Hacettepe Journal of Mathematics and Statistics
Volume 35 (1) (2006), 111 – 115

IMPROVEMENT IN VARIANCE ESTIMATION USING AUXILIARY INFORMATION

Cem Kadılar* and Hülya Çıngı*

Received 07 : 06 : 2005 : Accepted 25 : 01 : 2006

Abstract

We propose a new estimator for the population variance using an auxiliary variable in simple random sampling. We obtain the mean square error (MSE) equation of the proposed estimator and show that the proposed estimator is more efficient than the traditional ratio and regression estimators, suggested by Isaki [2], under certain conditions. In addition, we support this theoretical result with the aid of a numerical illustration.

Keywords: Variance estimators, Simple random sampling, Mean square error, Auxiliary variable, Efficiency.

2000 AMS Classification: 62D 05, 62 G 05

1. Introduction

Ratio-type estimators take advantage of the correlation between the auxiliary variable, x and the study variable, y. When information is available on the auxiliary variable that is positively correlated with the study variable, the ratio estimator is a suitable estimator to estimate the population variance. For ratio estimators in sampling theory, population information of the auxiliary variable, such as the coefficient of variation or the kurtosis, is often used to increase the efficiency of the estimation for a population variance.

Isaki [2] presented the ratio estimator for the population variance using the auxiliary information, S_x^2, as

$$s_{\text{ratio}}^2 = s_y^2 \frac{S_x^2}{s_x^2}, \tag{1}$$

where s_x^2 and s_y^2 are unbiased estimators of the population variances S_x^2 and S_y^2, respectively. The MSE of this estimator is

$$\text{MSE}(s_{\text{ratio}}^2) \cong \lambda S_y^4[\beta_2(y) + \beta_2(x) - 2\theta] \tag{2}$$

*Hacettepe University, Faculty of Science, Department of Statistics, 06800 Beytepe, Ankara, Turkey. E-mail: (C. Kadılar) kadilar@hacettepe.edu.tr, (H. Çıngı) hcingi@hacettepe.edu.tr

[7], where

$$\lambda = \frac{1}{n},\ \beta_2(y) = \frac{\mu_{40}}{\mu_{20}^2},\ \beta_2(x) = \frac{\mu_{04}}{\mu_{02}^2},\ \theta = \frac{\mu_{22}}{\mu_{20}\mu_{22}} \text{ and } \mu_{st} = \frac{1}{N}\sum_{j=1}^{N}(y_j - \overline{Y})^s(x_j - \overline{X})^t$$

[6]. Here N is the number of units in the population, n is the sample size, $\overline{X}, \overline{Y}$ are the population means and $\beta_2(x), \beta_2(y)$ are the population kurtosis of the auxiliary variate x_i and the variate of interest y_i, respectively.

Isaki [2] also considered the regression estimator for the population variance using an auxiliary variable

$$s_{\text{reg}}^2 = s_y^2 + b(S_x^2 - s_x^2), \tag{3}$$

where b is a constant, which makes the MSE of the estimator a minimum when $b = B = \frac{V_R(\theta-1)}{\beta_2(x)-1}$. Here $V_R = \frac{S_y^2}{S_x^2}$ [1]. The MSE of this estimator is given by

$$\text{MSE}(s_{\text{reg}}^2) \cong \lambda S_y^4 \left\{ [\beta_2(y) - 1] - \frac{(\theta-1)^2}{\beta_2(x) - 1} \right\}. \tag{4}$$

Comparing equations (4) and (2),

$$\begin{aligned} -1 - \frac{(\theta-1)^2}{\beta_2(x)-1} - \beta_2(x) + 2\theta < 0 &\iff -\frac{(\theta-1)^2}{\beta_2(x)-1} + 2(\theta-1) - [\beta_2(x) - 1] < 0 \\ &\iff -\left[\frac{(\theta^*)^2}{\beta_2^*(x)} - 2\theta^* + \beta_2^*(x)\right] < 0 \\ &\qquad (\text{letting } \theta^* = \theta - 1,\ \beta_2^*(x) = \beta_2(x) - 1) \\ &\iff -\left[\frac{\theta^*}{\sqrt{\beta_2^*(x)}} - \sqrt{\beta_2^*(x)}\right]^2 < 0 \end{aligned}$$

provided that $\beta_2^*(x) = \beta_2(x) - 1 > 0$. We deduce that the regression estimator given in (3) is more efficient than the ratio estimator given in (1) when $\beta_2(x) > 1$.

