International Journal of Pure and Applied Mathematics
Volume 85 No. 2 2013, 285-295
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
doi: http://dx.doi.org/10.12732/ijpam.v85i2.8

A CLASS OF PREDICTIVE ESTIMATORS IN TWO-STAGE SAMPLING WHEN AUXILIARY CHARACTER IS ESTIMATED AT SSU LEVEL

Monika Saini

School of Mathematics & Computer Applications
Thapar University
Patiala, 147001, INDIA

Abstract: This paper presents a class of predictive estimators for two stage sampling with unequal first and second stage units to the case where information on auxiliary character is not available. A double sampling procedure is proposed as alternative under such a situation. The proposed class consists of mainly two estimators namely ratio and regression. The mean square error (MSE) and minimum mean square of this class have been derived. In addition, we support these theoretical results by an empirical study.

AMS Subject Classification: 62D05
Key Words: double sampling, finite population, predictive estimators, two-stage sampling

1. Introduction

Auxiliary information has always been effective in increasing the precision of estimates in survey sampling. It is common practice to make use of auxiliary information on a auxiliary character x for the estimation of the finite population mean or total of a character under study. Laplace (1820) was the first to use

Received: November 18, 2012

the auxiliary information in ratio type estimator. Watson (1937) used regression method of estimation to estimate the average area of the leaves on a plant. The auxiliary information has been effectively used in two-phase sampling to estimate a population characteristic. A variety of approaches are available to construct more efficient estimators for the population mean and total, including design based and model based methods. In a predictive approach a model is specified for the population values and is used to predict the non sampled values. Prediction theory for sampling surveys (or model-based theory) can be considered as a general framework for statistical inferences on the character of finite population. Well-known estimators of population totals encounter in the classical theory, as expansion, ratio, regression another estimators can be predictors in a general prediction theory under some special model. Srivastave (1983) used the predictive concept for the product estimator. Scott and Smith (1968) described the predictive for two stage sampling. More recently Hossian and Ahmed (2001) suggested some predictive estimators with the use of auxiliary variable. In this paper we propose a class of predictive estimators for two stage design when auxiliary character is estimated to construct ratio and regression estimators by utilizing the prediction criterion.

2. Two Stage Sampling Set-Up and Notations

Consider a finite population U be partitioned into N first stage units (fsu) denoted by $(U_1, \ldots, U_i, \ldots, U_N)$ such that the number of second stage units (ssu) in U_i is M_i and $M = \sum_{i}^{N} M_i$. Let y_{ij} and x_{ij} be the value of the survey variable y and a correlated auxiliary variable x respectively for the jth ssu of U_i $(j = 1, 2, \ldots M_i,\ i = 1, 2, \ldots, N)$. We want to use the auxiliary information at estimation stage to improve the precision of an estimator but some time the required auxiliary information is not available at ssu level, a double sampling procedure is proposed as an alternative under such a situation for the estimation of population mean. Assume a sample s of n fsu is drawn from N fsu by using SRSWOR, then we select a sample s_i' of m_i' ssu from the ith selected fsu U_i by using SRSWOR and information of an auxiliary variable is collected. Further from the selected m_i' ssu a sample of m_i ssu is selected by using SRSWOR.

According to the sampling set-up we define symbolically

$$\bar{y} = \frac{1}{n}\sum_{i\in s} W_i \bar{y}_i, \quad \bar{x} = \frac{1}{n}\sum_{i\in s} W_i \bar{x}_i \quad \text{and} \quad \bar{x}' = \frac{1}{n}\sum_{i\in s} W_i \bar{x}_i'$$

where

$$\bar{y}_i = \frac{1}{m_i}\sum_{j\in s_i} y_{ij}, \quad \bar{x}_i = \frac{1}{m_i}\sum_{j\in s_i} x_{ij} \quad \text{and} \quad \bar{x}'_i = \frac{1}{m_i}\sum_{j\in s_i} x_{ij}$$

also

$$w_i = \frac{M_i}{M} = \frac{NM_i}{M}, \quad M = \sum_i^N M_i \quad \text{and} \quad \bar{M} = \frac{1}{N}\sum_i^N M_i$$

