Preliminary Test Regression Estimator in Double Sampling Based on Two-Stage and Ranked Set Sampling with Two Auxiliary Variables

Worthing Yanglem[1] 🄳 · Phrangstone Khongji[2]

Abstract

In this paper a preliminary test regression estimator for estimating the population mean of study variable Y is suggested for the two-stage sampling using ranked set sampling in the second stage when there is partial information on both of the two auxiliary variables X and Z. The variables (X, Y, Z) are considered to follow a trivariate normal distribution. Bias and mean square error of the proposed estimator are computed and comparison is made with the usual regression estimator under the same sampling scheme with two auxiliary variables analytically and numerically.

Keywords Two-stage sampling · Ranked set sampling · Partial information · Preliminary test · Regression estimator

1 Introduction

The use of an auxiliary variable, say X which is correlated with the study variable Y is well known to increase the efficiency of an estimator. When there is a linear relationship between them but the line does not pass through the origin, the linear regression estimator could be used under the assumption that the population mean μ_x is known. However, μ_x is usually unknown in practice so, double sampling is used where a preliminary sample is taken to estimate it.

There is also a situation where partial information about μ_x might be available, then we can perform a preliminary test and construct an estimator based on the result of the test. Han (1973a) assumes $(x_i, y_i); i = 1, 2, ..., n$ being a bivariate normal

✉ Worthing Yanglem
  yanglemthing2@gmail.com

1 Department of Statistics, North Eastern Hill University, Shillong, Meghalaya 793022, India

2 Department of Basic Science and Social Sciences, North Eastern Hill University, Shillong, Meghalaya 793022, India

sample with means μ_x, μ_y (unknown), known variances σ_x, σ_y and correlation coefficient ρ and consider a preliminary test for

$$H_0 : \mu_x = \mu_0 \text{ against } H_1 : \mu_x \neq \mu_0$$

where μ_0 is the value of μ_x obtained from the partial information on it. Assuming $\mu_0 = 0$ without loss of generality, the preliminary test estimator (PTE) is constructed as

$$\bar{y}* = \begin{cases} \bar{y} - \rho\bar{x} & \text{if } |\sqrt{n}\bar{x}| \leq Z_\alpha \\ \bar{y} & \text{if } |\sqrt{n}\bar{x}| > Z_\alpha \end{cases}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y respectively, α being the preliminary test level and Z_α is the $100(1 - \alpha/2)\%$ of $N(0, 1)$. The estimator with μ_0 will be used when H_0 is accepted. If H_0 is rejected, $\bar{y}$ is used as an estimator.

In another paper of Han (1973b), double sampling is used to estimate the unknown μ_x. When there is partial information on μ_x also, he performs the same preliminary test and the estimator is defined as

$$\bar{y}** = \begin{cases} \bar{y} - \rho\bar{x} & \text{if } |\sqrt{n'}\bar{x}'| \leq Z_\alpha \\ \bar{y} + \rho(\bar{x}' - \bar{x}) & \text{if } |\sqrt{n'}\bar{x}'| > Z_\alpha \end{cases}$$

where $\bar{x}'$ is the sample mean of n' observations from double sampling. If H_0 is accepted, the prior value μ_0 is used; otherwise the sample mean $\bar{x}'$ based on the preliminary sample in double sampling is used.

For two auxiliary variables, Das and Bez (1995) have constructed a preliminary test regression estimator with double sampling which is found to be more efficient than the ordinary regression estimator. Khongji and Das (2012) have also worked on some preliminary test estimators in double sampling using stratification.

In this paper, we will construct a preliminary test regression estimator in two-stage cluster sampling using ranked set sampling (RSS) in the second stage using double sampling. RSS was introduced by McIntyre in 1952 for estimating yield where measurement of the study variable is difficult but judgment ranking is possible. Stokes (1977) proved that the variance of RSS sample mean, ranking based on concomitant variables is smaller than that of SRS.

Sud and Mishra (2006) focused on the estimation of finite population mean in two-stage RSS designs. Nematollahi et al. (2008) used RSS in the second stage with replacement of two-stage sampling and showed that the two-stage cluster sampling with RSS (TSCRSS) is more efficient than SRS in two-stage sampling. Ozturk (2019) also worked on RSS with and without replacement in two-stage cluster sampling. He used sample mean as an estimator and found that the efficiency depends on the intra-cluster correlation coefficient and the sampling designs also.

