Calibrated Maximum Likelihood Design Weights in Survey Sampling

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

Abstract

In this paper, we propose a new technique to calibrate the design weights in survey sampling by the method of maximum likelihood. We show that the design weights used in the Narain (1951) and the Horvitz and Thompson (1952) estimators are in fact maximum likelihood design weights. Later, we discuss two different situations: (a) when the variance of the calibrated weights is assumed to be known; and (b) when the variance of the calibrated weights is assumed to be unknown. Under situation (a), we obtain the linear regression estimator as a special case of it, and under situation (b) we obtain a new estimator, slightly different than the linear regression estimator. The calibrated estimators available since Deville and Särndal (1992) belong to the former case (a) whereas case (b) is a new development in this area. A simulation study has been carried out to investigate the performance of the resultant estimators. At the end, an application based on a real dataset from the biosciences is given.

Key Words: Maximum likelihood function, calibrated weights, design weights, simulation study.

1. Introduction

It is a well known fact that in the theory of sampling the precision of an estimate is usually increased by the use of some auxiliary variables correlated with the variables under investigation. It was Professor W. G. Cochran who, in 1940, discovered the ratio estimator of the population mean of a study variable by making use of an appropriate auxiliary variable. Today, ratio, product, regression estimators and their several generalizations have been discussed in the literature. These estimators use information in the form of known population parameters of the auxiliary variables. Statisticians are often interested in the precision of survey estimates. Currently, the most commonly used estimator of population total or population mean is the generalized linear regression (GREG) estimator.

Professor W. G. Cochran (1909-1980)
Printed with permission

Consider a population, $\Omega = \{1,\ 2,..,i,..,N\}$, from which a probability sample s $(s \subset \Omega)$ is drawn with a given sampling design $p(.)$. The inclusion probabilities $\pi_i = P(i \in s)$ and $\pi_{ij} \in P(i \in s,\ j \in s)$ are assumed to be strictly positive and known. Let y_i be the value of the study variable, y, for the i^{th} population unit. The well-known Horvitz and Thompson (1952) estimator of the population total Y, based on a sample s, is defined as:

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

The Horvitz-Thompson (1952) estimator was independently considered by Narain (1951). For a detailed history, one may refer to “Five decades of the Horvitz-Thompson estimator and furthermore” by T.J. Rao (2004).

Together with y_i, the value of the variable of interest, y, for the i^{th} unit of the population, let there also be associated values of the auxiliary variables $\mathbf{x}_i^* = (x_{1i}, x_{2i}, ..., x_{pi})$. Thus for a selected unit $i \in s$, we observe $(y_i, \mathbf{x}_i^*)$. The population totals $\mathbf{x}_j^* = \sum_{i \in \Omega} x_{ji}$, $j = 1,2,...,p$ of the auxiliary variables are assumed to be known. Deville and Särndal (1992) considered the problem of calibrating the design weights, $d_i = 1/\pi_i$, that appear in the HT estimator by making use of known benchmarks of auxiliary information. Before 1992, there were a few papers related to calibration of design weights, for example see Bethlehem and Keller (1987) and Haung and Fuller (1978), but Deville and Särndal’s (1992) methodology was very clear and easy to understand. The GREG (Generalized regression estimator, Sarndal, Swensson and Wretman, 1992) assisted by a multivariate linear superpopulation regression model ξ for which:

$$y_i = \mathbf{x}_i \mathbf{B} + \varepsilon_i \ = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + \varepsilon_i \tag{1.2}$$

where $\mathbf{x}_i = (1, \mathbf{x}_i^*)$, $\mathbf{B} = (\beta_0, \beta_1, ..., \beta_p)^T$ is a column vector, the errors terms ε_i such as $E_\xi(\varepsilon_i) = 0$ and $V_\xi(\varepsilon_i) = \sigma_\varepsilon^2 v_i$ with v_i known can be considered as a special case of calibration by setting:

(i) Chi Square distance

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

(ii) the set of auxiliary variables $\mathbf{x}_i = (1, \mathbf{x}_i^*)$.

(iii) $q_i = 1/v_i$.

Note the use of $\mathbf{x}_i^*$ and $\mathbf{x}_i$. Then the GREG-x estimator of the total in this case takes the form

$$\hat{Y}_{\mathbf{GREG-x}} = \sum_s d_i y_i + \left(\sum_\Omega \mathbf{x}_i - \sum_s d_i \mathbf{x}_i\right)\hat{\mathrm{B}} \tag{1.4}$$

where $\hat{\mathbf{B}} = \left(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p\right)^T = \left(\sum_s \frac{\mathbf{x}_i \mathbf{x}_i^T}{v_i \pi_i}\right)^{-1} \left(\sum_s \frac{\mathbf{x}_i^T y_i}{v_i \pi_i}\right)$. Now following Wu and Sitter (2001), the estimator (1.4) can easily be written as:

