ISSN:
Print - 2277 - 0593
Online - 2315 - 7461

Journal of Natural Science, Engineering and Technology

A MODIFIED GENERALIZED CHAIN RATIO IN REGRESSION ESTIMATOR

***F. S. APANTAKU, O. M. OLAYIWOLA, A. O. AJAYI AND O. S. JAIYEOLA**

Department of Statistics, College of Physical Sciences, Federal University of Agriculture, PMB 2240, Abeokuta, Ogun State, Nigeria
***Corresponding Author:** fsapantaku@yahoo.com **Tel:**

ABSTRACT

Generalized Chain ratio in regression type estimator is efficient for estimating the population mean. Many authors have derived a Generalized Chain ratio in regression type estimator. However, the computation of its Mean Square Error (MSE) is cumbersome based on the fact that several iterations have to be done, hence the need for a modified generalized chain ratio in regression estimator with lower MSE. This study proposed a modified generalized chain ratio in regression estimator which is less cumbersome in its computation. Two data sets were used in this study. The first data were on tobacco production by tobacco producing countries with yield of tobacco (variable of interest), area of land and production in metric tonnes as the auxiliary variables. The second data were the number of graduating pupils (variable of interest) in Ado-Odo/Ota local government, Ogun state with the number of enrolled pupils in primaries one and five as the auxiliary variables. The mean square errors in the existing and proposed estimators for various values of alpha were derived and relative efficiency was determined. The MSE for the existing estimator of tobacco production gave six values 0.0080, 0.0079, 0.0080, 0.0082, 0.0087 and 0.0093 with 0.0079 as the minimum while the proposed estimator gave 0.0054. The MSEs for the existing estimator for the graduating pupils were 20.73, 11.08, 7.49, 9.96, 18.50 and 33.10 with 7.49 as the minimum while the proposed was 6.52. The results of this study showed that the proposed estimator gave lower MSE for the two data sets, hence it is more efficient.

Keywords: Chain ratio; regression estimator: relative efficiency

INTRODUCTION

Regression estimation is a type of model assisted survey estimation approach. Model-assisted estimation (Sârndal et al., 1992) provides inferences and the asymptotic framework which are design based, with the working model only used to improve efficiency. Thus, the regression estimators are model assisted and design based, but not model dependent. Typically, the linear models are used as a working model in regression estimation. Generalized regression (GREG) estimators (e.g., Cassel et al., 1976; Sàrndal, 1980, 1982) including ratio and linear regression estimators (Cochran, 1977), best linear unbiased estimators (Brewer, 1963; Royall, 1970), and post-stratification estimators (Holt and Smith, 1979), are all based on assumed linear models. Generalized Chain ratio in regression type estimator is efficient for estimating the population mean. This estimator uses two auxiliary variables. Several authors have worked on related topic. The regression type estimators of the population mean or total y of assume advance knowledge of either population

mean $\bar{X}$ or total X of the auxiliary variable x. In the absence of such information a large one of size n' is selected to observe x and thereby to estimate X while a subsample of size n is drawn to measure y. Thus the two-phase regression type estimator of population mean Y is $\bar{y}_{lr} = \bar{y} + \hat{B}(\bar{x}' - \bar{x})$

Suppose that information on yet another auxiliary variable z is available on all units of the population, with population mean $\bar{Z}_N$. Mohanty (1967) suggested the following regression in ratio estimator assuming that the population mean of the second auxiliary variable (z) is known; x being the first auxiliary variable.

$$T_c = \left[\bar{y} + b_{yx}(\bar{x}' - \bar{x})\right]\frac{\bar{Z}}{\bar{z}} \qquad (1)$$

Motivated by Chand (1975) and Kiregyra (1984), Khare et al (2013) now proposed a generalized chain ratio in regression estimator for population mean by using auxiliary characters which is given as follows:

$$T_{l3} = \bar{y} + b_{yx}\left[\bar{x}'\left(\frac{\bar{Z}}{\bar{z}'}\right)^{\alpha} - \bar{x}\right] \qquad (2)$$

The obtained estimator uses an iterative procedure and continued with the process until it converges. This continuous iteration will take a lot of time, hence the need to develop another estimator which will not require an iterative procedure and will satisfy all conditions regardless of the population.

