Global Scientific JOURNALS

GSJ: Volume 7, Issue 3, March 2019, Online: ISSN 2320-9186
www.globalscientificjournal.com

# APPLICATION OF REGRESSION TYPE ESTIMATOR IN DOUBLE SAMPLING SKILLS TO STUDENTS' ENROLLMENT IN OYO STATE

**C. G. Udomboso[1], O. B. Akanbi[2], S. A. Afolabi[3]**

Department of Statistics,
University of Ibadan,
Ibadan, Nigeria.
E-mail: [1]cg.udomboso@gmail.com, [2]muhdbashola@yahoo.com, [3]Afsab4ever@yahoo.com

**Abstract** - This research derived the precision using Regression Estimation technique with the application of secondary data obtained using the number of students enrollment in 2015 (Auxiliary variable "x") and 2016 (response variable "y") respectively in secondary schools of Ibadan, Oyo State, Nigeria for the purpose of obtaining average enrollment figures in the selected state in order to know the bright future of secondary schools in Oyo State in general and to establish the empirical comparison of the optimum variances in obtaining the most efficient estimator in order to satisfy the condition; $\rho^2 \geq \frac{4C_1C_2}{(C_1+C_2)^2}$ based on the coefficients of Variation for the validity and reliability, the relative efficiency was also determined based on the conditions attached to the supremacy in terms of the estimated mean square error (variance) whereby the regression line does not pass through the origin from the graph of Relative Efficiency (R.E) against Correlation Coefficients ($\rho$) that maintain inverse relation. Proper conclusions and recommendations are made based on findings from the analysis in terms of adequate record keeping among the contemporary states within.

**Keywords:** Sampling, Regression Estimation, Precision, Relative Efficiency.

## 1. Introduction

Double Sampling can also be referred to as the Two-phase Sampling. Auxiliary information has always been seems effective in increasing the precision of estimates in survey sampling in which the precision of estimates of the mean of the variable of interest is increased by the presence of highly correlated auxiliary variables. There are situations when auxiliary information is available at the population level and the cost of collecting the variable of interest per unit is affordable, then single-phase sampling is more appropriate. But when prior information on auxiliary variable is lacking, then it is neither practical nor economical to conduct a census for this purpose. Therefore, an appropriate technique employed to get estimates of auxiliary variables on the basis of samples is Double-phase sampling. This technique is used when the cost of obtaining estimates of the variable of interest directly from the population is expensive or impracticable.

The theory of Double sampling is presented under the assumption that one of the sample is a subsample of the other. This type of sampling technique is called Double Sampling Technique or Sampling followed by sub-sampling. When the mean $\overline{X}$ of an auxiliary variable is completely unknown, double sampling mechanism can be adopted. Cochran, (1977).

Watson (1973) was the first to present an early theory of Double sampling with regression. This was immediately followed by Neyman (1938) who developed the theory for double sampling with stratification after a problem was posed to him in a meeting at the United States Department of Agriculture. The basic problem was that a survey was to be undertaken to determine the population value of a character of interest, say y.

Sukhatme (1962) presented several ratio estimators in two-phase sampling. He showed that one of them is an unbiased estimate and follows direct from the one presented by Hartley and Ross for single-phase sampling. He further discussed the efficiency of two-phase sampling with respect to single-phase sampling under a single cost function. Rao (1973) applied the scheme to Stratification, non-response and analytic comparisons.

Senapati et al, (2006) proposed a general class of estimators in two-phase sampling for finite population mean when the mean of the main auxiliary variable X is unknown but that of an additional auxiliary variable z is known. They concluded that proposed class of estimators is superior to some of the previously studied class of estimators under minimum variance criterion.

Naqvi et al, (2013) proposed a regression type estimator for two phase sampling when prior knowledge of auxiliary variables is not at all available at population level.

This research aimed at obtaining the average secondary school enrollment figures in Oyo state using regression estimator to demonstrate the relative efficiency of double sampling for regression estimation when population parameter of the auxiliary variable is unknown in order to satisfy the existing condition.

## 2. Methodology

### 2.1 Sample Size Determination

In many applications of the regression method of estimation, the population mean $\overline{X}$ of the auxiliary variable is unknown. In double sampling, we take a first preliminary sample of size $n'$ where $x_i$ is a measure, in second sampling we take sample of size $n$ from $n'$ such that $n = k n'$ where $k < 1$ and specified in advance so that $n$ will not be a random variable. From the units involved in $n$ sample, we sample $x_i$ and $y_i$.

In the first (large) sample size $n'$ or $n_1$, measure of $x_i$ is assumed; In the second, a random sub-sample of size $n_2 = kn_1$ ($k < 1$) where k is chosen in advance so that n will not be a random variable. The measures of both $x_i$ and $y_i$ were known.

