Al-Qadisiyah Journal for Administrative and Economic Sciences ISSN$_{Online}$: 2312-9883
QJAE, Volume 25, Issue 3 (2023) ISSN$_{Print}$: 1816-9171

Robust Variance inflation factor for multiple linear regression having multicolinearity and outlier with application

Bahr Kadhim Mohammed
Asraa Khudair
Al-Qadisiyah University - College of Administration and Economics
Corresponding Author: Asraa Khudair

Abstract :In this paper, we proposed new estimation methods called (RVIF) to address the common problem of multicollinearity and outliers. For this we propose to use strong coefficient determination (RR2) instead of the classical R2 to obtain the strong variance inflation vector (RVIF). In order to evaluate the performance of the proposed method, we compared it with the existing methods. The results indicated that the proposed method has a high performance compared to other methods for all cases of sample size, linear relationship ratios, and contamination percentage

Keywords- Multicollinearity, outliers, VIF, robust VIF , Multiple linear regression model

INTRODUCTION: The regression model that contains more than one variable is called multiple regression, and the multiple regression method often depends on the estimates of the individual regression coefficients. In which the predictor variables are themselves strongly correlated, the multiple linear relationship is a case of multiple regression. Interference negatively affects the estimates (Smith 1974; Belsley 1984), to diagnose this type of problem there are several methods, including the VIF method, which is considered one of the most common methods for diagnosing multicollinearity as it refers to the amount of variance estimated for regression targeting normal or weighted least squares regression Basic . The most widely used diagnostic factor to detect *multicollinearity* in regression applications. Its function is to measure the inflation of variances in the regression coefficients when the predictors are not correlated. The method of treating the multiple correlation between the independent variables depends on the degree of this correlation, if the correlation is of a low degree, we can accept it, but in the case where the correlation is high, we must treat it through several methods, including that we increase the sample size, which leads to a decrease in the values of Standard errors of parameters, as well as using priori information about the relationship between independent variables, and reducing the number of independent variables whose correlation is high using the principal components analysis method, as this method converts variables into a smaller number called principal components and other methods that help us get rid of the high correlation between independent variables.

Therefore, a measure must be developed to diagnose the overlap in the linear relationship in a way that enables researchers to use the correct estimates to solve the problem of the multiple linear relationship. Therefore, we propose a strong diagnostic relationship, which is Robust Vif.

I. Linear regression model:

Modern statistics uses the linear regression (LR) model extensively in a wide range of applications. The multiple linear regression model is used when there are several independent variables .One of the often employed statistical techniques is the multiple linear regression model, which calculates the relationship between a quantitative variable and a limit, which is the dependent variable and a set of quantitative variables that are intended to be the independent variables. The multiple linear regression (MLR) model, Know as follows:

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i} \qquad \ldots(\mathbf{1})$$

The matrix notation for the model in (2.1) is given by

$$\mathbf{y = X\beta + \varepsilon} \qquad \ldots(\mathbf{2})$$

where y be an($n \times 1$)vector of response variable, X be an $n \times (p+1)$ matrix of independent variables, β be an $(p+1) \times 1$ vector of the unknown coefficients to be estimated and ε be an ($n \times 1$) random vector assumed to be independently identically distributed (*iid*) normal with constant variance and mean zero. As a matrs notations, the model in (2.2) can expressed as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & & x_{nj} & & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{bmatrix} \quad \dots \quad (3)$$

The least square (OLS) approach is a common applying to investigate regression coefficient of MLR model. The OLS technique goals to minimize the sum of squares of residuals, given by

$$\sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^t \boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \qquad \dots (4)$$

The solution for OLS estimators is given by,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \qquad \dots (5)$$

The variance of the estimates,

$$\mathbf{Var}[\widehat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1} \qquad \dots (6)$$

The model parameters of the regression are estimated using the OLS, which is one of the methods that should be used in regression analysis. Where this method has good characteristics and ease of calculation, which made it the most powerful and broadest method in the event that the conditions of the regression model are met. The ordinary least squares (OLS) coefficients are considered the best unbiased linear estimator if the error tearms are independent and symmetrically distribution, but despite these characteristics, it is not a robust estimation method in the case of unusual data as it has a very low breakdown point. One of the main assumptions of regression models, the independent variables should be not correlated. When this assumption is violated, we face the problem of Multicollinearity

