Detection and Control of Multicollinearity in Multiple Regression using Farrah-Glauber and Variance Inflation Function (VIF) Methods

Chikezie Onyeka Okonkwo[1], Victoria Anita Oware[2]

[1]Department of Mathematical Sciences, Ajayi Crowther University, Oyo, Nigeria

[2]Guaraty Trust Bank PLC, Ibadan, Nigeria

Corresponding author: co.okonkwo@acu.edu.ng

Abstract

Multicollinearity occurs when predictor variables in a regression model are highly correlated, leading to challenges in interpreting regression coefficients and unreliable p-values. This study investigates two classical methods for detecting multicollinearity: the Farrah-Glauber test and VIF. The Farrah-Glauber test assesses the linear dependence between predictor variables, while VIF quantifies the correlation strength. In addition, the backward elimination variable selection method was applied to choose the more significant explanatory variables among the three explanatory variables. The data was obtained from the Federal Trade Commission (FTC). The data is on the annual varieties of domestic cigarettes according to Tar, Nicotine and Carbon-monoxide content present. Multiple linear regression model was employed the parameter estimates were obtained. Although the Farrah- Glauber test showed that two of the three variables considered were correlated, VIF went beyond. The VIF method showed that although there was multicollinearity among all three explanatory variables. We illustrate these methods using empirical data and discuss their implications for model stability and coefficient estimation. Researchers and practitioners can apply these techniques to enhance the reliability of regression analyses by identifying and addressing multicollinearity effectively.

Keywords: Multicollinearity, Farrah-Glauber Test, Variance Inflation Function (VIF), Regression, Backward Elimination.

INTRODUCTION

Multicollinearity is the term accorded by econometricians to denote strong linear relationships (e.g. collinearity) amongst explanatory variables in a multiple linear regression analysis. Indeed, multicollinearity and its effects on statistical inference are well-explored concepts in Statistics, Economics and many engineering fields. economist engaged in an applied study and most other social science-related statistical research (Ayinde *et al.,* 2015). Multicollinearity refers to the existence of a high degree of correlation among two or more independent variables in a multiple linear regression model. In simpler terms, it occurs when one independent variable can be reasonably predicted from a linear combination of other independent variables. This redundancy in information creates a scenario where it becomes difficult to isolate the unique effect of each independent variable on the dependent variable (Akintunde *et al.* 2021).

When an empirical study yields a particular regression parameter that is statistically insignificant (i.e. one that fails to reject the appropriate hypothesis test), there are two reasons why this occurs: (i) the true value of the parameter of interest is, in fact, zero, or (ii) the data sample is not informative enough to conclusively distinguish this parameter as statistically significantly different from zero. Since the latter is strongly related to the degree of collinearity in explanatory variables, a researcher particularly interested in demonstrating a statistically significant relationship may, therefore, be motivated not only to employ techniques intended to alleviate (ii), but also to search for a reasonable justification of rejecting (i) informally by arguing that the failure of the significance test is more attributable to (ii) through the presence of multicollinearity.

Let there exist a multiple linear regression model given by

$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i$ (1)

Where y_i is a collection of dependent variables that are normally distributed with a given mean and variance, $y \sim N(\mu_y, \sigma_y^2)$, each X is an exogenous independent or source variable, hence X is an $n \times (k+1)$ matrix of independent variables, $\mathrm{B}' = (\beta_0, \beta_1, \beta_2, \dots, \beta_k)$ is a vector of regression parameters and ϵ_i is the ith error term associated with y_i. Given the model in (1) above, some assumptions suffice and the execution or relaxation of these assumptions that determine the type of estimation procedure that will be most efficient for the regression parameters.

