ON THE SENSITIVITY OF STUDY VARIABLE TO DATA TRANSFORMATION EXCLUSION IN CORRECTING FOR LINEARITY ASSUMPTION VIOLATION IN TWO-PHASE SAMPLING

Supported by tetfund TERTIARY EDUCATION TRUST FUND

P. I. Ogunyinka[1]*, T. J. Odule[1], O. O. Solanke[1], E. F. Ologunleko[2] and K. A. Adeleke[3]
[1]Department of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye, Nigeria
[2]Department of Statistics, University of Ibadan, Oyo State, Nigeria
[3]Department of Mathematics, Obafemi Awolowo University, Ile-Ife, Nigeria
*Corresponding author: ogunyinka.peter@gmail.com

Received: November 30, 2019 **Accepted:** February 14, 2020

Abstract: Two-phase sampling for regression has proven to be efficient in estimation especially when there is high correlation coefficient between the study and the auxiliary variable(s). However, the presence of extreme value makes the distribution to violate the basic statistical assumptions. Violation of linearity assumption by the concerned distribution, among other assumptions, may lead to type-I or type-II error in Two-phase sampling for regression. This study applied non-linear data transformation to the study variable and/or auxiliary variables. It was confirmed that data transformation is an efficient empirical tool to correct the effect of linear assumption violation in Survey Statistical Inference. However, when such data transformation is applied only to the auxiliary variable (that is, not transforming the study variable) even in the presence of high correlation coefficient, less efficient estimate would be obtained. It was concluded that simultaneous application of data transformation on both the study and auxiliary variables rather than correlation coefficient should be the condition for selecting efficient estimator in Two-phase sampling in Survey Statistical Inference.

Keywords: Data transformation, two-phase sampling, correlation level, linearity, extreme observation

Introduction

The use of inexpensive and readily available auxiliary variable(s) has improved estimation in survey statistical inference. Two-phase sampling uses auxiliary information in estimation, hence, making two-phase sampling efficient over single-phase sampling. Ogunyinka and Sodipo (2013) empirically established the superiority of two-phase sampling for regression over two-phase sampling for ratio when the linear relationship line between the study variable (y) and the auxiliary variable (x) has a zero interception on the y axis. Leys *et al.* (2013) and Inhyeok and Un (2019) concluded that the presence of extreme value or outlier in a distribution has been confirmed to lead to violation of basic statistical assumptions. Linearity assumption is one of the basic statistical assumptions that may be violated. Two-phase sampling for regression requires the confirmation of the non-violation of the linearity assumption between y and x variables. Osborne (2002) ascertained that the violation of linearity assumption may increase the probability of committing type-I or type-II error, hence, recommending non-linear data transformation as a solution to the violation of linearity assumption in linear regression model. Agunbiade and Ogunyinka (2013) established the effect of correlation levels on the precision of estimates in two-phase sampling.

There is need to know the reaction of the two-phase sampling estimator when non-linear data transformation is applied in the presence of varying correlation coefficient levels. This research empirically ascertains the justification of data transformation in correcting the effect of extreme values in the distribution. It also investigates to know the priority between data transformation and high correlation level in obtaining précised estimates in two-phase sampling. Further enquiry will also be carried out to ascertain the reaction of the study and auxiliary variables to data transformation in two-phase sampling for regression.

Materials and Methods

Data transformation in linear regression

The linear regression model is presented as

$$y = \hat{\alpha} + \hat{\beta}x + e. \qquad (1)$$

Where $\hat{\alpha}$ represents the estimated interception of the model line on the y axis of the scattered plot and $\hat{\beta}$ is the estimated regression coefficient of the model.

