MODEL VIOLATION AND MULTICOLINEARITY: A PRELIMINARY STUDY

Reuben Friday Osemeke[1,*], John Nwabueze Igabari [2], Nwabenu Dominic Christian[3]

[1]*Department of Mathematics and Statistics,*
University of Delta, Delta State Nigeria.

[2] *Department of Mathematics,*
Delta State University, Delta State Nigeria.

[3]*Department of Mathematics and Statistics,*
Ogwashi-Uku Delta State Polytechnic. Nigeria

**Corresponding Author Email: osemekereuben@gmail.com*

Abstract

Sometimes regression results are non-optimal, biased estimates, increase in Type 1 error rates, poor standard errors, untrustworthy confidence intervals, and insignificance F and T Test. This is as a result of assumption violations emanating from multicollinearity among the independent variables. The variance inflation factor (VIF) of predictors are appropriately high (X1= 50.603) and (X2= 49.286). They are above the threshold value of 10+. The bivariate scatter plots shows evidence of 45 degree linear relationship between Trademark(X_1) and Bread Texture(X_2). The Bivariate Correlation suggests that a strong correlation exist between X1 and X2 of 0.990 (r= 0.990) exceeding the set requirement of 0.8+. The above diagnostics are signs of multicollinearity present in the linear model. Furthermore, the significance value for each of the independent variable is greater 0.05 level of significance showing evidence of multicollinearity and non-normality in the model. The multicollinearity detention gave birth to other violations hidden in the model. A new diagnostics measures called adjusted coefficient of determination($Adjr^2$) instead of ordinary coefficient of determination(r^2) were use in the computation of variance inflation factor. However, in all trial transformations, increase in sample size$(n \leq 69)$ with linear and quadratic features seems very satisfactorily because its variance inflation factor (VIF) of each independent variable was less than 2.500 specified-justified thresholds. These are also accompanied with satisfaction of four classical regressions, optimal statistical power, feasible partial regression, and residuals plots as well as regression statistics been improved upon.Therefore, the variance inflation factor of 2.500 below should be used a as a stopping rule to optimal models, predictive power ability, strong statistical power and all assumptions satisfaction.

Keywords: *Model violation, Multicollinearity, Preliminary study, Transformations, Classical Assumption, Other assumption violations.*

Introduction

Janah & Kartini(2022) asserted that multiple linear regression analysis is a statistical techniques with multiple independent variables $(x_1, x_2, x_3, \dots \dots .. x_p)$. Hermawan (2021) said that regression analysis is an analysis whose aim is to determine whether they exist a statistically significant relationship between the dependent variable and the set of independent variables. Another point worth mentioning according to Sartika et al (2020) is that the purpose of multiple regression analysis is to predict the effects of relationship between two or more independent variables on one response variable. Ningsih et al(2019) said that multiple linear regression analysis is an equation expressed in mathematical form between set of independent variables and the response variable.

The most frequently employed model to express the influence of several predictors $(x_1, x_2, x_3, \dots \dots .. x_p)$ on a continuous outcome variable(y_i) is the linear regression model which is mathematically defined by:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + e_i \qquad (1)$$

The model in equation (1) is optimal when it satisfies the regression assumption of linearity, normality, independence and constant variance of the error term and partial regression plots. The model is term optimal, predictive and has significant improvement in the regression coefficients with positive tendency.

The challenge that often arises in multiple regression diagnostics is the problem arising from violations of four classical regression assumptions of normality, homoscedasticity, independence, and linearity. The model violation of multicollinearity in the multiple linear regression models gave birth to other violations. Multicollinearity is the birth rock or the preliminary study to other violations hidden in the model. Sartika et al (2020) said that multicollinearity problems occur when predictor variable and other predictor variables are correlated with each other. Senaviranta & Cooray(2019) is of the opinion that in research data, sometimes multicollinearity data is accidentally found where the independent variables are related to each other, so it becomes a serious problem in analysing research data .

In addition, if there is multicollinearity present among the independent variables, the estimated regression model parameters using the least square method will produce an unbiased estimator with a large variance**.** In a linear regression analysis, multicollinearity refers to the presence of a significant correlation between two or more independent variables. In view of the above, Lavery et al (2019) said that this make it difficult to interpret the predicted regression coefficients effectively when employing the least square method since they become unstable.

Literature has it that multicollinearity can be detected using correlation coefficient(CC) between pairwise variables, tolerance value(TV), and high correlation(HC), high coefficient of determination(HCD) between the pairwise independent variables, variance inflation factor

(VIF), conditional index,(CI) and bivariate scatter plots(BSC) with forty five degree linear relationship. Bayman & Dexter (2021) asserted that multicollinearity can be done informally with linear correlation coefficients between independent variables, or in a formal way with the variance inflation factor (VIF).

