

On the efficient use of regression-in-ratio estimator in simple random sampling

Sull'uso efficiente dello stimatore regressione-in-rapporto nel campionamento casuale semplice

Pier Francesco Perri

Dipartimento di Economia e Statistica

Università della Calabria

pierfrancesco.perri@unical.it

Riassunto: Al fine di stimare la media incognita di una variabile oggetto di studio quando si dispone di due variabili ausiliarie correlate a quella di indagine, il metodo di stima della regressione viene impiegato congiuntamente a quello del rapporto per definire, nel campionamento casuale semplice senza reimmissione, lo stimatore regressione-in-rapporto. Dopo averne ricavato l'espressione della distorsione e dell'errore quadratico medio, lo stimatore proposto viene confrontato con la media campionaria e con gli stimatori rapporto e regressione nel campionamento a due fasi. Vengono, quindi, stabilite le condizioni sotto le quali lo stimatore in esame si rivela più efficiente degli altri.

Keywords: auxiliary variables, efficiency, ratio estimator, regression estimator, two-phase sampling.

1. Introduction

Auxiliary information is common used in sampling surveys practice in order to achieve higher precision in the estimates. This information may be used at planning stage of a survey leading, for example, to stratification or πps sampling designs, or directly at estimation stage through the ratio, product or regression methods, or at both phases.

In this paper we will concerned with the utilization of auxiliary information in the estimation stage. In simple random sampling without replacement ($srsWOR$) we will combine the ratio and regression methods when two auxiliary variables are available, but the mean of the main one is unknown.

In section 2, with the purpose of estimating the population mean of a study variable, the *regression-in-ratio estimator* is defined and the approximate expressions for its bias and mean square error are obtained under the hypothesis that the sample is sufficiently large. In section 3, the proposed estimator is compared with the sample mean in $srsWOR$ and with the ratio and regression estimators in *two-phase srsWOR*. Hence, the conditions that make the sample strategy based on the proposed estimator more efficient than the others are established.

2. The regression-in-ratio estimator

Consider a finite population $U = \{1, 2, \dots, N\}$ of size N and let $\mathcal{Y}$ and $\mathcal{X}$ be, respectively, the study and the auxiliary variable, taking positive values Y_i and X_i on the i -th population unit. Let $\bar{Y}$ and $\bar{X}$ be the population means of the two variables and suppose that $\bar{Y}$ is

unknown. Without taking into account the auxiliary variable, the unknown mean may be estimated in *srswor* using the *sample mean*, $\bar{y} = \sum_{i=1}^n y_i/n$.

If the two variables are highly correlated and the mean $\bar{X}$ is known before selecting the sample, the precision of the estimate may be considerably improved using the *ratio estimator*, $\hat{Y}_R = \bar{y}\bar{X}/\bar{x}$, if the correlation is positive, the *product estimator*, $\hat{Y}_P = \bar{y}\bar{x}/\bar{X}$, in case of negative correlation, or the *regression estimator*, $\hat{Y}_{lr} = \bar{y} + \hat{\beta}_{yx}(\bar{X} - \bar{x})$, being $\bar{x} = \sum_{i=1}^n x_i/n$ and $\hat{\beta}_{yx}$ the least squares estimate of β_{yx} , the population regression coefficient of $\mathcal{Y}$ on $\mathcal{X}$. In particular, focusing on ratio and regression methods of estimation, it is well known Cochran (1977) that in a large sample the strategy $(srswor, \hat{Y}_{lr})$ is always more efficient than the strategy $(srswor, \bar{y})$ (unless $\beta_{yx} = 0$, in which case the two strategies are equivalent), while the strategy $(srswor, \hat{Y}_R)$ is more efficient than $(srswor, \bar{y})$ if $\rho_{xy} > C_x/2C_y$, being ρ_{xy} the correlation coefficient between $\mathcal{X}$ and $\mathcal{Y}$, $C_w = S_w/\bar{W}$ the coefficient of variation of $\mathcal{W}$ and $S_w^2 = \sum_{i=1}^N (W_i - \bar{W})^2/(N-1)$, $\mathcal{W} = \mathcal{X}, \mathcal{Y}$.

Sometimes, however, the mean $\bar{X}$ is *a priori* unknown. In such a case, a common procedure is to use the *two-phase sampling* (or *double sampling*). This scheme requires the selection, according to any specified sample design, of a preliminary large sample s' of size n' to collect information on $\mathcal{X}$; a second(-phase) sample s of size n ($n < n'$) is then drawn from s' and used to observe $\mathcal{Y}$ and $\mathcal{X}$ and to estimate $\bar{Y}$ and $\bar{X}$. If the two samples are drawn according to *srswor*, the ratio estimator is defined as $\hat{Y}_{R.d} = \bar{y}\bar{x}'/\bar{x}$, where $\bar{x}'$ is the sample mean on the first sample s' while $\bar{y}$ and $\bar{x}$ represent the sample means on s . In a similar way, the regression estimator becomes $\hat{Y}_{lr.d} = \bar{y} + \hat{b}_{yx}(\bar{x}' - \bar{x})$, being $\hat{b}_{yx}$ the least squares estimate of the population regression coefficient of $\mathcal{Y}$ on $\mathcal{X}$ computed on the second-phase sample.

