Detection of Collinearity Effects on Explanatory Variables and Error Terms in Multiple Regressions

Alhassan Umar Ahmad, U.V. Balakrishnan, Prem Shankar Jha

Abstract—In this work we investigates the effects and consequences of multicollinearity on both standard error and explanatory variables in multiple regression, the correlation between X1 to X6 (independent variables) measure their individual effect and performance on Y (Response variable) and it is carefully observes how those explanatory variables intercorrelated with one another and to the response variable. There are many procedures available in literature for detecting presence, degree and severity of multicollinearity in multiple regression analysis here we used correlation analysis to discover it is presence; we use variance inflation factors, tolerance level, indices number, eigenvalues to access fluctuation and influence of multicollinearity present in the model. Multicollinearity was discovered in this research work with a severe proportion using arrays of correlation analysis procedure which affects the performance of the explanatory variables present in the model by making it less independent and more redundant as it should not be. Collinearity inflates variance of estimates and brings change in direction and signs of the co-efficient of the estimates leading to unrealistic erroneous inference, wrong interpretation and instability among the predictor variables. standard error is discover to be slightly height which directly affect the accuracy and precision of the final result from the analysis, it bring type 1 error during and after the hypothesis testing and finally undermined the overall inference of the entire analysis interest and is in good agreements with the finding of Complete elimination of collinearity is not possible but in this work we reduce it is degree of intensity to enhance the performance of independent variables and error term in the model.

Index Terms—Multicollinearity, predictor variable, standard error and multiple regression

1. INTRODUCTION

Multicollinearity is an area that recently getting attention from researchers gradually as a results of the development and level of advancement from many recent statistical software which is always simplifying the analysis of multiple and complicated data to more easier, precise and accurate level than before. recently scholars are trying to minimize the errors in their research which exactly affect the accuracy and precession of their final work and multicollinearity is playing a vital role on this problem from use of simple linear regression and multiple regression model, it also have many unfavorable effects on the estimated coefficients in a multiple regression model despite it is very essential that researchers are trains in such a way that they can always check multicollinearity to avoid it is deficiency in the analysis more especially by examining the latent roots and latent vectors, tolerance level, correlation matrix and variance inflation factors are all use to access multicollinearity in a given date to insure the independency of the explanatory variable before analysis Edward R. Mansfield and Billy P. Helms (1982). Investigation has it that serious increase in multicollinearity level among the explanatory variable is making it unrealistically erroneous and highly vulnerable and bring a serious consequences to the coefficient of the parameter estimate during analysis Graham (2003).

Multicollinearity as a statistical phenomenon where by it exist when two or more of the explanatory variables in regression model are moderately, severely or highly correlated with one another. Where by a predictor variable is linearly predicted from or by one of the explanatory variables present in the same regression model with a nontrivial degree of accuracy which directly alter it is independent nature and seriously affect the coefficient of estimates in multiple regression negatively. Small difference from the data base source due to poor experimental designed produce a serious multicollinearity issue Judge (1988). It has been carefully observed that multicollinearity does not affect the reliability or productive power of multiple regression models at all. But now discover to be seriously disturbing if is to acknowledging the individual contributions of performances of the independent variables present in the model to the response variable. Correlation analysis always here explain which explanatory variable is redundant than others and which predictor variable is responsible for such redundancy. If one of the predictor variables is directly a linear function of another explanatory variable as result of a higher correlation in between two variables then collinear effect is a problem in the model D. Stephen Voss (2004). Currently researchers are making good effort to specify and classify the methods of identification of collinear effects base on the type of data source and nature of multicollinearity expected be it data base or as a result of mathematical artifact against the most well-known general method of testing both type by variance inflation factor and Tolerance level index number most of the time Johnston (1972) and Kroll et al. (2004). Multicollinearity hinders some computer software from

Revised Version Manuscript Received on April 12, 2019.

Alhassan Umar Ahmad, Department of Mathematics, School of Basic Science and Research Sharda University, Greater Noida, U.P, Delhi, India.(E-mail: umarcomrade855@gmail.com)

U.V Balakarishnan, Department of Mathematics, School of Basic Science and Research Sharda University, Greater Noida, U.P, Delhi, India

Prem Shankar Jha, Department of Mathematics, School of Basic Science and Research Sharda University, Greater Noida, U.P, Delhi, India

performing effectively more especially if the task to be done has to do with purely independent variables. Indication of multicollinearity is showing lack of pure independency among the predictor variables due to the little difference in the data which bring a wider range among parameter estimates, slightly higher value for the error term coefficients, law significant levels values, higher R^2 and typically wrong sign of coefficients (Greene 2000).

