JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 967 - 974
https://publishoa.com
ISSN: 1309-3452

Investigating Multicollinearity in Factors Affecting Number of Born Children in Iraq

Salisu Ibrahim[1, a*)], Mowafaq Muhammed Al-Kassab [1, b)], Muhammed Qasim Al-Awjar [2, c)]

[1] Department of Mathematics Education, Tishk International University-Erbil, Kurdistan Region, Iraq
[2] Department of Statistics and Informatics, College of Computers and Mathematics, University of Mosul, Iraq.

[a*]salisu.ibrahim@tiu.edu.iq, ibrahimsalisu46@yahoo.com
[a*]https://orcid.org/ **0000-0002-1467-5426**
[b]mowafaq.muhammed@tiu.edu.iq
[b*]https://orcid.org/ **0000-0002-9471-528x**
[c]mqy.alawjar@uomosul.edu.iq

ABSTRACT

The occurrence of multicollinearity in several multiple regression models leads to major problems that can affect the entire multiple regression model outcomes, among the problems are a reduction in the precision of the estimated coefficients, which decreases the statistical power of the model. The effect of sensitivity on the estimated coefficients is due to a small swing in the model. This paper considers the two fundamental approaches for identifying multicollinearity. The first approach is the correlation coefficient (CC) and the second one is the variance inflation factor (VIF). The ridge regression method, principal components regression, intent root regression, and weighted regression are advanced regression models for investigating the existence of multicollinearity, these findings would tackle, reduce, and fixed the multicollinearity among the independent variables, and help to predict the best-fitted model. Lastly, we came up with the best-fitted model.

Keywords: Multiple regression, multicollinearity, correlation coefficient, variance inflation factor Smoking mother.

1. Introduction

Multicollinearity occurs when explanatory variables in a regression model are correlated. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results [1]. A key goal of regression analysis is to isolate the relationship between each independent variable and the dependent variable [2]. Multicollinearity makes it hard to interpret your coefficients, and it reduces the power of your model to identify independent variables that are statistically significant. These are serious problems. However, the good news is that you don't always have to find a way to fix multicollinearity [3]. Several studies examined and discussed the problems of multicollinearity for the regression model and also emphasized that the major problem related to multicollinearity comprises uneven and biased standard errors and impractical explanations of the results [4, 5, 6].

In this paper, we considered the correlation coefficient (CC) and the variance inflation factor (VIF) approaches for identifying the multicollinearity amongst the independent variables, in the year 2015. Multiple regression is considered for the prediction of the best models. Based on the results, we discovered that there is multicollinearity among the factors, these necessitate the use of CC and the VIF approaches to tackle, reduce, and fixed the multicollinearity among the independent variables. Lastly, we came up with the best-fitted model. This paper is scheduled as: Section 2 provides the methods for investigating multicollinearity. The results and diagnosed multicollinearity are presented in Section 3 and Section 4, respectively. The conclusion follows in Section 5.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 967 - 974
https://publishoa.com
ISSN: 1309-3452

2. Materials and Methods for Investigating Multicollinearity

In this section, we present the materials and methods used for investigating the multicollinearity within the independent variables. The dataset was selected at random from 100 women records, moreover, the dataset used in this study is collected from the Babil Governorate health center [7]. The independent variables (IVs) are, husband age, mother weight, the mother age, years of marriage, smoking mother, number of dead children, the mother age when married, number of sports hours per week, the mother with the thyroid gland, the mother sleeping hours per week, the mother taking medicine, breastfeeding duration per month, and mother job, while the dependent variable (DV) is the number of born children. Other factors like financial assistance, chronic illness (breast cancer), stress due to job, illegal drug, and house activities can be among the leading risk factors affecting the number of born children. These factors lead to serious health conditions that make one vulnerable to covid 19, see [8]. When it comes to the application perspective, the authors in [9, 10, 11] make use of commutativity to study the relation and the sensitivity between systems, the idea can be extended to investigates the commutativity and sensitivity between the independent variables, The main aim of this research is to investigate multicollinearity using some techniques such as i) correlation coefficient and ii) variance inflation factor.

