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 \( E(y_i|x_i) \) [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 and combined ratio estimator using a 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 \( n_k \) be the number of units drawn by simple random sampling without replacement (SRSWOR) from the \( h \) th stratum consisting of \( N_k \) units, and \( N = \sum_{k=1}^{L} n_k \) and \( N = \sum_{k=1}^{L} N_k \) give the total sample size and the population size. For the \( h \) th strata, let \( W_h = N_k/N \) 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 \( y_{hij} \) and \( x_{hij} \) denote the values of the \( j \) th population (sampled) unit of the study variable (\( Y \)) and the
Assume that the strata means \( \bar{x}_j = \sum_{h=1}^{L} W_h \bar{x}_{hp} \) 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:

\[
\bar{y}_n = \sum_{h=1}^{L} W_h \bar{y}_h .
\]

In the presence of more than one auxiliary information, we suggest a multivariate calibrated estimator of the population mean under stratified sampling given by

\[
\bar{y}^c = \sum_{h=1}^{L} W^*_h \bar{y}_h
\]

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

\[
\sum_{j=1}^{p} \sum_{h=1}^{L} \frac{(W^*_h - W_h)^2}{W_h q_{hi}}
\]

is minimum, subject to the calibration constraints

\[
\sum_{h=1}^{L} W^*_h \bar{x}_{hi} = \bar{x}_j ; \quad j = 1, 2, \ldots, p.
\]

Note that \( q_{hi} > 0 \) 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

\[
W^*_h \geq 0 ; \quad h = 1, 2, \ldots, L.
\]

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

Minimize

\[
Z = \sum_{h=1}^{L} \frac{(W^*_h - W_h)^2}{W_h Q_h}.
\]

subject to

\[
\sum_{h=1}^{L} W^*_h \bar{x}_{hi} = \bar{x}_1, \quad \sum_{h=1}^{L} W^*_h \bar{x}_{h2} = \bar{x}_2, \quad \ldots
\]

and

\[
W^*_h \geq 0 ; \quad h = 1, 2, \ldots, L.
\]

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:

\[
L(W^*_h, \lambda_j) = \sum_{h=1}^{L} \frac{(W^*_h - W_h)^2}{W_h Q_h} - 2\sum_{j=1}^{p} \lambda_j \left( \sum_{h=1}^{L} W^*_h \bar{x}_{hi} - \bar{x}_j \right)
\]

The necessary conditions for the solution of the problem are

\[
\frac{\partial L}{\partial W^*_h} = 0, \quad \frac{\partial L}{\partial \lambda_j} = 0.
\]

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

\[
W^*_h = W_h + W_h Q_h \left( \lambda_1 \bar{x}_{h1} + \lambda_2 \bar{x}_{h2} \right)
\]

where,

\[
\lambda_1 = \frac{-C(\bar{x}_1 - \hat{x}_1) + B(\bar{x}_2 - \hat{x}_2)}{B^2 - AC},
\]

\[
\lambda_2 = \frac{B(\bar{x}_1 - \hat{x}_1) - A(\bar{x}_2 - \hat{x}_2)}{B^2 - AC},
\]

\[
\hat{x}_1 = \sum_{h=1}^{L} W^*_h \bar{x}_{h1}, \quad \hat{x}_2 = \sum_{h=1}^{L} W^*_h \bar{x}_{h2}.
\]
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:

\[
L = \sum_{k=1}^{L} W_k^* \ln \left( \frac{W_k^*}{W_k} \right) - \lambda_1 \left( \sum_{k=1}^{L} W_k^* x_{k1} - \bar{X}_1 \right) - \lambda_2 \left( \sum_{k=1}^{L} W_k^* x_{k2} - \bar{X}_2 \right).
\]

