Digital Innovations
& Contemporary Research in SCIENCE, ENGINEERING & TECHNOLOGY
Vol. 10. No. 3, September 2022

Article Citation Format
AnosikeJ,U, & Ogoke, U.P. (2022): A Comparative Analysis On The Model Adequacy Of Four Transformation Techniques. Journal of Digital Innovations & Contemporary Research in Science, Engineering & Tech. Vol. 10, No. 3. Pp 137-147
DOI: dx.doi.org/10.22624/AIMS/DIGITAL/V10N1P10x

Article Progress Time Stamps

Article Type: Research Article
Manuscript Received: 14th July, 2022
Review Type: Blind
Final Acceptance: 20th August, 2022

A Comparative Analysis on the Model Adequacy of Four Transformation Techniques

Anosike Joseph Ugonna & Ogoke Uchenna Petronilla
Department of Mathematics and Statistics
University of Port Harcourt
Choba, Port Harcourt, Nigeria
Email: juanohugo@gmail.com; uchenna.ogoke@uniport.edu.ng

ABSTRACT

This study compared four different transformation techniques by applying a simple linear regression on raw and transformed data. The R^2 of each model was obtained and a test on the significance of these R^2 was carried out. Also, the r_{xy}(coefficient of correlation) were also obtained. The data used is a secondary data consisting of 53years (1964-2016) of the infant mortality rate in Nigeria (https://www.ceicdata.com/en/nigeria/health-statistics/ng-mortality-rate-infant-per-1000-live-births). The r_{xy} were also compared and the results, 95.8%, 95.8%, 96.2%, 93.0%, and 92.9% respectively. The R^2 obtained for the raw data, logarithm, square-root, square and inverse are as follows: 91.8%, 91.7%, 92.5%, 86.6% and 86.4% respectively. However, the R^2 obtained for the raw data, logarithm, square-root, square and inverse compete favourably but the performance of inverse transformation suits the data most in terms of model accuracy.

Key Words: Transformation, Raw data, logarithm, square-root, square, inverse

1. INTRODUCTION

Data transformation is a process of converting data or information from one format to another, usually from the format of source system into the required format of a new destination system. There are a great variety of possible data transformations, from adding constants to multiplying, squaring, or raising to a power, converting to logarithmic scales, inverting, taking the square root of the values and even applying trigonometric transformation such as sine wave transformation. Simple linear regression captures the linear relationship between the expected value of Y and an independent variable say X. If linearity fails to hold, even approximately, it is sometimes possible to transform either the independent or dependent variables in the regression model to improve linearity. When fitting a linear regression model, one assumes that there is a linear relationship between the response variable and each of the explanatory variable. However, in many situations there may instead be a non-linear relationship between the variables.

This can sometimes be remedied by applying suitable transformation to some (or all) of the variables. Transformations can be used to correct violations of assumptions such as constant error variance and normality. The primary reasons for data transformation, as they are used for improving the compatibility of data with assumptions underlying a modelling process include viz: to stabilize the variance of the dependent variable, to normalize and linearity. Many statistical procedures assume that the variables are normally distributed. A significant violation of the assumption of normality can seriously increase the chance of the researcher committing either a type I or II error (depending on the nature of the analysis and the non-normality). However, Mecceri (1989) points out that true normality is exceedingly rare in education and psychology. Thus, one reason researchers utilize data transformations is improving the normality of variables. Zimmerman (1995, 1998) pointed the importance of normality in all statistical analysis whether parametric or non-parametric tests.

Two other reasons for non-normality are the presence of outliers and the nature of the variable itself. Judd and Clelland (1989) argued that outliers' removal is desirable, honest, and important. However, not all researchers feel that way (Orr, etal, 1991). Transformation can be useful to a researcher needing to know whether a variable's distribution is significantly different from a normal (or other) distribution (Rosenthal (1968), and Wilcox (1997)). Most people find it difficult to accept the idea of transforming data. Turkey (1977) probably had the right idea when he called data transformation calculation "re-expression" rather than "transformations Tabachnich and Fidell (2001) recommended transformation as a remedy for outliers and for failures of normality, linearity, and homoscedasticity, they are not universally recommended. Different techniques of transformation has been treated in different forms (see Mc Neil (1977),Velleman and Hoaglin, (1981) but this research aims to select the best technique to use in terms of model adequacy.

