Investigation of the Effects of Box-Cox transformation and Interaction Term on the R^2 And VIF: A Regression Approach

[1]Onu, Obineke Henry, [1]Ejukwa, Justin Odadami and [2] Nwanneako, Sabinus Nnamdi
1: Mathematics & Statistics department, Ignatius Ajuru University of Education, Rumuolumeni, Port Harcourt, Rivers State.
2. Mathematics department, Rivers State University, Oroworukwo, Port Harcourt.
Corresponding email: onuobinekehenry@gmail.com

DOI: 10.56201/ijcsmt.v9.no4.2023.pg69.79

Abstract

The study investigated the effects of Box-Cox Transformation and the introduction of interaction term on the coefficient of determination (R^2) and the variance inflation factor (VIF) using regression approach. The regression equation used was built with employment and insecurity as the predictors while the small and medium scale enterprises (SMEs) was used as the response variable. There were three different analysis known as analysis 1, 2 and 3. Analysis 1 was the regression analysis without interaction and transformation (RAWoIT), 2 was the regression analysis with transformation and without interaction (RAWTWoI) while 3 was the regression analysis with transformation and interaction (RAWTI). The p-values, t-values, analysis of variances and Pareto charts among others were employed in this work. It was found that the Box-Cox transformation introduced in analysis 2 had no visible effects on the VIF and p-values of the employment and the constant term but decrease the $R^2$and adjusted R^2 by 4.34% and 4.91% respectively. In analysis 3, when interaction term and Box-Cox transformation was included in the model, it was observed that the VIF increased from 1.03 recorded for analyses 1 and 2 to 25.06, 613.64 and 595.50 for employment, insecurity and the interaction term, respectively. It was recommended that: to reduce VIF and or multicollinearity in a system, interaction term should not be included and that the Box-Cox transformation reduces the coefficient of determination.

Keywords*: SMEs, Employment, Insecurity, Box-Cox transformation, Variance Inflation Factor and Coefficient of determination.*

1. Introduction

Homoscedasticity is an assumptions of linear statistical model which is stated as, the variance of each disturbance term μ_i conditional on the chosen values of the explanatory variables which is a constant equal to δ^2. i.e $Var(\mu_i) = E[(\mu_i)^2] - E[(\mu_i)]^2 = E[(i)^2] = \delta_u^2$. In most cases, this assumption fails, leading to the problem of heteroscedasticity, which is the direct opposite of homoscedasticity. When this happens the usual ordinary least square regression coefficients

becomes less efficient than some alternative estimators and it causes the standard errors to be biased as seen in Nwakuya & Nwobueze, (2018) and Lyon & Tsai (1996).

Box-Cox transformation according to Henriques-Rodrigues and Gomes (2022) is a method used to make more suitable statistical data for analysis. It is a transformation of dependent variables that are not normally distributed into normally distributed type. Obviously, normality is a very crucial assumption for many statistical techniques, especially the linear regression technique, if the response data is found not to be normal, it is wise to apply a Box-Cox transformation.

This work is aimed at investigating the effects of Box-Cox transformation and the introduction of interaction term on the coefficient of determination and the variance inflation factor using regression approach.

The effects of Box-Cox transformation and the introduction of interaction term on the coefficient of determination and the variance inflation factor have not been clearly stated in the literature. It has been known from the works of Nwakuya & Nwobueze, (2018) and Lyon & Tsai (1996) that if the response variables are not normally distributed, the assumptions of homoscedasticity is always violated, hence leading to heteroscedasticity. To correct this, Box-Cox transformation is carried out on such data to make it be normally distributed. In this work, the authors want to know how Box-Cox transformation and the introduction of interaction term will affect the coefficient of determination and the variance inflation factor, irrespective of whether the data is normal or not. It was as a result that this study considered an ordinary least square analysis for a regression model without transformation and without interaction term to see how the coefficient of determination and variance inflation factor will be and, the analysis of the model with transformation and without interaction is done to also see how the introduction of transformation without interaction will affect the coefficient of determination and variance inflation factor and finally, the analysis of the model where transformation and interaction are introduced to investigate the effect on the same R^2 and VIF is carried out. All these results will be compared and recommendations will be given according to the findings.

