Efficacy of Auxiliary Information for Determining Robust Estimators

Abba Badamasi, Babagana Modu

Abstract—Auxiliary variables plays a critical role in sampling surveys towards estimating population parameter with high degree of precision. In this study, we proposed multivariate auxiliary information and explore the effect of auxiliary variables to estimate the population mean of the study variable. The simple random sampling scheme was employed and selected the sample considered in this study. To achieve this, we used a data from 2010-2012 of students enrolled for admission into Senior Secondary School per local government in Kano State, Nigeria. The 2012 admission was considered the study variable Y, while 2010 and 2011 are the auxiliary variables X1 and X2 respectively. A very strong correlation was observed between each of the auxiliary variables and Y. The estimators such as sample mean and multivariate ratio estimators were computed from the samples selected. Subsequently, the results showed that multivariate ratio estimator outperformed the classical sample mean estimator.

Index Terms— Multivariate ratio estimator, simple random sampling, auxiliary information, correlation, study variable, population parameter, efficiency.

---------- ◆ ----------

1 INTRODUCTION

In our everyday activities, we are always encountering estimation problems. This has been a serious challenges, coming from almost all the sectors of our society be it economy or social. The estimation of country standard of living; life expectancy; fertility rate; birth rate; gross domestic product among other. However, statistics as a discipline provided solutions to this problems by deployment of the methodology called sampling theory.

This paper focuses on application of an auxiliary variables towards increasing precision of an estimator. The auxiliary variables have been utilized by many reseachers to increase precision of estimators, and especially for regression and ratio estimators [1]. However, doing so would mitigate the systematic bias that could occurred to the survey, when simple random sampling (SRS) scheme was used [2].

In a sample survey, researchers usually make use of supplementary information called auxiliary variables to obtain an improved designs or more efficient estimator. There were vast and rich litereature on the application of an auxiliary variables. But, in this paper, we are considering very few that are closely relevant to this work. The effect of correlation level on the use of auxiliart variable in double sampling for regression estimation was discussed [3]. Accuracy of estimators for the population parameter must be depends on how best the linear relationship between response and the set of an auxiliary variables [4]. In other words, the ratio estimator is used when the variable of interest is positively correlated with each of the auxiliary variables. The chain ratio estimator was obtained by [5] using auxiliary information and deduced precised mean squared errors of the population mean.

- *Badamasi Abba is currently pursuing PhD in Statistics at Ahmadu Bello University Zaria, Nigeria. E-mail: badamasiabba@gmail.com*
- *Babagana Modu is currently pursuing PhD in Data Science at University of Bradford, United Kindgom. E-mail: b.modu@bradford.ac.uk*

New ratio estimators using coefficient of variation and kurtosis of auxiliary information was presented by [6]. This work was aimed to justify the efficacy of supplementary information. However, it is also reach to enlighten researchers, policy makers and others interested in determining precise estimators of population parameter are encouraged to always use auxiliary information.

The remaining parts of this paper is structured as follows. Section 2, describes the materials and methods comprises of population studied, sampling procedure and the sample selected and illustration of the mathematical framework. In Section 3, we presents the data analysis and the discussion of results. We conclude this work in Section 4, and unravel prospective study.

2 MATERIALS AND METHODS

This section presents step by step descriptions of the methodologies of this paper. The aim of this work is to explore the efficacy of the auxiliary variables for the determination of precise estimator of population parameter. We therefore, choose to obtain information on student's entrollment for admission into senior secondary schools across Kano State, Nigeria. Thereafter, we explicitly illustrate the mathematical framework to support the data analysis.

2.1 Source of data

The data used for this study was collected from Kano State Senior Secondary Schools Management Board (KSSSSMB), Nigeria. Department of planning, research and statistics of the board are responsible for keeping record of that nature. In the study area, that is Kano State, there are 44 local government areas clustered around the state. The data is for three (3) years students admission into Senior Second-ary Schools (SSS), that is, the number of students enrolled for admission into senior secondary schools (SSS) by the Board from 2010-2012 per each local government (see the details in Table I).

Table I. The distribution of Students admission into Senior Secondary Schools (SSS) per Local Government, across the Kano State, starting from 2010 to 2012.

S/No	Local Govt	2010	2011	2012
1.	Bagwai	234	343	432
2.	Bichi	1565	1701	1723
3.	Kunchi	196	198	266
4.	Tsanyawa	448	477	555
5.	Dala	3442	3882	3697
6.	Gwale	6638	7061	7322
-	-	-	-	-
-	-	-	-	-
-	-	-	-	-
42.	Garko	7322	7061	6638
43.	Sumaila	1413	1374	1115
44.	Wudil	1089	1431	1443

Source: Kano State Senior Secondary School Management Board (KSSSSMB).

