EFFICIENCY OF DATA TRANSFORMATION AND CORRECTION FACTOR METHODS ON THE CORRECTION OF EXTREME VALUE EFFECT IN SAMPLE SURVEY THEORY

Peter I. Ogunyinka [1], Emmanuel F. Ologunleko [2], B. T. Efuwape [1], O. M. Olayiwola [3], Dawud A. Agunbiade [1]

[1] **Department of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye, Nigeria**
[2] **Department of Statistics, University of Ibadan, Ibadan, Nigeria**
[3] **Department of Statistics, Federal University of Agriculture, Abeokuta, Nigeria**

Corresponding Author: Peter I. Ogunyinka, ogunyinka.peter@gmail.com

ABSTRACT: Extreme value, in Sample Survey Theory, is termed outlier in General Statistical Theory. Extreme value data analysis would yield over-estimation or under-estimation in statistical estimation process. Non-linear data transformation method and Sarndal's correction factor method in Sample Survey Theory have been confirmed to correct extreme value effect in an estimate. However, since the two methods work towards the same objective, there is need to ascertain the efficient estimate between the two methods. This study has empirically compared the two extreme value correction methods. Results revealed that non-linear data transformation method, though associated with back transformation challenge, had lower Percentage Coefficient of Variation (PCV) over correction factor method. Hence, non-linear data transformation method proved efficient over correction factor method. It was recommended that Survey Statisticians should improve on the Sarndal's developed correction factor method and/or developed new improved estimators for the correction of extreme value in Sample Survey Theory.

KEYWORDS: Extreme value, non-linear data transformation, correction factor, regression estimator, percentage coefficient of variation.

1. INTRODUCTION

The Survey Statisticians leverage on the presence of auxiliary information in any data set to improve on any concerned estimators. Single-phase sampling estimator like ratio ([Coc40)], regression ([Coc42]), product ([HHM53]) and difference ([Rob57]) estimators and two-phase sampling estimators ([Ney38] and [Kee05]) have used auxiliary information (either auxiliary variable or attribute) to produce different Uniformly Minimum Variance (UMV) estimators in Sample Survey Theory. The use of single-phase sampling requires positive and high correlation coefficient between the study variable (y) and the auxiliary variable (x) ([AO13]). Among the statistical assumptions required for the use of single-phase sampling estimator are linearity assumption between the study and the auxiliary variable and the normality assumption. However, the violation of these assumptions will lead to over-estimation or under-estimation of the concerned parameter ([Sar72]), therefore, increasing the Mean Square Error (MSE) of the estimate. The presence of extreme value in a data set has been confirmed as the major source of the violation of statistical assumption ([A+18]).

Extreme value in Sample Survey Theory is also called outlier in General Statistics Theory. Extreme value could be identified in the data set as outrageously high observation(s) or outrageously low observation(s) in the data set. The prior detection of the presence of extreme value in the data set, to the application of estimators, is very germane to obtaining efficient estimate. Ratio, regression, difference and product estimators are also affected by the presence of extreme value in the data set which produce less efficient estimate. This study has identified two solutions for the correction of extreme value(s) in the data set. The use of non-linear data transformation on the data set prior to the estimation of the parameter and the use of correction factor in the estimator ([Sar72]) are the solutions considered in this study.

2. METHODOLOGY

2.1 The theory of data transformation and Sarndal's correction factor

Data transformation is the use of mathematical tool to change the variable scale or unit of measurement of any data set. [OB14] had reported that both linear and non-linear data transformation methods are considered to be legitimate. Non-linear data transformation will be necessary to decrease or increase (change) the integrity of variable relationships. While the validity and importance of data transformation (either linear or non-linear) has

been ascertained by [TF07], some authors have reported the set-back or challenge of *back transformation* to be associated with non-linear data transformation. [OB14] has applied non-linear data transformation to the estimation of online software repository variables. Table 1 shows the varieties of non-linear transformation tools used by [OB14]. Those tools include exponential, quadratic, reciprocal, logarithm, power and square transformations. Back transformation is a major challenge associated with non-linear data transformation. Fortunately, while it cannot be avoided, it can be minimized. [Mil84] and [JR08] are among the major theoretical back transformation solutions that are recognized by authors to have reduced the bias introduced by back transformation.

This study has classified data transformation as a *pre-estimation extreme value control method* since it is applied prior to the estimation of the parameter. [Sar72] had developed an improved estimator using optimum correction factor (C_{opt}) to minimize the effect of extreme value within the data set. Sarndal assumed that this correction factor (C_{opt}) should be added to the estimator if the extreme value causes under-estimation but subtracted if the extreme value causes over-estimation. This study has classified Sarndal's correction factor method as *estimation extreme value correction method* since the correction factor is embedded in the estimator. Figure 2 shows the flowchart on the procedural application of data transformation and control factor methods. It also explains the pre-estimation and estimation extreme value correction methods. The flowchart illustration on the use of pre-estimation (non-linear data transformation) is also applicable to General Statistical Theory.