2. Materials and Methods

The study will first build a linear regression of the form

$$SMEs = \beta_0 + \beta_1 Insec + \beta_2 Empoly + \mathcal{E} \quad (1)$$

where,

SMEs=Small and Medium Scale emterprises, Insec=Insecurity level, Employ=Employment rate in Nigeria. The $\beta_0$is the grand mean or the value of the SMEs in Nigeria, when Insecurity level and Employment rate are zero (0). β_1 is used as a measure of the Insecurity level, it is also known as the gradient or slope of the Insecurity level, while β_2 measures the rate of employment generation in Nigeria, (Victor-Edema & Onu, 2023). It can also be called the gradient or slope of the Employment generation in Nigeria. In the model in equation (1), the ordinary least square estimation method will be applied, to obtain the parameters as seen in the equation

$$\underline{\beta} = (X'X)^{-1}X'SMEs \quad (2)$$

where X is the design matrix obtained from the data of Insecurity and Employment, put in matrix form. Hence X is an $M \times N$ matrix of the predictors. Equation (2) is known as the least square equation applied when the response data (SMEs) obeys homoscedacisity, where $\underline{\beta}=\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\end{pmatrix}$. This application of least square will be carried out on the data without transformation and without interaction, thereafter, the Box-Cox transformation will be introduced, which is to say generalized linear model will be applied on the data with interaction term included.

In Box-Cox transformation, an exponent λ which varies in the range, $-5 \leq \lambda \leq 5$ are considered while the optimal value for the applied data is chosen. This optimal value of the response data is one which results in the best approximation of a normal distribution curve. The transformation Y comes in the form as proposed by Box and Cox, (1964)

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda} \\ \log y\ ,\ if\ \lambda = 0 \end{cases}, \text{if } \lambda \neq 0 \qquad (3)$$

The equation in (3) as seen in Atkinson (2020) can work for only positive data which is the basis of this work. The data are all positive.

Rather, for negative data,

$$y(\lambda) = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1-1}}{\lambda} \\ \log(y+\lambda_2)\ ,\ if\ \lambda_1 = 0 \end{cases}, \text{if } \lambda_1 \neq 0 \qquad (4)$$

Yeo and Johnson, (2000), extended the idea of Box and Cox as expressed in (4) to observations that are mixture of negative and positive. It is given as

$$y \geq 0: = \frac{(y+1)^{\lambda}-1}{\lambda_{\dot{y}Y_J}^{\lambda-1}},\ (\lambda \neq 0);\ \dot{y}Y_j \log(y+1)\ (\lambda = 0) \qquad (5)$$

$$y < 0: -\frac{\{(-y+1)^{2-\lambda}-1}{(2-\lambda)\dot{y}_{YJ}{}^{\lambda-1}}\ (\lambda \neq 2);\ -\log(-y+1)\ /\dot{y}_{YJ}\ (\lambda = 2). \qquad (6)$$

Minitab software was used for this computation, since manual computation will be tedious.

3. Results and Discussion

Analysis 1

Regression Analysis: SMEs versus INSECURITY, EMPLOYMENT WITHOUT TRANSFORMATION AND INTERACTION

Regression Equation

SMEs = -1459962 - 3.0 INSECURITY + 441151 EMPLOYMENT

Table 1: Regression Coefficients with VIF for RAWoIT

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-1459962	282872	-5.16	0.000	
INSECURITY	-3.0	12.3	-0.24	0.810	1.03
EMPLOYMENT	441151	44962	9.81	0.000	1.03

Table 2: Model Summary for RAWoIT

S	R-sq	R-sq(adj)
433896	86.98%	85.24%

Table 3: Analysis of Variance for RAWoIT

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	1.88611E+13	9.43055E+12	50.09	0.000
INSECURITY	1	11255670449	11255670449	0.06	0.810
EMPLOYMENT	1	1.81241E+13	1.81241E+13	96.27	0.000
Error	15	2.82399E+12	1.88266E+11		
Total	17	2.16851E+13			

Analysis 2

Regression Analysis: SMEs versus EMPLOYMENT, INSECURITY WITH TRANSFORMATION AND WITHOUT INTERACTION

Method

Box-Cox transformation	
Rounded λ	0
Estimated λ	0.146067
95% CI for λ	(-0.282433, 0.513567)

Note that the Confidence interval values for λ lie within -5 and +5.