Consequently, violation of classical regression assumptions due to multicollinearity present causes various types of problematic situations to the linear model. Firstly, coefficients estimates may become poorly estimated, that is not estimating the true value on average. Secondly, F-test and T-test can lead to increase in Type 1 error rates or decrease in statistical power. Andrew (2019) suggested that the rejection of null hypotheses too often lead to rejection of varying variations, standard errors of estimated coefficients and results to statistically insignificant coefficients. Thirdly, confidence intervals might become untrustworthy, becoming too narrow or too wide. Auqino et al (2019) in another effects said that multicollinearity can cause standard error of the regression parameters to increase. Adeboye et al (2015) said that multicollinearity affects standard error of the regression coefficients, so the estimation results may not be accurate. The effect of multicollinearity obscures the interpretation of structure of the model equation. In real sense, it affects the statistical power of coefficient of determination, and adjusted coefficient of determination

However, to eliminate the problem arising from multicollinearity Nyrhinen & Leskine(2014) said that correlated independent variable's should be removed from the model. In the least square method, the main objective is to the find the regression coefficient value that best estimates the value of the dependent variable from the independent variables. Obite et al(2020) asserted that when there is multicollinearity between the independent variables, the least square method cannot produce accurate coefficients values.

Finally, to overcome the problem of multicollinearity in our regression model, several feasible transformation are tried until the variance inflation factor(VIF) for each of the given independent variables are each less than 2.500 specified benchmark. The variance Inflation factor less than 2.500 for each of the independent variables is predictive and has the following optimal features.

(i) Removing the problem of multicollinearity in the multiple linear regression model

(ii) Satisfies the four classical regression assumptions.

(iii) The scatter plots of residuals and partial regression plots will be even distributed around the mean of zero and standard deviation of 1.

(iv) The Durbin Watson test statistic of residuals will be around specified benchmark of 1.5 to 2.5 to show independence of the observation.

(v) The graph of probability to probability plots (P-P) will follow a 45 degree linearity pattern along the diagonal line of (0, 0) to (1, 1).

(vi) The residual scatter plots against the predicted value will be cluster around zero (0) mean and evenly distributed without outliers.

In the light of the above, Weaving et al(2019) said that the results of linear regression model can becomes more accurate and meaningful in explaining the relationship between the dependent variable and independent variables.

Statement of Problem

Apart from the purpose of making predictions, regression can be used to find a model. Before the fitted linear multiple model is used for other things or prediction purposes, the set of given independent variables, the dependent variable and analysis of residuals should be explode thoroughly to check for the existence of inherent non-optimal regression assumptions emanating from multicollinearity. The problem of multicollinearity sheds bad light on fitted linear regression model and adversely provides negative insights to other model assumptions of non-normality, heterosceadasticity, serial correlation, curvilinear effects and non-independence of the observations.

Furthermore, this research work seems to use various statistical tools and graphical plots to detect multicollinearity but its creates more emphasizes on the variance inflation factor(VIF) which is used to detect multicollinearity violations and used as a great guide to stop trial and error transformation with a specified-justified value of 2.500. The set of independent variables below 2.500 of variance inflations factor (VIF) seems very optimal, highly predictive, and satisfies the four classical regression assumptions and should be used for future research purpose

Aims and Objectives of the Study

The general objective of this research is to investigate and analyse the detentions of violations of linear model assumptions of multicollinearity by means of independent variables and residuals through its statistical- graphical test and providing feasible transformation with variance inflation factor (VIF) of specified and justified- benchmark of 2.500 with a view to remove multicollinearity and other model violations present.

The specific objectives of this study are to:

1. Establish evidence of bivariate scatter plots of 45 degree linear relationship between the independent variables and evidence of high bivariate independent correlation matrix of 0.8+.
2. Establish the given significance value of each of the independent variables greater than 0.05 level of significance and establishing high conditional index of 15 to 30 and above
3. Detecting other model assumption violations of serial correlation, non-constant variance and non-normality with its significance in comparing with 0.05 level of significance and specified benchmark.
4. Providing trial and feasible transformation to remove multicolinearity problem and satisfies the model assumptions of the four classical regression assumption’s
5. Use variance inflation factor (VIF) greater than 10+ to detect multicollinearity in each independent variable.
6. Use variance inflation factor (VIF) below 2.500 as the best satisfactory transformation of optimal regression models and satisfaction of classical regression assumptions

Materials and Methods

Several methods of detecting multicollinearity had been proposed in literature. They are the correlation coefficient(R), correlation matrix diagnostics (CMD), the pairwise scatter plots (PSP), variance Inflationary factor (VIF), tolerance value (TV), Eigen value (EV), condition indices (CI), F test and a host of others not mentioned. This research work will examined the optimality and efficiency of each of the statistical test and bivariate graphical plots use to detect multicollinearity in the multivariate linear model.