II. Multicollinearity with Multiple linear regression model:

Multicollinearity is a condition where there is an approximately linear relationship among independent variables. When the Gaussian Markov hypotheses are satisfied, the OLS coefficients have best unbiased linear estimators. However, in the absence of independency among predictor variables, the OLS estimators will be imprecised and have large standard errors. The departure of independency in the model is also referred to as multicollinearity problem. This problem of multicollinearity occurs when the predicted variables are strongly correlated. Mathematically, the data set has the multicollinearity problem, if the design matrix (X^tX) is ill-conditioned (not invertible). Hoerl and Kennard (1970a) showed that a solution to the OLS does not always available and there is no unique solution when the matrix $(X^tX)^{-1}$does not exist. By considering the regression model given in Equation (1), for the p-predictor variables $x_1, x_2, \dots, x_p$perfect multicollinearity is said to exist if the following condition is achieved (Gujarati, 2003):

$$\mathbf{a_1x_1 + a_2x_2 + \cdots + a_px_p = 0} \qquad \dots \quad (7)$$

where $a_1, a_2, \dots, a_p$ are constants such that not all of them are zero simultaneously. However, we say imperfect multicollinearity exists if the following is achieved:

$$\mathbf{a_1x_1 + a_2x_2 + \cdots + a_px_p + \varepsilon_i = 0} \qquad \dots \quad (8)$$

where ε_i is a random error term. The preceding algebraic approach to multicollinearity can be described in Figure (1). In this figure, the degree of multicollinearity can be measured by the district of the overlapping (shaded area) of x_1 and x_2 circles. In strong relationship, if x_1 and x_2 were to overlap completely (or if $x_1$were completely inside x_2, or vice versa), multicollinearity would be perfect. This obviously leads to problems if (X^tX) is not reversible. Correspondingly, the variance of the estimates in equation (6) will explode in the case of a single (X^tX). If this matrix is not perfectly singular, but is close to being non-invertible, the variances are inflated.

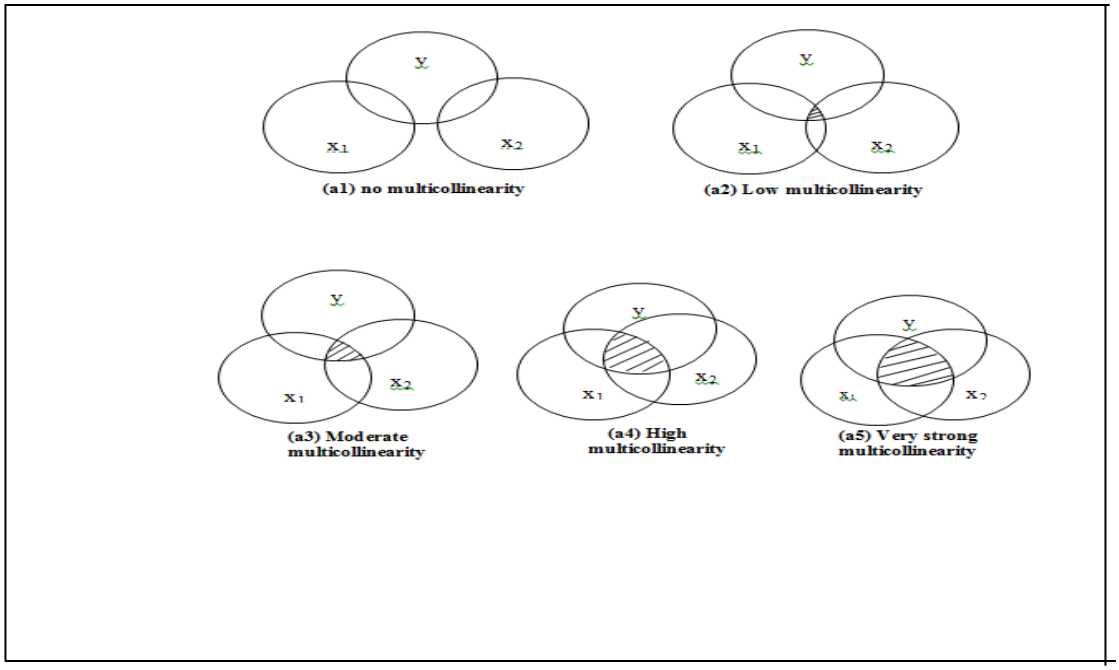

Figure 1: The Venn diagram of multicollinearity (Source: Gujarati, 2003)