Regression Equation

ln(SMEs) = 10.192 - 0.000008 INSECURITY+ 0.5082 EMPLOYMENT

Table 4: Regression Coefficients with VIF for RAWTWoI

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	10.192	0.388	26.23	0.000	

EMPLOYMENT	0.5082	0.0617	8.23	0.000	1.03
INSECURITY	-0.000008	0.000017	-0.46	0.651	1.03

Table 5: Model Summary for RAWTWoI

S	R-sq	R-sq(adj)
0.595893	82.64%	80.33%

Table 6: Analysis of Variance for Transformed Response

Source	DF	Seq SS	Seq MS	F-Value	P-Value
Regression	2	25.3564	12.6782	35.70	0.000
EMPLOYMENT	1	25.2808	25.2808	71.20	0.000
INSECURITY	1	0.0756	0.0756	0.21	0.651
Error	15	5.3263	0.3551		
Total	17	30.6827			

Analysis 3

Regression Analysis: SMEs versus EMPLOYMENT, INSECURITY WITH TRANSFORMATION AND WITH INTERACTION

Method

Box-Cox transformation	
Rounded λ	0
Estimated λ	0.22518
95% CI for λ	(-0.226320, 0.618680)

Note that the Confidence interval values for λ lie within -5 and +5.

Regression Equation

ln(SMEs) = 9.48+ 0.000253 INSECURITY+ 0.698 EMPLOYMENT
- 0.000069 EMPLOYMENT*INSECURITY

Table 7: Regression Coefficients with VIF for RAWTI

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	9.48	1.21	7.83	0.000	
EMPLOYMENT	0.698	0.311	2.24	0.042	25.06

INSECURITY	0.000253	0.000420	0.60	0.556	613.64
EMPLOYMENT*INSECURITY	-0.000069	0.000110	-0.62	0.544	595.50

Table 8: Model Summary for RAWTI

S	R-sq	R-sq(adj)
0.608451	83.11%	79.49%

Table 9: Analysis of Variance for RAWTI

Source	DF	Seq SS	Seq MS	F-Value	P-Value
Regression	3	25.4997	8.4999	22.96	0.000
EMPLOYMENT	1	25.2808	25.2808	68.29	0.000
INSECURITY	1	0.0756	0.0756	0.20	0.658
EMPLOYMENT*INSECURITY	1	0.1434	0.1434	0.39	0.544
Error	14	5.1830	0.3702		
Total	17	30.6827			

Discussion of Results

Analysis 1: Regression analysis without interaction and transformation (RAWoIT)

The study reveals that the Variance Inflation Factor (VIF) was found to be 1.03 for both the insecurity and employment while the coefficient of determination and the adjusted were found to be 86.98% and 85.24% respectively and the effect of employment on Small and Medium scale Enterprises (SMEs) was seen to be significant with p-value of 0.000 in the analysis, while the insecurity was not significant with p-value of 0.810 at 5% level of significant. The analysis of variance shows that employment has significant contribution on SMEs while insecurity has no significant contribution on SMEs. It was further shown that employment has positive contribution while insecurity has negative contribution on the SMEs. The Pareto chart shows that Employment has the highest contribution.

Analysis 2: Regression analysis with transformation and without interaction (RAWTWoI)

It was revealed that when Box-Cox transformation was applied, the p-values for the constant term and the employment term remained unchanged (unaffected), likewise the variance inflation factor from the result of analysis 1, while the T-value for the constant term increased from -4.56 to 26.23 and the T-value for insecurity changed from -0.24 to -0.46, where that of employment changed from 9.89 to 8.23. The coefficient of determination and the adjusted changed from 86.98% and 85.24% to 82.64% and 80.33% respectively, showing a 4.34% and 4.91% decrease in $R^2$and adjusted R^2 respectively. This result was different from what was obtained by Nwakuya and Nwabueze 2018), where they reported an increase in R^2 and adjusted R^2after Box-Cox transformation. The reason for this discrepancy may be linked to mere inflation of values of $R^2$and

its adjusted by the ordinary least square method, but, their point of agreement was that, the regression model was significant. Also, the positive effect of employment reduced drastically to 0.5082 ton from 44115 ton and that of insecurity reduced to -0.0000008 from -3.0. The Pareto chart shows that Employment has the highest contribution.