The use of equation (1) above requires the absence of extreme values in the distribution. Extreme value can lead to the violation of statistical assumptions (Huxley, 2016; Machado, 2018). Among these assumptions is the linearity assumption. Linearity assumption indicates that the relationship between y (dependent or study variable) and x (independent or auxiliary variable) must be linear. Hence, the relationship must product a straight line on the scattered plot of y against x. Osborne (2002) confirmed that the violation of this assumption may increase the probability of committing type-I or type-II error. Marija (2004) established that data transformation of variables can strengthen non-linear relationship to a linear (or an assumed linear) relationship. O'Hara and Hotze (2010) emphasis that the main purpose of data transformation is to get a sample data to conform to the assumptions of parametric statistics such as ANOVA, t-test and linear regression or to manage outliers or extreme values in a distribution. Marija (2004) established that data transformation technique is neither a cheating nor distortion of the true picture of the data under consideration; rather, it is a legitimate statistical tool. Literatures established that among the methods to detect an efficient transformed dataset are to establish linearity, obtain the coefficient of determination (R^2) and to conduct a significant test of the independent variable on the response variable. It is expected that coefficient of determination of the transformed data will be higher than the non-transformed or original data (though, may not be possible in all cases) and the independent variables will be significant to the response variable. Among the non-linear data transformation tools are Logarithm (of any desired based number), inverse, power, quadratic, square, square root, cube and cube root transformations. Table 1 shows some common data transformation models with the corresponding back transformation models.

Table 1: The common statistical non-linear transformation techniques

S/N	Method	Transformation	Regression Equation	Predicted/Back transformation value $(\hat{Y})$
01	Standard linear regression	None	$Y = b_0 + b_1X$	$\hat{Y} = b_0 + b_1X$
02	Exponential model	Dependent variable $(log_{10}Y)$	$log_{10}Y = b_0 + b_1X$	$\hat{Y} = 10^{(b_0+b_1X)}$
03	Quadratic model	Dependent variable $(Sqrt(Y))$	$Sqrt(Y) = b_0 + b_1X$	$\hat{Y} = (b_0 + b_1X)^2$
04	Reciprocal model	Dependent variable (y^{-1})	$Y^{-1} = b_0 + b_1X$	$\hat{Y} = 1/(b_0 + b_1X)$
05	Logarithm transformation	Independent variable $(log_{10}X)$	$Y = b_0 + b_1log_{10}X$	$\hat{Y} = b_0 + b_1 10^X$
06	Power model	Dependent variable $log_{10}Y$ and independent variable $log_{10}X$	$log_{10}Y = b_0 + b_1log_{10}X$	$\hat{Y} = 10^{(b_0+b_1log_{10}X)}$
07	Square model	Independent variable (X^2)	$Y = b_0 + b_1X^2$	$\hat{Y} = b_0 + b_1\sqrt{X}$

Back transformation is used to return a transformed predicted value to its original scale. Miller (1984) confirmed that back transformation of the predicted value gives value for median response but not the mean response as it is expected. He (Miller) concluded that back transformation on the mean of the dependent variable results to a serious bias. Hence, he established a solution to minimize this bias. Similarly, Jia and Rathi (2008) established a more efficient solution that almost removes the bias introduced by back transformation on the dependent variable.

Two-phase sampling for regression

Two-phase sampling for regression becomes necessary, among other assumptions, when $\hat{\alpha} \neq 0$ in (equation (1)). The two-phase sampling for regression estimator for estimating the population mean is presented as;

$$\overline{y}_{dl} = \overline{y} - \hat{\beta}\left(\overline{x} - \overline{x}^{'}\right). \qquad (2)$$

Where $\overline{x}^{'}$ = First phase sample mean of the auxiliary variable, $\overline{x}$= Second phase sample mean of the auxiliary variable, and$\overline{y}$= Second phase sample mean of the study variable. Okafor (2002) presented the estimated variance of $\overline{y}_{dl}$ as;

$$\hat{V}(\overline{y}_{dl}) = \left[\frac{1}{n^{'}} - \frac{1}{N}\right]s_y^2 + \left[\frac{1}{n} - \frac{1}{n^{'}}\right](s_y^2 + \hat{\beta}^2 s_x^2 - 2\hat{\beta}s_{xy}). \qquad (3)$$

In two-phase sampling for regression, in addition to the confirmation of linearity assumption, there is need for increase in the coefficient of determination (R^2) (if possible) and significance of the independent variable(s). Data transformation is expected to establish a positive correlation (relationship) between the study variable (y) and the auxiliary variable (x).