In a data set with $\overline{y}$ as the unbiased estimator of the population mean, y_{min} as the minimum extreme value and y_{max} as the maximum extreme value, [Sar72] has suggested an efficient sample mean with the correction factor (C_{opt}) as

$$\overline{y}_s = \begin{cases} \overline{y} + c & if\ sample\ contains\ y_{min}\ only \\ \overline{y} - c & if\ sample\ contains\ y_{max}\ only \\ \overline{y} & for\ all\ other\ samples, \end{cases} \quad (1)$$

where $0 < c < (y_{max} - y_{min})/n$ and n is the sample size.

The optimum value of c is presented as $c_{opt} = \frac{(y_{max} - y_{min})}{2n}$ and the optimum variance of $\overline{y}_s$ is presented as

$$V(\overline{y}_s)_{min} = V(\overline{y}) - \frac{\theta \Delta_y^2}{2(N-1)}. \quad (2)$$

where $V(\overline{y}) = \theta S_y^2$, $\Delta_y = (y_{max} - y_{min})$, $\theta = \left(\frac{1}{n} - \frac{1}{N}\right)$, N = population size and n = Sample size.

[Coc42] developed regression estimator of the study variable (y) in relationship with the auxiliary variable (x) as

$$\overline{y}_{lr} = \overline{y} + b(\overline{X} - \overline{x}). \quad (3)$$

The corresponding minimized variance of $\overline{y}_{lr}$ is presented as:

$$V(\overline{y}_{lr})_{min} = \theta S_y^2 (1 - \rho_{yx}^2), \quad (4)$$

where ρ_{yx} = Correlation coefficient.

[KS13] seems to be the first to improve regression estimators for estimating the population mean in the presence of extreme value in the data set using the correction factor of [Sar72]. The improved estimators is presented as

$$\overline{y}_{lrc} = \begin{cases} (\overline{y} + c_1) + b\left(\overline{X} - (\overline{x} + c_2)\right), & if\ sample \\ & contains\ y_{min}\ only; \\ (\overline{y} - c_1) + b\left(\overline{X} - (\overline{x} - c_2)\right), & if\ sample \\ & contains\ y_{max}\ only \\ \overline{y} + b(\overline{X} - \overline{x}), for\ all & \\ other\ samples. & \end{cases} \quad (5)$$

The corresponding variance is presented as

$$V(\overline{y}_{lrc}) = V(\overline{y}_{lr}) - \frac{2\theta n}{(N-1)}\left[(c_1 - \beta c_2)\{\Delta_y - \beta \Delta_x - 2n(c_1 - \beta c_2)\}\right]. \quad (6)$$

The optimum values of c_1 and c_2 are presented as $c_{1opt} = \frac{(y_{max} - y_{min})}{2n}$ and $c_{2opt} = \frac{(x_{max} - x_{min})}{2n}$.

The minimized variance is presented as

$$V(\overline{y}_{lrc})_{min} = V(\overline{y}_{lr}) - \frac{\theta}{2(N-1)}\left[\Delta_y - \beta \Delta_x\right]^2, \quad (7)$$

where $\Delta_y = (y_{max} - y_{min})$, $\Delta_x = (x_{max} - x_{min})$ and β is the regression coefficient of y on x.

Data transformation and correction factor have proved to be efficient in the correction of extreme values in Sample Survey Theory. However, the former is pre-estimation extreme value correction method while the latter is an estimation extreme value correction method. This classification is done based on the stage of application of the extreme value correction method in the estimation process. Since these two correction methods serve the same objective but at different application stages, then it will be of interest to know which method produces efficient estimate. It is observed that efficiency of the data transformation method cannot be examined

theoretically unlike Sarndal's correction method. Therefore, this study shall examine the efficiency of these two correction methods empirically.
Table 1 shows different transformation tools that will lead to different units of measurement of the study variable and the variance. Consequently, all the estimates and variances that will be compared will be in different units of measurement. Hence, there is need for a statistical measure that can be used for comparison in this case. The Coefficient of Variation (CV) will be a better choice.

Table 1: Some of the mathematical tools used in non-linear data transformation (extracted from [OB14])