Analysis 3: Regression analysis with transformation and interaction (RAWTI)

It was observed that when Box-Cox transformation and interaction term were included in the analysis, the p-value of the constant was not affected, while, the p-value of employment changed from 0.000 to 0.0042, while that of insecurity changed from 0.651 to 0.658 while the T-value of the constant changed from 26.23 to 7.83 and that of employment changed from 8.23to 2.24 and insecurity changed from -0.46 to 0.60. The coefficient of determination and the adjusted were found to be 83.11% and 79.49% respectively. The initially unchanged VIF of 1.03 in the first and second analyses changed to 25.06 for employment, 613.64 for insecurity and 595.50 for interaction term. The Pareto chart shows that Employment has the highest contribution followed by the interaction factor.

Conclusion

The research studied the effect of transforming the response variable using the Box-Cox method on the coefficient of determination and its adjusted and the variance inflation factor. It also investigated the combined inclusion of Box-Cox transformation and the interaction term on the coefficient of determination and its adjusted and the variance inflation factor. The study was done in three phases, where phase 1 was the analysis of a linear regression with employment and insecurity as predictors and SMEs as response, the model has no transformation and no interaction term. In phase 2 transformation was introduced to the model of phase1, while in phase 3, both transformation and interaction were introduced. The p-values, T-values, sum of square and Pareto chart were employed. For analysis one, variance inflation factor remains constant for all the three variables studied and the effect of employment on on SMEs in Nigeria was found to be significant with a p-value of 0.000 while insecurity was not significant. In analysis two, when Box-Cox transformation was applied on the SMEs data, without interaction term in the model, the p- values we're not affected, which means that, transformation does not affect the variance inflation factor and p- values of the employment term and that of the constant term. The coefficient of determination and it's adjusted decreased by 4.34 and 4.19% respectively. In analysis three, when both Box-Cox transformation and interaction term we're carried out together in the model, **it was also the p- value of the constant was not affected but that of employment change from 0.000 to 0.0042. The coefficient of determination increased slightly from what was obtained in analysis two to 83.11 percent and it's adjusted decreased to 79.49 percent.**

Recommendations

The study recommends the following to Statistics practitioners, Captains of Industries, Government at all levels, that:

1. To reduce the relationship among the predictors measured by the VIF, model without interaction should be built. This is because the study found that interaction increases VIF indiscriminately.
2. To obtain higher values of coefficient of determination and it's adjusted, transformation should not be considered in the model.

Contribution to Knowledge

The study contributes the following to knowledgebase:

1. Box-Cox Transformation has no effect on the VIF in a model without interaction.
2. Interaction term increases the VIF in among the predictors.
3. Box-Cox Transformation decreases the coefficient of determination and its adjusted.

References

Atkinson, A. B. (2000). The Box-Cox transformation: review and extensions. Statistical Science. ISSN 0883-4237 (In Press).

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). Journal of the Royal Statistical Society, Series B, 26, 211=252.

Henriques-Rodrigues, L. and Gomes, M. I. (2022). Box-Cox Transformations and Bias Reduction in Extreme Value Theory. *Computational and Mathematical Methods, 2022*, 1-15.

