IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN:2319-765X. Volume 10, Issue 3 Ver. I (May-Jun. 2014), PP 00-00
www.iosrjournals.org

Estimation of the Effect of Multicollinearity on the Standard Error for Regression Coefficients

[1]N.O Adeboye, I. S Fagoyinbo and [2]T.O Olatayo
[1.] Department of Mathematics & Statistics, Federal Polytechnic Ilaro. PMB 50, Nigeria
[2.] Department of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye, Nigeria

Abstract: *This research was set to examine the effect Multicollinearity has, on the standard error for regression coefficients when it is present in a Classical Linear Regression model (CLRM). A classical linear regression model was fitted into the GDP of Nigeria ,and the model was examined for the presence of Multicollinearity using various techniques such as Farrar-Glauber test, Tolerance level, Variance inflation factor, Eigen values etc and the result obtained shows that Multicollinearity has contributed to the increase of the standard error for regression coefficients, thereby rendering the estimated parameters less efficient and less significant in the class of Ordinary Least Squares estimators. Tolerance levels of 0.012, 0.005, 0.002 and 0.001 for*β_1*,* β_2*,* β_3 *,and* β_4 *respectively clearly shown a very low tolerance among all the explanatory variables with very high Variance Inflation Factors of 84.472, 191.715,502.179 and 675.633 respectively. A Coefficient of determination (R- Square) of 99%, though signaled a very high validity for the CLRM but it is equally an indications of a very high degree of Multicollinearity among the explanatory variables. The Eigen values of 0.431, 0.005, 0.002 and 0.000 for* $\beta_{0,}$ β_1*,* β_2*,* β_3 *,and* β_4 *respectively clearly shown a very low Eigen value among the explanatory variables, which are closer to zero with very high Condition index of 30.983, 49.759 and 100.810 for* β_2*,* β_3 *,and* β_4 *respectively which indicate that the Multicollinearity present is due greatly to the influence of regressors* X_2*,* X_3*, and* X_4*.*

Keywords: *Eigen values, Multicollinearity, Standard Errors , Tolerance Level ,Variance Inflation Factor*

I. Introduction

Multicollinearity is one of the important problems in multiple regression analysis. It is usually regarded as a problem arising out of the violation of the assumption that explanatory variables are linearly independent. However, just satisfaction of this assumption does not preclude the possibility of an approximate linear dependence among the explanatory variables and hence the problem of multicollinearity. Though no precise definition of multicollinearity has been firmly established in the literature. According to (Belsely, Kuh and Welsh, 1980), Multicollinearity is generally agreed to be present if there is an approximate linear relationship (i.e. shared variance) among some of the predictor variables in the data. In theory, there are two extremes: Perfect Multicollinearity and No Multicollinearity. In practice, data typically are somewhere between those extremes. Thus, multicollinearity is a matter of degree. Though some multicollinearity is almost present, the real issue is to determine the point at which the degree of multicollinearity becomes “harmful”. The econometric literature typically takes the theoretical position that predictor variable construct are not collinear in the population. Hence, any observed multicollinearity in empirical data is construed as a sample based “problem” rather than as representative of the underlying population relationship (Kmenta, 1986). In many marketing research situations, however it is unrealistic to assume that predictor variables will always be strictly orthogonal at the population level (especially when one is working with behavioral constructs). Regardless of whether Multicollinearity in data is assumed to be a sampling artifact or true reflection of population relationships, it must be considered when data are analyzed with regression analysis because it has several potential undesirable consequences parameters estimates that fluctuate dramatically with negligible changes in the sample, parameter estimates with signs that are wrong in terms of theoretical considerations, theoretically important variables with insignificant coefficients, and the inability to determine the relative important of multicollinearity variables. The regression coefficients though determinate, posses large standard errors which implies that the coefficients cannot be estimated with great accuracy (Gujarati and Porter, 2009). Hawking (1983), Bowerman and O’Connell (2006) states that the term multicollinearity refers to a situation in which there is an exact (or nearly exact) linear relation among two or more of the explanatory variables .Exact relations usually arise by mistake or lack of understanding. We can define multicollinearity through the concept of orthogonality; when the predictors are orthogonal or uncorrelated, all eigenvalues of the design matrix are equal to one and the design matrix is full rank. If at least one eigenvalue is different from one, especially when equal to zero or near zero, then non-orthogonality exists, meaning that multicollinearity is present (Vinod and Ullah, 1981).