The necessary conditions given in (8) are

\[
\frac{\partial L}{\partial W_k^*} = 1 + \ln \left( W_k^* \right) - \ln \left( W_k \right) - \lambda_1 x_{k1} - \lambda_2 x_{k2} = 0,
\]

\[
\frac{\partial L}{\partial \lambda_1} = -\left( \sum_{k=1}^{L} W_k^* x_{k1} - \bar{X}_1 \right) = 0,
\]

and

\[
\frac{\partial L}{\partial \lambda_2} = -\left( \sum_{k=1}^{L} W_k^* x_{k2} - \bar{X}_2 \right) = 0.
\]

From (17) we have

\[
W_k^* = \exp \left[ \ln \left( W_k \right) + \lambda_1 x_{k1} + \lambda_2 x_{k2} - 1 \right].
\]

Using (20) and solving the equations (18) and (19), we obtain the values of \( \lambda_1 \) and \( \lambda_2 \).

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 \( Q_k = \sum_{j=1}^{L} q_{kj} = 1 \).

Table I shows the following sample information:

\[
\bar{X}_1 = \sum_{k=1}^{L} W_k^* x_{k1} = 59811.28,
\]

\[
\bar{X}_2 = \sum_{k=1}^{L} W_k^* x_{k2} = 80.
\]
Thus, the proposed multivariate calibrated weights are: $A = \sum_{k=1}^{L} W_k \bar{x}_i = 14211940497.2$, $B = \sum_{k=1}^{L} W_k \bar{x}_i \bar{x}_2 = 80340.21$ and $C = \sum_{k=1}^{L} W_k \bar{x}_i \bar{x}_2 = 2.63050$.

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), $\lambda_1 = -2.10924 \times 10^4$ and $\lambda_2 = 0.0573$. Thus, the multivariate calibrated weights $W_k^*$ in stratified sampling proposed in (9) is reduced to

$$W_k^* = W_k + A \left( -2.10924 \times 10^4 \bar{x}_1 + 0.0573 \bar{x}_2 \right)$$

which are obtained and presented in Table II.

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

$$\bar{Y}_p = \sum_{k=1}^{L} W_k \bar{Y}_k = 94666.73. \text{ (21)}$$

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

$$\bar{Y}_n = \sum_{k=1}^{L} W_k^* \bar{Y}_k = 53952.56. \text{ (22)}$$

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 $Q_k = \sum_{j=1}^{2} q_{kj} = 1$. 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.

$$\hat{\bar{Y}}_i = \sum_{k=1}^{L} W_k \bar{Y}_k = 170020.6,$$

$$\hat{\bar{Y}}_2 = \sum_{k=1}^{L} W_k \bar{X}_2 = 1.562,$$

$$A = \sum_{k=1}^{L} W_k Q_k \bar{X}_1 = 93926357687.4,$$

$$B = \sum_{k=1}^{L} W_k Q_k \bar{X}_1 \bar{X}_2 = 326031.7,$$

and $C = \sum_{k=1}^{L} W_k Q_k \bar{X}_1 \bar{X}_2 = 2.493$.

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), $\lambda_1 = -2.59723 \times 10^4$ and $\lambda_2 = 0.34163$. Thus, the multivariate calibrated weights $W_k^*$ 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 $\lambda_1 = -8.76261 \times 10^4$ and $\lambda_2 = 1.12097$. Thus, the proposed multivariate calibrated weights $W_k^*$ 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 mean $\bar{Y}$ 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**

