IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 17, Issue 2 Ser. II (Mar. – Apr. 2021), PP 28-43
www.iosrjournals.org

Using Different Methods to Overcome Modeling Problems

Ahmed Mohamed Mohamed Elsayed
Al-Obour High Institute For Management & Informatics.
Department of Basic Science.
Kilo 21 Cairo-Belbies Road, P.O. Box 27 Al-Obour City, Egypt.

Abstract: *There are many problems of modeling process. Multicollinearity phenomenon one is happened when there are high collinearity between the independent variables. It makes hard to interpret the coefficients, and reduces the power of the model. In this paper, we tried to solve this problem using two methods. The first one used the ridge Regression model (RRM). It is compared with a traditional linear regression model (LRM). The second one modified the original dataset by differencing (using the function "diffM" in "MTS" package), and scaling (using the function "scale" in "base" package) processes. We supposed three cases of the independent variables for this justify this purpose. Independent, Dependent, and Combination linear cases. The simulation study is used to generate the dataset, with 500 observation for each variable, using R program. The "MASS" and the "ridg" packages, and their functions "lm.ridge", "check_collinearity()", and "Linear.Ridge" all are used to determine the variance inflation factor (VIF) for each independent variable to know whether the Multicollinearity is absent or not. The ridge parameter (RP) is chosen automatically from the "Linear.Ridge" function. From studying this simulation, we insured from that knowledge: If the RP is small or moderate value, then there is no need to use the RRM to modify the obtained results. Also, the presence of strong collinearity between the independent variables, increases the VIF as well as RP. The strong collinearity between the independent variables does not reflect the Multicollinearity.Finally, the second method for overcoming the Multicollinearity is effective way to eliminate the Multicollinearity phenomenon and make the regression model results well.*

Keywords: *Linear Regression Model; Ridge Regression Model; Ridge parameter; Multicollinearity; MASS package; Generate dataset.*

Date of Submission: 18-03-2021 Date of Acceptance: 01-04-2021

I. Introduction

Multicollinearity criteria occur when the independent variables in a regression model are correlated. This collinearity is a problem, because the independent variables should be independent. If the degree of collinearity between variables is high enough, it can cause problems when fitting the model.

The potential solutions of multicollinearity include the following: Remove some of highly correlated independent variables. Add some independent variables together with linear combine. Perform an analysis designed for highly correlated variables, such as principal components analysis (PCA) or partial least squares (PLS) regression [1]. Fortunately, there is a very simple test to assess multicollinearity by identifying the collinearity between independent variables. The variance inflation factors (VIFs) can be used for this purpose. The VIFs start from "1" and have no upper limit. A value of "1" indicates that there is no collinearity between these independent variables. The range between "1" and 5 suggests that there is a moderate collinearity. The VIFs greater than 5 represent critical levels of multicollinearity, where the coefficients are poorly estimated, and the p-values are questionable. There are also situations where high VFIs can be safely ignored without suffering from multicollinearity: When high VIFs only exist in control variables, the regression coefficients are not impacted. When high VIFs are caused as a result of the inclusion of the products or powers of other variables, multicollinearity does not cause negative impacts. When a dummy variable that represents more than two categories has a high VIF, multicollinearity does not necessarily exist [2:5]. Multicollinearity is generally more severe in small samples, Goldberger[6] called it "micronumerosity". The are some tests to detect the multicollinearity phenomenon such as Farrar–Glauber test [7]. If the variables orthogonal, then there is no multicollinearity. Wichers[8] has argued that Farrar–Glauber test is ineffective in a given partial correlation with different multicollinearity patterns. This test has been criticized by other researchers[9:10]. The condition number (also considered as a test) is computed by finding square root of the maximum eigenvalue divided by the minimum eigenvalue of the design matrix.

If the condition number is above 30, then the regression may have high Multicollinearity. One advantage of this method, it shows which variables are causing this problem [11]. Multicollinearity can be detected also by adding random noise to the data and re-running the regression many times and seeing how much the coefficients change [12]. Multicollinearity may affect the results of fitting regression application data [13], econometrics data [14], business data and industry data [15], and financial data [16]. Lateral collinearity and misleading results are devoted by Kock and Lynn [17]. Reviews and provides examples of the different ways in which multicollinearity can affect a research project will be found in [18].

This paper is organized as: Section II presents the materials and methods that are used in this paper. Section III presents the numerical study. Section VI presents the discussion of the obtained results from Section III. Finally, Section V presents the conclusions.

II. Material and Methods

In this section, we present some methods, criteria and algorithms that are used in this article. Namely, the linear regression model (LRM), Multicollinearity phenomenon, the VIF, the ridge regression model (RRM), some methods for normalizing (scaling) and differencing the simulation dataset, and the used packages from R program.

II.1 Linear Regression Model (LRM)

As we know, the LRM is a powerful tool for predicting numerical values. The "lm()" function in R program creates a regression model. Use the "summary()" function to review the weights and performance measures. The residuals can be examined by pulling on the "$resid" variable from the model. Residuals present the errors between the predicted and the actual values. Smaller residuals are better. Residual Conditions of LRM: (1) Mean of the errors is zero. (2) Distributions of the errors are normal. (3) Errors are independent. (4) Variance of errors is constant. Standard Error: is the standard deviation of the residuals. Smaller is better. R-squared : shows the amount of variance explained by the model. Adjusted R-Square : most useful for multiple-regression. The adjusted R-squared shows whether adding additional predictors improve a regression model or not. Negative Adjusted R-square appears when Residual sum of squares approaches to total sum of squares, that means the explanation towards response is very low or negligible. This means that insignificance of explanatory variables. F-test checks if at least one variable’s weight is significantly different from zero. Histogram and QQ-plot is used to explain if the residual fit is a normal distribution or not. Null hypothesis of the Jarque-Bera test is that the errors are normally distributed. Null hypothesis of the Durbin-Watson test is that the errors are serially uncorrelated. Constant variance can be checked by looking at the “Studentized” residuals – normalized based on the standard deviation. The presence of multicollinearity has a negative impact on the analysis as a whole, and can severely limit the conclusions of the research study.

II.2 Multicollinearity Criterion

When the independent variables are correlated, this leading to the notation of Multicollinearity. It indicates that changes in one variable are associated with shifts in another variable. The stronger collinearity, the more difficult it is to change one variable without changing another. Multicollinearity makes it hard to interpret coefficients, and it reduces the power of the model to identify a significant independent variables. There are two kinds of multicollinearity. Structural multicollinearity: this type occurs when we create a model term using other terms. Data multicollinearity: this type is presented in the data itself rather than being an artifact of the model.

Multicollinearity causes:

(1)Estimates can swing wildly based on which other independent variables are in the model. (2)Coefficients become very sensitive to small changes in the model. The severity of the problems increases with the degree of the multicollinearity. Therefore, if we have low or moderate multicollinearity, we may not need to resolve it. Multicollinearity affects the coefficients and P-values, but it does not influence the predictions, precision of the predictions, and the goodness-of-fit statistics. Multicollinearity inflates the variance of coefficients and causes type II errors, it is essential to detect and correct it. There are simple and commonly ways to correct multicollinearity: (1)Remove one or more of the highly correlated variables. (2)Use principal components analysis (PCA) or partial least square (PLS) regression instead of ordinary least squares (OLS) regression. PLS regression can reduce the variables to a smaller set with no collinearity among them. (3)We can use Robust regression analysis instead of OLS, such as Ridge regression and Lasso regression. (4)Statistical learning regression is also a good method, like regression Tree, Bagging regression, Random-Forest regression, Neural network and Support-Vector Regression (SVR).

II.3 Variance Inflation Factor (VIF)

If there is perfect multicollinearity among the independent variables, X **.** So the strong collinearity will cause computational instability and the OLS estimator is no longer the BLUE (best linear unbiased estimator). VIF can be calculated by equation (1):

$$\mathrm{VIF}_j = \frac{1}{1-R_j^2} = \frac{1}{\mathrm{Tolerance}} \qquad (1)$$

Where R_j^2 represents a coefficient of determination for regressing the j-th independent variable on the remaining ones. VIF or Tolerance can be used to detect multicollinearity. If R_j^2 is equal to 0, the variance of the remaining independent variables cannot be predicted from the j-th independent variable. Therefore, when VIF or tolerance is equal to 1, the j-th independent variable is not correlated . In this case, the variance of the j-th regression coefficient is not inflated. Generally, VIF above 4 or Tolerance below 0.25 indicates that multicollinearity might exist. When VIF is higher than 10 or Tolerance is lower than 0.1, there is significant multicollinearity that needs to be corrected [5].