Now, suppose that a second positive auxiliary variable, say $\mathcal{Z}$, closely related to $\mathcal{X}$, is available. Suppose also that its mean, $\bar{Z}$, is known and let $\bar{z}$ be the sample mean. Then, when the mean $\bar{X}$ is unknown, it seems intuitive to use the relation between the two auxiliary variables to improve the performance of the ratio and regression estimator. This possibility has been widely discussed in *two-phase sampling*, among others, by Chand (1975), Kiregyera (1980), Kiregyera (1984), Sahoo and Sahoo (1993), Sahoo *et al.* (1994), Mishra and Rout (1997), Singh and Gangele (1999), Diana and Tommasi (2003).

In this paper we suggest to chain together the variables $\mathcal{Y}$, $\mathcal{X}$ and $\mathcal{Z}$ not through the *two-phase srswor* but drawing only a simple random sample (without replacement) having the same size, n , of a second-phase sample. Firstly, we assume that the variable $\mathcal{Z}$ is used to estimate $\bar{X}$ by means of regression method, then the regression estimator is chained with the ratio one involving $\mathcal{Y}$ and $\mathcal{X}$.

Let, therefore, $\hat{X}_{lr} = \bar{x} + \hat{\beta}_{xz}(\bar{Z} - \bar{z})$ be the regression estimator of $\bar{X}$, where $\hat{\beta}_{xz} = s_{xz}/s_z^2$ with $s_z^2 = \sum_{i=1}^n (z_i - \bar{z})^2/(n-1)$ and $s_{xz} = \sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})/(n-1)$. Replacing in the ratio estimator expression $\bar{X}$ by $\hat{X}_{lr}$, we define the *regression-in-ratio* estimator as:

$$\hat{Y}_{lr,R} = \frac{\bar{y}}{\bar{x}} \hat{X}_{lr}. \quad (1)$$

In order to obtain the expression for the Bias (B) and Mean Square Error (MSE) of the estimator, it is convenient to re-write (1) as:

$$\hat{Y}_{lr,R} = \frac{\bar{Y}(1 + \delta_y)}{\bar{X}(1 + \delta_x)} \left[\bar{X}(1 + \delta_x) - \bar{Z}\beta_{xz} \frac{1 + \delta_{s_{xz}}}{1 + \delta_{s_z^2}} \right] \quad (2)$$

where:

$$\delta_w = \frac{\bar{w} - \bar{W}}{\bar{W}}, \quad \mathcal{W} = \mathcal{X}, \mathcal{Y}, \mathcal{Z}; \quad \delta_{S_z^2} = \frac{s_z^2 - S_z^2}{S_z^2}; \quad \delta_{S_{xz}} = \frac{s_{xz} - S_{xz}}{S_{xz}}.$$

Assuming that for all the $\binom{N}{n}$ samples $|\delta_x| < 1$ and $|\delta_{S_z^2}| < 1$, the terms $(1 + \delta_x)^{-1}$ and $(1 + \delta_{S_z^2})^{-1}$ can be expanded in Taylor's series. If the sample size is sufficiently large these assumptions can be considered reliable. Now, taking in the expanded expression the expectation term-by-term up to including terms of order n^{-1} , we can obtain, to the first order of approximation, the following expressions for the Bias and Mean Square Error:

$$B(\hat{Y}_{lr,R}) = \frac{1-f}{n} \frac{\bar{Y}}{C_{002}} \left[C_{101}(C_{101} - C_{011}) + \frac{N-1}{N-2} (C_{003}C_{101} - C_{102}) \right] \quad (3)$$

$$MSE(\hat{Y}_{lr,R}) = \frac{1-f}{n} \bar{Y}^2 \left[C_{020} + \frac{C_{101}}{C_{002}} (C_{101} - 2C_{011}) \right] \quad (4)$$

where $f = \frac{n}{N}$ and $C_{rst} = \frac{1}{N-1} \frac{\sum_{i=1}^N (X_i - \bar{X})^r (Y_i - \bar{Y})^s (Z_i - \bar{Z})^t}{\bar{X}^r \bar{Y}^s \bar{Z}^t}.$

3. Efficiency comparisons

The efficiency of the proposed estimator is firstly analysed with respect to the sample mean. Using the previous notation, the variance of the sample mean can be expressed as $Var(\bar{y}) = (1-f) \bar{Y}^2 C_{020}/n$. Comparing $Var(\bar{y})$ with $MSE(\hat{Y}_{lr,R})$, it immediately follows that the proposed estimator is more efficient than $\bar{y}$ if $C_{101} (C_{101} - 2C_{011}) < 0$. In particular, we can observe that $\hat{Y}_{lr,R}$ is more efficient than $\bar{y}$ if $0 < \rho_{xz} < 2\rho_{yz}C_y/C_x$ or if $2\rho_{yz}C_y/C_x < \rho_{xz} < 0$, being ρ_{xz} the correlation coefficient between the two auxiliary variables.