Since we are using regression estimator of μ_y, it is required to know the population mean of the auxiliary variable. However, the population mean of the auxiliary variable is usually unknown. An experimenter may have partial information on it from other sources and believes the population mean of the auxiliary variable is some value, which we call a prior value but not known for certain. So, hypothesis testing

is needed to test the significance of the prior value. Under the assumption that we have partial information on auxiliary variables, we will study whether the proposed estimator can be better than the conventional regression estimator in the two-stage sampling by using RSS without replacement in the second stage, even though there is error in ranking. Throughout this paper we are going to take a situation where ranking is imperfect. When the ranking is no better than random, the sample obtained through RSS can be considered an almost as simple random sample.

2 Development of the Proposed Estimator

Suppose, X and Z are two auxiliary variables and we have partial information on both. According to Han (1973b), we can construct a preliminary test estimator using double sampling to utilize the partial information. Let (X, Y, Z) follows a trivariate normal distribution with means μ_x, μ_y, μ_z, variances $\sigma_x^2, \sigma_y^2, \sigma_z^2$ respectively and correlation coefficients $\rho_{yx}, \rho_{yz}, \rho_{xz}$. Considering μ_{0x} and μ_{0z} being the prior values of μ_x and μ_z respectively, we can perform a preliminary test on the hypotheses

$$H_{01} : \mu_x = \mu_{0x} \text{ and } H_{02} : \mu_z = \mu_{0z} \tag{1}$$

For selecting samples we will implement RSS in the second stage of the two-stage sampling:

Assume that a population has N primary sampling units (PSUs) or first stage units, each having $M_i; i = 1, 2, ..., N$ secondary stage units (SSUs). In the first stage sampling, n units are selected from N PSUs with SRS without replacement (SRSWOR).

In the second stage sampling, we will select sample of SSUs from each selected PSU. However, we are assuming μ_x and μ_z to be unknown. Thus, double sampling is used to first estimate μ_x and μ_z by selecting m_i' SSUs from each selected PSU using SRSWOR. Let $\bar{x}'$ and $\bar{z}'$ be the sample means obtained from the preliminary samples with sample size $m' = \sum_i^n m_i'$ on X and Z respectively through double sampling. Since we are using RSS, all preliminary samples within each selected PSUs are to be ranked according to the auxiliary variable. Here, we have two auxiliary variables, so, the variable which is more highly correlated with the study variable is used for ranking. Select the largest unit from the first PSU, the second largest from the second PSU and continue till the smallest of the ith PSU $; i = 1, 2, ..., n$ is chosen. We will use different sample sizes for each cluster.

Suppose we select $m_i (< m_i')$ units using RSS such that $m_i = r_i m_i''$ where m_i'' is the number of samples selected in r_i cycles from the ith selected PSU and $m = \sum_{i=1}^n m_i$. If X is used for ranking, then

$X_{il(j)} =$ value of X for the jth rank in the lth cycle of the ith selected PSU;

$i = 1, 2, ..., n, l = 1, 2, ..., r_i, j = 1, 2, ..., m_i''$.

Then, $Y_{il[j]}$ and $Z_{il(j)}$ are the value of Y and Z respectively corresponding to $X_{il(j)}$. Let $\bar{x}_{2srss}, \bar{y}_{2srss}$ and $\bar{z}_{2srss}$ denote the corresponding sample means of observations obtained on X, Y and Z such that

$$\bar{y}_{2srss} = \frac{1}{n\bar{M}} \sum_{i=1}^{n} \sum_{l=1}^{r_i} \sum_{j=1}^{m_i''} \frac{M_i}{r_i m_i''} Y_{il[j]}, \quad \bar{x}_{2srss} = \frac{1}{n\bar{M}} \sum_{i=1}^{n} \sum_{l=1}^{r_i} \sum_{j=1}^{m_i''} \frac{M_i}{r_i m_i''} X_{il(j)}$$

$$\text{and } \bar{z}_{2srss} = \frac{1}{n\bar{M}} \sum_{i=1}^{n} \sum_{l=1}^{r_i} \sum_{j=1}^{m_i''} \frac{M_i}{r_i m_i''} Z_{il(j)} \quad \text{where } \bar{M} = \frac{1}{N} \sum_{i=1}^{N} M_i$$

