Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.4, No.9, 2014

Optimizing Transformation for Linearity between Online Software Repository Variables.

Ogunyinka, Peter I[1*]and Badmus, Nofiu Idowu[2]

1. Department of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye, Nigeria
2. Department of Statistics, Abraham Adesanya Polytechnic, Ijebu-Igbo, Nigeria

* E-mail of the corresponding author: pixelgoldprod@yahoo.com

Abstract

Online Software Repositories (OSR) like sourceforge.net and google code contain a wealth of valuable data about software projects but these data violate the linearity and normal assumptions, hence making the data impossible for use in most statistical data analysis. To prepare these data for statistical data analysis, the data were non-linearly transformed, hence, these research established the best twelve (12) transformed model that obey linearity assumptions, higher coefficient of determination (R^2), positive and negative relationship and gained variable significance over the original data. Similarly, the back transformation or interpretation was provided about each of these twelve (12) best ranked linear models to solve the challenges of data transformation encountered by researchers.

Keywords: Data transformation, Linear regression model, OkikiSoft, Online Software Repository and Sourceforge.net.

1. Introduction

Online Software repository (OSR) is a web based storage of Computer software. OSR contains a wealth of valuable information about software projects. Ahmed (2008) gave types of software repository as historical repositories, run-time repositories and Code repositories. This research focuses on the code repositories (CR). CR, such as www.sourceforge.net and google code (www.code.google.com), host the source codes of various applications developed by several developers. According to Ahmed (2008), very often, data available on OSR exhibit large amount of noise and skew. The use of such data may lead to incorrect results and conclusions. Ahmed (2008) recommends that software repository researchers should closely study the noise and skew in the data and better understand the effect on the analysis. Statistical visualization is essential to spot the noise and skewness. He concluded in his recommendation that OSR researchers should provide guidelines and tools to improve the quality of repository data. This research uses mining software repository (MSR) software called OKIKISOFT. Okikisoft is the authors developed artificial intelligence software for automatic mining of data on the webpages of sourceforge.net. The statistical analysis of the data mined on the repository revealed the violation of linearity assumption between the repository variables. This violation of statistical assumption can lead to type-I or type-II error, hence, calling for data transformation for the improvement of the quality of the repository data for subsequent analysis.

Why Sourceforge.net?

Wikipedia (2014) established sourceforge.net as the first free web based source code repository that hosts free and open source software. Among its competitors are Github (www.github.com), Google Code (www.code.google.com) and Javaforge (www.javaforge.com). Alexa (2014) rates sourceforge as 162nd world best website and the first best repository among the aforementioned competitors. These characteristics have called the attention of this research to take a study of the repository for the hundreds of researchers and millions of visitors visiting the website.

2. Methodology

Regression analysis requires the satisfaction of linearity, normality, homoscedasticity and independence. Osborne (2002) established that the violation of the conditions can increase the probability of committing type-I or type-II error. Over decades, data transformation has been recommended as the solution to non-linearity, outliers, among others. Data transformation involves using a mathematical operation to change the measurement scale of variable(s). Linear and nonlinear data transformations are the types of data transformation available. Linear transformation retains the relationship between variables while non-linear transformation changes (increases or decreases) the integrity of the relationship between variables. Tabanick and Fidell (2007) acknowledging the importance of transformation stated that it may not be generally acceptable by authors as a results of difficulty of interpreting the transformed variables but remains a legitimate statistical tool for the realization of linear assumption. This research intends to apply transformation technique to online software repository variables to establish relationship between these variables which was found absent in the original data.