II.4 Ridge Regression Model (RRM)

The RRM is used to analyze any data that suffers from multicollinearity. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, these results in predicted values to be far away from the actual values. The cost function for RRM can be denoted by equation (2):

$$\mathrm{Min}\left(\| Y - X\,(\theta) \|^2 + \lambda \| \theta \|^2 \right) \qquad (2)$$

Lambda λ is the penalty term is denoted by α parameter in the ridge function. So, by changing the values of α , we are controlling the penalty term λ. Higher values of α , bigger the penalty λ, and therefore the magnitude of coefficients is reduced. Therefore, it is used to prevent multicollinearity. If we add the λ parameter to the regression $Y = X\beta + e$ equation, then the variance not evaluated by the general model is considered. The RRM addresses the problem by estimating regression coefficients by equation (3):

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y \qquad (3)$$

The λ is the ridge parameter (RP) , and I is the identity matrix. Small positive values of λ reduce the variance of the estimates. Bias increases as λ increases. The variance decreases as λ increases. While the biased causes smaller mean square error compared to least-squares estimates. The assumptions of RRM are the same as that of LRM: linearity, constant variance, and independence. However, as ridge regression does not provide confidence limits, the distribution of errors to be normal need not be assumed[19:24].

II.5 Normalizing and Scaling Dataset

Objective is to improve predictive accuracy and not allow a particular feature of dataset impact the prediction due to large numeric value range. We may need to normalize or scale values under different features such that they fall under common range. Data frame could be normalized using Min-Max normalization technique that specifies the following formula to be applied to each value of features to be normalize. This technique can be formulated as equation (4):

$$\textbf{Scaled data} = \frac{X - Min(X)}{Max(X) - Min(X)} \qquad (4),$$

Disadvantage with min-max normalization technique is that it tends to bring data towards the mean. In order to achieve z-score standardization, we use R’s built-in "scale()" function [25].

II.6 R packages

In R language, we used the "lm" function in "stats" package for linear regression model (LRM), the function "lm.ridge()" in package MASS [26] for implement ridge regression model (RRM). Furthermore, package "ridge" provides a function called "linearRidge()" [27], which also can fits a RRM, and optionally, the ridge parameter λ is chosen automatically using the method proposed by Cule et al. [28]. In this case, the function choose 0.01 as λ, so the result is little different from the output of "lm.ridge()" function. For differencing data we will use the function "diffM()" in "MTS" package. Also, for scaling the dataset we will use the function "scale" in the package "base".

III. NUMERICAL STUDY

In this section, we will generate the dataset using some methods of univariate and multivariate samples with 500 observations in whole dataset frame. Then we classify the dataset into separate variables independent (correlated) and dependent variables. In the next subsections, we study the multicollinearity in different cases:

III.1 Multicollinearity in Independence Case

In this subsection, we generate all variables data as standard normal variables with 500 observations. This is to get variables with absent the multicollinearity. The correlation matrix explains that the pairwise correlations are low:

Variable	X_1	X_2	X_3	X_4
X_1	1	-0.008	-0.035	-0.054
X_2	-0.008	1	0.004	0.019
X_3	-0.035	0.004	1	0.034
X_4	-0.054	0.019	0.034	1

The LRM is:

$$y = 0.005876 - 0.022609x_1 - 0.009309x_2 + 0.001819x_3 + 0.153541x_4$$

Residuals	Min	1Q	Median	3Q	Max
	-2.88881	-0.74867	0.06363	0.67903	2.93159

We can test LRM as:

Coefficients	Estimate	Std. Error	t-statistic	P-value
Intercept	0.005876	0.045365	0.130	0.896998
X_1	-0.022609	0.045840	-0.493	0.622080
X_2	-0.009309	0.045300	-0.205	0.837272
X_3	0.001819	0.044994	0.040	0.967768
X_4	0.153541	0.043532	3.527	0.000459 ***

Residual standard error	Multiple R-squared	Adjusted R-squared	F-statistic	P-value
1.006	0.02549	0.01762	3.238	0.01225*

The LRM has a much lower standard error, meaning the residuals have a small variance. The R-squared is too low. F-test is statistically significant (at 5%). This means that the LRM model has at least one variable, X_4, that is significantly different from zero.

Test for independence of residuals: Durbin-Watson test

H_0: Errors are serially uncorrelated.
H_1: Errors are serially correlated.
DW = 1.9594, P-value = 0.3274.
We accept H_0. Errors are serially uncorrelated.

Test residuals for normality (Jarque - Beranormalality test)

H_0: Errors are normally distributed.
H_1: Errors are not normally distributed.
JB= 1.4382, P-value = 0.4872.
We accept H_0: Errors are normally distributed.

Figure 1 displays the plots of residuals in independence case:

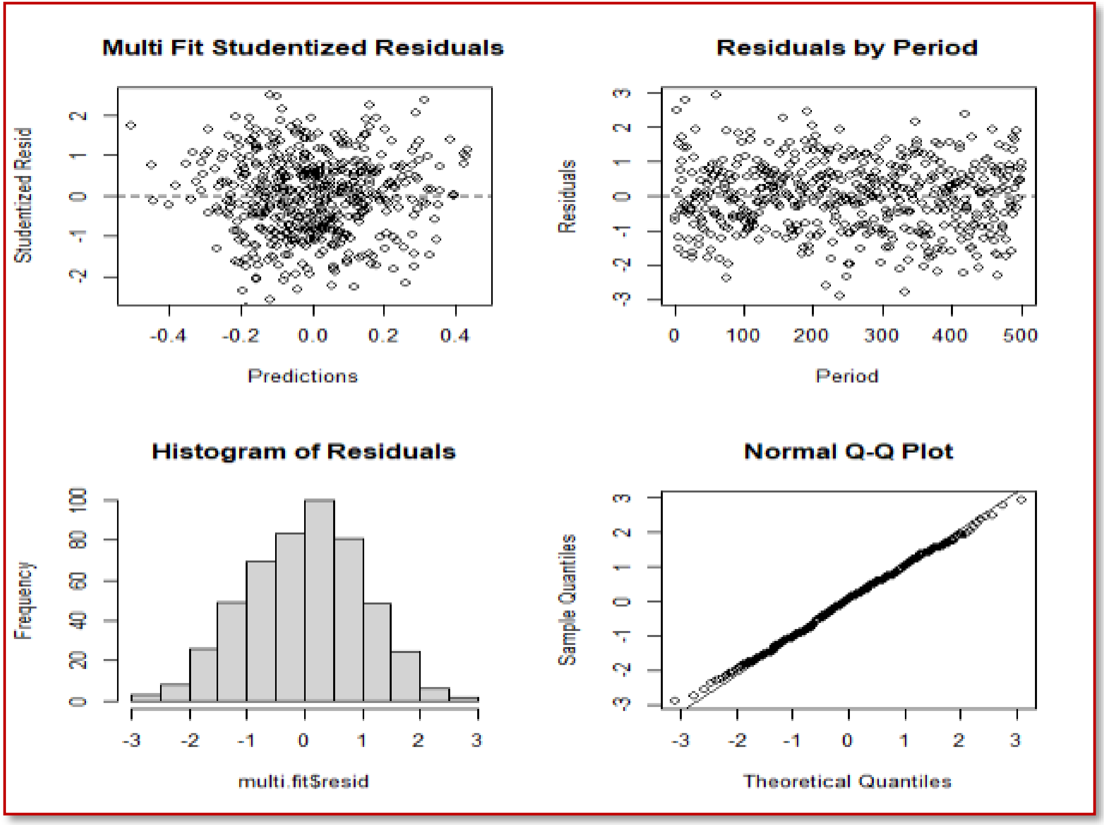

Figure 1:Plots of residuals in independence case

Multi Fit Studentized Residuals plot shows that there is not any obvious outliers.
Histogram of residuals look normally distributed. The QQ-Plot shows all points around the normal line.We can check for Multicollinearity using the function "check_collinearity()" of fitted model. We get a low collinearity, VIF = 1, SE = 1, for all independent variables as shown in Figure 2:

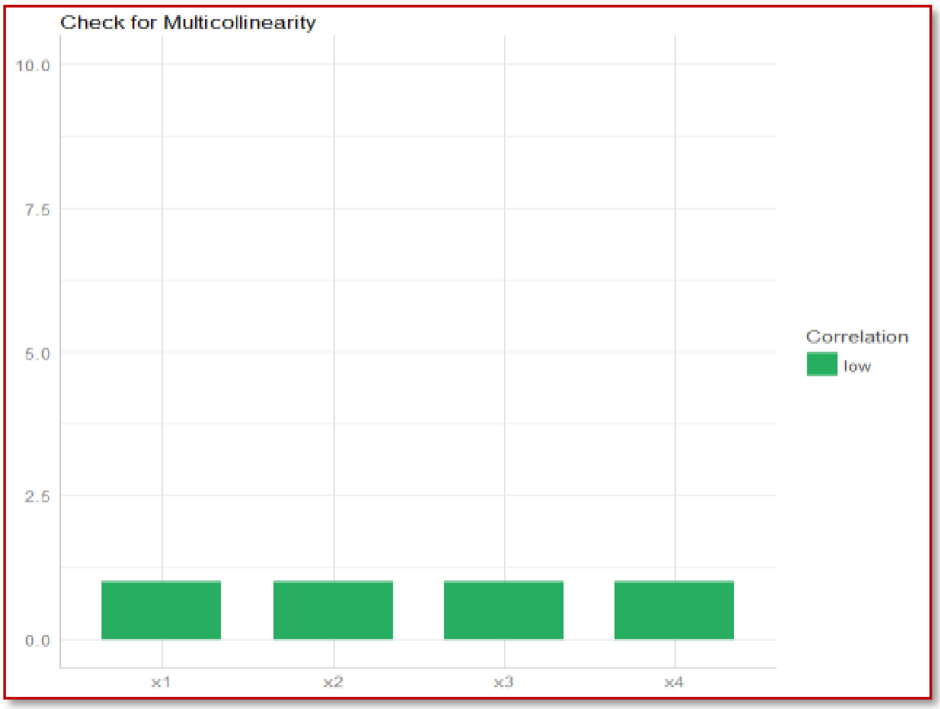

Figure 2: Check for Multicollinearity in independent case

Using the RRM to fit the data. Here, we omit the intercept from the model.
The RRM using the "lm.ridge()" function with RP = 0.1 is:

$$y = -0.022606313x_1 - 0.009258321x_2 + 0.001163623x_3 + 0.158883915x_4$$

Also, we can use the function "linearRidge()" in the "ridge" package, without specified λ, the RRM is:

Coefficients	Estimate	Scaled estimate	Std. Error (scaled)	t statistic	P-value
X_1	-0.018397	-0.404525	0.746532	0.542	0.587907
X_2	-0.006307	-0.140074	0.746298	0.188	0.851119
X_3	0.002477	0.055430	0.746421	0.074	0.940803
X_4	0.114060	2.641221	0.746545	3.538	0.000403 ***

RP = 0.3473546, chosen automatically, computed using 2PCs, Variance = 2.201, residual = 3.733. This means that the RRM also has at least one variable, X_4, is significantly different from zero.

III.2 Multicollinearity in Dependence Case

With "MASS" package using the function "mvrnorm()" with mu = 0, and Sigma matrix all off-diagonal = 0.7, with n=500 observations, to generate the correlated (independent variables X's) , and standard normal distribution to generate the (dependent variable Y). The correlation matrix for correlated variables is:

Variables	X_1	X_2	X_3	X_4
X_1	1	0.699	0.694	0.681
X_2	0.699	1	0.696	0.673
X_3	0.694	0.696	1	0.669
X_4	0.681	0.673	0.669	1

The LRM is:

$$y = 0.031641 - 0.031379x_1 - 0.015836x_2 + 0.007256x_3 + 0.080905x_4$$

Residuals	Min	1Q	Median	3Q	Max
	-3.1011	-0.6576	-0.0214	0.6530	3.3271

We can test LRM as:

Coefficients	Estimate	Std. Error	t-statistic	P-value
Intercept	0.031641	0.045756	0.692	0.490
X_1	-0.031379	0.073357	-0.428	0.669
X_2	-0.015836	0.076246	-0.208	0.836
X_3	0.007256	0.073935	0.098	0.922
X_4	0.080905	0.071961	1.124	0.261

.

Residual standard error	Multiple R-squared	Adjusted R-squared	F-statistic	P-value
1.021	0.00329	-0.004764	0.4085	0.8026

F-test is not statistically significant (at 5%). This means that the LRM has no one variable is significantly different from zero.
DW = 2.0151, P-value = 0.5663.
We accept H_0. Errors are serially uncorrelated.
JB= 0.2863, P-value = 0.8666.
We accept H_0: Errors are normally distributed.

Figure 3 displays the residual plots in dependence case:

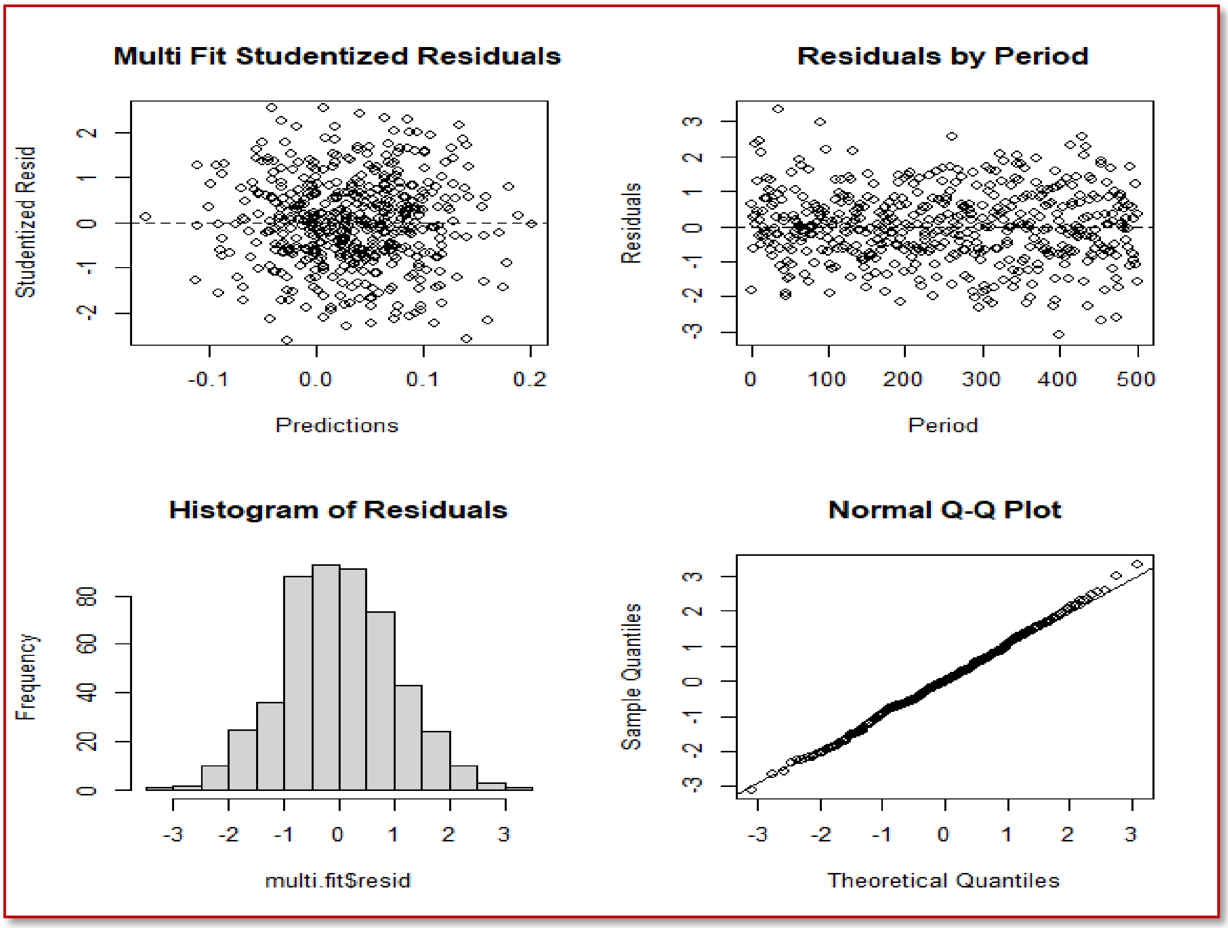

Figure 3:Plots of residuals in dependence case

We can check the multicollinearity in dependence case, we have low multicollinearity, as we see in Figure 4, and from VIF values.