The presence of multicollinearity can wreak havoc on the interpretation and reliability of regression results, some of the ways that it does this is that it leads to an inflated standard error of the regression coefficients, hence it affects the statistically significant effect of the dependent variable. Furthermore, in cases where the independent variable even appears to be significant, the inflated standard error casts doubt on the precision of its estimated coefficient. In addition to these, p-values become unreliable in the presence of multicollinearity. A seemingly insignificant variable might become significant with minor changes to the model, or vice versa. This instability undermines our confidence in the model's ability to accurately represent the true relationships between variables (Olanrewaju *et al.,* 2017).

There are sources of multicollinearity and these are attributable to the following factors: The method employed for data collection, the presence of constraint on the model or in the population being sampled, wrong specification of the model and over-determination of a model. Overspecification occurs when there is more explanatory (independent) variable than the number of observations. This study aims to identify and control the multicollinearity problem in a regression model the objectives are to detect the presence of multicollinearity and to obtain models that give better estimates of the regression parameters (coefficients). Also, the paper will check which method is the best to detect multicollinearity.

METHODOLOGY

Several estimation methods have been developed to detect and control Multicollinearity. Among such methods to detect multicollinearity and methods to control it are Farah-Glauber test (chi-square, F-distribution and t-distribution).

One of the most reliable tests for detecting multicollinearity in a data is the Farrah-Glauber test. The test is conducted following three stages: (i) Conduct the chi-square test to locate the existence and severity of multicollinearity, (ii) Carry out the F-test to locate the variance(s) that are intercorrelated if the chi-square test is positive, and (iii) Conduct t-test to detect the variable(s) that are responsible for multicollinearity if the F-test is positive.

The chi-square test

This is computed using the pairwise correlation matrix r_{ij} for the explanatory variables X_1 , X_2 and X_3 and then their determinant D.

$$r_{ij} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{pmatrix}, \quad (2)$$

$\rho_{12} = \rho_{21} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$, $\rho_{13} = \rho_{31} = \frac{\sigma_{13}}{\sqrt{\sigma_1^2 \sigma_3^2}}$ and $\rho_{23} = \rho_{32} = \frac{\sigma_{23}}{\sqrt{\sigma_2^2 \sigma_3^2}}$

And $D = |r_{ij}|$ is the determinant of the correlation matrix.

distribution with ½ k (k $-$ 1) degrees of freedom. The chi-square test is given as

$$\chi^2 = -\left[T - 1 - \frac{1}{6}(2k + 5)\right] \log_e D \quad (3)$$

where T is the number of observations, k is the number of explanatory variables, and D is the determinant of the correlation matrix r_{ij}. A significant difference in the χ^2 values indicate the presence of multicollinearity among the variables.

The F-test

Having established the presence of multicollinearity among the variables. The F-test follows that for one of the variables, say X1 which is suspected to be intercorrelated with other X variables. Thus, $X_1 = f(X_2, X_3, \dots, X_k)$, for $X = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}$

So, $X_1 = \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$

Using least square estimation B $= \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (X'X)^{-1}X'X_1$

The F-test is a test for collinear regressors. The test is achieved using the coefficients of determination $R^2 = \frac{\hat{\beta}' X' X_1 - \frac{(\sum X_1)^2}{T}}{X_1' X_1 - \frac{(\sum X_1)^2}{T}}$

The test statistic ω is given by

$$\omega = \frac{R_i^2}{1 - R_i^2}\left(\frac{T-k}{k-1}\right) \quad (4)$$

where ω is F-distributed with $k-1, T-1$ degrees of freedom. The result is compared with $F_{0.05, k-1, T-k}$.

