Research Journal of Mathematical and Statistical Sciences ISSN 2320–6047
Vol. **3(10),** 1-3, October **(2015)** *Res. J. Mathematical and Statistical Sci.*

A Note on the Performance of Some Two-Stage Regression Estimators in Two-Phase Sampling

N.R. Das and L.N. Sahoo
Department of Statistics, Utkal University, Bhubaneswar 751004, INDIA

Available online at: www.isca.in, www.isca.me
Received 25th July 2015, revised 1st September 2015, accepted 9th October 2015

Abstract

In this paper, we study and compare the performance of three two-stage sampling regression estimators, viz., classical, chain and predictive regression estimators considered in the survey sampling literature under a two-phase sampling set-up. The study leads to a conclusion that the chain regression estimator has a better performance than others. Numerical studies have also been reported for illustration.

Keywords: Auxiliary variable, chain estimator, predictive estimator, regression estimator, two-phase sampling, two-stage sampling. MSC 2010 Subject Classification: 62D05.

Introduction

Suppose that a finite population U is divided into N clusters, denoted by $U_1, U_2, \dots, U_N$, called first stage units (*fsu*) such that the number of second stage units (*ssu*) in U_i is M_i and $M = \sum_{i=1}^{N} M_i$. Let y_{ij} and x_{ij} denote values of the study variable y and an auxiliary variable x respectively, for the jth *ssu* of U_i $(j = 1, 2, \dots, M_i; i = 1, 2, \dots, N)$. Define

$\bar{Y}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij}$ and $\bar{X}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} x_{ij}$

as the means of U_i, and

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} u_i \bar{Y}_i$ and $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} u_i \bar{X}_i$

as the overall means of U, where $u_i = \frac{NM_i}{M}$.

To estimate $\bar{Y}$, assume that a sample s of n *fsu*s is drawn from U and then a sample s_i of m_i *ssu*s from the selected U_i according to the design simple random sampling without replacement (SRSWOR). Let us define the following statistics:

$\bar{y}_i = \frac{1}{m_i}\sum_{j\in s_i} y_{ij}, \bar{x}_i = \frac{1}{m_i}\sum_{j\in s_i} x_{ij}$, $\bar{y} = \frac{1}{n}\sum_{i\in s} u_i \bar{y}_i$, $\bar{x} = \frac{1}{n}\sum_{i\in s} u_i \bar{x}_i$ and $\bar{x}' = \frac{1}{n}\sum_{i\in s} u_i \bar{X}_i$.

When $\bar{X}$ is known accurately, the classical regression estimator

$t_{RG} = \bar{y} - b_{byx}(\bar{x} - \bar{X})$

considered in Sukhatme *et al.*[1], where

$b_{byx} = \frac{\sum_{i\in s}(u_i\bar{y}_i - \bar{y})(u_i\bar{x}_i - \bar{x})}{\sum_{i\in s}(u_i\bar{x}_i - \bar{x})^2}$,

is well known in the literature. But, under the assumption that the means of x for the selected $U_i, i\epsilon s$, *i.e.*, $\bar{X}_i$'s are known, Sahoo[2] suggested a chain regression estimator of the form

$t_{CRG} = \frac{1}{n}\sum_{i\in s} u_i\left[\bar{y}_i - b_{iyx}(\bar{x}_i - \bar{X}_i)\right] - b_{byx}(\bar{x}' - \bar{X})$

where

$b_{iyx} = \frac{\sum_{j\epsilon s_i}(y_{ij}-\bar{y}_i)(x_{ij}-\bar{x}_i)}{\sum_{j\epsilon s_i}(x_{ij}-\bar{x}_i)^2}$.

Under the same assumption on the availability of auxiliary information and motivated by the usual predictive approach of Basu[3], Sahoo and Panda[4] developed a predictive regression estimator, defined by

$t_{PRG} = \frac{1}{n}\sum_{i\in s} u_i\left[\bar{y}_i - \{b_{byx} - \emptyset(b_{byx} - b_{iyx})\}(\bar{x}_i - \bar{X}_i)\right] - b_{byx}(\bar{x}' - \bar{X})$,

where $\emptyset = \frac{n}{N}$.

Now, we see that the three regression estimators t_{RG}, t_{CRG} and t_{PRG} need prior information on the auxiliary variable x at different survey operations. Both t_{CRG} and t_{PRG} require that the cluster means $\bar{X}_i, i\epsilon s$, as well as the overall population mean $\bar{X}$ must be known in advance. On the other hand, t_{RG} requires that the prior information on $\bar{X}$ must be available. But, in many surveys, such extensive information is unavailable. Thus, the common procedure in such a situation is to use a two-phase sampling or sampling followed by sub-sampling. The topic of this paper is to study relative efficiencies of the classical, chain and predictive regression estimators of the population mean $\bar{Y}$ under a two-phase sampling scheme.