2. The Suggested Estimator

Shabbir and Yaab [8] proposed a ratio-type estimator for the population mean given by

$$\bar{y}_{\text{ST}} = (1 - J)\bar{y} + Jt_b\overline{X}, \tag{5}$$

where J is a constant and $t_b = \frac{\bar{y}}{\bar{x}}\left[\frac{1 + \Psi C_{xy}}{1 + \Psi C_x^2}\right]$. Here $C_x = \frac{S_x}{\overline{X}}$ is the population coefficient of variation of the x_i, $C_{xy} = \rho_{xy}C_xC_y$, $\Psi = \frac{N-n}{Nn}$, ρ_{xy} is the population coefficient of correlation between the x_i and the y_i,$\bar{x}$ and $\bar{y}$ are the sample means of the x_i and y_i, respectively. Note that the optimum value of J is $J^* = \rho C_y/C_x$.

Adapting the estimator of Shabbir and Yaab [8], given in (5), to the estimator for the population variance, we develop the following estimator:

$$s_{pr}^2 = \omega_1 s_y^2 + \omega_2 \frac{s_y^2}{s_x^2}\tau S_x^2, \tag{6}$$

where ω_1 and ω_2 are weights that satisfy the condition: $\omega_1 + \omega_2 = 1$, and $\tau = \frac{1 + \Psi C_{xy}}{1 + \Psi C_x^2}$ is a constant for a fixed sample size. In applications, we suggest to take τ between 0 and 1.

The MSE of the proposed estimator can be found using the first degree approximation in the Taylor series method defined by

$$\text{MSE}(s_{pr}^2) \cong \mathbf{d}\,\mathbf{\Sigma}\,\mathbf{d}', \tag{7}$$

where

$$\mathbf{d} = \left[\left.\frac{\partial h(a,b)}{\partial a}\right|_{S_y^2,\,S_x^2} \quad \left.\frac{\partial h(a,b)}{\partial b}\right|_{S_y^2,\,S_x^2} \right]$$

$$\mathbf{\Sigma} = \begin{bmatrix} V(s_y^2) & \text{cov}(s_y^2, s_x^2) \\ \text{cov}(s_x^2, s_y^2) & V(s_x^2) \end{bmatrix}$$

[9]. Here $h(a,b) = h(s_y^2, s_x^2) = s_{pr}^2$. According to this definition, we obtain $\mathbf{d}$ for the proposed estimator as follows:

$$\mathbf{d} = \begin{pmatrix} \omega_1 + \omega_2\tau & -\omega_2 V_R \tau \end{pmatrix}.$$

We obtain the MSE of the proposed estimator using (7) as

$$\text{MSE}(s_{pr}^2) \cong (\omega_1 + \omega_2\tau)^2 V(s_y^2) - 2(\omega_1 + \omega_2\tau)\omega_2 V_R \tau \text{cov}(s_y^2, s_x^2) + \omega_2^2 V_R^2 \tau^2 V(s_x^2), \tag{8}$$

where

$$V(s_y^2) = \lambda S_y^4[\beta_2(y) - 1]$$
$$V(s_x^2) = \lambda S_x^4[\beta_2(x) - 1]$$
$$\text{cov}(s_y^2, s_x^2) = \lambda S_y^2 S_x^2(\theta - 1)$$

[6]. From (8), we can write

$$\text{MSE}(s_{pr}^2) \cong \lambda S_y^4\{(\omega_1 + \omega_2\tau)^2[\beta_2(y) - 1] - 2(\omega_1 + \omega_2\tau)\omega_2\tau(\theta - 1) + \omega_2^2\tau^2[\beta_2(x) - 1]\}. \tag{9}$$