Parameter	VIF	Increased SE
X_1	2.55	1.6
X_2	2.53	1.59
X_3	2.48	1.58
X_4	2.32	1.52

Figure 4 displays the plots of ckeck of multicollinearity in dependence case:

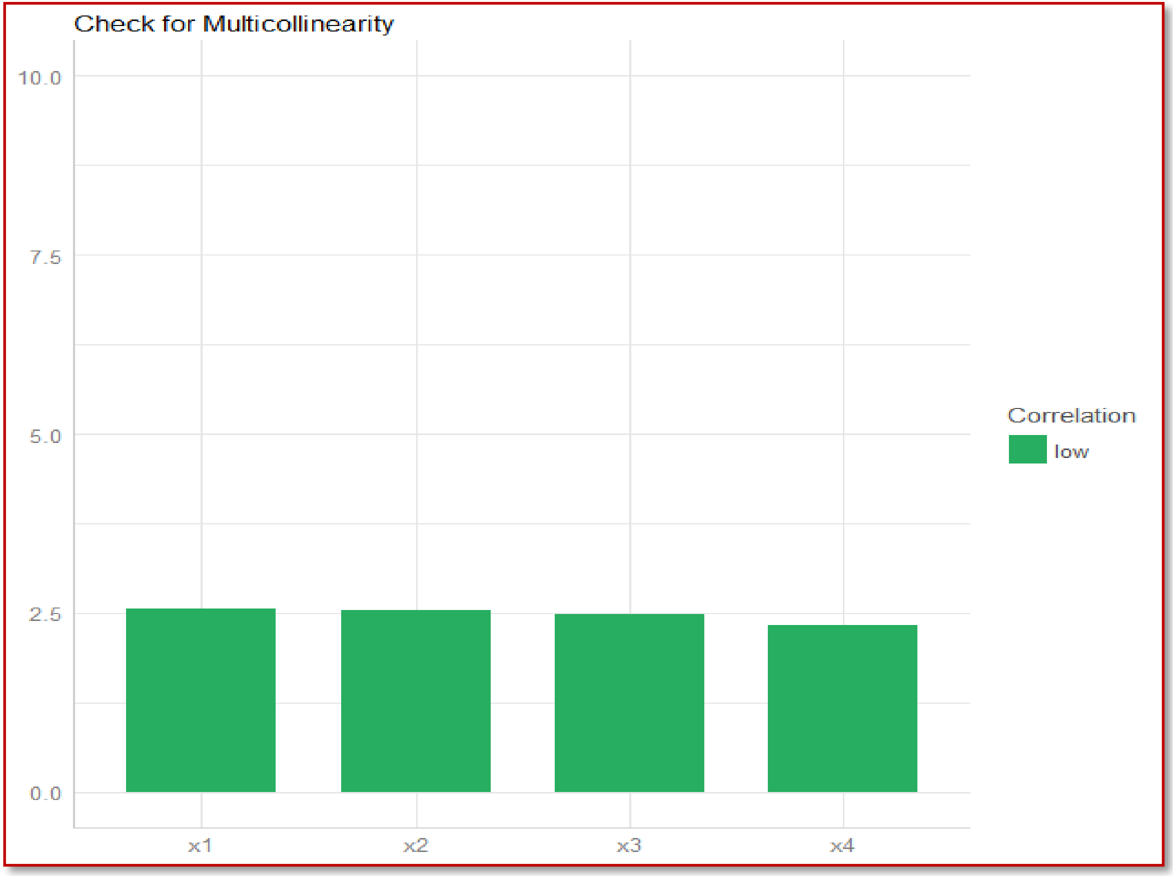

Figure 4: Check of Multicollinearity in correlated case

The RRM using the "lm.ridge()" function with RP = 0.1 is:

$$y = -0.03262302x_1 - 0.01576830x_2 + 0.01005597x_3 + 0.07757007x_4$$

Also, we can use the function "**linearRidge**()" in ridge package, without specified λ, the RRM is:

Coefficients	Estimate	Scaled estimate	Std. Error (scaled)	t statistic	P-value
X_1	0.001221	0.027172	0.128467	0.212	0.832
X_2	0.002010	0.042796	0.128631	0.333	0.739
X_3	0.003080	0.067046	0.128871	0.520	0.603
X_4	0.007214	0.156076	0.129809	1.202	0.229

RP = 5.803454, chosen automatically, computed using 1 PCs, Variance = 0.1269 , residual = 0.8715.
This means that the RRM also has no one variable significantly different from zero.

From independence and dependence case we desire to insure that: if λ is small value , then there is no need to use RRM, because the results do not change much. The presence of a strong collinearity between the independent variables increases somewhat the VIF as well as λ .

III.3 Multicollinearity in Linearity Combinations Case

In this subsection, we will suppose some linear combinations between the independent variables, and then we can check the multicollinearity. These combinations and generation data process can be presented using R program as :

X_1 = rnorm(500),
X_2 = rnorm(500)+X_1,
X_3 = rpois(500,3)+X_1+X_2,
X_4 = rexp(500)+X_3+X_2+X_1,Y = rnorm(500).

Of course, Multicollinearity will be there. We can use two methods to adjust the dataset. The first one is differencing these data. The second is normalizing (scaling) the data, and then we test whether if the multicollinearity disappeared or not?

III.3.1 Analyzing the Original Data

Before differencing process, we will analyze the original data, to find the impact of relations between the independent variables. The LRM is:

$$y = 0.01885 + 0.09361x_1 + 0.05423x_2 + 0.04657x_3 - 0.03899x_4$$

The residuals are:

residuals	Min	1Q	Median	3Q	Max
	-2.71896	-0.63534	-0.04729	0.69282	2.86544

We can test the LRM as:

Variables	Estimate	Std. Error	t statistic	P-value
Intercept	0.01885	0.10232	0.184	0.854
X_1	0.09361	0.08624	1.086	0.278
X_2	0.05423	0.07113	0.762	0.446
X_3	0.04657	0.05470	0.851	0.395
X_4	-0.03899	0.04873	-0.800	0.424

Residual standard error = 0.965 , Multiple R-squared = 0.009666, Adjusted R-squared = 0.001663. F-statistic = 1.208 , P-value = 0.3065.
DW = 2.0273, P-value = 0.6221.
We accept H_0. Errors are serially uncorrelated.
JB = 0.6724, P-value: 0.7145.
We accept H_0: Errors are normally distributed.