In this research, $n' = 120$; $N_1 = 50$ and $N_2 = 50$ such that:

$$C = c_1 n_2 + c_2 n_1 \qquad (1)$$

Since $N_1 = N_2 = 50$, therefore;

$$n_1 = n_2 \equiv n = \frac{C}{c_1 + c_2} \qquad (2)$$

Hence $n = 80$ and $n_1 = 40$ while $n_2 = 40$

Where;

$C = C_0 = 50,000$; Overall Initial Cost.

$c_1 = C_1 = 60$; Cost per unit for response variable.

$c_2 = C_2 = 600$; Cost per unit for auxiliary variable.

### 2.2 Estimation and Derivation

**Estimate of Beta**

$$\beta = b = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{S_x^2} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} \qquad (3)$$

Where;

$S_y^2 = 9033.179$; $S_x^2 = 9983.266$ and $S_{xy} = 8971.619$

**Estimate of Mean and Variance**

Mean

$$\bar{y}_m = \bar{y} - b\left(\bar{x} - \bar{X}'\right) \qquad (4)$$

Variance

$$V\left(\bar{y}_m\right) = V_1 E_2\left(\bar{y}_m\right) + E_1 V_2\left(\bar{y}_m\right)$$

$$V\left(\bar{y}_m\right) = \left(\frac{1}{n_1} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n_2} - \frac{1}{n_1}\right)\left(1-\rho^2\right) S_y^2$$

$$V\left(\bar{y}_m\right) = \frac{1}{n_1} S_y^2 - \frac{1}{N} S_y^2 + \frac{1}{n_2}\left(1-\rho^2\right) S_y^2 - \frac{1}{n_1}\left(1-\rho^2\right) S_y^2$$

Hence;

$$V\left(\bar{y}_m\right) = \frac{1}{n_1}\rho^2 S_y^2 - \frac{1}{N} S_y^2 + \frac{1}{n_2}\left(1-\rho^2\right) S_y^2 \qquad (5)$$



**Estimate of Optimum Variance**

The cost function of taking the first preliminary sample of size $n'$ and second sub-sample of size $n$ from $n'$ is given as;

$$C_0 = c_1 n_2 + c_2 n_1$$

Using the method of Lagrange multiplier on equation (5) to derive optimum values

$$G(\lambda, n_1, n_2) = V\left(\bar{y}_m\right) + \lambda(\mathrm{C}_1 n_2 + \mathrm{C}_2 n_1 - \mathrm{C}_0)$$

$$G(\lambda, n_1, n_2) = \frac{\rho^2 S_y^2}{n_1} + \frac{\left(1-\rho^2\right) S_y^2}{n_2} - \frac{S_y^2}{N} + \lambda(\mathrm{C}_1 n_2 + \mathrm{C}_2 n_1 - \mathrm{C}_0)$$

By appropriate partial derivatives;

$$\frac{\partial G(\lambda, n_1, n_2)}{\partial n_1} = \frac{-\rho^2 S_y^2}{(n_1)^2} + \lambda C_2 \qquad (*)$$

$$\frac{\partial G(\lambda, n_1, n_2)}{\partial n_2} = \frac{-\left(1-\rho^2\right)S_y^2}{(n_2)^2} + \lambda C_1 \qquad (**)$$

Setting (*) and (**) to zero, we have

$$\frac{\lambda C_1 n_2^2}{\lambda C_2 n_1^2} = \frac{\left(1-\rho^2\right)S_y^2}{\rho^2 S_y^2}$$

Multiplying the above by $\frac{C_2}{C_1}$

$$\left(\frac{n_2}{n_1}\right)^2 = \frac{C_2\left(1-\rho^2\right)}{C_1\rho^2}$$

$$\left(\frac{n_2}{n_1}\right) = \left[\frac{C_2\left(1-\rho^2\right)}{C_1\rho^2}\right]^{1/2}$$

$$n_1 = n_2 = \left[\frac{C_2\left(1-\rho^2\right)}{C_1\rho^2}\right]^{1/2}$$



Substituting in the cost function; $C_0 = c_1 n_2 + c_2 n_1$

$$n_1 = \frac{\rho C_0}{\sqrt{\left(1-\rho^2\right)C_1 C_2} + \rho C_2}$$

Assuming N is large in equation (5), it becomes;

$$V\left(\bar{y}_m\right) = \frac{\rho^2 S_y^2}{n_1} + \frac{\left(1-\rho^2\right)S_y^2}{n_2}$$

So that;

$$V_{opt}\left(\bar{y}_m\right) = \frac{S_y^2}{C_0}\left[\sqrt{\left(1-\rho^2\right)C_1} + \sqrt{\rho^2 C_2}\right]^2 \qquad (6)$$