2.2 Population Parameter

This is a real value function say $\theta(Y_i : i = 1,2,\dots,N)$ which describes the characteristic of the population. For example populations mean, variance, coefficient of variation etc.

2.3 Parameter Estimation

This is the process of using representative part of a population (i.e. sample) to estimate the unknown population parameter. An estimate of the population parameter given by a single value number is called point estimate. While an estimate of a population parameter given by two numbers which the parameter is consider to lie is called an interval estimate.

2.4 Unbiased Estimator

We are interested to show that $\hat{\bar{Y}}_{MR}$ is an unbiased estimator for $\bar{Y}$.

Therefore, $\hat{\bar{Y}}_{MR} = \sum_{i=1}^{p} W_i r_i \bar{X}_i$ $\frac{\bar{Y}}{\bar{Y}} = 1$

Multiplying the right hand side (RHS) by $\frac{\bar{Y}}{\bar{Y}}$ the quantity we have:

$$\hat{\bar{Y}}_{MR} = \sum_{i=1}^{p} W_i r_i \bar{X}_i \frac{\bar{Y}}{\bar{Y}} = \sum_{i=1}^{p} W_i r_i \frac{\bar{Y}}{R_i}$$

By the taking expectation of both sides, we can obtain the resultant equation

$$E\left(\hat{\bar{Y}}_{MR}\right) = E\left(\sum_{i=1}^{p} W_i r_i \bar{X}_i \frac{\bar{Y}}{\bar{Y}}\right) = E\left(\sum_{i=1}^{p} W_i r_i \frac{\bar{Y}}{R_i}\right)$$

$$= \sum_{i=1}^{p} W_i E(r_i) \frac{\bar{Y}}{R_i} = \sum_{i=1}^{p} W_i R_i \frac{\bar{Y}}{R_i} = \bar{Y} \sum_{i=1}^{p} W_i$$

The constraint regarding the sum of p-weight can be given as

$$\sum_{i=1}^{p} W_i = 1$$

Then, we have the following results of the expected value of the estimate of population mean using the direct methods as:

$$E\left(\hat{\bar{Y}}_{MR}\right) = \bar{Y}$$

Therefore, the $\hat{\bar{Y}}_{MR}$ is the required unbiased estimator of the population mean.

2.5 Multivariate Ratio Estimator

This is concern with the estimation of population parameter using ratio estimator. It involes using multivariate auxiliary variables to increases the precision of the estimate. The generalized multivariate ratio estimator for the population mean is given by [2] as:

$$\hat{\bar{Y}}_{MR} = W_1 r_1 \bar{X}_1 + W_2 r_2 \bar{X}_2 + \cdots + W_p r_p \bar{X}_p = \sum_{i=1}^{p} W_i r_i \bar{X}_i$$

Then, estimation of the population mean or total using multi-auxiliary scenario increases the precision of the estimator. Thereby, the more number of auxiliary information for estimating the population parameter, higher precision could be achieved. In multivariate estimation we have the following model, suppose

$$Y_1, Y_2, \dots\dots\dots, Y_N \quad \bar{Y} \text{ is unknown}$$

Considering the matrix of order p by N auxiliary variables and the corresponding p-means column vector given as

$$\begin{pmatrix} X_{11} & X_{12} & \cdots X_{1N} \\ X_{21} & X_{22} & \cdots X_{2N} \\ \vdots & \vdots & \\ X_{p1} & X_{p2} & \cdots X_{pN} \end{pmatrix} \qquad \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{pmatrix}$$

We defined R1, R2… Rp to represents the ratios of response variable mean to the respective means of the auxiliary variables given below.

$$\begin{pmatrix} R_1 = \frac{\bar{Y}}{\bar{X}_1} \\ R_2 = \frac{\bar{Y}}{\bar{X}_2} \\ \vdots \\ R_p = \frac{\bar{Y}}{\bar{X}_p} \end{pmatrix}$$

The variance-covariance matrix S is given as

$$S = \begin{pmatrix} S_{00} & S_{01} & \cdots S_{0p} \\ S_{10} & S_{11} & \cdots S_{1p} \\ \vdots & \vdots & \\ S_{p0} & X_{p1} & \cdots X_{pp} \end{pmatrix}$$

The subscripts 0, 1, 2, 3…, p refers to $y, x_1, x_2, \dots, x_p$ respectively are the p auxiliary variables. The Multivariate Ratio Estimator of $\bar{Y}$ is given by

$$\bar{Y}_{MR} = W_1 r_1 \bar{X}_1 + \cdots + W_n r_n \bar{X}_n$$

$$\hat{\bar{Y}}_i = \frac{\bar{y}}{\bar{x}_i}\bar{X}_i$$

Represents the components of the population mean ratio estimate affiliated to the i^{th} auxiliary variable and W_i is the weight which maximize the precision of $\bar{Y}_{MR}$ subject to constraint

