International Journal of Applied Mathematics and Theoretical Physics
2021; 7(3): 62-67
http://www.sciencepublishinggroup.com/j/ijamtp
doi: 10.11648/j.ijamtp.20210703.11
ISSN: 2575-5919 (Print); ISSN: 2575-5927 (Online)

Comparing Two Classical Methods of Detecting Multicollinearity in Financial and Economic Time Series Data

Mutairu Oyewale Akintunde, Abolade Oludayo Olawale, Ajitoni Simeon Amusan, Adeyinka Ismail Abdul Azeez

Department of Statistics, School of Applied Sciences, Federal Polytechnic, Ede, Nigeria

Email address:
waleakintunde2004@gmail.com (M. O. Akintunde)

To cite this article:
Mutairu Oyewale Akintunde, Abolade Oludayo Olawale, Ajitoni Simeon Amusan, Adeyinka Ismail Abdul Azeez. Comparing Two Classical Methods of Detecting Multicollinearity in Financial and Economic Time Series Data. *International Journal of Applied Mathematics and Theoretical Physics.* Vol. 7, No. 3, 2021, pp. 62-67. doi: 10.11648/j.ijamtp.20210703.11

Received: May 12, 2021; **Accepted**: June 7, 2021; **Published**: August 18, 2021

Abstract: Multicollinearity is an unavoidable problem being faced by researchers in financial and Economic data. It refers to a situation where the degrees of correlations between two or more independent variables are high. This is to say, one explanatory variable can be used in forecasting the other variable. This creates redundant information in a series under study, skewing the results in regression models. There is need to search for the source of the problem and proffering solution to this problem in Economics and Financial data. The data used was extracted from the record of Federal trade commission (FTC), 2019. The commission usually ranks annually arrays of locally made cigarettes in relation to Tar, nicotine and carbon monoxide components that was made available. Farrah-Glauber test and variance inflation factor were used as methods of detection multicollinearity in this paper. SPSS and J-muliti packages were used to analyse the data collected for empirical illustration. The results of analysis indicated that variance inflation factor of X_1 and X_2 (Tar and Nicotine) are far above 10 (21.63 and 21.90) must be removed or collapsed from the model in order to correct multicollinearity. So, the preciseness of VIF made it to be preferred to Farrah-Glauber test. In line with the analysis, the use of Variance Inflation Factor is more preferred to Farrah-Glauber method. As VIF not only detected but also pointed to the direction of the problem.

Keywords: Multicollinearity, Farrah-Glauber, Predictor, Variance Inflation Factor, Financial and Economic Data, Regression Model

1. Introduction

Multicollinearity refers to the circumstances where two or more independent variables in a statistical model are linearly related they are sometimes called collinearity: [1]. It is an important economic problem that has received several attentions globally but unfortunately the problem of resolving it has not yielded desire result. Of recent authors like [18, 15, 8, 2, 10, 11] researched into this econometric problem and established the danger the problem posed to the forecast ability of regression models. It is also regarded as economic problem that can lead to poor judgmental error and lead to poor economic policy formulation in financial time series the error is assumed to be independent and identically distributed whereas in the real-life situation most of the time is not so. Multicollinearity among predictor variables has been attended to severally in econometric theory and in econometric texts (for examples., [6, 7, 19]. [9] Determines how collinearity upshots parameter coefficient instability in a measurement error situation. Many statistical models, notably those that are commonly use in ecology, finance, marine and Economics are liable to collinearity [3, 4, 17]. This occurs when too many variables have been pulled together in the model and a number of them measure similar phenomena. The existence of multicollinearity in a variable under study affects both the estimation of the parameters of the model and also gives rise to wrong interpretation of the results. Regression parameters estimates so obtained are compromised and may lead to instability, the estimated errors are extremely stretched and as a result inferences made based on these statistics are biased and lead to wrong policy formulation. However, for the models that are not robust enough two problems are bound to happen under multicollinearity: any effects arising in the variable cannot be put apart variable effects cannot be separated and extrapolation or out of sample forecast is likely to be seriously erroneous and give a very wrong judgmental decision (s) [12].