1.1 Why Is Multicollinearity a Problem

Prediction of response variable from predictor variable in multiple regression models has very little or no much trouble with the term multicollinearity. This is because it always predict with higher accuracy and precision due to the effect of linearity on classical linear regression model and the overall values of R^2 and adjusted value of R^2 is always quantify how well the response variable is been estimated. If the goal of the analysis is to estimate how each individual explanatory variable affect the response variable in the model then multicollinearity is a serious issue because a particular p-value will tend to be higher than necessary and a very wider confidence interval which may include zero. This will give no confident on some statistics values after the analysis because weather an increases in the explanatory variables will produce the necessary increases or decreases of the respond variable Ranjit kumar paul (2008). It has been proved that the collinear effect that result in multicollinearity is more severe on a small sample size than larger sample size data statistically this is because smaller sample always produce higher correlation and a large standard error Saman Babaie-Kafaki & Mahdi Roozbeh (2017).

1.2 Consequences of Data Base Multicollinearity

Collinearity can give wrong estimators for the regression coefficients, inflates both the error term and partial t-tests produce inaccurate inference and non-significant p-value at all and finally alter the aim of the analysis Shalabh (2016). Collinearity increases parameter variance estimates unnecessarily and undermine the aims of the analysis Greene (1990). Investigation on both predictor variables and explanatory variables is all for the aims to excess the collinearity in between all the predictor variables and detects which one has the most alarming collinear relationship that must be reduce to increase the productivity and enhance independency among the predictor variables. Therefore higher correlation in between predictor variables is finally making the predictor variables to be not properly independent due to presence of multicollinearity Johnston (1984).

2. METHOD AND MATERIAL

In our effort to makes investigation on multicollinearity we obtained and study the data of about 200 samples to excess multicollinearity presence and degree of intensity in data base multicollinearity to be able to upper practical solution to the collinearity problem.

Therefore from regression equation we have;

$Y= \beta_0 +\beta_1X_{1i}+ \beta_2X_{2i}+ \beta_3X_{3i}+ \dots + B_kx_{ki}+ \varepsilon_i$ …… (1)

Where β_0 = intercept, β_1 to β_k is the partial slope of the coefficients

ε_i= error term

i=i^{th} is the observation n. being the size of our sample from a population

Let our response variable be Y

Let our Independent variables be X1, X_2, X_3, X_4, X_5, X_6 respectively.

Let the Standard error be ε_i

From our effort to investigate multicollinearity effects on both standard error and predictor variables

A model is now specify as

$Y=f(X_1, X_2, X_3, X_4, X_5, X_6)+ \varepsilon_i\beta_1$ ……..(2)

Re-writing (2) in more of the explicit form it will now exactly be

$Y= \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3 X_3+ \beta_4 X_4+ \beta_5 X_5 + \beta_6X_{6 +} \varepsilon_i$. (3)

Now we have

$Y= \beta_0 +\beta_1X_{1i}+ \beta_2X_{2i}+ \beta_3X_{3i}+ \beta_4X_{4i}+ \beta_5X_{5i}+ \beta_6X_{6i} \dots + B_kx_{ki}+ \varepsilon_i$ ……..(4)

From the relationship above Equation (4) is identifying k_i predictors variables from multiple regression models obtain from the explanatory variable such as X_1….X_6 and constant terms that always assumed to influence the respond variable regressed. The linear relationship from equation (1) hold for all hypertensive patients only if we could have a reasonable value of the explanatory variables and respond variable from the standard error which is the error term due to disturbances in multiple regression model, This is done by fitting a regression line to the observed sample data as an approximation to the true line. If then the true relationship between X_1 to X_6 and Y is as given in Equation (3). The raw data exhibits a strong collinear relationship those covariates and hence requires some treatment to reduce the effect from collinearity. We then implement both the ridge method, with two different choices of shrinkage parameters X_1 to X_6 and the perturbation method on the data to obtain alternative estimators Blaze, T. J., and Ye, F. (2012).

2.1 Covariance Method

British scientist who happen to be a good biometrician as well a good statistician by profession developed what we called Karl Pearson's co-efficient of correlation which is one of the recognized way of checking the interrelationship between two variables X and Y today is been used widely in science, social science and management. it is usually been obtained by r(X,Y) or r_{xy} or it can simply be written as r. and is defined as the numerical measure of the linear relationship between two or more variables and can also be in a given relationship as the ration of the covariance between X and Y which is written by Cov(X,Y)to the product of the standard deviation of X and Y. it can be written as ;

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad \dots\dots (5)$$

That means if (x_1,y_1), $(x_{2,} y_2)$, $(x_{3,}y_3)$,…, (x_n ,y_n) are the n-pairs of observation of the variable X and Y in a Bivariate Distribution so it will be

$$Cov(XY) = \frac{1}{2}\sum_{i=1}^{n}(X-\bar{X})(Y-\bar{Y}) \quad \dots\dots (6)$$

