A MODIFIED POST-STRATIFIED REGRESSION ESTIMATOR FOR TWO OCCASIONS AND USING TWO VARIABLES

*Olayiwola, Olaniyi M., Ajayi, Abayomi O. & Okunlade, Abiola A.*
Department of Statistics, College of Physical Sciences, Federal University of Agriculture, Abeokuta, Ogun State, Nigeria
*Corresponding Author’s E-mail: laniyimathew@yahoo.com*

ABSTRACT
Post stratification is grouping of selected samples into sub-homogeneous strata and it is appropriate when there is lack of prior knowledge of stratifying variable. Many authors have worked on post stratified estimators by using one variable but very often this leads to high Mean Square Error (MSE) and low response rate. This work proposed a modified post stratified regression estimator by using two variables which are highly correlated for the purpose of minimizing the mean square error and increasing the response rate. In this study we proposed two post stratified estimators, a separate estimators and a combined regression estimators, which were then extended to estimate the un-matched part of the second occasion in two occasions sampling. The variances for both estimators were derived. The optimum variance for the second occasion in two occasions sampling were also obtained. These estimators were compared with existing estimators. The variances for the separate and combined estimators were 186,962,721 and 122,368,755 respectively, while for the existing were 336,189,459. The optimum variance for the second occasion of the two occasion sampling for the separate and combined were 186,205,360 and 116,814,981 respectively while for the existing were 278,884,253. For the two data sets, it can be seen that the combine post stratified estimator were found to be more efficient (ratio 3:2) than separate post stratified regression estimators. It can be seen that the variances for the proposed estimators were lesser, hence it is more efficient.

INTRODUCTION
Post stratification is considered desirable in sample surveys for two reasons - it reduces the mean squared error when averaged over all possible samples, and it reduces the conditional bias when conditioned on stratum sample sizes. In finite population sampling designed based inference, it is well established that, when an auxiliary variable is available, proper use of stratified random sampling (STRS) reduces the variance of the estimate of the population mean compared to the variance of the estimate obtained from simple random sampling (SRS). When stratified random sampling is not used, but stratum information is available, a post stratified estimator may be used. In one aspect post stratification is potentially more efficient than stratification before selection, since after sampling the stratification factors can be chosen different ways for different sets of variables in order to maximize the gains in precision. It might be predicted that such a simple practical scheme would feature prominently in most texts on sampling, but this is not the case. The use of known auxiliary information at the estimation stage as well as at the selection stage leads to improved estimation strategies in survey sampling. When such information is not completely known or lacking and it is relatively cheaper to obtain information on the auxiliary variable(s), one can consider taking a large preliminary sample for estimating population mean(s) of the auxiliary variable(s) to be used at the estimation or selection stage of the ultimate estimation strategies. Many Authors have worked on post stratified regression estimator by considering only one variable (variable of Interest). It was observed that there is always a high mean square error and low response rate in the process of collecting data on the variable of interest in the existing estimator because of some sensitive questions. Based on this there is the need to develop a modified Post Stratified regression estimator which will minimize the mean square error and improve response rate for the purpose of having high precision by introducing another auxiliary character which is highly correlated with the variable of interest.

LITERATURE REVIEW
Post-stratification is widely recognized as an effective method for obtaining more accurate estimates of population quantities in the context of survey sampling. Although post-stratification is not always used in academic studies, it is a commonplace tool in commercial public
opinion polls. One of the greatest practical limitations to the use of post-stratification is the need to know the proportion of the population in each stratum. Smith (1991), Holt and Smith (1979), Jagers et al., (1985) & Jagers (1986) have done a lot of inspired work in the area of post-stratification in sample survey. Further the theory of Post-stratification in sample survey was extended by Cochran (1977). Data regarding changing properties of the populations of cities or counties, such as unemployment statistics, are collected regularly on a sample basis, to estimate the changes from one occasion to the next or to estimate the average over a certain period. An important aspect of continuous surveys is the structure of the sample on each occasion. To meet these requirements, successive sampling provides a strong tool for generating the reliable estimates at different occasions. Manish & Shukla (2008) looked into the efficient tool for generating the reliable estimates at different requirements, successive sampling provides a strong tool for generating the reliable estimates at different occasions. They aimed at providing the optimum estimates of population variance in successive sampling on two occasions and analyzed its properties.