Multicollinearity can lead to increasing complexity in the research results, thereby posing difficulty for researcher interpretation. Multicollinearity complicates interpretation as a function of its influence on the magnitude of regression weights and the potential inflation of their standard error, thereby negatively influencing statistical significance tests of these coefficients. Multicollinearity is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated. In this situation the Regression coefficient estimates may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data themselves; it only affects calculations regarding Dependent and independent variables used in statistics individual predictors. A high degree of multicollinearity can also prevent computer software packages from performing the matrix inversion required for computing the regression coefficients, or it may make the results of that inversion inaccurate. Though multicollinearity does not affect the goodness of fit or the goodness of prediction, it can be a problem if our purpose is to estimate the individual effects of each explanatory variable. Once multicollinearity is detected, the best and obvious solution to the problem is to obtain and incorporate more information. Other procedures have been developed instead, for instance, model re-specification, biased estimation, and various variable selection procedures. Greene, (2000) states that multicollinearity may be observed in the following situation: small changes in the data produce wide swings in the parameter estimates. Coefficients may have very high standard errors and low significance levels even though they are jointly significant and the R^2 for the regression is quite high and the coefficients may have the wrong sign magnitude.

Neter, (1989) said that in the process of fitting regression model, when one independent variable is nearly combination of other independent variables, the combination will affect parameter estimates. This problem is defined as multicollinearity. Basically, multicollinearity may cause serious difficulties. Variances of parameter estimates may be unreasonably large, parameter estimates may not be significant and a parameter estimate may have a sign different from what is expected.

The standard error of an estimate can be defined as the square root of the estimated error variance (δ^2) of the quantity. i.e. $S_E = \delta^2 = \sqrt{\delta^2} = \delta$.The larger the sample size, the smaller the standard error and the smaller the standard error the more representative the sample will be.

In our attempt to research into the challenges of Multicollinearity as it affects the standard errors of regression coefficients, a Gross Domestic Product (GDP) Model shall be specified as;

$GDP = f(CP, L, FR, FI) + \varepsilon$ (1)

When this model is written in explicit form it becomes

$GDP_i = ß_o + ß_1CPi + ß_2L_i + ß_3FR_i + ß_4FI_i + \varepsilon_i$ (2)

Where $ß_o$, $ß_1$, $ß_2$, $ß_3$ and $ß_4$ are estimable parameters

CP= Crop Production, L= Livestock, FR= Forestry, FI= Fishing, GDP= Gross domestic product income of Agriculture and ε = Error term.

This model shall be critically examined for Multicollinearity and hence establish reasonable inferences on its effects on the standard errors of regression coefficients.

II. Materials And Methods

To analyze the data collected, the classical linear regression model shall be fitted, the standard error for regression coefficients computed and the methods to be adopted in detecting multicollinearity will also be illustrated.

The CLRM model can be written in terms of the k-variable population regression function (PRF) model involving the dependent variable Y and k-1 explanatory variables X_2, X_3 , ..., X_k as:

$Y_i = \beta_1 + \beta_2X_{2i} + \beta_3X_{3i} + \dots\dots\dots + \beta_kX_{ki} + u_i$, i = 1,2,3,……,n (3)

Where, β_1 = the intercept

β_2 to β_k = partial slope coefficients

u = stochastic disturbance term

And i = i^{th} observation, n' being the size of the population.

This equation identifies k-1 explanatory variables (regressors) namely X_1, X_2, …….X_k and a constant term that assumed to influence the dependent variable (regressand).

The essence of regression in econometrics is to generalized for the population from what we get from the sample. For instance, the linear relationship from Equation (3) holds for the population only if we could obtain considerable values of Xs, Y and u which form the population values of these variables. Since this is impossible in practice, the alternative is to get sample observations for Xs and Y, specify the distribution of the u's and try to get satisfactory estimate of true parameters of the relationship.