Correlation level in two-phase sampling for regression

Agunbiade and Ogunyinka (2013) established that different correlation levels in linear regression havesignificant effect on the precision of estimate in two-phase sampling for regression. It was ascertained that the higher the correlation coefficient, the better the precision of estimates in two-phase sampling. They amended the correlation coefficient boundaries for the correlation coefficient classification of Mukaka (2012). The updated correlation coefficient is presented in Table 2.

Table 2: Correlation Coefficient interpretations proposed for this investigation

Size of Correlation	Interpretation
0.90 to 1.0	Very high Positive (Negative) Correlation
0.70 to <0.90	High Positive (Negative) Correlation
0.50 to <0.70	Moderate Positive (Negative) Correlation
0.30 to <0.50	Low Positive (Negative) Correlation
0.00 to <0.30	Negligible Correlation

By further simplification, equation (3) can be expressed with respect to the correlation coefficient (ρ) giving that $\left(\hat{\beta} = \hat{\rho}\frac{s_y}{s_x}\right)$.

$$\hat{V}(\overline{y}_{dl}) = \left[\frac{1}{n^{'}} - \frac{1}{N}\right]s_y^2 + \left[\frac{1}{n} - \frac{1}{n^{'}}\right]s_y^2[1 - \hat{\rho}^2]. \qquad (4)$$

This research empirically ascertains priority between the correlation level and the data transformation inincreasing the precision of estimates in two-phase sampling for regression. Coefficient of Variation (CV) is statistical tool that compares the precision of estimates. The lower the CV, the higher the precision of the estimate under consideration. The no-scale or no-unit comparison is one major advantage of using coefficient of variation. Sequel to this advantage, this analysis uses CV for comparison and decision making. Equation (5) presents the percentage coefficient of variation used in this study.

$$CV(\overline{y}_{dl}) = \frac{\sqrt{\hat{V}(\overline{y}_{dl})}}{\overline{y}_{dl}} * 100\% \qquad (5)$$

Results and Discussion

This section uses data mined by Okikisoft software from www.sourceforge.net about the properties of the repository. Okikisoft is a source forge software repository data miner. Total software rater per software is used as the study variable (y) while the total download per software is used as the auxiliary variable (x), see Ogunyinka and Badmus (2014) for details. This section analyses data in two different categories. The first part analyses the data and compares results to justify the importance of data transformation in correcting the effect of extreme values in a distribution. The second part analyses data and compares results to establish the priority between correlation level and data transformation in order to obtain efficient estimates.

Analyses based on data transformation

Figure 1 reveals the violation of linearity assumption by the original data through the scattered plot. Fig. 2 reveals the satisfaction of linearity assumption using log_{10} transformation on both the study and auxiliary variables.

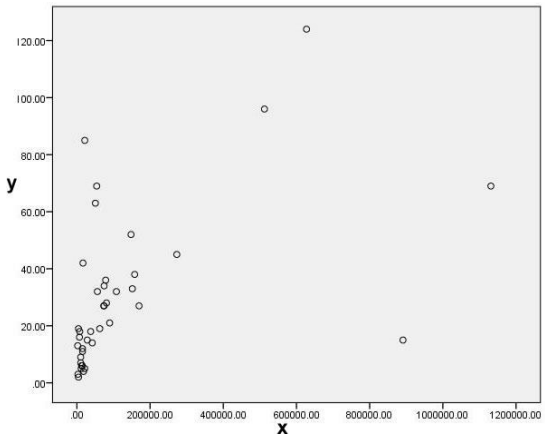

Fig. 1: Graph of y against x usingthe original data

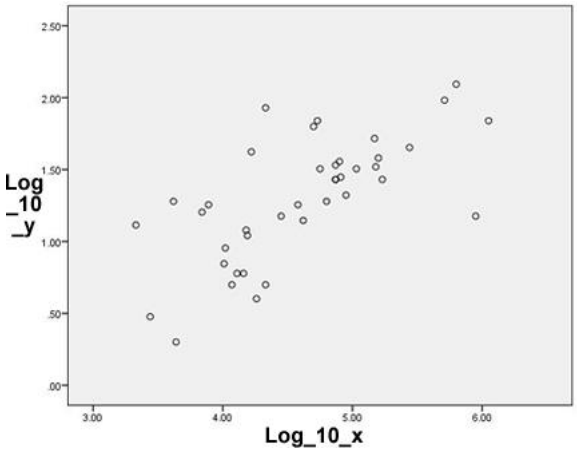

Fig. 2: Graph of Log_{10}y against Log_{10}x using transformed data

Analysis for setting priority between correlation level and data transformation in two-phase sampling