Housila et al., (2007) looked into the problem of estimating a finite population quantile in successive sampling on two occasions. They aimed at providing the optimum estimates by combining.

(i) Three double sampling estimator viz. ratio-type, product-type and regression-type estimator, from the matched portion of the sample and.

(ii) A sample quantile based on a random sample from the unmatched portion of the sample on the second occasion.

A simulation study was carried out in order to compare the three estimators and it was found that the performance of the regression-type estimator was the best among all the estimators discussed.

**METHODOLOGY**

Notations

Y- Variable of interest

X, Z- Auxiliary Variables

\( \bar{y}_{ps} , \bar{x}_{ps}, \bar{z}_{ps} \) - Post stratified samples mean

\( \bar{y}_{2m} \) - The population mean on the matched sample on the second occasion.

\( \bar{y}_{2u} \) is the mean based on the post stratification of the second occasion

\( \frac{N_h}{n} \) - The proportion \( \frac{n_{1h}}{n_i} \).

\( S^2 = \) The population mean square error for the \( i\)-th occasion, \( i=1, 2, \ldots, n \).

Consider a finite population consisting of \( N \) units which can be partitioned into \( L \) strata of size \( N, N_1, \ldots, N_L \) respectively such that the strata \( W_h = \frac{N_h}{N} \) weights are known from previous investigation but the identity of the unit in various strata is not known. The problem considered is the expectation of the population mean of the characteristics of interest \( Y \).

A SRS of fixed size \( n \) is drawn without a replacement from the population and the sample unit identified from different strata on the basis of information collected on the stratification variable \( X \). The number of sample units \( n \), falling into stratum \( h \) is a random variable which in extreme cases can take zero value therefore we assumed that \( n \) is large enough so that \( Pr(n=0)=0 \)

**PROPOSED ESTIMATOR**

\[
\bar{y}_{pstl} = \frac{1}{N} \sum_{h=1}^{L} W_h \bar{y}_{lh} \tag{1}
\]

Equation 1 is the mean based on the post stratified which is an adaptation of Mostafa (1990) separate post stratified ratio estimator. This method is adapted to regression estimator to have a “separate post stratified regression estimator”. Thus we have,

\[
\bar{y}_{lh} = \bar{y}_h + b_{h\bar{x}} (\bar{z}_h - \bar{z}_h) \tag{2}
\]

and

\[
\bar{y}_{pstc} = \bar{y}_{ps} + b (\bar{z} - \bar{z}_{ps}) \tag{3}
\]

Equation 3 is a combined post stratified regression estimator, with

\[
\bar{y}_{ps} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{4}
\]

\[
\bar{z}_{ps} = \sum_{h=1}^{L} W_h \bar{z}_h \tag{5}
\]

\[
b = \sum_{h=1}^{L} W_h^2 \left( \frac{1 - f_h}{n_h} \right) S_{h\bar{x}} \sqrt{\sum_{h=1}^{L} W_h^2 \left( \frac{1 - f_h}{n_h} \right) S_{h\bar{x}}^2} \]

\[
b_{h\bar{x}} = S_{h\bar{x}} \sqrt{S_{h\bar{x}}^2}
\]

\[
W_h = \frac{N_h}{N}
\]
The unconditional variance of the proposed estimator was derived to be

\[ V(\bar{Y}_{\text{psl}}) = V\left(\frac{1}{N} \sum_{h=1}^{l} W_h \bar{Y}_{bh}\right) = \]

\[ \left(\frac{1-f}{n}\right) \sum_{h=1}^{l} W_h \left( S_{yh}^2 - \frac{S_{zh}^2}{S_{zh}} \right) + \frac{1}{n^2} \sum_{h=1}^{l} (1-W_h) \left( S_{yh}^2 - \frac{S_{zh}^2}{S_{zh}} \right) \]  

The unconditional variance for \( \bar{Y}_{\text{psl}} \) is

\[ V(\bar{Y}_{\text{psl}}) = V\left(\bar{Y}_{ps} + b(\bar{Z} - \bar{Z}_{ps})\right) = \]

\[ \left(\frac{1-f}{n}\right) \sum_{h=1}^{l} W_h \left( S_{yh}^2 + b^2 S_{zh}^2 - 2b S_{yzh} \right) + \frac{1}{n^2} \sum_{h=1}^{l} (1-W_h) \left( S_{yh}^2 + b^2 S_{zh}^2 - 2b S_{yzh} \right) \]  