SN	Dependent Variable (T)	Independent Variable (D)	Linear Regression Model	Back Transformation
1	T	$log_{10}D = D^*$	$T = \alpha + \beta D^*$	$\hat{T} = \alpha + \beta D^*$
2	$log_{10}D = D^*$	$\sqrt[3]{T} = T^*$	$D^* = \alpha + \beta T^*$	$\hat{D} = 10^{(\alpha+\beta T^*)}$
3	$log_{10}T = T^*$	$log_{10}D = D^*$	$T^* = \alpha + \beta D^*$	$\hat{T} = 10^{(\alpha+\beta D^*)}$
4	$\sqrt{D} = D^*$	$\sqrt[3]{T} = T^*$	$D^* = \alpha + \beta T^*$	$\hat{D} = (\alpha + \beta T^*)^2$
5	$log_2 T = T^*$	$log_{10}D = D^*$	$T^* = \alpha + \beta D^*$	$\hat{T} = 2^{(\alpha+\beta D^*)}$
6	$\sqrt{D} = D^*$	$T^2 = T^*$	$D^* = \alpha + \beta T^*$	$\hat{D} = (\alpha + \beta T^*)^2$
7	T	$\sqrt{D} = D^*$	$T = \alpha + \beta D^*$	$\hat{T} = \alpha + \beta D^*$
8	$\sqrt{T} = T^*$	$\sqrt{D} = D^*$	$T^* = \alpha + \beta D^*$	$\hat{T} = (\alpha + \beta D^*)^2$
9	$log_{10}T = T^*$	$D^{-1} = D^*$	$T^* = \alpha + \beta D^*$	$\hat{T} = 10^{(\alpha+\beta D^*)}$
10	$T^3 = T^*$	D	$T^* = \alpha + \beta D$	$\hat{T} = \sqrt[3]{(\alpha + \beta D)}$
11	D	$T^2 = T^*$	$D = \alpha + \beta T^*$	$\hat{D} = \alpha + \beta T^*$
12	$T^3 = T^*$	$\sqrt[3]{D} = D^*$	$T^* = \alpha + \beta D^*$	$\hat{T} = \sqrt[3]{(\alpha + \beta D^*)}$
13	$log_{10}D = D^*$	$T^{-1} = T^*$	$D^* = \alpha + \beta T^*$	$\hat{D} = 10^{(\alpha+\beta T^*)}$
14	$T^{-1} = T^*$	$D^{-1} = D^*$	$T^* = \alpha + \beta D^*$	$\hat{T} = (\alpha + \beta D^*)^{-1}$

Coefficient of Variation (CV) is a statistical measure of variability for any experiment with different units of measurement. The CV has the major advantage because it converts experiment to a dimensionless output. Hence, it makes the comparison of experiment with different measurement of unit possible. [Man13] reported that [CCK67] developed coefficient of variation while [M+017] reported on the classification of CV. However, this study will not be concerned with the classification of CV. The lower the CV, the more efficient is the concerned estimate or estimator. The higher the CV, the least efficient is the concerned estimate or estimator. The Percentage Coefficient of Variation (PCV) is presented as

$$PCV = \frac{SE}{Mean} * 100\%, \quad (8)$$

where SE= Standard error.

3. RESULTS AND DISCUSSION

3.1 Results

Table 2 explains the classification of correlation coefficient based on the recommendations of [AO13]. M represents Moderate correlation coefficient and H represents High correlation coefficient. Tables 3 through 7 show analysis results for different population sizes (N) and sample sizes (n). Table 3 has $N = 150$ and $n = 30$, table 4 has $N = 200$ and $n = 30$, table 5 has $N = 200$ and $n = 50$, table 6 has $N = 250$ and $n = 70$ and table 7 has $N = 250$ and $n = 90$. Analysis in table 3 through 7 shows the results of the correlation coefficient, the classification of the correlation coefficient, the confirmation of the outlier in the data set, confirmation of the violation of linearity assumption (represented in A1), independence assumption (represented in A2), normality assumption (represented in A3) and homoscedasticity assumption (represented in A4).

Table 2: Correlation Coefficient interpretation extracted from [AO13]

Size range of Correlation	Interpretation	Notation
0.90 to 1.0	Very high Positive (Negative) Correlation	V
0.70 to <0.90	High Positive (Negative) Correlation	H
0.50 to <0.70	Moderate Positive (Negative) Correlation	M
0.30 to <0.50	Low Positive (Negative) Correlation	L
0.00 to <0.30	Negligible Correlation	N

They also show the population mean estimates obtained after applying non-linear data transformation and using correction factor in [KS13] regression estimator. Finally, the corresponding variance, Percentage Coefficient of Variation (PCV) and the rank of the PCV were presented in the five tables. The analyses measure the rank of correlation coefficient, confirmed the presence of outlier and the violation of statistical assumptions. The ranking results (in the last column of tables 3 through 7) ranked the correction method and estimators based on results of the PCV in ascending order since lower PCV signifies higher efficiency of the estimates or estimator and vice versa.

