Journal, Advances in Mathematical & Computational Sciences
Vol. 8 No. 2 June 2020
www.mathematics-computationaljournal.info

Article Citation Format
Ilo H.O, Awariefe, C., Afolabi, N. (2020): Investigating the Diagnostic Factors of Multicollinearity in a Multiple Linear Regression Model. Journal of Advances in Mathematical & Computational Sc. Vol.8, No. 2. Pp 9-18

Article Progression Time Stamps

Article Type: Research Article
Manuscript Received 13th April, 2020
Final Acceptance: 2nd June, 2020
Article DOI Prefix: dx.doi.org/10.22624

Investigating the Diagnostic Factors of Multicollinearity in a Multiple Linear Regression Model

Ilo H.O[1], Awariefe, C.[2] & Afolabi, N.[3]
[1 & 3] Department of Statistics, Ogun State Institute of Technology, Igbesa, Ogun State, Nigeria.
[2]Department of Statistics, Delta State Polytechnic, Ozoro, Delta State, Nigeria
E-mails: [1]alameda44@yahoo.com; [2]awariefec@gmail.com; [3]nasimotafolabi@gmail.com
Phones: [1]+2347041112438; [2]+2348033471856; [3]+2348097337963

ABSTRACT

A high degree of linear dependency among several explanatory variables in a multiple regression causes multicollinearity, usually when the correlation coefficients among explanatory variables is either very close to 1 or to -1 then the problem of multicollinearity arises. This problem inflates or overestimates the standard error of regression coefficients of multiple regression models, rendering the standard errors to be abnormally large, the larger the regression coefficient, the less likely it is that this coefficient will be statistically significant. This study is focused on investigating the diagnostic factors of multicollinearity in a multiple regression model. This paper used real data to check for multicollinearity via variance inflation factor (VIF), tolerance value, condition number, eigenvalue and examination of the correlation matrix. Based on the empirical analysis of the econometric data employed, we infer that variance inflation factor (VIF), tolerance value, examination of correlation matrix and the condition index are effective tools in detection of multicollinearity problem in a regression model.

Keywords: Multicollinearity, Variance Inflation Factor, Tolerance Value, Correlation Matrix, Condition Number, Eigenvalue.

1. INTRODUCTION

Multicollinearity in multiple regression models is a condition whereby there is high degree of correlation or dependency among several independent (explanatory) variables. This problem happens when the correlation coefficients among explanatory variables is either very close to 1 or to -1. Ranjit (2014) presented his opinion that the presence of multicollinearity can render the least-squares analysis of the regression model inadequate. Sometimes, multiple regression outcomes may look inconsistent. Though the overall p-value is very low, the individual p values are high. Hawkins (1983) explains the term multicollinearity as a situation in which there is an exact or nearly exact linear relation among two or more independent variables. The exact relation commonly arises due to errors or lack of understanding of the input variables.

There is no clear-cut threshold for evaluating multicollinearity of linear regression models. Computation of correlation coefficients of independent variables can be obtained. But high correlation coefficients do not necessarily imply multicollinearity. A judgment by checking related statistics can be made, such as tolerance value or variance inflation factor (VIF), Eigenvalue, and condition number. Belsley et al (1980). Asterou and Hall (2015) posited that if near linear dependency exists, the auxiliary regression will display a small equation standard error, a large R^2 and statistically significant *F*-value. Jeeshim (2003) defines multicollinearity as a high degree of correlation (linear dependency) among several independent variables. It commonly occurs when a large number of independent variables are incorporated in a regression model. It is because some of them may measure the same concepts or phenomena. Freund and Littell (2000) identified Variance inflation factor (VIF) as just the reciprocal of a tolerance value, thus low tolerances correspond to high VIF.VIF shows how multicollinearity has increased the instability of the coefficient estimates.

Greene (2000) identified that Multicollnearity has following the consequences.

(i) Variance of the model and variances of coefficients are inflated. As a result, any inference is not reliable and the confidence interval becomes wide.

(ii) Estimates remain BLUE, so does coefficient of determination (R^2)

2. DIAGNOSTIC FACTORS OF MULTICOLLINEARITY

There is no definite measure for evaluating multicollinearity of linear regression models; however, judgement about Multicollinearity of regression models can be checked by using statistics, such as variance inflation factor (VIF), tolerance value, correlation matrix (correlation coefficient), condition number, and eigenvalue.

2.1 Variance Inflation Factor And Tolerance Value.

Farrar and Glauber(1967) proposed the Variance Inflation Factor (VIF) measures the inflation of the parameter estimates being computed for all explanatory variables in regression models.