Considering that the covariance matrix Σ is known and $\sigma_x^2 = \sigma_y^2 = \sigma_z^2 = 1$ without loss of generality, the joint distribution of $(\bar{x}', \bar{x}_{2srss}, \bar{y}_{2srss}, \bar{z}', \bar{z}_{2srss})$ follows a multivariate normal distribution with mean $(\mu_x, \mu_x, \mu_y, \mu_z, \mu_z)$ and covariance matrix given by

$$\Sigma = \begin{pmatrix} \frac{1}{m'} & \frac{1}{m'} & \frac{\rho_{yx}}{m'} & \frac{\rho_{xz}}{m'} & \frac{\rho_{xz}}{m'} \\ \frac{1}{m'} & \frac{1}{m} & \frac{\rho_{yx}}{m} & \frac{\rho_{xz}}{m} & \frac{\rho_{xz}}{m} \\ \frac{\rho_{yx}}{m'} & \frac{\rho_{yx}}{m} & \frac{1}{m_i} & \frac{\rho_{yz}}{m'} & \frac{\rho_{yz}}{m} \\ \frac{\rho_{xz}}{m'} & \frac{\rho_{xz}}{m} & \frac{\rho_{yz}}{m} & \frac{1}{m'} & \frac{1}{m} \\ \frac{\rho_{xz}}{m'} & \frac{\rho_{xz}}{m} & \frac{\rho_{yz}}{m} & \frac{1}{m'} & \frac{1}{m} \end{pmatrix}$$

The test statistic to perform a preliminary test significance of the hypotheses (1) at α level of significance is as follows. The null hypothesis H_{01} can be accepted when

$$\left| \frac{\bar{x}' - \mu_x}{SE(\bar{x}')} \right| \le Z_\alpha$$

and H_{02} may be accepted when

$$\left| \frac{\bar{z}' - \mu_z}{SE(\bar{z}')} \right| \le Z_\alpha$$

If the null hypotheses are accepted, μ_{0x} and μ_{oz} are used instead of μ_x and μ_z in the proposed estimator; otherwise, sample means based on the preliminary samples $\bar{x}'$ and $\bar{z}'$ are used. Thus, it follows that

$$P_{H_{01}} \left[\bar{x}' - \frac{Z_\alpha}{\sqrt{m'}} \le \mu_x \le \bar{x}' + \frac{Z_\alpha}{\sqrt{m'}} \right] = 1 - \alpha \tag{2}$$

and

$$P_{H_{02}} \left[\bar{z}' - \frac{Z_\alpha}{\sqrt{m'}} \le \mu_z \le \bar{z}' + \frac{Z_\alpha}{\sqrt{m'}} \right] = 1 - \alpha \tag{3}$$

Since we have partial information on μ_x and μ_z, we can also assume $\mu_{0x} = \mu_{0z} = 0$ without loss of generality such that

$$H_{01} : \mu_x = 0 \text{ and } H_{02} : \mu_z = 0$$

Equations (2) and (3) imply that

$$P_{H_{01}}\left[\bar{x}' \le \frac{Z_\alpha}{\sqrt{m'}}, \bar{x}' \ge \frac{Z_\alpha}{\sqrt{m'}}\right] = 1 - \alpha$$

and

$$P_{H_{02}}\left[\bar{z}' \le \frac{Z_\alpha}{\sqrt{m'}}, \bar{z}' \ge \frac{Z_\alpha}{\sqrt{m'}}\right] = 1 - \alpha$$

Hence, it follows that the hypothesis H_{01} can be accepted if $|\bar{x}'| \le \frac{Z_\alpha}{\sqrt{m'}}$ and H_{02} will be accepted if $|\bar{z}'| \le \frac{Z_\alpha}{\sqrt{m'}}$.