2. METHOD

The four (4) methods of data transformation techniques adopted is briefly explained using their mathematical formula.

2.1 Different Transformation Techniques

The linear relationship assumed in the preceding analysis may be inappropriate in some problems. Indeed, non-linearities may be expected in the real world situations. Bearing in mind the complexity in analysis, model transformation becomes inevitable. This is to be able to use a regression model of simple forms in the transformed variables, rather than a more complicated one in the original variables. Some of the most common forms (and transformations) of non-linear models as used in this research are presented by the following polynomials.

i) $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + e$ (1)

This is usually called the **curvilinear regression model**.

Let $z_i = x^i$, $(i=1,2,\ldots,k)$. Then

$Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_k z_k + e$ (2)

ii) $y = \beta_0 + \beta_1 x_1^{-1} + \beta_2 x_2^{-1} + \beta_k x_k^{-1} + e$ (3)

let $z_i = x_i^{-1}$ (called **inverse or reciprocal transformation**), then

$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_k z_k + e$ (4)

iii) $$y = \beta_0 x_1^{\beta 1} x_2^{\beta 2} \ldots x_k^{\beta_k} e \quad (5)$$

Taking logarithm of both sides (**called the log-transformation**), we have

$$\ln y = \ln\beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \ldots + \beta_k \ln x_k + e \quad (6)$$

Or equivalently $$y' = \beta_0' + \beta_1' z_1 + \beta_2' z_2 + \ldots + \beta_k' z_k + e \quad (7)$$

iv) $$y = \beta_0 + \beta_1 x_1^{1/2} + \beta_2 x_2^{1/2} + \ldots + \beta_k x_k^{1/2} + e \quad (8)$$

$z_i = x_i^{1/2}$ (**called the square root transformation**)

$$y' = \beta_0 + \beta_1 z_1 + \beta_2 z_2 = \ldots + \beta_k z_k + e \quad (9)$$

v) Given a production function (say),

$$y = \left(\alpha_1 x_1^{\rho} + \alpha_2 x_2^{\rho}\right)^{v/\rho}$$

on transformation, this gives the equation

$$y^{\rho/v} = \alpha_1 x_1^{\rho} + \alpha_2 x_2^{\rho} \quad (10)$$

Thus each observation on output (y) should be raised to the power of ρ/v and each observation on capital (x_1) and labour input (x_2) are raised to power p. This is an example of **power transformation of the variables.**

Note: After transforming the variables, the usual method of estimating the parameters is employed

3. RESULTS

The results of this research will be presented in a tabular form which will be followed by the discussions on the tables listed.

Table 1: Infant Mortality Rate for Nigeria, Number per 1,000 Live Births, Annual (From 1964 to 2016)