And

$$\sigma_X = \sqrt{\frac{1}{n}\sum(X-\bar{X})^2}$$

Therefore similarly

$$\sigma_Y = \sqrt{\frac{1}{n}\sum (Y-\bar{Y})^2} \quad \text{..... (7)}$$

By taking the summation over n pairs of observation and subtracting equation (2) to (4) in to equation (1) we have

$$\frac{\frac{1}{n}\sum (X-\bar{X})(Y-\bar{Y})}{\left(\sqrt{\frac{1}{n}\sum (X-\bar{X})^2}\right)\left(\sqrt{\frac{1}{n}\sum (Y-\bar{Y})^2}\right)} \quad \text{..... (8)}$$

Therefore

$$\frac{\frac{1}{n}\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\frac{1}{n}\sum (X-\bar{X})^2 \frac{1}{n}\sum (Y-\bar{Y})^2}} \quad \text{..... (9)}$$

$$\frac{\frac{1}{n}\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\frac{1}{n}\{\sum (X-\bar{X})^2 * \sum (Y-\bar{Y})^2\}}}$$

$$\frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 * \sum (Y-\bar{Y})^2}} \quad \text{..... (10)}$$

Therefore

$$r = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 * \sum (Y-\bar{Y})^2}} \quad \text{..... (12)}$$

Than equation (9) can be written as

$$r = \frac{\sum dx.dy}{\sqrt{\sum dx^2 \sum dy^2}} \quad \text{..... (13)}$$

Whereas dx and dy donate the actual deviation of variable x and y, the value from their arithmetic means $\bar{X}$ and $\bar{Y}$ respectively such that;

dx= X- $\bar{X}$, dY= Y- $\bar{Y}$

Now by simplifying equation (2) which is the covariance of X and Y. Cov(x,y).

$$\text{Cov (X, Y)} = \frac{1}{n}\sum (X-\bar{X})(Y-\bar{Y}) \quad \text{..... (14)}$$

$$\frac{1}{n}\sum \left(XY - X\bar{Y} - \bar{X}Y + \bar{X}\bar{Y}\right) \quad \text{..... (15)}$$

$$\frac{1}{n}\sum XY - \bar{X}.\frac{1}{n}\sum Y - \bar{Y}\frac{1}{n}\sum X + \frac{1}{n}.n\bar{X}\bar{Y} \quad \text{..... (16)}$$

Since $\bar{X}$ and $\bar{Y}$ are constant term here and from

$\sum CX = C\sum X$ And $\sum C = nC$

Where, and a. is a constant

Therefore from (14)

$$\frac{1}{n}\sum XY - \bar{X}.\frac{1}{n}\sum Y - \bar{Y}\frac{1}{n}\sum X + \frac{1}{n}.n\bar{X}\bar{Y} \quad \text{..... (17)}$$

Than we have

$$\frac{1}{n}\sum XY - \bar{X}\bar{Y} \quad \text{..... (20)}$$

Therefore covariance of X and Y is now given by

$$\text{Cov (x,y)} = \frac{1}{n}\sum XY - \bar{X}\bar{Y}$$

$$\frac{1}{n}\sum XY - \frac{1}{n}\sum X.\frac{1}{n}\sum Y$$

$$\frac{1}{n}\sum XY\left(\frac{\sum X}{n}\right)\left(\frac{\sum Y}{n}\right) \quad \text{..... (21)}$$

$$\frac{1}{n}\sum XY\left(\frac{\left(\frac{\sum X}{n}\right)\left(\frac{\sum Y}{n}\right)}{n^2}\right)$$

$$\frac{1}{n^2}\left[n\sum XY - \left(\sum X\right)\left(\sum Y\right)\right] \quad \text{..... (22)}$$

But

$$\sigma_X^2 = \frac{1}{n}\sum (X_i-\bar{X})^2 = \frac{1}{n}\sum X^2 - \bar{X}^2$$

$$\frac{1}{n}\sum X^2 - \left(\frac{\sum X}{n}\right)^2 = \frac{1}{n^2}\left[n\sum X^2 - \left(\sum X\right)^2\right] \quad \text{..... (23)}$$

Therefore substituting equation (22), (23) and (20) in to (5) we have

$$r = \frac{Cov(XY)}{\sigma_X \sigma_Y}$$

$$r = \frac{\frac{1}{n^2}\left[n\sum XY - \left(\sum X\right)\left(\sum Y\right)\right]}{\sqrt{\frac{1}{n^2}\left[n\sum X^2 - \left(\sum X\right)^2\right]}.\sqrt{\frac{1}{n^2}\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}} \quad \text{..... (24)}$$

$$r = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y - \left(\sum Y\right)^2\right]}} \quad \text{..... (25)}$$

This is what we called Karl-peason's correlation co-efficient is also known as the productive moment correlation co-efficient which we used I finding the relationships between explanatory variable and response variable to discover the exact relationship.

