)^{-1} d_k q_k \mathrm{z_k}$ and resulting functional-form calibration estimator for population total is:

$$\hat{t}_{CALF} = \sum_s w_{CALF} y_k = \sum_s d_k y_k + b'_{w,CALF}\left(t_x - \hat{t}_{x\pi}\right)$$

where $\mathrm{b_{w,CALF}} = \left(\sum_s d_k q_k \mathrm{z_k x'_k} \right)^{-1} d_k q_k \mathrm{z_k} y_k$

It defines a family of calibration estimator that included the family of GREG estimators such that:

$$\hat{t}_{GREG} \subseteq \hat{t}_{CALF} \subseteq \hat{t}_{yw}.$$

The functional-form calibration weights allow to create a variety of calibration weights for different values of q_k and z_k. For $m = 2$, we have $z_k = x_k$ therefore, the weights w_k are a special case of weights $w_{k,CALF}$.

q_k (for $k=1,2,....,n$) are relative parameters to be estimated and can be chosen to improve calibration weights and relative statistical efficiency. Estevo & Särndal (2000) have shown that arbitrary positive random values of q_k do not harm the asymptotic unbiasness of the estimator. Since $N^{-1}\left(t_x - \hat{t}_{x\pi}\right)' R_k$ is $O\left(n^{-1/2}\right)$, therefore, the calibration weight system (1.3) remains asymptotically design-unbiased for random values of q_k. These are not regression weights but weights of individual observations for each of the kth term that can be attached to values of auxiliary variable.

In determining calibration weights, many times the question arises about the most favorable value for q_k. The most popular value for $q_k = 1$. This gives $b_{w1} = \dfrac{\sum_s d_k x_k y_k}{\sum_s d_k x^2{}_k}$. The other values of q_k can also be used, as Deville and Särndal (1992) supposed $q_k = 1/x_k$ and obtained a ratio estimator.

Brewer (1999) suggested to use $q_k = 1 - \pi_k$ to avoid the weights which are less than 1 or negative. For same purpose, Bankier (2002) recommended to set $q_k = \dfrac{1-\pi_k}{\pi_k}$ or $q_k = d_k - 1$. The previously considered values of q_k are mostly a function of survey design weights. We consider some other values of q_k that are still unexplored, and use them to show equivalence of instrumental calibration and linear calibration approach.

Estevo and Särndal (2000) gave the detailed results of a Monte Carlo simulation in which they consider various values of $m \geq 2$ in a functional-form calibration approach. They show that estimators having values $m > 2$ give more efficient results. Kim (2010) has shown asymptotic equivalence between functional-form calibration and instrumental variable calibration estimators. Also, Park and Kim (2014) proposed an optimal instrumental calibration estimator and extended it under two phase. We demonstrated equivalence of instrumental variable and classical linear calibration approach by using different values of q_k and corresponding values of m.

We use values $1 \leq m \leq 2$ and conduct a simulation study to check the performance of the resulting estimators. In section 2, different cosmetic estimators have been obtained by using different values of q_k for $k=1,2,.....,n$ and in section 3, the performance of these estimators has been evaluated by a simulation study.

2. COSMETIC CALIBRATION BY USING OBSERVATION'S INDIVIDUAL WEIGHTS

We obtained cosmetic estimators by considering different sets of weights q_k, for $k=1,2,.....,n$. We do not consider q_k to be a function of the survey weights but that of the available auxiliary information.

For simplicity, for every $k \in s$, we consider only one auxiliary variable that is linearly related to the study variable with a known population total. In this case vector x_k has only one component and the product of $x_k x_k'$ can be written as x_k^2.