Preliminary study

Auqino et al(2019) said that one of the methods of estimating multiple linear regression parameters is the least square methods. In general, the multiple linear regression models is mathematically defined as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} +, \dots, + \beta_p x_{pi} + \varepsilon_i \quad \text{For i = 1, 2,…..,j} \qquad \textbf{(2)}$$

Where $y_i$the observations of the dependent variable is, $x_{1i,}x_{2i,} \dots x_{pi}$ is the values of the independent variables, β_o is the constant value, $\beta_{1,}\ \beta_{2,}\ \beta_{3,} \dots \dots . . \beta_p$ is the regression parameters and ϵ_i represent the standard errors. Multiple fitted linear regression models (least square method) is used to estimate the regression coefficients by minimizing the sum of squared errors. The ordinary least square estimators of β is obtained using

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \widehat{\beta_2} x_{2i} + \cdots \dots + \widehat{\beta_p} x_{pi} + \epsilon_i \quad \text{For i = 1, 2, 3,…, n} \qquad \textbf{(3)}$$

For instance, short linear model of equation (3) is express as

$$Y = X\hat{\beta} + e \qquad (4)$$

In vector and matrix approach, we have

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & . & . & x_{1p} \\ 1 & x_2 & . & . & x_{2p} \\ . & . & . & . & . \\ \vdots & & & & \\ 1 & x_n & . & . & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \qquad (5)$$

The vector Y is the n*1 observation vector, X matrix is called n*(p+1) design matrix, β is called (p+1)* 1 unknown parameter vector and e is the n*1 error vector. The model makes certain assumptions to enable the analyst perform certain important statistical test. These assumptions are:

(i) Normality of the error term distribution
(ii) Independence of the error terms
(iii) Constant variance of the error terms
(iv) Linearity of the error terms

Residuals are used to check for violations of these assumptions mostly multicollinearity and for other diagnostics checks. In other words to check model accuracy

Instead of writing out the n equations, using matrix notation, our simple linear regression function reduces to a short and simple statement

$$Y = X\hat{\beta} + e \quad \textbf{(6)}$$

Making e the subject gives

$$e = Y - X\hat{\beta} \quad (7)$$

Therefore $e^T e = \left(Y - X\hat{\beta}\right)^T \left(Y - X\hat{\beta}\right)^T$ (8)

$$e^T e = Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\beta \quad (9)$$

From Equation (6) we can be differentiate $e^T e$ with respect to $\hat{\beta}$

To have

$$\frac{\delta e^T e}{\delta \hat{\beta}} = \frac{\delta Y^T Y - 2\hat{\beta}^T Y + \hat{\beta}^T X^T X\beta)}{\delta \hat{\beta}} \quad (10)$$

This is finally decomposed to give

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (11)$$

The present of multicollinearity reduces each of the regression coefficients of predictor variables, and reduces the adjusted coefficient of determination on the dependent variable. This paper focus on using a new method called the adjusted coefficient of determination instead of the ordinary coefficient of determination as used in literature as a new method of multicollinearity diagnoses.

Pairwise Visual Assessment and Correlation Matrix Diagnostics

Gujarati (2003) is of the opinion that multicollinearity can be diagnosed by calculating the correlation coefficient among the given set of independent variables. Young (2017) is of the opinion that correlation coefficient between the pairwise variables greater than 0.80 and above, shows sign of multicollinearity. In addition, the pairwise visual assessment between independent variables shows evidence of multicollinearity present if the graph shows existence of 45 degree linear relationship. This method is seemed easy but quite subjective.

The correlation coefficient emanating from pairwise correlation matrix is the most prominent method as used in SPSS to determine the present of collinear variables. The assigning of variables for correlation coefficient are

$$ni * \sum x_i y_i = \text{Re} \quad (12)$$

$$\left(\sum x_i\right)\left(\sum y_i\right) = \text{Rm} \quad (13)$$

$$n_i \sum x^2 - \left(\sum x\right) \sum x = \text{Rk} \quad (14)$$

$$n_i \sum y^2 - (\sum y_i) \sum y_i = \text{Rp} \tag{15}$$

This is formulated into a single mathematical definition of correlation coefficient to get

$$r = \frac{Re - Rm}{Rk - Rp} \tag{16}$$

Variance Inflationary Factor (VIF)

Variance inflationary factor (VIF) arises when some of the independent variables in a multiple settings are highly correlated with some independent variable. In such situations, collinear variables do not provide new information, and it becomes difficult to separate the effects of such variables on the dependent variable. Rawling et al (1998) and Levery et al (2019) defines VIF as

$$VIF = \frac{1}{1-R_j^2} = \frac{1}{tolerance_{R_j^2}} \tag{17}$$

Where R_j^2 is the coefficient of multiple determinations? The Tolerance value is simply the inverse of VIF. VIF greater than 5 to 10 above shows very high correlated variables.

$$VIF_{adj} = \frac{1}{1 - R_{adjusted}^2} = \frac{1}{tolerance_{adj}} \tag{18}$$

Where $R_{adjusted}^2$ is the adjusted coefficient of determination?

