Graphical Diagnostic and Minimum Variance Criteria for Comparing Transformed and Untransformed Linear Regression Models

Yusuf O. Afolabi[1] and Peter A. Oluwagunwa[2]

[1,2]Department of Mathematics and Statistics, Rufus Giwa Polytechnic, Owo, Nigeria
[1]ayusufolasunkanmi@gmail.com, [2]petergunwa@gmail.com

Abstract: Embarking on various transformations of linear regression model was investigated with a keen interest on the core difference between transformed and untransformed variables whose interdependence is closely related. Mean square error and error diagnostic analysis were adopted as basis for adjudging the best linear regression model. It was deduced from available results that transformations of logarithm (bases 10 and 7), square root, reciprocal and inverse square were found to have among others possessed the minimum mean square errors of 0.001 and lesser. In contrary, when compared with the qq-plots and other residual plots from the diagnostic analysis, logarithm transformation of base 10 was acknowledged to have performed better among other transformation competitors. Therefore, error diagnostic analysis should form part of reliable yardsticks apart from the minimum condition of least mean square error for selecting best linear regression model when transformations of closely related variables with same number of observations are involved.

[Afolabi, Y. O. and Oluwagunwa, P. A. **Graphical Diagnostic and Minimum Variance Criteria for Comparing Transformed and Untransformed Linear Regression Models.** *N Y Sci J* 2017;10(8):183-193]. ISSN 1554-0200 (print); ISSN 2375-723X (online). http://www.sciencepub.net/newyork. 21. doi:10.7537/marsnys100817.21.

Keywords: Transformed and untransformed variables, Linear regression model, Mean square error, Quantile-quantile plot

1. Introduction

Regression model techniques are centred on the field of econometrics for determining any functional relationship that may exist between two or more variables of interest which need not to be based only on a priori economic rationale but on the form or shape of the scatter plot thereby giving rooms for transformations that might be required for linearity.

Cochran (1947) identified some difficulties which can best be appreciated by considering the main assumptions regarding the nature of the observations that are necessary before an analysis of variance can be considered valid. it is often assumed that observation $y_1, y_2, ..., y_n$ are independently normally distributed with constant variance and with expectations specified by a model which is linear in a set of parameters.

Transformation is applied in context of regression analysis as the most powerful tool that is widely used and one of the most abused as well. It is used to offer important set of tools for understanding the association between two or more variables as many important results in statistical analyses follow from the assumption that the population being sampled or investigated is normally distributed with a common variance and additive error structure.

Simplification of the model is achieved through transforming the dependent or independent variable in a regression model which often reduce the complexity of the model required to fit the data. This simplicity is often seen as reducing the degree of the polynomial required to fit a curve as relevant theoretical assumptions relating to a selected method of analysis are approximately satisfied making the usual procedures applicable in order to make inferences about the unknown parameters of interest.

Data transformations are commonly useful tools that serve many functions in quantitative analysis including improving the normality of a distribution and equalizing variance to meet assumptions and improve effect sizes, thus constituting important aspects of data cleaning and preparations for statistical analyses. There are many potential types of data transformations as mathematical functions of some of the commonly discussed traditional transformations include adding constants, square root, converting to logarithms (e.g. base 10, natural logarithm) scales and applying, inverting and reflecting trigonometric transformations such as sine wave transformations (Berry, 1990; Cleveland, 1984; Draper and Smith, 1998; Keene, 1995 and Osborne, 2002).

Moreover, there are many reasons to utilize transformations as the focus of this paper is on transformations that improve normality of data as both parametric and non-parametric tests tend to benefit from normally distributed data (Zimmerman, 1998). While transformations are important tools, they should be utilized thoughtfully as they fundamentally alter the nature of the variable making the interpretation of the results somehow more complex.

Keene (1995) pointed out that many of the approaches to decisions on transformations are essentially subjective and have led to a widespread suspicion of the use of any transformation as unnecessary data transformation should be avoided and should a data transformation performed in an event, the rationale for the choice of data transformation along with interpretation of the estimates of treatment effects based on transformed data should be provided. Thus, some authors suggest reversing the transformation once the analyses are done for reporting of means, standard deviations, graphing, etc in which this decision ultimately depends on the nature of the hypotheses and analyses which is best left to the discretion of the researcher.

However, when presence of collinearity is pronounced among the variables of interest, the value of the estimated coefficients in the sample may differ markedly from the true value in the population as this is often seen as a core problem by social scientists. The univariate objective is generally to create a transformed variable that is more normally distributed (Osborne, 2002 and 2008) to cater for the presence or degree of collinearity as failure to give serious consideration leads to regression coefficients with large standard errors and resulted to faulty conclusion.

The aim of this research is to distinctly make a comparison between transformed and untransformed variables of regression models using mean square error (MSE) and error graphical diagnostic criteria to determine a better fit and recommend an appropriate measure of transformation that will checkmate the presence of closely related variables.

2. Estimation Of Parameters In Multiple Linear Regression Model

The multiple linear regression model is examined by transforming its variables on bases (10, 7, 5, 2 and exponent), square root, reciprocal, inverse square root, inverse square, sine and deviates compared to its unadulterated variables (that is, untransformed variables). While mean square errors (MSE) of the analysis of variance (ANOVA) for both transformed and untransformed variables of the regression models will be adjudged as the basis of comparison to determine the most appropriate working model since number of variables and observations are equal.

$$Y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

indicating observational form

$$Y_1 = \beta_o + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \ldots + \beta_k x_{1k} + \varepsilon_1$$
$$Y_2 = \beta_o + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23} + \ldots + \beta_k x_{2k} + \varepsilon_2$$
$$Y_3 = \beta_o + \beta_1 x_{31} + \beta_2 x_{32} + \beta_3 x_{33} + \ldots + \beta_k x_{3k} + \varepsilon_3$$
$$\vdots$$
$$Y_n = \beta_o + \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + \ldots + \beta_k x_{nk} + \varepsilon_n$$

rewritten in matrix observational form

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & x_{33} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

represented in matrix form

$$Y = X\beta + \varepsilon \tag{1}$$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & x_{33} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

X is called the design matrix.