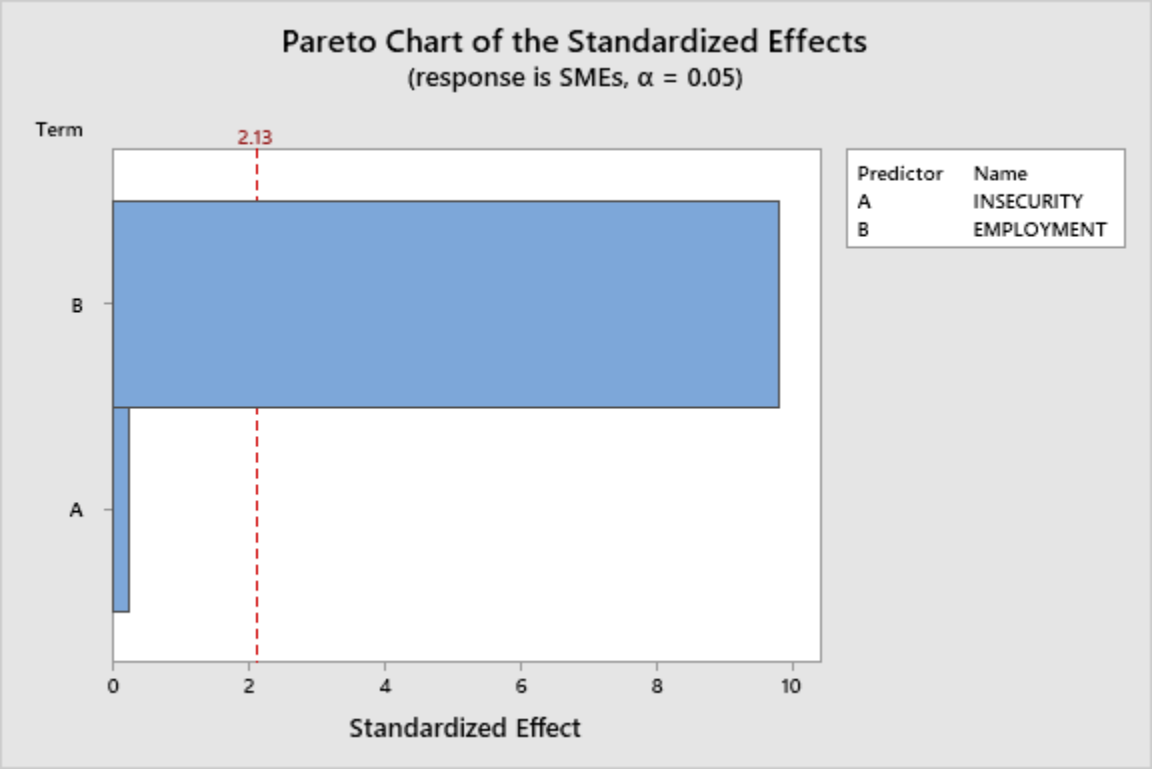
Lyon & Tsai (1996): A Comparison of Tests for Heteroscedasticity; *The Statistician, 45, 3*, 337-349.

Nwakuya, M.T. & Nwabueze, J. (2018). Application of Box-Cox Transformation as a Corrective Measure to Heteroscedasticity Using an Economic Data. *American Journal of Mathematics and Statistics, 8*(1): 8-12.

Victor-Edema, U. A. and Onu, O. H. (2023). Small and Medium scale Enterprises as Panacea to Employment Generation, Insecurity and Educational Development among Nigerian Youths. *Faculty of Natural and Applied Sciences, Journal of Scientific Innovation.* Conference Paper.

Yeo, I. K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. Biometrika, 87, 954-959.