III. Brief Overview of Outlier

Outliers are the observation that appears different from the rest of the other components of the sample and is inconsistent with other data. The location of this observation is far from the regression line and has a large amount of error compared to other natural observations, which greatly affects the linear model and its capabilities Beckman and Cook (1983) pointed out that "no observation can be guaranteed to constitute an entirely reliable manifestation of a phenomenon under examination". It is obvious that by identifying the "faulty" points in a dataset, one can gain a good understanding of the phenomenon being studied. However, it is much easier to spot inconsistent observations, which helps with inference and future predictions. Francis Bacon (1620) is credited

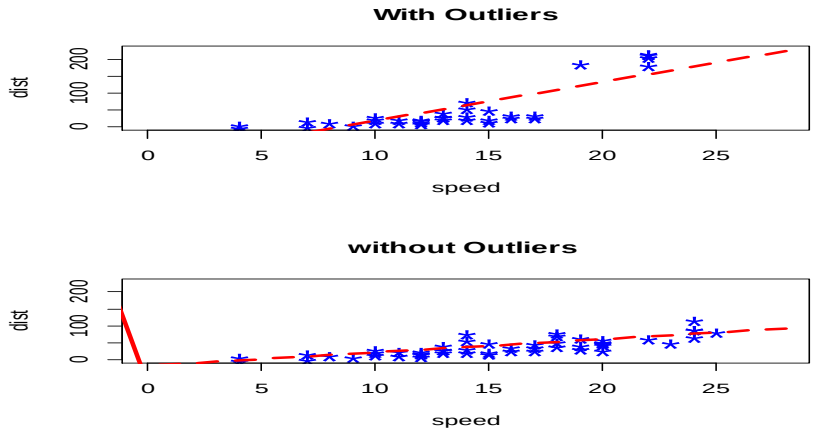

Figure (2): Shows the improvement of the fitting when the outliers are removed

with the following quote, which was derived from Billor, Hadi, and Velleman (2000): "Whoever understands the methods of Nature will more readily discern her flaws; while, on the other hand, whoever is aware of her deviations will be able to characterize her better".

Figure (1) shows the regression line is affected strongly by outliers. We see clearly that when we remove outliers, observe the change in the slope of the best fit line. If we had trained the model using the outliers (left chart), our predictions for higher values of speed would have been inflated (high error) due to the higher slope.

There are two types of values the first type is known as outliers, which is a set of observations called residual outliers based on the extent of failure of the regression equation that fits its absorption. The extreme value whose location in the independent variable is called the levering points, while the values related to the dependent variable are called the vertical deviation. There are several reasons for abnormal values to appear, including:

- The researcher may make a set of errors while monitoring the measurements or as a result of damage to the measurement equipment used in the experiments, as well as errors resulting from inaccuracy in the calculations, all of which lead to the emergence of abnormal values.
- Asymmetric data distribution, where the distributions are divided into two distributions, the first is affected by anomalous values, where the ends approach zero slowly, while the second type is a distribution that is absolutely resistant to anomalies and the end approaches zero faster than the first type.
- The data are divided based on the resulting distributions into a basic distribution that generates good observations and a polluted distribution that generates abnormal observations.

The high leverage points (HLPs) can control the behavior of majority of data and they may change the pattern of independency among the predictor variables. Therefore, the leverage points may enhance (collinearity enhancing observation) or reduce (collinearity reducing observations) the degree of multicollinearity problem in a data set (Hill, 1977; Sengupta and Bhimasankaran, 1977).

Moreover, in the presence of HLPs, most of the classical diagnostic methods such as the variance inflation factor (VIF) fail to correctly detect the multicollinearity problem. Statistical practitioners usually will employ the commonly used estimator such as the ridge regression and latent root regression when the classical VIF shows the existence of multicollinearity. However, when multicollinearity is caused by HLP, those estimators are not appropriate (Mason and Gunst, 1985; Walker, 1985; Walker, 1989). In this respect, it is very important to develop an appropriate diagnostic measure to diagnose multicollinearity so that statistics practitioners can employ the correct estimators to solve multicollinearity problems which depend on the source of multicollinearity.