X	Y (per 1000)	Y	LOG10(Y)	SQRT(Y)	SQR(Y)	INV(Y)
1	195.7	195700	5.29159083	442.3799	38298490000	0.00000511
2	191.2	191200	5.28148789	437.2642	36557440000	0.00000523
3	186.9	186900	5.2716093	432.3193	34931610000	0.00000535
4	182.8	182800	5.26197619	427.5512	33415840000	0.00000547
5	178.4	178400	5.25139485	422.3742	31826560000	0.00000561
6	174.1	174100	5.24079877	417.2529	30310810000	0.00000574
7	169.4	169400	5.22891341	411.5823	28696360000	0.00000590
8	164.6	164600	5.21642983	405.7093	27093160000	0.00000608
9	159.6	159600	5.20303289	399.4997	25472160000	0.00000627
10	154.6	154600	5.18920949	393.1921	23901160000	0.00000647
11	149.5	149500	5.17464119	386.6523	22350250000	0.00000669
12	144.7	144700	5.16046853	380.3945	20938090000	0.00000691
13	140.1	140100	5.14643814	374.2993	19628010000	0.00000714
14	135.9	135900	5.13321946	368.6462	18468810000	0.00000736
15	132.2	132200	5.12123146	363.5932	17476840000	0.00000756
16	129.3	129300	5.11159852	359.5831	16718490000	0.00000773
17	127.0	127000	5.10380372	356.3706	16129000000	0.00000787
18	125.4	125400	5.09829754	354.1186	15725160000	0.00000797
19	124.4	124400	5.09482038	352.7038	15475360000	0.00000804
20	124.0	124000	5.09342169	352.1363	15376000000	0.00000806
21	124.1	124100	5.09377178	352.2783	15400810000	0.00000806
22	124.5	124500	5.09516935	352.8456	15500250000	0.00000803
23	125.1	125100	5.09725731	353.6948	15650010000	0.00000799
24	125.6	125600	5.09898964	354.4009	15775360000	0.00000796
25	126.0	126000	5.10037055	354.9648	15876000000	0.00000794
26	126.2	126200	5.10105935	355.2464	15926440000	0.00000792
27	126.2	126200	5.10105935	355.2464	15926440000	0.00000792
28	126.0	126000	5.10037055	354.9648	15876000000	0.00000794
29	125.6	125600	5.09898964	354.4009	15775360000	0.00000796
30	125.3	125300	5.09795107	353.9774	15700090000	0.00000798
31	124.6	124600	5.09551804	352.9873	15525160000	0.00000803
32	123.6	123600	5.09201847	351.5679	15276960000	0.00000809
33	122.2	122200	5.08707121	349.5712	14932840000	0.00000818
34	120.2	120200	5.07990447	346.6987	14448040000	0.00000832
35	117.8	117800	5.07114529	343.22	13876840000	0.00000849
36	115.2	115200	5.06145248	339.4113	13271040000	0.00000868
37	112.3	112300	5.05037976	335.1119	12611290000	0.00000890

X	Y (per 1000)	Y	LOG10(Y)	SQRT(Y)	SQR(Y)	INV(Y)
38	109.2	109200	5.03822264	330.4542	11924640000	0.00000916
39	106.1	106100	5.02571538	325.7299	11257210000	0.00000943
40	102.9	102900	5.01241537	320.7803	10588410000	0.00000972
41	99.8	99800	4.99913054	315.9114	9960040000	0.00001002
42	96.5	96500	4.98452731	310.6445	9312250000	0.00001036
43	93.2	93200	4.96941591	305.2868	8686240000	0.00001073
44	90.0	90000	4.95424251	300	8100000000	0.00001111
45	87.0	87000	4.93951925	294.9576	7569000000	0.00001149
46	83.9	83900	4.92376196	289.655	7039210000	0.00001192
47	81.1	81100	4.90902085	284.7806	6577210000	0.00001233
48	78.3	78300	4.89376176	279.8214	6130890000	0.00001277
49	75.7	75700	4.87909588	275.1363	5730490000	0.00001321
50	73.3	73300	4.86510397	270.7397	5372890000	0.00001364
51	71.0	71000	4.85125835	266.4583	5041000000	0.00001408
52	69.0	69000	4.83884909	262.6785	4761000000	0.00001449
53	66.9	66900	4.82542612	258.6503	4475610000	0.00001495

3.1 Test of R^2 Using F-Statistics

We test for the significance of R^2 using the F- statistics.

$F = \frac{R^2/k}{1-R^2/n-k-1}$ ~ $F_{k,\ n-k-1\ (\alpha)}$; where, k is the number of regression coefficient or parameter, n is the number of observation., α is the level of significance.

3.2 Hypothesis

H_0: $R^2 = 0$ (R^2 Not significant) vs H_1: $R^2 > 0$ (R^2 is significant)

Using the given hypothesis, the raw data and the transformed data will be tested at 5% level of significance and the results presented on Table 2 below.