2.1. Correlation Coefficient

The Pearson's correlation coefficient (also called Pearson's *R*) is a relationship coefficient regularly utilized in direct relapse. The formula of the Pearson correlation coefficient is given as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{1}$$

where n is a sample size, r is the correlation coefficient, y_i and x_i are dependent and independent variables indexed in i respectively. If the correlation coefficient value is higher with the pairwise variables, it indicates the possibility of collinearity. In general, if the absolute value of the Pearson correlation coefficient is close to 0.8, collinearity is likely to exist [12].

2.2. Variance Inflation Factor (VIF)

Variance Inflation Factor (VIF) is a simple way to detect multicollinearity in a regression model, it is used to determine the correlation between independent variables. The VIF measures how much the variance is inflated. VIF is calculated as

$$VIF_j = \frac{1}{1 - R_{ij}^2} = \frac{1}{Tolerance}. \tag{2}$$

Please observe that the higher the tolerance, the lower the VIF, and the limited possibility for multicollinearity among the variables. The VIF with the value of 1 clearly shows that there is no correlation between the independent variables. But if the VIF has a value within 1< VIF < 5, it suggest that there is a moderate correlation between the variables, with VIF between $5 \leq$ VIF ≤ 10, it indicates multicollinearity that needs corrective action and VIF > 10 are indications of severe correlation between the variables, with critical levels of the multicollinearity [13].

2.3. Multiple linear regression

The multiple linear regression model is given as.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 967 - 974
https://publishoa.com
ISSN: 1309-3452

$$\sum_{j=1}^{13} \beta_0 + \beta_j x_{ij} + e_i. \quad (3)$$

where β_0, β_i are the unknown constants, x_i are the IVs, y is the DV and e_i is the error term that has a normal distribution with mean o and variance σ^2. The mother age (x_1), the mother age when married (x_2), mother weight (x_3), smoking mother (x_4), husband age (x_5), years of marriage (x_6), number of dead children (x_7), number of sports hours per week (x_8), the mother with the thyroid gland (x_9), mother sleeping hours per week (x_{10}), the mother taking marriage (x_{11}), breastfeeding duration per week (x_{12}), and mother job (x_{13}) are the IVs and also the number of born child (y) is the DV.

3. Results

The author in [14] discusses some primary techniques for detecting multicollinearity using the questionnaire survey data on customer satisfaction. In this section, we statistically detect the multicollinearity among the independent variables using the correlation coefficient method in Eq . (1), VIF in Eq. (2), and lastly with the help of multiple linear regression in Eq. (3).

3. 1. Investigating Multicollinearity Using Pairwise Scatterplot

The scatterplot is one of the methods used for detecting multicollinearity by observing the relationship between the variables**.** The dots depicted in Figure 1 represents the values of two variables.

