International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:6, No:12, 2012

Mathematical Programming on Multivariate Calibration Estimation in Stratified Sampling

Dinesh Rao, M.G.M. Khan, and Sabiha Khan

***Abstract*—**Calibration estimation is a method of adjusting the original design weights to improve the survey estimates by using auxiliary information such as the known population total (or mean) of the auxiliary variables. A calibration estimator uses calibrated weights that are determined to minimize a given distance measure to the original design weights while satisfying a set of constraints related to the auxiliary information. In this paper, we propose a new multivariate calibration estimator for the population mean in the stratified sampling design, which incorporates information available for more than one auxiliary variable. The problem of determining the optimum calibrated weights is formulated as a Mathematical Programming Problem (MPP) that is solved using the Lagrange multiplier technique.

***Keywords*—**Calibration estimation, Stratified sampling, Multivariate auxiliary information, Mathematical programming problem, Lagrange multiplier technique.

I. Introduction

CALIBRATION is commonly used in survey sampling to increase the precision of the estimators of population parameter when auxiliary information is available. The method works by modifying the original design weights incorporating the known population characteristics, in practice population totals or population means, of the auxiliary variables. Deville and Särndal (1992) first used the calibration estimators in survey sampling [3]. Wu and Sitter (2001) suggested the model-calibration estimator that uses an explicit working model for $E\left(y_i \left| x_i\right.\right)$ [8]. Instead of using an explicit parametric model, Briedt and Opsomer (2000) adopted a local polynomial regression model to derive a non-parametric regression estimator [1]. Singh, Horn and Yu (1998), and Kim, Sungur and Heo (2007) introduced the calibration estimation in stratified sampling. They suggested the calibration estimators, respectively, for combined generalized regression estimator and combined ratio estimator using a single auxiliary information [4], [7]. Chen and Qin (1993) suggested a calibrated estimator that makes an efficient use of auxiliary variables for equal probability sampling by maximizing the constrained empirical likelihood [2]. Kim (2009) extended this technique to unequal probability sampling and also implemented the result in stratified sampling [5].

In surveys, when more than one auxiliary information is available, the precision of the estimate can further be increased by adjusting the design weights based on all the auxiliary information. In this paper, we propose a multivariate calibration estimator for the population mean with the aid of several auxiliary information in stratified random sampling for improving the precision of the estimate. The problem of determining the optimum calibrated weights is formulated as a Mathematical Programming Problem (MPP) that minimizes the chi-square type distance subject to the p calibration constraints and the non-negativity restrictions on the calibrated weights, where p is the number of available auxiliary variables. Ignoring the non-negativity restrictions a solution procedure is developed to solve the MPP using Lagrange multiplier technique. The closed form expression for the solution of the calibrated weight is derived in the presence of two auxiliary variables for a stratified random sampling design. The MPP is solved completely if the non-negativity restrictions on calibrated weights are satisfied. If the restrictions are violated, another solution procedure is developed by extending the procedure proposed by Singh (2003) that minimizes a distance function subject to the p calibration constraints, which guarantees the non-negativity of the weights [6]. Two numerical examples are presented to illustrate the application and computational details of the proposed techniques to determine the multivariate calibrated estimator. The examples reveal that the proposed multivariate calibrated estimator is more efficient than the usual estimator of the population mean in stratified sampling.

II. Formulation of the Problem as an MPP

Let the population be divided into L non-overlapping strata and n_h be the number of units drawn by simple random sampling without replacement (SRSWOR) from the hth stratum consisting of N_h units, and $n=\sum_{h=1}^{L} n_h$ and $N=\sum_{h=1}^{L} N_h$ give the total sample size and the population size. For the hth strata, let $W_h = N_h/N$ be the strata weights and $\bar{y}_h, \bar{Y}_h$ are the sample and population means, respectively, for the study variable.