Table 2: TEST FOR THE SIGNIFICANCE OF R^2

TEST FOR THE SIGNIFICANCE OF R^2				
DATA SET	R^2	F_{cal}	F_{tab}	DECISION
Raw data	0.918	279.88	3.18	Significant
Log	0.917	276.20	3.18	Significant
Square root	0.925	308.33	3.18	Significant
Square	0.866	161.57	3.18	Significant
Inverse	0.864	160.00	3.18	Significant

4. DISCUSSION

Table I shows the raw and transformed data which was fitted using SPSS software. The fits were examined using linear regression models, coefficient of determination (see appendix).
Table 2 shows the R^2 of the original data, log transform data, square-root transform data, square transform data data were 91.8%, 91.7%, 92.5%, 86.6%, and 86.4% respectively. These results showed a clear picture in terms of the competitiveness in modeling (fitting) data. The linear regression model was obtained for the raw data set and the four different transforms. The respective R^2 of data set were then examined to determine the total variation of the mortality rate explained by the changes in period (year). Obviously, the square-root transform had the highest of the R^2. Also considered was the test of normality for all the data set (see appendix). It was seen clearly that the square-root transform did not differ significantly from normality. It gave a more normal distribution from a skewness value very close to zero as required for a normal distribution than the rest of the data set. Furthermore, the correlation coefficient r_{xy} was also considered for the raw data, log transform data, square-root transform data, square transform data, and the inverse transform data, the result were as follows; 95.8%, 95.8%, 96.2%, 93%, 92.9% respectively. Consequently, upon the results obtained from the foregoing, the square-root transformation is the best method of data transformation for the data used in this research.

REFERENCES

1. Goodman, L.A. (1954). Kolmogorov-smirnov tests for psychological Research. Psychological-Bulletin, 51, 160-168.
2. https://www.tibco.com/reference-center/what-is-data-transformation
3. Judd, C.M. and Mc Clelland, G.H. (1989). Data analysis: A model-comparison approach. San Diego, CA: Harcourt Brace Jovanovich.
4. Micceri, T. (1989).The Unicorn, the Normal Curve and Other Improbable Creatures. Psychological Bulletin, 105, 156-166.
5. Mc Neil, Donald (1977). Interactive data Analysis: A practical primer. wiley-interscience.
6. Orr, J. M., Sackett, P.R., Dubois, C.L.Z. (1991). Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. Personnel psychology, 44: 473-486.
7. Rosenthal, R. (1968). An application of the kolmogorov-smirnov test for normality with estimated mean and variance, psychological-reports, 22,570.
8. Turkey, John (1977). Exploratory data analysis (Addison-wesley).
9. Tabachnick, B.G., and Fidell, L.S. (2001). Using multivariate statistics. New York: Harper Collins.
10. Velleman, Paul F. and David C. Hoaglin (1981). Applications, Bases and computing of exploratory data analysis (Duxbury press).
11. Wilcox, R.R. (1997). Some practical reasons for re-considering the kolmogorov-smirnov test. British journal of mathematical and statistical psychology, 50(1), 9-20.
12. Zimmerman, D.W. (1995). Increasing the power of non-parametric tests by detecting and down weighting outliers. Journal of experimental education, 64, 71-78.
13. Zimmerman, D.W. (1998). Invalidation of parametric and non-parametric statistical tests byconcurrent violation of two assumptions, journal of experimental education, 67, 55-68.

APPENDIX 1
SPSS OUTPUT OF REGRESSION OF Y AGAINST X (ORIGINAL DATA SET)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.958[a]	.918	.916	9489.259	.918	567.238	1	51	.000

a. Predictors: (Constant), X
b. Dependent Variable: Y

ANOVA[a]

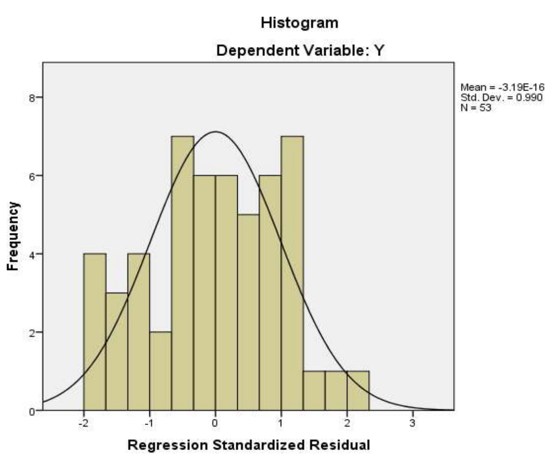
Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	51077524567.812	1	51077524567.812	567.238	.000[b]
	Residual	4592347507.660	51	90046029.562		
	Total	55669872075.472	52			