IV. Diagnose of the Multicollinearity

With applying LR model, the true variables that used in the study may associated in nature. The multiple relationship happen when more than one of regressor is associated with each other, and this multiple leads to Multicollinearity problem. Multicollinearity do to decrease in the reliability of the statistical inferences and this creates a weakness of unique information for the LR model. As a solution of this problem, the problematic variables must be removed from the study. The common test for diagnostic the multicollinearity is the variance inflation factor (VIF). The VIF is a measurement of a power of relationship among the explanatory variables in the LR model. The VIF value that less than one indicate there is no correlation, whereas, if the VIF value is between 1-5, indicate that moderate correlation. The VIF value that greater than 5 indicate strong correlation among regressors. The high VIF value means high the possibility of mulicollinearity problem. With VIF that higher 10, there is a significant mulicollinearity which needs more consideration

The formula of VIF is:

$$\mathbf{VIF_i} = \frac{\mathbf{1}}{\mathbf{1 - R_i^2}} \qquad \dots \ (\mathbf{9})$$

Whare:

R_i^2 *=Unadjusted coefficient of determination for regressing that computed for* i^{th} *explanatory* variable on the remaining ones. *When* R_i^2 *is equal to 0 , the VIF is equal to 1indicates the multicollinearity does not exist*.

V. A New Robust Variance Inflation Factors

In the presence of outliers in x's say high leverage points (HLP's), the classical VIF (CVIF) fails to correctly detect the multicollinearity problem due to it dependency on classical coefficient of determination R^2 which are highly sensitive to high leverage points. In order to reduce this problem a robust diagnostic method can be used. However, the calculation of VIF is highly dependent on R^2, we suggest using robust coefficient determination (RR^2) rather than

classical R^2 to get robust variance inflation vector (RVIF) .The procedure for our proposed diagnostic multicollinearity method RVIF_MRCD is as follows

1- By using the robust regression (RR) estimation method which is explained previously, compute the coefficients of $RR(\hat{\beta}_{RR})$, as follows (Hadi, 1988; Imon, 1996, 2002, 2005)

$$\mathbf{r_i = z - x\,\hat{\beta}_{RR}}, \qquad \mathbf{i = 1, 2, \dots, n} \qquad \dots \qquad (\mathbf{10})$$

$$\hat{\sigma} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{r_i}}{\mathbf{n - p - 1}} \qquad \dots \qquad (\mathbf{11})$$

2- Calculate the robust R^2 based on JRR-MGM, namely $RR^2(RR)$, as follows.

$$RR^2(RR) = 1 - \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} w_{i(r)} r_{i(r)}^2}{\sum_{\mathbf{i=1}}^{\mathbf{n}} w_{i(r)} (z_i - \bar{z}_{(r)})^2} \qquad \dots \qquad (12)$$

where z_i an independent variable that regress with the rest of independent variables and $w_{i(r)}$ and $r_{i(r)}^2$ are the robust weights and residuals for RR respectively. $\bar{z}_{(r)}$ is the weighted average of z, calculated as

$$\bar{z}_{(r)} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} w_{i(r)} z_i}{\sum_{\mathbf{i=1}}^{\mathbf{n}} w_{i(r)}} \qquad \dots \qquad (13)$$

3- Finally, the robust VIF based on robust R^2 is formulated as

$$RVIF_j(RR\_mrcd) = \frac{1}{1 - RR^2_{j.(RR\_mrcd)}}, \qquad j = 1,2,\dots, p \qquad (14)$$

VI. simulation study

In this section, a Monte-Carlo simulation study is conducted in order to investigate the performance of our new proposed diagnostic multicollinearity measures RVIF. The multiple linear regression model in Equation (3.1) is considered and the predictor variables (x_1, x_2 and x_3) were generated according Equation (3.4). The ρ is the degree of correlation between $x's$ .Three values of correlation are consider ($\rho = 0.90, 0.95$ and 0.99), with four different sizes of samples (n =40, 70, 100 and 200). The contamination is done by replacing the clean data in x variable by a huge amount equal to 5 with different percentages of contamination ($\alpha = 0.5, 0.10$). Two types of data were generated, as follows;

1- Generate non-correlated data, with different percentage of high leverage points, where, the first 100(α) percent of observation for x_1 and x_2 are replaced by a huge amount equal to 5.
2- Generate correlated data. The correlation coefficient (ρ) between the predictors are chosen to be very high at 0.98. To generate high correlated data, the first 100(α) percent of observations for x_1 and x_2 have been replaced by a huge amount equal to 5.

