

Multivariate Regression-Type Estimators

Stimatori Multivariati Tipo Regressione

Giancarlo Diana¹, Pier Francesco Perri²,

¹ Dipartimento di Scienze Statistiche, Università di Padova
e-mail: giancarlo.diana@unipd.it

² Dipartimento di Economia e Statistica, Università della Calabria

Riassunto: Le informazioni ausiliarie ricoprono un ruolo di notevole importanza nella pratica campionaria in quanto consentono, se opportunamente impiegate, di ottenere strategie campionarie altamente efficienti. In questo lavoro viene proposta una classe di stimatori per la media incognita di una variabile di studio in presenza di più variabili ausiliarie. Nell'ambito del campionamento casuale semplice senza reimmissione viene determinato, al primo ordine di approssimazione, l'errore quadratico medio minimo per gli stimatori appartenenti alla classe, mostrando che esiste almeno uno stimatore, quello per regressione, che lo raggiunge. Lo stimatore ottimale nella classe viene, infine, confrontato con due stimatori multivariati tipo-rapporto.

Keywords: auxiliary variables, mean square error, multivariate ratio-type estimators.

1. Introduction

The estimation of the population mean is a persistent issue in sampling practice and many efforts have been done to obtain improved designs and more efficient estimators. In large scale sample surveys it often happens to collect data on more than one auxiliary character. The literature on survey sampling describes a great variety of techniques for using auxiliary information. In particular, it is well known that when the auxiliary information is to be used at the estimation stage, the *ratio*, *product* and *regression* estimators are widely utilized in many situations.

This work is concerned with the problem of estimating the population mean of a study variable when two or more auxiliary variables are available. In this situation many estimators may be defined by linking together ratio, product or regression estimators, each one exploiting a single variable. Olkin (1958) was the first author to deal with the problem. He suggested the use of information on more than one supplementary characteristic, positively correlated with the study variable, considering a linear combination of ratio estimators based on each auxiliary variable separately. Singh (1967) gave a multivariate expression of product estimator similar to Olkin's multivariate ratio estimator, while Raj (1965) suggested a method of using multi-auxiliary variables considering a linear combination of single *difference estimators*. Many other contributions are present in sampling literature and, recently, some new estimators have appeared. Among others, we mention Ceccon and Diana (1996) who gave a multivariate extension of the Naik and Gupta (1991) univariate class of estimators, and Abu-Dayyeh *et al.* (2003).

In Section 2, we introduce a general class of estimators and derive its minimum mean square error up to the first order of approximation. In Section 3 we compare the efficiency of the best estimator in the class with that of the *multivariate ratio-type* estimators proposed in Olkin (1958) and Ceccon and Diana (1996).

2. A general class of estimators

Let $P = \{1, 2, \dots, N\}$ be a finite population, Y the study variable and $X_1, X_2, \dots, X_p$ p auxiliary variables taking values Y_j and X_{ij} for the j -th population unit, $j = 1, 2, \dots, N$. Let us suppose that the auxiliary variables are correlated to Y . Moreover, let $\bar{Y} = N^{-1} \sum_{j=1}^N Y_j$ be the unknown mean of the study variable and suppose that the population means of the p auxiliary variables, $\bar{X}_i = N^{-1} \sum_{j=1}^N X_{ij}$, are known.

In order to estimate $\bar{Y}$ a sample of size n is drawn from the population according to simple random sampling without replacement. Let us indicate with $\bar{y}$ and $\bar{x}_i$, the sample means of Y and X_i , $i = 1, 2, \dots, p$. Then, we define a general class of estimators as:

$$\bar{y}_g = g(\bar{y}, \bar{\mathbf{x}}^t) \quad (1)$$

where $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)^t$ and g is a function such that:

1. $g: S \rightarrow \mathbb{R}$ where $S \subseteq \mathbb{R}^{p+1}$ is a convex and bounded set containing the point $(\bar{Y}, \bar{\mathbf{X}}^t)$, with $\bar{\mathbf{X}} = \mathbb{E}(\bar{\mathbf{x}}) = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)^t$;
2. it is continuous and bounded in S ;
3. its first partial derivatives exist and are continuous and bounded in S ;
4. $g(\bar{y}, \bar{\mathbf{X}}^t) = \bar{y}$.

In order to determine the mean square error of the best estimator belonging to the class, let us indicate with

$$g_0 = \left. \frac{\partial g(\bar{y}, \bar{\mathbf{x}}^t)}{\partial \bar{y}} \right|_{(\bar{y}, \bar{\mathbf{x}}^t) = (\bar{Y}, \bar{\mathbf{X}}^t)} \quad \text{and} \quad g_i = \left. \frac{\partial g(\bar{y}, \bar{\mathbf{x}}^t)}{\partial \bar{x}_i} \right|_{(\bar{y}, \bar{\mathbf{x}}^t) = (\bar{Y}, \bar{\mathbf{X}}^t)}$$

the first partial derivatives of g with respect to $\bar{y}$ and $\bar{x}_i$, $i = 1, 2, \dots, p$.