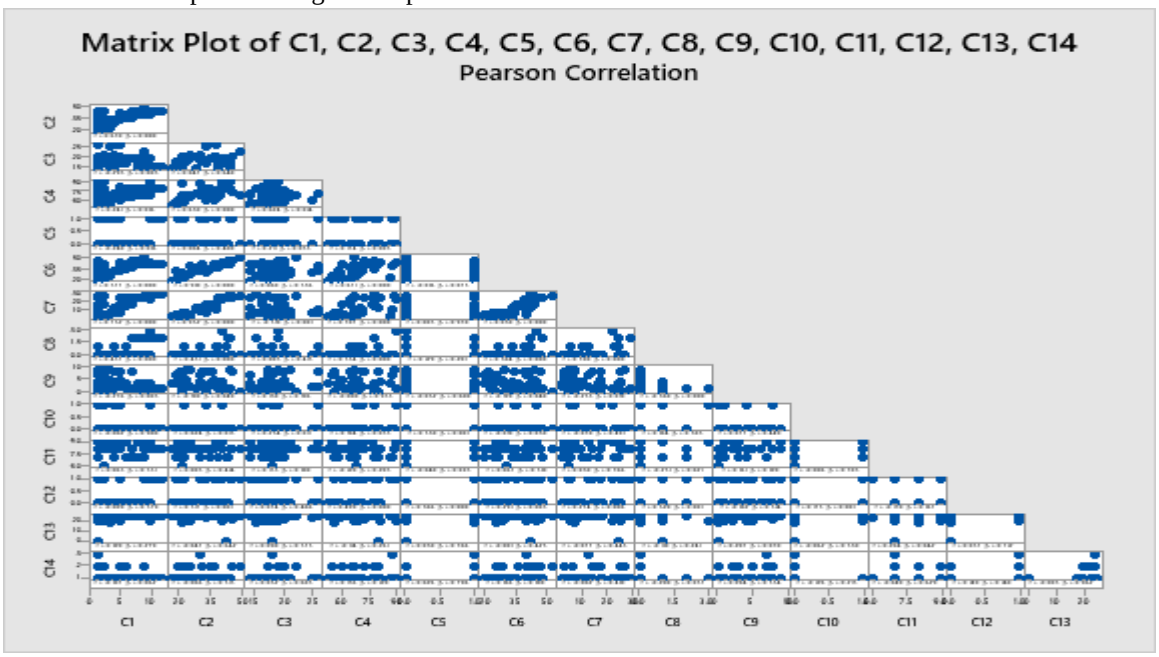

Figure 1. Scatterplot of pairwise variables

3. 2. Investigating Multicollinearity Using Pearson's Correlations

Pearson's correlations are very important method used to investigate collinearity between the independent variables. Table 2 shows the relationships in terms of collinearity between the independent variables. The results obtained from

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 967 - 974
https://publishoa.com
ISSN: 1309-3452

the overall correlation detected the collinearity between the variables, the most highly correlated variables are (x_1), (x_3), (x_5), and (x_6). The mother age (x_1) versus mother weight (x_3) has [r = 0.638, c.f = (0.505, 0.741), p < 0.05], the mother age (x_1) versus husband age (x_5) has no logical relation, the mother age (x_1) versus years of marriage (x_6) has [r = 0.932, c.f = (0.9, 0.954), p < 0.05], the mother weight (x_3) versus husband age (x_5) has no logical relation, the mother weight (x_3) versus years of marriage (x_6) has [r = 0.597, c.f = (0.451, 0.7101), p < 0.05], and husband age (x_5) versus years of marriage (x_6) has [r = 0.850, c.f = (0.784, 0.897), p < 0.05]. The Pearson correlation coefficient is close to 0.8, this shows the existence of collinearity between the variables.

Table 1: Descriptive Statistics.

Variable	N	Mean	SE Mean	Median	Mode
y	100	3.570	0.299	2.000	1
x_1	100	31.030	0.838	30.000	22
x_2	100	18.640	0.321	18.500	19
x_3	100	68.520	0.894	67.000	67
x_4	100	0.3800	0.0488	0.0000	0
x_5	100	34.170	0.829	33.500	42
x_6	100	12.390	0.883	9.500	3
x_7	100	0.3400	0.0655	0.0000	0
x_8	100	3.850	0.291	3.000	2
x_9	100	0.0900	0.0288	0.0000	0
x_{10}	100	8.0900	0.0740	8.0000	8
x_{11}	100	0.3000	0.0461	0.0000	0
x_{12}	100	23.210	0.276	24.000	24
x_{13}	100	1.1700	0.0428	1.0000	1

Table 2: Pearson's Correlations Coefficients

Variables	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
x_1	0.659												
x_2	-0.293	0.047											
x_3	0.241	0.638	0.024										
x_4	-0.240	0.084	0.211	0.174									
x_5	0.577	0.918	0.060	0.671	-0.016								
x_6	0.732	0.932	-0.319	0.597	0.003	0.850							
x_7	0.437	0.451	0.083	0.354	0.129	0.384	0.398						
x_8	-0.276	-0.198	0.130	-0.008	-0.052	-0.199	-0.235	-0.360					
x_9	-0.002	0.024	0.254	0.194	0.330	-0.019	-0.070	0.104	0.077				
x_{10}	0.063	0.083	0.133	-0.128	-0.040	0.067	0.030	-0.231	0.161	-0.086			
x_{11}	-0.089	0.317	0.074	0.499	0.566	0.291	0.274	0.328	-0.102	0.175	-0.139		
x_{12}	0.109	-0.047	0.090	-0.114	0.030	-0.081	-0.077	-0.118	0.207	-0.062	0.204	0.037	
x_{13}	-0.187	-0.066	0.052	0.156	0.026	0.166	-0.082	-0.208	0.094	-0.126	-0.049	0.149	-0.005

