Estimation of Population Mean in Two Phase Sampling using Attribute Auxiliary Information

Nadeem Shafique Butt[1], Muhammad Qaiser Shahbaz[2]

[1] College of Statistical and Actuarial Sciences, University of the Punjab, Pakistan
nadeemshafique@hotmail.com

[2] Department of Mathematics, COMSATS Institute of Information Technology, Pakistan
qshahbaz@gmail.com

Abstract. A new estimator for population mean has been proposed in two phase sampling by using information of multiple auxiliary attributes. The minimum variance of the proposed estimator has been obtained.

1. Introduction

The auxiliary information has always been a source of improvement in estimation of certain population characteristics. Several estimators have been developed in single and two phase sampling which utilizes information on auxiliary variables as well as auxiliary attributes. The classical estimators which use information on auxiliary variables are the ratio and regression estimators as given in Hansen, Hurwitz, & Madow (1953). The classical regression estimator of population mean is given as:

$$\bar{y}_{lr} = \bar{y} + \beta\left(\bar{X} - \bar{x}\right) \tag{1.1}$$

The value of β for which the mean square error of (1.1) is minimum is $\beta = \dfrac{S_{xy}}{S_x^2}$. The minimum mean square error of (1.1) is given as

$$MSE(\bar{y}_{lr}) = \theta S_y^2\left(1 - \rho_{yx}^2\right) \tag{1.2}$$

Where $\theta = n^{-1} - N^{-1}$ and ρ_{yx} is the correlation coefficient between X and Y. The estimator (1.1) in case of auxiliary attribute is discussed by Naik & Gupta (1996), and the estimator in this case is given as:

$$\mathrm{t}_{1(1)} = \bar{\mathrm{y}} + \mathrm{b}\left(\mathrm{p}_1 - \mathrm{P}_1\right) \tag{1.3}$$

where P_1 is sample proportion for auxiliary variables. The mean square error of (1.3) is:

$$\mathrm{MSE}\left(\mathrm{t}_{1(1)}\right) = \theta\left(1 - \rho_{\mathrm{Pb}_1}^2\right)\mathrm{S}_{\mathrm{y}}^2 \tag{1.4}$$

where $\rho^2_{Pb_1}$ is the squared point bi-serial correlation coefficient. Jhajj, Sharma, & Grover (2006) has proposed a family of estimators in single and two phase sampling using information on a single auxiliary attributes. The proposed family is based upon a general function and is given as:

$$t_{2(1)} = g_\omega\left(\bar{y}, v_1\right) \tag{1.5}$$

where $v_1 = \frac{p_1}{P_1}$ and $g_\omega\left(\bar{y}, v_1\right)$ is a parametric function of $\bar{y}$ and v_1 such that $g_\omega\left(\bar{Y}, 1\right) = \bar{Y}$, for all $\bar{Y}$.

The mean square error of each estimator; to the terms of order $1/n$; of this family is,

$$\mathrm{MSE}\left(T_{2(1)}\right) \approx \theta\left(1-\rho^2_{Pb_1}\right)S_y^2 . \tag{1.6}$$

The mean square error of the proposed family depends upon the point bi-serial correlation coefficient.

Shabbir & Gupta (2007) have also proposed an estimator for population mean in single phase sampling using information of single auxiliary attribute. The estimator is given as:

$$t_{3(1)} = \left[d_1\bar{y} + d_2\left(P_1 - p_1\right)\right]\frac{P_1}{p_1}, \text{ for } \quad p_1 > 0 \tag{1.7}$$

where d_1 and d_2 are unknown constants. The mean square error of (1.7) is:

$$\mathrm{MSE}\left(t_{3(1)}\right) \approx \frac{\theta\left(1-\rho^2_{Pb_1}\right)S_y^2}{1+\theta\left(1-\rho^2_{Pb_1}\right)C_y^2} \tag{1.8}$$

In this paper we have proposed a modified regression type estimator using information on several auxiliary attributes.

2. Notations

In this section we define the notations used for the development of the estimator and its variance. Let $\boldsymbol{\delta}$ be a vector of auxiliary attributes with covariance matrix $\mathbf{S}_{\boldsymbol{\delta}}$, τ be another auxiliary attribute and Y be the variable of interest.