3. Prediction Criterions and Proposed Class of Estimators

When information on auxiliary character for U_i is unknown and is collected by using double sampling, then the classical two-stage ratio and regression estimators of $\bar{y}$ and their respective first order approximate mean square errors (MSE) are given below:
Consider the Ratio estimator

$$(d_1) = \frac{\bar{y}}{\bar{x}}\bar{x}' \tag{1}$$

Then MSE

$$\begin{aligned}(d_1) = \left(\frac{1-f}{n}\right)\{S_y^2 - 2RS_{yx} + R^2S_x^2\} + \frac{1}{nN}\sum_i^N w_i^2\left(\frac{1-f'_i}{m'_i}\right)S_{iy}^2 \\ + \frac{1}{nN}\sum_i^N w_i^2\left(\frac{1-f_i}{m_i}\right)\{S_{iy}^2 - 2RS_{iyx} + R^2S_{ix}^2\}\end{aligned} \tag{2}$$

Consider the Regression estimator

$$\begin{aligned}(d_2) &= \bar{y} + b(\bar{x}' - \bar{x}) \\ &\approx \bar{y} + B(\bar{x}' - \bar{x})\end{aligned} \tag{3}$$

Then MSE

$$\begin{aligned}(d_2) = \left(\frac{1-f}{n}\right)\{S_y^2 - 2BS_{yx} + B^2S_x^2\} + \frac{1}{nN}\sum_i^N w_i^2\left(\frac{1-f'_i}{m_i}\right)S_{iy}^2 \\ + \frac{1}{nN}\sum_i^N w_i^2\left(\frac{1-f_i}{m_i}\right)\{S_{iy}^2 - 2BS_{iyx} + B^2S_{ix}^2\}\end{aligned} \tag{4}$$

where

$$R = \frac{\bar{y}}{\bar{x}}, \quad B = \frac{S_{yx}}{S_x^2}, \quad f = \frac{n}{N} f_i' = \frac{m_i'}{M_i} \quad \text{and} \quad f_i = \frac{m_i}{m_i}$$

$$S_{yx} = \frac{1}{N-1} \sum_{i}^{N} (w_i \bar{Y}_i - \bar{Y})(w_i \bar{X}_i - \bar{X})$$

$$S_{iyx} = \frac{1}{M_i - 1} \sum_{j}^{M_i} (y_{ij} - \bar{Y}_i)(x_{ij} - \bar{X}_i)$$

and S_y^2, S_x^2, S_{iy}^2 and S_{ix}^2 can be obtained from S_{yx} and S_{iyx} using $y = x$.
Under the usual predictive set-up, it possible to express, for a given non empty set s, we can partition

$$\begin{aligned}
\bar{Y} &= \frac{\sum_{i}^{N} \sum_{j}^{M_i} y_{ij}}{\sum_{i}^{N} M_i} = \frac{\sum_{i}^{N} M_i \bar{y}_i}{\sum_{i}^{N} M_i} \\
&= \frac{1}{M} \left[\sum_{i}^{N} M_i \bar{y}_i \right] \\
&= \frac{1}{M} \left[\sum_{i \in s} M_i \bar{Y}_i + \sum_{i \in \bar{s}} M_i \bar{y}_i \right] \\
&= \frac{1}{M} \left[\sum_{i \in s} \left\{ \sum_{j \in S_j} y_{ij} + \sum_{j \in s_i} y_{ij} \right\} + \sum_{i \in \bar{s}} M_i \bar{Y}_i \right] \\
&= \frac{1}{M} \left[\sum_{i \in s} \left\{ \sum_{j \in S_j'} \left\{ \sum_{j \in s_i} y_{ij} + \sum_{j \in \bar{s}_i} y_{ij} \right\} + \sum_{j \in \bar{s}_j} \right\} + \sum_{i \in \bar{s}} M_i \bar{y}_i \right] \qquad (5)
\end{aligned}$$

where $\bar{s}$ denotes the set of $(N - n)$ fsu's in U which are not included in s, $\bar{s}_i'$ denote the set of $(M_i - m_i')$ first phase of ssu in U_i which are not included in s_i', $\bar{s}_i$ the set of $(m_i' - m_i)$ second phase of U_i which are not included in s_i, $i = 1, 2, \ldots, n$.