The VIF formula is as follows: Wooldridge (2000),

$$\text{Var}(\hat{\beta}_i) = \frac{\sigma^2}{S_{ii}(1-R_i^2)} \qquad \text{.....2.1}$$

Where

$S_{ii} = \sum_i^n (X_{ij} - \bar{X}_i)^2$ and R_i^2 is the unadjusted R^2

Procedure:

X_i is regressed against all the other explanatory variables in the model, that is, against a constant, X_2 , X_3, ...,X_{i-1}, , X_{i+1}, ...,X_k. Suppose there is no linear relation between X_i and the other explanatory variables in the model. Then, R_i^2 will be zero and the variance of $\hat{\beta}_i$ will be $\frac{\sigma^2}{S_{ii}}$.

Dividing this into the above expression for Var($\hat{\beta}_i$), we obtain the variance inflation factor and tolerance as

VIF($\hat{\beta}_i$) = $\frac{1}{(1-R_i^2)}$ 2.2

Tolerance ($\hat{\beta}_i$) = $\frac{1}{VIF} = 1 - R_i^2$ 2.3

As It is shown in 2.2 and 2.3 the higher VIF or the lower the tolerance index, the higher the variance of $\hat{\beta}_i$ and the greater the chance of finding β_i insignificant, which means that severe multicollinearity effects are present. Hence, these measures can be useful in identifying multicollinearity.

The practice is to choose each right hand side variable (explanatory variable) as the dependent variable and regress it against a constant and the remaining explanatory variables. We will then obtain k–1 values for VIF. If any of them is high, then multicollinearity is indicated. Unfortunately, however, there is no theoretical way to say what the threshold value should be to judge that VIF is "high." However, Marquardt and Snee (1975) indicate that if any of the VIFs exceeds 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity. Specifically, if the overall F statistic is significant but the individual t statistics are all non significant, multicollinearity is present.

2.2 Correlation Matrix of Input Variables.

Reddy et al (2013) posited that a very simple measure of multicollinearity is inspection of the off-diagonal elements, if the regressors are nearly linearly dependent, then r_{ij} in $X'X$ will be near unity. If there is high multicollinearity between any two predictor variables, then the correlation coefficient between these two variables will be near to unity. By using correlation matrix, we can identify the close relationships between the input variables and further investigate them to decide about including them in the final model. Generally, a correlation of more than 0.6 can be treated as variable that cause the multicollinearity problem.

2.3 Condition Number and Eigenvalue

Condition number is used as measure for detecting the existence of multicollinearity in regression models. This measure is based on the eigenvalues of the explanatory variable matrix, measuring the sensitivity of small estimators to small variations in the variances. Condition number can be computed using the formula below: Yong-Wei (2008)

C.N = $\sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$ 2.31

where λ_{max} is the largest eigen-value of the matrix $X'X$
λ_{min} is the smallest eigen-value of the matrix $\mathbf{X'X}$

Yong-Wei (2008) suggested that if CN is between 20 and 30 as an indicator for a high linear multicollinearity. In case of no multicollinearity all eigenvalues would be unity. Eigenvalues smaller or larger than unity would indicate presence of multicollinearity.

3. DATA ANALYSIS AND DISCUSSION OF RESULTS

This study utilizes secondary data obtained from the website of Central Bank of Nigeria. The data consist of econometric data of Gross domestic product which represent the dependent variable(Y) and independent variables of exchange rate (Dollar) (X_1), bureau d-change rate (BDC) (X_2), Inflation (X_3), Interbank Rate (X_4), Unemployment (X_5), Domestic crude Production(X_6) from 2004-2017.The data obtained in this study will be processed using the statistical packages software, SPSS.

SPSS OUTPUT

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Domestic crude production, Unemployment, Interbank rate, Inflation, BDC, Exchange rate[b]	.	Enter

a. Dependent Variable: GDP
b. All requested variables entered.