2.1. Method of Maximum Likelihood for Estimating βs

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_k X_{ik} + \varepsilon_i$$
$$\mathrm{E}(Y_i) = \mu = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_k X_{ik}$$
$$V(Y_I) = \delta^2$$
$$Y_i \sim N(\mu, \delta^2)$$

that is,

$$Y_i \sim N(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_k X_{ik}, \delta^2)$$
$$f(Y_i) = \frac{1}{\sqrt{2\pi\delta^2}} \exp\left[-\frac{1}{2\delta^2} (Y_i - \mu)^2 \right]$$

the likelihood function becomes

$$L\left(Y_1, Y_2, \ldots, Y_n; \beta_0, \beta_1, \ldots, \beta_k, \delta^2\right)$$
$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\delta^2}} \exp\left[-\frac{1}{2\delta^2}\left(Y_i - \mu\right)^2\right]$$
$$= \left(2\pi\delta^2\right)^{-n/2} \exp\left[-\frac{1}{2\delta^2}\sum_{i=1}^{n}\left(Y_i - \mu\right)^2\right]$$

taking the natural logarithm of the likelihood function

$$l = -\frac{n}{2}\ln\left(2\pi\delta^2\right) - \frac{\sum_{i=1}^{n}\left(Y - \mu\right)^2}{2\delta^2} \quad (2)$$

and thereafter the partial derivatives with respect to

$$\beta_0, \beta_1, \cdots, \beta_k$$

$$\frac{\partial l}{\partial \beta_0} = \frac{\sum_{i=1}^{n}\left(Y_i - \mu\right)}{\delta^2}$$

$$= \frac{1}{\delta^2}\sum_{i=1}^{n}\left(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \beta_3 X_{i3} - \cdots - \beta_k X_{ik}\right)$$

setting $\frac{\partial l}{\partial \beta_0}$ to zero and rearranging

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1\sum_{i=1}^{n} X_{i1} + \beta_2\sum_{i=1}^{n} X_{i2} + \beta_3\sum_{i=1}^{n} X_{i3} + \cdots + \beta_k\sum_{i=1}^{n} X_{ik} \quad (3)$$

$$\frac{\partial l}{\partial \beta_1} = \frac{\sum_{i=1}^{n} X_{i1}\left(Y_i - \mu\right)}{\delta^2}$$

$$= \frac{1}{\delta^2}\sum_{i=1}^{n} X_{i1}\left(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \beta_3 X_{i3} - \cdots - \beta_k X_{ik}\right)$$

setting $\frac{\partial l}{\partial \beta_1}$ to zero and rearranging

$$\sum_{i=1}^{n} X_{i1} Y_i = \beta_0\sum_{i=1}^{n} X_{i1} + \beta_1\sum_{i=1}^{n} X_{i1}^2 + \beta_2\sum_{i=1}^{n} X_{i1}X_{i2} + \cdots + \beta_k\sum_{i=1}^{n} X_{i1}X_{ik} \quad (4)$$

it therefore follows subsequently,

$$\frac{\partial l}{\partial \beta_k} = \frac{\sum_{i=1}^{n} X_{ik}\left(Y_i - \mu\right)}{\delta^2}$$

$$= \frac{1}{\delta^2}\sum_{i=1}^{n} X_{ik}\left(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \beta_3 X_{i3} - \cdots - \beta_k X_{ik}\right)$$

setting $\frac{\partial l}{\partial \beta_k}$ to zero and rearranging

$$\sum_{i=1}^{n} X_{ik} Y_i = \beta_0\sum_{i=1}^{n} X_{ik} + \beta_1\sum_{i=1}^{n} X_{ik}X_{i1} + \beta_2\sum_{i=1}^{n} X_{ik}X_{i2} + \cdots + \beta_k\sum_{i=1}^{n} X_{ik}^2 \quad (5)$$

Combining equations (3), (4) and (5) called **normal system of linear equations**

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1\sum_{i=1}^{n} X_{i1} + \beta_2\sum_{i=1}^{n} X_{i2} + \beta_3\sum_{i=1}^{n} X_{i3} + \cdots + \beta_k\sum_{i=1}^{n} X_{ik}$$

$$\sum_{i=1}^{n} X_{i1} Y_i = \beta_0\sum_{i=1}^{n} X_{i1} + \beta_1\sum_{i=1}^{n} X_{i1}^2 + \beta_2\sum_{i=1}^{n} X_{i1}X_{i2} + \cdots + \beta_k\sum_{i=1}^{n} X_{i1}X_{ik}$$

$$\vdots$$

$$\sum_{i=1}^{n} X_{ik} Y_i = \beta_0\sum_{i=1}^{n} X_{ik} + \beta_1\sum_{i=1}^{n} X_{ik}X_{i1} + \beta_2\sum_{i=1}^{n} X_{ik}X_{i2} + \cdots + \beta_k\sum_{i=1}^{n} X_{ik}^2$$

which can be represented in matrix as

$$\begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{i1} Y_i \\ \vdots \\ \sum_{i=1}^{n} X_{ik} Y_i \end{bmatrix} =$$

$$\begin{bmatrix} n & \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i2} & \sum_{i=1}^{n} X_{i3} & \cdots & \sum_{i=1}^{n} X_{ik} \\ \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i1}^2 & \sum_{i=1}^{n} X_{i1}X_{i2} & \sum_{i=1}^{n} X_{i1}X_{i3} & \cdots & \sum_{i=1}^{n} X_{i1}X_{ik} \\ & & \vdots & & & \\ \sum_{i=1}^{n} X_{ik} & \sum_{i=1}^{n} X_{ik}X_{i1} & \sum_{i=1}^{n} X_{ik}X_{i2} & \sum_{i=1}^{n} X_{ik}X_{i3} & \cdots & \sum_{i=1}^{n} X_{ik}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

where,

$$X^T Y = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{i1} Y_i \\ \vdots \\ \sum_{i=1}^{n} X_{ik} Y_i \end{bmatrix}$$