Most introductory textbooks on statistics recognized multicollinearity as a problem principally associated with finance and Economics data. It is regarded as a situation where the model is not identified. As terrible as it is, several approaches for investigating it and working with it have been mapped out. Regardless of the peculiarities of the problem and the several available methods of solving them, most ecological, finance and Economics research have not made efforts to address this ubiquitous problem of multicollinearity [5, 16]: Non- addressing of these problem are directly linked to a very erroneous belief that statistical methods are not affected by multicollinear problems, ambiguity that surrounded the method to use couple with incompatible of a method in relation to the available data to be analysed, inability to interpret the results as a result of usage of approaches that incorporate variables or software that cannot be accessed. This problem is not only limited to ecology, finance and Economics [10, 13, 14].

The central objective of this paper is to provide a better perception of multicollinearity and to compare two methods (Farah-Glauber test and variance inflation method) of detecting its presence and determine the better one.

2. Mathematical Preliminaries

2.1. The Farrar and Glauber Test

This is a test to determine the presence as well as the degree of Multicollinearity in an equation. To achieve this objective a matrix of pair wise correlation coefficients is formed from the explanatory variables.

$$r_{ij} = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1k} \\ r_{21} & 1 & r_{23} & \cdots & r_{2k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ r_{1k} & r_{2k} & r_{3k} & \cdots & 1 \end{bmatrix}$$

This test is performed in three stages

i. Chi-square test to determine or ascertain the existence and degree of multicollinearity.

ii. F-test to locate the variable (s) that are intercorrelated, provided the test appeared positive.

iii. t-test is use to determine the variable (s) that is (are) causing the multicollinearity problem provided the F-test is positive.

2.1.1. Chi-Square Test

$H_0 : X's$ are orthogonal, is a statistic predicated on the determinant $|X'X|$ and could give a valuable measure of of the existence of multicollinearity in the expanatory variables. Bartlett (1937) obtained a transformation of $|X'X|$

$$\chi^2 = -\left[N - 1 - \tfrac{1}{6}(2K+5)\ln D\right]$$

This is distributed approximately as chi-square with $v = \tfrac{1}{2}K(K-1)$ degrees of freedom; where K is the number of explanatory variables present in the series. S

2.1.2. $F-Test$

If the Chi-square test confirmed the presence of Multicollinearity, we therefore, have no choice than to proceed to F – test using the following steps:

i. List out the x_i considered to be inter-correlated with other xs as a function of xs. Therefore, $x_i = f(x_1, x_2, \cdots, x_{i-1}, x_{i+1}, x_k)$. Using data, we can write $x_1 = \beta_2 X_2 + \beta_3 X_3 + U$

ii. Compute the parameter $b_i = \begin{bmatrix} b_2 \\ b_3 \end{bmatrix}$

As $b = (X'X)^{-1} X'X_1$ $\quad$ *where* $X = (X_2 X_3)$

iii. Compute $R_i^2 = \dfrac{b_2 \sum X_1 X_2 + b_3 X_1 X_2}{\sum X_1^2}$

Where $X_i X_j = \sum X_i X_j - \dfrac{\sum X_i \sum X_j}{N}$

iv. Compute the $F-Statistic$

$F_{Calc.} = \dfrac{R_i^2 / (K-1)}{(1-R_i^2)/(T-K)}$ Check $F_\alpha(K-1, N-K)$ F – distribution table

2.1.3. t – Test

Having discovered that F test is positive. The t test is thereafter conducted to detect/examine which pair of variables are responsible for the multicollinearity. Suppose x_1 is intercorrelated with x_2 and x_3, under this condition, the t test is conducted as follows:

1. Define the hypothesis
   $H_0 : x_2$ and x_3 *are* is not responsible for multicollinearity against the
   $H_1 : x_2$ and x_3 *are* responsible for multicollinearity.
2. Compute the partial coefficient of determination $r_{12.3}^2 = \dfrac{(r_{12} - r_{13} r_{23})^2}{(1 - r_{13}^2)(1 - r_{23}^2)}$.
3. Define $H_0 : r_{123} = 0$ against $H_1 : r_{123} \neq 0$.
4. Compute the statistic $t_{12} = \dfrac{r_{12.3}\sqrt{N-K}}{\sqrt{1 - r_{12.3}^2}}$ and check $t_\alpha(N-K)$ from t – distribution table.
5. Repeat the test for x_1 and x_3.