Result and discussion

Tables 3.4 demonstrate the VIF values for non-correlated and correlated data without outliers. For non-correlated data, all VIF measures indicate that there is no multicollinearity in the data whereas, in correlated data, the CVIF can correctly indicate the multicollinearity in the dataset, while the VIF_MM method fails to detect the multicollinearity . Moreover, seen in Table 3.5 , for uncorrelated data, the CVIF immediately increases drastically when certain percentage of HLPs are induced in the data. This situation implies wrongly that there is a multicollinearity in the data which is caused by HLPs. Nonetheless, it is clearly seen that the RVIF_MM and our proposed diagnostic RVIF show

no multicollinearity problem in the dataset regardless of size of samples, percentage of contaminants and magnitude of contamination values.

In addition, it can be observed from Table 3.6 that for correlated data, when HLPs are added to the data, the CVIF and RVIF(MM) show wrongly no multicollinearity in the data as pointed by small values of *VIF* values that is less than 5. On the other hand, the RVIF still can identify correctly the multicollinearity problem is existing in the simulated data.

Table (1) vif for diagnostic method for clean data

		non-correlated data			correlated data		
		CVIF	RVIF-MM	RVIF-MRCD	CVIF-	RVIF-MM	RVIF-MRCD
20	X_1	1.4211	1.5936	1.9027	37.3159	5.7738	46.1255
	X_2	1.7616	1.2275	1.9595	38.6388	5.3424	51.3653
	X_3	1.9478	1.2478	1.9285	34.7268	5.4829	41.5433
30	X_1	1.7659	1.6552	1.7547	43.0484	6.2766	60.8514
	X_2	1.7706	1.4240	1.5267	46.8411	5.9258	62.3798
	X_3	1.3309	1.7795	1.2851	40.3900	5.9680	56.0223
50	X_1	1.2061	1.2829	1.3062	36.7208	5.0489	61.6731
	X_2	1.7553	1.4348	1.8112	33.7107	5.0614	58.4510
	X_3	1.0754	1.8077	1.4040	34.7445	5.2049	59.2158
100	X_1	1.4138	1.2946	1.1781	36.4896	5.8302	72.1547
	X_2	1.1635	1.0605	1.1790	38.0653	5.4489	74.9578
	X_3	1.4180	1.8317	1.9043	38.0674	6.2943	75.9331
200	X_1	1.6695	1.0230	1.3605	34.3334	5.9660	75.8075
	X_2	1.9242	1.3182	1.7054	34.1899	6.0192	71.3551
	X_3	1.0714	1.2797	1.8674	34.9318	5.8245	74.8087

Table (2) vif for diagnostic method for uncorrelated data with high leverage observation ($\alpha = 0.05$ and 0.10)

		$\alpha = 0.05$			$\alpha = 0.10$		
		CVIF	RVIF-MM	RVIF-MRCD	CVIF-	RVIF-MM	RVIF-MRCD
20	X_1	3948.88	1.2579	1.8962	3695.75	2.4506	1.8721
	X_2	3946.48	1.2537	2.1571	3699.18	2.8315	2.1666
	X_3	2.07	1.6631	2.1150	1.56	1.6242	2.1033
30	X_1	2657.42	1.5978	2.0133	6407.64	1.1800	1.2521
	X_2	2654.03	1.2267	2.2539	6405.99	1.1978	1.3237
	X_3	2.07	1.9074	2.2300	1.24	2.0506	1.4141
50	X_1	2603.91	1.8778	1.4478	6561.32	1.4226	2.1138
	X_2	2603.80	1.8990	1.2535	6562.50	1.6525	1.5135
	X_3	1.69	1.8349	1.8365	1.94	1.7994	1.7144
100	X_1	3104.00	1.3829	1.4350	6158.92	1.1789	1.8058
	X_2	3104.32	1.7288	1.9669	6158.09	1.1674	1.8885
	X_3	1.44	1.2875	1.6759	1.31	1.0381	1.7313
200	X_1	3048.68	1.3007	1.9682	6164.81	1.5457	1.1650
	X_2	3048.78	1.9374	1.8814	6163.46	1.9371	2.0331
	X_3	1.34	1.3388	1.1586	1.05	1.1641	1.2741

Table (3) vif for diagnostic method for correlated data with high leverage observation ($\alpha = 0.05$ and 0.10)