2.2 Tolerance level

Tolerance as an important factor and measure of multicollinearity present in which all researcher expect and work towards having higher tolerance level statistically to avoid linear relation in between independent variables. This is because law tolerance level affect the final result of the research and undermine it is final result. Tolerance level is obtained by calculating auxiliary individual value of R^2-value against other explanatory variables in multiple regression models during analysis against other predictor variables in the model and subtracts R^2- value from 1. (1- R^2) this will give a tolerance level of an individual variable with a range normally from 0 to 1 it is well known that a tolerance value of 0.50 and above is for no serious concern because it indicate law multicollinearity and a value around 0.20 indicate a serious and higher multicollinearity. Tolerance level is also the same as the reciprocal of the variance inflation factor.

2.3 Variance Inflation Factor

Variance inflation factor indicate how far the variances of estimation for the regression confidents are inflated compared to when the explanatory variables are not orthogonal Neter, wasserman and Kutner (1989).Variance inflation factor is a phenomenon for measure the amount of multicollinearity in a given multiple regression model, A multiple regression model is used to test the effect of multiple variables on a particular outcome. The dependent variable is the outcome that is being acted upon by the independent variables, which are the inputs into the model. Multicollinearity exists when there is a linear relationship, or reasonable correlation between some independent variables or inputs from the predictor variables. Multicollinearity creates a problem in the multiple regressions because since the inputs are all influencing each other, they are not actually independent and it is difficult to test how much the combination of the independent variables affects the dependent variable, or outcome, within the regression model. Variance inflation factor indicate how is the variance and standard error of the coefficient estimate in collinearity is being inflate, the variance inflation factor for the estimated coefficient b_k – donate VIF_k (Variance inflation factor) is a factor by which the variance is been change;-

Let us consider estimation in which any X_k is a prediction such that

$Y_i = \beta_0 + \beta_k X_{ik} + \varepsilon_{ij}$ ……..….(26)

Therefore the variance of estimated coefficient b_k is given by;-

$$\text{var}(b_k) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_{ik} - X_k)^2} \quad \text{…….. (27)}$$

Note that we put (min) in order to indicate it is the smallest value that variance can be and based on keeping the track of this base line variance it can show how much the variance of b_k is been changed if we add correlated predictor to our multiple regression model, now we shall consider such model with many predictors.

$$Y_i = \beta_0 + \beta_i X_i +,....,+\beta_k X_K +,....,+\beta_{P-1} X_{I-P} + \xi_I. \quad \text{…… (28)}$$

Now if Predictors are collected with some predictor variables X_k than the variance b_k is changed, therefore it will be shows as the variance of b_k below;-

When R^2_k is the R^2-value which is obtained by taking regression of K^{th} predictor on the remaining predictors Note that the greater the linear dependency among the predictors X_k and the other predictors, the larger the R^2_k value and as the above formula suggested then the larger the R^2_k value the larger the variance of b_k .

To express this more clearly we take the ratio

$$\frac{Var(b_k)}{Var(b_k)_{\min}} = \frac{\frac{\sigma^2}{\sum_{i=1}^{n}(X_{ik} - \bar{X}_k)^2} * \frac{1}{1 - R_K^2}}{\frac{\sigma^2}{\sum_{i=1}^{n}(X_{ik} - X_k)^2}} \quad \text{……..… (29)}$$

$$= \frac{1}{1 - R_K^2} \quad \text{…..…… (30)}$$

$$VIF_k = \frac{1}{1 - R_K^2} \quad \text{……. (31)}$$

Above is what we called variance inflation factor for the K^{th} predictors that is

Whereas R^2_k is the R^2 – value obtained by regressing the K^{th} predictor on the remaining predictors note that the variance inflation factor exist for each of the K predictor predictors in multiple regression model.

The variance inflation factor shows how the estimate variance of K^{th} regression coefficient is changed above what it should be more spatially if the value of the coefficient of estimation R^2 is totally equal to zero, from the other hand in a situation where K^{th} independent variable is perfectly has no any correlation in between each other than it will achieved orthogonally to the other independent variable in the analysis, therefore VIF will definitely provide a reasonable and serious indication of the effect of collinearity on some sample fluctuations of the parameters Robert M. O'brien (2007).

2.4 Farrar-Glauber Test

The procedures for detecting multicollinearity such as t-test which is a type of inferential *statistic*s used to determine if there is a significant difference between the means of two or more groups of the predictor variables, which may be related in certain features. Chi Square statistic is commonly used for testing relationships between categorical variables. The null hypothesis of the Chi-Square test is that no relationship exists on the categorical variables only if they are independent "F Test" is a catch-all term for any test that uses the F-distribution. In most cases when people talk about the F-Test, what they are actually talking about is The F-Test to Compare Two Variances. However the *f-statistic* is used in a variety of tests including *regression analysis*. Adeboye N.O, Fagoyinbo I. S and Olatayo T.O (2014).

3. RESULTS AND DISCUSSION

A sample of about 200 data was collected to investigate and test the present of multicollinearity and provide a general solution on how to reduce the negative effect of multicollinearity on both explanatory variables and error term from multiple regression analysis models. this is making an independent variables present in multiple regression model to become not fully or moderately independent due to presence of collinear effects in between predictor variables during analysis.