The calibration weights when $q_k = 1$, can be written as:

$$w_k = d_k\left(1+q_k x_k \lambda_s\right),$$

where,

$$\lambda = \frac{1}{\sum_s d_k q_k x_k^2}\left(t_x - \hat{t}_{x\pi}\right).$$

We obtained,

$$w_k = d_k\left(1+\frac{q_k\ x_k}{\sum_s d_k q_k x_k^2}\left(t_x - \hat{t}_{x\pi}\right)\right), \tag{2.1}$$

And resulting calibration estimator is a GREG estimator that is

$\hat{t}_{GREG} = \hat{t}_{y\pi} + b'_w\left(t_x - \hat{t}_{x\pi}\right)$, where

$$b_w = \frac{\sum\limits_s d_k q_k x_k y_k}{\sum\limits_s d_k q_k x^2{}_k}. \tag{2.2}$$

We construct nine cosmetic estimators by attaching the weights q_{kl} (for $l=1,.....,9$) to the auxiliary-variable's observations, as follows:

$$q_{kl}\ :1, \frac{1}{\left|x_k^{1/4}\right|}, \frac{1}{\left|x_k^{1/3}\right|}, \frac{1}{\left|x_k^{2/5}\right|}, \frac{1}{\left|x_k^{1/2}\right|}, \frac{1}{\left|x_k^{3/5}\right|}, \frac{1}{\left|x_k^{3/4}\right|}, \frac{1}{\left|x_k^{9/10}\right|}, \frac{1}{|x|} \tag{2.3}$$

We start with the classical choice of value of q_k that is 1. The resulting estimator is the classical linear calibration or generalized regression (GREG) estimator with $b_{w1} = \frac{\sum\limits_s d_k x_k y_k}{\sum\limits_s d_k x^2{}_k}$. We note that using these values of q_{kl}, result in calibration estimators having different regression coefficients b_{wl}, that is:

$$\hat{t}_{ywl} = \hat{t}_{y\pi} + b'_{wl}\left(t_x - \hat{t}_{x\pi}\right) \text{ for } l=1,.....,9$$

Let, b_w be an instrumental variable regression-coefficient estimator, of the form:

$$b_w = \left(h'_s x_s\right)^{-1} h_s y \tag{2.4}$$

where, $h'_s = z_s' d_k$ and z_k is an instrumental variable vector such that

$$z'_k = \left(x_{1k}^{\,m-1}, x_{2k}^{\,m-1}, ..., x_{pk}^{\,m-1}\right) \text{ for } x_{jk} > 0 .$$

Hence, we consider nine values of m within the interval $1 \le m \le 2$ for one auxiliary variable $z_k = x_k^{m-1}$. The b_{wm} for $m = 1, 11/10, 5/4, 7/5, 3/2, 8/5, 5/3, 7/4$ and 2, in (2.4) yield the same regression weights that are produced by using q_{kl} weights (for $l = 1, ..., 9$) in (2.2) and result in equivalent estimators of population total but the estimators obtained by using q_k weights in linear calibration are of wider class than instrumental-variable estimators. We noted that many choices of instrumental-variable vectors restrict $x_{jk} > 0$, but in case of attaching q_k weights with auxiliary-variable, no condition is imposed on the x_k values.

Table 1 shows the q_k weights and the corresponding values of m, that generate the same regression coefficient estimators.

Result 1:

The calibration estimators $\hat{t}_{ywl} = \hat{t}_{y\pi} + b'_{wl}\left(t_x - \hat{t}_{x\pi}\right)$ for $l = 1, ..., 9$ are asymptotically design unbiased**.**

Proof:

To prove asymptotically un-biased, we follow the results given by Fuller and Isaki (1981):

$$\frac{1}{N}\left(t_x - \hat{t}_{x\pi}\right) = O_p\left(\frac{1}{\sqrt{n}}\right)$$

Also for a sequence of finite populations and sampling design when $N \to \infty$ with corresponding sample size $n \to \infty$ we assume that,

$$\frac{1}{N}\left(t_x - \hat{t}_{x\pi}\right) \to 0 \text{, Provided } \lim_{N\to\infty} \frac{t_x}{N} \text{ exists.}$$

where, $t_x = \sum_U x_k$ and $\hat{t}_{x\pi} = \sum_s d_k x_k$.