Therefore, we propose the following estimator under the above assumptions as

$$T = \begin{cases} \bar{y}_{2srss} - B_{yx}\bar{x}_{2srss} - B_{yz}\bar{z}_{2srss} & \text{if } |\bar{x}'| \le \frac{Z_\alpha}{\sqrt{m'}}, |\bar{z}'| \le \frac{Z_\alpha}{\sqrt{m'}} \\[2mm] \bar{y}_{2srss} + B_{yx}(\bar{x}' - \bar{x}_{2srss}) - B_{yz}\bar{z}_{2srss} & \text{if } |\bar{x}'| > \frac{Z_\alpha}{\sqrt{m'}}, |\bar{z}'| \le \frac{Z_\alpha}{\sqrt{m'}} \\[2mm] \bar{y}_{2srss} - B_{yx}\bar{x}_{2srss} + B_{yz}(\bar{z}' - \bar{z}_{2srss}) & \text{if } |\bar{x}'| \le \frac{Z_\alpha}{\sqrt{m'}}, |\bar{z}'| > \frac{Z_\alpha}{\sqrt{m'}} \\[2mm] \bar{y}_{2srss} + B_{yx}(\bar{x}' - \bar{x}_{2srss}) + B_{yz}(\bar{z}' - \bar{z}_{2srss}) & \text{if } |\bar{x}'| > \frac{Z_\alpha}{\sqrt{m'}}, |\bar{z}'| > \frac{Z_\alpha}{\sqrt{m'}} \end{cases}$$

Here, $B_{yx} = \frac{\rho_{yx} - \rho_{yz}\rho_{xz}}{1 - \rho_{xz}{}^2}$ and $B_{yz} = \frac{\rho_{yz} - \rho_{yx}\rho_{xz}}{1 - \rho_{xz}{}^2}$ are known population regression coefficients of Y on X and Z respectively.

2.1 Bias and Mean Square Error of T

We know that

$$E(T) = E\left(\bar{y}_{2srss} - B_{yx}\bar{x}_{2srss} - B_{yz}\bar{z}_{2srss}\right) + B_{yx}E\left(\bar{x}'/|\bar{x}'| > \frac{Z_\alpha}{\sqrt{m'}}\right)$$

$$P\left(|\bar{x}'| > \frac{Z_\alpha}{\sqrt{m'}}\right) + B_{yz}E\left(\bar{z}'/|\bar{z}'| > \frac{Z_\alpha}{\sqrt{m'}}\right).P\left(|\bar{z}'| > \frac{Z_\alpha}{\sqrt{m'}}\right)$$

$$= \mu_y - Bias(T)$$

where, $Bias(T) = B_{yx}\mu_x\{\phi(a) - \phi(b)\} - B_{yx}.\frac{\varphi(a) - \varphi(b)}{\sqrt{m'}} + B_{yz}\{\phi(A) - \phi(B)\}$

$$- B_{yz}.\frac{\varphi(A) - \varphi(B)}{\sqrt{m'}}$$

where $\varphi(.)$ denotes the density function and $\phi(.)$ is the cumulative distribution function of $N(0, 1)$

and $\;a = Z_\alpha - \mu_x\sqrt{m'}; b = -Z_\alpha - \mu_x\sqrt{m'}; A = Z_\alpha - \mu_z\sqrt{m'}; B = -Z_\alpha - \mu_z\sqrt{m'}$

Then, the mean square error can be obtained as

$$MSE(T) = Var(T) + \{Bias(T)\}^2$$
$$= E(T^2) - \{E(T)\}^2 + 2\mu_y.Bias(T)$$

Using multivariate normal distribution and differentiation under integral sign,

$$MSE(T) = \left(\frac{1-B}{m} + \frac{B}{m'}\right) + H$$

where $B = B_{yx}^2 + B_{yz}^2 - 2B_{yx}\rho_{yx} - 2B_{yz}\rho_{yz} + 2B_{yx}B_{yz}\rho_{zx};$