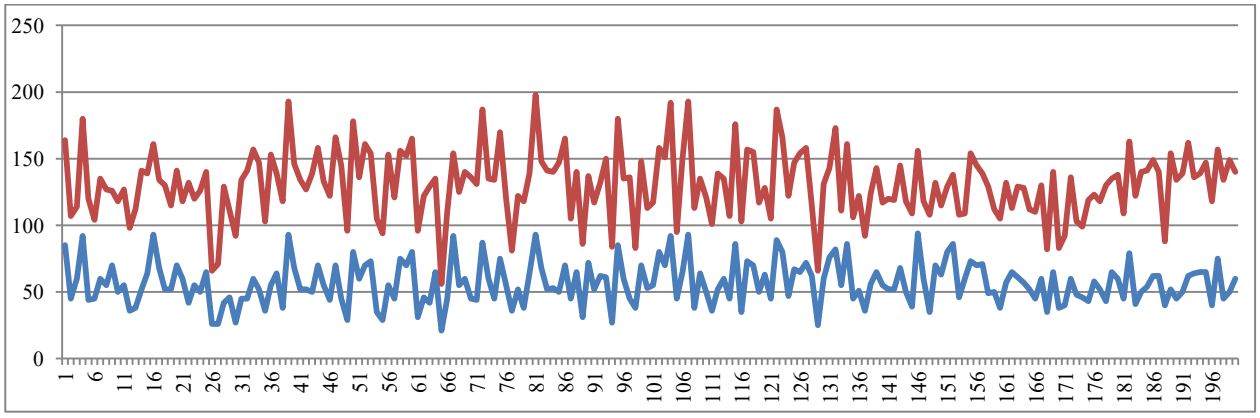

Fig. 1 Graphical Illustration Of X_1 (in Blue) and X_2 (in Red) Variables

From the graphical illustration above figure 1 expresses x_1 and x_2 visually as the two variables with higher correlation which indicates relationship and shows the same pattern for both x_1 and x_2 which means an element of similarity or a copy from one independent variable to another exist after the first analysis.

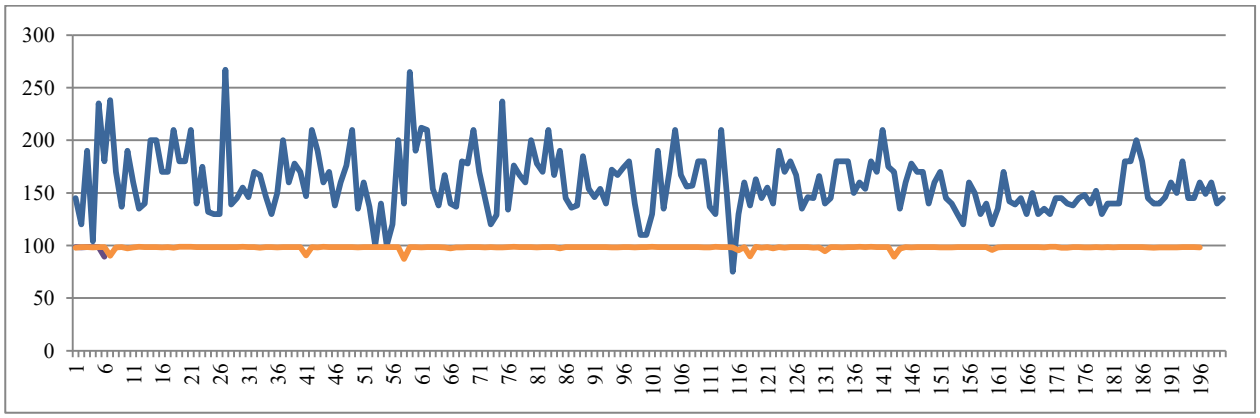

Fig. 2 Graphical Illustration Of X_3 (in Yellow) and X_5 (in Blue) Variables

From the graphical illustration above figure 1 where it expresses x_1 and x_2 visually as the two variables with higher relationship and shows the same pattern for both x_1 and x_2 which means an element of similarity or a copy from one independent variable to another that is to say some similarities exist in between which indicate little or no much independency among the explanatory variables in the model that is to say multicollinearity is present but after the treatment to reduce it to minimum level now we are successfully been able to reduced multicollinearity after the second analysis, from figure 2 both x_3 and x_5 with higher relationship now visually shows different pattern which indicate individual independency among the explanatory variables by showing different behavior on the graphical illustration.

A Correlation analysis was conducted and a severe correlation in between X_1 and X_2 was discover to have a significance value of 0.344 at 0.01 level (2tailed) which purely indicates multicollinearity level present in the date and a treatment and procedure must be conducted to bring it to a lower level than what it is right now to reduce collinear effects that will increases the independence of all explanatory variables present in the model for better outcome from the data.