This is done by fitting a regression line to the observed sample data as an approximation to the true line. If then the true relationship between Xs and Y is as given in Equation (3), the true regression line is

$E(Y_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots\dots\dots + \beta_n X_{ki}$ (4)

And the estimated relationship is:

$Y_i = b_1 + b_2X_{2i} + b_3X_{3i} + \dots\dots\dots + b_nX_{ki} + e_i$ (5)

Equation (3) is a shorthand expression for the following set of n simultaneous equations:

$$
\begin{aligned}
Y_1 &= \beta_1 + \beta_2X_{21} + \beta_3X_{31} + \dots\dots\dots + \beta_nX_{k1} + u_1 \\
Y_2 &= \beta_1 + \beta_2X_{22} + \beta_3X_{32} + \dots\dots\dots + \beta_nX_{k2} + u_2 \qquad (6) \\
. \quad & \quad . \qquad . \qquad . \qquad . \\
Y_2 &= \beta_1 + \beta_2X_{2n} + \beta_3X_{3n} + \dots\dots\dots + \beta_nX_{kn} + u_n
\end{aligned}
$$

We can write the system of equations (6) in matrix form as shown below:

$$
\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{21} & X_{31} \dots & X_{k1} \\ 1 & X_{22} & X_{32} \dots\dots & X_{k2} \\ \dots & \dots & \dots & \dots \\ 1 & X_{2n} & X_{3n} \dots\dots & X_{kn} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \qquad (7)
$$

$$
\begin{array}{ccccccc} Y & = & X & \beta & + & u \\ n \text{ x } 1 & & n \text{ x } k & k \text{ x } 1 & & n \text{ x } 1 \end{array}
$$

where,

Y = n x 1 column vector of observations on the dependent variable Y.

X = n x k matrix giving 'n' observations on k – 1 variables X_2 to X_k, the first column of 1's representing the intercept term.

β = k x 1 column vector of the unknown parameters $\beta_1, \beta_2, \dots\dots\dots, \beta_k$.

u = n x 1 column vector of n disturbances u_i.

Equation (3.1.5) can be written more compactly as:

$Y = X\beta + u$ (8)

To obtain the consistent estimators of β, we minimise the residual sum of square (SSE) which is normally given as $ESS = u'u$ (9)

But $u = Y – X\beta$ (10)

$$
\begin{aligned}
\text{Hence,} \qquad u'u &= (Y – X\beta)'(Y – X\beta) \\
&= Y'Y – \beta'X'Y – Y'X\beta + \beta'X'X\beta \\
&= Y'Y – 2\beta'X'Y + \beta'X'X\beta
\end{aligned}
$$

Since the transpose of a scalar is a scalar, thus;

$Y'X\beta = (Y'X\beta)' = \beta'X'Y$

$$
\begin{aligned}
\text{Thus, } \frac{\partial u'u}{\partial B/\hat{\beta}} &= - 2 X'Y – 2X'X\hat{\beta} = 0 \\
2X'X\hat{\beta} &= 2 X'Y \\
\hat{\beta} &= (X'X)^{-1} X'Y \qquad (11)
\end{aligned}
$$

Where equation (11) is the least square estimates for the parameters of a classical linear regression model.

The standard error for the regression coefficients are as

$SE(ß) = \quad \delta^2(X^IX)^{-1} \quad \text{or} \quad \delta \quad (X^IX)^{-1}$ (12)

Thus, Equation (3.3) is derived as follows

$$
\begin{aligned}
& V(ß)= E \quad (ß- ß)\,(ß- ß)^I \\
& E\,((X^IX)^{-1} X^IY – ß)(\,(X^IX)^{-1} X^IY – ß)^I \\
& \text{Where } Y= Xß+ \varepsilon \\
& E(X^IX)^{-1} X^I(Xß+ \varepsilon) – ß((X^IX)^{-1} X^I(Xß+ \varepsilon) – ß) \\
& E((X^IX)^{-1} X^IXß+(X^IX)^{-1} X^I \varepsilon – ß)((X^IX)^{-1} X^IXß+(X^IX)^{-1} X^I \varepsilon) \\
& E((X^IX)^{-1} X^I \varepsilon)(\,\varepsilon^I X\,(X^IX)^{-1}) \\
& E\,((X^IX)^{-1} X^I\varepsilon\varepsilon^I X\,(X^IX)^{-1}) \\
& ((X^IX)^{-1} X^IX)E(\varepsilon\varepsilon^I\,(X^IX)^{-1}) \\
& \text{Where } E(\varepsilon\varepsilon^I)= \delta^2 \\
& V(ß)= \delta^2\,(X^IX)^{-1} \qquad (13)
\end{aligned}
$$

In detecting the presence of multicollinearity on the standard error for regression coefficients, the following techniques shall be adopted.