2.2. Variance Inflation Factor (VIF)

The aftermath of the multicollinearity is the rise in variance inflation factor. For the jth independent variable, the Variance Inflation Factor is given as

$$VIF_j = \frac{1}{1-R_j^2}.$$

where R_j^2 is the coefficient of determination when variable X_j is regressed on the j-1 remaining explanatory variables, these factors are useful indicator in adjudging which of the variables may caused multicollinearity. Rule of thumb seems to be that we should be suspicious if any $VIF_j > 5$ and positively horrified if $VIF_j > 10$. If this kind of result is obtained, a variable should be dropped or the model should be changed. if multicollinearity is discovered, theory and practical judgement should be used to pick the best variables to be kept in the model.

The sampling variance of the jth coefficient $\hat{\beta}_j$ is

$$V\left(\hat{\beta}_j\right) = \frac{1}{\left(1-R_j^2\right)} \frac{\sigma^2}{\left(T-1\right)S_j^2}.$$

Where $S_j^2 = \frac{\sum_{i=1}^{T}\left(X_{ij}-\bar{X}_j\right)^2}{\left(T-1\right)}$ is the variance of X_j and σ^2 variance (Fox, 1997). The term $\frac{1}{1-R_j^2}$ indicates the impact of multicollinearity on $\hat{\beta}_j$. It can be explained as the ratio of variance of $\hat{\beta}_j$ to a supposed variance if X_j were uncorrelated with the remaining X_i.

3. Specification and Analysis of Data Used

The data used for this study was obtained from The Federal Trade commission (FTC), 2018, annually ranks varieties of domestic cigarettes according to their tar, nicotine and carbon monoxide contents.

3.1. Descriptive Statistics

Table 1. *Descriptive Statistics.*

STATISTIC	TAR	NICOTINE	WEIGHT
Mean	12.216	0.8764	0.9703
Standard Error	1.1332	0.0708	0.0175
Median	12.8	0.9	0.9573
Standard Deviation	5.6658	0.3541	0.0877
Sample Variance	32.1014	0.1254	0.0077
Kurtosis	2.9515	4.1604	0.4234
Skewness	0.7567	0.9690	0.4623
Sum	305.4	21.91	24.2571
Count	25	25	25

Table 1 describes reveal hidden statistics about the data used for the study, such statistics include, the mean, variance standard error kurtosis. Skewness and so on just to mention the few. The importance of all this information is to enrich would be policy makers, investors and academia on the associated properties of the data used. For example, Tar and carbon could be both regarded as being approximately normal as their Kurtosis is less than 3, Nicotine is non-normal (Kurtosis greater than 3) as it possesses heavier tails compared to normal distribution. The skewness analysis show that the data are moderately skewed.

Table 2. *Autocorrelation function (ACF).*

Autocorrelations					
Series: y					
Lag	**Autocorrelation**	**Std. Error[a]**	**Box-Ljung Statistic**		
			Value	**df**	**Sig.[b]**
1	-.567	.192	8.724	1	.003
2	.145	.188	9.319	2	.009
3	-.180	.183	10.283	3	.016
4	.230	.179	11.929	4	.018
5	-.216	.174	13.455	5	.019
6	.099	.170	13.797	6	.032
7	-.005	.165	13.798	7	.055
8	-.042	.160	13.865	8	.085
9	.137	.155	14.648	9	.101
10	-.216	.150	16.724	10	.081
11	.222	.144	19.082	11	.060

Autocorrelations					
Series: y					
Lag	**Autocorrelation**	**Std. Error**[a]	**Box-Ljung Statistic**		
			Value	**df**	**Sig.**[b]
12	-.043	.139	19.177	12	.084
13	-.084	.133	19.580	13	.106
14	-.088	.127	20.064	14	.128
15	.156	.120	21.746	15	.115
16	-.018	.113	21.771	16	.151
a. The underlying process assumed is independence (white noise).					
b. Based on the asymptotic chi-square approximation.					

Table 3. *Partial Autocorrelation function (PACF).*

Partial Autocorrelations		
Series: y		
Lag	**Partial Autocorrelation**	**Std. Error**
1	-.567	.204
2	-.260	.204
3	-.355	.204
4	-.077	.204
5	-.194	.204
6	-.205	.204
7	-.094	.204
8	-.218	.204
9	.060	.204
10	-.188	.204
11	.003	.204
12	.242	.204
13	-.008	.204
14	-.095	.204
15	-.096	.204
16	-.006	.204

Critical evaluation of the features exhibited by both ACF and PACF reveal that they contain 16 lags each, they slowly reduced exponentially. these features could be linked to the presence of multicollinearity or long memory in the series under study.