$$\hat{Y}_{\mathbf{GREG-1x}} = \sum_s d_i y_i + \left(N - \sum_s d_i\right)\hat{\beta}_0 + \left(\sum_\Omega \mathrm{x}_i^* - \sum_s d_i \mathrm{x}_i^*\right)\hat{\mathbf{B}}_{\mathbf{ols}} \tag{1.5}$$

where $\hat{\beta}_0$ and $\hat{\mathbf{B}}_{\mathbf{ols}} = \left(\hat{\beta}_1, ..., \hat{\beta}_p\right)^T$ and are the ordinary least squared estimators of the intercept β_0 and the partial regression coefficients $\mathrm{B}_{\mathbf{ols}} = \left(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_p\right)^T$ respectively, and $\mathbf{x}_{\mathbf{i}}^* = (x_1, x_2, ..., x_p)$ is the same variable vector $\mathbf{x}_i$ except it lacks the column with 1s.

Following Singh (2003, 2004, 2006), Stearns and Singh (2008), and Singh and Arnab (2011), we consider another interesting subclass of Deville and Särndal (1992) with a calibration criterion as:

(i) Chi square distance
(ii) $\sum_{i \in s} w_i = \sum_{i \in s} d_i$
(iii) the set of auxiliary variables $\mathbf{x}_{\mathbf{i}}^* = (x_1, x_2, ..., x_p)$,
(iv) $q_i = 1/v_i$

The resultant calibrated GREG-dx estimator is given by:

$$\hat{Y}_{\text{GREG}-\text{dx}} = \sum_s d_i y_i + \left(\sum_\Omega \mathbf{x}_i^* - \sum_s d \mathbf{x}_i^*\right)\hat{\mathbf{B}}_{\text{ols}} \tag{1.6}$$

Singh and Arnab (2011) have studied these estimators using extensive simulation study. In sub-section 1.1, we consider a special case with one auxiliary variable.

1.1 Special Cases

Consider the estimator of the population total Y,

$$\hat{Y}_{\text{G}} = \sum_{i \in s} w_i y_i \tag{1.7}$$

that was proposed by Deville and Särndal (1992), with weights w_i chosen to be as close as possible in an average sense, to the d_i for a given measure and subject to the calibration constraint:

$$\sum_{i\in s} w_i x_i = X . \tag{1.8}$$

Minimization of chi squared (CS) distance,

$$D = \sum_{i\in s} (w_i - d_i)^2 (d_i q_i)^{-1} , \tag{1.9}$$

where q_i are suitably chosen constants such that the estimator depends upon its choice leads to a general regression type estimator (GREG) of population total, Y, given by:

$$\hat{Y}_{\text{G}} = \hat{Y}_{HT} + \hat{\beta}_{\text{ds}}\left(X - \hat{X}_{HT}\right) \tag{1.10}$$

where

$$\hat{\beta}_{\text{ds}} = \frac{\sum_{i\in s} d_i q_i x_i y_i}{\sum_{i\in s} d_i q_i x_i^2} \tag{1.11}$$

The choice $q_i = 1/x_i$, makes it into the ratio estimator due to Cochran (1940). To develop the linear regression estimator, Singh (2003) suggested making use of an additional constraint:

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

Minimization of the distance function (1.9) subject to (1.8) and (1.12) leads to a linear regression type estimator due to Hansen, Hurwitz and Madow (1953) given by:

$$\hat{Y}_{\text{LR}} = \sum_{i\in s} d_i y_i + \hat{\beta}_{\text{ols}}\left(X - \hat{X}_{HT}\right) \tag{1.13}$$

where

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

Further discussion on this topic can be had from Stearns and Singh (2008) and for the use of multi-auxiliary information refer to Singh and Arnab (2011). A related open invitation to a fresh survey methodology was advertised by Singh and Sedory (2012).

2. Motivation

The authors were motivated to develop the notion of calibrating maximum likelihood design weights by the following consideration. Note that the design weight d_i is fixed for the i^th^ unit in a given sample. The calibrated weight w_i is a function of d_i and x_i for the i^th^ unit in the sample, thus must have expected value and variance, and in particular, we assume $E(w_i \mid x_i) = d_i$ and $V(w_i \mid x_i) = \sigma_i^2$. A pictorial representation of such a situation is shown in Figure 2.1. The values of x_1, $x_2, \ldots., x_n$ assigned to the x-axis are in the random order as the sample is selected, and are **not** ranked from lowest to highest value in the sample. Note that the design weights d_i are known and fixed by the design, but the values of σ_i^2 may be known or may be unknown, thus leading to two different situations discussed in this paper. This motivated the authors to think along these lines. It is also worth mentioning that a new approach to modeling the survey weights is used by Beaumont (2008).