2.1 Linear Regression and Data Transformation

The interpretation of data based on the analysis of variance (ANOVA) is valid if the assumptions of normality, homogeneity, independence and addition assumptions are satisfied. Similarly, regression analysis acknowledges linearity and the first three aforementioned assumptions for implementation. O'Hara and Hotze (2010) emphases that the main purpose of data transformation is to get a sample data to conform with the assumptions of parametric statistics such as ANOVA, t-test and linear regression or to manage outliers in a dataset. Marija (2004) established that data transformation technique is neither a cheating technique nor distortion of the true picture of the data under consideration, rather, it is a legitimate statistical tool. Literatures have established that results interpretation of transformed data analysis is the major challenge facing this statistical technique. However, one added benefit about most transformation technique is that when data are transformed to meet a certain assumption, we often come closer to satisfy another assumptions as well. For instance, square root transformation may help to equate group variances by compressing the upper of the distribution end more than it compresses the lower end. It may also have effect of making a positively skewed distribution more nearly normal in shape. Howel (2007) recommended, as a solution to interpretation problem of data transformation, that researchers should look at both the transformed and original data means and make sure that they are telling the same basic story. Table 1 presents the common ways to transform variables to achieve literatures for regression analysis.

Table 1: The common statistical transformation techniques

SN	Method	Transformed Variable	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 transformation	Dependent variable $(log_{10}Y)$	$log_{10}Y = b_0 + b_1X$	$\hat{Y} = 10^{(b_0+b_1X)}$
03	Quadratic transformation	Dependent variable $(Sqrt(Y))$	$Sqrt(Y) = b_0 + b_1X$	$\hat{Y} = (b_0 + b_1X)^2$
04	Reciprocal transformation	Dependent variable (y^{-1})	$y^{-1} = b_0 + b_1X$	$\hat{Y} = 1/(b_0 + b_1X)$
05	Logarithm transformation	Independent variable $(log_{10}X)$	$Y = b_0 + b_1log_{10}X$	$\hat{Y} = b_0 + b_1log_{10}X$
06	Power transformation	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 transformation	Independent variable (X^2)	$Y = b_0 + b_1X^2$	$\hat{Y} = b_0 + b_1X^2$

In practise, these methods need to be tested on the data to which they are applied for the confirmation that they increase rather than decrease the linearity strength of the relationship. Among the methods to detect the efficiency of the transformed data 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 R^2 of the transformed variables will be higher than the non-transformed variables and that the independent variables will be significant to the response variable. Back transformation is used to return a transformed predicted value to its original scale. Back transformation predicted values give values for the medium response but not the mean response as it is expected. Miller (1984) established that back transformation on the mean of the dependent variable results to serious bias. He, further, established a solution that minimizes the bias. Jia and Rathi (2008), confirming the bias, established a more efficient solution that almost removed the bias. Researchers are advised to consult this aforementioned literature for implementation.

2.2 Mining of data

Visitors to sourceforge have the privilege to find, create, publish, rate and download free and open source software from the repository. Rating of software is done based on two criteria viz. Stars rating and laid down criteria rating. The user can rate 5,4,3,2 or 1 star(s) after which the average stars rating is computed automatically. Computation of the average rating is shown in table 2 below.

Table 2: Average rating computation for filezilla on sourceforge.net as at April 26, 2014.

(a)	***(b)***	***(c=a*b)***
Rating Star	***Number of Raters***	***Total***
5	*843*	*4215*
4	*11*	*44*
3	*6*	*18*
2	*4*	*8*
1	*113*	*113*
Total	***977***	***4398***
Average $\left(\frac{\sum_{i=1}^{5} c_i}{\sum_{i=1}^{5} b_i}\right)$		***4.5***

Similarly, laid-down criteria including design, ease, feature and support are considered in rating software by the user. The repository takes account of total number of people that rate software and download software on it. Perhaps, knowing to the visitors, the repository system collects some information about the users' operating system maker and users' country locations.

Manually collecting data on sourceforge can be a tedious task spanning through days and weeks depending on the volume of the data under concern. This research uses mining software repository (MSR) software named *Okikisoft* which was specially developed for this research purpose. Okikisoft automatically and invisibly visits the pages of sourceforge to extract the required data. It compiles the mined data into CSV files and save it in the application folder. Okikisoft can act as a server-side or client-side mining system.