Fararr-Glauber Test

3.1.1. Chi-square Test

$$r_{ij} = \begin{pmatrix} 1 & 0.98 & 0.49 \\ 0.98 & 1 & 0.50 \\ 0.49 & 0.50 & 1 \end{pmatrix}$$

$$Det\left(r_{ij}\right) = 0.05$$

$$\log_e 0.05 = -3.00$$

$$\chi^2_{calculated} = 66.51 \;\; \chi^2_{0.05}3 = 7.82$$

Since $\chi^2_{calculated} > \chi^2_{0.05}3$ i.e 66.51>7.83 . Therefore, multicollinearity exists.

3.1.2. F – Test

Next is to carry out $F-$ test to determine variable (s) causing multicollinearity

$$X_1 = f\left(X_2, X_3\right) \;\; X_1 = \beta_2 X_2 + \beta_3 X_3 + U$$

Thus, $\begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} = \left(X'X\right)^{-1} X'Y$.

The values obtained from the analysis for β_2 and β_3 are

$$\begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 15.28 \\ 2.91 \end{bmatrix}$$

The $F-$ computed is 206.52 and the tabulated value of $F_{0.05}2,22 = 4.30$

Since $F-$ computed is greater than $F-$ tabulated we reject H_0 and conclude that X_1 is inter-correlated with X_2 and X_3 .

3.1.3. $t-$Test

The following hypothesis were set up

H_0 : X_2 and X_3 are not responsible for multicollinearity

H_1 : X_2 and X_3 are responsible for multicollinearity

The value of T obtained from the analysis is 17.31, while the table value was 2.07.

Since computed value of T is greater than table value of T , we can conclude that X_2 and X_3 are responsible for multicollinearity.

3.2. Variance Inflation Factor

The following values were obtained for VIF based on the computer analysis.

Table 4. *Parameters estimation and variance inflation factor.*

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	3.202	3.462		.925	.365					
	Tar	.963	.242	1.151	3.974	.001	.957	.655	.247	.046	21.631
	Nicotine	-2.632	3.901	-.197	-.675	.507	.926	-.146	-.042	.046	21.900
	Weight	-.130	3.885	-.002	-.034	.974	.464	-.007	-.002	.750	1.334

a. Dependent Variable: Carbon Monoxide

Table 4 above shows the value for β_0, β_1, β_2 and β_3 with its VIF value. From VIF value for Tar and Nicotine i.e X_1 and X_2 the value on the table is more than 10 which means multicollinearity exist.

Variable Selection Method

Existence of multicollinearity was established as shown in the above table 1, the next thing is how to correct it. To correct the existence of multicollinearity in the study, variable 3, X_2 (Nicotine) was removed and the new VIF checked.

Table 5. *New results obtained after excluding* X_2

Model		Unstandardized Coefficients		Standardized Coefficients	T	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	3.114	3.416		.912	.372					
	Tar	.804	.059	.961	13.622	.000	.957	.946	.838	.759	1.317
	Weight	-.423	3.813	-.008	-.111	.913	.464	-.024	-.007	.759	1.317

a. Dependent Variable: Carbon Monoxide

Table 5 showing the values for β_0, β_1 and β_3 with its VIF value. From VIF value for X_1 and X_3 the values obtained indicated that multicollinearity has vanished as none of VIF is up to 10.

Table 6. *Compared the results obtained before and after dropping variable* X_2

	β_0	β_1	β_2	β_3	Remark
Multicollinearity with all variables Standard error Variance Inflation factor	3.202 (3.416)	0.963 (0.242) *21.631	-2.632 (3.901) *21.900	-0.130 (3.885) *1.334	Presence of multi-collinearity (X_1 and X_2)
Multicollinearity When X_2 was excluded Standard error Variance Inflation factor	3.114 (3.416)	0.804 (0.059) *1.317		-0.130 (3.813) *1.317	Multicollinearity disappeared.