Consider now the strategies based on ratio and regression estimators which make use of the single auxiliary variable $\mathcal{X}$ in *two-phase srswor*. It may be interesting to establish whether the use of the second auxiliary variable $\mathcal{Z}$ by means of $\hat{Y}_{lr,R}$ provides better estimates than the estimators $\hat{Y}_{R,d}$ and $\hat{Y}_{lr,d}$.

In order to compare the proposed strategy $S_0 = (srswor, \hat{Y}_{lr,R})$ with the two strategies $S_1 = (srswor, \hat{Y}_{R,d})$ and $S_2 = (srswor, \hat{Y}_{lr,d})$, the differences $MSE(\hat{Y}_{R,d}) - MSE(\hat{Y}_{lr,R})$ and $MSE(\hat{Y}_{lr,d}) - MSE(\hat{Y}_{lr,R})$ need to be studied. The expressions for $MSE(\hat{Y}_{R,d})$ and $MSE(\hat{Y}_{lr,d})$ can be found, among others, in Cochran (1977).

For large N , the following results can be proved after some calculations:

Theorem 1 Let $A = \rho_{xz} - 2\rho_{yz}C_y/C_x$, $B = 1 - 2\rho_{xy}C_y/C_x$ and assume $B > 0$. Then the strategy S_0 is more efficient than S_1 if one of the two cases occurs:

(1.a) $A\rho_{xz} < 0$

(2.a) $0 < A\rho_{xz} < B$ and $n/n' < 1 - (A/B)\rho_{xz}$.

Conversely, the strategy S_1 is more efficient than S_0 if one of the two cases occurs:

(3.a) $A\rho_{xz} > B > 0$

(4.a) $0 < A\rho_{xz} < B$ and $n/n' > 1 - (A/B)\rho_{xz}$.

Assume now $B < 0$. Then the strategy S_0 is more efficient than S_1 if one of the two cases occurs:

(1.b) $A\rho_{xz} < B < 0$

(2.b) $B < A\rho_{xz} < 0$ and $n/n' > 1 - (A/B)\rho_{xz}$.

Conversely, the strategy S_1 is more efficient than S_0 if one of the two cases occurs:

(3.b) $A\rho_{xz} > 0$

(4.b) $B < A\rho_{xz} < 0$ and $n/n' < 1 - (A/B)\rho_{xz}$.

Theorem 2 Let $C = 1 - 2(\rho_{yz}C_y/\rho_{xz}C_x)$ and $D = (\rho_{xz}C_x/\rho_{xy}C_y)^2$. Then the strategy S_0 is more efficient than S_2 if $n/n' > 1 + CD$

The two theorems show the conditions under which the use of the proposed strategy based on $\hat{Y}_{lr,R}$ is recommended. Nevertheless, the decision about the strategy to use depends on different quantities, such as correlation or variation coefficients, which may be unknown in certain practical situations. In this paper, we assume that these values are known or available on the basis of previous data, pilot survey, past experience or efficient estimates.

In future, we hope to extend the comparisons to different estimators which employ the auxiliary variables in two-phase sampling.

References

- Chand L. (1975) *Some Ratio-Type Estimators Based on Two or More Auxiliary Variates*, Unpublished Ph. D. Dissertation, Iowa State University, Ames, IOWA, USA.
- Cochran W.G. (1977) *Sampling Techniques*, 3rd ed, John Wiley & Sons, New York.
- Diana G. and Tommasi C. (2003) Optimal estimation for finite population mean in two-phase sampling, *Statistical Methods & Applications*, 12, 41–48.
- Kiregyera B. (1980) A chain ratio-type estimator in finite population double sampling using two auxiliary variables, *Metrika*, 27, 217–223.
- Kiregyera B. (1984) Regression-type estimators using two auxiliary variables and the model of double sampling from finite population, *Metrika*, 31, 215–226.
- Mishra G. and Rout K. (1997) A regression estimator in two-phase sampling in presence of two auxiliary variables, *Metron*, 55, 177–186.
- Sahoo J. and Sahoo L.N. (1993) A class of estimators in two-phase sampling using two auxiliary variables, *Journal of the Indian Statistical Association*, 31, 107–114.
- Sahoo J., Sahoo L.N. and Mohanty S. (1994) An alternative approach to estimation in two-phase sampling using two auxiliary variables, *Biometrical Journal*, 36, 293–298.
- Singh H.P. and Gangele R. (1999) Classes of almost unbiased ratio and product-type estimators in two phase sampling, *Statistica*, 59, 109–124.