<table>
<thead>
<tr>
<th>h</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$W_e$</th>
<th>$W_e x_1$</th>
<th>$W_e x_2$</th>
<th>$W_e x_3$</th>
<th>$W_e Q_1 x_1$</th>
<th>$W_e Q_1 x_2$</th>
<th>$W_e Q_1 x_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1304.7</td>
<td>1.940</td>
<td>2592.0</td>
<td>0.05660</td>
<td>73.85</td>
<td>0.10981</td>
<td>96348.4</td>
<td>0.21303</td>
<td>143.27</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>29075.0</td>
<td>1.377</td>
<td>26763.0</td>
<td>0.05660</td>
<td>1645.75</td>
<td>0.07792</td>
<td>47850318.4</td>
<td>0.10728</td>
<td>2265.66</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5191.7</td>
<td>2.793</td>
<td>14766.3</td>
<td>0.07547</td>
<td>391.82</td>
<td>0.21082</td>
<td>4423584.9</td>
<td>0.19653</td>
<td>2954.75</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>21700.0</td>
<td>1.443</td>
<td>29900.0</td>
<td>0.09434</td>
<td>2047.17</td>
<td>0.13616</td>
<td>164575</td>
<td>0.07792</td>
<td>47850318.4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6808.0</td>
<td>1.788</td>
<td>12462.5</td>
<td>0.11321</td>
<td>770.72</td>
<td>0.20236</td>
<td>5247041.2</td>
<td>0.36172</td>
<td>1377.66</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1800.0</td>
<td>1.785</td>
<td>3375.0</td>
<td>0.03774</td>
<td>67.92</td>
<td>0.06736</td>
<td>122264.2</td>
<td>0.12023</td>
<td>121.25</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>24481.5</td>
<td>1.323</td>
<td>38411.8</td>
<td>0.28302</td>
<td>6928.74</td>
<td>0.37436</td>
<td>169626245.6</td>
<td>0.49517</td>
<td>9164.83</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>294809.2</td>
<td>1.320</td>
<td>473455.2</td>
<td>0.16038</td>
<td>47280.72</td>
<td>0.21170</td>
<td>1393878309.1</td>
<td>0.27944</td>
<td>62410.54</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>6303.7</td>
<td>1.327</td>
<td>7480.3</td>
<td>0.09434</td>
<td>594.69</td>
<td>0.12516</td>
<td>3748699.4</td>
<td>0.16604</td>
<td>788.95</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>350.0</td>
<td>1.900</td>
<td>822.5</td>
<td>0.03774</td>
<td>9.91</td>
<td>0.05377</td>
<td>3467.0</td>
<td>0.10217</td>
<td>18.82</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>59811.28</td>
<td>1.56942</td>
<td>14211940497.2</td>
<td>2.63050</td>
<td>80340.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## TABLE II

**OPTIMUM CALIBRATED WEIGHTS**

<table>
<thead>
<tr>
<th>h</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_e$</td>
<td>0.06274</td>
<td>0.05760</td>
<td>0.08673</td>
<td>0.09782</td>
<td>0.12318</td>
<td>0.04145</td>
<td>0.28986</td>
<td>0.07278</td>
<td>0.10026</td>
<td>0.03136</td>
</tr>
</tbody>
</table>

## TABLE III

**SAMPLE INFORMATION**

<table>
<thead>
<tr>
<th>h</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$W_e$</th>
<th>$W_e x_1$</th>
<th>$W_e x_2$</th>
<th>$W_e x_3$</th>
<th>$W_e Q_1 x_1$</th>
<th>$W_e Q_1 x_2$</th>
<th>$W_e Q_1 x_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>719082.2</td>
<td>2.037</td>
<td>14707.1</td>
<td>0.16667</td>
<td>119847</td>
<td>0.339</td>
<td>86179873719.3</td>
<td>0.691</td>
<td>244088.5</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>13190.3</td>
<td>1.640</td>
<td>19935.7</td>
<td>0.20833</td>
<td>2748</td>
<td>0.342</td>
<td>36246591.1</td>
<td>0.560</td>
<td>4506.7</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>20992.1</td>
<td>1.394</td>
<td>33021.5</td>
<td>0.33333</td>
<td>6997.4</td>
<td>0.465</td>
<td>146889097.8</td>
<td>0.648</td>
<td>9753.2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>162587.8</td>
<td>1.427</td>
<td>262896.1</td>
<td>0.29167</td>
<td>47421.4</td>
<td>0.416</td>
<td>7710149597.8</td>
<td>0.594</td>
<td>67683.3</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td>170020.6</td>
<td>1.562</td>
<td>93926357687.4</td>
<td>2.493</td>
<td>326031.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## REFERENCES