Let the estimation of unknown population means $\bar{Y}$ be of interest using the information from p auxiliary variables X_j, $j=1,2,\ldots,p$. Let y_{hi} and x_{hij} denote the values of the ith population (sampled) unit of the study variable (Y) and the

Dinesh Rao and M.G.M Khan are with the School of Computing, Information and Mathematical Sciences, Faculty of Science, Technology and Environment, The University of the South Pacific, Suva, Fiji Islands (e-mail: rao_di@usp.ac.fj and khan_mg@usp.ac.fj).

Sabiha Khan is with the Department of Public Health, Fiji School of Medicine, Fiji National University, Suva, Fiji Islands (e-mail: sabiha.khan@fnu.ac.fj).

j th auxiliary variable (X_j) respectively, in the h th stratum. Assume that the strata means $\bar{X}_j = \sum_{h=1}^{L} W_h \bar{X}_{hj}$ are accurately known. The purpose is to estimate the population mean $\bar{Y} = \sum_{h=1}^{L} W_h \bar{Y}_h$ by using the auxiliary information X_j .

The usual estimator of population mean $\bar{Y}$ is given by

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \ . \tag{1}$$

In the presence of more than one auxiliary information, we suggest a multivariate calibrated estimator of the population mean $\bar{Y}$ under stratified sampling given by

$$\bar{y}_{st}^{*} = \sum_{h=1}^{L} W_h^{*} \bar{y}_h \tag{2}$$

with new weights W_h^{*}. When more than one auxiliary variables X_j , $j = 1, 2, \ldots, p$ is available, the new weights W_h^{*} are so chosen such that the sum of the chi-square type distances given by

$$\sum_{j=1}^{p} \sum_{h=1}^{L} \frac{\left(W_h^{*} - W_h\right)^2}{W_h q_{hj}} \tag{3}$$

is minimum, subject to the calibration constraints

$$\sum_{h=1}^{L} W_h^{*} \bar{x}_{hj} = \bar{X}_j ; \quad j = 1, 2, \ldots, p. \tag{4}$$

Note that $q_{hj} > 0$ in (3) are suitability chosen weights which determine the form of estimator. One of the challenges in calibration approach of estimation is that sometimes the calibrated weights do not satisfy the desired constraint of weights being non-negative. To avoid such situation one needs to impose the restrictions

$$W_h^{*} \geq 0 ; \ h = 1, 2, \ldots, L. \tag{5}$$

Thus, the problem of determining the optimum calibrated weights W_h^{*} may be formulated as a Mathematical Programming Problem (MPP) as given below:

$$\text{Minimize } Z = \sum_{h=1}^{L} \frac{\left(W_h^{*} - W_h\right)^2}{W_h Q_h}$$

$$\text{subject to } \sum_{h=1}^{L} W_h^{*} \bar{x}_{h1} = \bar{X}_1 ,$$

$$\sum_{h=1}^{L} W_h^{*} \bar{x}_{h2} = \bar{X}_2 ,$$

$$\vdots$$

$$\sum_{h=1}^{L} W_h^{*} \bar{x}_{hp} = \bar{X}_p ,$$

$$\text{and } W_h^{*} \geq 0 ; \ h = 1, 2, \ldots, L \tag{6}$$

where $Q_h = \sum_{j=1}^{p} q_{hj}$.

III. Determining Optimal Calibrated Weights: The Solution Procedure

Ignoring the restrictions in (5), we can use Lagrange multipliers technique to solve the MPP (6) for determining the optimum values of W_h^{*}, since the constraints are equality constraints. If the values W_h^{*} satisfy the ignored restrictions, the MPP in (6) is solved completely.