Equation (18) has a better predictive power because the adjusted coefficient of determination select only significant optimal variables into the model than the ordinary coefficient of determination which increases as the sample size increases, so it is better to use the adjusted coefficient of determination as a better method for determining the variation of the independent variables over the dependent variable.

Furthermore, Belsey (2018) and Weaving et al (2019) is of the opinion that VIF>10 is an indication that the regression coefficients are feebly estimates with the presence of multicollinearity. Adeboye et al (2015) further said that VIF far more than 10 indicates a relatively high level of multicollinearity in the regression model.

Tabl1 1: VIF Presentation and Interpretation

VIF value	**Decision**	**Conclusion**
VIF = 1	Not Correlated	No Multicollinearity present
$1 < VIF \leq 5$	Moderately Correlated	Partial Multicollinearity
VIF>5	High Correlated variables	Full Present of Multicollinearity
VIF>10(More than 10)	100% Highly Correlated variables	100% Full present of Multillinearity

Belsley et al (1980) and Cohen et al (2003) proposed that the variance inflation factor (VIF) is one of the most popular measures of detecting multicollinearity, although several other diagnostics are available.

ANOVA F statistic: An Insight Guide to Multicollinearity Present

The ANOVA f test is used to test for the overall significance of multiple independent variables; its null and alternative hypotheses are as follows

H_0: Estimated regression model are non-normal: An insight guide to multicollinearity present
H_1: Estimated regression model is normal: No multicollinearity present
F test statistic for the entire regression model is

$$F = \frac{MSR}{MSE} \tag{19}$$

Where MSR = Mean Square Regression, MSE = Mean Square Error.

The decision is to rejects the null hypotheses if the f test calculated is greater than F critical. The significance value(P-value) can be used to assess the stated hypotheses

Eigen value: A Statistical Tool for Detecting Multicollinearity

The sum of the eigenvalue (λ) must be equal to the number of independent variables. A very small Eigen value, approximately close to 0.05 is signs of multicollinearity. Sometimes, it is difficult to interpret the eigenvalue close to 0.05, so conditional index can be an alternative. Moreover, condition index or conditional number of the correlation can be used to measure the overall multicollinearity of the variables. The conditional index (C1) is a function of eigenvalues.

$$C1 = \frac{\lambda_{\max}}{\lambda_{\min}} \tag{20}$$

The value of C1 is always greater than one (1), so a higher value of C1 indicates the present of multicollinearity in the multiple regression models.

Generally, C1<15 is an indication of weak multicollinearity. 15<C1<30 is evidence of moderate multicollinearity. Young (2017) and Knoke et al (2002) says that condition index greater than (>) 30 is an indication of high multicollinearity.

Hidden possible indicators of Multicollinearity in Multivariate linear Model

It is also possible that the adjusted R^2 for a model is pretty good and even the overall – F Test statistic is also significant but some of the individual coefficients are statistically insignificant. This scenario can be possible indication of the presence of multicollinearity as multicollinearity affects the coefficients and corresponding P-values, but it does not affect the goodness of fit statistics or the overall model significance.

The Coefficient of Determination (r^2) and Adjusted r^2 Criterion

Coefficient of determination ($\mathbf{r^2}$) is a statistic that gives some information about the goodness of fit of the model. In a regression, $\mathbf{r^2}$ is a statistical measure of how well the regression line approximates the real data points. Everitt(2002) defines $\mathbf{r^2}$ as a coefficient of determination. $\mathbf{r^2}$ is the square of the correlation between two variables. Nagelkere(1991) indicates that $\mathbf{r^2}$ in the range of -0 to 1 perfectly fits the regression line.

The model with the highest value of $\mathbf{r^2}$ provides the closet fit. However the major drawback of $\mathbf{r^2}$ is that as the model increased, $\mathbf{r^2}$ goes high whether the extra variable provides any important information about the dependent variable(y) or not. Therefore, it makes no sense to define the best model as the model with the largest $\mathbf{r^2}$ value. A common way to avoid this problem is to use the adjusted version of $\mathbf{r^2}$ instead of the ordinary $\mathbf{r^2}$ itself. The adjusted $\mathbf{r^2}$ statistic for a model with P independent variables is given by

$$r_{adj}^2 = \left[1 - r_{y.n.......p}^2 \frac{n-1}{n-p-1}\right] \quad \textbf{(21)}$$

where P denotes the number of independent variables in the regression equation and $r_{y.n....p}^2$ denotes the coefficient of determination from full regression model.