This implies,

$$\hat{t}_{ywl} = \hat{t}_{y\pi} + b'_{wl}\left(t_x - \hat{t}_{x\pi}\right) \to \hat{t}_{y\pi} \text{ for } n \to \infty$$

Therefore, the constructed estimators are asymptotically design unbiased, for arbitrary values of q_k. The similar results have shown by Estevo & Särndal (2000) that

different values of q_k do not harm the asymptotic un-biasness of the calibration estimator.

Table 1
Values of q_k, Corresponding m Values and Resulting Regression Coefficient

q_k	m	b_w
1	2	$\frac{\sum_s d_k x_k y_k}{\sum_s d_k x^2_k}$
$1/x^{1/4}$	7/4	$\frac{\sum_s d_k x_k^{3/4} y_k}{\sum_s d_k x_k^{7/4}}$
$1/x^{1/3}$	5/3	$\frac{\sum_s d_k x_k^{2/3} y_k}{\sum_s d_k x_k^{5/3}}$
$1/x^{2/5}$	8/5	$\frac{\sum_s d_k x_k^{3/5} y_k}{\sum_s d_k x_k^{8/5}}$
$1/x^{1/2}$	3/2	$\frac{\sum_s d_k y_k \sqrt{x_k}}{\sum_s d_k x_k^{3/2}}$
$1/x^{3/5}$	7/5	$\frac{\sum_s d_k x_k^{2/5} y_k}{\sum_s d_k x_k^{7/5}}$
$1/x^{3/4}$	5/4	$\frac{\sum_s d_k x_k^{1/4} y_k}{\sum_s d_k x_k^{5/4}}$
$1/x^{9/10}$	11/10	$\frac{\sum_s d_k x_k^{9/10} y_k}{\sum_s d_k x_k^{11/10}}$
$1/x$	1	$\frac{\sum_s d_k y_k}{\sum_s d_k x_k}$

3. SIMULATION STUDY

To check the performance of the constructed estimators, we conduct a Monte Carlo simulation study. The finite populations consisting of pair values (y_k, x_k) are generated such that there is linear relationship between x_k and y_k. For simplicity, we take equal design weights for all the n observations, that is $d_k = \frac{N}{n}$.

We carried out two simulations to check the performance of the constructed nine cosmetic estimators for small and large sample sizes. For this purpose, two sample sizes $n = 50$, and $n = 500$ are used.

3.1 Simulation 1

In simulation (1), 10 finite populations $i = 1, 2, ..., 10$ of $N = 1000$ of pair values (y_k, x_k) k $= 1, 2, ...1000$ are generated for each set of q_{kl} described in (2.2) for $l = 1, 2, ..., 9$ where, y_k are normally distributed with mean= 10, and standard deviation = 2.6. From each population 100 samples of $n = 50$ are selected for each case.

The 9 cases of finite populations of $N = 1000$ are generated under the following specifications.

(i) The generated finite populations consist of data set (y_k, x_k) that show high correlation between study and auxiliary variable for the cases when

Case I: the associated x variable values are approximately in the same band as of y values, that is $y \approx x$.

Case II: the x values are generated such that $x_k \approx a_k + y_k^2$ where $a_k \le y_k$.

Case III: the x values are generated corresponding to the same y values such that $x_k \approx a_k + y_k^3$ where a is any number, such that , $a_k \le y_k$ .Also,

(ii) The generated finite populations show a Medium correlation between study and auxiliary variable for the case (a), (b) and (c) and

(iii) The generated populations prove a low correlation between y and x for the above mentioned three cases (a), (b) and (c).