ANOVA[a]

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	3081202571.241	6	513533761.874	5.625	.020[b]
	Residual	639039937.480	7	91291419.640		
	Total	3720242508.721	13			

a. Dependent Variable: GDP
b. Predictors: (Constant), Domestic crude production, Unemployment, Interbank rate, Inflation, BDC, Exchange rate

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	27798.624	51770.157		.537	.608		
	Exchange rate	-65.201	394.126	-.238	-.165	.873	.012	84.328
	BDC	156.234	253.835	.831	.615	.558	.013	74.344
	Inflation	-2221.239	1019.284	-.525	-2.179	.066	.423	2.366
	Interbank rate	641.933	1048.058	.177	.612	.560	.294	3.406
	Unemployment	686.050	590.273	.320	1.162	.283	.324	3.085
	Domestic crude production	-12167.585	21389.242	-.167	-.569	.587	.283	3.532

a. Dependent Variable: GDP

Correlations

		Exchange rate	BDC	Inflation	Interbank rate	Unemployment	Domestic crude production
Exchange rate	Pearson Correlation	1	.991**	.484	.782**	.718**	-.581*
	Sig. (2-tailed)		.000	.079	.001	.004	.029
	N	14	14	14	14	14	14
BDC	Pearson Correlation	.991**	1	.503	.736**	.711**	-.608*
	Sig. (2-tailed)	.000		.067	.003	.004	.021
	N	14	14	14	14	14	14
Inflation	Pearson Correlation	.484	.503	1	.225	.375	-.689**
	Sig. (2-tailed)	.079	.067		.440	.187	.006
	N	14	14	14	14	14	14
Interbank rate	Pearson Correlation	.782**	.736**	.225	1	.543*	-.335
	Sig. (2-tailed)	.001	.003	.440		.045	.242
	N	14	14	14	14	14	14
Unemployment	Pearson Correlation	.718**	.711**	.375	.543*	1	-.172
	Sig. (2-tailed)	.004	.004	.187	.045		.557
	N	14	14	14	14	14	14
Domestic crude production	Pearson Correlation	-.581*	-.608*	-.689**	-.335	-.172	1
	Sig. (2-tailed)	.029	.021	.006	.242	.557	
	N	14	14	14	14	14	14

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions						
				(Constant)	Exchange rate	BDC	Inflation	Interbank rate	Unemployment	Domestic crude production
1	1	6.668	1.000	.00	.00	.00	.00	.00	.00	.00
	2	.197	5.824	.00	.00	.00	.00	.02	.00	.00
	3	.085	8.875	.00	.00	.00	.30	.14	.00	.00
	4	.035	13.754	.00	.00	.01	.23	.54	.04	.00
	5	.014	22.209	.03	.00	.01	.01	.04	.69	.01
	6	.001	71.082	.96	.00	.01	.46	.00	.27	.97
	7	.001	88.326	.01	.99	.97	.00	.26	.00	.02

a. Dependent Variable: GDP

From the SPSS regression coefficient output above it is seen that the independent variable (Exchange rate) and (BDC rate) have high correlation coefficients and VIF, VIF more than 10 indicate presence of serious multicollinearity while Inflation rate, Inter-Bank rate, Unemployment rate and Domestic production(X_6) has very low VIF. Though the model has serious multicollinearity, from the ANOVA output the model is adequate since p-value 0.003 is less than the level of significance (0.05) with high coefficient of determination R^2 = 0.950.

BDC, Interbank and Unemployment and Exchange rate explanatory variables are very highly correlated (r = 0.99, 0.78, 0.72). Of course, the tolerances for these variables are therefore also very low. The collinearity diagnostic SPSS output does not explicitly reports the condition number but it reports the largest condition indices of 72.082 and88.326. This falls within our "rule of thumb" range for concern. The collinearity diagnostic SPSS output has all eigenvalues less than and greater than unity, which indicate presence of multicollinearity.

It can been seen from SPSS regression coefficient output that the independent variables, exchange rate, BDC, interbank rate and unemployment are all insignificant with p-values less than 0.05 due to the high presence of multicollinearity that affected the ordinary least square estimates. Yet, the overall F is significant. All of these checks are notification of multicollinearity. A change of one or two of the explanatory variables causing problem could completely reverse the estimates of the effects.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.874[a]	.764	.659	9881.82004

a. Predictors: (Constant), Domestic crude production, Unemployment, Interbank rate, Inflation

ANOVA[a]

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	2841389202.408	4	710347300.602	7.274	.007[b]
	Residual	878853306.313	9	97650367.368		
	Total	3720242508.721	13			

a. Dependent Variable: GDP

b. Predictors: (Constant), Domestic crude production, Unemployment, Interbank rate, Inflation

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	63716.213	47213.938		1.350	.210		
	Inflation	-2406.043	1046.553	-.569	-2.299	.047	.429	2.331
	Interbank rate	1309.653	754.717	.361	1.735	.117	.606	1.651
	Unemployment	1301.684	457.826	.607	2.843	.019	.576	1.735
	Domestic crude production	-33008.931	17677.025	-.454	-1.867	.095	.443	2.255

a. Dependent Variable: GDP

From table ANOVA and Coefficient output above it is seen that when BDC and Exchange rate cases were eliminated from the model since they have very high VIF's. The multiple regression model no longer exhibit multicollinrarity and the ordinary least square estimates is now precise with some of the independent variables, that is, Inflation rate, Unemployment rate and Domestic crude production significant since their respective p-values are less than ($\alpha = 0.05$) and the model remains adequate from the ANOVA table returning p-value less than ($\alpha = 0.05$), this implies that a change of one or two of the explanatory variables causing the problem could completely reverse the estimates of the effects.