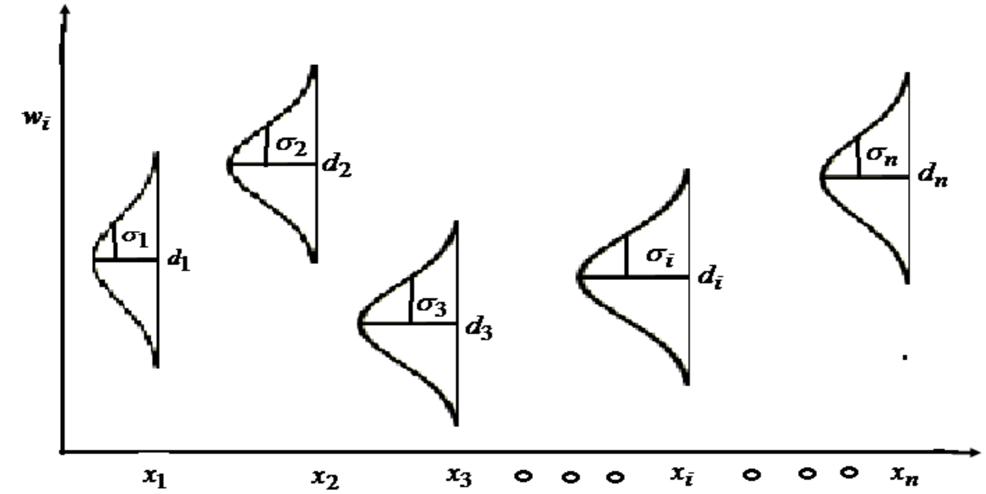

Fig. 2.1. Motivation for calibrated maximum likelihood design weights.

3. Calibrated Maximum Likelihood Design Weights

We artificially impose a probability density function on w_i for each $i \in s$; in particular we take,

$$f(w_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{1}{2}\left(\frac{w_i - d_i}{\sigma_i}\right)^2\right\}, \tag{3.1}$$

where $-\infty < w_i < +\infty$. Then we consider the corresponding synthetic likelihood function given by:

$$L = \prod_{i=1}^{n} f(w_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{1}{2}\left(\frac{w_i - d_i}{\sigma_i}\right)^2\right\} \tag{3.2}$$

The synthetic log-likelihood function can then be written as:

$$\ln(L) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\ln(\sigma_i^2) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{w_i - d_i}{\sigma_i}\right)^2 \tag{3.3}$$

The natural approach to choosing the weights w_i is to maximize this log-likelihood function. On setting

$$\frac{\partial \ln(L)}{\partial w_i} = 0\,, \tag{3.4}$$

we have,

$$w_i = d_i \tag{3.5}$$

This means that in the absence of any auxiliary information, the maximum likelihood weights w_i are the design weights d_i. It is worth pointing out here that the Narain (1951) and the Horvitz and Thompson (1952) estimators are in fact using maximum likelihood design weights, an observation that seems to have gone unnoticed during the past 60 years.

In this paper, we consider a new calibration criterion as:

(i) Log-likelihood function is to be maximized,
(ii) $\sum\limits_{i\in s} w_i = \sum\limits_{i\in s} d_i$ is the constraint,
(iii) the set of auxiliary variables is $\mathbf{x}_\mathbf{i}^{*} = (x_1, x_2, ..., x_p)$.

We construct a new estimator of the population total based on the calibrated maximum likelihood design weights denoted by

$$\hat{Y}_{\text{MLE}} = \sum_s w_i y_i \tag{3.6}$$

We discuss here two situations:

Case-I (when σ_i^2 is known): If the variance σ_i^2 is known, then the resultant calibrated estimator is given by:

$$\hat{Y}_{\mathbf{MLE(1)}} = \sum_s d_i y_i + \left(\sum_\Omega \mathbf{x}_i^{*} - \sum_s d\mathbf{x}_i^{*}\right)\hat{\mathrm{B}}_{\mathbf{mle}} \tag{3.7}$$

where $\hat{\mathbf{B}}_{\mathbf{mle}} = (\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)^T$ are estimators of the partial regression coefficients $\beta_1, \beta_2, \ldots, \beta_p$ obtained by minimizing the weighted least squared error defined by $\sum\limits_{i\in s} \sigma_i^2 e_i^2$, where $e_i = (y_i - \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi})$ denotes the ith residual term. Note that if we set $\sigma_i^2 = d_i q_i$ then $\hat{\mathbf{B}}_{\text{mle}} = \hat{\mathbf{B}}_{\text{ols}}$.