3. Presentation and Analysis of the Mined Data

The extracted data variables are represented as follows: **D** represents Download total for software, **F** represents Filesize of the software **,** **R** represents Average Rating for the software and **T** represents Total number of visitors that rate the software. Okikisoft mined 1802 software data ranging between February 1 through February 28, 2014. However, a sample size of $n = 50$ was randomly selected for this research. The scatter plots matrix of the original data is presented in figure 1.

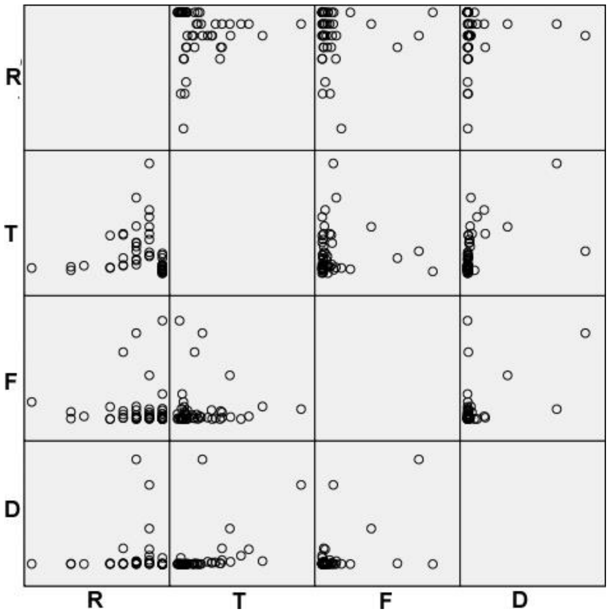

Figure 1: Scatter plot matrix of the original data.

Figure 1 reveals that none of the plots in the scatter plots matrix can be assumed to be linear. Similarly table 3 shows the analysis results done on the original data. Very low coefficient of determination (R^2) and insignificance of the independent variable were experienced between (R and T), (R and F), (R and D) and (T and F) and negative relationship was experienced between (T and D) and (F and D). This result coincides with Ahmed (2008) that the original online repository data may violate fundamental assumptions required by researchers, hence, we recommend that researchers should study the data before further analysis.

Table 3: Analysis results of original data.

Sn.	$\boldsymbol{Var1} \\ \boldsymbol{= y}$	$\boldsymbol{Var2} \\ \boldsymbol{= x}$	**Linear Regression Model**	**+/ - r**	$\boldsymbol{R^2}$	$\boldsymbol{H_0: \beta = 0}$ $\boldsymbol{\alpha = 0.05}$
1	R	T	$R = \alpha + \beta T$	+	0.03%	$P = 0.976$: x *insignificant*
2	R	F	$R = \alpha + \beta F$	+	0.02%	$P = 0.879$: x *insignificant*
3	R	D	$R = \alpha + \beta D$	-	0.02%	$P = 0.880$: x *insignificant*
4	T	F	$T = \alpha + \beta F$	+	0.0%	$P = 0.992$: x *insignificant*
5 (*)	T	D	$T = \alpha + \beta D$	-	23.6%	$P = 0.000$: x *significant (*)*
6 (*)	F	D	$F = \alpha + \beta D$	-	21.4%	$P = 0.001$: x *significant (*)*

3.1 Data transformation and analysis

To prevent values between 0 and 1 in the original data (Osborne (2002)), a linear transformation was done by adding 2 to all data in the four variables. The linearly transformed variables were further nonlinearly transformed with the seven tools viz: $log_2 y$, $log_{10} y$, $\sqrt{y}$, y^2, $\sqrt[3]{y}$, y^3 and y^{-1}. Table 4 shows the results of the data transformation and figure 2 shows the scatter plot matrix for the data.

Table 4: Analysis results of transformed data.