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 967 - 974
https://publishoa.com
ISSN: 1309-3452

The model is given as

$$y = -1.03 + 0.294\,x_1 - 0.338\,x_2 - 0.0614\,x_3 - 1.04\,x_4 - 0.0310x_5 + 1.38\,x_7 - 0.0904\,x_8 + 2.19\,x_9 + 0.165\,x_{10} - 1.75\,x_{11} + 0.246\,x_{12} + 0.521\,x_{13}.$$

The overall significance of the model is given in table (3).

Table: 3: Analysis of Variance

Model	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	12	697.9	58.154	27.40	0.000
Residual Error	87	184.656	2.122		
Total	99	882.5			

3. 3. Investigating Multicollinearity Using Variance Inflation Factor (VIF)

The variance inflation factor (VIF) identifies the correlation between independent variables and the strength of that correlation. The regression analysis illustrated in Table 4 detected multicollinearity by identifying variables with p-value > 0.05 and VIF > 5. These results show that the mother age (x_1), the mother age when married (x_2), mother weight (x_3), smoking mother (x_4), years of marriage (x_6), number of dead children (x_7), the mother with the thyroid gland (x_9), mother sleeping hours per week (x_{10}), and mother job (x_{13}) are statistically significant while husband age (x_5), number of sports hours per week (x_8), the mother taking marriage (x_{11}), breastfeeding duration per week (x_{12}) are not statistically significant. Moreover, the model indicates that the mother age (x_1) and husband age (x_5) has the highest VIFs of 10.8 and 11.7 respectively. This indicates serious multicollinearity that requires removal.

Table 4: Regression Analysis

Predictor	**Coef**	**SE Coef**	**T-Value**	**P-Value**	**VIF**
Constant	**-1.030**	**2.658**	**-0.39**	**0.699**	
β_1	**0.29400**	**0.05732**	**5.13**	**0.000**	**10.8**
β_2	**-0.33848**	**0.05007**	**-6.76**	**0.000**	**1.2**
β_3	**-0.06140**	**0.02669**	**-2.30**	**0.024**	**2.7**
β_4	**-1.0369**	**0.4119**	**-2.52**	**0.014**	**1.9**
β_5	**-0.03101**	**0.06052**	**-0.51**	**0.610**	**11.7**
β_7	**1.3775**	**0.2919**	**4.72**	**0.000**	**1.7**
β_8	**-0.09044**	**0.05865**	**-1.54**	**0.127**	**1.4**
β_9	**2.1872**	**0.5818**	**3.76**	**0.000**	**1.3**
β_{10}	**0.1649**	**0.2215**	**0.74**	**0.459**	**1.3**
β_{11}	**-1.7490**	**0.4640**	**-3.77**	**0.000**	**2.1**
β_{12}	**0.24604**	**0.05639**	**4.36**	**0.000**	**1.1**
β_{13}	**0.5215**	**0.4549**	**1.15**	**0.255**	**1.8**

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 967 - 974
https://publishoa.com
ISSN: 1309-3452

The R-square is 79 %.

4. Diagnosed Multicollinearity

There are several methods to remove multicollinearity, the authors in [15, 16] studied the application of latent roots regression to multicollinear data, but in this research, we will consider i) removal of variables with high VIF and ii) removing non-significant variables.

4. 1. Diagnosed Multicollinearity by Removing High VIF

In our model, the mother age (x_1) and husband age (x_5) has the highest VIFs of 10.8 and 11.7 respectively. The correlation between the mother age (x_1) and husband age (x_5) is significant with r = 0.918, see Table 2. So instead of removal both of them, we keep the mother age (x_1) with a VIF of 10.8 and remove the husband age (x_5) with VIF 11.7, we obtained a new model in Table 6. We can see that all the VIFs are down to satisfactory values with (VIFs < 5). The model is given as

$$y = -0.86 + 0.2678\,x_1 - 0.3415x_2 - 0.0648\,x_3 - 0.974\,x_4 + 1.379\,x_7 - 0.085\,x_8 + 2.186x_9 + 0.157x_{10} - 1.745\,x_{11} + 0.2477x_{12} + 0.394\,x_{13}$$

The overall significance of the model is given in table (5).