$$\bar{y} = \frac{1}{M} \left[\sum_{i \in s} \left\{ \sum_{j \in S_j'} \left\{ \sum_{j \in s_i} y_{ij} + \sum_{j \in \bar{s}} j y_{ij} \right\} + \sum_{j \in \bar{S}_i} y_{ij} \right\} + \sum_{i \in \bar{s}} M_i \bar{Y}_i \right]$$

Let

$$\bar{Y}_r = \frac{1}{N-n} \sum_{i \in \bar{s}} W_i \bar{Y}_i, \quad \bar{Y}_{ir}' = \frac{1}{M_i - m_i'} \sum_{j \in \bar{s}_i} Y_{ij},$$

$$\bar{Y}_{ir} = \frac{1}{m_i - m_i} \sum_{j \in \bar{s}_{\mathrm{i}}} Y_{ij}$$

we have

$$\bar{Y} = \frac{1}{M}\Bigg[\sum_{i \in s}\Bigg\{\sum_{j \in S'_{\mathrm{i}}}\Bigg\{\sum_{j \in s_{\mathrm{i}}} m_i \bar{y}_i + (m'_i - m_i)\bar{Y}_{ir} + (M_i - m'_i)\bar{Y}_{ir}\Bigg\} + \frac{N-n}{N}\bar{Y}_r\Bigg\}\Bigg] \quad (6)$$

where $\bar{Y}_r = \dfrac{N\bar{Y} - n\bar{y}}{N-n}$, $\bar{Y}'_{ir} = \dfrac{M_i\bar{Y}_i - m'_i, \bar{y}'_i}{M_i - m'_i}$ and $\bar{Y}_{ir} = \dfrac{m'_i\bar{y}'_i - m_i\bar{y}_i}{m'_i - m_i}$.

To estimate $\bar{Y}$ we have to predict the quantities $\bar{Y}_{ir}$, $\bar{Y}'_{ir}$ and $\bar{Y}_r$ from the sample data because the first component of the right hand side of (3.6) is already known. Using $\bar{Z}_{ir}$, $\bar{Z}'_{ir}$ and $\bar{Z}_r$ as their respective predictors.

Then the predictive estimator of the population mean $\bar{Y}$ is

$$(\hat{\bar{Y}})_{\mathrm{pre}} = \frac{1}{M}\Bigg[\sum_{i \in s}\Bigg\{\sum_{j \in s_{\mathrm{i}}}\Bigg\{\sum_{j \in s_{\mathrm{i}}} m_i \bar{y}_i + (m'_i - m_i)Z_{ir} + (M_i - m_i)Z'_{ir}\Bigg\} + \frac{N-n}{N}Z_r\Bigg\}\Bigg] \quad (7)$$

where Z_r as the predictor of $\bar{Y}_r$ for first stage unit, Z'_{ir} and Z_{ir} are the predictors for second stage unit using double sampling respectively.

In equation (3.7) we combine the last second and third terms i.e. Non sampled part of first phase and second phase in second stage unit because in the second stage units using double sampling when we go to from one phase to second sampling unit is not changed and non sampled units are $M_i - m'_i + m'_i - m_i = M_i - m_i$.

Equation (3.7) can be written as

$$(\hat{\bar{Y}})_{\mathrm{pre}} = \frac{1}{M}\Bigg[\sum_{i \in s}\Bigg\{\sum_{j \in s_{\mathrm{j}}}\Bigg\{\sum_{j \in s_{\mathrm{i}}} m_i \bar{y}_i + (M_i - m_i)Z^*_{ir} + \frac{N-n}{N}Z_r\Bigg\}\Bigg] \quad (8)$$

Define Z_r and Z^*_{ir} as the class of estimators for first stage unit and second stage unit using single auxiliary variable x for second stage unit using double sampling.