Sn	$Var1 = y$	$Var2 = x$	Linear Regression Model	-/+ r	Back Transformation	R^2	Ranking
1	T	$log_{10}D = D^*$	$T = \alpha + \beta D^*$	++	$\hat{T} = \alpha + \beta D^*$	55.7%	1st
2	$log_{10}D = D^*$	$\sqrt[3]{T} = T^*$	$D^* = \alpha + \beta T^*$	++	$\hat{D} = 10^{(\alpha+\beta T^*)}$	52.2%	2nd
3	$log_{10}T = T^*$	$log_{10}D = D^*$	$T^* = \alpha + \beta D^*$	++	$\hat{T} = 10^{(\alpha+\beta D^*)}$	46.4%	3rd
4	$\sqrt{D} = D^*$	$\sqrt[3]{T} = T^*$	$D^* = \alpha + \beta T^*$	++	$\hat{D} = (\alpha + \beta T^*)^2$	41.6%	4th
5	$log_2T = T^*$	$log_{10}D = D^*$	$T^* = \alpha + \beta D^*$	++	$\hat{T} = 2^{(\alpha+\beta D^*)}$	41.2%	5th
6	$\sqrt{D} = D^*$	$T^2 = T^*$	$D^* = \alpha + \beta T^*$	++	$\hat{D} = (\alpha + \beta T^*)^2$	40.6%	6th
7	T	$\sqrt{D} = D^*$	$T = \alpha + \beta D^*$	++	$\hat{T} = \alpha + \beta D^*$	40.5%	7th
8	$\sqrt{T} = T^*$	$\sqrt{D} = D^*$	$T^* = \alpha + \beta D^*$	++	$\hat{T} = (\alpha + \beta D^*)^2$	35.5%	8th
9	$log_{10}T = T^*$	$D^{-1} = D^*$	$T^* = \alpha + \beta D^*$	- -	$\hat{T} = 10^{(\alpha+\beta D^*)}$	31.1%	9th
10	$T^3 = T^*$	D	$T^* = \alpha + \beta D$	++	$\hat{T} = \sqrt[3]{(\alpha + \beta D)}$	31.0%	10th
11	D	$T^2 = T^*$	$D = \alpha + \beta T^*$	++	$\hat{D} = \alpha + \beta T^*$	29.0%	11th
12	$T^3 = T^*$	$\sqrt[3]{D} = D^*$	$T^* = \alpha + \beta D^*$	++	$\hat{T} = \sqrt[3]{(\alpha + \beta D^*)}$	27.1%	12th
13	$log_{10}D = D^*$	$T^{-1} = T^*$	$D^* = \alpha + \beta T^*$	- -	$\hat{D} = 10^{(\alpha+\beta T^*)}$	17.5%	
14	$T^{-1} = T^*$	$D^{-1} = D^*$	$T^* = \alpha + \beta D^*$	++	$\hat{T} = (\alpha + \beta D^*)^{-1}$	13.9%	

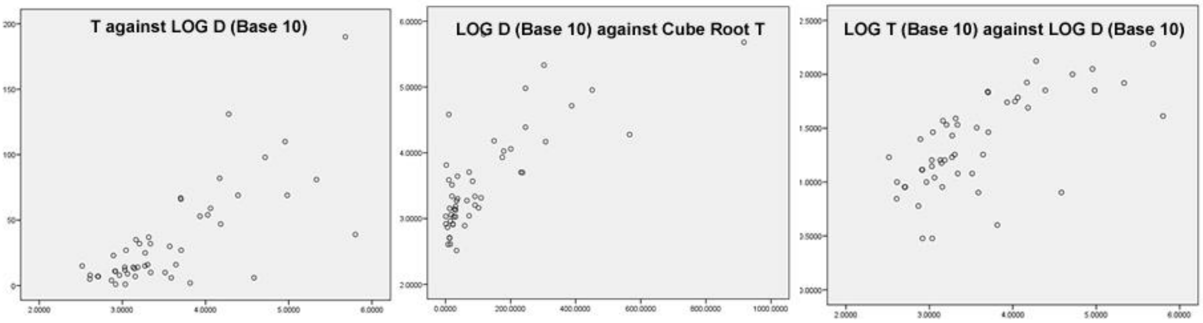

Figure 2: Scatter plot matrix of the first 3 rated models.