3.2 Simulation 2

In simulation (2), again, 10 finite populations consisting of 1000 units are generated for each of the nine cases but from these populations, 100 sample of size 500 are selected by using SRS with replacement. The generated populations consist of data set (y_k, x_k) where y_k are generated such that they are normally distributed with mean = 25 and standard deviation = 4.2. The corresponding x_k values are generated under the nine specifications against the same y_k values such that,

(i) The generated data set (y_k, x_k) show high correlation

(ii) The generated x_k values have medium correlation with y_k, and

(iii) The generated data set show low correlation.

The data set (y_k, x_k) described in (i) is generated under the three specifications such that the x_k values are

Case I: Approximately in the same band as of y_k values

Case II: Fall within the interval (0,1000)

Case III: Fall within the interval (0,40000)

Similarly the data sets in (ii) and (iii) are generated under the cases I,II and III.

The results are based on $10 \times 100 = 1000$ samples for each case. We computed the Mean square error (RSE), R^2 and sum of square of the error (SSE) by using the following formulae, where $n_1 = 50$ and $n_2 = 500$

$$Avg\,MSE_{n1} = \frac{1}{10}\sum_{i=1}^{10} \bar{\hat{\sigma}}_{in1}^2 \quad i = 1, 2, ..., 10$$

$$Avg\,MSE_{n2} = \frac{1}{10}\sum_{i=1}^{10} \bar{\hat{\sigma}}_{in2}^2$$

where $\bar{\hat{\sigma}}_i = \frac{1}{100}\sum_{j=1}^{100} \hat{\sigma}_{ij}$ is calculated for each sample by using the formula given below,

$$MSE_{n1} = \hat{\sigma}_{n1} = \sqrt{\frac{\sum_{k=1}^{50} e_k^2}{n_1 - 2}}$$

and for $k = 1, 2, ..., 50$

$$MSE_{n2} = \hat{\sigma}_{n2}^2 = \sqrt{\frac{\sum_{k=1}^{500} e_k^2}{n_2 - 2}}$$

for $e_k^2 = \left(y_k - \hat{y}_j\right)^2$

Similarly,

$$Avg\,SSE_{n1} = \frac{1}{10 \times 100}\sum_{i=1}^{10}\sum_{j=1}^{100}\sum_{k=1}^{50}\left(y_k - \hat{y}\right)^2,$$

and

$$Avg\,SSE_{n2} = \frac{1}{10\times 100}\sum_{i=1}^{10}\sum_{j=1}^{100}\sum_{k=1}^{500}\left(y_k - \hat{y}\right)^2 .$$

$$R_{n1}^2 = \frac{1}{10}\sum_{i=1}^{10}\bar{\hat{R}}_{n1}^2$$

where,

$$\bar{\hat{R}}_{n1}^2 = \frac{1}{100}\sum_{j=1}^{100}\frac{\sum_{k=1}^{50}\left(\hat{y}_k - \bar{y}\right)^2}{\sum_{k=1}^{50}\left(y_k - \bar{y}\right)^2}$$

Similarly,

$$R_{n2}^2 = \frac{1}{10}\sum_{i=1}^{10}\bar{\hat{R}}_{n2}^2 \text{ and } \rho = \frac{\sum_{k=1}^{1000}\left(y_k - \bar{y}_i\right)\left(x_k - \bar{x}_i\right)}{\sqrt{\sum_{k=1}^{1000}\left(y_k - \bar{y}_i\right)^2\left(x_k - \bar{x}_i\right)^2}}.$$

The value of correlation coefficient is used to check the strength of linear relationship between generated x_k and y_k, and is obtained by averaging the correlation coefficient of 10 populations for each of 18 cases and denoted by $\bar{\rho}$.

where $\bar{\rho} = \frac{1}{10}\sum_{i=1}^{10}\rho_i$.

The results are summarized in Table 1 and 2. The results of GREG or classical linear calibration estimator (for $q_k = 1$) are shaded in pink and the blue shaded row shows the results that are more efficient.