Case-II (when σ_i^2 is unknown): If the variance σ_i^2 is unknown, then the resultant calibrated estimator takes the form:

$$\hat{Y}_{\mathbf{MLE(2)}} = \sum_{s} d_i y_i + \left(\sum_{U} \mathbf{x}_i^* - \sum_{s} d\, \mathbf{x}_i^* \right) \hat{\mathbf{B}}_{\mathbf{ols}} + \sum_{i \in s} \frac{y_i - \mathbf{x}_i^* \hat{B}_{\mathbf{ols}}}{\mathbf{x}_i \boldsymbol{\lambda}} \tag{3.8}$$

where $\mathbf{x}_i^* = (x_{1i}, x_{2i}, ..., x_{pi})$, $\hat{\mathbf{B}}_{\text{ols}} = (\hat{\beta}_1, \hat{\beta}_2, ...\hat{\beta}_p)^T$, $\mathbf{x}_i = (1, x_{1i}, x_{2i}, ..., x_{pi})$, and $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, ..., \lambda_p)^T$. The values of the Lagrange's multipliers $\boldsymbol{\lambda}$ are given by a set of $(p+1)$ non-linear equations:

$$\sum_{i \in s} \frac{1}{\mathbf{x}_i\ \boldsymbol{\lambda}} = 0 \tag{3.9}$$

and

$$\sum_{i \in s} \frac{\mathbf{x}_i^*}{\mathbf{x}_i\ \boldsymbol{\lambda}} = (\sum_{\Omega} \mathbf{x}_i^* - \sum_{s} d_i \mathbf{x}_i^*) \tag{3.10}$$

Clearly the estimator $\hat{Y}_{\text{MLE(2)}}$ is a new estimator in the field of survey sampling, and needs investigations. We explain the derivations of these estimators with special cases considered in the following sub sections:

4 Special Cases

Let us first consider the simple case of using a single auxiliary variable. We maximize the synthetic log likelihood function subject to the two constraints (1.8) and (1.12). In case of one auxiliary variable, the Lagrange's function is given by:

$$\mathrm{L}_1 = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i \in s}\ln(\sigma_i^2) - \frac{1}{2}\sum_{i \in s}\left(\frac{w_i - d_i}{\sigma_i}\right)^2 + \lambda_0\left(\sum_{i \in s} w_i - \sum_{i \in s} d_i\right) + \lambda_1\left(\sum_{i \in s} w_i x_i - X\right) \tag{4.1}$$

where λ_0 and λ_1 are Lagrange multipliers. We discuss separately the two cases:
(a) when σ_i^2 is known (b) when σ_i^2 is unknown in the following sections.

4.1 Case-I (when σ_i^2 is known): Upon maximizing the Lagrange's function (4.1) by setting:

$$\frac{\partial L_1}{\partial w_i} = 0, \tag{4.2}$$

we get:

$$w_i = d_i + \lambda_0 \sigma_i^2 + \lambda_1 \sigma_i^2 x_i, \tag{4.3}$$

On substituting (4.3) into (1.7) we get:

$$\lambda_0 \sum_{i\in s}\sigma_i^2 + \lambda_1 \sum_{i\in s}\sigma_i^2 x_i = 0\,, \tag{4.4}$$

and, on substituting (4.3) into (1.3) we get:

$$\lambda_0 \sum_{i\in s}\sigma_i^2 x_i + \lambda_1 \sum_{i\in s}\sigma_i^2 x_i^2 = (X - \hat{X}_{HT})\,. \tag{4.5}$$

Upon solving (4.4) and (4.5) for λ_0 and λ_1 and then substituting these into (4.3), the calibrated weights are given by:

$$w_i = d_i + \frac{(\sigma_i^2 x_i)(\sum_{i\in s}\sigma_i^2) - \sigma_i^2(\sum_{i\in s}\sigma_i^2 x_i)}{(\sum_{i\in s}\sigma_i^2)(\sum_{i\in s}\sigma_i^2 x_i^2) - (\sum_{i\in s}\sigma_i^2 x_i)^2}\left(X - \hat{X}_{HT}\right) \tag{4.6}$$

The calibrated estimator (3.6) thus leads to a new estimator of the population total Y given by:

$$\hat{Y}_{\text{new}(1)_1} = \hat{Y}_{HT} + \hat{\beta}_{\text{new}}\left(X - \hat{X}_{HT}\right) \tag{4.7}$$

where the regression coefficient $\hat{\beta}_{\text{new}}$ is given by:

$$\hat{\beta}_{\text{new}} = \frac{(\sum_{i\in s}\sigma_i^2 x_i y_i)(\sum_{i\in s}\sigma_i^2) - (\sum_{i\in s}\sigma_i^2 y_i)(\sum_{i\in s}\sigma_i^2 x_i)}{(\sum_{i\in s}\sigma_i^2)(\sum_{i\in s}\sigma_i^2 x_i^2) - (\sum_{i\in s}\sigma_i^2 x_i)^2}\,. \tag{4.8}$$

Remark: If $\sigma_i^2 = d_i q_i$ is assumed to be known, then the new estimator $\hat{Y}_{\text{new}(1)}$ reduces to the estimator studied by Stearns and Singh (2008) and Singh and Arnab (2011). Forcing $\sigma_i^2 = d_i q_i$ may not be a good idea, because the variance of w_i is also a function of x_i, which may affect the resulting estimator.