		$\alpha = 0.05$			$\alpha = 0.10$		
		CVIF	RVIF-MM	RVIF-MRCD	CVIF-	RVIF-MM	RVIF-MRCD
20	X_1	1.7197	4.3536	2.0077	1.8887	4.3786	1.4849
	X_2	1.8567	3.5614	2.1741	1.5965	3.9101	1.6755
	X_3	2.0064	3.7573	35.0953	1.8372	3.9576	28.0656
30	X_1	1.5220	3.9088	2.1214	1.3575	4.3905	1.6014
	X_2	1.0781	3.7343	1.9630	1.4916	4.5328	1.9729
	X_3	1.8905	3.8930	27.4267	1.4000	4.2816	31.7377
50	X_1	1.9420	4.2182	2.5819	1.8349	4.2055	1.7437
	X_2	1.6811	4.8518	2.4543	1.7747	3.9966	2.2774
	X_3	1.6174	4.4343	30.1851	1.7582	3.7955	29.0799
100	X_1	1.5096	4.6828	1.9366	1.0948	4.0583	1.4750
	X_2	1.6817	4.1195	2.0827	1.2783	4.2524	1.8834
	X_3	1.8106	4.6424	34.1520	1.9894	3.9243	31.5179
200	X_1	1.8395	4.2525	2.0828	1.3843	4.4740	2.3057
	X_2	1.7234	4.2973	2.4785	1.3955	4.2232	2.0724
	X_3	1.9683	4.7971	35.5766	1.1365	4.4551	37.0133

VII.

VIII. R

eal Data

Thalassemia is a genetic blood disorder that is passed from parents to children. It arises due to mutations in The DNA of the cells responsible for the production of hemoglobin (a substance in red blood cells responsible for transporting oxygen and nutrients to cells throughout the body, and for getting rid of waste and carbon dioxide), which leads to a decrease in its percentage in the body from the normal rate that affects .Thalassemia is a serious life-threatening disease that negatively affects the functions of other organs. It may cause serious complications if it is not treated and controlled, so it is called fatal anemia.

x_1: Age - Most thalassemia patients live until the age of 30-25 years, and some of them can live until the age of 60

x_2: Marital status - Thalassemia results from the marriage of two carriers of the disease's genes, which leads to the infection of their children with thalassemia of all kinds.

x_3: Blood group - The blood groups were divided according to the presence or absence of antibodies on the surface of the red blood cells, and the blood groups are inherited from the parents to the children, so the children will inherit one of the blood groups.

x_4: Weight - If the body fails to produce enough of these proteins, it affects the development of red blood cells and leads to anemia that begins in early childhood and continues throughout life. Which makes the patient lose weight.

x_5: Hemoglobin - Iron overload. People with plasmemia get too much iron, either because of the disease or because of frequent blood transfusions. Too much iron in your body can damage your heart, liver, and endocrine system, which contains the hormone-secreting glands that regulate your entire body's operations.

x_6: Residence (countryside–city) - The environment may affect the incidence of diseases or the weakness of the defensive cells in the blood, and therefore we find it an influential factor that may cause many problems.

x_7: Consanguinity - Thalassemia results from a defect in the genes that affect the production of hemoglobin. A child cannot be born with thalassemia unless they inherit these defective genes from both parents.

Figure (1) shows that there is HLPs in the x_1 and x_2, this leads to false correlation (0.923) between x_1 and x_2 as shown in Figure (2) When we remove the HLPs from x_1 and x_2 as shown in Figure (3) we see clearly there is no correlation

(0.169) between these variable as shown in Figure (4). In addition, Figures (3) and (4) there is a high correlated between x_6 and x_7 in the original data set (between Residence and Consanguinity). To investigate the performance of diagnose of multicollinearity, we compute the VIF and R^2 for method of study. Table (4) present the VIF and R^2 for methods. The results show that the CVIF identify wrongly there is a high correlation between x_1 and x_2 due to existing HLPs, whereas the RVIF-MM and RVIF declare correctly there is no correlation between x_1 and x_2 for original dataset. On the other hand, the CVIF fails to identify collinearity between x_6 and x_7 while RVIF-MM and RVIF declare correctly there is high correlation between x_1 and x_2 for original dataset.