4. CONCLUSION

Based on the empirical analysis of the econometric data we can infer that variance inflation factor (VIF), tolerance value, examination of correlation matrix and the condition index are effective tools in detection of multicollinearity problem in a regression model. When the presence of multicollinearity is severe in a multiple regression model then the ordinary least square estimators are imprecisely estimated. Multicollinearity problem undermines the statistical significance of an explanatory variable. It overestimate the standard error of regression coefficients of multiple regression, making the standard errors to be unusually large, the larger the regression coefficient, the less likely it is that corresponding coefficient(s) will be statistically significant.

REFERENCES

1. Asterou, D and Hall, G (2015). Performance of a new Ridge Regression Estimator. *Journal of the Association of Arab Universities for Basic and Applied Sciences*, 9(2), pp. 43-50.
2. Belsley, David. A., Edwin. Kuh, and Roy. E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley and Sons.
3. Freund, J. and Littell.C (2000). *SAS System for Regression* (Third edition). Cary, NC: SAS Institute.
4. Greene, William H. 2000. *Econometric Analysis* (Fourth edition). Upper Saddle River, NJ: Prentice-Hall.
5. Hawking, R. R. and Pendleton, O. J. (1983). The regression dilemma, *Commun*. *Stat*.- *Theo. Meth*, 12, 497-527.
6. Jeeshim and KUCC625. (*2003)* Multicollinearity in Regression Models http://php.indiana.edu/~kucc625 *8*
7. Marquardt, D. W. and Snee, R. D. (1975) Ridge regression in practice. *Amer*. *Statis*t., 29, 3-19.
8. Ranjit, K.P (2014) Multicollinearity, Causes, Effects and Remedies. *Indian Agricultural Research Institute*, I.A.S.R.I, Library Avenue, New Delhi-110012
9. Reddy, M., Balasubramanyam, P., and Subbarayudu, P (2013) An Effective Approach to Resolve Multicollinearity in Agriculture Data. *International Journal of Research in Electronics and Computer Engineering*. IJRECE Vol. 1 Issue 1 : 2348-2281
10. Yong-Wei, G., Willan, A.R. and Watts, D.G. (2008) A method to measure and test the damage of multicollinearity to parameter estimation. *Science of Surveying and Mapping*, 2, pp.1-44.20
11. Wooldridge, J. M (2000) *Introductory Econometrics: A Modern Approach*, South Western. California.

APPENDIX

Table 1: Econometric data of Nigeria

Year	GDP (Y) 'Trillion'(=N=)	Exchange rate (Dollar) X_1	BDC (X_2)	Inflation (X_3)	Interbank rate (X_4)	Unemployment X_5	Domestic crude Production (X_6) 'Billion'
2004	4,725	132.5	140.69	15.38	6.45	25.6	2.03
2005	6,912.40	128.5	143.94	17.85	7.26	38	2.12
2006	8,487	126.4	129.82	8.38	7.38	32.3	2.43
2007	11,411.10	124	123.8	5.42	7.93	32.2	2.36
2008	14,572.20	115.5	118.1	11.53	11.86	32.1	2.2
2009	18,564.60	147	152.03	12.59	11.87	27.5	2.04
2010	20,657.30	148	153.13	13.76	4.02	39	2.05
2011	24,294.20	151	160.35	10.85	10.57	43.5	2.59
2012	24,794.20	154	159.32	12.24	13.94	42.7	2.27
2013	29,205.80	154.5	167.14	8.52	12.08	25.6	2.21
2014	44,725.10	154	175.85	7.18	11.67	38	2.11
2015	51,345.34	196	232.4	9.12	11.87	42.12	2.43
2016	50,123.34	305.18	415.36	18.34	15.67	48.6	1.69
2017	50,102.12	305.90	362.41	16.12	22.95	49.15	1.93

Source: CBN Bulletin 2018