Two-Phase Sampling and the Estimators

Let us consider the following two-phase sampling mechanism under the assumption that sampling at each phase and each stage is done by SRSWOR: i. A first phase sample s' $(s' \subset U)$ of n' *fsu*s is drawn out of N in the first stage and a sample s_i' $(s_i' \subset U_i)$ of m_i' *ssu*s is drawn from M_i *ssu*s of $U_i, i \in s'$. The sample so selected consists of $\sum_{i=1}^{n'} m_i'$ *ssu*s used to gather inexpensive information on x. ii. A second phase sample (sub-sample) s $(s \subset s')$ of n *fsu*s is selected out of n' *fsu*s selected in the first phase sample s' and then in $U_i, i \in s$, a sub-sample s_i $(s_i \subset s_i')$ of m_i *ssu*s is selected out of the m_i' *ssu*s selected in

the first phase sample s_i'. The study variable y is then observed for the *ssu*s selected in the second phase sample.

Thus, our two-phase sampling classical, chain and predictive regression estimators are defined as

$t_{RG} = \frac{1}{n}\sum_{i\in s} u_i\left[\bar{y}_i - b_{byx}(\bar{x}_i - \bar{x}_{di})\right] - b_{byx}(\bar{x}_d' - \bar{x}_d)$,

$t_{CRG} = \frac{1}{n}\sum_{i\in s} u_i\left[\bar{y}_i - b_{iyx}(\bar{x}_i - \bar{x}_{di})\right] - b_{byx}(\bar{x}_d' - \bar{x}_d)$,

and

$t_{PRG} = \frac{1}{n}\sum_{i\in s} u_i\left[\bar{y}_i - \{b_{byx} - \emptyset(b_{byx} - b_{iyx})\}(\bar{x}_i - \bar{x}_{di})\right] - b_{byx}(\bar{x}_d' - \bar{x}_d)$

respectively, where

$\bar{x}_{di} = \frac{1}{m_i'}\sum_{j\epsilon s_i'} x_{ij}$, $\bar{x}_d = \frac{1}{n'}\sum_{i\epsilon s'} u_i\,\bar{x}_{di}$ and $\bar{x}_d' = \frac{1}{n}\sum_{i\epsilon s} u_i\,\bar{x}_{di}$.

The statistics $\bar{y}_i, \bar{x}_i, b_{iyx}$ and b_{byx} are computed using data on the second phase sample as defined earlier.

Comparison of Estimators

For the comparison purpose, we need exact variance expressions of the estimators. But, it is not possible to derive these expressions as the estimators have complex structure. Therefore, we consider approximate variance expressions of the estimators to a first order of approximation. However, to get them in an easier way, we first approximate $b_{iyx} \doteqdot \beta_{iyx}, i\epsilon s$, and $b_{byx} \doteqdot \beta_{byx}$, where the parameters β_{iyx} and β_{byx} are defined by

$\beta_{iyx} = \frac{S_{iyx}}{S_x^2}$ and $\beta_{byx} = \frac{S_{byx}}{S_{bx}^2}$

such that

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

and expressions for S_{iy}^2 , S_{ix}^2 , S_{by}^2 and S_{bx}^2 can be obtained by considering $y = x$. Then, we find that the resulting estimators are unbiased so that exact variance expressions of the same estimators can be easily obtained in the traditional way. These are obviously approximate variance expressions of the regression estimators t_{CRG}, t_{PRG} and t_{RG}. Now, omitting details of the derivations to save space, these expressions are presented below:

$$V(t_{RG}) = \frac{1-\gamma}{n}S_{by}^2(1-\rho_{byx}^2) + \frac{1-\emptyset'}{n'}S_{by}^2 + \frac{1}{n'N}\sum_{i=1}^{N}u_i^2\frac{1-\emptyset_i'}{m_i'}S_{iy}^2 + \frac{1-\gamma}{nN}\sum_{i=1}^{N}u_i^2\frac{1-\emptyset_i'}{m_i'}(S_{iy}^2 + \beta_{byx}^2S_{ix}^2 - 2\beta_{byx}S_{iyx}) + \frac{1}{nN}\sum_{i=1}^{N}u_i^2\frac{1-\gamma_i}{m_i}(S_{iy}^2 + \beta_{byx}^2S_{ix}^2 - 2\beta_{byx}S_{iyx}) \quad (1)$$

$$V(t_{CRG}) = \frac{1-\gamma}{n}S_{by}^2(1-\rho_{byx}^2) + \frac{1-\emptyset'}{n'}S_{by}^2 + \frac{1}{n'N}\sum_{i=1}^{N}u_i^2\frac{1-\emptyset_i'}{m_i'}S_{iy}^2 + \frac{1-\gamma}{nN}\sum_{i=1}^{N}u_i^2\frac{1-\emptyset_i'}{m_i'}(S_{iy}^2 + \beta_{byx}^2S_{ix}^2 - 2\beta_{byx}S_{iyx}) + \frac{1}{nN}\sum_{i=1}^{N}u_i^2\frac{1-\gamma_i}{m_i}S_{iy}^2(1-\rho_{iyx}^2) \quad (2)$$