Expanding $\bar{y}_g$ at the point $(\bar{Y}, \bar{\mathbf{X}}^t)$ in a first order Taylor's series, and considering that $g(\bar{Y}, \bar{\mathbf{X}}^t) = \bar{Y}$ and $g_0 = 1$, we get:

$$\bar{y}_g \cong g(\bar{Y}, \bar{\mathbf{X}}^t) + (\bar{y} - \bar{Y})g_0 + \sum_{i=1}^p (\bar{x}_i - \bar{X}_i)g_i = \bar{y} + (\bar{\mathbf{x}} - \bar{\mathbf{X}})^t \mathbf{g}_x \quad (2)$$

where $\mathbf{g}_x = (g_1, g_2, \dots, g_p)^t$. From (2), taking the expectation term-by-term up to include terms of order $O(n^{-1})$, we obtain the first order mean square error (variance) given by:

$$MSE(\bar{y}_g) \cong \frac{1-f}{n} (S_y^2 + \mathbf{g}_x^t \mathbf{S}_{xx} \mathbf{g}_x + 2\mathbf{g}_x^t \mathbf{S}_{yx}) \quad (3)$$

where $f = n/N$, S_y^2 is the variance of the study variable Y , $\mathbf{S}_{xx} = \{Cov(X_i, X_j)\}$, $i, j = 1, 2, \dots, p$, denotes the variance-covariance matrix of the auxiliary variables and $\mathbf{S}_{yx} = \{Cov(Y, X_i)\}$ is the vector of the covariances between Y and X_i , $i = 1, 2, \dots, p$.

Minimizing (3) with respect to $\mathbf{g}_x$ we get the optimum value $\mathbf{g}_x^* = -\mathbf{S}_{xx}^{-1} \mathbf{S}_{yx}$. We observe that $-\mathbf{g}_x^*$ denotes the vector of the *partial regression coefficients* of Y on X_i , $i = 1, 2, \dots, p$. Replacing the optimum $\mathbf{g}_x^*$ in (3) we obtain the minimum first order mean square error:

$$MSE^*(\bar{y}_g) \cong \frac{1-f}{n} S_y^2 (1 - \rho_{y.x}^2) \quad (4)$$

where $\rho_{y,x}^2 = \mathbf{S}_{yx}^t \mathbf{S}_{xx}^{-1} \mathbf{S}_{yx} / S_y^2$ denotes the *squared multiple correlation coefficient* of Y on $X_1, X_2, \dots, X_n$. Finally, substituting $\mathbf{g}_x^*$ in (2), we get the *regression estimator*:

$$\bar{y}_{reg} = \bar{y} + (\bar{\mathbf{X}} - \bar{\mathbf{x}})^t \mathbf{S}_{xx}^{-1} \mathbf{S}_{yx}.$$

This estimator is the best in the class g but it is not unique. Any other estimator which attains the minimum variance bound given in (4) is optimum as well and thus it is equivalent to $\bar{y}_{reg}$ in the sense of the efficiency. For instance, the multivariate estimator $\bar{y}_h$ suggested by Abu-Dayyeh *et al.* (2003) is equivalent to $\bar{y}_{reg}$. Moreover, if only two auxiliary variables are considered, $\bar{y}_{reg}$ is equivalent to the estimator $\bar{y}_{pr}$ proposed by Kadilar and Cingi (2005) and to the *ratio-cum-product type estimators* developed in Perri (2005). Finally, we observe that the estimators $\bar{y}_w$ and $\bar{y}_{reg2}$ given in Abu-Dayyeh *et al.* (2003) and Kadilar and Cingi (2005), respectively, are members of the class g but not equivalent to $\bar{y}_{reg}$ and thus, they are not optimum in the class.

Remark. In many circumstances the partial regression coefficients in $\mathbf{g}_x^*$ are unknown. Nevertheless, if we replace such coefficients with suitable estimates, we obtain an estimator that is equivalent to $\bar{y}_{reg}$ at the first order of approximation.

3. Efficiency comparison

In this section we compare the efficiency of the best estimator $\bar{y}_{reg}$ with that of the Olkin (1958) and Ceccon and Diana (1996) multivariate ratio-type estimators. The Olkin's multivariate ratio-type estimator, $\bar{y}_O$, and the multivariate Naik-Gupta estimator, $\bar{y}_{MNG}$, proposed by Ceccon and Diana (1996), are:

$$\bar{y}_O = \bar{y} \sum_{i=1}^p w_i \left(\frac{\bar{X}_i}{\bar{x}_i} \right) \quad , \quad \bar{y}_{MNG} = \bar{y} \sum_{i=1}^p w_i \left[\frac{\bar{X}_i + a_i(\bar{x}_i - \bar{X}_i)}{\bar{X}_i + b_i(\bar{x}_i - \bar{X}_i)} \right]^{k_i}$$

where w_i, a_i, b_i and k_i are real constants to be suitably chosen in such a way to minimize the mean square error.