The $\mathbf{r^2}$ adjusted does not necessarily increase when the numbers of predictor variables increase. It increased when the data has significant effects on the model.

Solution to the Problem of Multicoliinearity among the Independent Variables

(i) Linearly combine the independent variables by adding them together or eliminating one of the variable.

(ii) Increase sample size of both the dependent and independent variables to a desirable level with a linear and quadratic tendency. Vatcheva et al (2016) proposed that increasing the sample size reduces the standard errors of regression coefficients and decreases the degree of multicollinearity between or among the in dependent variables.

(iii) Simply drop one of the correlated predictor. From a practical point of view, there is no point in keeping two(2) very similar predictors in our model

Results and Discussion

Data Presentations

A Bread bakery is interested in establishing the effects of Trademark(X1), Bread Texture(X2) and Bread aroma(X3) on Consumers Attitude(Y). Twenty three customers were sampled, and their ratings are presented in Table 2.

Table 2: Consumer s Attitude Survey

N	Y	X1	X2	X3
1	**6.52**	**26.02**	**26.02**	**25.02**
2	**7.92**	**20.02**	**16.02**	**52.02**
3	**6.82**	**16.02**	**10.02**	**29.02**
4	**6.62**	**17.02**	**12.02**	**28.02**
.	**.**	**.**	**.**	**.**

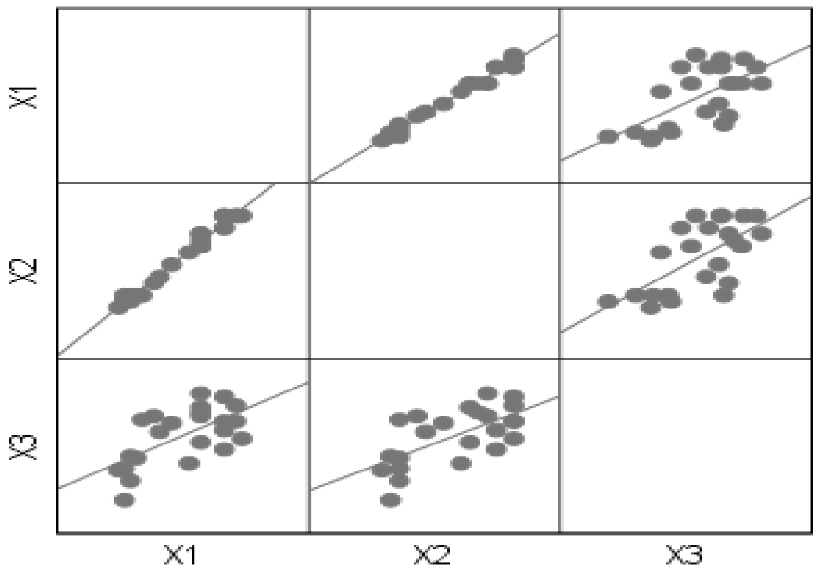
.	.	.	.	.
.	.	.	.	.
20	**6.82**	**35.02**	**38.02**	**39.02**
21	**6.52**	**15.02**	**12.02**	**22.02**
22	**8.62**	**28.02**	**32.02**	**65.02**
23	**8.62**	**32.02**	**38.02**	**49.02**

Table 3: **Bivariate Correlation Matrix**

	Y	**X1**	**X2**	**X3**
Y	1	.366	.367	.511**
X1	.366	1	.990*	.635**
X2	.367	.990*	1	.647**
X3	.511*	.635**	.647**	1

The Bivariate Correlation of Table 3 suggests that a strong correlation exist between Trademark(X1) and Bread Texture(X2) of 0.990 (r= 0.990) as mark * exceeding the previously set requirement of 0.8. This is a sign of multicollinearity present in the data set.

Figure 2: Bivariate Scatter Plots of Predictor variables



The matrix scatter plots of Figure 2 shows evidence of bivariate scatter plots of 45 degree linear relationship between Trademark(X_1) and Texture(X_2) which shows evidence of multicollinearity present.

Table 4: Regression Coefficients

	Coefficients	**Stand Error**	**T statistic**	**Sig value**	**Tolerance**	**VIF**	**Remark**
(constant)	5.691	1.454	3.915	**.001**			
X1	.028	.152	.185	**.855>0.05**	.020	50.603	Multicolliearity
X2	-.014	.101	-.137	**.892>0.05**	.579	1.726	Multicollinearity
X3	.023	.013	1.827	**.083>0.05**	.020	49.286	Multicollinearity

In Table 4, the variance inflationary factor (VIF) of Trademark(X1) and Bread Contentment(X3) are 50.603 and 49.286 which are appropriately high. It is above the rule of thumb ten (10) as said in literature. This shows great evidence of multicollinearity in the regression model. The Sig values of the three variables (X_1), (X_2) and (X_3) are all greater than 0.05. This shows evidence of no significant relationship in the variables. The model is inadequate and provides great insight to multicollinearity present.