To solve (6), we associate a multiplier $-2\lambda_j$ with the j th constraint in (6), Then, the Lagrangian function L is formed as:

$$L\left(W_h^{*}, \lambda_j\right) = \sum_{h=1}^{L} \frac{\left(W_h^{*} - W_h\right)^2}{W_h Q_h} - 2 \sum_{j=1}^{p} \lambda_j \left(\sum_{h=1}^{L} W_h^{*} \bar{x}_{hj} - \bar{X}_j \right). \tag{7}$$

The necessary conditions for the solution of the problem are

$$\frac{\partial L}{\partial W_h^{*}} = \frac{\partial L}{\lambda_j} = 0. \tag{8}$$

The determination of the optimum calibrated weights W_h^{*} using the Lagrange multiplier technique discussed above is illustrated in Theorem 1 when information on two auxiliary variables $X_j ; (j = 1, 2)$ is available.

Theorem 1: In stratified sampling, when $p = 2$, the optimum solution to the MPP (6), that is, the optimum calibrated weights W_h^{*} that minimize (3) subject to the conditions (4) is given by

$$W_h^{*} = W_h + W_h Q_h \left(\lambda_1 \bar{x}_{h1} + \lambda_2 \bar{x}_{h2} \right) \tag{9}$$

where,

$$\lambda_1 = \frac{-C\left(\bar{X}_1 - \hat{\bar{X}}_1\right) + B\left(\bar{X}_2 - \hat{\bar{X}}_2\right)}{B^2 - AC}, \qquad \lambda_2 = \frac{B\left(\bar{X}_1 - \hat{\bar{X}}_1\right) - A\left(\bar{X}_2 - \hat{\bar{X}}_2\right)}{B^2 - AC}, \tag{10}$$

$$\hat{\bar{X}}_1 = \sum_{h=1}^{L} W_h \bar{x}_{h1}, \quad \hat{\bar{X}}_2 = \sum_{h=1}^{L} W_h \bar{x}_{h2}, \tag{11}$$

$$A=\sum_{h=1}^{L} W_h Q_h \bar{x}_{h1}^2,$$
$$B=\sum_{h=1}^{L} W_h Q_h \bar{x}_{h1}\bar{x}_{h2} \quad \text{and} \tag{12}$$
$$C=\sum_{h=1}^{L} W_h Q_h \bar{x}_{h2}^2.$$

Proof of Theorem 1: Using Lagrange multiplier technique, the function to be minimized

$$L=\sum_{h=1}^{L}\frac{\left(W_h^*-W_h\right)^2}{W_h Q_h}-2\lambda_1\left(\sum_{h=1}^{L} W_h^*\bar{x}_{h1}-\bar{X}_1\right)$$
$$-2\lambda_2\left(\sum_{h=1}^{L} W_h^*\bar{x}_{h2}-\bar{X}_2\right).$$

The necessary conditions given in (8) are

$$\frac{\partial L}{\partial W_h^*}=2\frac{\left(W_h^*-W_h\right)}{W_h Q_h} \tag{13}$$
$$2\lambda_1\bar{x}_{h1}-2\lambda_2\bar{x}_{h2}=0,$$

$$\frac{\partial L}{\partial \lambda_1}=-2\left(\sum_{h=1}^{L} W_h^*\bar{x}_{h1}-\bar{X}_1\right)=0 \tag{14}$$

and

$$\frac{\partial L}{\partial \lambda_2}=-2\left(\sum_{h=1}^{L} W_h^*\bar{x}_{h2}-\bar{X}_2\right)=0. \tag{15}$$

Solving the necessary conditions (13) to (15) completes the proof.

Therefore, we obtain the new multivariate calibrated estimator of the population mean

$$\bar{y}_{st}^*=\sum_{h=1}^{L} W_h^*\bar{y}_h$$

stated in (2), where the optimum calibrated weights W_h^* is defined in (9).

If the calibrated weights in (9) violate the restrictions (5), we develop the following technique that minimizes a distance function given in Singh (2003) subject to constraints given in (4) [6].

The following theorem discusses a distance function which guarantees the non-negativity of the calibrated weights when information on two auxiliary variables $X_j;(j=1,2)$ is available.