$$X^T X = \begin{bmatrix} n & \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i2} & \sum_{i=1}^{n} X_{i3} & \cdots & \sum_{i=1}^{n} X_{ik} \\ \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i1}^2 & \sum_{i=1}^{n} X_{i1}X_{i2} & \sum_{i=1}^{n} X_{i1}X_{i3} & \cdots & \sum_{i=1}^{n} X_{i1}X_{ik} \\ & & \vdots & & & \\ \sum_{i=1}^{n} X_{ik} & \sum_{i=1}^{n} X_{ik}X_{i1} & \sum_{i=1}^{n} X_{ik}X_{i2} & \sum_{i=1}^{n} X_{ik}X_{i3} & \cdots & \sum_{i=1}^{n} X_{ik}^2 \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

rewritten in matrix form as

$$X^T Y = X^T X\beta \quad (6)$$

Considering equation (1)

$$Y = X\beta + \varepsilon$$
$$\mathrm{E}(Y) = \mu = X\beta$$
$$V(Y) = \delta^2 I$$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}, \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

X is an $n \times k$ matrix with i^{th} row given by $[f_1(x_{i1}), f_2(x_{i2}), f_3(x_{i3}),..., f_k(x_{ik})]$ and $\beta^T = [\beta_0, \beta_1, \beta_2,..., \beta_K]$, concentrating on estimating βs

The idea of least square estimation is that we find an estimator $\hat{\beta}$ of β which minimizes the sum of error squared

$$S = \sum_{i=1}^{n} (Y_i - \mu)^2$$

which can be rewritten in matrix form as

$$S = (Y - X\beta)^T (Y - X\beta)$$
$$S = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta$$

differentiating S with respect to β,

$$\frac{\partial S}{\partial \beta} = -2X^T T + 2\beta X^T Y$$

$\frac{\partial S}{\partial \beta}$ setting to zero and rearranging, the result becomes

$$X^T X\beta = X^T Y \quad (7)$$

called **normal system of linear equations**.

Resolving equations (6) and (7) from methods of maximum likelihood and least squares, the estimator $\hat{\beta}$ for β becomes

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (8)$$

2.2. Estimation of Sums of Square Error (SSE) and Mean Square Error (MSE)

Recall from equation (8) that

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

indicating

$$\hat{Y} = X\hat{\beta}$$
$$\hat{Y} = X(X^T X)^{-1} X^T Y$$
$$\hat{Y}_{n\times 1} = H_{n\times n} Y_{n\times 1}$$

$H_{n\times n}$ is $X(X^T X)^{-1} X^T$

$H_{n\times n}$ is called Hat matrix and is both symmetric and idempotent, that's $H^T = H$ and $HH = H$ respectively.

$$\varepsilon_{n\times 1} = Y_{n\times 1} - \hat{Y}_{n\times 1} = Y - X\hat{\beta} = Y - HY = (1 - H)Y$$
$$SSE = \varepsilon^T \varepsilon = \left(Y - X\hat{\beta}\right)^T \left(Y - X\hat{\beta}\right)$$
$$= Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta}$$
$$= Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T \hat{Y}$$
$$= Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X(X^T X)^{-1} X^T Y$$
$$= Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T IX^T Y$$

since $X^T X(X^T X)^{-1} = I$

$$= Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T \hat{Y}$$
$$= Y^T Y - \beta^T X^T Y$$
$$= Y^T (I - H)Y$$
$$SSE = Y^T Y - \hat{\beta} X^T Y$$
$$MSE = \frac{SSE}{n - k}$$

where

k is the number of estimated parameters excluding the regression intercept and n is the number of observations.

3. Analysis and Results

X_1, X_2, X_3, X_4 are considered in the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{4k} + \varepsilon_i$$

or

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

with closely related variables of interest that are positively correlated with one another as its data were simulated through economic rationale in R (a statistical programming software).

3.1. Regression Model of the Untransformed Variables

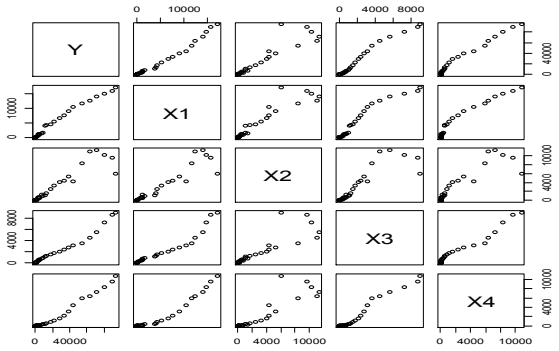

Figure 1: Matrix Plot of the Untransformed Variables of the Regression Model.

Consider the linear model

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

with estimated model

$$\hat{Y} = -2004813 + 2.0467 X_1 + 0.7389 X_2 + 3.2374 X_3 + 2.3090 X_4$$

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	2.7345e+10	2.7345e+10	57162.0765	< 2.2e-16 ***
X2	1	1.3028e+06	1.3028e+06	2.7234	0.1093
X3	1	4.2093e+08	4.2093e+08	879.9115	< 2.2e-16 ***
X4	1	4.5024e+07	4.5024e+07	94.1187	9.223e-11 ***
Res	30	1.4351e+07	4.7838e+05		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

MSE = 4.7838e+05

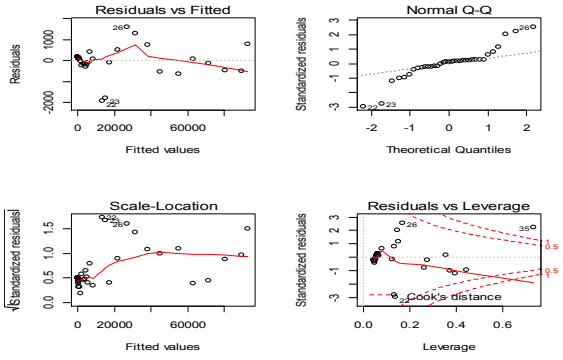

Figure 2: Error Diagnostic Analysis of the Untransformed Regression Model.