Table: 5: Analysis of Variance Table and Overall Significant of the Model

Model	**DF**	**Adj SS**	**Adj MS**	**F-Value**	**P-Value**
Regression	11	697.296	63.391	30.12	0.000
Residual Error	88	185.214	2.105		
Total	99	882.510			

Table 6: Regression Analysis

Term	**Coef**	**SE Coef**	**T-Value**	**P-Value**	**VIF**
Constant	-0.86	2.63	-0.33	0.745	
$\boldsymbol{\beta_1}$	0.2678	0.0259	10.34	0.000	2.22
$\boldsymbol{\beta_2}$	-0.3415	0.0495	-6.90	0.000	1.19
$\boldsymbol{\beta_3}$	-0.0648	0.0257	-2.52	0.014	2.49
$\boldsymbol{\beta_4}$	-0.974	0.391	-2.49	0.015	1.72
$\boldsymbol{\beta_7}$	1.379	0.291	4.74	0.000	1.70
$\boldsymbol{\beta_8}$	-0.0850	0.0574	-1.48	0.143	1.31
$\boldsymbol{\beta_9}$	2.186	0.579	3.77	0.000	1.31
$\boldsymbol{\beta_{10}}$	0.157	0.220	0.71	0.478	1.25
$\boldsymbol{\beta_{11}}$	-1.745	0.462	-3.78	0.000	2.13
$\boldsymbol{\beta_{12}}$	0.2477	0.0561	4.42	0.000	1.12
$\boldsymbol{\beta_{13}}$	0.394	0.379	1.04	0.302	1.24

The R-square is 79 %.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 967 - 974
https://publishoa.com
ISSN: 1309-3452

4. 2. Diagnosed Multicollinearity by Removing Non-Significant Variables

Removing the husband age (x_5) with VIF 11.7 is not enough to predict the best model since, we still have some variables such as number of sports hours per week (x_8), mother sleeping hours per week (x_{10}), and mother job (x_{13}) that are not statistically significant in Table 7. This necessitates the removal of this variable. We can see that after removing the non-significant variables, the p-values of all the variables are down to satisfactory values with ($p < 0.05$) in Table 8. The model is given as

$$y = 0.89 + 0.2746\,x_1 - 0.3407x_2 - 0.0690\,x_3 - 0.950\,x_4 + 1.382\,x_7 + 1.985x_9 - 1.661\,x_{11} + 0.2349x_{12}$$

The overall significance of the model is given in table (7).

Table: 7: Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	8	689.73	86.216	40.70	0.000
Residual Error	91	192.78	2.118		
Total	99	882.51			

Table 8: Regression Analysis

Term	Coef	SE Coef	T-Value	value	VIF
Constant	0.89	2.10	0.42	0.673	
$\boldsymbol{\beta_1}$	0.2746	0.0242	11.32	0.000	1.93
$\boldsymbol{\beta_2}$	-0.3407	0.0481	-7.08	0.000	1.12
$\boldsymbol{\beta_3}$	-0.0690	0.0242	-2.85	0.005	2.18
$\boldsymbol{\beta_4}$	-0.950	0.392	-2.42	0.017	1.71
$\boldsymbol{\beta_7}$	1.382	0.260	5.32	0.000	1.35
$\boldsymbol{\beta_9}$	1.985	0.566	3.50	0.001	1.24
$\boldsymbol{\beta_{11}}$	-1.661	0.457	-3.63	0.000	2.07
$\boldsymbol{\beta_{12}}$	0.2349	0.0545	4.31	0.000	1.06

The R-square is 78 %.