Let $\mathbf{s}_{\tau\boldsymbol{\delta}}$ be the vector of covariances between τ and $\boldsymbol{\delta}$, $\mathbf{s}_{y\boldsymbol{\delta}}$ be the vector of covariances between Y and $\boldsymbol{\delta}$. Using these notations we define $\boldsymbol{\gamma} = \mathbf{S}_{\boldsymbol{\delta}}^{-1}\mathbf{s}_{\tau\boldsymbol{\delta}}$ as

vector of regression coefficients between τ and $\boldsymbol{\delta}$, and $\boldsymbol{\gamma} = \mathbf{S}_{\boldsymbol{\delta}}^{-1}\mathbf{s}_{y\boldsymbol{\delta}}$ as vector of regression coefficients between Y and $\boldsymbol{\delta}$. We also define $\beta_{y\tau.\boldsymbol{\delta}} = S_{\tau y.\boldsymbol{\delta}}/S^2_{\tau.\boldsymbol{\delta}}$ as partial regression coefficient between Y and τ keeping the $\boldsymbol{\delta}$ at constant level. Also $S_{y\tau.\boldsymbol{\delta}} = S_{y\tau} - \mathbf{s}'_{\tau\boldsymbol{\delta}}\mathbf{S}_{\boldsymbol{\delta}}^{-1}\mathbf{s}_{\tau\boldsymbol{\delta}}$ is partial covariance between Y and τ after removing the effect of $\boldsymbol{\delta}$, $\mathbf{S}^2_{y.\boldsymbol{\delta}} = S^2_y - \mathbf{s}'_{\tau\boldsymbol{\delta}}\mathbf{S}_{\boldsymbol{\delta}}^{-1}\mathbf{s}_{\tau\boldsymbol{\delta}}$ is the partial variance of Y, and $S^2_{\tau.\boldsymbol{\delta}} = S^2_{\tau} - \mathbf{s}'_{\tau\boldsymbol{\delta}}\mathbf{S}_{\boldsymbol{\delta}}^{-1}\mathbf{s}_{\tau\boldsymbol{\delta}}$ is the partial variance of τ. We also define $\rho^2_{y\tau.\boldsymbol{\delta}} = S^2_{y\tau}/\left(S^2_{\tau.\boldsymbol{\delta}}S^2_{y.\boldsymbol{\delta}}\right)$ as partial correlation coefficient between Y and τ after removing the effect of $\boldsymbol{\delta}$, $\rho^2_{y\tau.\boldsymbol{\delta}}$ as squared multiple bi-serial correlation coefficient between Y and combined effects of τ and $\boldsymbol{\delta}$, $\rho^2_{y.\boldsymbol{\delta}}$ as squared multiple correlation coefficient between Y and combined effects of $\boldsymbol{\delta}$.
Using the above notations we proposed the new estimators in the section 3.

3. The Proposed Estimator

We propose following unbiased estimator of population mean in two phase sampling using information of several auxiliary attributes:

$$t_{nss(A)} = \bar{y}_2 + k\left[p_{\tau_1} + \boldsymbol{\gamma}'\left(\mathbf{p}_{\delta} - \mathbf{p}_{\delta_1}\right) - \left\{p_{\tau_2} + \boldsymbol{\eta}'\left(\mathbf{p}_{\delta} - \mathbf{p}_{\delta_2}\right)\right\}\right] \tag{3.1}$$

Using
$\bar{y}_2 = \bar{Y} + \bar{e}_{y_2}$, $p_{\tau_1} = p_{\tau} + \bar{e}_{\tau_1}$, $p_{\tau_2} = p_{\tau} + \bar{e}_{\tau_2}$, $\mathbf{p}_{\delta_1} = \mathbf{p}_{\delta} + \bar{\mathbf{e}}_{\delta_1}$ and $\mathbf{p}_{\delta_2} = \mathbf{p}_{\delta} + \bar{\mathbf{e}}_{\delta_2}$ in (3.1) we have:

$$t_{nss(A)} - \bar{Y} = \bar{e}_{y_2} + k\left[\left(\bar{e}_{\tau_1} - \bar{e}_{\tau_2}\right) - \boldsymbol{\gamma}'\bar{\mathbf{e}}_{\delta_1} + \boldsymbol{\eta}'\bar{\mathbf{e}}'_{\delta_2}\right]$$