3.2. Regression Model of the Logarithm (Base Exponent) Transformed Variables

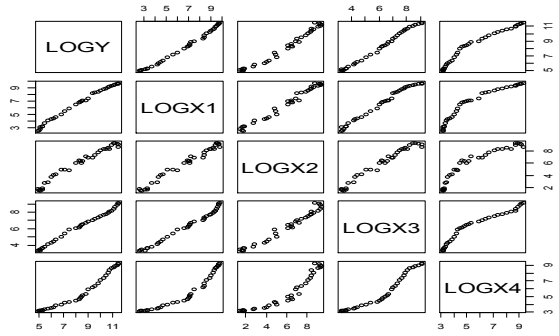

Figure 3: Matrix Plot of the Logarithm (Base Exponent) Transformed Variables of the Regression Model.

Consider the linear model

$$\log_e Y = \beta_0 + \beta_1 \log_e X_1 + \beta_2 \log_e X_2 + \beta_3 \log_e X_3 + \beta_4 \log_e X_4$$

with estimated model

$$\hat{Y} = 1.9840 + 0.3641 \log_e X_1 + 0.0946 \log_e X_2 + 0.3413 \log_e X_3 + 0.2159 \log_e X_4$$

Analysis of Variance Table

Response: LOGY

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	168.956	168.956	34823.2226	< 2.2e-16 ***
X2	1	0.010	0.010	1.9991	0.1677
X3	1	0.837	0.837	172.5939	5.619e-14 ***
X4	1	0.533	0.533	109.9036	1.506e-11 ***
Res	30	0.146	0.005		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

MSE = 0.005

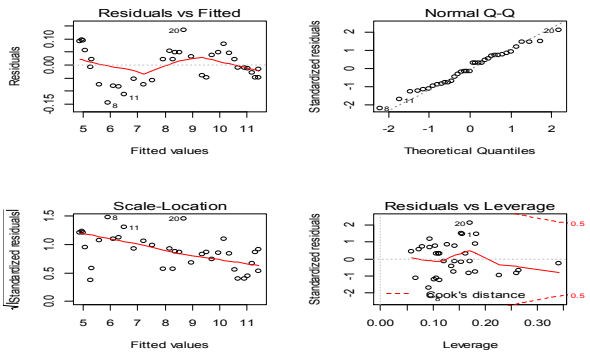

Figure 4: Error Diagnostic Analysis of the Transformed Logarithm (Base Exponent) Regression Model.

3.3. Regression Model of the Logarithm (Base 10) Transformed Variables

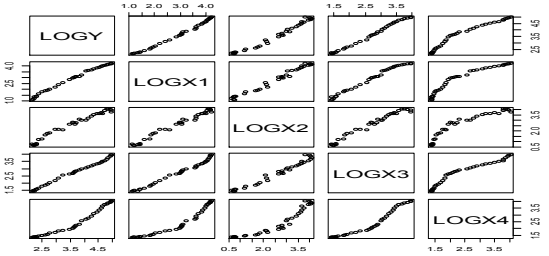

Figure 5: Matrix Plot of the Logarithm (Base 10) Transformed Variables of the Regression Model.

Consider the linear model

$$\log_0 Y = \beta_0 + \beta_1 \log_0 X_1 + \beta_2 \log_0 X_2 + \beta_3 \log_0 X_3 + \beta_4 \log_0 X_4$$

with estimated model

$$\hat{Y} = 0.8617 + 0.364 \log_0 X_1 + 0.0946 \log_0 X_2 + 0.3413 \log_0 X_3 + 0.2159 \log_0 X_4$$

Analysis of Variance Table

Response: LOGY

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	31.867	31.867	34823.2226	< 2.2e-16 ***
X2	1	0.002	0.002	1.9991	0.1677
X3	1	0.158	0.158	172.5939	5.619e-14 ***
X4	1	0.101	0.101	109.9036	1.506e-11 ***
Res	30	0.027	0.001		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

MSE = 0.001

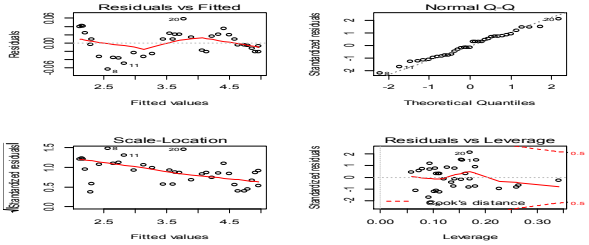

Figure 6: Error Diagnostic Analysis of the Transformed Logarithm (Base 10) Regression Model.

3.4. Regression Model of the Logarithm (Base 7) Transformed Variables

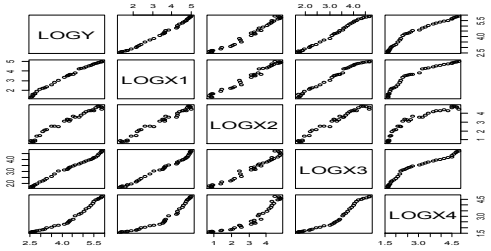

Figure 7: Matrix Plot of the Logarithm (Base 7) Transformed Variables of the Regression Model.

Consider the linear model

$$\log_7 Y = \beta_0 + \beta_1 \log_7 X_1 + \beta_2 \log_7 X_2 + \beta_3 \log_7 X_3 + \beta_4 \log_7 X_4$$

with estimated model

$$\hat{Y} = 0.8617 + 0.364 \log_7 X_1 + 0.0946 \log_7 X_2 + 0.3413 \log_7 X_3 + 0.2159 \log_7 X_4$$

Analysis of Variance Table

Response: LOGY

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	31.867	31.867	34823.2226	< 2.2e-16 ***
X2	1	0.002	0.002	1.9991	0.1677
X3	1	0.158	0.158	172.5939	5.619e-14 ***
X4	1	0.101	0.101	109.9036	1.506e-11 ***
Res	30	0.027	0.001		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

MSE = 0.001

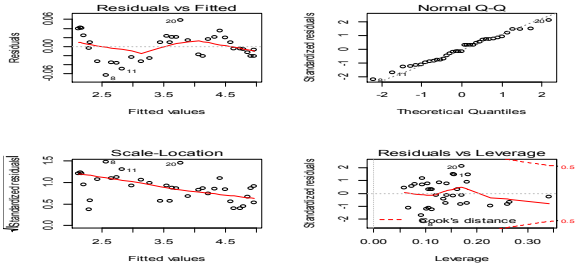

Figure 8: Error Diagnostic Analysis of the Transformed Logarithm (Base 7) Regression Model.