4.2 Case-II (when σ_i^2 is unknown): We next consider the problem of estimating σ_i^2 by maximum likelihood approach. On setting:

$$\frac{\partial L_1}{\partial \sigma_i^2} = 0\,, \tag{4.9}$$

we get:

$$\sigma_i^2 = (w_i - d_i)^2\,. \tag{4.10}$$

On substituting w_i from (4.3) into (4.10), an estimator of the variance σ_i^2 is given by:

$$\hat{\sigma}_i^2 = \frac{1}{(\lambda_0 + \lambda_1 x_i)^2} \tag{4.11}$$

where, from equations (4.4) and (4.5) after replacing σ_i^2 by $\hat{\sigma}_i^2$, the values of λ_0 and λ_1 are given by solutions to the following non-linear equations:

$$\sum_{i \in s} \frac{1}{\lambda_0 + \lambda_1 x_i} = 0 \tag{4.12}$$

and

$$\sum_{i \in s} \frac{x_i}{\lambda_0 + \lambda_1 x_i} = (X - \hat{X}_{HT}) \tag{4.13}$$

Such non-linear equations can be solved by following Owen (2001). The calibrated estimator (3.6) then leads to a new estimator of the population total Y given by:

$$\hat{Y}_{\text{new}(2)_1} = \hat{Y}_{HT} + \sum_{i \in s} \frac{y_i}{\lambda_0 + \lambda_1 x_i} \tag{4.14}$$

Note carefully that the new estimator $\hat{Y}_{\text{new}(2)_1}$ can be written as:

$$\hat{Y}_{\text{new}(2)_1} = \hat{Y}_{HT} + \sum_{i \in s} \frac{(y_i - \hat{\beta}_{ols} x_i) + \hat{\beta}_{ols} x_i}{\lambda_0 + \lambda_1 x_i}$$

or equivalently

$$\hat{Y}_{\text{new}(2)_1} = \hat{Y}_{HT} + \hat{\beta}_{ols}\left(X - \hat{X}_{HT}\right) + \sum_{i \in s} \frac{(y_i - \hat{\beta}_{ols} x_i)}{\lambda_0 + \lambda_1 x_i} \tag{4.15}$$

which is clearly a new estimator in the field of survey sampling. It can easily be extended for stratified sampling, non-response, two-phase sampling, and small or medium sized area estimation etc.

In the next section, we perform a small scale simulation study where the proposed estimator is shown to remain more efficient than the linear regression estimator.

5. Simulation Study

We generated several synthetic populations of size $N = 5001$ by following Singh and Horn (1998) for different values of the population correlation coefficient ρ as follows:

$$y_i = 0.008 + \sqrt{(1-\rho^2)}\, y_i^* + \rho \frac{\sigma_y}{\sigma_x} x_i^* \tag{5.1}$$

and

$$x_i = 0.004 + x_i^* \tag{5.2}$$

where $x_i^* \sim \text{Gamma}(0.3, 0.2)$ and $y_i^* \sim \text{Gamma}(0.1, 1.0)$. We considered $\sigma_x^2 = (0.3)(0.2)^2$, $\sigma_y^2 = (0.1)(1.0)^2$, and different values of the population correlation coefficient ρ in the range 0.5 to 0.9. From a given population of size $N = 5001$ and for a given value of ρ, we selected $T = 50{,}000$ samples each of size n in the range of 30 to 80. For simplicity, let

$$\hat{\theta}_0 = \bar{y}_n\,, \tag{5.3}$$

$$\hat{\theta}_1 = \bar{y}_n \left(\frac{\bar{X}}{\bar{x}_n} \right), \tag{5.4}$$

$$\hat{\theta}_2 = \bar{y}_n + \hat{\beta}_{\text{ols}} (\bar{X} - \bar{x}_n)\,, \tag{5.5}$$

and

$$\hat{\theta}_3 = \bar{y}_n + \hat{\beta}_{\text{ols}} (\bar{X} - \bar{x}_n) + \sum_{i=1}^{n} \frac{(y_i - \hat{\beta}_{\text{ols}} x_i)}{\lambda_0 + \lambda_1 x_i} \tag{5.6}$$

where the values of λ_0 and λ_1 are obtained by solving the non-linear equations:

$$\sum_{i=1}^{n} \frac{1}{\lambda_0 + \lambda_1 x_i} = 0 \text{ and } \sum_{i=1}^{n} \frac{x_i}{\lambda_0 + \lambda_1 x_i} = (\bar{X} - \bar{x}_n)\,. \tag{5.7}$$

(we solved these equations by using IMSL subroutine NEQNF in FORTRAN).