Theorem 2: The optimum calibrated weights obtained by minimizing the distance function

$$Z=\sum_{h=1}^{L} W_h^*\ln\left(\frac{1}{W_h}\right)-\sum_{h=1}^{L} W_h^*\ln\left(\frac{1}{W_h^*}\right) \tag{16}$$

subject to the calibration constraints (4) for $p=2$, leads to non-negative weights.

Proof of Theorem 2: In this situation, the Lagrange function to be minimized is as follows:

$$L=\sum_{h=1}^{L} W_h^*\ln\left(\frac{W_h^*}{W_h}\right)$$
$$-\lambda_1\left(\sum_{h=1}^{L} W_h^*\bar{x}_{h1}-\bar{X}_1\right)-\lambda_2\left(\sum_{h=1}^{L} W_h^*\bar{x}_{h2}-\bar{X}_2\right).$$

The necessary conditions given in (8) are

$$\frac{\partial L}{\partial W_h^*}=1+\ln\left(W_h^*\right)-\ln\left(W_h\right) \tag{17}$$
$$-\lambda_1\bar{x}_{h1}-\lambda_2\bar{x}_{h2}=0,$$

$$\frac{\partial L}{\partial \lambda_1}=-\left(\sum_{h=1}^{L} W_h^*\bar{x}_{h1}-\bar{X}_1\right)=0 \tag{18}$$

and

$$\frac{\partial L}{\partial \lambda_2}=-\left(\sum_{h=1}^{L} W_h^*\bar{x}_{h2}-\bar{X}_2\right)=0. \tag{19}$$

From (17) we have

$$W_h^*=\exp\left[\ln\left(W_h\right)+\lambda_1\bar{x}_{h1}+\lambda_2\bar{x}_{h2}-1\right]. \tag{20}$$

Using (20) and solving the equations (18) and (19), we obtain the values of λ_1 and λ_2.

Thus (20) shows that the calibrated weights are always non-negative if the distance function (16) is minimized, satisfying the calibration constraint (4) Hence the theorem.

IV. Numerical Illustrations

Example 1: In order to illustrate and demonstrate the determination of the proposed multivariate calibrated estimator, we use a tobacco population data of $N=106$ countries with three variables: area (in hectares), yield (in metric tons) and production (in metric tons). The data are obtained from the Agriculture Statistics 1999 reported in Singh (2003) [6]. The countries were divided into $L=10$ strata and a sample of $n=40$ countries using proportional allocation was selected. Suppose that an estimate of average production ($\bar{Y}$) of tobacco crop is of interest using the two auxiliary variables X_1 = area and X_2 = yield. Assume that $\bar{X}_1$ and $\bar{X}_2$ in different countries are known. To compute the multivariate calibrated weights in stratified sampling and the value of the estimate of $\bar{Y}$, we use the same sample units as obtained in Singh (2003) [6]. Assuming $Q_h=\sum_{j=1}^{2} q_{hj}=1$, Table I shows the following sample information:

$\hat{\bar{X}}_1=\sum_{h=1}^{L} W_h\bar{x}_{h1}=59811.28$,

$\hat{\bar{X}}_2 = \sum_{h=1}^{L} W_h \bar{x}_{h2} = 1.56942$,

$A = \sum_{h=1}^{L} W_h Q_h \bar{x}_{h1}^2 = 14211940497.2$,

$B = \sum_{h=1}^{L} W_h Q_h \bar{x}_{h1} \bar{x}_{h2} = 80340.21$

and $C = \sum_{h=1}^{L} W_h Q_h \bar{x}_{h2}^2 = 2.63050$.