Then we proposed as

$$\left.\begin{aligned} Z_r &= \bar{y}_r + t(\bar{X}_r - \bar{x}) \\ Z^*_{ir} &= \bar{y}_{ir} + t_i(\bar{x}'_{ir} - \bar{x}_i) \end{aligned}\right\} \quad (9)$$

Estimator	Different values			
	t	t_1	T	T_{i}
Ratio	$\frac{\bar{y}}{\bar{x}}$	$\frac{\bar{y}_{\text{i}}}{\bar{x}_{\text{i}}}$	$\frac{\bar{Y}}{\bar{X}}$	$\frac{\bar{Y}_{\text{i}}}{\bar{X}_{\text{i}}}$
Regression	b	b_{i}	B	B_{i}

Table 1

and t and t_i are suitably chosen statistics for defining ratio and regression estimators. Assume that $E(t) = T$ or $E(t) \approx T$ and $E(t_i) = E_1E_2(t_i) = T_i$ or $E(t_i) \approx T_i$. The choices of t, t_i, T and T_i for different predictive estimators are presented in the following Table 1.

Now for first order approximation the MSE (3.9) are given respectively

$$M(Z_r) = \left(\frac{1-f}{n}\right)\{S_y^2 - 2T\rho S_y S_x + T^2 S_x^2\} \tag{10}$$

$$M(Z_{ir}^*) = \left(\frac{1-f_i'}{m_i'}\right)S_{iy}^2 + \left(\frac{1-f_i}{m_i}\right)\{S_{iy}^2 - 2T_i\rho S_{iy}S_{ix} + T_i^2 S_{ix}^2\} \tag{11}$$

where equation (3.10) is the mean square error for first stage unit and equation (3.11) is the mean square error for second stage unit double sampling respectively.

If we put different value of T and T_i are given below then we get the various types of estimators.

Ratio Estimator:

$$M(Z_r) = \left(\frac{1-f}{n}\right)\{S_y^2 - 2R\rho S_y S_x + R^2 S_x^2\}$$

$$M(Z_{ir}^*) = \left(\frac{1-f_i'}{m_i}\right)S_{iy}^2 + \left(\frac{1-f_i}{m_i}\right)\{S_{iy}^2 - 2R_i\rho S_{iy}S_{ix} + R_i^2 S_{ix}^2\}$$

Regression Estimator:

$$M(Z_r) = \left(\frac{1-f}{n}\right)\{S_y^2 - 2R\rho S_y S_x + R^2 S_x^2\}$$

$$M(Z_{ir}^*) = \left(\frac{1-f_i'}{m_i}\right)S_{iy}^2 + \left(\frac{1-f_i}{m_i}\right)\{S_{iy}^2 - 2R_i\rho S_{iy}S_{ix} + R_i^2 S_{ix}^2\}$$

From equation (3.8) the proposed class of predictive estimators is

$$(\bar{Y})_{\text{pre}} = \frac{1}{M}\left[\sum_{i\in s}\sum_{j\in s_{\text{i}}}\sum_{j\in s} M_i(\bar{y}_i + t_i(\bar{x}'_i - \bar{x}_i))\right] + (\bar{y} + t(\bar{X} - \bar{x}))$$

From table I after simple algebraic manipulations $(\bar{Y})_{\text{pre}}$ turns out to be ratio and regression estimators respectively:

$$d_{01} = d_1 + \frac{1}{N}\sum_i w_i\left(\frac{\bar{y}_i}{\bar{x}_i} - \frac{\bar{y}}{\bar{x}}\right)\bar{x}'_i$$

$$d_{02} = d_2 = \frac{1}{N}\sum_i w_i[(\bar{y}_i - b_i(\bar{x}'_i - \bar{x}_i)) + (\bar{y} - b(\bar{X} - \bar{x}))]$$