We computed the empirical relative bias in the jth estimator given by:

$$\text{B}(\hat{\theta}_j) = \frac{\frac{1}{T} \sum_{k=1}^{T} \hat{\theta}_{j|k} - \bar{Y}}{\bar{Y}} \times 100\%, \quad j = 1, 2, 3\,. \tag{5.8}$$

where $\hat{\theta}_{j|k}$ is the value of the estimator $\hat{\theta}_j$ obtained from the k^th sample. We computed the relative efficiency of the estimators $\hat{\theta}_j$, $j = 1,2,3$ over the sample mean estimator $\hat{\theta}_0$, as:

$$RE(\hat{\theta}_j) = \frac{\sum_{k=1}^{T} (\hat{\theta}_{0|k} - \bar{Y})^2}{\sum_{k=1}^{T} (\hat{\theta}_{j|k} - \bar{Y})^2} \times 100\%, \quad j = 1, 2, 3\,. \tag{5.9}$$

We also computed average values of λ_j, $j = 0,1$, along with their standard deviations, as:

$$\bar{\lambda}_j = \frac{1}{T}\sum_{k=1}^{T}\lambda_{j|k} \text{ and } \sigma_{\lambda_j} = \sqrt{\frac{1}{T}\sum_{k=1}^{T}(\lambda_{j|k} - \bar{\lambda}_j)^2} \qquad (5.10)$$

The results so obtained are presented in Table 5.1. We observe that the mean and standard deviation of λ_j values, $j = 0,1$, remain the same for each fixed sample size, as the value of the correlation coefficient varies. The average values of λ_j, $j = 0,1$ along with their standard deviations are presented in the last four columns of the table. The value of $\rho_{xy} = 0.4959$ is the observed value of the correlation coefficient in the population generated with $\rho = 0.5$ in Singh and Horn (1998) method. The percent relative bias in all four of the estimators considered here is negligible. For $\rho_{xy} = 0.4959$ the ratio estimator $\hat{\theta}_1$ remains less efficient than the sample mean for any sample size between 30 and 80. For $\rho_{xy} = 0.5968$, the relative efficiency of the ratio estimator with respect to the sample mean estimator increases from 86.04% to 105.69% as sample size increases from 30 to 80. For $\rho_{xy} = 0.6977$, or above, the ratio estimator always remains more efficient than the sample mean estimator. For $\rho_{xy} = 0.4959$, the percent relative efficiency of the regression estimator increases from 127.47% to 131.52% as the sample size increases from 30 to 80; whereas the percent relative efficiency of the proposed estimator increases from 130.63% to 140.30%. For $\rho_{xy} = 0.8995$, the percent relative efficiency of the ratio estimator increases from 302.54% to 369.08%; that of the regression estimator increases from 509.13% to 521.49% and that of the proposed estimator increases from 521.98% to 579.20% as the sample size increases from 30 to 80. Thus this simulation study shows that there could be situations in real practice where the proposed estimator can perform better than the linear regression estimator.

Table 5.1. Performance of the proposed estimator with respect to the sample mean, ratio and regression estimator.

n	B($\hat{\theta}_1$)	B($\hat{\theta}_2$)	B($\hat{\theta}_3$)	RE($\hat{\theta}_1$)	RE($\hat{\theta}_2$)	RE($\hat{\theta}_3$)	$\bar{\lambda}_0$	σ_{λ_0}	$\bar{\lambda}_1$	σ_{λ_1}
$\rho_{xy} = 0.4959$ ($\rho = 0.5$)										
30	4.995	-0.047	-1.172	72.81	127.47	130.63	-882.6	1088.5	-5928.1	19373.1
40	3.906	0.039	-1.865	77.43	129.22	136.29	-941.9	698.6	-2618.1	12413.8
50	2.927	-0.134	-2.770	83.21	131.31	141.49	-948.8	470.1	-1221.7	7690.9
60	2.420	-0.094	-3.384	84.91	130.20	142.31	-944.3	329.6	-587.4	5100.4
70	2.232	-0.032	-3.964	88.09	131.95	140.30	-943.5	239.4	-251.7	3581.3
80	1.752	-0.062	-4.519	89.69	131.52	144.28	-950.8	183.2	52.7	2627.3
$\rho_{xy} = 0.5968$ ($\rho = 0.6$)										
30	4.292	-0.041	-1.007	86.04	149.70	153.42	-882.6	1088.5	-5928.1	19373.1
40	3.359	0.035	-1.603	91.18	151.25	159.56	-941.9	698.6	-2618.1	12413.8
50	2.515	-0.115	-2.383	98.15	153.97	166.00	-948.8	470.1	-1221.7	7690.9