$W = (W_1, \ldots, W_n)$ Such that $\sum_{i=1}^{p} W_i = 1$

$$r_i = \frac{\bar{y}}{\bar{x}_i}$$

Also, represents the estimate of R_i, the population ratio for $i = 1,2,3, \ldots, p$ and $\bar{X}_i$ is the known population mean for the p-auxiliary variables.

$$W_1 = \frac{(a_{22} - a_{12})}{\{(a_{11} - a_{12}) + (a_{22} - a_{12})\}}$$

$$W_2 = \frac{(a_{11} - a_{12})}{a_{11} + a_{22} - 2a_{12}}$$

2.5 Multi-auxiliary variables

In a survey research, there are times when information is available on every unit in the population. If a variable that is known for every unit of the population is not a variable of interest but instead employed to improve the sampling plan or to enhance the estimation of the variable of interest. It is called an Auxiliary variable. The term Multi-auxiliary variables are most commonly associated with the use of such variables, available for all units in the population, in ratio estimation, regression estimation and extensions (calibration estimation). The ratio or multivariate ratio estimator is most widely used estimator that takes advantage of an Auxiliary or Multi-auxiliary variable(s) [7].

2.6 Simple Random Sampling (SRS)

This is a technique for selecting n (sample) out of the N (Population), such that every one of the distinct sample has an equal chance of being drawn [8]. In this approach, the sampling can be done with or without replacement. However, the selection of the samples when the population is larges. Therefore, the appropriate scheme to follow and achieve this is the application of random number tables or statistical software packages. But in this work, we are deploying random number tables, and the sampling scheme is without replacement.

3 DATA ANALYSIS AND DISCUSSION OF RESULTS

In this section, we presents the analysis of data starting with selecting the appriopriate units from the target population, see Table II.

3.1 Simple Random Sampling

Seclecting sample from population, particularly large, a random number tables or computer software packages are basically in this case [9]. In selecting the samples, we used random numbers generated through a computer program. Using the random numbers tables, each of the digits from 0 to 9 has a 1 in 10 chance of appearing. We select a sample of size 11 from a population of 44 local governments. Doing so, we label the 44 local governments with digits starting from 0 to 43. Then, we start with any three columns together anywhere in the table and select eleven 2-digit numbers. We reject a number that exceeds 43, and then we discarded those numbers repeated, move to another set of three columns. We select again some more numbers in the same manner, and continue the procedure until 11 numbers are selected. The numbers can also be selected, three at a time, from the rows or both from the rows and columns of the random numbers. In Table II, we presents the results of using SRS for selecting the appropriate samples to be used in the analysis

Table II. The results of SRS, showing the number of units selected to represents the population under study.

S/No	SRS Number	Y	X1	X2
1.	18	1796	2115	1483
2.	42	678	611	517
3.	05	3697	3882	3442
4.	26	1695	1912	1479
5.	38	685	645	571
6.	14	525	524	561
7.	22	883	824	951
8.	16	913	734	634
9.	09	551	558	446
10.	06	7322	7061	6638
11.	08	1413	1374	1115
Mean		1832.55	1840	1621.55

SRS: Simple Random Sampling Y: Represents 2012 admission, X1: Represents 2011 admission and X1: Represents 2010 admission

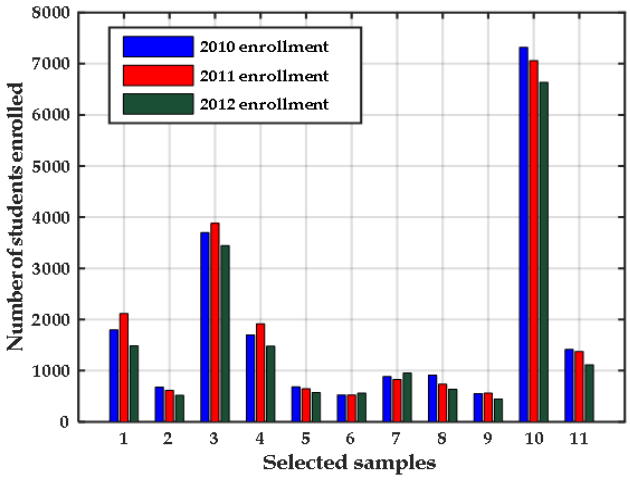

Figure 1. We presents the samples selected

To understand the distribution of the selected sample, we therefore represents the information in Table II, using a suitable chart (see Figure 1). We used the Matlab software and produced the plot. Depicted from the plot, the enrollment for admission was in spike in sample 10 and relatively similar on average for 2, 5, 6, 7, 8 and 9. While, 1, 3, 4 and 11 could be classified using the same cluster.