4. The Efficiency of Estimators

After some simplification, we have the first order of mean square error of $(\hat{\bar{Y}})_{\text{pre}}(d_{01}, d_{02})$ is as follows:

$$M(\hat{\bar{Y}})_{\text{pre}} = \left(\frac{1-f}{n}\right)\{S_y^2 + T^2S_x^2 - 2TS_{yx}\} + \frac{1}{nN}\sum_i^N w_i^2\left(\frac{1-f'_i}{m_i}\right)S_{iy}^2$$
$$+\frac{1}{nN}\sum_i^N w_i^2\left(\frac{1-f_i}{m_i}\right)\{S_{iy}^2 + \alpha_i^2S_{ix}^2 - 2\alpha_iS_{iyx}\} \qquad (12)$$

where $\alpha_i = T - f(T - T_i)$.

The optimum values of T and T_i which minimize $M(\hat{\bar{Y}})_{\text{pre}}$ are given respectively as

$$T_{\text{opt}} = \frac{Syx}{S_x^2} = B_b \quad \text{and} \quad T_{\text{opt}} = \frac{S_{iyx}}{S_{ix}^2} = B_{ib}$$

Then the minimum mean square error of

$$M(\hat{\bar{Y}})_{\text{opt}} = \left(\frac{1-f}{n}\right)\{S_y^2 + T_{\text{opt}}^2S_x^2 - 2T_{\text{opt}}S_{yx}\} + \frac{1}{nN}\sum_i^N w_i^2\left(\frac{1-f'_i}{m'_i}\right)S_{iy}^2$$
$$+\frac{1}{nN}\sum_i^N w_i^2\left(\frac{1-f_i}{m_i}\right)\{S_{iy}^2 + \alpha_{iopt}^2S_{ix}^2 - 2\alpha_{iopt}S_{iyx}\}$$

$$M(\hat{\bar{Y}})_{\text{opt}} = \left(\frac{1-f}{n}\right)\{S_y^2 + T_{\text{opt}}^2 S_x^2 - 2T_{\text{opt}} S_{yx}\} + \frac{1}{nN}\sum_i^N w_i^2 \left(\frac{1-f_i'}{m_i'}\right) S_{iy}^2$$
$$+ \frac{1}{nN}\sum_i^N w_i^2 \left(\frac{1-f_i}{m_i}\right)\{S_{iy}^2 + \alpha_{iopt}^2 S_{ix}^2 - 2\alpha_{iopt} S_{iyx}\}$$

where $\alpha_{\text{opt}} = A_\alpha$.

where is the mean square error of regression estimator. Hence regression estimator is the optimum estimator of this class.

The minimum mean square error of d_{01} and d_{02} up to the first order approximation are given respectively as follows:

Ratio Estimator:

$$M(d_{01}) = \left(\frac{1-f}{n}\right)\{S_y^2 + R^2 S_x^2 - 2RS_{yx}\}$$
$$+ \frac{1}{nN}\sum_i^N w_i^2 \left(\frac{1-f_i'}{m_i'}\right) S_{iy}^2$$
$$+ \frac{1}{nN}\sum_i^N w_i^2 \left(\frac{1-f_i}{m_i}\right)\{S_{iy}^2 + \alpha_i^2 S_{ix}^2 - 2\alpha_i S_{iyx}\} \tag{13}$$

Regression Estimator:

$$M(d_{02}) = \left(\frac{1-f}{n}\right)\{S_y^2 + R^2 S_x^2 - 2RS_{yx}\}$$
$$+ \frac{1}{nN}\sum_i^N w_i^2 \left(\frac{1-f_i'}{m_i'}\right) S_{iy}^2$$
$$+ \frac{1}{nN}\sum_i^N w_i^2 \left(\frac{1-f_i}{m_i}\right)\{S_{iy}^2 + \alpha_i^2 S_{ix}^2 - 2\alpha_i S_{iyx}\} \tag{14}$$

where $\alpha_i = R - f(R - R_i)$ and $\gamma_i = B - f(B - B_i)$.