TABLE 1. Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.196[a]	.039	.009	28.54913

a. Predictors: (Constant), X_1, X_2, X_4, X_5, X_6, and X_7

From the above table 1shows the co-efficient of determination R-value to be 0.196 for the entire explanatory variables with R^2 value 0.039 and adjusted R^2 value of 0.009 having the overall standard error of estimation 0f 28.54913 indicating the level of instability due to error term.

TABLE 2. ANOVA

Model		Sum of Squares	df	Mean Square	F.	Sig.
1	Regression	6312.583	6	1052.097	1.291	.263[b]
	Residual	157305.172	193	815.053		
	Total	163617.755	199			

a. Dependent Variable: Y
b. Predictors: (Constant), X_1, X_2, X_4, X_5, X_6, and X_7

Above table 2 showing the real analysis of variance from the model where alpha value is 0.05 that is 5% level of significant with the sum of the squares, degree of freedom and mean square for regression and residual respectively. It providing F-statistics calculated at 1.291 while the F-statistics from F-table at 6 and infinity for the degree of freedom stand at 2.191 showing the acceptance of Ho; at 5% level of significant. Therefore since calculated F-value is less than the table F- value our assumption is within the range.

TABLE 3. Coefficients

Model		Unstandardized Coefficients					Collinearity Statistics	
1	(Constant)	B	Std. Error	Standardized Coefficients Beta	t	Sig.	Tolerance	VIF
		172.111	132.984		1.294	.197		
	X1	-.050	.143	-.028	-.350	.727	.805	1.243
	X2	.110	.154	.055	.712	.477	.831	1.204
	X4	.152	.137	.082	1.107	.270	.910	1.099
	X5	-.134	1.289	-.008	-.104	.917	.960	1.042
	X6	-.370	.150	-.188	-2.466	.015	.855	1.169
	X7	.151	.302	.036	.502	.616	.979	1.022

a. Dependent Variable: Y

From above table 3 analyzing variance inflation factor as a measure of collinearity in between the explanatory variables in the model indicates presence of multicollinearity from it is values of x1, x2, x3, x4, x5, and x6 which are 1.243, 1.204, 1.099, 1.042, 1.169 and 1.022 respectively with the regression coefficients for instance from the VIF value of x1 equal to 1.243 it means that the variance of estimate coefficient of x_1 is perfectly inflated by a factor of 1.243 and this is so because x_1 and x_2 has higher correlation among all the predictor variables in the model and the lower tolerance level from first analysis on table 3 with the range around 0.979 to 0.805. This is the percentage of the variance in the independent variable that has never been accounted for by other independent variables in the model.

TABLE 4. Collinearity Diagnostics

Mode	Dimension	Eigenvalue	Condition Index	Variance Proportions						
				(Constant)	X1	X2	X4	X5	X6	X7
1	1	6.867	1.000	.00	.00	.00	.00	.00	.00	.00
	2	.063	10.463	.00	.65	.03	.03	.00	.01	.00
	3	.032	14.699	.00	.22	.57	.07	.00	.10	.00
	4	.018	19.509	.00	.09	.07	.23	.00	.82	.01
	5	.017	20.262	.00	.04	.27	.65	.00	.00	.06
	6	.003	44.562	.01	.00	.05	.00	.02	.02	.91
	7	.000	236.283	.99	.00	.01	.00	.98	.04	.01

a. Dependent Variable: Y

Condition indices as above the threshold value is usually found to be in the range of 15 to 30 and with 30 as the most commonly used value indicating the level of the combined collinearity effect among the explanatory variables present in the model By applying treatments and procedure of reducing multicollinearity in the data base source we have the following improvement again**.**

TABLE 5. Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.195[a]	.038	.013	28.48449

a. Predictors: (Constant); X2, X4, X5, X6, X7.[X_1, X_2, X_4, X_5, X_6, and X_7]

From table 1 the standard error of estimation now slightly improved to 28.48449 from the previous one of 28.54913 which means after the reduction of collinear affect now the explanatory variable are much more independent than before.

TABLE 6. ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6212.699	5	1242.540	1.531	.182[b]
	Residual	157405.056	194	811.366		
	Total	163617.755	199			

a. Dependent Variable: Y

b. Predictors: (Constant); X_2, X_4, X_5, X_6, and X_7

Above table 6 showing the analysis of variance from the model after removing the most violating of the explanatory variable in the model where alpha value is 0.05 that is 5%

level of significant with the sum of the squares, degree of freedom and mean square for regression and residual respectively. It provide F-statistics calculated at 1.531 with significance at 0.82 while the F-statistics from F-table at 6 and infinity for the degree of freedom stand at 2.231 showing the rejection of H_1; at 5% level of significant. Therefore since calculated F-value is less than the table F- values our assumption is within the range and indicates reduction in multicollinearity.