The required hypothesis is given as

H_0: X_1 is not intercorrelated with X_2 and X_3

H_1: X_1 is intercorrelated with X_2 and X_3

The process is repeated the test for other $X's$ suspected to be intercorrelated with others

The t-test

This is carried out when the F-test is positive. The t - test is conducted to detect which pair of variables is responsible for multicollinearity. According to the F-test performed to test for intercorrelation between X_1 and X_2, X_3. Our F –test, is found to be intercorrelated with X_2 and X_3 then we conduct the t – test by stating the hypothesis

H_0: The variables X_2 and X_3 are NOT responsible for the multicollinearity

H_1: The variables X_2 and X_3 are responsible for the multicollinear

the partial coefficient of variables $r_{12.3}$ is given by

$$r_{12.3} = \frac{(r_{12} - r_{13}r_{23})^2}{(1 - r_{13}^2)(1 - r_{23}^2)} \quad (5)$$

The test can be carried out using the hypotheses

H_0: $r_{12.3} = 0$

H_1: $r_{12.3} \neq 0$

The test statistics is given by

$$t_{12} = \frac{r_{12.3}\sqrt{T-k}}{\sqrt{1-r_{12.3}^2}} \quad (6)$$

Which follows the t-distribution with $T - k$ degrees of freedom

Variance Inflation Factor

Variance Inflation Factor (VIF) is a measure of the of impact of collinearity on the precision of the estimates of the degree of inflation.

Variance Inflation Factor $VIF(X) = \frac{1}{1-R_j^2}$ is from the explanatory variables of the regression equation. When the VIF is greater than 10 is shows that multicollinearity exist. Asymptotic tests and simultaneous confidence bands for the parameter function have been obtained by using this dimension reduction approach. An estimation procedure based on B-splines expansion maximizing the penalized log-likelihood has been studied in Marx and Eilers (1999) for a functional binomial response model and in Cardot and Sarda (2005) for the general case of functional generalized linear models.

Variable Selection Procedures

Consider a time series with functional predictor $(X\ (t): t\ \in T)$ *;* whose sample curves belong to the space $L2(T)$ of square integral functions on *T;* and a categorical response random variable *Y* with *s* categories. Given a sample of observations of the functional predictor $(x_i(t) : t \in T, i = (1 \dots \dots n)$, the sample of observations of the response associated to them is a set of *n* vectors $(y_{i1}, \dots, y_{in})'$ of dimension *s* is defined by

$$y_{is} = \begin{cases} 1, if\ category\ is\ observed\ X_i = x_i(t) \\ 0, otherwise \end{cases}$$

Hence each observation is given by the multinomial distribution, that is to say $X_i \sim M_t(s; p_i)$, with each $p_i = P(y_i = s)$,

So that

$$P\big(X_i = x_i(t)\big) = \frac{n!}{x_1!\, x_2! \dots x_t!} p_1^{x_1} p_2^{x_2} \dots\ p_t^{x_t}$$

Where $\frac{n!}{x_1! x_2! \dots x_t!} = \binom{n}{x_1 x_2 \dots x_t}$ and $\sum_{i=1}^{t} p_i = 1$

Let us observe that $y_i's$ are redundant. Then, if we denote by$y_i = (y_{i1}, y_{i2}, \dots, y_{is})'$ then *the* vector response for subject i, with mean vector $\mu_i = E(y_i)$ is the multinomial response model in a particular case of generalized linear model.

Results and Discussions

Description of the Variables

The first step to observing the variables would be to take a look at the description of the variables.

Table 1: Descriptive Statistics

	Mean	Std. Deviation	N
Carbon-monoxide	12.53	4.740	25
Tax_X1	12.22	5.666	25
Nicotine_X2	.88	.354	25
Weight_X3	.97	.088	25

Table 2: ANOVA Table

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	495.258	3	165.086	78.984	.000b
	Residual	43.893	21	2.090		
	Total	539.150	24			

a. Dependent Variable: Carbon-monoxide
b. Predictors: (Constant), Weight_X3, Tax_X1, Nicotine_X2

Table 3: Significance Table

Model		Unstandardized Coefficients		t	Sig.		
		B	Std. Error			Tolerance	VIF
1	(Constant)	3.202	3.462	.925	.365		
	Tax_X1	.963	.242	3.974	.001	.046	21.631
	Nicotine_X2	-2.632	3.901	-.675	.507	.046	21.900
	Weight_X3	-.130	3.885	-.034	.974	.750	1.334

Table 1 simply shows the mean and standard deviation of the independent variables (Tax, Nicotine and Weight for 25 observations, while table 2 is a description of the regression variables. Table 3 shows that among the variables only Tax seems to significantly contribute to volume of carbon-monoxide being dispensed to the public. The first step in the discussion of Farrah-Glauber test is to obtain the pairwise correlation of the explanatory variables. The results are as shown in table 4 below.