The computation of enteries in the variance-covariance matrix S can be obtain using the following:

$$S_{00} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = 4155564$$

$$S_{00} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = 4155564$$

$$S_{22} = \frac{\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2}{n-1} = 3503018$$

Hence, the subsequent enteries can be obtain using similar treatment, and the variance-covariance is S given by

$$= \begin{pmatrix} 4155564 & 4077889 & 3808777 \\ 4077889 & 4029591 & 3738508 \\ 3803777 & 3738508 & 3503018 \end{pmatrix}$$

Using the enteries of variance-covariance S, and compute a_{11}, a_{12} and a_{22} as follows: a_{11} = 30745.9190, a_{12} = -3195.2204 and a_{22} = 20731.6641 and then the weights are W_1 = 0.413 and W_2 = 0.587. Therefore, the estimate of mean using multivariate ratio estimate is defined as follows:

$$\hat{\bar{Y}}_{MR} = W_1 r_1 \bar{X}_1 + W_2 r_2 \bar{X}_2$$

= 1576 students / local government

3.4 Variance of the estimate

The computation of the variance is explicitly presented using the formula as:

$$V\left(\hat{\bar{Y}}_{MR}\right) = (W_1^2 a_{11} + W_2^2 a_{22} + 2W_1 W_2 a_{12})\frac{(1-f)}{n}$$

= 738.992, hence the standard deviation can be obtain by taking the square root given as:

$$S.D\left(\hat{\bar{Y}}_{MR}\right) = \sqrt{V\left(\hat{\bar{Y}}_{MR}\right)} = 27.184$$

At 5% level of significance, the confidence interval for mean of Y, is [1559.97, 1594.03]. Hence, the actual value falls within the interval and thus, we say that the estimator is robust.

The quality of good relationship between the response variable and the auxiliary information, as the pre-condition to obtain more precisie estimator [2]. We therefore computes, the correlation between the Y against X1 and X2, and presents the result using the following correlation matrix. The result shows a strong positive relationship between the variables as: corr(Y, X1) = 0.997, corr(Y, X2) = 0.998 and corr(X1, X2) = 0.995 respectively.

Correlation Matrix	Y	X1	X2
Y	1.000	0.997	0.998
X1	0.997	1.000	0.995
X2	0.998	0.995	1.000

We compare the method by Olkin's against the direct method for estimating true population mean. The results shows that Olkin's estimator gives an estimate of the population mean with a very small variance and very small width of the confidence interval as presented in the Table III below.

Table III. The results presents summary statistics for selecting best performing method.

Approach	Direct Method	Olkin's Estimator
Mean	1832	1576
Variance	4155564	738.992
SD	2038.52	27.184
95% CI	[463.03-3202.04]	[1516.20-1635.80]

SD: Standard Deviation **CI:** Confidence Interval

4. Conclusion

From the results of the analysis, it can be concluded that multivariate ratio method of estimation is more efficient than the direct sample mean method of estimation under simple random sampling scheme. Thus we suggested multivariate ratio type estimator is preferable over the direct sample mean technique when each of the auxiliary variables are highly correlated with the reponse variable. In the future work, we proposed to apply stratified sampling scheme to multivariate auxiliary variables and compare it with simple random sampling using simulation.

References

[1] Cochran, W. G. (2007). Sampling techniques. John Wiley & Sons.
[2] Olkin, I. (1958). Multivariate ratio estimation for finite populations. Biometrika, 45(1/2), 154-165.
[3] Agunbiade, D. A., & Ogunyinka, P. I. (2013). Effect of correlation level on the use of Auxiliary variable in Double sampling for regression estimation. Open Journal of Statistics, 3(05), 312.
[4] O.O. Ngesa (2012) Multivariate Ratio Estimator of a Population Total under Stratified Random Sampling, *Open Journal of Statistics*, 2, 300-304.
[5] Lu, J., Yan, Z., Ding, C., & Hong, Z. (2010). The chain ratio estimator using two auxiliary information. In International Conference on Computer and Communication Technologies in Agriculture Engineering (pp. 586-589).
[6] Lu, J., Yan, Z., Ding, C., & Hong, Z. Some new ratio estimators using coefficient of variation and Kurtosis of auxiliary variate. In 2010 International Conference on Computer and Communication Technologies in Agriculture Engineering.
[7] M. P. Cohen, Auxiliary Variable, Encyclopedia of Survey Research Methods
[8] Singh, H. P., & Solanki, R. S. (2013). A new procedure for variance estimation in simple random sampling using auxiliary information. Statistical Papers, 54(2), 479-497.
[9] Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.