5. Conclusion

This paper investigates the multicollinearity relation among the independent variables, the mother age, the mother age when married, husband age, mother weight, years of marriage, smoking mother, number of sports hours per week, number of dead children, the mother with the thyroid gland, the mother sleeping hours per week, the mother taking medicine, breastfeeding duration per month, and mother job, the model obtained proves to be not significant since some variables have *p* less than 0.05. These were as a result of multicollinearity among the variables. The two methods; correlation coefficient and variance inflation factor proposed in this work were used to detect the multicollinearity among the variables. Among several methods to removed collinearity, we consider two methods; removing variables with high VIF and removing variables that are not statistically significant ($p < 0.05$). Lastly, we obtained the best-fitted model that predicts the factors affecting the number of born children in Iraq. Moreover, the Anova obtained

from table 7 shows that the model is more fitted since we observed a monotone increment in the f-value, from 27.40 in table 3 to 40.70 in table 7. Furthermore, more advanced research techniques such as the ridge regression method, latent root regression, weighted regression method, and principal components regression can be used to detect collinearity [17, 18]. The results are validated with Minitab version 19.

Funding: No funding.

References.

1. Young, D.S., Handbook of regression methods, CRC Press, Boca Raton, FL, 2017, 109-136.
2. Frank, E.H. Jr. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis, Springer, New York, 2001, 121-142.
3. Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X., *Applied Logistic Regression,* John Wiley & Sons, New Jersey, 2013.
4. Pedhajur, E.J., Multiple regression in behavioral research*: explanation and prediction (3rd edition),* Thomson Learning, Wadsworth, USA, 1997.
5. Keith, T.Z., Multiple regression and beyond: An introduction to multiple regression and structural equation modeling (2nd edition), Taylor and Francis, New York, 2015.
6. Aiken, L.S. and West, S.G., Multiple regression: Testing and interpreting interactions, Sage, Newbury Park, 1991.
7. Majid, S., Alsabah, S. Parameters Estimation of the Multiple Linear Regression
The mode under Multicollinearity problem. Iraqi Academic Scientific Journals. 12(1), pp 1-28, (2020).
8. Ibrahim, S., and Al-Kassab, M.M., Using Linear Regression Analysis to Study the Recovery Cases of COVID 19 in Erbil, Kurdistan Region. *Drugs and Cell Therapies in Hematology, 10(1), 1226-1239,* (2021).
9. Ibrahim, S., and M. E. Koksal, "Commutativity of sixth-order time-varying linear systems," *Circuits, Systems, and Signal Processing*, vol. 40, no. 10, pp. 4799–4832, 2021.View at: Publisher Site | Google Scholar
10. S. Ibrahim and M. E. Koksal, "Realization of a fourth-order linear time-varying differential system with nonzero initial conditions by cascaded two second-order commutative pairs," *Circuits, Systems, and Signal Processing*, vol. 40, no. 6, pp. 3107–3123, 2021.View at: Publisher Site | Google Scholar
11. Salisu Ibrahim, Abedallah Rababah, "Decomposition of Fourth-Order Euler-Type Linear Time-Varying Differential System into Cascaded Two Second-Order Euler Commutative Pairs", *Complexity*, vol. 2022, ArticleID 3690019, 9 pages, 2022. https://doi.org/10.1155/2022/3690019.
12. Gunst, R.F. and Webster, J.T., "Regression analysis and problems of multicollinearity," *Communications in Statistics*, 4 (3). 277-292. 1975. Gunst, R.F. and Webster, J.T., "Regression analysis and problems of multicollinearity," *Communications in Statistics*, 4 (3). 277-292. 1975.
13. Belsley, D.A., Conditioning diagnostics*:* Collinearity and weak data in regression, John Wiley & Sons, Inc., New York, 1991.
14. Noora Shrestha**.** Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and Statistics, 2020, Vol. 8, No. 2, 39-42.*
15. Al-Kassab, M.M., Adnan, M.A., Dilnas, S.Y. Studying the Effect of Some Variables on the economic Growth Using Latent Roots Method. (11), pp 1-10, (2019).
16. Al-kassab, M.M., Dilnas S.Y. Application of Latent Roots Regression to Multicollinear Data. Journal of Advanced Research in Computer Science & Engineering. 4(12), pp 1-11., (2017).
17. Ibrahim, S.: 'Numerical Approximation Method for Solving Differential Equations'. *Eurasian Journal of Science and Engineering*. 6(2): 157-168, (2020). Doi: 10.23918/eajse.v6i2p157
18. Rababah, A., Ibrahim, S.: 'Weighted G^1-Multi-Degree Reduction of Bézier Curves'. International Journal of Advanced Computer Science and Applications, 7(2): 540-545, (2016).