Table 4: Correlations Matrix

		Tax_X1	Nicotine_X2	Weight_X3
Tax_X1	Pearson Correlation	1	.977**	.491*
	Sig. (2-tailed)		.000	.013
	N	25	25	25
Nicotine_X2	Pearson Correlation	.977**	1	.500*
	Sig. (2-tailed)	.000		.011
	N	25	25	25
Weight_X3	Pearson Correlation	.491*	.500*	1
	Sig. (2-tailed)	.013	.011	
	N	25	25	25

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

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

The result show to a strong correlation between Tax and Nicotine of 0.977 with a strong significance of 0.00, which is significant even at a p-value of 0.01 hence we see that using the two variables in estimation will yield an erroneous conclusion. The correlation matrix is given as

$r_{ij} = \begin{pmatrix} 1 & 0.977 & 0.491 \\ 0.977 & 1 & 0.500 \\ 0.491 & 0.500 & 1 \end{pmatrix}$ and the determinant of the matrix $D = |r_{ij}| = 0.034097$ the size of D is indicative of the presence of multicollinearity among the variables.

The following hypothesis holds

H_0: There is no multicollinearity among the explanatory variables

H_1: There is multicollinearity among the explanatory variables

The closer the value D is to zero the stronger the presence of multicollinearity among the variables.

The chi-square test $\chi^2 = 74.8911$ when compared with $\chi^2_{0.05}$ with $v = \frac{1}{2}k(k-1)$ degrees of freedom. $\chi^2_{0.05, v=\frac{1}{2}k(k-1)} = 7.83$

Since $\chi^2 > \chi^2_{0.05,v}$ we reject the null hypothesis and conclude that multicollinearity does exist among the explanatory variables.

- **F-Test**

Following after the chi-square test is the F-test for testing X_1 (tax) as a response variable against X_2 (Nicotine) and X_3 (Weight). The results show that $R^2 = 0.95$

Table 5: R squared results

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.977[a]	.954	.950	1.2723920

Table 6: ANOVA table

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	734.816	2	367.408	226.938	.000[b]
	Residual	35.618	22	1.619		
	Total	770.434	24			

a. Dependent Variable: Tax_X1

b. Predictors: (Constant), Weight_X3, Nicotine_X2

Table 7: Variable Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.650	3.026		-.545	.591
	Nicotine_X2	15.604	.847	.975	18.419	.000
	Weight_X3	.197	3.419	.003	.058	.955

a. Dependent Variable: Tax_X1

From the foregoing, the test statistic $\omega = \frac{R_i^2}{1-R_i^2}\left(\frac{T-k}{k-1}\right) = 238.5$, and from the F-distribution table, $F_{0.05, k-1, T-k} = 3.49$. Given that $\omega > F_{0.05, k-1, T-k}$, we reject the null hypothesis and conclude that X_1 is intercorrelated with X_2 and X_3.

- **t-Test**

The t-test is also carried out in the same fashion and the result

Table 8: Partial Correlations

Control Variables			Tax_X1	Nicotine_X2
Weight_X3	Tax_X1	Correlation	1.000	.969
		Significance (2-tailed)	.	.000
		df	0	22
	Nicotine_X2	Correlation	.969	1.000
		Significance (2-tailed)	.000	.
		df	22	0

The table 8 above shows that partial correlation using X_3 (weight) as the controlling variable. The result shows that there is strong partial correlation of 0.969 among the variables, $r_{12.3} = 0.96$

As with any other functional regression model, estimating the parameters of a functional multinomial response model is an ill-posed problem due to the infinite dimension of the predictor space. See Ramsay and Silverman (2005) for a discussion on the functional linear model. In addition, the functional predictor is not observed continuously in time, so sample curves xi(t) are observed in a set of discrete time points {tik: k = 1…….mi} that could differ for each sample individual. The most used solution to these problems is to reduce dimension by performing a basis expansion of the functional predictor.