Using the large sample approximation, the expression for bias and mean square error (MSE) of Khare et al (2013) estimator was given as follows:

$$Bias(T_{l3}) = \theta\left[-\mu_{14} + \mu_{15} + \mu_{24} - \mu_{25} - \alpha\mu_{34} + \alpha\mu_{35} + \frac{f'}{n'}\left\{\frac{\alpha(\alpha+1)}{2}C_z^2 - \alpha\rho_{xz}C_xC_z\right\}\right] \qquad (3)$$

$$MSE(T_{l3}) = \bar{Y}^2\left(\frac{1}{n} - \frac{1}{N}\right)C_y^2 - \bar{Y}^2\left(\frac{1}{n} - \frac{1}{m}\right)\rho_{xy}^2C_y^2 + \left(\frac{1}{m} - \frac{1}{n}\right) + \bar{Y}^2\left[\frac{\alpha^2\rho_{xy}^2C_y^2C_z^2}{C_x^2} - \frac{2\alpha\rho_{xy}\rho_{yz}C_y^2C_z}{C_x}\right] \qquad (4)$$

At various transformation of α say
$\alpha = 1$, the MSE becomes a chain ratio in regression estimator
$\alpha = -1$, the MSE becomes a product in regression estimator
$\alpha = 0$, the MSE becomes a regression estimator

METHODOLOGY

Double Sampling Procedure

Double sampling is a sampling method which makes use of auxiliary data where the auxiliary information is obtained through sampling. More precisely, we first take a sample of units strictly to obtain auxiliary information, and then take a second subsample where the variable(s) of interest were observed. It will often be the case that this second sample is a subsample of the preliminary sample used to acquire auxiliary information.

Notations

m – Population Size
n – first phase sample size

Y – second phase sample size
X, Z – variable of interest
X, Y, Z – Auxiliary variables
$\bar{x}, \bar{y}, \bar{z}$ - population means
C_x, C_y, – sample means
β – coefficient of variations
ρ - regression coefficient
- correlation coefficient

Derivation of the Proposed Estimator

Based on the notations and derivation of Khare et al. (2013), $\alpha = \frac{1}{2}$, we have a ratio type square root transformation

If $\alpha = 0$, we have a ratio type square transformation

At various transformation of α, certain conditions about the population must be met,

For example if $\alpha = 1$, the relationship between y and z must be highly positively correlated

$\alpha = \frac{1}{2}$, the relationship between y and z must be linearly and negatively correlated and if $\alpha = 2$, the population must be skewed. Testing for various value of α in the population will take time and this condition about the population may be hypothetical.

Prasad (1989) proposed a ratio estimator in simple random sampling by introducing a shrinkage constant, the estimator was obtained to be

$$\bar{y}_p = k\bar{y}_r = k\frac{\bar{y}}{\bar{x}}\bar{X} \quad (5)$$

Also Kadilar and Cingi (2005) also suggested the use of Shrinkage constant for ratio estimator in stratified random sampling the obtained estimator was

$$\bar{y}_{stp} = K^*\bar{y}_{RC} \quad (6)$$

this work thereby introduced the constant to Khare et al (2013) estimator. Now the obtained

$$T_P = K\left[\bar{y} + b_{yx}\left[\bar{x}'\left(\frac{\bar{Z}}{\bar{z}'}\right)^{\alpha} - \bar{x}\right]\right] \quad (7)$$

$T_P = K(T_{l3})$ estimator for is

$$\in_0 = \frac{\bar{y}}{\bar{Y}} - 1, \in_1 = \frac{\bar{x}}{\bar{X}} - 1, \in_2 = \frac{\bar{x}'}{\bar{X}} - 1, \in_3 = \frac{\bar{z}}{\bar{Z}} - 1$$

and $\in_4 = \frac{\bar{z}'}{\bar{Z}} - 1$

Singh (2003) defined

$$E(\in_j) = 0,\ j = 0,1,2,3,4.$$

$$E(\in_0^2) = \left(\frac{1}{n} - \frac{1}{N}\right)C_y^2,\ E(\in_1^2) = \left(\frac{1}{n} - \frac{1}{N}\right)C_x^2,\ E(\in_2^2) = \left(\frac{1}{m} - \frac{1}{N}\right)C_x^2,\ E(\in_3^2) = \left(\frac{1}{n} - \frac{1}{N}\right)C_z^2,$$

$$E(\in_4^2) = \left(\frac{1}{m} - \frac{1}{N}\right)C_z^2,\ E(\in_0\in_1) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{xy}C_xC_y,\ E(\in_0\in_2) = \left(\frac{1}{m} - \frac{1}{N}\right)\rho_{xy}C_xC_y,$$