**Single Sampling for Regression Estimator**

If all resources are devoted to a single sample then C = $C_1n_2$ or $C_2n_1$ with no regression adjustment, the sample has size $n_1 = n_2 = \frac{C_0}{C}$, the variance of a simple sample is;

$$V\left(\bar{y}_m\right) = (1-f)\frac{S_y^2}{n_2}$$

$$V\left(\bar{y}_m\right) = \left(1-\frac{n_2}{N}\right)\frac{S_y^2}{n_2}$$

$$V\left(\bar{y}_m\right) = \left(\frac{1}{n_2}-\frac{1}{N}\right)S_y^2$$

By applying $n_1 = n_2 = \frac{C_0}{C}$

$$V_{\min}\left(\bar{y}_m\right) = \left(\frac{1}{C_0/C}-\frac{1}{N}\right)S_y^2$$

$$V_{\min}\left(\bar{y}_m\right) = \frac{C}{C_0}S_y^2 - \frac{S_y^2}{N}$$



Using $C_2$;

$$V_{\min}\left(\bar{y}_m\right) = \frac{C_2}{C_0}S_y^2 - \frac{S_y^2}{N} \qquad (7)$$

**Double Sampling for Regression Estimator**

It follows then that double sampling for regression mean will be more precise than sample mean (in terms of their variances) if $V_{\min}\left(\bar{y}_m\right) > V_{\min}\left(\bar{Y}\right)$.

Therefore;

$$V_{\min}\left(\bar{y}_m\right) - V_{\min}\left(\bar{Y}\right) > 0$$

$$\frac{C_1}{C_0}S_y^2 - \frac{1}{N}S_y^2 - \left\{\frac{S_y^2}{C_0}\left[\sqrt{\left(1-\rho^2\right)C_1} + \sqrt{\rho^2 C_2}\right]^2 - \frac{S_y^2}{N}\right\} > 0$$

$$\frac{C_1}{C_0}S_y^2 - \frac{S_y^2}{C_0}\left[\sqrt{\left(1-\rho^2\right)C_1} + \sqrt{\rho^2 C_2}\right]^2 > 0$$

$$C_1\left[\sqrt{\left(1-\rho^2\right)C_1} + \sqrt{\rho^2 C_2}\right]^2 > 0$$

$$\sqrt{C_1} - \sqrt{\rho^2 C_2} > \sqrt{\left(1-\rho^2\right)C_1}$$

Squaring both sides;

$$\left(\sqrt{C_1} - \sqrt{\rho^2 C_2}\right)^2 > \left(\sqrt{\left(1-\rho^2\right)C_1}\right)^2$$

$$2\rho\sqrt{\rho^2 C_1 C_2} < (C_1 + C_2)\rho^2$$

Squaring both sides;

$$4\rho^2 C_1 C_2 < (C_1 + C_2)^2 \rho^4$$

$$\frac{4C_1 C_2}{(C_1 + C_2)^2} < \rho^2$$



Therefore;

$$\rho^2 \geq \frac{4C_1 C_2}{(C_1 + C_2)^2} \qquad (8)$$

**Relative Efficiency of Double Sampling**

The correlation is given thus;

$$\rho = \frac{S_{xy}}{\sqrt{S_x^2}\sqrt{S_y^2}} \qquad (9)$$

By the application of efficiency of Double sampling for regression estimator, the relative efficiency of double sampling is given by;

$$R.\mathrm{E} = \frac{1}{\left[\left[1-\rho^2\right] + \rho\sqrt{\frac{C_2}{C_1}}\right]^2} \qquad (10)$$

## 3. Results and Discussions

Table 1: Estimation Output

| Statistics | Random | Optimum | Single Sampling | Double Sampling |
|---|---|---|---|---|
| Mean | 167 | 167 | 167 | 167 |
| Variance | 135.498 | 119.151 | 18.066 | 0.331 |
| Standard Error | 11.64 | 10.916 | 4.25 | 0.575 |
| Coefficient of Variation | 6.97% | 6.54% | 2.5% | 0.34% |

**Optimality Condition**

$$\rho^2 \geq \frac{4C_1C_2}{(C_1+C_2)^2}$$

$$0.945^2 \geq \frac{4\times 60\times 600}{(60+600)^2}$$

$$0.893 \geq 0.331$$

Table 2: Estimation of Relative Efficiency and Correlation Coefficients

| $\rho$ | R. E. |
|---|---|
| 0.1 | 58.17 |
| 0.2 | 38.47 |
| 0.3 | 27.62 |
| 0.4 | 21.02 |
| 0.5 | 16.7 |
| 0.6 | 13.74 |
| 0.7 | 11.67 |
| 0.8 | 10.21 |
| 0.9 | 9.28 |