3.5. Regression Model of the Logarithm (Base 5) Transformed Variables

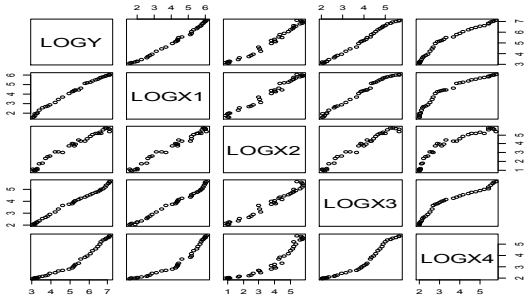

Figure 9: Matrix Plot of the Logarithm (Base 5) Transformed Variables of the Regression Model.

Consider the linear model

$$\log_5 Y = \beta_0 + \beta_1 \log_5 X_1 + \beta_2 \log_5 X_2 + \beta_3 \log_5 X_3 + \beta_4 \log_5 X_4$$

with estimated model

$$\hat{Y} = 2.8623 + 0.364 \log_5 X_1 + 0.0946 \log_5 X_2 + 0.3413 \log_5 X_3 + 0.2159 \log_5 X_4$$

Analysis of Variance Table

Response: LOGY

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	31.867	31.867	34823.2226	<2.2e-16 ***
X2	1	0.002	0.002	1.9991	0.1677
X3	1	0.158	0.158	172.5939	5.619e-14 ***
X4	1	0.101	0.101	109.9036	1.506e-11 ***
Res	30	0.027	0.001		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

MSE = 0.01

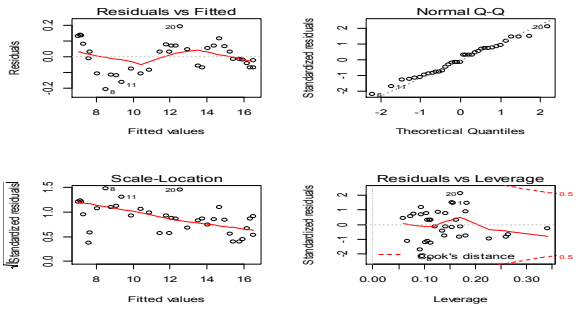

Figure 10: Error Diagnostic Analysis of the Transformed Logarithm (Base 5) Regression Model.

3.6. Regression Model of the Logarithm (Base 2) Transformed Variables

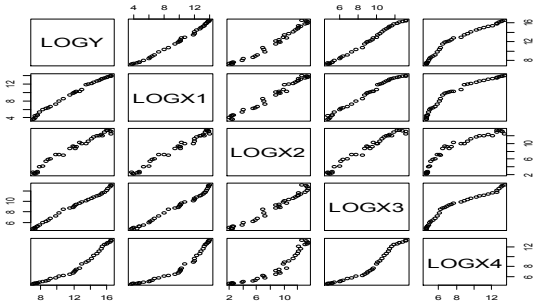

Figure 11: Matrix Plot of the Logarithm (Base 2) Transformed Variables of the Regression Model.

Consider the linear model

$$\log_2 Y = \beta_0 + \beta_1 \log_2 X_1 + \beta_2 \log_2 X_2 + \beta_3 \log_2 X_3 + \beta_4 \log_2 X_4$$

with the estimated model

$$\hat{Y} = 2.8623 + 0.3641 \log_2 X_1 + 0.0946 \log_2 X_2 + 0.3413 \log_2 X_3 + 0.2159 \log_2 X_4$$

Analysis of Variance Table

Response: LOGY

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	351.66	351.66	34823.2226	< 2.2e-16 ***
X2	1	0.02	0.02	1.9991	0.1677
X3	1	1.74	1.74	172.5939	5.619e-14 ***
X4	1	1.11	1.11	109.9036	1.506e-11 ***
Res	30	0.30	0.01		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

MSE = 0.01

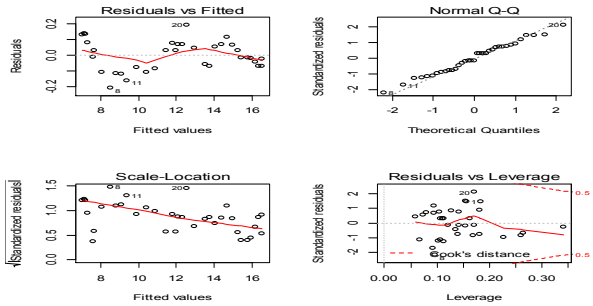

Figure 12: Error Diagnostic Analysis of the Transformed Logarithm (Base 2) Regression Model.

3.7. Regression Model of the Square Root Transformed Variables

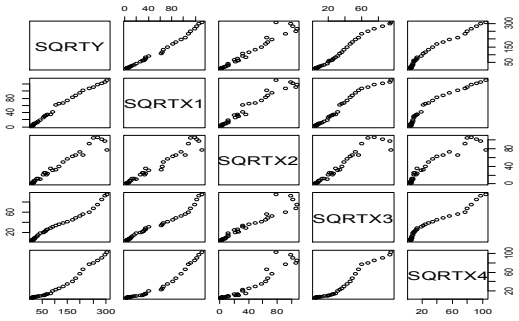

Figure 13: Matrix Plot of the Square Root Transformed Variables of the Regression Model.