Squaring and applying expectation, the mean square error of (3.1) is given as:

$$\begin{aligned} S = MSE\left(t_{nss}\right) = \theta_2 s_y^2 + k^2\left[\left(\theta_2 - \theta_1\right)s_{\tau}^2 + \theta_1\boldsymbol{\gamma}'\mathbf{S}_{\boldsymbol{\delta}}\boldsymbol{\gamma} + \theta_2\boldsymbol{\eta}'\mathbf{S}_{\boldsymbol{\delta}}\boldsymbol{\eta} + 2\left(\theta_1 - \theta_2\right)\boldsymbol{\eta}'\mathbf{s}_{\tau\boldsymbol{\delta}}\right. \\ \left. -2\theta_1\boldsymbol{\gamma}'\mathbf{S}_{\boldsymbol{\delta}}\boldsymbol{\eta}\right] + 2k\left[\left(\theta_1 - \theta_2\right)s_{y\tau} - \theta_1\boldsymbol{\gamma}'\mathbf{s}_{y\boldsymbol{\delta}} + \theta_2\boldsymbol{\eta}'\mathbf{s}_{y\boldsymbol{\delta}}\right] \end{aligned} \tag{3.2}$$

The optimum values of $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$ and k are obtained by minimizing (3.2). These values are obtained by solving following three equations, obtained by partially differentiating (3.2) and setting the derivative to zero

$$2k\left[\left(\theta_2-\theta_1\right)s_\tau^2+\theta_1\boldsymbol{\gamma}'\mathbf{S}_{\boldsymbol{\delta}}\boldsymbol{\gamma}+\theta_2\boldsymbol{\eta}'\mathbf{S}_{\boldsymbol{\delta}}\boldsymbol{\eta}+2\left(\theta_1-\theta_2\right)\boldsymbol{\eta}'\mathbf{s}_{\tau\boldsymbol{\delta}}-2\theta_1\boldsymbol{\gamma}'\mathbf{S}_{\boldsymbol{\delta}}\boldsymbol{\eta}\right]$$
$$+2\left[\left(\theta_1-\theta_2\right)s_{y\tau}-\theta_1\boldsymbol{\gamma}'\mathbf{s}_{y\boldsymbol{\delta}}+\theta_2\boldsymbol{\eta}'\mathbf{s}_{y\boldsymbol{\delta}}\right]=0 \quad (1)$$

$$k\mathbf{S}_{\boldsymbol{\delta}}\left(\boldsymbol{\gamma}-\boldsymbol{\eta}\right)-\mathbf{s}_{y\boldsymbol{\delta}}=0 \quad (2)$$

$$k\mathbf{S}_{\boldsymbol{\delta}}\left(\theta_2\boldsymbol{\eta}-\theta_1\boldsymbol{\gamma}\right)-k\left(\theta_2-\theta_1\right)\mathbf{s}_{\tau\boldsymbol{\delta}}+\theta_2\mathbf{s}_{y\boldsymbol{\delta}}=\mathbf{0} \quad (3)$$

Solving the above equations simultaneously, the optimum values of $\boldsymbol{\alpha},\boldsymbol{\beta}$ and k are:

$$\boldsymbol{\gamma}=S_{\boldsymbol{\delta}}^{-1}s_{x\boldsymbol{\delta}},\ \boldsymbol{\eta}=\boldsymbol{\gamma}-k^{-1}\boldsymbol{\gamma} \text{ and } k=\boldsymbol{\eta}_{y\tau.\boldsymbol{\delta}}=\frac{s_{y\tau.\boldsymbol{\delta}}}{s_{\tau\tau.\boldsymbol{\delta}}}$$