For this population the known population means for the auxiliary variables are:

$$\bar{X}_1 = 34438.61 \text{ and } \bar{X}_2 = 1.5507$$

Using (10), $\lambda_1 = -2.10924 \times 10^{-6}$ and $\lambda_2 = 0.0573$. Thus, the multivariate calibrated weights W_h^* in stratified sampling proposed in (9) is reduced to

$$W_h^* = W_h + W_h \left(-2.10924 \times 10^{-6} \bar{x}_{h1} + 0.0573 \bar{x}_{h2} \right)$$

which are obtained and presented in Table II.

The usual estimator of population mean $\bar{Y}$ given in (1) under the proportional allocation is

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h = 94666.73. \quad (21)$$

Whereas an estimate of the average production of tobacco using the proposed generalized multivariate estimator in (2) is

$$\bar{y}_{st}^* = \sum_{h=1}^{L} W_h^* \bar{y}_h = 53952.56. \quad (22)$$

The true average production of the tobacco crop for this population is 52444.56. Thus from (21) and (22), it is evident that the proposed multivariate calibration estimator is more closed to true population mean as compared to usual estimator.

Example 2: In this illustration we use an artificial population data with the auxiliary variables X_1 and X_2 and the main variable Y, where the data were divided into $L = 4$ strata and a sample using proportional allocation was selected. Suppose that an estimate of $\bar{Y}$ is of interest using the two auxiliary variables X_1 and X_2. Assume that $\bar{X}_1$ and $\bar{X}_2$ are known and $Q_h = \sum_{j=1}^{2} q_{hj} = 1$. To compute the multivariate calibrated weights in stratified sampling and the value of the estimate of $\bar{Y}$, the following sample information shown in Table III was used.

$\hat{\bar{X}}_1 = \sum_{h=1}^{L} W_h \bar{x}_{h1} = 170020.6$,

$\hat{\bar{X}}_2 = \sum_{h=1}^{L} W_h \bar{x}_{h2} = 1.562$,

$A = \sum_{h=1}^{L} W_h Q_h \bar{x}_{h1}^2 = 93926357687.4$,

$B = \sum_{h=1}^{L} W_h Q_h \bar{x}_{h1} \bar{x}_{h2} = 326031.7$

and $C = \sum_{h=1}^{L} W_h Q_h \bar{x}_{h2}^2 = 2.493$.

For this population the known population means for the auxiliary variables are

$$\bar{X}_1 = 37453.78 \text{ and } \bar{X}_2 = 1.5671.$$

Using (10), $\lambda_1 = -2.59723 \times 10^{-6}$ and $\lambda_2 = 0.34163$. Thus, the multivariate calibrated weights W_h^* in stratified sampling proposed in (9) are $W_1^* = -0.02864$, $W_2^* = 0.31792$, $W_3^* = 0.47389$ and $W_4^* = 0.31072$. These calibrated weights violate the restrictions (5) as one of the calibrated weights is negative. Thus we use the second technique developed in Section 3 to compute the non-negative calibrated weights.

Using (20) and solving the equations (18) and (19), we obtain the constants to be $\lambda_1 = -8.76261 \times 10^{-6}$ and $\lambda_2 = 1.12097$. Thus, the proposed multivariate calibrated weights W_h^* given in (20) are $W_1^* = 0.00110$, $W_2^* = 0.42924$, $W_3^* = 0.48673$, and $W_4^* = 0.12782$.

From this data the usual estimator of population mean $\bar{Y}$ under the proportional allocation using (1), is 94289.66. Whereas the estimate of population means $\bar{Y}$ using the method of minimizing distance is 58249.34. The true population mean is 49299.73. Therefore it is evident that the proposed multivariate calibration estimator is more efficient than the usual estimator of population mean.

V. Conclusion

In this paper, we propose the techniques of determining the multivariate calibrated estimator to improve the survey estimates when more than one auxiliary variable is available. The problem of determining optimum calibrated weights is formulated as an MPP, which is solved using Lagrange multiplier technique.