The expressions for the minimum mean square error of $\bar{y}_O$ and $\bar{y}_{NGM}$ are more complex than the expression for $\bar{y}_{reg}$ defined in (4) because the coefficients of variation of all the considered variables appear. For this reason, the analytical comparison among the estimators is not easy and some simplifications are requested. Following Olkin (1958, p. 158), let us assume some working hypotheses. Let C_0 and C_i denote the coefficients of variation of Y and $X_i, i = 1, 2, \dots, p$. Moreover, let ρ_{0i} be the correlation coefficient between Y and X_i and ρ_{ij} be the correlation coefficient between X_i and $X_j, i \neq j = 1, 2, \dots, p$. Let us assume that $C_0 = C_i = C$ and $\rho_{0i} = \rho_{ij} = \tilde{\rho} > 0$.

With these hypotheses the minimum mean square errors of $\bar{y}_{reg}, \bar{y}_O$ and $\bar{y}_{NGM}$ are:

$$\begin{aligned} MSE(\bar{y}_{reg}) &= \frac{1-f}{n} C^2 \bar{Y}^2 \frac{(1-\tilde{\rho})(1+\tilde{\rho}p)}{1+\tilde{\rho}(p-1)} \\ MSE(\bar{y}_O) &= \frac{1-f}{n} C^2 \bar{Y}^2 \frac{(1-\tilde{\rho})(1+p)}{p} \\ MSE(\bar{y}_{MNG}) &= \frac{1-f}{n} C^2 \bar{Y}^2 \frac{(1-\tilde{\rho})[p(1+\tilde{\rho}) - \tilde{\rho}^2(p-1)]}{p}. \end{aligned}$$

Comparing the mean square errors it is easy to prove that, for $\tilde{\rho} > 0$, the most efficient estimator is always $\bar{y}_{reg}$. In particular, for $\tilde{\rho} > (p - 1)^{-1}$ the following ordering holds:

$$MSE(\bar{y}_{MNG}) > MSE(\bar{y}_O) > MSE(\bar{y}_{reg}).$$

Without the previous hypotheses, the comparison among the estimators is now carried out assuming two auxiliary variables and using the real data set discussed in Perri (2005). For each of the Italian provinces, three correlated variables are considered: the *total amount (tons) of recyclable-waste collection in 2003* (Y), the *total amount of recyclable-waste collection in 2002* (X_1) and the *number of inhabitants in 2003* (X_2). Using the whole data set the following values are obtained: $\bar{Y} = 62621$, $\bar{X}_1 = 55719$, $\bar{X}_2 = 556554$, $C_0 = 1.452$, $C_1 = 1.461$, $C_2 = 1.091$, $\rho_{01} = 0.994$, $\rho_{02} = 0.730$, $\rho_{12} = 0.677$. Thus, for a sample of size $n = 25$, the mean square errors of the estimators are: $MSE(\bar{y}_{MNG}) = 2188413$, $MSE(\bar{y}_O) = 1944566$ and $MSE(\bar{y}_{reg}) = 1490284$, which are consistent with the ordering found for the analytical comparison. Actually, from a simulation study in which the efficiency of the considered estimators was investigated over a wide spectrum of values for both the correlation and the variation coefficients, the estimator $\bar{y}_{reg}$ was always the most efficient one.

We conclude pointing out what seems to be the main result of the paper. At the first order of approximation, if only the auxiliary vector of means, $\bar{\mathbf{X}}$, is used to improve the estimation of the unknown mean $\bar{Y}$, no estimator can be more efficient than $\bar{y}_{reg}$. More efficient estimators may be found only considering a different order of approximation or using a different vector of auxiliary information, such as the vector $(\bar{\mathbf{X}}, \mathbf{S})^t$, where $\mathbf{S} = (S_1^2, S_2^2, \dots, S_p^2)^t$ denotes the variances of the auxiliary variables.

References

- Abu-Dayyeh W.A., Ahmed M.S., Ahmed R.A., Muttalak H.A. (2003) Some estimators of finite population mean using auxiliary information, *Applied Mathematics and Computation*, 139, 287–298.
- Ceccon C., Diana G. (1996) Classi di stimatori di tipo rapporto in presenza di più variabili ausiliarie, *Atti della XXXVIII Riunione Scientifica della SIS*, Rimini, 9-13 Aprile 1996, 119–126.
- Kadilar C., Cingi H. (2005) A new estimator using two auxiliary variables, *Applied Mathematics and Computation*, 162, 901–908.
- Naik V.D., Gupta P.C. (1991) A general class of estimators for estimating population mean using auxiliary information, *Metrika*, 38, 11–17.
- Olkin I. (1958) Multivariate ratio estimation for finite populations, *Biometrika*, 45, 154–165.
- Perri P.F. (2005) Combining two auxiliary variables in ratio-cum-product type estimators, *Atti del Convegno Intermedio della SIS "Statistica e Ambiente"*, Messina, 21-23 Settembre 2005, 193–196.
- Raj D. (1965) On a method of using multi-auxiliary information in sample surveys, *Journal of the American Statistical Association*, 60, 154–165.
- Singh M.P. (1967) Multivariate product method of estimation for finite populations, *Journal of the Indian Society of Agricultural Statistics*, 31, 375–378.