Appendix A: **Figure 1: Pareto Chart for RAWoIT**



Appendix B: **Figure 2: Pareto Chart for RAWTWoI**

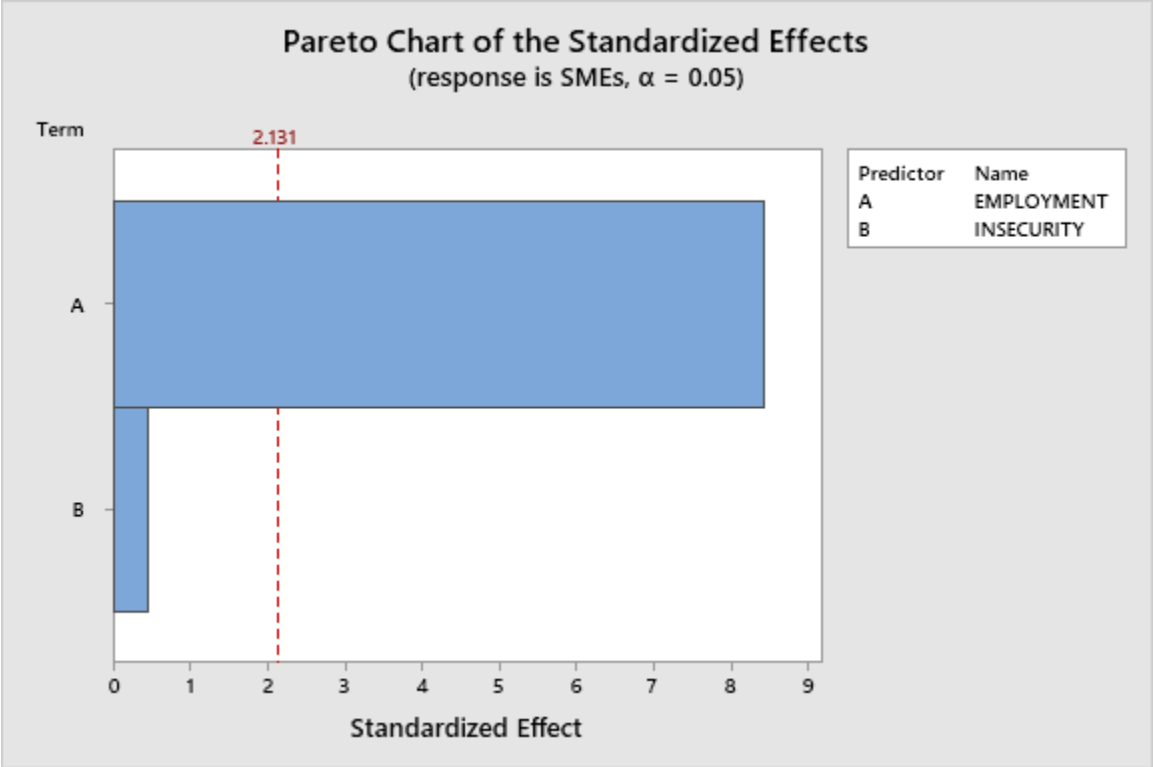


Figure 2: Pareto Chart for RAWTWoI

Appendix C: **Figure 3: Pareto Chart for RAWTI**

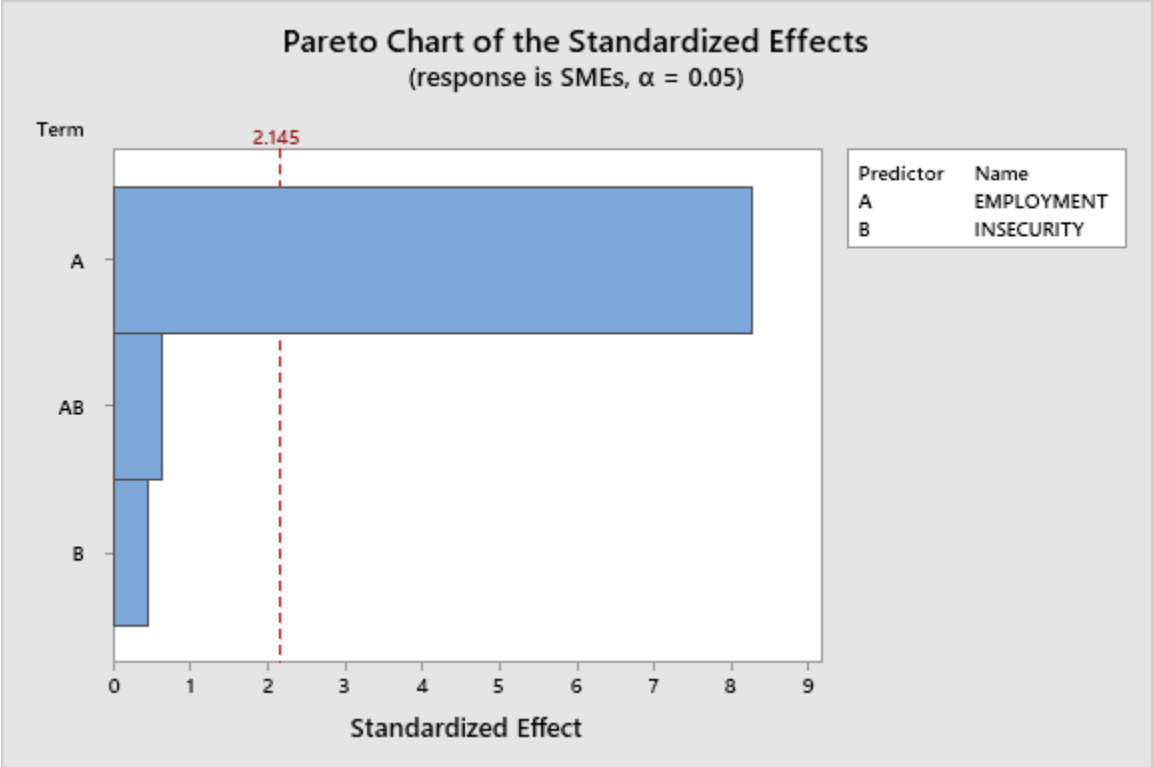


Figure 3: Pareto Chart for RAWTI