Table 6 above pooled together and compared the results obtained before and after variable X_2 was excluded coupled with parameters estimate and standard errors. It is glaring that after the exclusion of the variable X_2 the model becomes multi-collinearity free which fulfills the mission of the study.

4. Summary and Conclusion

So far, so good the study examines the descriptive nature of the series. Both the ACF and PACF decay exponentially establishing the fact that the series contain element of multicollinearity or long memory. Farrar-Glauber and variance information confirm the existence of multicolllinearity. Having established this variable X_2 was excluded and test re-conducted which after the analysis indicated that the multicollinearity earlier noticed had disappeared the preciseness of VIF made it to be preferred to Farrah-Glauber test. In line with the above assertion the use of Variance Inflation Factor is more preferred to Farrah-Glauber method. As VIF not only detected but also pointed to the direction of the problem.

References

[1] Alin, A. 2010. Multicollinearity. Journal of WIREs Computational Statistics How does Climate Change affect biodiversity, Vol. 2 No. 3.

[2] Ayinde, K; Lukman, AF; Arowolo, OT (2015). Combined parameters estimation methods of linear regression model with multicollinearity and autocorrelation. *Journal of Asian Scientific Research 5* (5), 243–250.

[3] Belsley, D. A. 1991. Conditioning Diagnostics: Collinearity and Weak Data Regression. John Wiley, New York.

[4] Chatfield, C. 1995. Model uncertainty, data mining and statistical inference. Journal of Royal Statistical Society Vol. 158, No. 3: 419-466.

[5] Graham, M. H. 2003. Confronting multicollinearity in ecological multiple regression. journal of Ecology society of America Vol. 84, No. 11: 2809-2815.

[6] Goldberger, Arthur, 1991. *A course in Econometrics* (Cambridge, MA: Harvard University Press).

[7] Greene, William H., 2003. *Econometric Analysis* (5th ed. Upper Saddle River, NJ: Prentice Hall).

[8] Ismail, B; Manjula, S (2016). Estimation of linear regression model with correlated regressors in the presence of autocorrelation. *International Journal of Statistics and Applications 6* (2), 35–39.

[9] Kalnins, Arturs, 2018. “Multicollinearity: How common factors cause Type 1 errors in multivariate regression,” *Strategic Management Journal* 22, issue 10.

[10] Kiers, H. A. L. and Smilde, A. K. 2007. A comparison of various methods for multivariate regression with highly collinear variables. Journal of Statistical Methods and Applications Vol. 4, No. 6: 193-228.

[11] Lukman, AF; Osowole, O. I; Ayinde, K (2015). Two stage robust method in a linear regression model. *Journal of Modern Applied Statistical Methods 14* (2), 53–67.

[12] Meloun, M., Militiky, J. and Hilli, M. 2002. Crucial problems in regression modelling and their solutions. Journal of Analyst Vol. 12, No. 7: 433-450.

[13] Mikolajczyk, DiSilvestro A, Zhang J. 2008. Evaluation of logistic regression reporting in current obstetrics and gynecology literature. Journal of Obstetrics & Gynecology Vol. 111 No. 10: 413-419.

[14] Murray, C. J. L., Kulkarni SC, Michaud C, Tomijima N. and Bulzacchelli. MT, 2006. Eight Americas: Investigating mortality disparity across races, counties, and race-counties in the United States. - PLoS Medicine 3: e260.

[15] Olanrewaju, S. O.; Yahaya, H. U. and Nasiru, M. O., (2017). Effects of multicollinearity on some estimators in a system of regression equation. *European Journal of Statistics and Probability 5* (3), 1–15.

[16] Smith, A. C., Nicola, K., Charles, M. F. and Lenore, F. 2009. Confronting collinearity: comparing methods for disentangling the effects of habitat loss and fragmentation. Journal of Landscape Ecology Vol. 24, No. 2: 1271-1285.

[17] Stewart, G. W. (1987). Collinearity and least squares regression. *Statistical Science*, 68–84.

[18] Tyagi, G; Chandra, S (2017). Note on the performance of biased estimators with autocorrelated errors. *International Journal of Mathematics and Mathematical Sciences*. Volume *2017*, 12 Pages, Article ID 2045653.

[19] Wooldridge, J., (2010). Econometric Analysis of Cross Section and Panel Data. The MIT Press, Cambridge, Mass.