3.2 Discussion

In table 3, the outlier column confirms that the original data contains significant outlier, violated linearity, normality and homoscedasticity assumptions. However, it was, further, revealed that non-linear data transformation on both the study and auxiliary variables corrected the effects of outliers and the violation of statistical assumptions. This justifies the importance of data transformation to the correction of violation of statistical assumptions. [KS13] regression estimator used the original data for analysis. The ranking result (last column in the table) rated the estimate obtained from the original data analysis to be the least (8th with PCV of 11.389%). It was confirmed that estimates obtained from non-linearly transformed data analyses proved efficient (with 1st to 6th ranking of the PCV) over the estimate from the original data. This ascertained the importance of data transformation method in the correction of extreme value in the data set. Estimate from Khan and Shabbir ([KS13]) regression estimator was ranked efficient (7th with PCV of 11.037%) over estimate from the original data analysis (8th with PCV of 11.389%). This shows that Sarndal's correction factor controls the extreme value effect. Contrarily, Khan and Shabbir ([KS13]) regression estimate proved least efficient (7th with PCV of 11.037%) over the estimates from the non-linear data transformation analysis (with 1st to 6th PCV ranking). This means that the extreme value effect (variability) control of [Sar72] correction method was not efficient compared to non-linear data transformation extreme value effect control method. Analysis results in tables 4 through 7 confirmed the above emphased results. Hence, it was confirmed that pre-estimation extreme value correction (non-linear data transformation) method proved efficient in correcting the violation of statistical assumptions in the original data. It was also confirmed that non-linear data transformation proved efficient over the use of correction factor in the [KS13] regression estimator. The summary of this discovery is presented in figure 1.

In figure 1, $\overline{Y}$ is the parameter to be estimated. The over-estimated and under-estimated estimators, $\overline{y}_{max}$ and $\overline{y}_{min}$, respectively, were obtained as a result of the presence of extreme value(s) in the data set. The adjusted maximum and minimum estimators, $\overline{y}_{1*}$ and $\overline{y}_{2*}$, respectively, were obtained based on the developed correction factor method by [Sar72]. Finally, $\overline{y}^{*}_{max}$ and $\overline{y}^{*}_{min}$ are the maximum and minimum estimates, respectively, obtained based on the non-linear data transformation technique. The aim is to obtain an estimate that will be very closed to the parameter, $\overline{Y}$. Figure 1 explains that estimates from non-linear data transformation have close proximity to the parameter $\overline{Y}$ than estimates from [Sar72] correction factor method but with associated back transformation challenge.

This study has discovered that efficiency of non-linear data transformation method over correction factor of [Sar72] using proposed regression estimator of [KS13]. Data transformation has the challenge of back transformation and it is classified as pre-estimation extreme value correction method. However, the use of correction factor of [Sar72] method is classified as estimation extreme value correction method. This result has identified the need for improvement on the Sarndal's correction factor method. Alternatively, Survey Statisticians may need to develop another improved estimation extreme value correction method(s) that will prove efficient over non-linear data transformation method in the correction of extreme value effect in Sample Survey Theory.

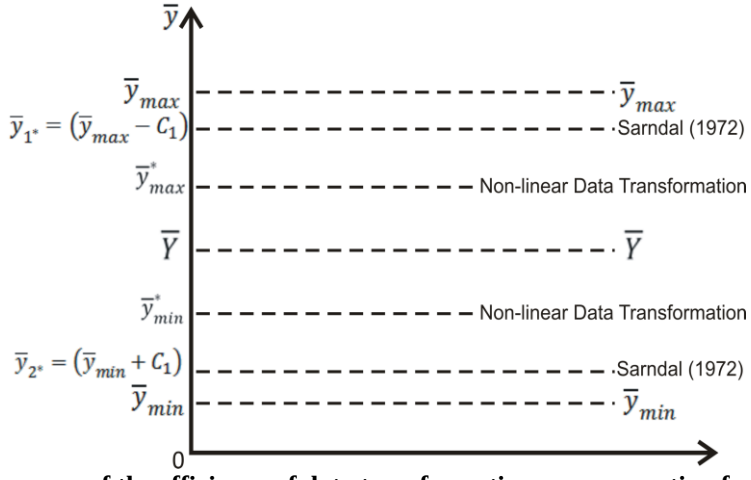

Figure 1: Summary of the efficiency of data transformation over correction factor methods

Table 3: Empirical analysis result when $N = 150$, $n = 30$, $C_{1opt} = 0.88$ and $C_{2opt} = 4867.00$

SN	Data Trans-formation	ρ_{yx}		Out-lier	A1	A2	A3	A4	$\bar{y}_{lr}$	$V(\bar{y}_{lr})$	$CV(\bar{y}_{lr})$	Rank
1	y/x	0.5625	:M	yes	D	O	D	D	16.1843	3.3973	11.389%	*8*
2	$\sqrt[3]{y}/\sqrt{x}$	0.5812	:M	no	O	O	O	O	2.3671	0.0082	3.818%	*3*
3	$\sqrt[3]{y}/\sqrt[3]{x}$	0.5828	:M	no	O	O	O	O	2.3774	0.0081	3.796%	*2*
4	$\sqrt[3]{y}/log_{10}x$	0.5732	:M	no	O	O	O	O	2.4002	0.0083	3.791%	*1*
5	$log_{10}y/\sqrt{x}$	0.5580	:M	no	O	O	O	O	2.5741	0.0116	4.186%	*6*
6	$log_{10}y/\sqrt[3]{x}$	0.5586	:M	no	O	O	O	O	2.5856	0.0116	4.165%	*5*
7	$log_{10}y/log_{10}x$	0.5468	:M	no	O	O	O	O	2.6110	0.0118	4.164%	*4*
8	**[KS13] Regression Estimator**								**16.5777**	**3.3475**	**11.037%**	*7*