These results are indicated by Figure 5:

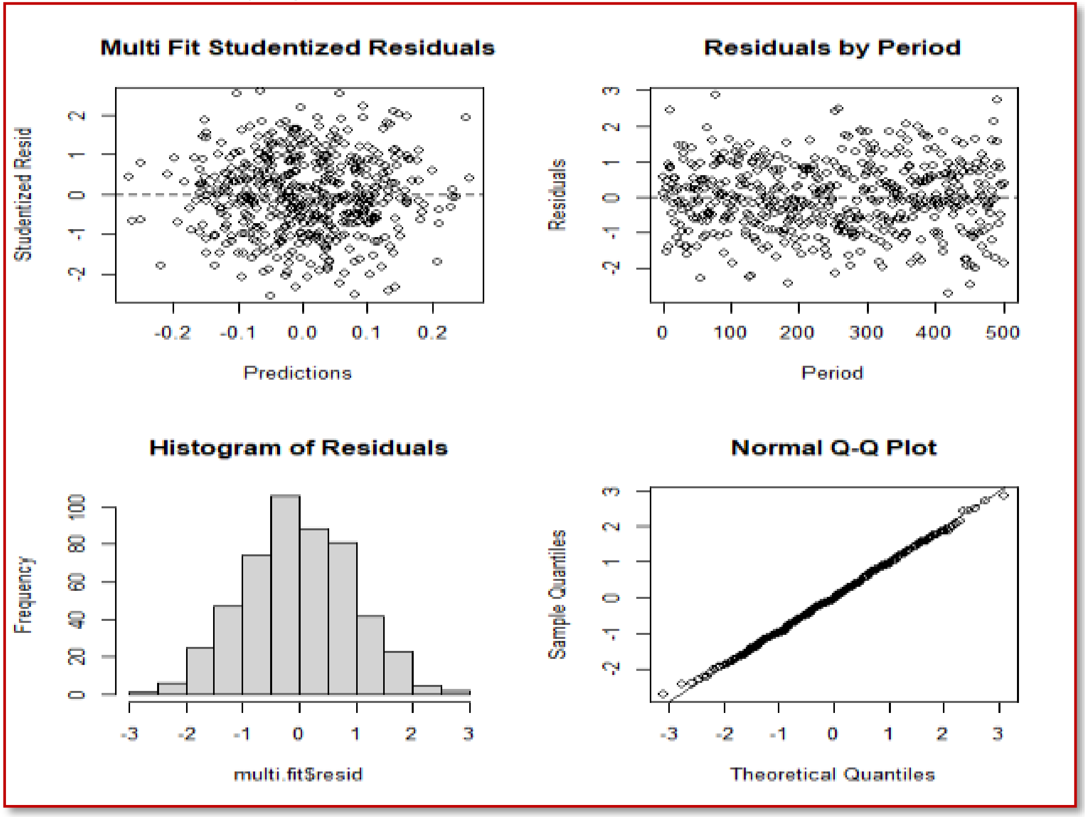

Figure 5: The plots of residuals in linearity case - original data

Check for Multicollinearity

Parameter	VIF	Increased SE	Multicollinearity
X_1	3.71	1.93	Low
X_2	5.36	2.32	Moderate
X_3	12.89	3.59	High
X_4	30.08	5.48	High

These results are indicated by Figure 6:

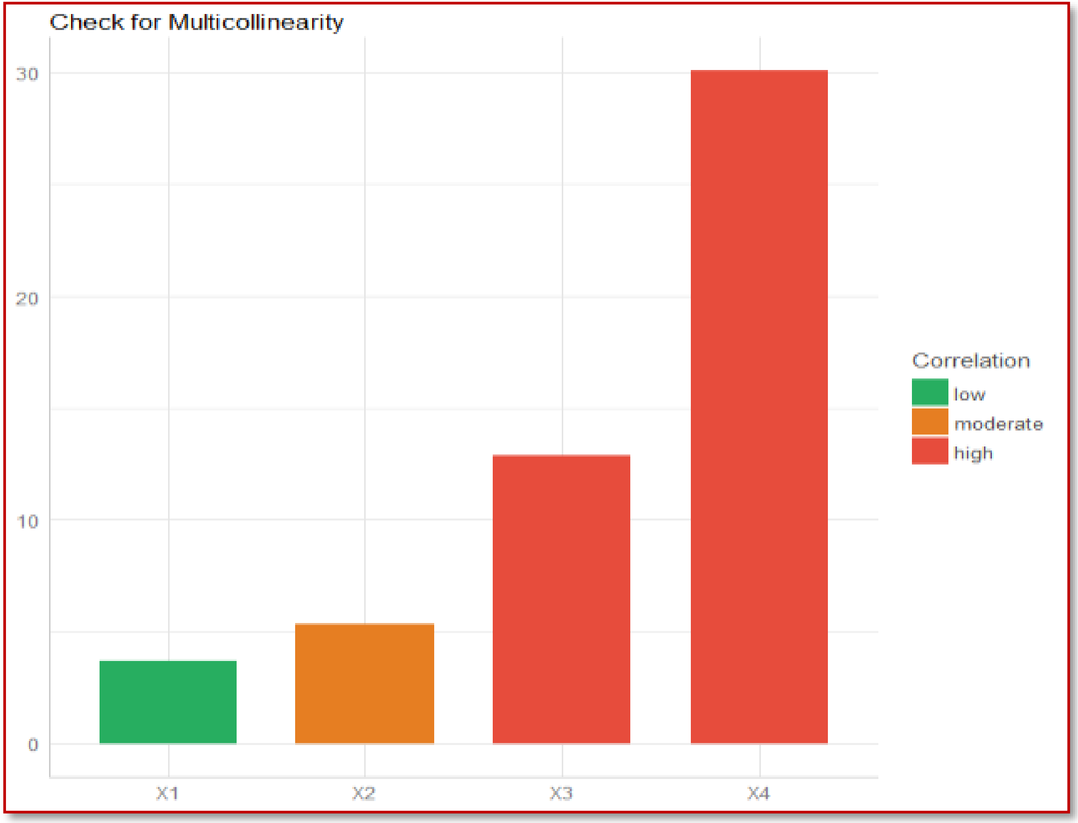

Figure 6: Check multicollinearity in linearity case - original data

3.3.2 Differencing the Original Data

After differencing process, the dataset will be 499 observations. So, we get the LRM as:

$$y = -0.0006085 - 0.0300114x_1 + 1.0728980x_2 + 0.9849307x_3 + 0.9863133x_4$$

The residuals are:

Residuals	Min	1Q	Median	3Q	Max
	-5.3058	-0.6016	0.0368	0.6449	5.3332

We can test the LRM as:

Variables	Estimate	Std. Error	t statistic	P-value
intercept	-0.0006085	0.0581376	-0.010	0.992
X_1	-0.0303114	0.0427593	-0.709	0.479
X_2	1.0728980	0.0661380	16.222	2e-16 ***
X_3	0.9849307	0.0485142	20.302	2e-16 ***
X_4	0.9863133	0.0244765	40.296	2e-16 ***

Residual standard error: 1.299, Multiple R-squared = 0.9644, Adjusted R-squared = 0.9641, F-statistic = 3341, P-value = 2.2e-16.
DW = 3.131, P-value = 1.

We accept H_0. Errors are serially uncorrelated.
JB = 67.9122, P-value: 1.776e-15.
We reject H_0. Errors are not normally distributed.

These results are indicated by Figure 7:

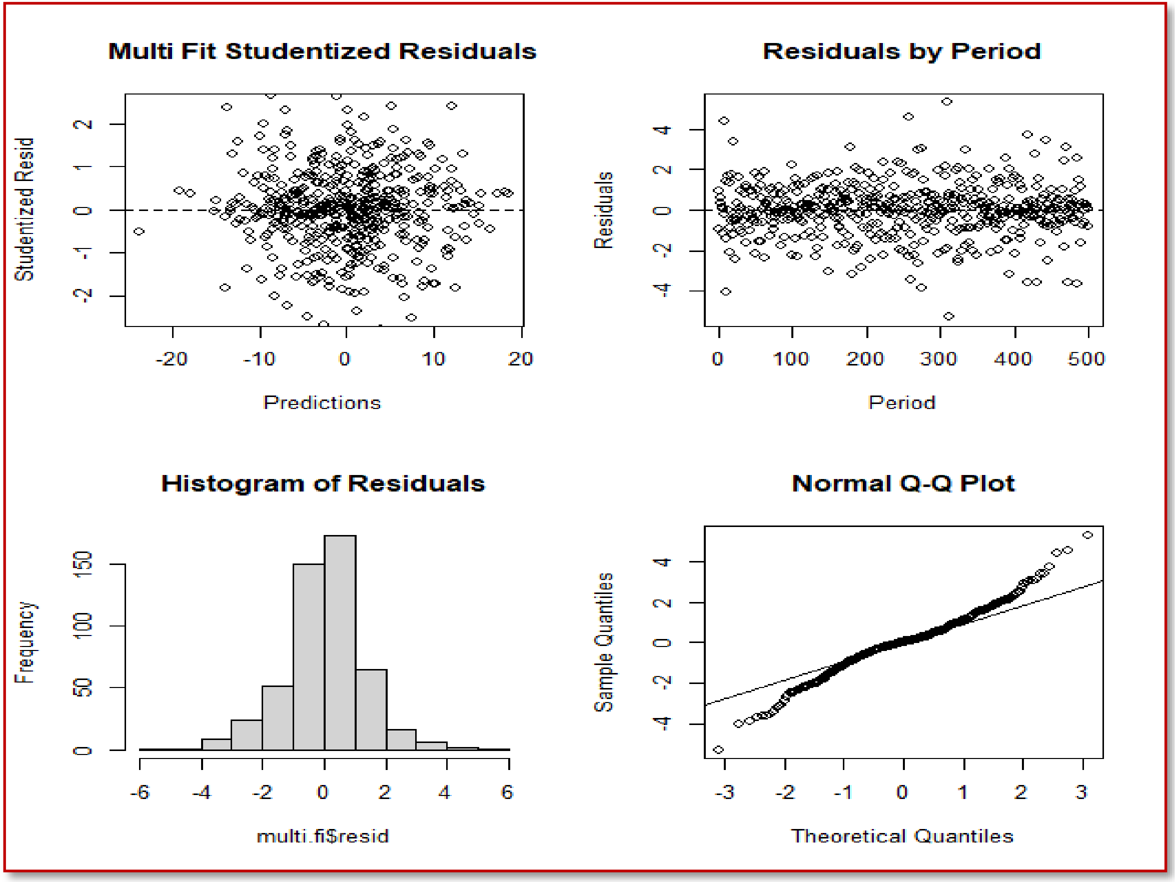

Figure 7: The plots ofresiduals in linearity case - differenced data

Check for Multicollinearity

Parameter	VIF	Increased SE	Multicollinearity
X_1	1.03	1.01	Low
X_2	2.40	1.55	Low
X_3	2.76	1.66	Low
X_4	2.79	1.67	Low

These results are indicated by Figure 8:

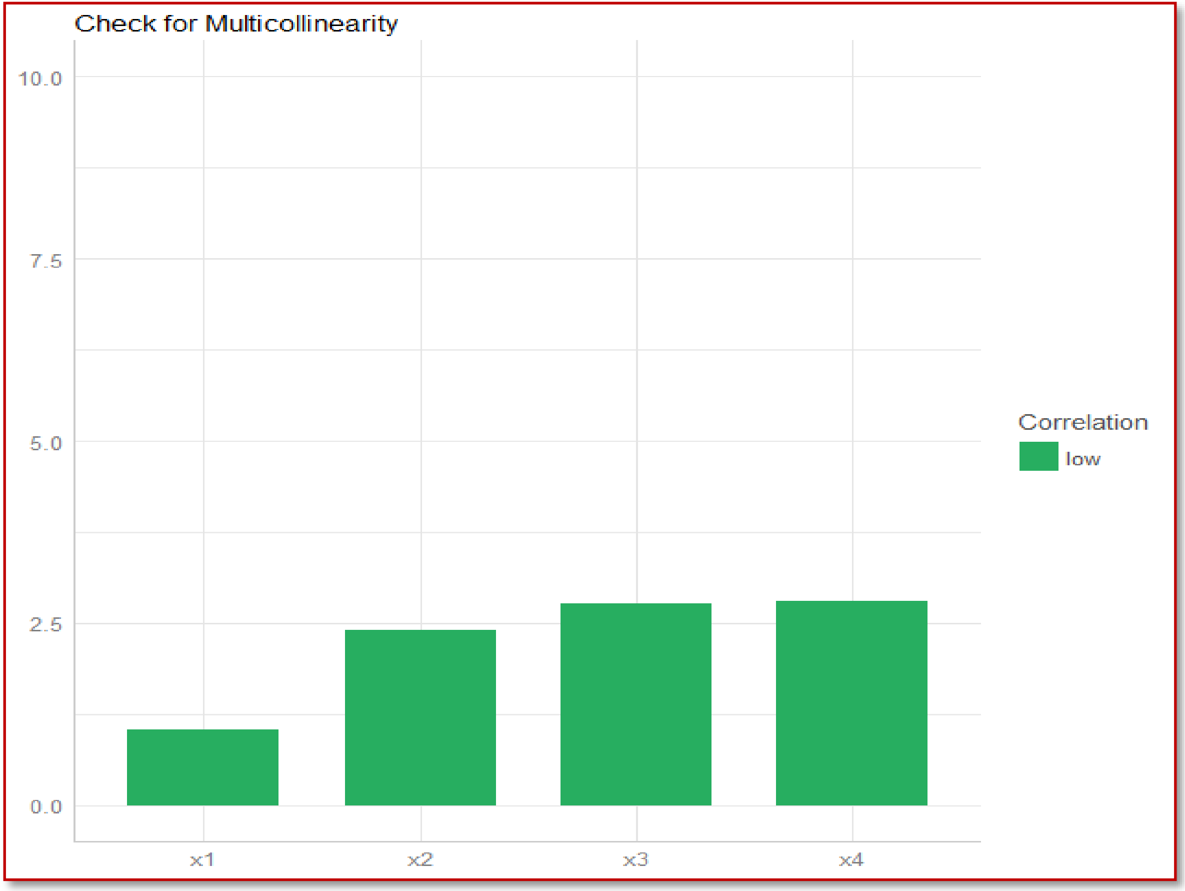

Figure 8: Check of Multicollinearity in linearity case - differenced data

The RRM using the "lm.ridge()" function with RP = 0.1 is:

$$y = -0.92217x_1 + 32.69305x_2 + 43.81672x_3 + 78.33978x_4$$

The RRM using the "linearRidge" function is:

Variables	Estimate	Scaled estimate	Std. Error (scaled)	t statistic	P-value
X_1	-0.02994	-0.92217	1.31481	0.701	0.483
X_2	1.07402	32.69305	2.00684	16.291	2e-16 ***
X_3	0.98543	43.81672	2.14946	20.385	2e-16 ***
X_4	0.98516	87.33978	2.16216	40.395	2e-16 ***

RP = 0.0007, chosen automatically, computed using 3 PCs, Variance = 3.988 , residual = 4.

III.3.3 Normalizing (Scaling) the Original Data

In this subsection, we will use the normalizing scale process:

$$\text{Scaled data} = \frac{x\text{-Min}(x)}{\text{Max}(x)\text{-Min}(x)}$$

After normalizing scale process, we get the LRM as:

$$y = -0.006898 - 0.004992x_1 + 0.197984x_2 + 0.260839x_3 + 0.570787x_4$$

The residuals are:

Residuals	Min	1Q	Median	3Q	Max
	-0.03707	-0.02267	-0.01002	0.01361	0.20308

We can test the LRM as below:

Variables	Estimate	Std. Error	t statistic	P-value
intercept	-0.006898	0.007102	-0.971	0.332
X_1	-0.004992	0.007627	-0.654	0.513
X_2	0.197984	0.015481	12.789	2e-16 ***
X_3	0.260839	0.014933	17.468	2e-16 ***
X_4	0.570787	0.014503	39.357	2e-16 ***

Residual standard error = 0.03312 , Multiple R-squared = 0.9578, Adjusted R-squared = 0.9575, F-statistic = 2810 , P-value = 2.2e-16.
DW = 1.8601, P-value = 0.05843.
We accept H_0. Errors are serially uncorrelated.
JB = 1376.2855, P-value = 2.2e-16.
We reject H_0. Errors are not normally distributed.