Consider the linear model

$$\sqrt{Y} = \beta_0 + \beta_1 \sqrt{X_1} + \beta_2 \sqrt{X_2} + \beta_3 \sqrt{X_3} + \beta_4 \sqrt{X_4}$$

with the estimated model

$$\hat{Y} = -1.1987 + 0.8750 \sqrt{X_1} + 0.3251 \sqrt{X_2} + 0.9734 \sqrt{X_3} + 0.6995 \sqrt{X_4}$$

Analysis of Variance Table

Response: SQRTY

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	313321	313321	59489.419	< 2.2e-16 ***
X2	1	369	369	70.103	2.407e-09 ***
X3	1	2536	2536	481.557	< 2.2e-16 ***
X4	1	776	776	147.312	4.198e-13 ***
Res	30	158	5		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

MSE = 5

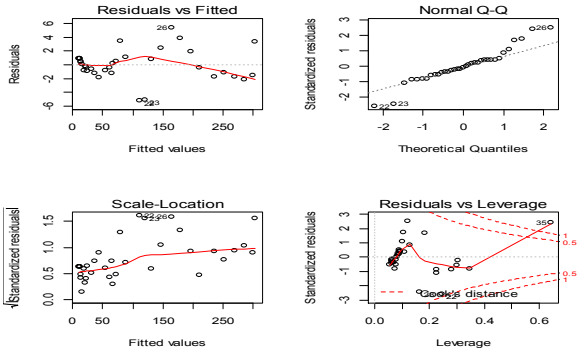

Figure 14: Error Diagnostic Analysis of the Square Root Transformation Regression Model.

3.8. Regression Model of the Reciprocal Transformed Variables

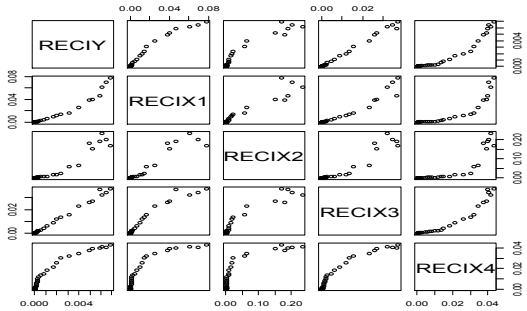

Figure 15: Matrix Plot of the Reciprocal Transformed Variables of the Regression Model

Consider the linear model.

$$1/Y = \beta_0 + \beta_1 1/X_1 + \beta_2 1/X_2 + \beta_3 1/X_3 + \beta_4 1/X_4$$

with estimated model

$$\hat{Y} = -6.83e{-}5 + 2.130e{-}2\ 1/X_1 + 3.310e{-}3\ 1/X_2 + 9.58e{-}2\ 1/X_3 + 2.723e{-}2\ 1/X_4$$

Analysis of Variance Table
Response: RECIY

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1.6338e-04	1.6338e-04	11386.518	< 2.2e-16 ***
X2	1	1.7320e-06	1.7320e-06	120.679	4.905e-12 ***
X3	1	7.0580e-06	7.0580e-06	491.856	< 2.2e-16 ***
X4	1	4.1800e-07	4.1800e-07	29.097	7.654e-06 ***
Res	30	4.3000e-07	1.4000e-08		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

MSE = 1.4000e-08

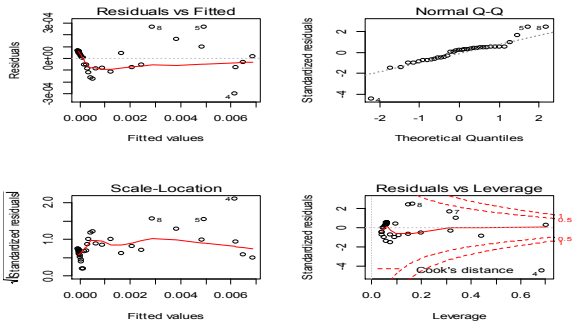

Figure 16: Error Diagnostic Analysis of the Reciprocal Transformation Regression Model.

3.9. Regression Model of the Inverse Square Root Transformed Variables

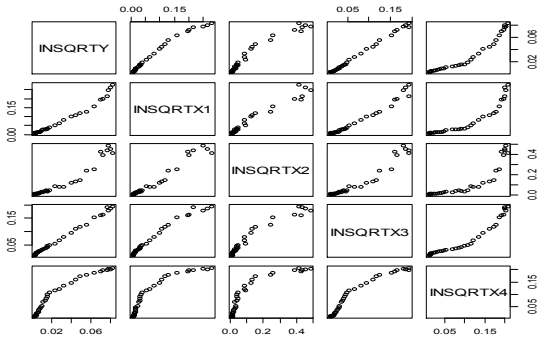

Figure 17: Matrix Plot of the Inverse Square Root Transformed Variables of the Regression Model.

Consider the linear model

$$1/\sqrt{Y} = \beta_0 + \beta_1 1/\sqrt{X_1} + \beta_2 1/\sqrt{X_2} + \beta_3 1/\sqrt{X_3} + \beta_4 1/\sqrt{X_4}$$

with estimated model

$$\hat{Y} = -0.0020 + 0.0559\ 1/\sqrt{X_1} + 0.0128\ 1/\sqrt{X_2} + 0.2796\ 1/\sqrt{X_3} + 0.0524\ 1/\sqrt{X_4}$$

Analysis of Variance Table
Response: INSQRTY

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	0.0251725	0.0251725	15684.9431	< 2.2e-16 ***
X2	1	0.0000003	0.0000003	0.2001	0.6578555
X3	1	0.0006043	0.0006043	376.5688	< 2.2e-16 ***
X4	1	0.0000260	0.0000260	16.1825	0.0003587 ***
Res	30	0.0000481	0.0000016		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

MSE = 0.0000016

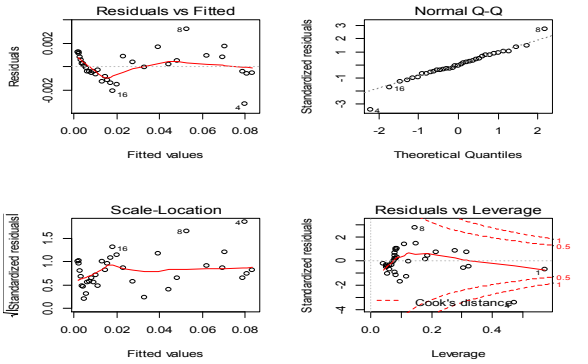

Figure 18: Error Diagnostic Analysis of the Inverse Square Root Transformation Regression Model.