The optimum values of ω_1 and ω_2 which minimize the MSE of s_{pr}^2 can easily be shown to be:

$$\omega_1^* = \frac{[\beta_2(y) - 1](\tau - 1) + (\theta - 1)(1 - 2\tau) + \tau[\beta_2(x) - 1]}{[\beta_2(y) - 1]\frac{(1-\tau)^2}{\tau} + 2(\theta - 1)(1 - \tau) + \tau[\beta_2(x) - 1]},$$
$$\omega_2^* = 1 - \omega_1^*.$$

Let $\beta_2^*(y) = \beta_2(y) - 1$, $\beta_2^*(x) = \beta_2(x) - 1$ and $\theta^* = \theta - 1$. Using these notations, we can also write

$$\omega_1^* = \frac{\tau[\beta_2^*(y) + \beta_2^*(x) - 2\theta^*] + \theta^* - \beta_2^*(y)}{\tau[\beta_2^*(y) + \beta_2^*(x) - 2\theta^*] + 2\theta^* - 2\beta_2^*(y) + \frac{\beta_2^*(y)}{\tau}}.$$

Note that when $\tau = 1$, or in other words when $\rho_{xy} = C_x/C_y$, we obtain $\omega_1^\dagger = \dfrac{\beta_2^*(x) - \theta^*}{\beta_2^*(x)}$, and $\omega_2^\dagger = \dfrac{\theta^*}{\beta_2^*(x)}$. When we use these expressions instead of ω_1 and ω_2 in (9), we see that the MSE equation of the proposed estimator is as same as the MSE equation of the traditional regression estimator, given in (4).

3. Efficiency Comparison

In this section, firstly we compare the MSE of the proposed estimator with the MSE of the traditional ratio estimator given in (2). We have the condition as follows:

$$\text{MSE}_{\min}(s_{pr}^2) < \text{MSE}(s_{\text{ratio}}^2)$$

holds if

$$W^2\beta_2^*(y) - \beta_2(y) - 2W\omega_2^*\tau\theta^* + 2\theta + (\omega_2^*)^2\tau^2\beta_2^*(x) - \beta_2(x) < 0, \tag{10}$$

where $W = \omega_1^* + \omega_2^*\tau$. When the condition (10) is satisfied, the proposed estimator is more efficient than the traditional ratio estimator given in (1).

Secondly, we compare the equations of the MSE for the proposed estimator and the traditional regression estimator given in (4) as follows:

$$\mathrm{MSE}_{\min}(s_{pr}^2) < \mathrm{MSE}(s_{\mathrm{reg}}^2)$$

holds if

$$(W^2-1)\beta_2^*(y) - 2W\omega_2^*\tau\theta^* + (\omega_2^*)^2\tau^2\beta_2^*(x) + \frac{(\theta^*)^2}{\beta_2^*(x)} < 0. \tag{11}$$

When the condition (11) is satisfied, the proposed estimator is more efficient than the traditional regression estimator given in (3). Note that when $\tau = 1$, the value of the expression on the left hand side of (11) as 0. This means that when $\tau = 1$, there is no difference between the MSE of the proposed estimator and the MSE of the traditional regression estimator, as mentioned in Section 2.

4. Numerical Illustration and Main Results

In this section we use the same data set that was used by Kadilar and Cingi [3]. However, we consider data of only the East Anatolia Region of Turkey, as we are interested in simple random sampling here. We apply the proposed and traditional estimators to data concerning the level of apple production (in 100 tones) (as the variate of interest), and the number of apple trees (as auxiliary variate) in 104 villages in the East Anatolia Region in 1999 (Source: Institute of Statistics, Republic of Turkey). These data, whose statistics are given in Table 1, are used to compute the MSE of the proposed and traditional estimators.

