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:

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-multii 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.

2.1.1. Chi-Square Test

Bartlett (1937) obtained a transformation of the existence of multicollinearity in the explanatory variables.

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

This is distributed approximately as chi-square with \( v = \frac{1}{2}K(K - 1) \) 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 \( x_j \) as a function of \( x_j \). Therefore, \( x_i = f(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, x_k) \). Using data, we can write \( x_i = \beta_0 x_2 + \beta_1 x_3 + U \).

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

As \( b = (XX)^{-1}X'X \)

where \( X = (x_2 x_3) \)

iii. Compute \( R_i^2 = \frac{b_2 X_i x_2 + b_3 X_i x_3}{\sum X_i^2} \)

Where \( X_i x_j = \sum X_i X_j - \frac{\sum X_i \sum X_j}{N} \)

iv. Compute the F-Statistic

\[ F_{calc} = \frac{R_i^2 / (K - 1)}{1 - R_i^2} / (T - K) \]

Check \( F_{\alpha} (K - 1, N - K) \) F-distribution table

2.1.3. t-Test

Having discovered that the F test is positive. The t-test is thereafter conducted to detect/examine which pair of variables are responsible for the multicollinearity. Suppose \( x_i \) 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 not responsible for multicollinearity against the \( H_1 : x_2 \) and \( x_3 \) are responsible for multicollinearity.

2. Compute the partial coefficient of determination

\[ \eta_{123}^2 = \frac{(\eta_{12}^2 - \eta_{13}^2 \eta_{23}^2)}{(1 - \eta_{12}^2)(1 - \eta_{23}^2)} \]

3. Define \( H_0 : \eta_{123} = 0 \) against \( H_1 : \eta_{123} \neq 0 \).

4. Compute the statistic

\[ t_{12} = \frac{\eta_{12} \sqrt{N - K}}{\sqrt{1 - \eta_{123}^2}} \] and check \( t_{\alpha} (N - K) \) from t-distribution table.

5. Repeat the test for \( x_i \) and \( x_j \).
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 indicators in adjudging which of the variables may cause 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

\[ V(\hat{\beta}_j) = \frac{1}{(1 - R_j^2)(T-1)S_j^2}. \]

where \( S_j^2 = \frac{1}{T-1} \sum_{i=1}^{T}(X_{ij} - \bar{X}_j)^2 \) is the variance of \( X_j \) and \( \sigma^2 \) 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.

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

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>
<thead>
<tr>
<th>Lag</th>
<th>Autocorrelation</th>
<th>Std. Error*</th>
<th>Box-Ljung Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Value</td>
</tr>
<tr>
<td>1</td>
<td>-.567</td>
<td>.192</td>
<td>8.724</td>
</tr>
<tr>
<td>2</td>
<td>.145</td>
<td>.188</td>
<td>9.319</td>
</tr>
<tr>
<td>3</td>
<td>-.180</td>
<td>.183</td>
<td>10.283</td>
</tr>
<tr>
<td>4</td>
<td>.230</td>
<td>.179</td>
<td>11.929</td>
</tr>
<tr>
<td>5</td>
<td>-216</td>
<td>.174</td>
<td>13.455</td>
</tr>
<tr>
<td>6</td>
<td>.099</td>
<td>.170</td>
<td>13.797</td>
</tr>
<tr>
<td>7</td>
<td>-.005</td>
<td>.165</td>
<td>13.798</td>
</tr>
<tr>
<td>8</td>
<td>-.042</td>
<td>.160</td>
<td>13.865</td>
</tr>
<tr>
<td>10</td>
<td>-.216</td>
<td>.150</td>
<td>16.724</td>
</tr>
<tr>
<td>11</td>
<td>.222</td>
<td>.144</td>
<td>19.082</td>
</tr>
</tbody>
</table>

Table 2. Autocorrelation function (ACF).
Comparing Two Classical Methods of Detecting Multicollinearity in Financial and Economic Time Series Data