This analysis uses non-linear data transformation tools (as presented in Table 1) to create over forty (40) tables of different transformed distributions among which five (5) major analysis tables were sampled for further emphases and explanation. Transformation and analysis on each dataset produced a summary table each, hence, spanning from tables 4.0 through 8.0.

$$y^* = \alpha + \beta x^* \qquad (6)$$

Equation (6) is used in Tables 4 through 8. However, the true representation of y^* and x^* (either in transformed or original state) in equation (6) is presented in the column titled "*Combination*" in each of the five (5) selected tables.

Discussion of Results

This study aimed to investigate the justification of data transformation and prioritise between data transformation and the correlation level in obtaining more precise estimate in two-phase sampling for regression. Results, shown in Table 3, revealed that sample mean estimate obtained with transformed data has lower CV of 4.32% over that of original data of 7.32%. Hence, it makes transformed data model and estimate, in two-phase sampling, more efficient over that of original data model and estimate, respectively. Table 5 revealed that transformed data estimates with high (0.804) and moderate (0.645) correlation levels have rated $CV = 5.21\%$ and $CV = 6.51\%$ precision, respectively, over original (untransformed) data estimate with very high correlation level of 0.952. The presence of high correlation (in the original data) did not yield high précised estimate. This is contradictory to the conclusion of Agunbiade and Ogunyinka (2013), Cochran (1940) and Okafor (2002). This established that when correlation level is high but the concerned data violate linearity assumption, such data will produce a high CV, leading to low précised estimate. However, if data transformation yields in a lower correlation level, estimates from such distribution will produce a lower CV, hence, making such estimates of high precision.

Table 6 revealed data transformed estimates with high ($r = 0.745$) correlation level and moderate ($r = 0.692$) correlation level have high rated precisions of $CV = 5.9\%$ and $CV = 8.35\%$, respectively as against the original (untransformed) dataset estimate with moderate correlation level of $r = 0.643$ but with least rated precision of $CV = 23.97\%$. Hence, it is established that less précised estimates will be obtained from original (untransformed) data with low correlation or the same correlation level with transformed data. Once again, significance of data transformation is established either at the lowest or highest correlation level.

In Table 7, all the five (5) estimates belong to the same moderate correlation level. The first rated estimate of $CV = 6.7\%$ was obtained at $r = 0.647$ under data transformation over the second précised rated estimate of $CV = 10.7\%$ at $r = 0.564$ of the original (untransformed) data. This conforms to the earlier results. However, the remaining three estimates at $r = 0.591$, 0.591 and 0.657 estimates contradicted the earlier results. Further examination revealed that in these three estimates, only the auxiliary variable (x) was transformed while the study variable (y) was not transformed. It was observed that estimates that has transformed auxiliary variables but original (untransformed) study variable performed less efficient to estimates with both transformed study and auxiliary variables. In fact, it was observed that estimates with untransformed study variable but transformed auxiliary variable, though yielded high correlation coefficient but performed less efficient to the estimate with untransformed (original) study and auxiliary variables.

Table 3: Analysis to justify the importance of data transformation

Combination	Linearity Assump.	Significance	β^*	N	n′	n	$\bar{y}_{dl}$	$SE(\bar{y}_{dl})$	CV	Rating based on CV
y/x	Violated	Significant	0.0000571	17226	128	40	53.8037	3.9363	7.32%	2nd
log_{10}y/log_{10}x	Satisfied	Significant	0.432	17226	128	40	1.2888	0.0556	4.32%	1st