Table (4) VIF values for diagnostic methods for real data

Variables	Diagnostic Methods					
	CVIF		RVIF (MM)		RVIF(MRCD)	
	VIF	R^2	RVIF	R^2	RVIF	R^2
x_1	11.128	0.910	1.066	0.062	1.218	0.177
x_2	12.071	0.917	1.275	0.216	1.133	0.116
x_3	1.113	0.100	1.112	0.100	1.164	0.141
x_4	1.068	0.064	1.068	0.064	1.193	0.162
x_5	1.041	0.040	1.126	0.112	0.986	-0.014
x_6	3.277	0.522	12.195	0.918	11.363	0.915
x_7	3.438	0.557	12.345	0.919	11.627	0.914

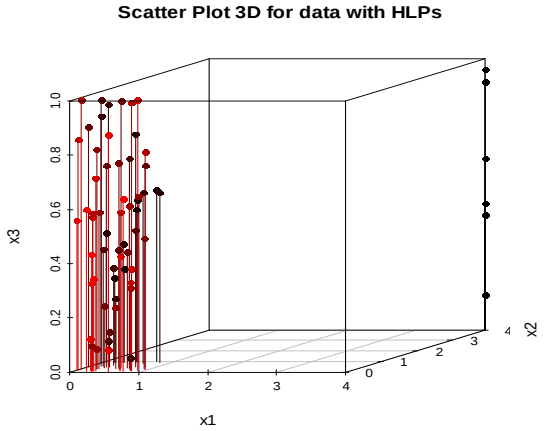

Figure (1) The scatter plot for Thalassemia dataset (original data)

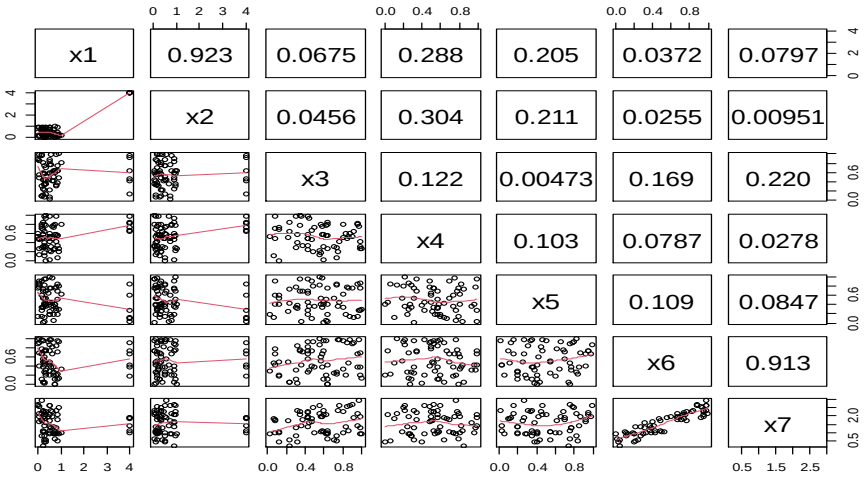

Figure (2) The correlation matrix for Thalassemia dataset (original data)

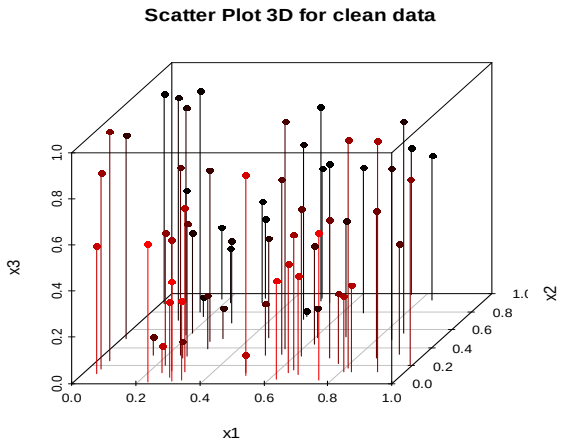

Figure (3) The 3-dimantion scatter plot for clean data (after remove the HLPs)

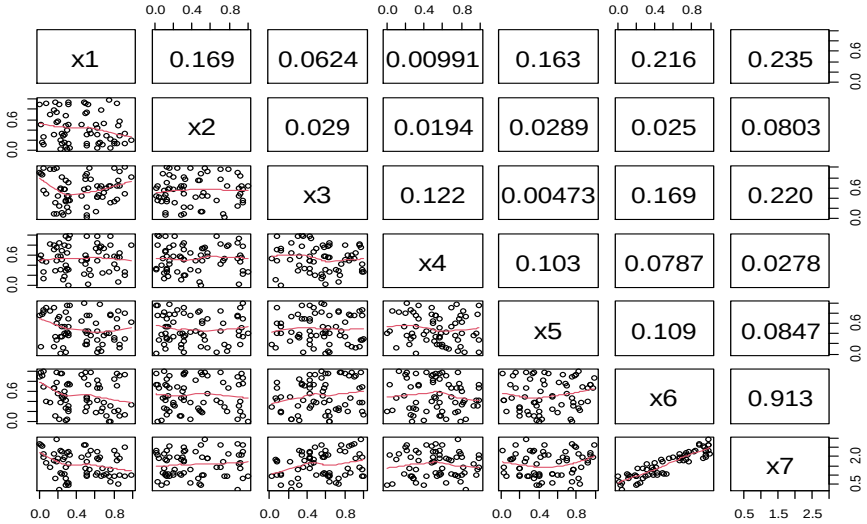

Figure (4) The correlation matrix for clean data (after remove the outliers)