Keys: A1 = Linearity assumption; A2 = Independence assumption; A3 = Normality assumption, $A4$ = Homoscedasticity assumption, O= Obey assumption, D= Disobey assumption, H= High Correlation Coefficient and M= Medium Correlation Coefficient

Table 4: Empirical analysis result when$N = 200$, $n = 30$, $C_{1opt} = 0.85$ and $C_{2opt} = 7136.6$

SN	Data Trans-formation	ρ_{yx}		Out-lier	A1	A2	A3	A4	$\bar{y}_{lr}$	$V(\bar{y}_{lr})$	$CV(\bar{y}_{lr})$	Rank
1	y/x	0.6895	:M	yes	D	O	D	D	18.0741	4.4840	11.716%	*14*
2	$y/\sqrt{x}$	0.7575	:H	yes	O	O	D	O	17.7489	3.6428	10.753%	*12*
3	$y/\sqrt[3]{x}$	0.7713	:H	yes	O	O	D	O	17.8234	3.4629	10.441%	*11*
4	$y/log_{10}x$	0.7742	:H	yes	O	O	D	O	18.2639	3.4239	10.131%	*10*
5	$\sqrt{y}/\sqrt{x}$	0.7277	:H	no	O	O	O	O	3.7904	0.0500	5.899%	*9*
6	$\sqrt{y}/\sqrt[3]{x}$	0.7447	:H	no	O	O	O	O	3.7976	0.0473	5.728%	*8*
7	$\sqrt{y}/log_{10}x$	0.7547	:H	no	O	O	O	O	3.8441	0.0457	5.563%	*7*
8	$\sqrt[3]{y}/\sqrt{x}$	0.7105	:H	no	O	O	O	O	2.3700	0.0091	4.035%	*3*
9	$\sqrt[3]{y}/\sqrt[3]{x}$	0.7281	:H	no	O	O	O	O	2.3729	0.0087	3.925%	*2*
10	$\sqrt[3]{y}/log_{10}x$	0.7397	:H	no	O	O	O	O	2.3917	0.0084	3.823%	*1*
11	$log_{10}y/\sqrt{x}$	0.6779	:M	no	O	O	O	O	2.5535	0.0127	4.410%	*6*
12	$log_{10}y/\sqrt[3]{x}$	0.6962	:M	no	O	O	O	O	2.5565	0.0121	4.302%	*5*
13	$log_{10}y/log_{10}x$	0.7100	:H	no	O	O	O	O	2.5766	0.0116	4.187%	*4*
14	**[KS13] Regression Estimator**								**18.0851**	**4.4218**	**11.627%**	*13*

Keys: A1 = Linearity assumption; A2 = Independence assumption; A3 = Normality assumption, $A4$ = Homoscedasticity assumption, O= Obey assumption, D= Disobey assumption, H= High Correlation Coefficient and M= Medium Correlation Coefficient.

Table 5: Empirical analysis result when $N = 200$, $n = 50$, $C_{1opt} = 0.56$ and $C_{2opt} = 7261.1$

SN	Data Trans-formation	ρ_{yx}		Out-lier	A1	A2	A3	A4	$\overline{y}_{lr}$	$V(\overline{y}_{lr})$	$CV(\overline{y}_{lr})$	Rank
1	y/x	0.6874	:M	yes	D	O	D	D	16.0146	2.1702	9.199%	***14***
2	$y/\sqrt{x}$	0.7629	:H	yes	O	O	D	D	16.1309	1.7197	8.130%	***12***
3	$y/\sqrt[3]{x}$	0.7686	:H	yes	O	O	D	D	16.3012	1.6837	7.960%	***10***
4	$y/log_{10}x$	0.7425	:H	yes	O	O	D	D	16.7778	1.8465	8.099%	***11***
5	$\sqrt{y}/\sqrt{x}$	0.7061	:H	no	O	O	D	O	3.5732	0.0270	4.595%	***9***
6	$\sqrt{y}/\sqrt[3]{x}$	0.7150	:H	no	O	O	D	O	3.5910	0.0263	4.514%	***7***
7	$\sqrt{y}/log_{10}x$	0.6965	:M	no	O	O	D	O	3.6417	0.0277	4.569%	***8***
8	$\sqrt[3]{y}/\sqrt{x}$	0.6767	:M	no	O	O	O	O	2.2721	0.0052	3.184%	***3***
9	$\sqrt[3]{y}/\sqrt[3]{x}$	0.6860	:M	no	O	O	O	O	2.2793	0.0051	3.136%	***1***
10	$\sqrt[3]{y}/log_{10}x$	0.6692	:M	no	O	O	O	O	2.2999	0.0053	3.174%	***2***
11	$log_{10}y/\sqrt{x}$	0.6272	:M	no	O	O	O	O	2.4318	0.0079	3.646%	***6***
12	$log_{10}y/\sqrt[3]{x}$	0.6370	:M	no	O	O	O	O	2.4395	0.0077	3.597%	***4***
13	$log_{10}y/log_{10}x$	0.6229	:M	no	O	O	O	O	2.4617	0.0079	3.617%	***5***
14	**[KS13] Regression Estimator**								**16.1039**	**2.1672**	**9.141%**	***13***

Keys: A1 = Linearity assumption; A2 = Independence assumption; A3 = Normality assumption, $A4$ = Homoscedasticity assumption, O= Obey assumption, D= Disobey assumption, H= High Correlation Coefficient and M= Medium Correlation Coefficient.