TABLE 5: Variance Decomposition Proportions of Eigenvalue And Condition Index

Model	**Dimension**	**Eigenvalue**	**Condition Index**	**Variance Proportions**			
				(Constant)	**X1**	**X2**	**X3**
1	1	3.854	1.000	.00	.00	.00	.00
	2	.088	6.619	.08	.00	.01	.01
	3	.057	8.189	.01	.00	.01	.98
	4	.001	63.587	.91	1.00	.98	.01

The condition index of Table 5 is extremely high (63.597) which cause great problems to the other two variables. Although the condition index of (2) and (3) are a bit fair but the high value of (4) has much evidence to detect multicollinearity in the model.

Conclusively, the above analysis shows clear evidence that multicollinearity exist in every statistical test and visual assessment as used in the materials and methods.

Unfolding Other Violations Due to Multicollinearity Present Using Specified Benchmark and Sig Value

Table 6: Model Summary Decomposition

Model	R	R Square	Adjusted R Square	Durbin-Watson	Decision
1	.514	.265	.148	1.470<1.5	Serial correlation evident

The Durbin Watson test statistic of Table 6 is 1.470 is below 1.5 threshold rule of thumb, showing evidence of serial correlation.

Table 7: F Test for Detecting Heteroskedasticity Present

F	df1	df2	Sig.	Decision
5.991	1	21	.023<0.05	Heteroscedasticity Present

The Sig value of Table 7 is 0.023 is less than the 0.05, therefore, the null hypotheses of equal variances is rejected. There is a present of heteroscedasticity in the model

Transformation

The detention of model violation of multicollinearity and other model violations can be corrected using trial by error transformation. Different transformations are tried until the following is achieved.

(i) Removal of multicollinearity when the VIF of independent variables is less than 2.500

(ii)The residual scatter plots must satisfied the classical regression assumption of normality, homoscedasticity, linearity and normality of the error term.

Table 8: Vif Values at Various Transformation Level: An Insight Guide to Optimal Model

Variables	VIF of Log Transformation	VIF of Increase Sample Sizes ($n \leq 69$)	VIF of Square Root Transformation
X1	49.286	2.028<2.500:Optimal Model	52.434
X2	50.603	2.238<2.500: Optimal Model	52.706
X3	1.726	1.179<2.500:Optimal Model	1.776

The VIF values of increase in sample size ($n \leq 69$) seems feasible and satisfactory than the other transformation. Its Variance inflation factor is very minimal(less than 2.500). They are all below the benchmark value of 5 as indicated in literature. The VIF of Trademark(X1) and Bread Texture(X2) have reduced significantly and satisfactory. The analysis is good for prediction and descriptive purposes because the VIF below 2.500 satisfies the four classical regression assumptions.

Table 9: Values of Durbin Watson, F Statistic And P Value of ($n \leq 69$)