$$H = \{\phi(a) - \phi(b)\}\left[B_{yx}^2\left(\mu_x^2 + \frac{1}{m_i}\right) - 2B_{yx}\rho_{yx} + B_{yx}B_{yz}\left(\mu_x\mu_z + \frac{\rho_{xz}}{m'}\right)\right]$$
$$+ \frac{\varphi(a) - \varphi(b)}{\sqrt{m'}}\left[-2B_{yx}^2\mu_x + 2B_{yx}\rho_{yx}\mu_x - B_{yx}B_{yz}(\mu_z + \rho_{xz}\mu_x)\right] + \frac{a\varphi(a) - b\varphi(b)}{m'}$$
$$(-B_{yz}^2 + 2B_{yx}\rho_{yx} - B_{yx}B_{yz}\rho_{xz}) + \{\phi(A) - \phi(B)\}\left[B_{yz}^2\left(\mu_z^2 + \frac{1}{m'}\right) - 2B_{yz}\rho_{yz}\right.$$
$$\left. + B_{yx}B_{yz}\left(\mu_x\mu_z + \frac{\rho_{xz}}{m'}\right)\right] + \frac{\varphi(A) - \varphi(B)}{\sqrt{m'}}\left[-2B_{yz}^2\mu_z + 2B_{yz}\rho_{yz}\mu_z - B_{yx}B_{yz}\right.$$
$$\left. (\mu_x + \rho_{xz}\mu_z)\right] + \frac{A\varphi(A) - B\varphi(B)}{m'}(-B_{yz}^2 + 2B_{yz}\rho_{yz} - B_{yx}B_{yz}\rho_{xz})$$

3 Comparison

The first quantity of the $MSE(T)$ is the variance of the regression estimator in two-stage with RSS using double sampling for two auxiliary variables i.e..,

$$T_1 = \bar{y}_{2srss} + B_{yx}(\bar{x}' - \bar{x}_{2srss}) + B_{yz}(\bar{z}' - \bar{z}_{2srss})$$

Thus, the proposed estimator is compared with the regression estimator T_1. We define the relative efficiency of T with respect to T_1 as

$$e = \frac{\frac{1-B}{m} + \frac{B}{m'}}{\left(\frac{1-B}{m} + \frac{B}{m'}\right) + H} = \frac{G}{G + H} \quad \text{(say)}$$

As we can see from the expression of MSE, the values of e depend on $m, m', B_{yx}, B_{yz}, \alpha, \mu_x, \mu_z$.

The relative efficiency e are computed for different values of $m', m = rm'', \mu_x, \mu_z, \alpha$. Table 1 presents the values of e when $\alpha = 0.01$ and for different values of $\rho_{yx} = 0.7(\mathbf{0.3})$, $\rho_{xz} = 0.6(\mathbf{0.5})$ and $\rho_{yz} = 0.8(\mathbf{0.4})$.

As shown in table 1, the value of e is very close to 1 at $\mu_x = \mu_z = 0$ but decreases as the value of μ_x and μ_z increases which is typical behavior of a preliminary test estimator. The relative efficiency attains the maximum at $\mu_x = \mu_z = 1$, but the efficiency is higher when the correlation between X, Y and Z is high. Also when the sample size m increases, e increases for each values of μ_x and μ_z.

4 Conclusion

The proposed regression estimator is more efficient than the ordinary preliminary test regression estimator under two-stage sampling along with RSS under certain conditions. However, in general, the information on the population mean is not always available and the experimenter may have some prior information on it. In such a situation, we can utilize partial information, and therefore, the efficiency of an estimator based on the preliminary test is quite high. And it is also more convenient to employ the preliminary test. Moreover, the estimator on the two-stage sampling with RSS in the second stage is also more efficient than SRS.

5 Empirical Study

Consider the total number of agricultural labourers as the study variable Y from 2011 census of Imphal West District, Manipur, India. Let $X =$ total number of cultivators and $Z =$ total number of households be the two auxiliary variables. The whole district is divided into 13 subdivisions/ towns (PSUs), taking villages as SSUs. Those PSUs with very small SSUs are combined together.

Table 1 Relative efficiency of T to T_1

μ_x	$\mu_z = 0$	0.5	1.0	0	0.5	1.0
	$m' = 15$			$m' = 15$		
	$r = 2, m'' = 2$			$r = 2, m'' = 3$		
0	0.96(**0.98**)	0.47(**0.69**)	1.23(**1.08**)	0.93(**0.98**)	0.35(**0.58**)	1.43(**1.13**)
0.5	0.63(**0.86**)	0.40(**0.66**)	0.83(**1.02**)	0.51(**0.80**)	0.29(**0.56**)	0.74(**1.03**)
1.0	1.07(**0.99**)	0.53(**0.72**)	1.26(**1.08**)	1.12(**0.98**)	0.42(**0.63**)	1.50(**1.13**)
	$m' = 20$			$m' = 20$		
	$r = 2, m'' = 2$			$r = 2, m'' = 3$		
0	0.96(**0.98**)	0.30(**0.51**)	1.04(**1.02**)	0.94(**0.98**)	0.21(**0.41**)	1.07(**1.03**)
0.5	0.45(**0.76**)	0.24(**0.47**)	0.56(**0.88**)	0.34(**0.67**)	0.16(**0.37**)	0.43(**0.83**)
1.0	0.99(**0.99**)	0.33(**0.54**)	1.06(**1.02**)	0.99(**0.99**)	0.24(**0.44**)	1.10(**1.03**)