Two numerical examples are presented to illustrate the computational details of the proposed techniques and the performance of the proposed estimator. The results reveal that the proposed estimator performs better than the usual estimator

APPENDIX

TABLE I
SAMPLE INFORMATION FOR TOBACCO POPULATION

h	$\bar{x}_{h1}$	$\bar{x}_{h2}$	$\bar{y}_h$	W_h	$W_h\bar{x}_{h1}$	$W_h\bar{x}_{h2}$	$W_hQ_h\bar{x}_{h1}^2$	$W_hQ_h\bar{x}_{h2}^2$	$W_hQ_h\bar{x}_{h1}\bar{x}_{h2}$
1	1304.7	1.940	2592.0	0.05660	73.85	0.10981	96348.4	0.21303	143.27
2	29075.0	1.377	26763.0	0.05660	1645.75	0.07792	47850318.4	0.10728	2265.66
3	5191.7	2.793	14766.3	0.07547	391.82	0.21082	2034219.1	0.58888	1094.49
4	21700.0	1.443	29900.0	0.09434	2047.17	0.13616	44423584.9	0.19653	2954.75
5	6808.0	1.788	12462.5	0.11321	770.72	0.20236	5247041.2	0.36172	1377.66
6	1800.0	1.785	3375.0	0.03774	67.92	0.06736	122264.2	0.12023	121.25
7	24481.5	1.323	38411.8	0.28302	6928.74	0.37436	169626245.6	0.49517	9164.83
8	294809.2	1.320	473455.2	0.16038	47280.72	0.21170	13938788309.1	0.27944	62410.54
9	6303.7	1.327	7480.3	0.09434	594.69	0.12516	3748699.4	0.16604	788.95
10	350.0	1.900	822.5	0.02830	9.91	0.05377	3467.0	0.10217	18.82
Total					59811.28	1.56942	14211940497.2	2.63050	80340.21

TABLE II
OPTIMUM CALIBRATED WEIGHTS

h	1	2	3	4	5	6	7	8	9	10
W_h^*	0.06274	0.05760	0.08673	0.09782	0.12318	0.04145	0.28986	0.07278	0.10026	0.03136

TABLE III
SAMPLE INFORMATION

h	$\bar{x}_{h1}$	$\bar{x}_{h2}$	$\bar{y}_h$	W_h	$W_h\bar{x}_{h1}$	$W_h\bar{x}_{h2}$	$W_h\bar{x}_{h1}^2$	$W_h\bar{x}_{h2}^2$	$W_h\bar{x}_{h1}\bar{x}_{h2}$
1	719082.2	2.037	14707.1	0.16667	119847	0.339	86179873719.3	0.691	244088.5
2	13190.3	1.640	19935.7	0.20833	2748	0.342	36246591.1	0.560	4506.7
3	20992.1	1.394	33021.5	0.33333	6997.4	0.465	146889097.8	0.648	9753.2
4	162587.8	1.427	262896.1	0.29167	47421.4	0.416	7710149597.8	0.594	67683.3
Total				1	170020.6	1.562	93926357687.4	2.493	326031.7

REFERENCES

[1] Briedt, F.J. and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Ann. Statist.*, 28, 1026–1053.
[2] Chen, J. and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika* 80, 107-116.
[3] Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, 87, 376–382.
[4] J.-M. Kim, E.A. Sungur, and T.-Y. Heo (2007), "Calibration Approach Estimators in Stratified Sampling", *Statistics & Probability Letters;* Vol. 77, 1, 99-103.
[5] Kim, J.K. (2009). Calibration estimation using empirical likelihood in unequal probability sampling. *Statist. Sinica*., 19, 145–157.
[6] Singh, S. (2003). *Advanced Sampling Theory with Applications*. Dordrecht: Kluwer Academic Publishers.
[7] Singh, S., Horn, S., Yu, F. (1998). Estimation of variance of the general regression estimator: higher level calibration approach. Survey Methodology 24, 41–50.
[8] Wu, C. & Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *J. Amer. Statist. Assoc.*, 96, 185–193.