Table 6: Empirical analysis result when $N = 250$, $n = 70$, $C_{1opt} = 0.5929$ and $C_{2opt} = 7069.53$

SN	Data Trans-formation	ρ_{yx}		Out-lier	A1	A2	A3	A4	$\overline{y}_{lr}$	$V(\overline{y}_{lr})$	$CV(\overline{y}_{lr})$	Rank
1	y/x	0.6163	:M	yes	D	O	D	D	16.5025	2.6005	9.772%	***14***
2	$y/\sqrt{x}$	0.7612	:H	yes	O	O	D	D	17.4438	1.7638	7.613%	***12***
3	$y/\sqrt[3]{x}$	0.7872	:H	yes	O	O	D	D	17.2777	1.5950	7.310%	***11***
4	$y/log_{10}x$	0.7665	:H	yes	O	O	D	D	18.1679	1.7293	7.238%	***10***
5	$\sqrt{y}/\sqrt{x}$	0.7455	:H	no	O	O	D	O	3.6032	0.0193	3.853%	***9***
6	$\sqrt{y}/\sqrt[3]{x}$	0.7753	:H	no	O	O	D	O	3.6398	0.0173	3.614%	***8***
7	$\sqrt{y}/log_{10}x$	0.7689	:H	no	O	O	D	O	3.7298	0.0177	3.571%	***7***
8	$\sqrt[3]{y}/\sqrt{x}$	0.7234	:H	no	O	O	D	O	2.2787	0.0035	2.610%	***3***
9	$\sqrt[3]{y}/\sqrt[3]{x}$	0.7544	:H	no	O	O	D	O	2.2934	0.0032	2.465%	***2***
10	$\sqrt[3]{y}/log_{10}x$	0.7542	:H	no	O	O	D	O	2.3297	0.0032	2.428%	***1***
11	$log_{10}y/\sqrt{x}$	0.6763	:M	no	O	O	O	D	2.4319	0.0051	2.944%	***6***
12	$log_{10}y/\sqrt[3]{x}$	0.7087	:H	no	O	O	O	D	2.4473	0.0047	2.802%	***5***
13	$log_{10}y/log_{10}x$	0.7182	:H	no	O	O	O	O	2.4860	0.0046	2.721%	***4***
14	**[KS13] Regression Estimator**								**16.5960**	**2.5969**	**9.710%**	***13***

Keys: A1 = Linearity assumption; A2 = Independence assumption; A3 = Normality assumption, $A4$ = Homoscedasticity assumption, O= Obey assumption, D= Disobey assumption, H= High Correlation Coefficient and M= Medium Correlation Coefficient.

Table 7: Empirical analysis result when $N = 250$, $n = 90$, $C_{1opt} = 0.4611$ and $C_{2opt} = 5498.52$

SN	Data Trans-formation	ρ_{yx}		Out-lier	$A1$	$A2$	$A3$	$A4$	$\overline{y}_{lr}$	$V(\overline{y}_{lr})$	$CV(\overline{y}_{lr})$	Rank
1	y/x	0.6116	:M	yes	D	O	D	D	15.8365	1.5497	7.861%	***14***
2	$y/\sqrt{x}$	0.7519	:H	yes	O	O	D	D	16.3297	1.0760	6.352%	***12***
3	$y/\sqrt[3]{x}$	0.7730	:H	yes	O	O	D	D	16.6082	0.9965	6.011%	***10***
4	$y/log_{10}x$	0.7424	:H	yes	O	O	D	D	17.0972	1.1114	6.166%	***11***
5	$\sqrt{y}/\sqrt{x}$	0.7263	:H	no	O	O	D	D	3.5439	0.0129	3.202%	***9***
6	$\sqrt{y}/\sqrt[3]{x}$	0.7522	:H	no	O	O	D	D	3.5723	0.0118	3.045%	***7***
7	$\sqrt{y}/log_{10}x$	0.7365	:H	no	O	O	D	O	3.6236	0.0125	3.081%	***8***
8	$\sqrt[3]{y}/\sqrt{x}$	0.7002	:H	no	O	O	D	D	2.2532	0.0024	2.193%	***3***
9	$\sqrt[3]{y}/\sqrt[3]{x}$	0.7276	:H	no	O	O	D	D	2.2647	0.0023	2.097%	***1***
10	$\sqrt[3]{y}/log_{10}x$	0.7180	:H	no	O	O	D	O	2.2858	0.0023	2.108%	***2***
11	$log_{10}y/\sqrt{x}$	0.6491	:M	no	O	O	O	D	2.4018	0.0037	2.526%	***6***
12	$log_{10}y/\sqrt[3]{x}$	0.6781	:M	no	O	O	O	D	2.4142	0.0034	2.428%	***5***
13	$log_{10}y/log_{10}x$	0.6778	:M	no	O	O	O	O	2.4372	0.0034	2.406%	***4***
14	**[KS13] Regression Estimator**								**15.8981**	**1.5480**	**7.826%**	***13***