4. Discussion of findings

Linearity or almost linearity was ascertained between transformed variables. Table 4 shows R^2 values, linear regression models and the back transformation or interpretation between the respective variables that proved to have linear relationship after transformation was successfully executed. Similarly, twelve (12) models out of the fourteen transformed linear regression models claim to have positive relationship between variables while the independent variables proved to be significant (at 5% significant level) to the corresponding dependent variable. These transformations only discovered linearity between **T** and **D** variables while relationship between other variables failed to claim linearity. We hope attention will be focused on this in future research. Since the result in table 4 are linear and significant, hence, ranking of the result was done using the values of the R^2. Column 7 of table 4 shows the ranking result. The first 12 ranked models proved better with $R^2 > 23.6\%$ which was the highest R^2 value obtained between **T** and **D** in table 3. Fig. 2 shows the scatter plot matrix of the first 3 ranked models. This research only uses y to represent the dependent variable and x for the independent variable in table 4, it is important for researchers to note that these variables can be interchanged but with consequences on the back transformation of the model. To prevent the aforementioned problem of bias (Miller (1984)) from back transformation on the dependent variables, we recommend models 1, 7 and 11 since they do not include the transformation of the dependent variable.

5. Conclusion

In this research, we have used data transformation statistical tools in preparing mined data from online software repository, a case study of sourceforge.net, for researchers for subsequent analysis in improving the efficiency of repositories. It was established that data from online repositories disobey linearity assumptions and may not be significant as required for regression analysis. Hence, we recommend that researchers should study the mined data from online repositories before utilizing them for analysis. Combination of non-linear transformation tools proved to be effective in establishing linearity between repository variables. It was also established that after transformation, only Download total and total number of visitors that rate software on sourceforge can be linear. This research has provided list, in rank, of the first best 12 linear regression models for the direct use of researchers in future analysis.

References

Ahmed, E. Hassan (2008), "A road ahead for mining Software repositories", IEEE. Doi:10.1109/FOSM.2008.4659248, Pp. 48-57.

Alexa (2014) www.alexa.com. Retrieved on April 20, 2014.

Cleveland, W.S. (1984), "Graphical methods for data presentation. Full Scale breaks dot charts multibased logging", The American Statistician, Vol.**38**(4), Pp. 270-280.

Howel, D.C. (2007), "Statistical methods for psychology" Belmont, C.A. Thomson Wadsworth. 6th Edition.

Jia, Siwei and Rathi, Sarika (2008), "On predicting log-transformed linear models with heteroscedaticity", SAS Global Forum, Paper 370-2008.

Manikandan, S. (2010), "Data transformation", J. Pharmacol Pharmocother. Jul.-Dec, Vol.**1**(2), doi:10.4103/0976-500x.72373, Pp126-127.

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.

O'Hara, Robert B. And Hotze, Johan D. (2010), "Do not log-transform Count data", Methods in Ecology and Evolution, Doi:10.1111/j.2041-210x.2010.00021.x.

Osborne, Jason (2002), "Notes on the use of data transformations", Practical Assessment, Research and Evaluation. Vol.**8**(6).

Tabacknick, B.G. and Fidell, L.S. (2007), "Using Multivariate Statistics", 5th Edition, Baston, Allyn and Bacon.

Turke,y J.W. (1977), "Exploratory data analysis", Reading M.A, Addison-Wesley.

Wikipedia (2014), Sourceforge, www.en.wikipedia.org/wiki/sourceforge, Retrieved on April 20, 2014.

www.sourceforge.net, Mining Software (Okikisoft) retrieved on March 13, 2014.