Table 2
Results based on 10 Populations for each of Nine Cases and within each Population 100 Sample of Size 50

$y_k \sim N(10, 2.6)$

$n = 50$										
	ρ	High			Medium			Low		
	q_k	AVG MSE	R^2	AVG SSE	AVG MSE	R^2	SS	AVG MSE	R^2	AVG SSE
Case I	1	0.8191	0.8712	32.2215	1.1971	0.7819	68.7991	1.4141	0.5031	95.9380
	$1/x^{1/4}$	0.8197	0.8704	34.3210	1.1983	0.7815	68.9351	1.4183	0.5056	96.4562
	$1/x^{1/3}$	0.8256	0.8698	36.0012	1.1992	0.7809	68.8322	1.4187	0.5089	97.0432
	$1/x^{2/5}$	0.8321	0.8672	37.3455	1.2012	0.7804	68.3515	1.4199	0.4987	97.8791
	$1/x^{1/2}$	0.8387	0.8578	38.3241	1.2041	0.7793	69.6164	1.4241	0.4934	98.1201
	$1/x^{3/5}$	0.8437	0.8497	39.6723	1.2124	0.7760	69.9016	1.4243	0.4901	98.6871
	$1/x^{3/4}$	0.9052	0.8452	41.2378	1.2191	0.7735	69.9431	1.4244	0.4812	99.4521
	$1/x^{9/10}$	0.9366	0.8317	42.1034	1.2231	0.7723	70.2130	1.4456	0.4894	100.2469
	$1/x$	2.26	-	250.199	2.26	-	250.1999	2.26	-	250.1999
Case II	1	0.5258	0.8992	13.7201	1.607	0.5047	123.769	1.9530	0.2680	183.1549
	$1/x^{1/4}$	0.4216	0.9352	8.5312	1.5987	0.5353	120.032	1.8743	0.2778	176.5641
	$1/x^{1/3}$	0.4045	0.9403	7.8527	1.5301	0.5994	116.653	1.7643	0.3021	168.4320
	$1/x^{2/5}$	0.3781	0.9479	6.8381	1.4788	0.6212	108.671	1.6854	0.3312	160.5552
	$1/x^{1/2}$	0.3669	0.9509	6.4616	1.3921	0.6321	102.762	1.5162	0.3988	153.1957
	$1/x^{3/5}$	0.3756	0.9486	6.7725	1.5551	0.5332	114.003	1.5278	0.3897	159.4433
	$1/x^{3/4}$	0.4299	0.9326	8.8705	1.5991	0.5039	119.889	1.5304	0.3854	163.8644
	$1/x^{9/10}$	0.5267	0.8989	13.3134	1.6140	0.4998	128.370	1.5473	0.3745	168.378
	$1/x$	1.639	-	150.31	1.639	-	250.199	1.639	-	250.1999
Case III	1	0.76581	0.8875	28.1496	1.9030	0.3348	173.8641	1.8811	0.3209	169.9013
	$1/x^{1/4}$	0.6668	0.9012	22.4478	1.7761	0.4373	166.8921	1.8407	0.3400	165.5521
	$1/x^{1/3}$	0.4224	0.9443	16.7765	1.7392	0.4421	162.2310	1.7561	0.3789	161.0031
	$1/x^{2/5}$	0.3980	0.9567	12.8901	1.7282	0.4673	160.0091	1.7220	0.3941	159.5412
	$1/x^{1/2}$	0.3649	0.9745	7.3341	1.6984	0.4858	157.3654	1.6970	0.4163	155.3210
	$1/x^{3/5}$	0.3400	0.9798	5.9905	1.6700	0.5021	155.8734	1.6700	0.4566	150.3001
	$1/x^{3/4}$	0.3223	0.9801	4.9861	1.6321	0.5211	153.8451	1.6552	0.4877	143.6214
	$1/x^{9/10}$	0.4179	0.9567	8.9384	1.6212	0.5312	153.3200	1.6502	0.4888	140.8214
	$1/x$	1.943	-	201.739	1.943	-	201.739	1.943	-	201.739