60	2.080	-0.083	-2.911	99.92	152.30	166.62	-944.3	329.6	-587.4	5100.4
70	1.918	-0.028	-3.408	103.81	154.58	164.58	-943.5	239.4	-251.7	3581.3
80	1.505	-0.054	-3.886	105.69	154.09	169.37	-950.8	183.2	52.7	2627.3
$\rho_{xy} = 0.6977\,(\rho = 0.7)$										
30	3.601	-0.035	-0.846	109.04	188.36	193.05	-882.6	1088.5	-5928.1	19373.1
40	2.822	0.029	-1.345	115.14	189.66	200.13	-941.9	698.6	-2618.1	12413.8
50	2.109	-0.099	-2.002	124.15	193.41	208.64	-948.8	470.1	-1221.7	7690.9
60	1.746	-0.070	-2.443	126.10	190.89	209.04	-944.3	329.6	-587.4	5100.4
70	1.610	-0.024	-2.862	131.17	193.97	206.82	-943.5	239.4	-251.7	3581.3
80	1.263	-0.046	-3.263	133.59	193.46	213.13	-950.8	183.2	52.7	2627.3
$\rho_{xy} = 0.7986\,(\rho = 0.8)$										
30	2.871	-0.028	-0.675	156.45	267.69	274.39	-882.6	1088.5	-5928.1	19373.1
40	2.253	0.024	-1.073	164.64	268.66	283.61	-941.9	698.6	-2618.1	12413.8
50	1.681	-0.081	-1.599	177.77	274.39	296.21	-948.8	470.1	-1221.7	7690.9
60	1.392	-0.057	-1.952	180.19	270.29	296.38	-944.3	329.6	-587.4	5100.4
70	1.285	-0.020	-2.285	187.57	274.86	293.64	-943.5	239.4	-251.7	3581.3
80	1.008	-0.037	-2.605	191.23	274.45	303.26	-950.8	183.2	52.7	2627.3
$\rho_{xy} = 0.8995\,(\rho = 0.9)$										
30	1.995	-0.020	-0.471	302.54	509.13	521.98	-882.6	1088.5	-5928.1	19373.1
40	1.568	0.017	-0.746	317.42	509.64	538.32	-941.9	698.6	-2618.1	12413.8
50	1.168	-0.058	-1.114	342.96	520.97	563.12	-948.8	470.1	-1221.7	7690.9
60	0.967	-0.041	-1.360	347.21	512.64	563.38	-944.3	329.6	-587.4	5100.4
70	0.893	-0.014	-1.591	361.37	521.19	558.66	-943.5	239.4	-251.7	3581.3
80	0.700	-0.027	-1.814	369.08	521.49	579.20	-950.8	183.2	52.7	2627.3

The results obtained through simulation study are quite encouraging, thus we look for datasets where the proposed estimator can perform better than the linear regression estimator which we do in the next section.

6. Application to Real Dataset

Despres *et al.* (1991) found that the topography of adipose tissue is associated with metabolic complications that are considered at risk factors for cardiovascular disease. It is important, they mentioned, to measure the amount of intra-abdominal adipose tissue as part of the evaluation of the cardiovascular-disease risk of an individual. There is only one technique, called computer tomography that precisely and reliably measures the amount of deep abdominal adipose tissue; however it is costly and requires irradiation of the subject. Also they mentioned that this technique is not available to many physicians. Despres *et al.* (1991), in their study, took data from 109 men in the age range of 18 to 42 years who were free from any metabolic disease that would require treatment. Among the measurements taken on each subject were the study variable, Y, deep abdominal adipose tissue obtained by computer tomography, and an auxiliary variable, X, waist circumference. The value of the population correlation coefficient is $\rho_{xy} = 0.8186$. A graphical representation of the dataset is shown in Figure 6.1.

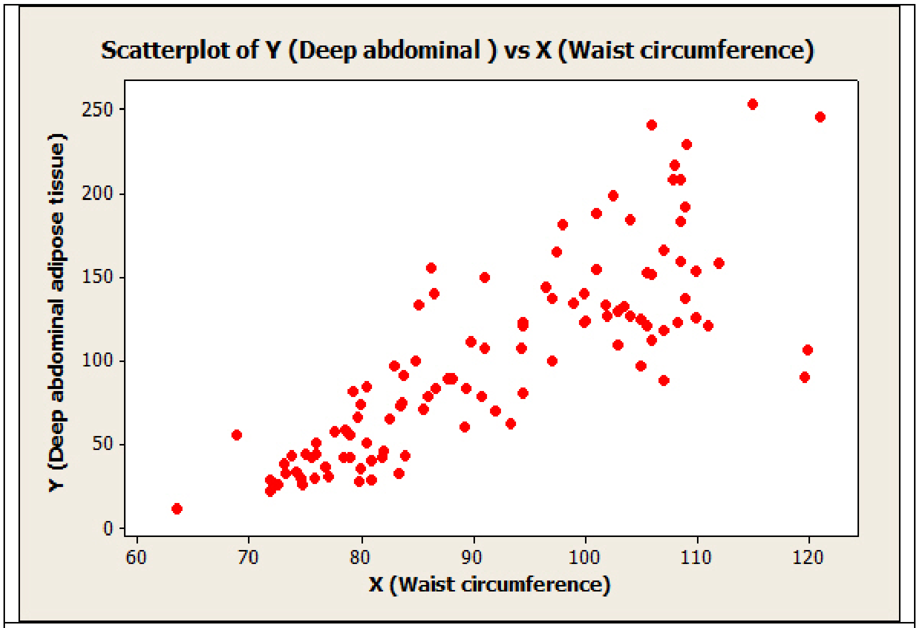

Fig. 6.1. The scatter plot of real data set.