TABLE 7. Coefficients

Model		Unstandardized Coefficients					Collinearity Statistics	
1		B	Std. Error	Standardized Coefficients Beta	t	Sig.	Tolerance	VIF
	(Constant)	169.782	132.517		1.281	.202		
	X2	.089	.142	.045	.626	.532	.979	1.021
	X4	.150	.137	.081	1.102	.272	.910	1.099
	X5	-.116	1.285	-.006	-.090	.928	.962	1.040
	X6	-.381	.147	-.194	-2.592	.010	.890	1.124
	X7	.155	.301	.037	.514	.608	.980	1.021

a. Dependent Variable: Y

after the second analysis From above 7 analyzing variance inflation factor as a measure of collinearity in between the explanatory variables in the model indicates presence of multicollinearity from it is values of X_2, X_4, X_5, X_6, and X_7 which are 1.021, 1.099, 1.040 1.124 and 1.021 respectively with the regression coefficients for instance from the VIF value of x_1 equal to 1.243 it means that the variance of estimate coefficient of x_1 is perfectly inflated by a factor of 1.021 against the initial value of 1.243 and this is so because x_3 and x_5 has higher correlation among all the predictor variables in the model and the lower tolerance level from first analysis on table 3 with the range around 0.979 to 0.805. Now improved after the second analysis. This is the percentage of the variance in the independent variable that has never been accounted for by other independent variables in the model.

multicollinearity now drastically reduced by removing the most violating of the explanatory variable with lower t-value, which is x_1 and run the second analysis after which it shows an improvement of VIF X_2, X_3, X_4, X_5, and X_6 which are 1.021, 1.099, 1.040, 1.124 and 1.021 respectively as shown on table 7 and finally indicate a serious reduction in multicollinearity level in the model and increases the tolerance level range from 0.980 to 0.890 on table 7 the tolerance now is moderately enhanced and improved.

Table 8. Collinearity Diagnostics

Mode	Dimension	Eigenvalue	Condition Index	(Constant)	Variance Proportions				
					X2	X4	X5	X6	X7
1	1	5.923	1.000	.00	.00	.00	.00	.00	.00
	2	.037	12.709	.00	.70	.11	.00	.06	.00
	3	.019	17.498	.00	.06	.04	.00	.82	.04
	4	.017	18.701	.00	.16	.85	.00	.04	.04
	5	.003	41.381	.01	.07	.00	.02	.02	.91
	6	.000	219.219	.99	.01	.00	.98	.05	.01

a. Dependent Variable; Y

Condition indices as above the threshold value is usually found to be in the range of 15 to 30 and with 30 as the most commonly used value indicating the level of the combined collinearity effect among the explanatory variables present in the model shows moderate improvement after the second analysis as it has indicated in above two tables 4 and 8.

4. CONCLUSION

Multicollinearity from is not a problematic some time especially if the aims of the analysis is to use multiple regression for prediction purposes, it will be accurate as it is supposed to be despite the presence of multicollinearity, where the problem lies is if to check the contribution of each individual independent variables In the model to the response variable in which it has to do with the co-efficient of correlations in between all of the explanatory variables. some of the factors are definitely a bit redundant in the model, the coefficients of regression and the standard error are increased unnecessarily in response it means co-efficient of some independent variables has turn to be unrealistic and insignificant, that is where multicollinearity is a serious problem because it is making some of the predictor variables more or less independent then as it is supposed to be, we accessed it is level, severity and degree, in which either dangerous not even the moderate but a severe multicollinearity was found in this work, we reduced it to minimum by dropping the most violating of the predictor variables in the model, runs the second analysis and

investigates VIF, tolerance level, index number, standard error as phenomenon which measure the level of multicollinearity among the explanatory variables in multiple regression mode, which is now find to be significantly enhanced and the VIF, tolerance level, index number, standard error are meritoriously improved, it has now been successfully reduced to a minimum level of multicollinearity and finally all independent variables are now much more better independent than before.