3.10. Regression Model of the Inverse Square Transformed Variables

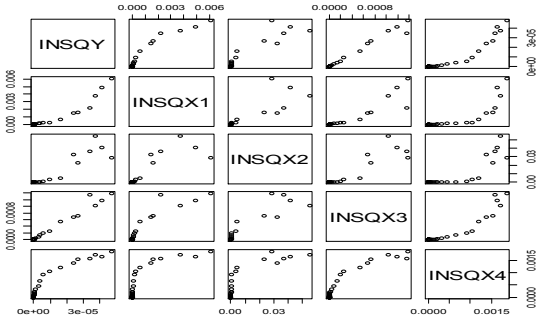

Figure 19: Matrix Plot of the Inverse Square Transformed Variables of the Regression Model.

Consider the linear model

$$1/Y^2 = \beta_0 + \beta_1 1/X_1^2 + \beta_2 1/X_2^2 + \beta_3 1/X_3^2 + \beta_4 1/X_4^2$$

with estimated model

$$\hat{Y} = -2.546e{-}07 + 2.790e{-}03\, 1/X_1^2 + 1.139e{-}04\, 1/X_2^2 + 1.286e{-}02\, 1/X_3^2 + 5.226e{-}03\, 1/X_4^2$$

Analysis of Variance Table
Response: INSQY

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	5.8481e-09	5.8481e-09	8342.889	< 2.2e-16 ***
X2	1	3.3180e-10	3.3180e-10	473.370	< 2.2e-16 ***
X3	1	2.6330e-10	2.6330e-10	375.684	< 2.2e-16 ***
X4	1	4.4600e-11	4.4600e-11	63.683	6.596e-09 ***
Res	30	2.1000e-11	7.0000e-13		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

MSE = 7.0000e-13

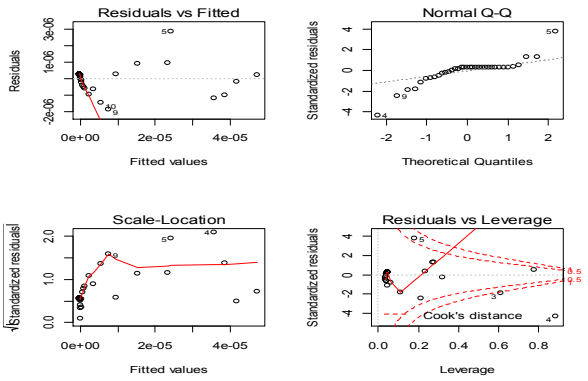

Figure 20: Error Diagnostic Analysis of the Inverse Square Transformation Regression Model.

3.11. Regression Model of the Sine Transformed Variables

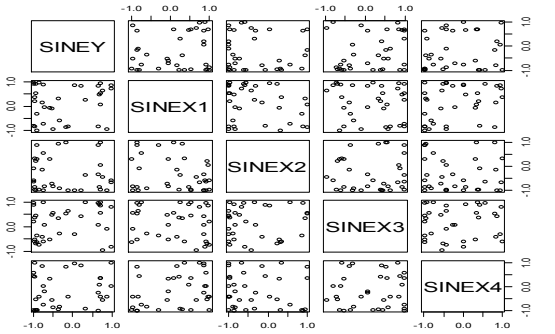

Figure 21: Matrix Plot of the Sine Transformed Variables of the Regression Model.

Consider the linear model

$$\sin Y = \beta_0 + \beta_1 \sin X_1 + \beta_2 \sin X_2 + \beta_3 \sin X_3 + \beta_4 \sin X_4$$

with estimated model

$$\hat{Y} = -0.2819 + 0.1765 \sin X_1 - 0.0033 \sin X_2 + 0.2975 \sin X_3 + 0.0845 \sin X_4$$

Analysis of Variance Table
Response: SINEY

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SINEX1	1	0.4648	0.46483	0.8618	0.3606
SINEX2	1	0.0000	0.00000	0.0000	0.9991
SINEX3	1	1.2744	1.27442	2.3629	0.1347
SINEX4	1	0.1047	0.10467	0.1941	0.6627
Res	30	16.1805	0.53935		

MSE = 0.53935

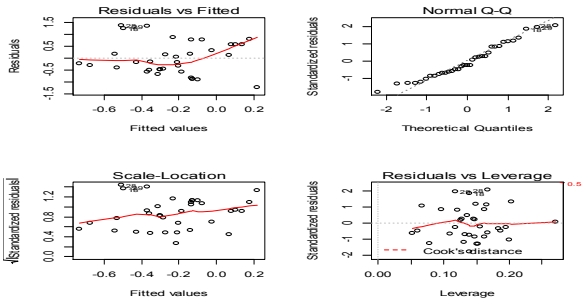

Figure 22: Error Diagnostic Analysis of the Sine Transformation Regression Model.

3.12. Regression Model of the Deviate Transformed Variables

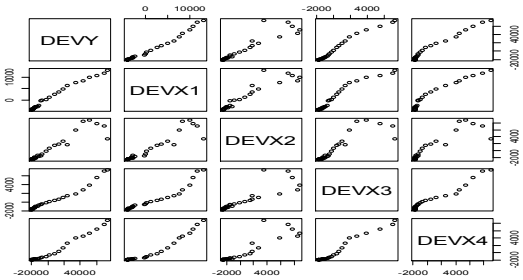

Figure 23: Matrix Plot of the Deviate Transformed Variables of the Regression Model.