Table 3
Results based on 10 Populations for each of Nine Cases and within each Population 100 Sample of Size 500

$y_k \sim N(25, 4.2)$

$N = 500$										
	ρ	High			Medium			Low		
	q_k	AVG MSE	R^2	AVG SS	AVG MSE	R^2	AVG SS	AVG MSE	R^2	AVG SS
Case I	1	0.9588	0.9416	457.7826	2.7040	0.5351	3642.416	3.5840	0.1835	6396.960
	$1/x^{1/4}$	0.9602	0.9401	464.8709	2.7051	0.5349	3644.678	3.5842	0.1823	6401.004
	$1/x^{1/3}$	0.9687	0.9356	469.8732	2.7058	0.5347	3647.880	3.5846	0.1817	6407.670
	$1/x^{2/5}$	0.9731	0.9377	472.5008	2.7065	0.5344	3648.998	3.5848	0.1811	6414.445
	$1/x^{1/2}$	0.9771	0.9393	475.4981	2.7071	0.5340	3650.544	3.5900	0.1807	6418.661
	$1/x^{3/5}$	0.9851	0.9356	487.0021	2.7132	0.5351	3661.456	3.5965	0.1778	6433.231
	$1/x^{3/4}$	0.9921	0.9344	501.6678	2.7201	0.5357	3669.126	3.5986	0.1766	6459.901
	$1/x^{9/10}$	1.0201	0.9339	517.9762	2.7261	0.5260	3676.569	3.6072	0.1731	6478.154
	$1/x$	3.962	-	7834.040	3.9621	-	7834.042	3.9621	-	7834.042
Case II	1	0.9338	0.9161	434.2426	2.0571	0.5931	2083.762	2.8190	0.2304	3958.513
	$1/x^{1/4}$	0.9013	0.9219	404.5169	2.0503	0.5950	2081.234	2.7994	0.2325	3890.563
	$1/x^{1/3}$	0.9021	0.9220	401.967	2.4981	0.5981	2080.862	2.7982	0.2394	3721.760
	$1/x^{2/5}$	0.8990	0.9227	398.651	2.0447	0.5982	2080.993	2.7951	0.2467	3487.001
	$1/x^{1/2}$	0.8914	0.9236	395.6902	2.0443	0.5983	2079.757	2.7911	0.2712	3278.562
	$1/x^{3/5}$	0.8974	0.9221	403.7210	2.0447	0.5950	2086.321	2.7989	0.2666	3376.023
	$1/x^{3/4}$	0.9083	0.9206	410.8263	2.0449	0.5948	2097.738	2.7998	0.2489	3489.003
	$1/x^{9/10}$	0.9231	0.9187	423.764	2.052	0.5917	2099.652	2.8231	0.2254	3431.890
	$1/x$	3.221	-	5177.192	3.221	-	5177.192	3.221	-	5177.192
Case III	1	1.1467	0.9123	653.6372	2.2582	0.5084	2538.881	3.071	0.2303	4697.029
	$1/x^{1/4}$	1.0021	0.9232	556.7432	2.2341	0.5567	2499.453	2.9904	0.2514	4487.046
	$1/x^{1/3}$	0.9445	0.9404	444.2560	2.2121	0.6007	2436.906	2.9767	0.2815	4384.632
	$1/x^{2/5}$	0.9134	0.9499	397.6071	2.2001	0.6066	2479.765	2.9034	0.2904	4199.324
	$1/x^{1/2}$	0.8556	0.9511	364.5977	2.1931	0.6075	2395.140	2.8930	0.3169	4285.321
	$1/x^{3/5}$	0.8348	0.9535	358.8901	2.1889	0.6082	2392.541	2.8122	0.3342	4032.947
	$1/x^{3/4}$	0.8424	0.9526	353.3962	2.1864	0.6099	2380.471	2.7993	0.3605	3902.881
	$1/x^{9/10}$	0.9021	0.9456	405.2753	2.1934	0.6074	2390.361	2.7451	0.3851	3752.319
	$1/x$	3.497	-	6102.748	3.497	-	6102.748	3.4971	-	6102.0748

4. DISCUSSION

We used different sets of weights, q_k, which result in cosmetic estimators having different regression coefficients, and carried out two simulations to illustrate and study the performance of these estimators.