- The Tolerance level
- The Variance inflation factor (VIF) and
- The Farrar- Glauber tests
- Eigen values and Eigen vectors

Tolerance Level

In multiple regressions, tolerance is used as an indicator of multicollinearity. Tolerance is estimated by $1-R^2$, where R^2 is calculated by regressing the independent variable of interest unto the remaining independent variables included in the multiple regression analyses. Researchers desire higher levels of tolerance, as low levels are known to affect adversely the result associated with a multiple regression analyses. The tolerance level is the $1-R^2$ value when each of the independent variables is regressed on the other independent variables.

Low tolerance levels indicate high levels of multicollinearity. Anytime a tolerance levels get somewhere below 0.40, then multicollinearity exist.

Variance Inflation Factor (VIF)

In multiple regressions, the VIF is used as an indicator of multicollinearity. Computationally, it is defined as the reciprocal of tolerance: $1\backslash1\text{-}R^2$. Researchers desire lower levels of VIF, as higher levels of VIF are known to affect adversely the result associated with a multiple regression analyses.

Infact, the utility of VIF, as distinct from tolerance is that VIF specifically indicates the magnitude of the inflation in the standard errors associated with a particular beta weight that is due to multicollinearity. VIF of over 2.50 start to indicate relatively high levels of multicollinearity.

Farrar-Glauber Test

Farrar and Glauber, (1967) also proposed a procedure for detecting multicollinearity which comprised of three tests (i.e Chi-square test, F-test and T-test) . The first one examines whether multicollinearity is present, the second one determines which regressors are collinear and the third one determines the form of multicollinearity.

Eigen values and Eigen vectors

When there are one or more close linear dependencies among the variables, one or more Eigen values (λ_1, λ_2, -----------,λ_p) of the correlation matrix will be smaller thanl the assessment of the matrix condition number (CN), given its symmetry, defining CN as the relation between the largest and smallest Eigen values. The authors point out that, if CN< 100, multicollinearity is not a serious problem. If 100 < CN < 1000, multicollinearity is moderate and if CN > 1000 there is severe multicollinearity (Montgomery and Peck, 1981). The analysis of the Eigen values can identify the approximate nature of the linear dependency existing between the variables Belsley, (1980). For these analysis, $R = V\Lambda V'$, where Λ is a diagonal matrix with dimensions p x p, (p is number of variables used to obtain the R correlation matrix), whose elements are the Eigen values λ_j(j = 1,2,…,p) of R, and V is an orthogonal matrix with p x p dimension whose columns (v_1,v_2, …,v_p) are the normalized Eigen vectors of R . An Eigen value (λj) close to zero indicates linear dependence among the observations. The elements of the Eigen vector (vj) associated with this Eigen value describe the nature of this independency.

III. Results And Discussion

The data collected is on the on the contribution of Agriculture products to the GDP of Nigeria for twenty years between year 1992 to 2011.

Fitting of Classical Linear Regression model

The required OLS model fitted into the collected data is given as

$$GDP = -487451.376 + 0.240CP + 8.929L + 279.598FR - 38.951FI \tag{14}$$

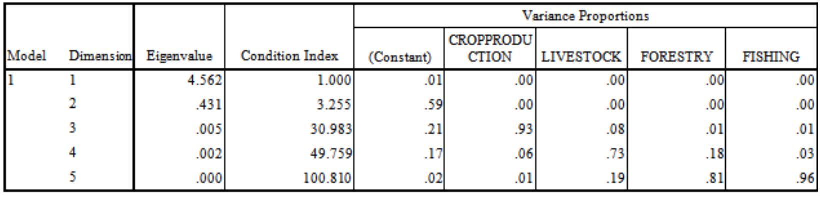
Table 1: Results of OLS Statistic

Statistic	Values
R	0.996
R^2	0.992
F	463.083 (with significant value of 0.000)
S.e (β_1)	0.691
S.e (β_2)	15.350
S.e (β_3)	125.177
S.e (β_4)	55.526

Table 2: Results of OLS Collinearity Statistics

REGRESSORS	**Tolerance Level**	**Variance Inflation Factor(VIF)**
Crop Production	0.012	**84.472**
Livestock	**0.005**	**191.715**
Forestry	**0.002**	**502.179**
Fishing	**0.001**	**675.633**

Table 3: Collinearity Diagonstics



Model	Dimension	Eigenvalue	Condition Index	Variance Proportions				
				(Constant)	CROPPRODUCTION	LIVESTOCK	FORESTRY	FISHING
1	1	4.562	1.000	.01	.00	.00	.00	.00
	2	.431	3.255	.59	.00	.00	.00	.00
	3	.005	30.983	.21	.93	.08	.01	.01
	4	.002	49.759	.17	.06	.73	.18	.03
	5	.000	100.810	.02	.01	.19	.81	.96