$$E(\in_0\in_3) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{yz}C_yC_z,\ E(\in_0\in_4) = \left(\frac{1}{m} - \frac{1}{N}\right)\rho_{yz}C_yC_z,\ E(\in_1\in_2) = \left(\frac{1}{m} - \frac{1}{N}\right)C_x^2$$

$$E(\in_1\in_3) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{xz}C_xC_z, E(\in_1\in_4) = \left(\frac{1}{m} - \frac{1}{N}\right)\rho_{xz}C_xC_z, E(\in_2\in_3) = \left(\frac{1}{m} - \frac{1}{N}\right)\rho_{xz}C_xC_z$$

$$E(\in_2\in_4) = \left(\frac{1}{m} - \frac{1}{N}\right)\rho_{xz}C_xC_z, \text{ and } E(\in_3\in_4) = \left(\frac{1}{m} - \frac{1}{N}\right)C_z^2$$

Using large sample approximation to derive the bias and MSE

$$T = \bar{Y}(1+\epsilon_0) + b_{yx}\left[(1+\epsilon_2)(1+\epsilon_4)^{-\alpha} - (1+\epsilon_1)\right]$$

$$\approx K\left[\bar{Y}(1+\epsilon_0) + \beta\bar{X}\left[\epsilon_2 - \alpha\epsilon_4 - \epsilon_1 + \frac{\alpha(\alpha+1)}{2}\epsilon_4^2 - \alpha\epsilon_2\epsilon_4\right]\right]$$

$$Bias(T_p) = E(T - \bar{Y}) = K[Bias(T_{l3})] + \bar{Y}(K-1) \quad (8)$$

$$MSE(T_p) = E(T_p - \bar{Y})^2 = K^2 MSE(T_{l3}) + 2K(K-1)\bar{Y}Bias(T_{l3}) + \bar{Y}^2(K-1)^2 \quad (9)$$

$$\frac{\partial MSE(T_p)}{\partial K} = K^2 MSE(T_{l3}) + 4K\bar{Y}Bias(T_{l3}) - 2\bar{Y}Bias(T_{l3}) + 2\bar{Y}^2(K-1) = 0 \quad (10)$$

$$K = \frac{\bar{Y}Bias(T_{l3}) + \bar{Y}^2}{MSE(T_{l3}) + 2\bar{Y}Bias(T_{l3}) + \bar{Y}^2} \quad (11)$$

Where $0 < K < 1$

RESULTS AND DISCUSSION

Two datasets were considered for numerical illustration in this study. The first dataset is the yield of tobacco in tobacco producing countries in specified countries in the world for the year 1998 (variable of interest), with area of land and production in metric tonnes as the auxiliary variables. The second data were the number of graduating pupils (variable of interest) in Ado-Odo/Ota local government for the year 2006 in Ogun state with the number of enrolled pupils in primaries one and five as the auxiliary variables. Table 1 and 2 shows the descriptive statistics of the datasets.

Table 1: Descriptive statistics on the production of tobac-

Parameters	Y (Yield in Metric tonnes)	X (Area in Hectares)	Z (Production in Metric tonnes)
N	106	106	106
m	80	80	80
n	50	50	50
Mean (1st phase)	1.54	19743.78	52239.60
Mean (2nd phase)	1.55	19948.94	20757.10
Population mean	1.55	22169.73	50184.13
Standard deviation	0.80	57916.08	253183.6
Covariance	0.51	2.61	5.05

$\rho_{yx} = -0.178$, $\rho_{yz} = -0.143$, $\rho_{xz} = 0.97$, $\beta(Y, X) = -2.824$

Table 2: Descriptive statistics on number of pupils in primary six, five and one

Parameters	Y	X	Z
N	116	116	116
m	80	80	80
n	50	50	50
Mean (1st phase)	88.73	77.66	70.05
Mean (2nd phase)	75.9	69.4	65.18
Population mean	88.10	76.03	69.58
Standard deviation	68.65	54.30	39.09
Covariance	0.78	0.71	0.56

$\rho_{yx} = 0.93, \rho_{yz} = 0.61, \rho_{xz} = 0.72, \beta(Y, X) = 1.08$

Table 3: MSE of the existing and the proposed estimator for the first dataset

Alpha	Existing	Proposed	Relative Efficiency(100%)
0.0	0.0080	0.0054	32.75
0.4	0.0079		31.99
0.8	0.0080		32.67
1.2	0.0082		34.78
1.6	0.0087		38.07
2.0	0.0093		42.12
Bias	2435.24	2035.25	