Table 4: Analysis result 1

Combination	R / Correlation Level	Linearity Assump.	β^*	N	n'	n	$\bar{y}_{dl}$	$SE(\bar{y}_{dl})$	CV	Rating based on CV
y/x	0.414: Low	Violated	0.000015	11448	128	40	14.06786	1.8691	13.29%	5th
$\log_2 y/\log_2 x$	0.488: Low	Satisfied	0.322	11448	128	40	3.16154	0.2181	6.90%	2nd
$y/\sqrt{x}$	0.515: Moderate	Satisfied	0.025	11448	128	40	13.98323	1.7977	12.86%	4th
$\sqrt{y}/\log_2 x$	0.591: Moderate	Satisfied	0.414	11448	128	40	3.273862	0.2211	6.75%	1st
$y/\log_{10} x$	0.629: Moderate	Satisfied	11.477	11448	128	40	13.83184	1.6956	12.26%	3rd

Table 5: Analysis result 2

Combination	R / Correlation Level	Linearity Assump.	β^*	N	n'	n	$\bar{y}_{dl}$	$SE(\bar{y}_{dl})$	CV	Rating based on CV
y/x	0.952: Very high	Violated	0.000015	13608	128	40	48.5253	5.7772	11.91%	3rd
$\log_{10} y/\sqrt[3]{x}$	0.645: Moderate	Satisfied	0.012	13608	128	40	1.1096	0.0723	6.51%	2nd
$\sqrt[3]{y}/\sqrt[3]{x}$	0.804: High	Satisfied	0.032	13608	128	40	2.5999	0.1356	5.21%	1st

Table 6: Analysis result 3

Combination	R / Correlation Level	Linearity Assump.	β^*	N	n'	n	$\bar{y}_{dl}$	$SE(\bar{y}_{dl})$	CV	Rating based on CV
y/x	0.643: Moderate	Violated	0.00012	13014	128	40	17.8727	4.2846	23.97%	3rd
$\log_{10} y/\log_{10} x$	0.692: Moderate	Satisfied	0.557	13014	128	40	0.9018	0.0753	8.35%	2nd
$\sqrt[3]{y}/\log_2 x$	0.745: High	Satisfied	0.327	13014	128	40	2.2219	0.1310	5.90%	1st

Table 7: Analysis result 4

Combination	R / Correlation Level	Linearity Assump.	β^*	N	n'	n	$\bar{y}_{dl}$	$SE(\bar{y}_{dl})$	CV (%)	Rating based on CV
y/x	0.564: Moderate	Violated	$5.581*10^5$	28674	128	40	45.433	4.871995	10.724	2nd
$y/\log_{10} x$	0.591: Moderate	Satisfied	30.493	28674	128	40	35.69253	4.804563	13.461	5th
$y/\log_2 x$	0.591: Moderate	Satisfied	9.189	28674	128	40	35.70783	4.804563	13.455	4th
$\sqrt{y}/\sqrt{x}$	0.647: Moderate	Satisfied	0.006	28674	128	40	5.41253	0.364839	6.741	1st
$y/\sqrt[3]{x}$	0.657: Moderate	Satisfied	0.951	28674	128	40	37.03364	4.621729	12.480	3rd

Table 8: Analysis result 5

Combination	R / Correlation Level	Linearity Assump.	β^*	N	n'	n	$\bar{y}_{dl}$	$SE(\bar{y}_{dl})$	CV (%)	Rating based on CV
y/x	0.498: Low	Violated	0.000164	10719	128	40	18.59269	2.9427	15.827	5th
$y/\sqrt{x}$	0.558: Moderate	Satisfied	0.101	10719	128	40	20.27614	2.8641	14.125	3rd
$y/\sqrt[3]{x}$	0.573: Moderate	Satisfied	0.925	10719	128	40	20.10558	2.8426	14.139	4th
$\sqrt[3]{y}/\sqrt{x}$	0.612: Moderate	Satisfied	0.005	10719	128	40	2.218786	0.1225	5.522	1st
$\log_{10} y/\log_{10} x$	0.703: High	Satisfied	0.700	10719	128	40	0.911318	0.0700	7.685	2nd

It was, also, observed in Tables 4, 7 and 8 that when only the auxiliary variable is transformed but the study variable is not transformed, the estimate will not be efficient compared to when both the study and auxiliary variables were transformed. It is very important to report that this result was confirmed to be true in all the data analyses including those analysis tables that could not be reported in this article. This significant discovery shows that while data transformation is an efficient empirical tool for the correction of the effect of extreme observations or outlier in any distribution, it is very important that both the study and the auxiliary variables must be transformed, simultaneously, in order to obtain the most efficient estimate in two-phase sampling. Though, this discovery is limited to survey statistical inference, it is recommended that the general statistical inference should also be cautious of this important discovery about data transformation method in statistics. Hence, it is recommended that similar investigation should be conducted in the general statistical inference.