$$V(t_{PRG}) = \frac{1-\gamma}{n}S_{by}^2(1-\rho_{byx}^2) + \frac{1-\emptyset'}{n'}S_{by}^2 + \frac{1}{n'N}\sum_{i=1}^{N}u_i^2\frac{1-\emptyset_i'}{m_i'}S_{iy}^2 + \frac{1-\gamma}{nN}\sum_{i=1}^{N}u_i^2\frac{1-\emptyset_i'}{m_i'}(S_{iy}^2 + \beta_{byx}^2S_{ix}^2 - 2\beta_{byx}S_{iyx}) + \frac{1}{nN}\sum_{i=1}^{N}u_i^2\frac{1-\gamma_i}{m_i}(S_{iy}^2 + B_i^2S_{ix}^2 - 2B_iS_{iyx}) \quad (3)$$

where

$\emptyset_i' = \frac{m_i'}{M_i}, \gamma_i = \frac{m_i}{m_i'}, \emptyset' = \frac{n'}{N}, \gamma = \frac{n}{n'}$, $\rho_{iyx} = \frac{S_{iyx}}{S_{iy}S_{ix}}$, $\rho_{byx} = \frac{S_{byx}}{S_{by}S_{bx}}$

and

$B_i = \beta_{byx} - \emptyset(\beta_{byx} - \beta_{iyx})$.

Hence, from the expressions (1), (2) and (3), it is clear that $V(t_{CRG}) \leq V(t_{PRG}) \leq V(t_{RG})$.

Thus, the two-phase chain regression estimator is the most efficient, whereas the two-phase predictive regression estimator has better performance than the two-phase classical regression estimator.

Numerical Study

As numerical illustrations of the gain in efficiency of different comparable estimators, we consider data on two natural populations as described below:

Population I: It consists of strip-wise complete enumeration data on timber volume $(= y)$ and length $(= x)$ for 176 strips (*ssu*s) divided into 10 $(= 10)$ blocks (*fsu*s) of the Black Mountain Experimental Forest given in Murthy [5]. For this population, we select $n' = 6, n = 3$, $m_i' = 5,5,5,5,4,5,4,5,8$ and 8, and $m_i = 2, \forall i \in s$.

Population II: This population (called as MU284 population) available in Sarndal, Swensson and Wretman [6] consists of 284 municipalities (*ssu*s) divided into 50 clusters (*fsu*s) with two variables 1985 population $(= y)$ and 1975 population $(= x)$. Here, we consider $n' = 20, n = 10, m_i' = 3, \forall i \in s'$, and $m_i = 2, \forall i \in s$.

Considering expressions (1), (2) and (3), relative precisions of t_{CRG}, t_{PRG} and t_{RG} compared to the direct estimator $\bar{y}$ are displayed in table 1. We use the following formula for the variance of $\bar{y}$:

$V(\bar{y}) = \frac{1-\gamma}{n}S_{by}^2 + \frac{1}{nN}\sum_{i=1}^{N}u_i^2\frac{1-\gamma_i}{m_i}S_{iy}^2$.

Table-1
Relative Precision of Different Estimators Compared to $\bar{y}$ (in %)

Population	Estimators			
	$\bar{y}$	t_{RG}	t_{CRG}	t_{PRG}
I	100	123	166	149
II	100	224	290	235

From the entries of table 1, it is clear that the gain in precision of t_{CRG} over other two regression estimators is considerably high for both the populations under consideration.

Conclusion

Our analytical as well as numerical comparisons of three regression estimators lead to the conclusion that the chain regression estimator is certainly better than the classical and predictive regression estimators. However, the numerical study undertaken, although confined to two populations only, shows that there are practical situations where chain regression estimator yields considerable efficiency gains compared to other two competitors.

References

1. Sukhatme P.V., Sukhatme B.V., Sukhatme S. and Asok C., *Sampling Theory of Surveys with Applications*, Indian Society of Agricultural Statistics, New Delhi, 233 (**1984**)

2. Sahoo L.N., A regression-type estimator in two-stage sampling, *Calcutta Statistical Association Bulletin*, **36**, 97-100 (**1987**)

3. Basu D., An essay on the logical foundations of survey sampling (Part I), *Foundations of Statistical Inference*, V.P. Godambe and D.A. Sprott (eds), Holt, Rinehart and Winston, Toronto, Canada, 203-242 (**1971**)

4. Sahoo L.N. and Panda P., A predictive regression-type estimator in two-stage sampling, *Journal of the Indian Society of Agricultural Statistics*, **52**, 303-308 (**1999**)

5. Murthy M.N., *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta, 131, (**1977**)

6. Sarndal C.E., Swensson B. and Wretman J., *Model Assisted Survey Sampling*. Springer-Verlag, 652 (**1992**)