Using the optimum values in (3.2) and simplifying, the mean square error of proposed estimator is:

$$MSE\left(t_{nss}\right)=s_{y.\boldsymbol{\delta}}^2\left[\theta_2\left(1-\rho_{\tau y.\boldsymbol{\delta}}^2\right)+\theta_1\rho_{\tau y.\boldsymbol{\delta}}^2\right] \quad (3.3)$$

Further, by using the fact that $s_{y.\boldsymbol{\delta}}^2=S_y^2\left(1-\rho_{\tau y.\boldsymbol{\delta}}^2\right)$ and utilizing the relationship that $1-\rho_{y.\tau\boldsymbol{\delta}}^2=\left(1-\rho_{y.\boldsymbol{\delta}}^2\right)\left(1-\rho_{y\tau.\boldsymbol{\delta}}^2\right)$, the mean square error of proposed estimator can be written as:

$$MSE\left(t_{nss}\right)=s_y^2\left\{\theta_2\left(1-\rho_{y.\tau\boldsymbol{\delta}}^2\right)+\theta_1\rho_{\tau y.\boldsymbol{\delta}}^2\left(1-\rho_{y.\boldsymbol{\delta}}^2\right)\right\} \quad (3.4)$$

From (3.4) we can see that the mean square error of (3.1) depends upon the squared multiple and partial correlation coefficients. The estimator and its mean square error for multiphase sampling can be analogously written from (3.1) and (3.4). Specifically if a sample of size n_h is taken at h^{th} phase and a sample of n_q is taken at q^{th} phase with $n_q<n_h$, the estimator of the population mean is:

$$t_{nss(A)}=\bar{y}_2+k\left[p_{\tau_h}+\boldsymbol{\gamma}'\left(\mathbf{p}_\delta-\mathbf{p}_{\delta_h}\right)-\left\{p_{\tau q}+\boldsymbol{\eta}'\left(\mathbf{p}_\delta-\mathbf{p}_{\delta_q}\right)\right\}\right] \quad (3.5)$$

The mean square error of (3.5) can be written from (3.4) as:

$$MSE\left(t_{nss}\right)=s_y^2\left\{\theta_q\left(1-\rho_{y.\tau\boldsymbol{\delta}}^2\right)+\theta_h\rho_{\tau y.\boldsymbol{\delta}}^2\left(1-\rho_{y.\boldsymbol{\delta}}^2\right)\right\} \quad (3.6)$$

For practical applicability, the proposed estimator can be easily modified by using the sample estimates in place of population parameters. The consistent estimate of population mean can be straight-away written as:

$$t_{nss}=\bar{y}_2+b_{y\tau.\boldsymbol{\delta}}\left(p_{\tau_1}-p_{\tau_2}\right)+b_{y\tau.\boldsymbol{\delta}}\mathbf{b}_{\tau\boldsymbol{\delta}}\left(\mathbf{p}_{\delta_2}-\mathbf{p}_{\delta_1}\right)+\mathbf{b}_{y\boldsymbol{\delta}}^{/}\left(\mathbf{p}_\delta-\mathbf{p}_{\delta_2}\right) \quad (3.7)$$

The estimated standard error of (3.1) is given as:

$$S.E\left(t_{nss}\right) = s_y\sqrt{\theta_2\left(1-r^2_{y.\tau\boldsymbol{\delta}}\right)+\theta_1 r^2_{y\tau.\boldsymbol{\delta}}\left(1-r^2_{y.\boldsymbol{\delta}}\right)} \quad (3.8)$$

Using (3.7) and (3.8), the confidence interval for true population mean can be constructed.

References:

1. Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample Survey Methods and Theory* (Vol. II): John Wiley.
2. Jhajj, H. S., Sharma, M. K., & Grover, L. K. (2006). A familty of estimators of population mean using information of auxiliary attribute. *Pak. J. Stat., 22*(1), 43-50.
3. Naik, V., & Gupta, P. (1996). A note on estimation of mean with known population proportion of an auxiliary character. *Jour. Ind. Soc. Agr. Stat, 48*(2), 151-158.
4. Shabbir, J., & Gupta, S. (2007). On estimating the finite population mean with known population proportion of an auxiliary variable. *Pak. J. Statist., 23*(1), 1-9.