Keys: A1 = Linearity assumption; A2 = Independence assumption; A3 = Normality assumption, $A4$ = Homoscedasticity assumption, O= Obey assumption, D= Disobey assumption, H= High Correlation Coefficient and M= Medium Correlation Coefficient.

4. CONCLUSIONS

Attention had been called to the sequency of extreme value and outlier in both Sample Survey Theory and General Statistical Theory, respectively. The presence of extreme value in a data set would cause over-estimation or under-estimation which consequently increases the Mean Square Error of the estimate. This study has considered non-linear data transformation and [Sar72] correction factor for the reduction or removal of extreme value effect in Sample Survey Theory. Non-linear data transformation was classified as a *pre-estimation extreme value control method* and [Sar72] correction factor was defined as an *estimation extreme value control method* in Sample Survey Theory. This study has compared the efficiency of these two methods using in-depth empirical approach. It was discovered that while non-linear data transformation method is associated with back transformation challenge, estimates from non-linearly transformed data set would significantly reduce the effect of extreme value than estimators with [Sar72] correction factor method. It was empirically ascertained that non-linear data transformation method will produce efficient estimates over [Sar72] correction factor method but not without its associated back transformation bias challenge. Consequently, this study has recommended improvement on the [Sar72] extreme value effect correction method and/or the need to develop another improved estimation extreme value correction method in Sample Survey Theory.

REFERENCES

[AO13] **D. A. Agunbiade, P. I. Ogunyinka** – *Effect of correlation level on the use of Auxiliary variable in Double sampling for regression estimation*, Open Journal of Statistics, Doi: http//dx.doi.org/10.4236/ojs.2013.35037, 2013.

[A+18] **N. Abbas, M. Abid, Tahir, M., Abbas, Z. Hussain** – *Enhancing ratio estimators for estimating population mean using maximum value of auxiliary variable*, J. Natn. Sci. Foundation Sri Lanka, vol. 46(3): 453-463, 2018.

[Coc40] **W. G. Cochran** – *Sampling Technique*, 3rd Edition, Wiley Eastern Limited, India, 1940.

[Coc42] **W. G. Cochran** – *Sampling theory when the sampling units are of unequal sizes*, J.Amer. Statistics Association, Vol. 37: 199-212, 1942.

[CCK67] **F. E. Croxton, D. J. Crowden, S. Klein** – *Applied General Statistics*, 3rd Edition, Prentice-Hall, New-York, 754, 1967.