Table 4: MSE of the existing and proposed estimator for second da-

α	Existing	Proposed	Relative Efficiency(100%)
0.0	20.73		68.54
0.4	11.08		41.11
0.8	7.49	6.52	12.88
1.2	9.96		34.52
1.6	18.50		64.74
2.0	33.10		80.30
Bias	0.99903	0.77	

Table 3 shows the Mean Square Error (MSE) and bias of the proposed and existing estimators for the first dataset. As explained earlier the existing mean square error was derived using iterative procedures thereby resulting in different values of 'alpha', (α) these values were used to obtained the range of MSEs, it was also plotted in figure 1. The proposed method introduced another constant which will not use iterative procedure and obtained a minimum MSE which is minimal than the minimum of the existing estimator. The minimum MSE is 0.0079, which was obtained for the existing when the value of alpha is 0.4. The proposed MSE is 0.0054. It can be seen that the proposed MSE is lesser than the least of the existing Khare et al (2013), thereby the pro-

posed estimator is more efficient. The results in table 4 (second dataset) shows that minimum MSE is 7.49, which was obtained for the existing when the value of alpha is 0.8. The proposed MSE is 6.52. It can be seen that the proposed MSE is lesser than the existing estimators; therefore the proposed estimator is more efficient.

CONCLUSION

A shrinkage generalized chain ratio in regression estimator for population mean has been proposed and its properties have been studied. Comparative study of the proposed estimator has been made with existing Khare et al. (2013). From the numerical examples, we infer that the proposed estimator is more efficient than the existing estimator by Khare et al. (2013), as the mean squared errors of the proposed is minimal than the mean squared errors of the existing estimator by Khare et al. (2013). The shrinkage constant "K" that was introduced eliminated the iterative procedures used to obtain "alpha" which was introduced by Khare et al. (2013). Proposed estimator since does not require iterative procedure.

REFERENCES

Brewer, K.R., Hanif, M. 1970. Durbin's new multistage variance estimator. *Journal of the Royal Statistical Society, Series B.* 32(2): 302-311.

Cassel, C.M., Särndal, C.E., Wretman, J.H. 1976. Some results on generalized difference estimation and generalized regression estimation for finite populations, *Biometrika* 63(3): 615–620. https://doi.org/10.1093/biomet/63.3.615.

Chand, L. 1975. Some Ratio-type Estimators based on two or more Auxiliary Variables. Unpublished Ph.D. Dissertation. Iowa State University, Iowa.

Cochran, W.G. 1977. Sampling Techniques (3rd Edition), New York, Wiley.

Holt, D., Smith, T.M.F. 1979. Post-stratification, *J.R. Stat. Soc., A* 142: 33-36.

Kadilar, C., Cingi, H. 2005. Ratio Estimators in Stratified Random Sampling. *Communications in Statistics-Theory and Methods* 34: 597-602.

Khare, B.B., Srivastava, U., Kumar, K. 2013. A generalized chain ratio in regression estimator for population mean using two auxiliary characters in sample survey. *J. Sci. Res.* 57: 147–153.

Kiregyera, B. 1984. Regression-type estimator using two auxiliary variables and model of double sampling from finite populations. *Metrika* 31: 215-226.

Mohanty, S. 1967. Combination of Regression and Ratio Estimate. *Journal of Indian Statistical Association* 5: 16-19.

Prasad, B. 1989. Some improved ratio type estimators of population mean and ratio in finite population sample surveys. *Communication Statistical theory Method* 18(1): 379-392.

Royall, R.M. 1970. On finite population theory under certain linear regression models. *Biometrika* 57: 377-387.

Särndal, C.E. 1980. On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika* 67:639–650.

Särndal, C.E. 1982. Implications of survey design for generalized regression estimation of linear functions. *J. Stat. Plan. Infer* 7: 155–

170.

Särndal, C.E., Swensson, B., Wretman J.H. 1992. Model assisted survey sampling. Springer, New York.

Sarndal, C.E., Swensson, B. 1987. A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistics Review* 55: 279-294.

Singh, S. 2003. Advanced sampling theory with applications: how Michael 'selected' Amy. 2nd Edition. Punjab Agricultural University Press.

Srivastava, S.R., Khare, B.B. 1990. A generalized chain ratio estimator for mean of finite population. *J. Ind. Soc. Agri. Statist.* 42 (1): 108-117.

(*Manuscript received: 24th June, 2018; accepted: 24th September, 2019)*