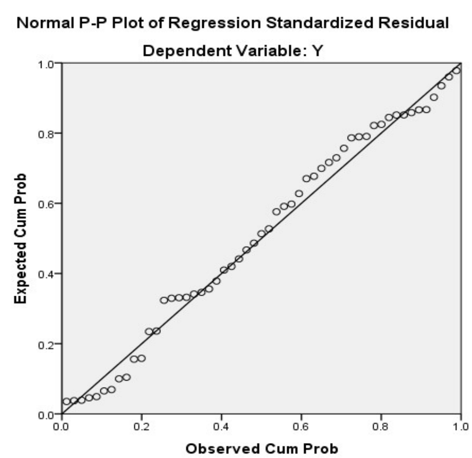
a. Dependent Variable: Y b. Predictors: (Constant), X

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	178646.807	2644.232		67.561	.000
	X	-2029.407	85.209	-.958	-23.817	.000

a. Dependent Variable: Y





Charts of the Original Data Set

APPENDIX 2

SPSS OUTPUT OF REGRESSION OF LOG Y AGAINST X (LOG TRANSFORMATION)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.958[a]	.917	.916	.03444	.917	566.838	1	51	.000

a. Predictors: (Constant), X
b. Dependent Variable: LOGY

ANOVA[a]

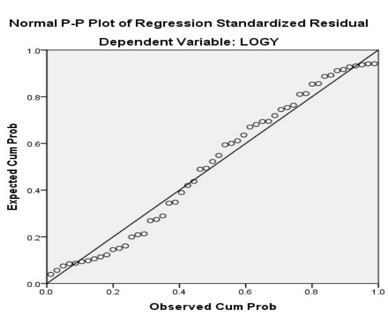
Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	.672	1	.672	566.838	.000[b]
	Residual	.060	51	.001		
	Total	.733	52			

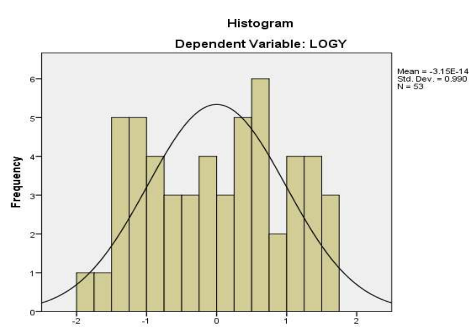
a. Dependent Variable: LOGY
b. Predictors: (Constant), X

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.276	.010		549.841	.000
	X	-.007	.000	-.958	-23.808	.000

a. Dependent Variable: LOGY





Charts of the Log transformation

Digital Innovations
& Contemporary Research in SCIENCE, ENGINEERING & TECHNOLOGY
Vol. 10. No. 3, September 2022

APPENDIX 3
SPSS OUTPUT OF REGRESSION OF SQUARE ROOT OF Y AGAINST X (SQUARE ROOT TRANSFORMATION)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.962[a]	.925	.923	12.95551	.925	626.435	1	51	.000

a. Predictors: (Constant), X
b. Dependent Variable: SQRTY

ANOVA[a]

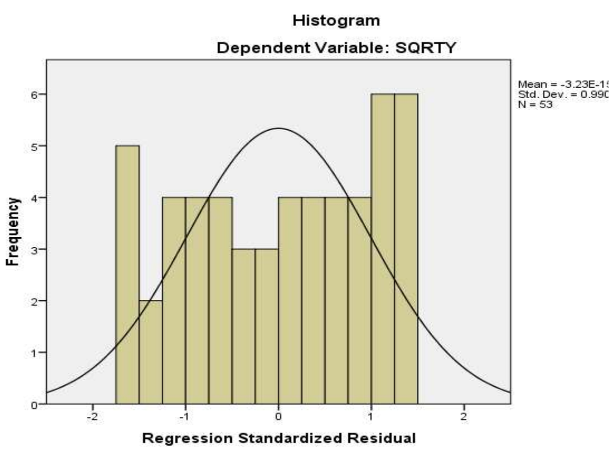
Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	105144.193	1	105144.193	626.435	.000[b]
	Residual	8560.108	51	167.845		
	Total	113704.301	52			