These results are indicated by Figure 9:

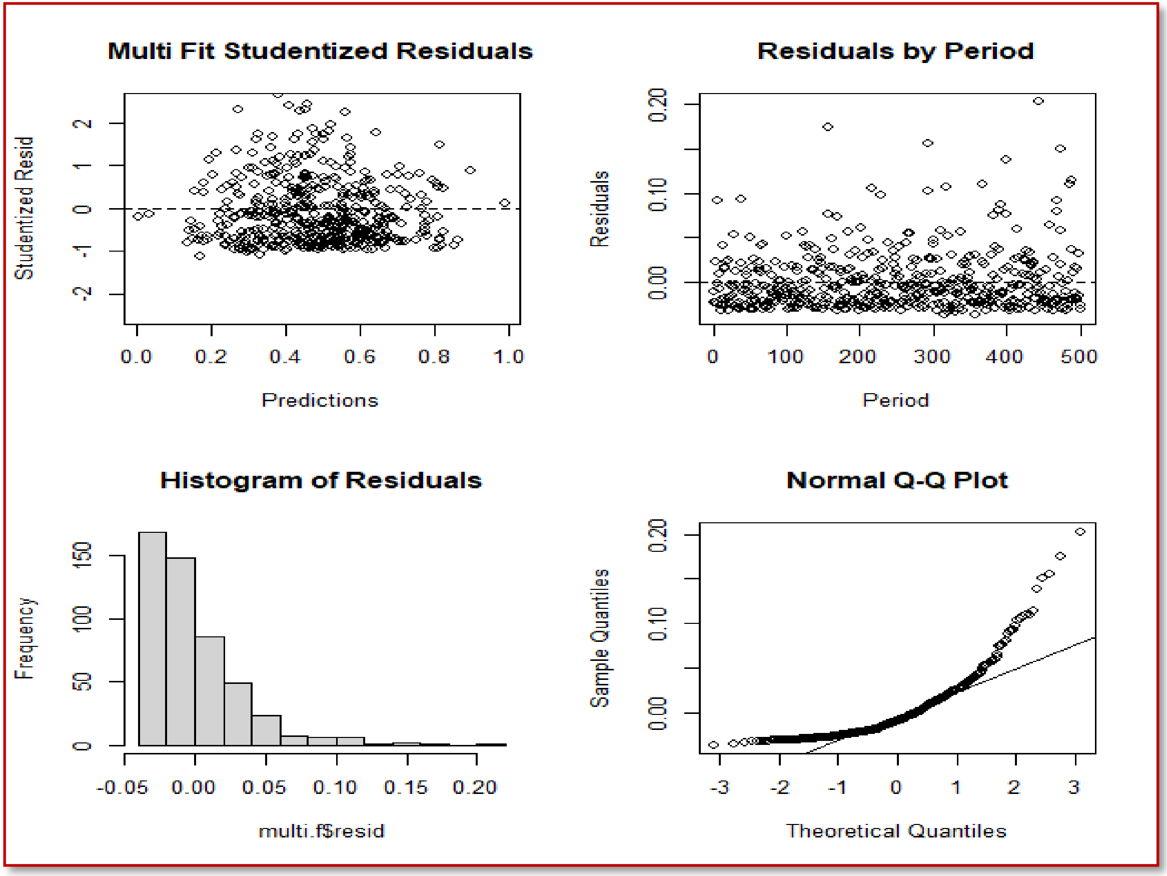

Figure 9: the plots of linearity case for normalizing scale data

Check for Multicollinearity

Parameter	VIF	Increased SE	Multicollinearity
X_1	1.01	1.00	Low
X_2	2.49	1.58	Low
X_3	3.08	1.76	Low
X_4	2.66	1.63	Low

These results are indicated by Figure 10:

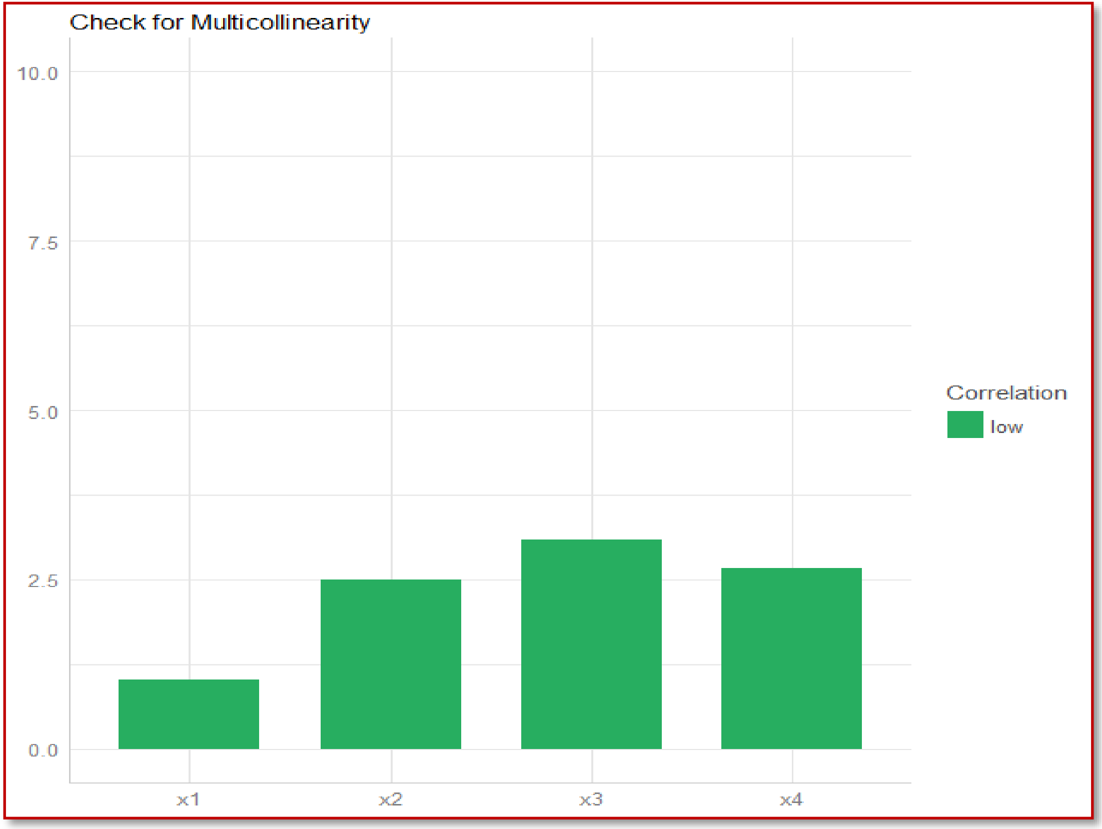

Figure 10:Check of multicollinearity of linearity case for normalizing scale data

The RRM using the "lm.ridge()" function with RP = 0.1 is:

$$y = -0.005257178x_1 + 0.109601471x_2 + 0.146630109x_3 + 0.262817013x_4$$

The RRM model using the "linearRidge" function is:

Variables	Estimate	Scaled estimate	Std. Error (scaled)	t statistic	P-value
X_1	-0.00501	-0.02183	0.03318	0.658	0.511
X_2	0.19821	0.66949	0.05213	12.843	2e-16 ***
X_3	0.26097	1.01587	0.05791	17.541	2e-16 ***
X_4	0.57018	2.12527	0.05388	39.445	2e-16 ***

RP = 0.0006, chosen automatically, computed using 3 PCs, Variance = 3.989 , residual = 4.

IV. DISCUSSION

In this section, we summarized the obtained results in section 3.

For the independence case:

F-test = 3.238, P-value=0.01225 is statistically significant (at 5%). This means that the LRM model has at least one variable (X_4) is significantly different from zero. Errors are serially uncorrelated. Errors are normally distributed. We checked Multicollinearity, we got Low collinearity, since, VIF = 1 and SE = 1 for all independent variables. With the function "Linear.Ridge()" RP = 0.3473546, chosen automatically, computed using 2PCs, Variance = 2.201 , residual = 3.733. This means that the RRM also has at least one variable(X_4) is significantly different from zero.

For the dependence case:

F-test = 0.4085, P-value=0.8026 is statistically not significant (at 5%). This means that the LRM has no one variable is significantly different from zero. Errors are serially uncorrelated. Errors are normally distributed. We checked Multicollinearity, we have Low collinearity Variance = 0.1269, residual = 0.8715.This means that the RRM also has no one variable significantly different from zero. From independence and dependence cases, we insure that: if RP is small value, then there is no need to use RRM, because the results do not change much. The presence of strong collinearity between the independent variables increases somewhat the VIF as well as RP. Also, the strong collinearity between the independent variable does not reflect Multicollinearity phenomenon.