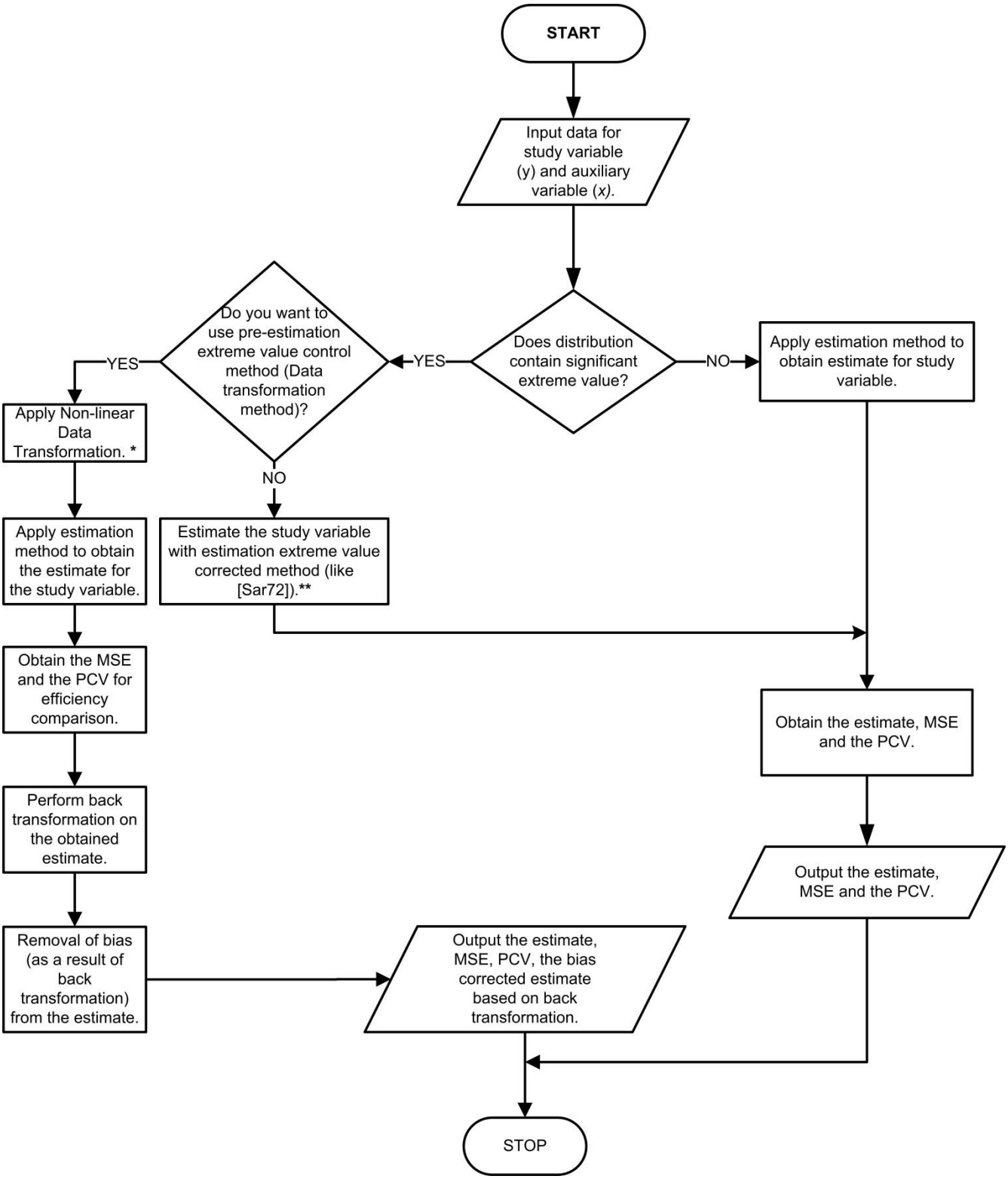

Figure 2: Flowchart for the use of pre-estimation and estimation extreme value correction methods in Sample Survey Theory

Notes to Figure 2.0

* Non-linear data transformation is applied prior to the estimation of the parameters. This is termed *pre-estimation extreme value correction method*.

** Extreme value correction is coming at the estimation stage. This is termed as *estimation extreme value correction method*.

PCV: Percentage Coefficient of Variation.

MSE: Mean Square Error

[HHM53] **M. H. Hansen, W. N. Hurwitz, W. G. Madow** – *Sample Survey Methods and Theory*, 2nd, Ed., New York: Wiley and Sons, 1953.

[JR08] **S. Jai, S. Rathi** – *On predicting log-transformed linear models with heteroscedaticity*, SAS Global Forum, Paper 370, 2008.

[Kee05] **K. J. Keen** – *Two Phase Sampling*. Encyclopedia of Bio-Statstics, 8, 1-4, 2005.

[KS13] **M. Khan, J. Shabbir** – *Some improved ratio, product and regression estimators of finite population mean when using minimum and maximum values*, The Scientific World Journal, Article ID: 431868, 1-7, 2013.

[Man13] **M. Manuel** – *A Coefficient of variability*, Journal of Mathematics and Statistics, vol. 9(11), 62-64, ISSN: 1549-3644, dou: 10.3844/jmssp.2013.62.64, 2013.

[Mil84] **D. Miller** – *Reducing transformation bias in Curve fitting*, The American Statistician, 30(2), 124-126, 1984.

[M+17] **A. B. V. Marcos, P. S. Pacheco, E. J. Seidel, P. A. Angela** – *Classification of the coefficient of variation to variables in beef Cattle experiments*. Ciencia Rural, Santa Maria, vol. 47(1), 1-4, http://dx.doi.org/10.1590/0103-8478cr20160946, ISSNe: 1678-4596, 2017.

[Ney38] **J. Neyman** – *Contribution to the theory of sampling human populations.* J. Amer. Statist. Assoc., Vol. 33, 101-116, 1938.

[OB14] **P. I. Ogunyinka, I. Badmus** – *Efficient linearity of online software repository variables.* 5th Series of Proceedings. International Conference on science, Technology, Education, Arts, Management and Social Sciences. Isteams Research Nexus Conference, Afe Babalola University, Ado-Ekiti, Nigeria. Pp. 453-458, 2014.

[Rob57] **D. Robson** – *Application of Multivariate Polykays to the Theory of unbiase Ratio Type Estimators.* Journal of American Statistical Association 52: 511-522, 1957.

[Sar72] **C. Sarndal** – *Sample Survey Theory Vs. General Statistical Theory: Estimation of the population mean*. Int. Stat. Rev., 40(1), 1-12, 1972.

[TF07] **B. G. Tabacknick, L. S. Fidell** – *Using Multivariate Statistics*. 5th Edition. Baston. Allyn and Bacon, 2007.