Durbin Watson	F Statistic	SigValue	Std Error	N
1.736	1.962	0.130	0.753	69

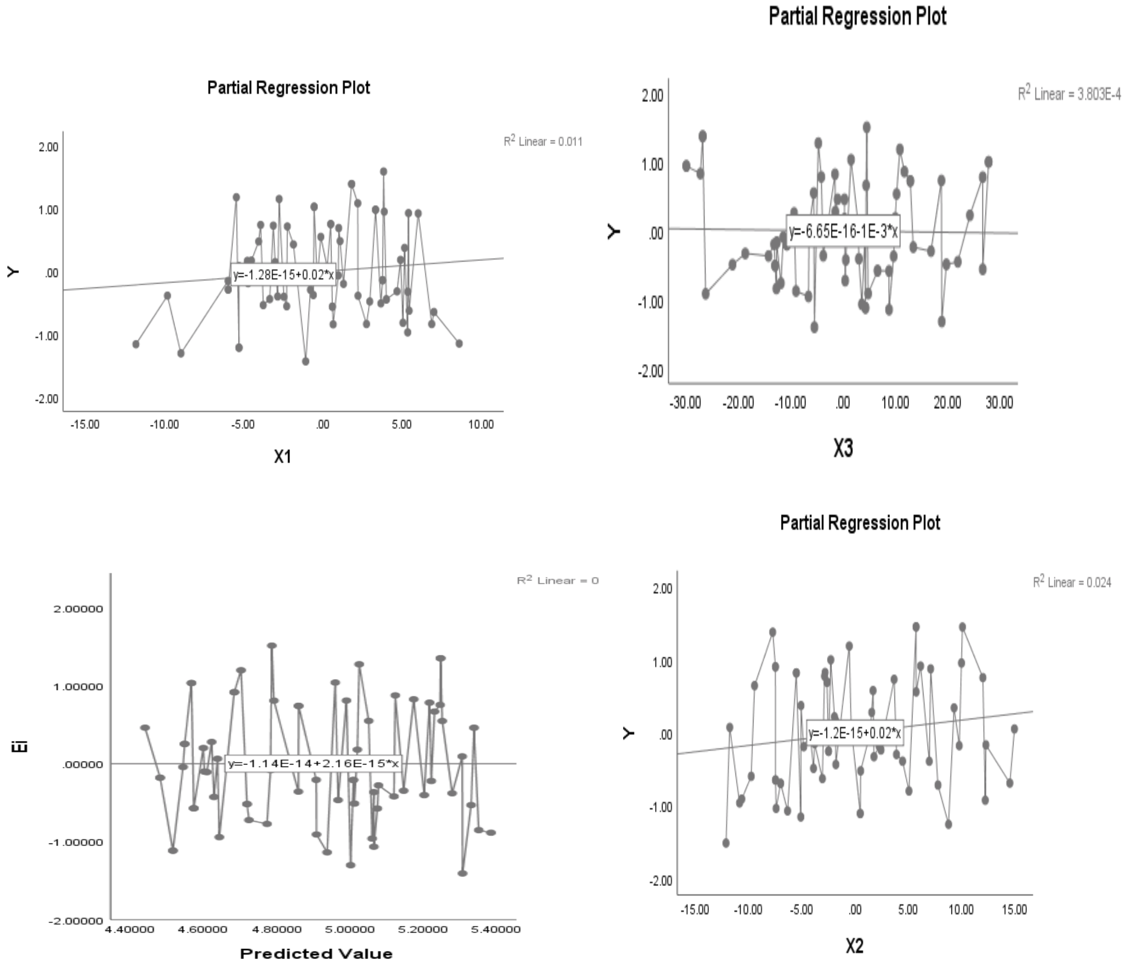
The Durbin Watson of Table 9 are 1.736 greater than the 1.5 benchmark. This shows evidence of Independent of the observation. The present of serial correlation has been removed.

Table 10: F Test For Detection Of Heteroscedasticity

F Statistic	SigValue
0.566	0.459

The value for F test for heteroscedasticity is 0.459 greater than the 0.05 level of significance. Evidence of homoscedasticity (constant- variance) is now present in the regression model.

Graphical Plots of Optimal Regression Assumption After Feasible Transformation of VIF less than 2.500



The feasible transformation of variance inflation factor (VIF) less than 2.500 for each of the independent variables satisfies the classical regression assumptions and has no present of multicollinearity in the regression model. Finally, the model is optimal, descriptive and has predictive ability.

Future Recommendation

Researchers in the areas of forecasting, transformations, variables selection and multivariate regression analysis should take note of the following

(1) The variance inflation factor (VIF) less than 2.500 should be used use as a stopping rule for

Optimal regression model. The reason is that variance inflation factor below 2.5000 satisfies the four classical regression assumptions of linearity, homoscedasticity, normality and independence of the observations.

(2) The bivariate correlation matrix is cumbersome as it requires a lot of computation.
The variance inflation factor provides for each of the independent variables specified benchmark for detecting multicollinearity. It also provides specified benchmark for feasible transformation. Transformation stops when the variance inflation factor is less than 2.500.

(3) The adjusted coefficient of determination should be used in place of ordinary coefficient of determination in computing variance inflation factor (VIF). The reason is that the adjusted coefficient of determination has a better predictive power than the ordinary coefficient of determination because its selects only significant optimal variables into the model than the ordinary coefficient of determination which increases as the sample size increases, so it is better to use the adjusted coefficient of determination as a better method for determining the variation of the independent variables over the dependent variable.

Conclusion

The statistical test, Sig Value and bivariate graphical diagnoses, bivariate correlation coefficient, bivariate correlation matrix, the pairwise scatter plots, variance Inflationary factor (VIF), tolerance value, Eigen value, condition indices , and F test used in this research work shows clear evidence of multicollinearity present in the regression model. In summary of all, model with the highest values of variance inflation factor shows convincing indicators of multicollinearity present than other tools. The high values of variance inflation factor provide great insight to other model assumptions violations. It also shows that there is present of heteroscedasticity, serial correlation, non-linearity and non-normality in the data. In addition, F test diagnostics also detect multicollinearity and also provides overall model violations. In all transformation, the increase in sample sizes $(n \leq 69)$ with linear and quadratic features have a variance inflation factor less than 2.500 and was used. Finally, variance inflation factor (VIF) and the F test diagnostics should be look upon when detecting multicollinearity because it provides an insight to other model violations present in the

Conflicts of interests.