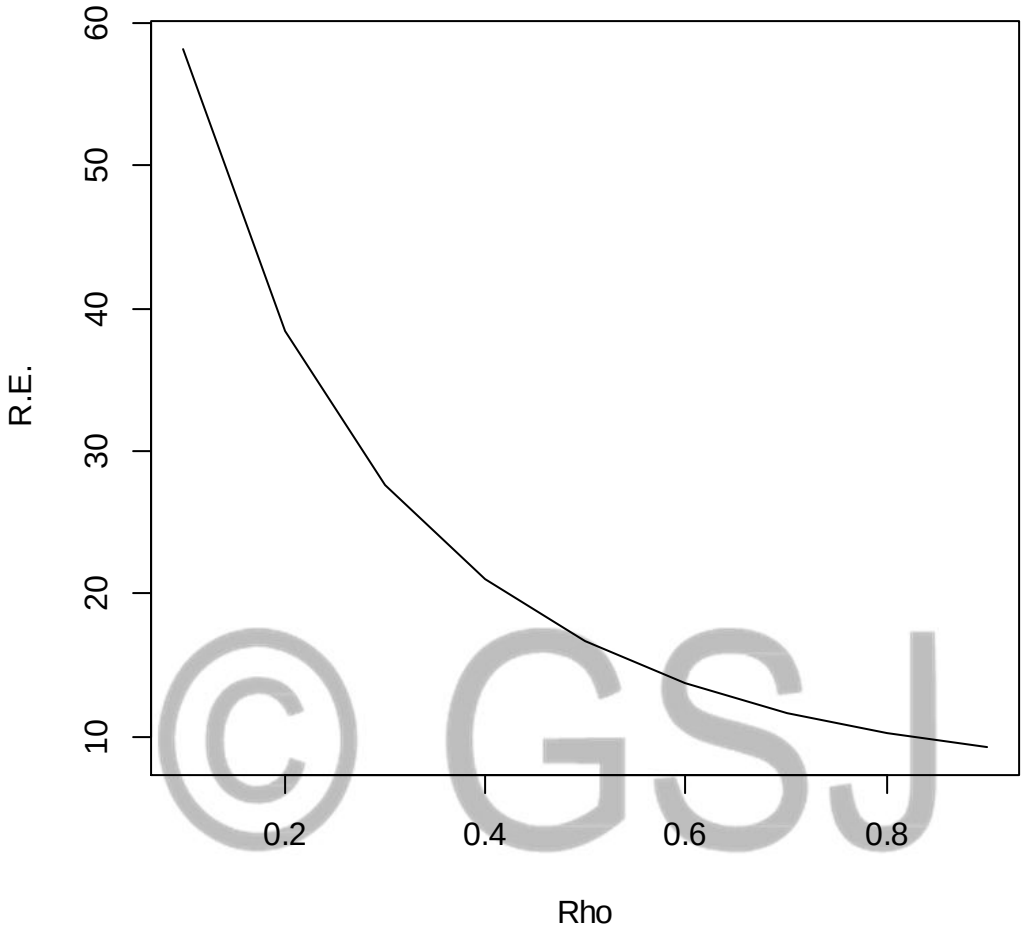


Figure 1: Graph of $\rho$ against R. E.

## 4. Concluding Remarks

In this research, the estimate of the mean for the first sub-sample is obtained with the value of 167.405, the variance estimate of simple random sampling was estimated to be 135.498 with the optimum variance value of 119.151 and the variance for double sample was estimated to be 18.06636. It was shown that the optimum variance of double sampling for regression estimator and its relative efficiency established more precision compare to the simple random sampling. Hence, the higher the ρ, the lesser the Relative Efficiency which is in inverse mode and established the reliability and efficiency of double sampling for regression estimation. It is now recommended that adequate record keeping and appropriate monitoring of enrollment figure should be the state concern among the contemporary states within.

## REFERENCES

Amahia, G. N., (2014). Unpublished lecture notes on design and analysis of sample survey. *University of Ibadan, Oyo state Nigeria.*

Cochran, W. G.,( 1977): Sampling Techniques, 3rd edition;New York:John Wiley.

Naqvi, H., Muhammad H. and Najeeb, H. (2013). A Regression Type Estimator with Two Auxiliary variables for two-phase Sampling. *Open journal of statistics, 3, pp. 74-78.*

Neyman, (1938): A Regression Type Estimator with Two Samples.

Rao, (1973). Double Sampling using Stratification Approach.

Senapati, S. C. and Sahoo, L. N. (2006): An alternative class of estimators in double sampling. Bulletin of the Malaysia mathematical society (2)29(1) (2006)" 23, 301-316.

Sodipo and Udomboso (2013). Unpublished lecture notes on Sample Survey. University of Ibadan, Oyo State, Nigeria.

Sukhatme, B. V. (1962).Some ratio-type estimators in two phase sampling. *Journal of American statistical association volume 57 pp 628-632.*

Watson, (1973). Concept of Double Sampling Regression Estimators.