The bold values in parentheses represent relative efficiency values for lower correlation coefficients

It is assumed that population means of X and Y are partially known. When μ_x and μ_z are unknown, we can estimate by using double sampling. In this data, we first select 8 out of 13 PSUs in the first stage sampling. In second stage, we will apply the double sampling procedure, where we select m_i' SSUs from each selected PSU. The estimates of μ_x and μ_z are given by $\bar{x}' = 161.8$ and $\bar{z}' = 303.1$ respectively.

Then, all m_i' SSUs are ranked based on the values of X. Samples are selected using RSS without replacement. We get

$$\bar{y}_{2srss} = 21.1, \ \bar{x}_{2srss} = 157.2, \ \bar{z}_{2srss} = 347.5.$$

Suppose we have partial information on auxiliary variables from the previous census data (2001 Census, Imphal West District, Manipur) where they are computed and given by $\mu_{ox} = 101.7$ and $\mu_{oz} = 329.5$. We need to utilize the preliminary samples to test the hypotheses

$$H_{01} : \mu_x = \mu_{0x} \text{ against } H_{11} : \mu_x \neq \mu_{0x}$$
$$\text{and } H_{02} : \mu_z = \mu_{0z} \text{ against } H_{21} : \mu_z \neq \mu_{0z}$$

At $\alpha = 0.01$, the test statistic, say, $|Z| > Z_\alpha$ under H_{01} . Therefore, H_{01} is rejected. Since reliable information on μ_x is not available, $\bar{x}'$ is used instead of μ_{0x} in the estimation of μ_x.

Under H_{02}, $|Z| < Z_\alpha$. When a reliable partial information of μ_x is available, H_{02} can be accepted and μ_{0z} is used in the estimator T. Therefore, 99% confidence interval is (243.3, 362.8).

Thus, the preliminary test regression estimator T which is the estimate of the population mean of Y is given by 14.8 and the usual regression estimator $T_1 = 4.3$. It is further observed that the $MSE(T) = 0.3, MSE(T_1) = 1.5$, hence, our proposed estimator T is more efficient than T_1.

References

Cochran WG (1977) Sampling techniques, 3rd edn. John Wiley and Sons, New York

Das G, Bez K (1995) Preliminary test estimators in double sampling with two auxiliary variables. Commun. Stat. - Theory Methods 24(5):1211–1226

Han CP (1973) Regression estimation for bivariate normal distributions. Ann Inst Stat Math 25(1):335–344

Han CP (1973) Double sampling with partial information on auxiliary variables. J Am Stat Assoc 68(344):914–918

Khongji P, Das G (2012) Studies on some preliminary test estimators in double sampling. J Indian Soc Agric Stat 66(3):381–390

McIntyre GA (1952) A method for unbiased selective sampling, using ranked sets. Aust J Agric Res 3(4):385–390. https://doi.org/10.1198/000313005X54180

Nematollahi N, Salehi M, Saba R (2008) Aliakbari Two-stage cluster sampling with ranked set sampling in the secondary sampling frame. Commun. Stat. - Theory Methods 37(15):2404–2415. https://doi.org/10.1080/03610920801919684

Ozturk O (2019) Two-stage cluster samples with ranked set sampling designs. Ann Inst Stat Math 71(1):63–91. https://doi.org/10.1007/s10463-017-0623-z

Stokes SL (1977) Ranked set sampling with concomitant variables. Commun. Stat. - Theory Methods 6(12):1207–1211

Sud Umesh, Mishra Dwijesh (2006) Estimation of finite population mean using ranked set two stage sampling designs, J Indian Soc Agric Stat., 60

Sukhatme PV, Sukhatme BV (1970) Sampling theory of surveys with applications. Iowa State University Press, Ames