Table 6.1. Performances of the proposed estimator with respect to the sample mean, ratio and regression estimators based on the real dataset.

n	B($\hat{\theta}_1$)	B($\hat{\theta}_2$)	B($\hat{\theta}_3$)	RE($\hat{\theta}_1$)	RE($\hat{\theta}_2$)	RE($\hat{\theta}_3$)	$\bar{\lambda}_0$	σ_{λ_0}	$\bar{\lambda}_1$	σ_{λ_1}
5	-3.950	0.310	-0.306	148.3	196.1	200.9	173.5	2000.7	144.5	203.9
6	-3.207	0.097	-0.652	149.9	228.5	234.2	-95.0	2049.4	121.4	162.4
7	-2.694	0.032	-0.767	150.3	241.2	246.9	85.1	2089.4	101.9	125.5
8	-2.313	-0.010	-0.897	151.0	253.9	259.6	-473.0	2058.6	94.8	114.8
9	-2.046	-0.040	-1.020	151.8	260.0	265.4	-476.8	2083.6	85.3	98.8
10	-1.770	0.048	-1.028	152.0	260.1	265.9	-150.9	2167.3	74.2	83.0
11	-1.586	0.031	-1.118	152.3	267.0	272.4	-434.2	2154.6	71.5	79.4
12	-1.366	0.079	-1.174	152.5	266.8	272.1	-113.2	2230.9	63.2	67.4
13	-1.272	0.051	-1.328	153.1	269.3	270.8	-443.4	2211.8	62.8	67.1
14	-1.170	0.043	-1.406	152.9	267.6	268.8	-240.6	2274.2	57.6	58.2
15	-1.084	0.021	-1.540	153.5	270.0	267.0	-364.3	2277.7	56.1	58.5

It is interesting to point out that the proposed estimator remains more efficient than both the ratio and the linear regression estimator in the case of small sample size. Again based on 50,000 iterations, for a sample of 5 subjects, the relative efficiency of the proposed estimator is 200.9% where as that of the linear regression estimator is 196.1%, and both estimators show negligible relative bias. Thus there is almost 4.8% gain in relative efficiency at the cost of solving two non-linear equations. For a sample of 6 subjects the gain in the relative efficiency becomes 5.7%, and up to a sample of 12 subjects the gain in percent relative efficiency remains more than 5% again at the cost of solving two non-linear equations. As the sample size becomes more than 15 then the proposed estimator becomes less efficient than the linear regression estimator. Thus in the case of small samples, the proposed estimator can be recommended for its applications in the biosciences to situations similar to the one considered here.

It is clearly a fact that experience is a must while choosing an estimator in a particular situation. The FORTRAN code used in the simulation study and in the real dataset application can be requested from the authors.

Acknowledgement: *The authors would like to thank Professor Maria del Mar Rueda, Department of Statistics, University of Granada, Spain for her valuable comments on the original version of this manuscript during her visit at Texas A&M University, Kingsville.*

References

Beaumont, J.F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, 95, 3, 539-553.

Bethlehem, J.G and Keller, W.J. (1987). Linear weighting of sample survey data. *J. Official Statist.*, 141-153.

Cochran, W.G. (1940). Some properties of estimators based on sampling scheme with varying probabilities. *Austral. J. Statist.,* 17, 22-28.

Despres, J.P., Prud'homme, D., Pouliot, M., Tremblay, A. and Bouchard, C. (1991). Estimation of Deep Abdominal Adipose-Tissue Accumulation from Simple Anthropometric Measurements in Men. *Amer. Jour. of Clinical Nutrition*, 54, 471-477.

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

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

Haung, E.T. and Fuller, W.A. (1978). Nonnenative Regression Estimation for Sample Survey Data. *Proc. of the Social Stat. Sec., Amer. Stat. Assoc., San Diego,* 300-305.

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

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *J. Ind. Soc. Agril. Statist.*, 3, 169-174.

Owen, A.B. (2001). *Empirical Likelihood*. Chapman & Hall.

Rao, T.J. (2004). Five decades of the Horvitz-Thompson estimator and furthermore. *J. Ind. Soc. Agril. Statist.*, 58(2), 177-189.

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

Singh, S. (2003). *Advanced Sampling Theory with Applications: How Michael Selected Amy.* Kluwer Academic Publishers. The Netherlands.

Singh, S. (2004). Golden and Silver jubilee year-2003 of the linear regression estimators. *Proc. of the Amer. Stat. Assoc., survey method section, Toronto,* pp. 4382-4389.

Singh, S. (2006). Survey statistician celebrate golden jubilee year-2003 of the linear regression estimator. *Metrika*, pp. 1-18.

Singh, S. and Arnab, R. (2011). On calibration of design weights. *Metron*, 69, 185-205.

Singh, S. and Horn, S. (1998). An alternative estimator for multi-character surveys. *Metrika,* 48, 99--107.

Singh, S. and Sedory, S.A. (2012). An open invitation to a fresh survey methodology. *Liaison, Statistical Society of Canada*, 26(2), page 54.

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

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