The constructed cosmetic estimators, $\hat{t}_{ywl}$ for $l = 1,.....,9$ can also be obtained by taking $m = 1, 11/10, 5/4, 7/5, 3/2, 8/5, 5/3, 7/4$ and 2 in (2.4).

The results of table 2 show that $q_k = 1$ or $m = 2$ is optimal only when the pair values (y_k, x_k) have approximately the same amount of heteroscedasticity; when that is not the case, other values of weights, q_k, give more efficient results.

The results are same for small and large sample sizes. Also, results are found to be more pronounced when the correlation between study and auxiliary variable is high or medium.

$q_k = \frac{1}{x_k^{1/2}}$ are the best weights, when X has variance approximately as the square of the variance of y. When data is more heterogeneous, such that x_k associated with y_k follow the pattern $x_k \approx a_k + y_k^3$, then $q_k = \frac{1}{x_k^{3/4}}$ or $\frac{1}{x_k^{9/10}}$ are the optimum weights to obtain minimum SSE.

Table 3 shows similar results for large sample size.

We noted that even for a simple linear relationship between x and y, $q_k = 1$ is not the best choice always, and when heteroscedasticity in x values corresponding to y values increases, other values of q_k may be used to reduce it.

5. CONCLUSION

The above simulation studies show that the weights, q_k, in linear calibration approach can be considered as an alternative of linear transformation to get minimum SSE. Therefore, these weights may be used to smoothen or linearize the data instead of using any linear transformation. The statisticians use $q_k = 1$ to get GREG estimator but this is not the best choice always, and when heteroscedasticity increases in data, a different value of q_k is more suitable for minimum SSE, as observed in tables 2 and 3. We noted that the weights q_{kl} in linear calibration estimator produce same regression estimators that are obtained from instrumental-variable approach. Furthermore, by using q_k in a linear calibration estimator generates a wider class of estimators than instrumental-variable approach. The instrumental-variable vectors restrict x_{jk} to be positive for various choices of z_k, but in case of attaching q_k with the auxiliary-variable, no condition is imposed on x_k and it can take any value on the real line.

ACKNOWLEDGEMENT

The authors are very thankful to James Knaub (America) and the referee for their valuable comments and guidance to improve this research paper.

REFERENCES

1. Bankier, M. (2002). Regression estimators for the 2001 Canadian Census. *Paper presented at the International Conference in Recent Advances in Survey Sampling*, Carlton University, Ottawa.
2. Brewer, K.R.W. (1979). A class of robust sampling designs for large scale surveys. *J. Amer. Statist. Assoc.*, 74, 911-915.
3. Cassel, C.M, Särndal, C.E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika, 63, 615-620.
4. Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, 87, 376-382.
5. Estevao, V.M. and Särndal, C.E. (2000). A functional form approach to calibration. *J. Off. Statist.*, 16, 379-399.
6. Fuller, W.A. and Isaki C.T. (1981). Survey design under the super population model. In *Current Topics in Survey Sampling*, New York Academic Press.
7. Kim, J.K. and Park, M. (2010). Calibration Estimation in Survey Sampling. *International Statistical Review*, 78(1), 21-39.
8. Kim, J.K. and Park, S. (2014). Instrumental-Variable Calibration Estimation in Survey Sampling. *Statistica Sinica*, 24, 1001-1015.
9. Särndal, C.E. and Wright, R.L. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian J. Statist.*, 11, 146-156.