From model (14), the value of regression constant β_0 is negative, this ordinarily should have been positive due to the background econometric principles of the model. This is as a result of the presence of multicollinearity as propounded by Neter, (1989) . From table 1, the standard error for the regression coefficients (0.691, 15.350, 125.177, 55.526) for β_1,β_2,β_3 ,and β_4 respectively clearly shown a very high degree of multicollinearity among the explanatory variables except for that of $\beta_1$that is moderately fair. The coefficient of determination (R- square) clearly shows a very high value (99%), as well as the correlation coefficient of 0.996 which also indicates a very high degree of multicollinearity among the explanatory variables. The significant value shows 0.000 which is less than the significant level (0.05), this indicate that Multicollinearity gave a false impression as to the significant of the regression coefficients, which is too perfect to be real.

According to table 2, the Tolerance levels of 0.012, 0.005, 0.002 and 0.001 for β_1, β_2, β_3 ,and β_4 respectively clearly shown a very low tolerance level among all the explanatory variables with very high Variance Inflation Factors of 84.472, 191.715,502.179 and 675.633 respectively, except for that of $\beta_1$which is less than 100, indicate the presence of multicollinearity.

The Eigen value of 4.562, 0.431, 0.005, 0.002 and 0.000 for β_0,β_1, β_2, β_3 ,and β_4 respectively clearly shown a very low Eigen values among the explanatory variables. The eigen-values are closer to zero with very high Condition index of 30.983, 49.759 and 100.810 for β_2, β_3 ,and β_4 respectively which indicate that the Multicollinearity present is due greatly to the influence of regressors X_2, X_3, and X_4.

IV. Conclusion

Going by the analyzed results explained above, Multicollinearity has exerted significant effect on the standard error of regression coefficients. This finding supported the view of Neter, (1989) who noted that multicollinearity may cause serious difficulties in Regression analysis. Standard error of parameter estimates may be unreasonably large, parameter estimates may not be significant and a parameter estimate may have a sign different from what is expected. However, researchers should be aware that complete elimination of multicollinearity is not possible, but we can reduce the degree of multicollinearity present in the data. Hence, researchers are expected to always check first, for the presence of multicollinearity before fitting a Classical Linear Regression Model (CLRM) and if present, should be made to minimize its effects either through Model re- specification, Addition of new variables etc.

References

[1] Belsley, D.A., Kuh, E. &Welsch, R.E.(1980), "Regression Diagnostics: Identifying influential Data and Sources of Collinearity". John Wiley & Sons., New York.

[2] Bowerman, B.L and O" Connell, R.T (2006), "Linear Statistical Models an Applied Approach" Boston: PWS-KENT Publishing

[3] Farrar and Glauber (1967), "Multicollinearity in regression analysis" review of economics and statistics, 49, pp. 92-107

[4] Greene (2000), "Econometric Analysis" .Fourth edition, Upper Saddle River, NJ:Prentice- Hall.

[5] Gujarati, D.N. and Porter, D.C. (2009**):** Basic Econometrics. 5th ed. Mc Graw-Hill, New York. Pp 320-351Haitovsky (1969), "Multicollinearity in Regression analysis Comment," review of economics and statistics, 50, pp. 486-489

[6] Hawking, R. R. and Pendleton, O. J. (1983), "The regression dilemma" Commun. Stat.- Theo. Meth, **12**, 497-527.

[7] Hoerl, A. E., and R. W. Kennard, (1970), "Ridge regression: biased" estimation for non-orthogonal problems. Technometrics, **12**, 55-67.

[8] Kmenta (1970) "Elements of econometrics" Macmillan publishing co. New York.

[9] Kumar (1975),"The problem of multicollinearity: A survey" unpublished mimeo, abt associates,Inc., Cambridge, Mass.

[10] Larsen&Marx(1986),"An Introduction To Mathematical Statistics and its Application" 2nd edition printice hall New Jersey.

[11] Mendenhall&Sincich(2003), "regression analysis a second course in statistics"printice hall New Jersey.

[12] Montgomry&peck(1981), "Introduction to Linear Regression Analysis", New York, NY: Wiley

[13] Neter,Wasserman&Kutner (1989), "Applied Linear Regression Models" 2nd edition.Irwin, Homewood IL.

[14] Weisberg (2005), "Applied Linear Regression" John Wiley &Sons. New York