The authors declare that they have no known competing financial interest or personal relationship that have appeared to influence the work reported in this paper

References

1. Adeboye, N., Fagoyinbo., I.S., & Olatayo, T. (2015). Estimation of the Effect of Multicollinearity on the Standard Error for Regression Coefficients: IOSR Journal of Mathematics, 10(4), 16-20. hpps://doi.org/10.9790/5728-10411620
2. Adenomen, M.O., and Micheal, V.A (2017): Model Violations Diagnostics. International Journal of Innovative Finance and Economics' Research 6(3):74-89
3. Andrew, F.H (2019):Using heteroscedasticity consistent standard error estimator in OLS regression. An introduction software. Journal of Statistical computation and simulation vol 6 No. 6, pp 787-803.
4. Auqino, S., Maiyastri, M., & Diana, R (2019). Perbandingan Metode Kuadrat Terkecil Dan Metode Bayes Pada Model Regresi Linier Berganda Yang Mengandung Multikolinieritas, Jurnal Matermatika UNAND, 8(1), 307 https://doi. org/10.25077/jmu.8.1.307-2019
5. Balita Gizi Buruk Di Kabupaten Bojonegoro. Jurnal Statistika Dan Komputasi, 1(2), 74-82. https://doi.org/10.32665/statkom. vli2.1170
6. Bayman, E. O., & Dexter, F.(2021). Multicollinearity in Logistic Regression Models. Anesthesia and Analgesia, 133(2), 362-365.
7. Belsley, D.A.(2018): Conditioning diagnostics: Collinearity and weak data in regression, John Wiley & Sons, Inc., New York, 1991.
8. Everitt, B.S.(2002).Thje Cambridge Dictionary of Statistics(2nd Ed.). Cambridge University Press publisher, New York
9. Nagelkere, N.J.(1991). A Note on a General Definition of the Coefficent of Determination. Biometrices: 78(3):pp 691-692
10. Hermawan, T.(2021). Estimasi Kurva Regresi Spline Pada Data longitudinal Dengan Metode kuadrat Terkecil. Intersecttions, 5(2), 17-25, https://doi.org/10.47200/intersection.v512.583
11. Janah, M., & Kartini, A.Y.(2022). Penerapan Metode Regresi Linier Berganda Pada Kasus
12. Knoke, et al. (2002): Statistics for social data analysis (4th edition), FOE. Peacock Publisher, illinios, 2002

13. Lavery, M.R., Acharya, P., Sivo, S.A & Xu, L.(2019). Number of Predictors and Multicollinearity: What are their effects on error and bias in regression? Communications in Statistics: Simulation and Computation, 48(1), 27-38 https://doi.org/10.1080/03610918.2017.137175
14. Ningsih, T., Herrhyanto, N., & Rachmatin, D.(2019). Analisis Regresi Linear Piecewise Dua Segmen Denga Menggunakan Metode Kuadrat Terkecil. Jurnal Eureka Matika, 7(2), 63-82
15. Nyrhinen, J.N., & Leskinen, E. (2014). Multicollinearity in Marketing Models: Notes on the application of Ridge Trace estimation in structural Equation Modelling. Electronic Journal of Business Research Methods, 12(1), 3-15
16. Obite, C.P., Olewuezi, N.P., Ugwuanyim, G.U., & Bartholomew, D.C.(2020). Multicollinearity Effect in Regression Analysis: A Feed Forward Artificial Neura Network Approach: Asian Journal of Probability and Statistics, 6(1), 22-33.
17. Sartika, L., Debataraja, N.N., & Imro 'ah, N.(2020). Analisis Regresi Dengan Metode Least Absolute Shrinkage and Selection Operator (Lasso) Dalam Mengatasi Multikolinearitas Bimaster: Buletin IImiah Matematika, Statistika Dan Terapannya, 9(1), 31-38 https://doi. org/10.26418/bbimst.v9iil.38029
18. Senavirata, N.A.M.R., & Cooray, T. M .J .A(2019).Diagnosing Multicollinearity of Logistic Regression model. Asian Journal of Probability and statistics, 5(20, 1-9. https://doi.org/10.9734/ajpas/2019/v5i230132
19. Vatcheva et al (2017): Multicollinearity in regression analyses conducted in epidemiologic studies. Epidemiology (Sunnyvale) 2016; 6:227. [PMC free article] [PubMed] [Google Scholar] [Ref list]
20. Weaving, D., Jones, B., Ireton, M., Whitehead, S., Till, K., & Beggs, C.B.(2019).Overcoming the problem of multicolliearity in sports performance data: A novel Application of Partial Least Squares Correlation Analysis. PLOS One, 14(2), 1-16.https://doi.org/10.1371/journal.pone.0211776
21. Young, D.S.(2017): Handbook of regression methods, CRC Press, Boca Raton, FL, 2017, 109-136.