Conclusion

The presence of extreme observation or outlier in the distribution will lead to the violation of statistical assumptions. Violation of basic assumptions like linearity in the use of two-phase sampling is a serious challenge even in the presence of high correlation coefficient between the study and auxiliary variables. Data transformation is a recommended efficient empirical tool for the correction of this violation. However, data transformation that is applied only to the auxiliary variable for the distribution to conform to linearity assumption in the presence of high correlation coefficient will reduce the efficiency of the estimate. This study strongly recommends that for good précised estimates to be obtained in two-phase sampling where concerned distributions disobey linearity assumption, non-linear data transformation must be performed on both the study and auxiliary variables, simultaneously, while the correlation coefficient attained after data transformation should be a secondary condition for the selection of efficient estimate.

Conflict of Interest

Authors declare that there is no conflict of interest related to this paper.

References

Agunbiade DA & Ogunyinka PI 2013. Effect of correlation level on the use of Auxiliary variable in Double sampling for regression estimation. *Open J. Statistics*, 3(5): 312-318. Doi:http//dx.doi.org/10.4236/ojs.2013.35037.

Cho H, Jeong ST, Ko DH & Son KP 2014. Efficient Outlier detection of the water temperature monitoring data. *J. Korean Soc. Coast. Ocean Engr.*, 26: 285–291 (In Korean).

Cochran WG 1940. The estimation of the yields of cereal experiments by sampling for the ratio of grains in total produce. *J. Agric. Sci.*, 30(2): 262-275.

Huxley TH 2016. Outing the Outliers–Tails of the Unexpected. In: Proceedings of the ICEAA International Training Symposium, Bristol, UK, 17–20 October, 2016.

Inhyeok B & Un J 2019. Outlier detection and smoothing process for water level data measured by ultrasonic sensor in stream flows. *Water*, 11: 951, doi:10.3390/w11050951.

Jia Siwei & Sarika Rathi 2008. On Predicting Log-transformed Linear Models with Heteroscedasticity, SAS Global Forum 2008 – Statistics and Data Analysis, paper 370-2008.

Leys C, Ley C, Klein O, Bernard P & Licata L 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.*, 49: 764–766.

Machado JMO 2018. Outlier Detection in Accounting Data.Master's Thesis, University of Porto, Porto, Portugal.

Marija J & Norusis 2004. SPSS 12.0 Guide to Data Analysis. Prentice Hall Inc. ISBN 0-13-147886-9.

Miller D 1984. Reducing transformation bias in curve fitting. *The American Statistician*, 30(2): 124-126.

Mukaka MM 2012. A guide to appropriate use of correlation coefficient in medical research. *MMJ*. 24(3): 69-71.

O'Hara Robert B & Hotze Johan D 2010. Do not log-transform count data. *Methods in Ecology and Evolution*. Retrieved on Sept. 13, 2014, from Doi:10.1111/j.2041-210x.2010.00021.x.

Ogunyinka PI & Sodipo AA 2014. Efficiency of Ratio and Regression Estimators using Double sampling. Application to household and online software Repository data sample surveys. LAP LAMBERT Academic Publishing. ISBN: 978-3-659-50156-2.

Ogunyinka PI & Badmus Idowu 2014. Efficient Linearity of Online Software Repository Variables. Proceedings of the 5th International Conference on Science, Technology, Education, Arts, Management and Social Sciences. 29-31st May 2014. Cjiemeka S.C., Longe O., Ekuobase G., Asani T. and Olaniyi T. K. Eds. Nigeria: AfeBabalola University (ABUAD) Ado-Ekiti, pp. 453-458.

Okafor FC 2002. Survey Theory with Applications. Afro-Orbis Publication ltd.

Osborne Jason 2002. Notes on the use of data transformations. *Pract. Assess., Res. and Eval.*, 8(6).