Autocorrelations

<table>
<thead>
<tr>
<th>Lag</th>
<th>Autocorrelation</th>
<th>Std. Error</th>
<th>Box-Ljung Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>-.043</td>
<td>.139</td>
<td>19.177</td>
</tr>
<tr>
<td>13</td>
<td>-.084</td>
<td>.133</td>
<td>19.580</td>
</tr>
<tr>
<td>14</td>
<td>-.088</td>
<td>.127</td>
<td>20.064</td>
</tr>
<tr>
<td>15</td>
<td>.156</td>
<td>.120</td>
<td>21.746</td>
</tr>
<tr>
<td>16</td>
<td>-.018</td>
<td>.113</td>
<td>21.771</td>
</tr>
</tbody>
</table>

a. The underlying process assumed is independence (white noise).
b. Based on the asymptotic chi-square approximation.

Table 3. Partial Autocorrelation function (PACF).

<table>
<thead>
<tr>
<th>Lag</th>
<th>Partial Autocorrelation</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-.567</td>
<td>.204</td>
</tr>
<tr>
<td>2</td>
<td>-.260</td>
<td>.204</td>
</tr>
<tr>
<td>3</td>
<td>-.355</td>
<td>.204</td>
</tr>
<tr>
<td>4</td>
<td>-.077</td>
<td>.204</td>
</tr>
<tr>
<td>5</td>
<td>-.194</td>
<td>.204</td>
</tr>
<tr>
<td>6</td>
<td>-.205</td>
<td>.204</td>
</tr>
<tr>
<td>7</td>
<td>-.094</td>
<td>.204</td>
</tr>
<tr>
<td>8</td>
<td>-.218</td>
<td>.204</td>
</tr>
<tr>
<td>9</td>
<td>.060</td>
<td>.204</td>
</tr>
<tr>
<td>10</td>
<td>-.188</td>
<td>.204</td>
</tr>
<tr>
<td>11</td>
<td>.003</td>
<td>.204</td>
</tr>
<tr>
<td>12</td>
<td>-.242</td>
<td>.204</td>
</tr>
<tr>
<td>13</td>
<td>-.008</td>
<td>.204</td>
</tr>
<tr>
<td>14</td>
<td>-.095</td>
<td>.204</td>
</tr>
<tr>
<td>15</td>
<td>-.096</td>
<td>.204</td>
</tr>
<tr>
<td>16</td>
<td>-.006</td>
<td>.204</td>
</tr>
</tbody>
</table>

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

\[
\begin{bmatrix}
1 \\
0.98 \\
0.49
\end{bmatrix}
\begin{bmatrix}
0.98 & 1 & 0.49 \\
0.49 & 0.50 & 1
\end{bmatrix}
\]

\[
\text{Det}(r) = 0.05
\]

\[
\log_e 0.05 = -3.00
\]

\[
\chi^2_{calculated} = 66.51 \quad \chi^2_{0.05 \ 3} = 7.82
\]

Since \(
\chi^2_{calculated} > \chi^2_{0.05 \ 3} \) i.e 66.51 > 7.82. Therefore, multicollinearity exists.

3.1.2. F-Test

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

\[
X_1 = f(X_2, X_3) \quad X_1 = \beta_2 X_2 + \beta_3 X_3 + U
\]

Thus, \[
\begin{bmatrix}
\beta_2 \\
\beta_3
\end{bmatrix} = (XX')^{-1} X Y
\]

The values obtained from the analysis for \( \beta_2 \) and \( \beta_3 \) 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
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 4 above shows the value for $\beta_0$, $\beta_1$, $\beta_2$ and $\beta_3$ 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

Table 5 showing the values for $\beta_0$, $\beta_3$ and $\beta_3$ 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 above pooled together and compared the results obtained before and after dropping variable $X_2$. 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 multicollinearity. 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