**TWO-OCCASIONS SAMPLING**

We try to use these proposed estimators to estimate the mean of the un-matched part of the second occasion in the presence of an auxiliary variable. Consider a population of size \( N \) that is sampled over two occasions. On the first occasion a sample of \( n \) units is selected from \( N \) units in the population by SRSWOR. On the second occasion a sub sample of \( u \) units, \( (0<\lambda<1) \), is selected from the first occasion sample by SRSWOR. This is supplemented by a fresh sample of \( \lambda n \) units from \( N \) or from a sample of \( n \) units from \( N \) or from \( (N-n) \) units. Let \( X \) and \( Y \) represent the variables measured in the first and second occasions respectively.

Manish & Shukla (2008) post-stratify the unmatched part of the second occasion and proposed an unbiased estimator for the mean of the second occasion. This work post-stratified the un-matched part but in the presence of an auxiliary variable.

**Theorem 1:** The variance of the estimator for the two-occasions sampling estimators is given as

\[ V(\bar{Y}_{bm}) = \phi^2 V(\bar{Y}_{bm}) + (1-\phi)^2 V(\bar{Y}_{2m}) + 2\phi(1-\phi) \text{COV}(\bar{Y}_{2m}, \bar{Y}_{2u}) \]

The successive sampling estimator is defined to be

\[ \bar{Y}_{bm} = \phi \bar{Y}_{2m} + (1-\phi) \bar{Y}_{2u} \]

then

\[ V(\bar{Y}_{bm}) = \phi^2 V(\bar{Y}_{bm}) + (1-\phi)^2 V(\bar{Y}_{2m}) \]

\[ 2\phi(1-\phi) \text{COV}(\bar{Y}_{2m}, \bar{Y}_{2u}) \]

Let estimator \( \bar{Y}_{2m} \) be a chain type ratio-regression type estimator as proposed by Singh and Karna (2009).

\[ \bar{Y}_{2m} = \frac{1}{N} \sum_{i=1}^{k} W_i \bar{Y}_{bh} \]

(10)

(11)

(12)

(13)

When \( \bar{Y}_{2m} = \bar{Y}_{psl} \) be the proposed post stratified mean, Then

\[ V\left(\bar{Y}_{2m}\right) = \left(\frac{1-f}{u}\right) \sum_{h=1}^{l} W_h \left( S_{yh}^2 - \frac{S_{zh}^2}{S_{zh}} \right) + \]

\[ \frac{1}{u^2} \sum_{h=1}^{l} (1-W_h) \left( S_{yh}^2 - \frac{S_{zh}^2}{S_{zh}} \right) \]

(14)

When \( \bar{Y}_{2u} = \bar{Y}_{psl} \) Then

\[ V\left(\bar{Y}_{2u}\right) = \left(\frac{1-f}{u}\right) \sum_{h=1}^{l} W_h \left( S_{yh}^2 + b^2 S_{zh}^2 - 2b S_{yzh} \right) + \]

\[ \frac{1}{u} \sum_{h=1}^{l} (1-W_h) \left( S_{yh}^2 + b^2 S_{zh}^2 - 2b S_{yzh} \right) \]

(15)

Now putting the equations 14, 15, 16 together in equation 9 we have

\[ V(\bar{Y}_{2m}) = \phi^2 V(\bar{Y}_{2m}) + (1-\phi)^2 V(\bar{Y}_{2m}) + 2\phi(1-\phi) \text{COV}(\bar{Y}_{2m}, \bar{Y}_{2u}) \]
OPTIMUM CHOICE
Getting optimum value of $\phi$ we have to differentiate 9 with respect to $\phi$ and equating to zero, then we have
\[
\phi_{opt} = \frac{\text{Var}(\bar{y}_w) - \text{Cov}(\bar{y}_{2m}, \bar{y}_m)}{\text{Var}(\bar{y}_{2m}) + \text{Var}(\bar{y}_m) - 2\text{Cov}(\bar{y}_{2m}, \bar{y}_m)}
\]  
(17)

Putting this value in the variance expression we have
\[
V(\bar{y}_{xx}) = \frac{\text{Var}(\bar{y}_w)\text{Var}(\bar{y}_{2m}) - \left[\text{Cov}(\bar{y}_{2m}, \bar{y}_w)\right]^2}{\text{Var}(\bar{y}_{2m}) + \text{Var}(\bar{y}_m) - 2\text{Cov}(\bar{y}_{2m}, \bar{y}_m)}
\]  
(18)

DATA PRESENTATION AND ANALYSIS
The data used in this study were the number of enrolment of primary school pupils and the number of available teachers in each state of Nigeria. The data was collected from publications of National bureau of Statistics in Nigeria.