a. Dependent Variable: SQRT
b. Predictors: (Constant), X

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.
		B	Std. Error	Beta		
1	(Constant)	427.482	3.610		118.412	.000
	X	-2.912	.116	-.962	-25.029	.000

a. Dependent Variable: SQRTY



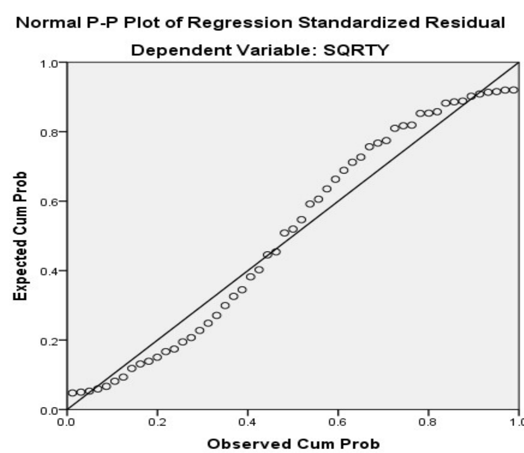


Chart of the Square root transformation

APPENDIX 4
SPSS OUTPUT OF REGRESSION OF SQUARE OF Y AGAINST X (SQUARE TRANSFORMATION)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.930[a]	.866	.863	3154956977.925	.866	328.955	1	51	.000

a. Predictors: (Constant), X
b. Dependent Variable: SQR

ANOVA[a]

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	3274341349815751000000.000	1	3274341349815751000000.000	328.955	.000[b]
	Residual	507641430160326300000.000	51	9953753532555418000.000		
	Total	3781982779976077300000.000	52			

a. Dependent Variable: SQRY
b. Predictors: (Constant), X

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	30263201661.829	879145427.183		34.423	.000
	X	-513826043.380	28330065.455	-.930	-18.137	.000

a. Dependent Variable: SQRY

Digital Innovations
& Contemporary Research in SCIENCE, ENGINEERING & TECHNOLOGY
Vol. 10. No. 3, September 2022

APPENDIX 5: SPSS OUTPUT OF REGRESSION OF INVERSE OF Y AGAINST X. (INVERSE TRANSFORMATION)

MODEL SUMMARY

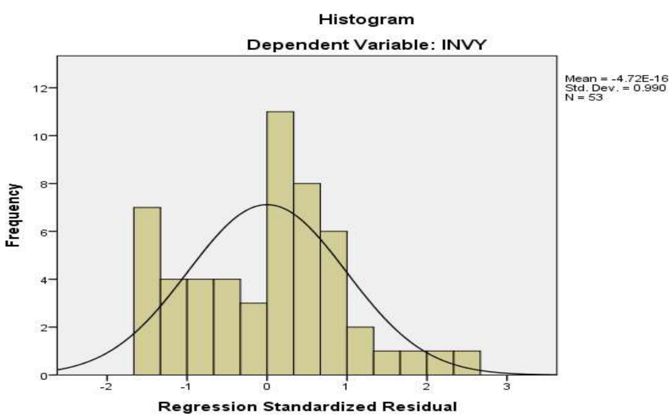
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.929[a]	.864	.861	.0000009294	.864	323.302	1	51	.000

a. Predictors: (Constant), X
b. Dependent Variable: INVY

ANOVA

Model		Sum of Squares	Df	Mean Square	F	Sig.
	Regression	.000	1	.000	323.302	.000[b]
1	Residual	.000	51	.000		
	Total	.000	52			

a. Dependent Variable: INVY
b. Predictors: (Constant), X



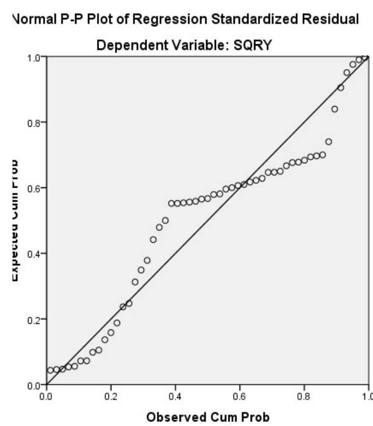


Chart of the Inverse transformation