Consider the linear model

$$\left(Y-\bar{Y}\right)=\beta_0+\beta_1\left(X_1-\bar{X}_1\right)+\beta_2\left(X_2-\bar{X}_2\right)+\beta_3\left(X_3-\bar{X}_3\right)+\beta_4\left(X_4-\bar{X}_4\right)$$

with the estimated model

$$\hat{Y}=-4.615e{-}12+2.047\left(X_1-\bar{X}_1\right)+7.389e{-}1\left(X_2-\bar{X}_2\right)+3.237\left(X_3-\bar{X}_3\right)+2.309\left(X_4-\bar{X}_4\right)$$

Analysis of Variance Table
Response: DEVY

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
DEVX1	1	2.7345e+10	2.7345e+10	57162.077	<2.2e-16 ***
DEVX2	1	1.3028e+06	1.3028e+06	2.7234	0.1093
DEVX3	1	4.2093e+08	4.2093e+08	879.9115	<2.2e-16 ***
DEVX4	1	4.5024e+07	4.5024e+07	94.1187	9.223e-11 ***
Res	30	1.4351e+07	4.7838e+05		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

MSE = 4.7838e+05

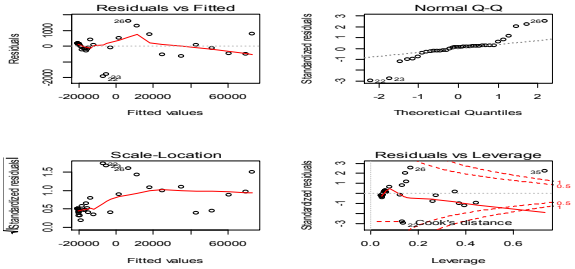

Figure 24: Error Diagnostic Analysis of the Deviate Transformation Regression Model.

4. Discussion of Results

The comparison of transformed and untransformed variables of a linear regression model was examined through various matrix plots, mean square errors and the normal quantile-quantile plots.

Figures (1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 and 23) reflected the relationship among individual variables of interest of the model for both transformed and untransformed as they all supported fitting a linear regression model through their scatter plots. However, the linear regression models fitted for the logarithm transformations of bases (10, 7, 5 & 2) resulted to same regression coefficients except for their regression intercepts that differed in values for base exponent, same for higher bases of ten & seven and lower bases of five & two whereas for untransformed, square root, reciprocal, inverse square root, inverse square, sine and deviate transformations were found to be negative in value as their regression coefficients were not in any form related to each other unlike that of logarithm transformations.

The essence of this research work as to which transformation is adjudged to fit a better linear regression model was deduced from the matrix plot of scatter diagram, mean square error and normal quantile-quantile plot for logarithm of bases (10, 7, 5 and 2), inverse square root, reciprocal and inverse square transformations. Results reflected logarithm (bases 7 and 10), reciprocal, inverse square root and inverse square transformations to be having comparatively small mean square errors with values ranging from 7.0000e-13 to 0.001 for figures (6, 8, 16, 18 and 20). It is therefore inferred that for economic variables of this interest in which they are closely related, the transformations with minimum mean square error should be preferred over others but further consultations from the transformation error diagnostic analyses in figures (6, 8, 16, 18 and 20) suggested that logarithmic transformations of bases (7 and 10) should be recognized over others as mean square error should not be the only minimum criteria for identifying best fitted transformation of variables for linear regression model. It was also discovered

that deviate transformation does not in any way differed from the untransformed model as both mean square errors and error diagnostic analyses in figures (2 and 24) were the same.

5. Conclusion

Transformations of linear regression model was investigated in this study to determine the transformation which can be recommended among logarithm, square root, reciprocal, inverse square root, inverse square, sine and deviate when closely related economic variables are involved. Findings revealed that all logarithmic transformations of different bases can be engaged as their mean square errors were relatively small compared to others except for inverse square root, reciprocal and inverse square transformations which also displaced higher sense of belonging in the transformations of linear regression model with its mean square errors being extremely very small but unfavoured by other regression model error diagnostic measures.

Transformation of variables for linear regression model should be encouraged among users of closely related economic variables considering its importance and usefulness in regression analysis but the choice of transformations to be considered should be guided by literatures and further diagnostics on linear models. Therefore, in line with this study, it is advisable to work with logarithmic transformation of base 10 considering the fact that all the closely related economic variables involved are positively correlated which is in line with the conclusion of (Cleveland, 1984).

We therefore recommended logarithmic transformations as being appropriate and suitable especially of base ten for econometricians, statisticians and other users of economic variables with incessant relationship over other transformations simply for the fact that the results of significance for untransformed linear regression model remained unaltered.

8/9/2017

Yusuf O. Afolabi and Peter A. Oluwagunwa

Both work as lecturers in the Department of Mathematics and Statistics of Rufus Giwa Polytechnic, Owo, Nigeria with interests in Mathematical & Environmental Statistics and Financial Mathematics respectively.

References

1. Berry, D. A. (1990). Basic Principles in Designing and Analyzing Clinical Studies in Statistical Methodology in the Pharmaceutical Sciences in Berry, D. A. (Ed.). *Marcel Dekker,* New York.
2. Cleveland, W. S. (1984). Graphical Methods for Data Presentation Full Scale Breaks, Dot Charts and Multi-Based Logging. *The American Statistician.* Vol.38(4), pp. 270-280.
3. Cochran, W. G. (1947). Some Consequences When the Assumptions for the Analysis of Variance are not Satisfied. *Biometrika.* Vol.3, pp. 22-38.
4. Draper, N. R. and Smith, H. (1998). Applied Regression Analysis. *Wiley-Interscience Publication John Wiley & Sons, Inc*. New York.
5. Keene, O. N. (1995). The Log Transformation is Special. *Statistics in Medicine.* Vol.14, pp. 811-819.
6. Osborne, J. W. (2002). Notes on the Use of Data Transformations. *Practical Assessment, Research & Evaluation.* Vol.8(6).
7. Osborne, J. W. (2008). Best Practices in Data Transformation; The Overlooked Effect of Minimum Values in Osborne, J. W. (Ed.) Best Practices in Quantitative Methods, Thousand Oaks, CA: SAGE Publishing.
8. Zimmerman, D. W. (1998). Invalidation of Parametric and Non-Parametric Statistical Tests by Concurrent Violation of Two Assumptions. *Journal of Experimental Education,* vol.67, pp. 55-68.