Table 1. Data Statistics

$N = 104$	$\overline{Y} = 6.254$	$\lambda = 0.050$
$n = 20$	$\overline{X} = 13931.683$	$\Psi = 0.040$
$\rho = 0.865$	$S_y = 11.670$	$\theta = 14.398$
$C_y = 1.866$	$S_x = 23026.133$	$V_R = 2.57E-07$
$C_x = 1.653$	$\beta_2(x) = 17.516$	$\omega_1^* = 0.188$
$C_{yx} = 2.668$	$\beta_2(y) = 16.523$	$\omega_2^* = 0.812$

These estimators have been compared with each other with respect to their MSE values for various sample sizes, as shown in Table 2.

Table 2. MSE values of variance estimators and the values of conditions (10) and (11) for various sample sizes.

n	Simple	Ratio	Regression	Proposed	Cond. (10)	Cond. (11)	τ
10	28792.39	9723.55	8632.17	8567.78	-0.623	-0.035	0.9954
20	14396.19	4861.77	4316.09	4299.93	-0.606	-0.017	0.9977
30	9597.46	3241.18	2877.39	2870.79	-0.599	-0.011	0.9986
40	7198.10	2430.89	2158.04	2154.76	-0.595	-0.007	0.9991
50	5758.48	1944.71	1726.43	1724.64	-0.593	-0.005	0.9994
60	4798.73	1620.59	1438.70	1437.67	-0.592	-0.003	0.9996

When we examine Table 2 in detail, we observe that the proposed estimator has the smallest MSE values for all sample sizes. This is an expected result because conditions (10) and (11) are always satisfied for this data set, as partially shown in Table 2. In addition, it is worth pointing out that when the sample size increases, τ gets closer to 1, and that this decreases the difference in the MSE values between the proposed estimator and the traditional regression estimator. It should be noted that when $\tau = 1$, or in other words, when $C_x^2 = C_{xy}$, the MSE of the proposed estimator is equal to the MSE of the traditional regression estimator, as shown theoretically in Section 2 and Section 3. As a result, we propose to take $\tau < 1$ in (6). By these simulation results, we can infer that the proposed estimator is more efficient than the traditional regression estimator when $C_x^2 > C_{xy}$.

5. Conclusion

We have developed a new estimator whose MSE value is smaller than the MSE values of the traditional ratio and of the regression estimators under the conditions (10) and (11), respectively. This theoretical inference has also been illustrated by the result of an application with original data given in Kadilar and Cingi [3]. It is worth pointing out that the proposed estimator is more efficient than traditional estimators in applications.

In future work, we hope to adapt the estimator, presented here, to stratified random sampling, as in Kadilar and Cingi [4], and hope to develop a variance estimator using two auxiliary variables as the estimator in Kadilar and Cingi [5].

References

[1] Garcia, M. R. and Cebrain, A. A. *Repeated substitution method: The ratio estimator for the population variance*, Metrika **43**, 101–105, 1996.

[2] Isaki, C. T. *Variance estimation using auxiliary information*, Journal of the American Statistical Association **78**, 117–123, 1983.

[3] Kadilar, C. and Cingi, H. *Ratio estimators in stratified random sampling*, Biometrical Journal **45**, 218–225, 2003.

[4] Kadilar, C. and Cingi, H. *A new ratio estimator in stratified random sampling*, Communications in Statistics: Theory and Methods **34**, 597–602, 2005.

[5] Kadilar, C. and Cingi, H. *A new estimator using two auxiliary variables*, Applied Mathematics and Computation **162**, 901–908, 2005.

[6] Kendall, M. and Stuart, A. *The Advanced Theory of Statistics: Distribution Theory, (Volume 1)* (Griffin, London, 1963).

[7] Prasad, B. and Singh, H. P. *Some improved ratio-type estimators of finite population variance in sample surveys*, Communications in Statistics: Theory and Methods **19**, 1127–1139, 1990.

[8] Shabbir, J. and Yaab, M. Z. *Improvement over transformed auxiliary variable in estimating the finite population mean*, Biometrical Journal **45**, 723–729, 2003.

[9] Wolter, K. M. *Introduction to Variance Estimation* (Springer-Verlag, 1985).