Subtracting the equation (4.2) and (4.3) from the equation (3.2) and (3.4) respectively, we get
Difference:

$$D_j = MSE(d_j) - MSE(d_{\circ j}), \quad j = 1, 2$$
$$D_1 = \frac{1}{nN}\sum_i^N w_i^2 \left(\frac{1-f-i}{m_i}\right)\{(R^2 - \alpha_i^2) S_{ix}^2 - 2(R - \alpha_i) S_{ixy}\}$$

$$D_2 = \frac{1}{nN}\sum_{i}^{N} w_i^2 \left(\frac{1-f-i}{m_i}\right)\{(B^2-\alpha_i^2)S_{ix}^2 - 2(B-\gamma_i)S_{ixy}\}$$

Thus d_{01} and d_{02} will be more efficient than d_1 and d_2 respectively

If $\beta_{iyx} \leq \dfrac{R+\alpha_i}{2}, \qquad i = 1, 2, \ldots, N$ for ratio

$\delta_{iyx} \leq \dfrac{B+\gamma_i}{2}$ for regression

where $\beta_{iyx} = \dfrac{S_{iyx}}{S_{ix}^2}$ and $\delta_{iyx} = \dfrac{S_{iyx}}{S_{ix}^2}$ be the regression coefficient of y on x in U_i.

5. Empirical Study

To show the usefulness of suggested methodology presented in this paper, numerically, an empirical study has been carried out.

Population: We consider the 2001 census data which relates to the total number of agricultural laboures and the total no. of cultivators of 444 villages of Bhiwani district of Haryana. We take no. of agricultural laboures in villages as y and the total no. of cultivators in villages as x. The whole population of Bhiwani district (444 villages) is divided into 9 blocks (fsus) where ith $(i = 1, 2, \ldots, 9)$ block consists of M_i villages (ssus) $i = 1, 2, \ldots 9$, $(M_1 = 25, M_2 = 56, M_3 = 32, M_4 = 45, M_5 = 36, M_6 = 78, M_7 = 66. M_8 = 55, M_9 = 51)$. The numerical values of the estimate of population mean, its variance were worked out under each of the proposed estimators from population values. Subsequently, the relative efficiency of the estimators is also calculated. Variance and percent relative efficiency of different estimators using auxiliary information in two stage design are obtained in Table 3.

6. Conclusion

Section 4 provides the conditions under which the estimators d_{01} and d_{02} have less mean squared error as compared to d_1 and d_2 when auxiliary character is estimated in two stage design.

Source	Village wise information of Bhiwani District of Haryana (2001 census data)
y	study variable (agricultural labourers)
x	auxiliary information (total number of cultivators)

Table 2: Description of population

(i)	(ii)	(iii)	(iv)
Estimators	Auxiliary variable used	Variance $\times 10^5$	R.E% w.r.t. d_1
d_1	x	2.55	100
d_{01}	x	2.45	104
			R.E% w.r.t. d_2
d_2	x	2.62	100
d_{02}	x	2.49	105

Table 3

Section 5, Table 3 shows that the estimators d_{01} and d_{02} has highest percent relative efficiency w.r.t. to d_1 and d_2. Thus the estimator d_{01} and d_{02} are recommended for used in practice for the estimation of population mean.

References

[1] M.I. Hossian, M.S. Ahmed, A class of Predictive estimators in Two-Stage sampling using Auxiliary information, *Information and Management Sciences*, 12(1) (2001), 49-55.

[2] P.S. Lapalace, *A Philosophical Essay on Probabilities*, English Translation, Dover (1951).

[3] L.N. Sahoo, B.C. Das and J. Sahoo, A class of predictive estimators in two-stage sampling, *Journal of the Indian Society of Agricultural Statistics*, **63**, No. 2 (2009), 175-180.

[4] A. Scott and T.M.P. Smith, Estimation in multistage surveys, *Jour. Amer. Stat. Assoc.*, **64** (1969), 830-840.

[5] S.K. Srivastava, Predictive Estimation of finite population mean using product estimator, *Metrika*, 30 (1983), 93-99.

[6] D.J. Watson, The estimation of leaf areas, *Jour. Agri. Sci.*, **27** (1937), 474.