REFERENCES

1. Abdalla, M. EL-Habil, Khaled I.A. Almghari (2011): Remedy of multicollinearity using Ridge regression. Journal of Natural Sciences, Al Azhar University-Gaza, vol. (13): pp.119-134.
2. Donald, E. Farrar and Robert R. Glauber (1967): the Review of Economics and Statistics Vol. 49, No. 1, pp. 92-107.
3. Neter, wasserman and Kutner (1989): Applied liner Regression Model, second edition. Irwin, Homeqood IL.
4. Jame lani (2019): What is Multiple Linear Regression? https://www.statisticssolutions.com/what-is-multiple-linear-regression/
5. Michael, H. Graham (2003): Confronting Multicollinearity in Ecological Multiple Regression, Ecological Society of America, Ecology, 84(11), pp. 2809–2815.
6. Tabachnick, B. G., and Fidell L. S. (1996): Using Multivariate Statistics (3rd Ed.). New York: Harper Collins.
7. Matthew Ryan Lavery, Parul Acharya, Stephen A. Sivo and Lihua Xu (2017); Number of Predictors and Multicollinearity What Are Their Effects on Error and Bias in Regression?, Communications in Statistics - Simulation and Computation.
8. https://en.wikipedia.org/wiki/Multicollinearity
9. Blaze, T. J., and Ye F. (2012): The Effect of Within- and Cross-Level Multicollinearity on Parameter Estimates and Standard Errors in Multilevel Modeling with Different Centering Methods, American Educational Research Association Conference, Vancouver, British Columbia, Canada.
10. Kroll, C. N. and Song P. (2013): Impact of multicollinearity on small sample hydrologic regression models, Water Resources Research, 49(6), pp. 3756--3769.
11. Robinson, C., & Schumacker, R. E. (2009): Interaction effects: Centering, variance inflation factor, and interpretation issues. Multiple Linear Regression Viewpoints, 35(1), pp. 6-11.
12. Kiers, H., and A. Smilde (2007): A comparison of various methods for multivariate regression with highly collinear variables, Statistical Methods Application, 16(2), pp. 193–228.
13. Saman, Babaie-Kafaki and Mahdi Roozbeh (2017): A revised Cholesky decomposition to combat multicollinearity in multiple regression models, Journal of Statistical Computation and Simulation.
14. García, C.B. García, J. López Martín M.M and Salmerón R. (2015): Collinearity: revisiting the variance inflation factor in ridge regression, Journal of Applied Statistics, 42:3, pp. 648-661.
15. Edward, R. Mansfield and Billy P. Helms (1982): Detecting Multicollinearity, The American Statistician, Vol; 36:3a, pp. 158-160.
16. Adeboye, N.O, Fagoyinbo, I. S and Olatayo T.O (2014): Estimation of the Effect of Multicollinearity on the Standard Error for Regression Coefficients. IOSR Journal of Mathematics (IOSR-JM), Vol. 10; pp. 2278-2284.
17. https://www.investopedia.com/terms/v/variance-inflation-factor.asp
18. Lisa, A. Weissfeld and Susan M. Sereika (1991): A Multicollinearity Diagnostic for Generalized Linear Models, Communications in Statistics - Theory and Methods, 20:4, pp. 1183-1198.
19. Chien-Chia, L. Huang, Yow-Jen Jou and Hsun-Jung Cho (2015): A New Multicollinearity Diagnostic for Generalized Linear Models, Journal of Applied Statistics.
20. Ranjit kumar Paul (2008): Multicollinearity: Causes, Effects and Remedies.
21. Harshada Joshi, Hrishikesh Kulkarni Cytel and Swapna Deshpande (2012): Multicollinearity Diagnostics in Statistical Modeling and Remedies to Deal with it using SAS, Statistical Software & Services Pvt Ltd, Pune, India.
22. Yoshio Takane and Elliot M. Cramer (1975): Regions of Significance in Multiple Regression Analysis, Multivariate Behavioral Research, Vol. 10:3, pp.373-383.
23. Tomáš Jurczyk (2012): Outlier Detection under Multicollinearity, Journal of Statistical Computation and Simulation, 82:2, pp. 261-278.
24. Chandrasekhar, C.K., Bagyalakshmi, H. Srinivasan M.R. and Gallo M. (2016): Partial Ridge Regression under Multicollinearity, Journal of Applied Statistics.
25. Carl, F., Mela and Praveen K. Kopalle (2002): The Impact Of Collinearity On Regression Analysis: The Asymmetric Effect Of Negative And Positive Correlations, Applied Economics, vol. 34; pp. 667-6677.
26. Robert, M. O'brien (2007): A Caution Regarding Rules of Thumb for Variance Inflation Factors, Quality and Quantity 41:pp. 673–690.
27. Ryuta Tamura, Ken Kobayashi, Yuichi Takano, Ryuhei Miyashiro, Kazuhide Nakata and Tomomi Matsui (2019): Mixed Integer Quadratic Optimization Formulations for Eliminating Multicollinearity Based on Variance Inflation Factor, Journal of Global Optimization, vol. 73, issue 2, pp. 431-446.
28. Richard Williams (2015): Multicollinearity, University of Notre Dame, https://www3.nd.edu/~rwilliam/ Last revised January 13.
29. Saman Babaie-Kafaki and Mahdi Roozbeh (2017): A Revised Cholesky Decomposition to Combat Multicollinearity in Multiple Regression Models, Journal of Statistical Computation and Simulation. pp. 1563-5163.
30. Peter Song and Chuck Kroll (2011): The Impact of Multicollinearity on Small Sample Hydrologic Regional Regression, World Environmental and Water Resources Congress 2011:pp. 3713-3723.
31. José Dias Curto and José Castro Pinto (2011): The corrected VIF (CVIF), Journal of Applied Statistics, 38:7, 1499-1507.
32. Goldberger, Arthur S. (1991): Multicollinearity. A Course in Econometrics. Cambridge: Harvard University Press. pp. 245–53.