For linear combinations case, we have three topics:

Before differencing process:
F-statistic = 1.208, P-value = 0.3065. This means that the LRM has no one variable significantly different from zero. Errors are serially uncorrelated. Errors are normally distributed. We checked Multicollinearity, we found that (X_1) has Low collinearity, (X_2) has Moderate collinearity, (X_3 , X_4) have High collinearity.

After differencing process:
F-statistic = 3341, P-value = 2.2e-16. This means that the LRM has three variables (X_2, X_3, X_4) are significantly different from zero. Errors are serially uncorrelated. Errors are not normally distributed. Check for Multicollinearity, we got Low collinearity, for all independent variables, after differencing. RP = 0.0007, chosen automatically, computed using 3 PCs, Variance = 3.988, Residual = 4.

After normalizing scale process:
F-statistic = 2810, P-value = 2.2e-16. This means also that the LRM has three variables (X_2, X_3, X_4) are significantly different from zero. Errors are serially uncorrelated. Errors are not normally distributed.Check for Multicollinearity become Low collinearity for all independent variables. RP=0.0006, chosen automatically, computed using 3 PCs, Variance = 3.989, Residual = 4.

V. Conclusions

Multicollinearity criteria occur when the independent variables in a regression model are correlated. This collinearity effects on the fitted results. In this paper, we tried to solve this problem using two methods. The first method used the ridge regression model (RRM), with eliminate the intercept of regression model. This model is compared with the traditional linear regression model (LRM). The second method modified the original dataset using differencing (the function "diffM" in "MTS" package), and scaling (the function "scale" in "base" package) processes. To make it, we supposed different cases to analysis the dataset. These cases are: independent case, dependent case, and combination linear case. In each case, we generated the dataset using R program from different distributions (univariate and multivariate). Sample size data consists of 500 observations that are classified to five variables (each one has 500 observations), four of them represented as independent variables (correlated or uncorrelated), the last one represented as dependent variable. For all cases, we fitted the data using LRM, and then check Multicollinearity with the function "check_collinearity()", with determining the variance inflation factor (VIF) for each independent variable. If Multicollinearity phenomenon is not absent, we try to remove it using the "lm.ridge" function in "MASS" package, or "Linear.Ridge" function in "ridge" package. Two functions give close results. The ridge parameter (RP) in the last one is chosen automatically. Then, we can re-check Multicollinearity again. From studying the independence and dependence cases, we insured from that: if the RP is small or moderate value, then there is no need to use the RRM, because of the results do not change much. The presence of strong collinearity between the independent variables increases somewhat the VIF as well as RP. Also, the strong collinearity between the independent variables does not reflect Multicollinearity phenomenon.Finally, when using the differencing or normalizing (scaling) dataset, for overcoming Multicollinearity, we found that these processes are effective ways to eliminate Multicollinearity and make the regression model results well, for estimates, variances, F-statistic or P-values.

Acknowledgement

Al-Obour High Institute for Management & Informatics.

References

[1]. S. Abubakari, Principal components to overcome multicollinearity problem, Oradea J. Busi. Econom. 4(1) (2019), 79-91.

[2]. R. B. Francoeur, Could sequential residual centering resolve low sensitivity in moderated regression? Simulations and cancer symptom clusters, Open J. Statist. 03(06) (2013), 24-44.

[3]. G. James, D. Witten, T. Hastie and R. Tibshirani, eds., An Introduction to Statistical Learning: with Applications in R, Springer, New York, 2013.

[4]. R. McElreath, Statistical Rethinking: A Bayesian Course with Examples in R and Stan, 2nd Edition, Chapman & Hall/CRC, 2020.

[5]. M. O'Brien, A Caution regarding rules of thumb for variance inflation factors, Quality & Quantity 41(5) (2007), 673-690.doi:10.1007/s11135-006-9018-6.

[6]. A. Goldberger, A Course in Econometrics, Harvard University Press, 1991, pp.248-250. ISBN 0-674-17544-1.

[7]. E. Farrar and R. Glauber, Multicollinearity in regression analysis: the problem revisited (PDF). Revi. Econom. Statist. 49(1) (1967), 92-107. doi:10.2307/1937887.

[8]. R. Wichers, The Detection of Multicollinearity: A Comment: The Review of Economics and Statistics, MIT Press, vol. 57(3)(1975), 366-368.

[9]. K. Kumar, Multicollinearity in regression analysis, Revi. Econom. Statist. 57(3) (1975), 365-366. doi:10.2307/1923925.

[10]. J. O'Hagan and B. McCabe, Tests for the severity of multicolinearity in regression analysis: a comment, Revi. Econom. Statist. 57(3) (1975), 368-370. doi:10.2307/1923927.

[11]. D. Belsley, Conditioning Diagnostics: Collinearity and Weak Data in Regression, Wiley, New York, 1991. ISBN 978-0-471-52889-0.

[12]. J. Hendrickx, Tools for Evaluating Collinearity, The package "perturb" v 2.10, 2019, R project.

[13]. S. Chatterjee, S. Hadi and B. Price, Regression Analysis by Example, Third Edition, John Wiley & Sons, 2000. ISBN 978-0-471-31946-7.

[14]. D. Gujarati, Multicollinearity: what happens if the regressors are correlated, Basic Econometrics, 4th edition, McGraw-Hill, 2009.

[15]. S. Lipovestky and M. Conklin, Analysis of regression in game theory approach, Appl. Stoch. Models in Busin. Indust. 17(4) (2001), 319-330. doi:10.1002/asmb.446.

[16]. D. Van Den Poel and B. Larivière, Customer attrition analysis for financial services using proportional hazard models, Europ. J. Operat. Res. 157 (2004), 196-217.doi:10.1016/S0377-2217(03)00069-9.

[17]. N. Kock and S. Lynn, Lateral collinearity and misleading results in variance-based SEM: an illustration and recommendations (PDF), J. Assoc. Informat. Syst. 13(7) (2012), 546-580. doi:10.17705/1jais.00302.

[18]. D. Schreiber-Gregory, Multicollinearity: What is It, Why should We Care, and How can It be Controlled? Henry M. Jackson Foundation for the Advancement of Military Medicine, Masters Graduate of National University, 2017. https://support.sas.com/resources/papers/proceedings17/1404-2017.pdf.

[19]. P. Allison, When can You Safely Ignore Multicollinearity? Statistical Horizons, 2012.

[20]. N. R. Draper and H. Smith, Applied Regression Analysis, 3rd Edition, Wiley, New York, 2003.

[21]. V. Dorugade and N. Kashid, Alternative method for choosing ridge parameter for regression, Appl. Math. Sci. 4(9) (2010), 447-456.

[22]. A. Ashok, Great learning blog: power ahead, Artificial Intelligence Machine Learning What is Ridge Regression?, 2020.

[23]. C. Montgomery, A. Peck and G. Vining, Introduction to Linear Regression Analysis, 3rd Edition, Wiley, New York, 2001.

[24]. R. Wicklin, Understanding Ridge Regression in SAS, 2012. http://blogs.sas.com/content/iml/2013/03/20/compute-ridge-regression.html.

[25]. A. Kumar, AI, BIG DATA, DATA SCIENCE, Data Analytics, 2014.

[26]. J. Brown, Measurement, Regression and Calibration Oxford, 1994. https://www.rdocumentation.org/packages/MASS/versions/7.3-53/topics/lm.ridge

[27]. S. Moritz, E. Cule, and M. De Iorio, Ridge Regression with Automatic Selection of the Penalty Parameter, ridge Package : Ridge Package, version 2.9, 2021 | R Documentation.

[28]. E. Cule and M. De Iorio, A Semi-automatic Method to Guide the Choice of Ridge Parameter in Ridge Regression, 2012. arXiv:1205.0686v1.

Ahmed Mohamed MohamedElsayed. "Using Different Methods to Overcome Modeling Problems." *IOSR Journal of Mathematics (IOSR-JM)*, 17(2), (2021): pp. 28-43.