DESCRIPTIVE ANALYSIS
From table 1, for the year 2007 it was gotten that the ratio of male pupils to female pupils is 1.067349 and male to female teachers is 1.0513. The ratio of male pupils to male teachers is 37.08924 which indicate that for every male teacher we have approximately 37 pupils, for the females the ratio of female pupils to female teachers is 35.04049 which indicate for every female pupil we have approximately 35 students. Then generally the ratio of total pupils to total teachers is 43.62757, so approximately for each teacher we have approximately 44 students. For the year 2008, it was gotten that the ratio of male pupils to female pupils is 1.133897 and male to female teachers is 1.071262. The ratio of male pupils to male teachers is 37.08924 which indicate that for every male teacher we have approximately 37 pupils, for the females the ratio of female pupils to female teachers is 35.04049 which indicate for every female pupil we have approximately 35 students. Then generally, the ratio of total pupils to total teachers is 43.62757, so for each teacher we have approximately 36 students.

Table 1 (Mean per sex for 2007)

<table>
<thead>
<tr>
<th>Sex</th>
<th>Enrolment of primary school pupils</th>
<th>Number of available teachers</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>288435</td>
<td>6563</td>
<td>43.94973</td>
</tr>
<tr>
<td>Female</td>
<td>270235</td>
<td>6243</td>
<td>43.28887</td>
</tr>
<tr>
<td>Total</td>
<td>558670</td>
<td>12805</td>
<td>43.62757</td>
</tr>
</tbody>
</table>

Table 2 (Mean per sex for 2008)

<table>
<thead>
<tr>
<th>Sex</th>
<th>Enrolment of primary school pupils</th>
<th>Number of available teachers</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>300654</td>
<td>8106</td>
<td>37.08924</td>
</tr>
<tr>
<td>Female</td>
<td>265151</td>
<td>7567</td>
<td>35.04049</td>
</tr>
<tr>
<td>Total</td>
<td>565806</td>
<td>15673</td>
<td>36.10012</td>
</tr>
</tbody>
</table>

Two Occasion Sampling Analysis
\[N=74, \ n=40, \ \mu=1.5, \ m=n\mu=20, \ u=n\mu=20, \ W_1=0.5, \ W_2=0.5, \ N_1=37, \ N_2=37, \]
Table 3 shows the parameters of the proposed estimators and Manish & Shukla (2008), and it can be seen that the proposed $\nu(\mu_1)$ is lesser (67% for separate, 42% for combined) than the estimator proposed by Manish & Shukla (2008). It can be said that the proposed estimator is more efficient than Manish & Shukla (2008). It should be noted $\bar{V}_2 = \bar{y}_2 = \mu_2$ which is the mean of the matched portion of the second occasion and $\nu(\mu_2)$ which is also the variance of the matched portion of the second occasion, also $\bar{V}_u = \bar{V}_{un} = \bar{V}_{dun}$ is the mean of the unmatched portion of the second occasion. Lastly $\nu(\mu_2)$ is the mean of the second occasion $\nu(\mu_2) = \nu(\bar{y}_{un}) = \nu(\bar{y}_{dun})$ is the variance of the second occasion obtained after inserting $\phi_{opt}$ into the variance of the second occasion.

**Conclusion**

From the analysis, the proposed variance of the un-matched part $\nu(\mu_2)$ (separate and combined post stratified regression estimator) was lesser than the one proposed by Travedi & Shukla (2008), it can be concluded that post stratification with auxiliary variable of the unmatched part of the second occasion gives a smaller variance. Hence it is more efficient. Since the variance of the unmatched part is lesser, the overall variance of the second occasion were found to be smaller.

**REFERENCES**


Accepted 7 July, 2016