IX. Concussion

In this study, we proposed new estimation methods called (RVIF) to address the common problem of multiple linearity and outliers. For this we propose to use strong coefficient determination (RR2) instead of the classical R2 to obtain the strong variance inflation vector (RVIF). In order to evaluate the performance of the proposed method, we compared it with the existing methods. The results indicate that the proposed method has superior performance compared to other methods for all cases of sample size, linear relationship ratios, and contamination percentage.

X. References

- H

oerl, A. E. and Kennard, R. W. (1970a). Ridge regression: applications to non-orthogonal problems. *Technometrics*. 12:69-82.

- H

ill, R. W. (1977). *Robust Regression When There Are Outliers in the Carriers*. Unpublished Ph.D. thesis. Harvard University, Boston, MA.

- Hill, R. W. and Holland, P. W. (1977). Two robust alternatives to robust regression. *Journal of the American Statistical Association.* 72: 828–833.
- Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York :Wiley.
- Beckman, R. J., & Cook, R. D. (1983). Outlier……. s. *Technometrics*, *25*(2), 119-149.
- Mason, R. L., & Gunst, R. F. (1985). Outlier-induced collinearities. *Technometrics*, *27*(4), 401-407.
- Walker, E. (1985). *Influence, Collinearity and Robust Estimation in Regression*. Unpublished Ph.D. thesis, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.
- Hadi, A.S. (1988). Diagnosing collineariy-influential observations. *Computational Statistics and Data Analysis*. 7:143-159.
- Walker, E. (1989). *Detection of collinearity-influential observations*. Communications in Statistics-Theory and Methodology.18:1675-1690.
- Billor, N., Hadi, A. S., & Velleman, P. F. (2000). BACON: blocked adaptive computationally efficient outlier nominators. *Computational statistics & data analysis*, *34*(3), 279-298.
- Gujarati, D.N. (2002). *Basic Econometrics*. 4th edition. New York: Macgraw-Hill
- Berry, W. D., Feldman, S., & Stanley Feldman, D. (1985). *Multiple regression in practice* (No. 50). Sage.
- Corlett, W. (1990). Multicollinearity. In *Econometrics* (pp. 158-159). Palgrave Macmillan, London.
- Belsley, D. A. (1991). A guide to using the collinearity diagnostics. *Computer Science in Economics and Management*, *4*(1), 33-50.
- Freund, R. J., Littell, R. C., & Creighton, L. (2003). *Regression using JMP*. SAS Institute.
- Stechman, R., & Allen, R. (2005). History of Ramjet Propulsion Development at the Marquardt Company-1944 to 1970. In *41st AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit* (p. 3538).
- Paul, R. K. (2006). Multicollinearity: Causes, effects and remedies. *IASRI*, *New Delhi*, *1*(1), 58-65.
- Slinker, B. K., & Glantz, S. A. (2008). Multiple linear regression: accounting for multiple simultaneous determinants of a continuous dependent variable. *Circulation*, *117*(13), 1732-1737
- Robinson, C., & Schumacker, R. E. (2009). Interaction effects: centering, variance inflation factor, and interpretation issues. Multiple linear regression viewpoints, 35(1), 6-11.
- Alin, A. (2010). Multicollinearity. *Wiley interdisciplinary reviews: computational statistics*, *2*(3), 370-374.
- Schirrmeister, B. E., Antonelli, A., & Bagheri, H. C. (2011). The origin of multicellularity in cyanobacteria. *BMC evolutionary biology*, *11*(1), 1-21.
- Liao, D., & Valliant, R. (2012). Variance inflation factors in the analysis of complex survey data. *Survey Methodology*, *38*(1), 53-62.
- Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, *8*(2), 39-42.
- Ekiz, O. U. (2021). An improved robust variance inflation factor: Reducing the negative effects of good leverage points. *Kuwait Journal of Science*.