Data Analysis

The second method to detect the existence of multicollinearity is:

Variance Inflation Factor

Table 9: The summary of the fitted model

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.958[a]	.919	.907	1.44573	.919	78.984	3	21	.000

a. Predictors: (Constant), Weight, Tar, Nicotine

From table 9, coefficient of determination was obtained to be 0.919 which means 91% of the dependent variable was explained by explanatory variables present in the model; which means with p value of 0.000, all the variable are significant because

Table 10: Parameter estimate with VIF value for the fitted model

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

Variable Selection Method

From the analysis above it has shown that multicollinearity exist, the next thing is how to correct it. In order to correct the existence of multicollinearity in this research work variable 3 i.e X_2 will first remove and check the result.

Table 11: The summary of the fitted model for selected variables with variable selection method

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.958a	.917	.909	1.42771	.917	121.251	2	22	.000

a. Predictors: (Constant), Weight, Tar

From table 11 coefficient of determination was obtained to be 0.917 which means 91.7% of the dependent variable was explained by the explanatory variables present in the model, which means with p value of 0.000, all the variable are significant because p<α.

Table 12: The analysis of variance of the fitted model when X_2 was removed from the model-by-model selection procedure.

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	494.306	2	247.153	121.251	.000a
	Residual	44.844	22	2.038		
	Total	539.150	24			

a. Predictors: (Constant), Weight, Tar

b. Dependent Variable: Carbon Monoxide

From table 12 Sum of Square Regression was obtained to be 494.306, Sum of Square for Residual to be 44.844 and Sum of Square for Total was obtained to be 539.150 with significant value 0. 000.Which informed us that the parameter of the model are significant, since P=0.000 < α.

Table 13: Parameter estimate value with Variance Inflation Factor values with the correlation and collinearity statistics

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 13 showing the value for β0, β1, β2 and β3with its VIF value. From VIF value for X1 and X3 the value on the table is normal which means multicollinearity is not exist again.

Table 14: The summary of the fitted model

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.926a	.857	.844	1.86954	.857	66.128	2	22	.000

a. Predictors: (Constant),
Nicotine, Weight

From table 14 above, coefficient of determination was obtained to be 0.857 which means 85.7% of the dependent variable was explained by explanatory variables present in the model, which means with p value of 0.000, all the variable are significant because $p<\alpha$.

Table 15: The analysis of variance for the fitted model after X1 was removed from the model by variable selection procedure.

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	462.256	2	231.128	66.128	.000a
	Residual	76.894	22	3.495		
	Total	539.150	24			

a. Predictors: (Constant), Nicotine, Weight

b. Dependent Variable: Carbon Monoxide

Table 15 showing the Sum of Square Regression to be 462.256, Sum of Square Residual to be 76.894 and Sum of Square Total be 539.150 with significant value 0.000. It tells us that the model are significant.

Table 16: Parameter estimate with VIF value for the fitted model

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	1.614	4.447		.363	.720					
	Weight	.059	5.024	.001	.012	.991	.464	.002	.001	.750	1.334
	Nicotine	12.388	1.245	.925	9.952	.000	.926	.905	.801	.750	1.334

a. Dependent Variable: Carbon Monoxide

Table 16 above shows the value for β_0, β_1, β_2 and $\beta_3$with its Variance Inflation Factor value. From VIF value for i.e X_2 and X_3 the value on the table is normal which means multicollinearity is not present again also.

Table 17: The summary of all analysis of the parameter coefficient, Standard error and variance inflation factor value.

	β_0	β_1	β_2	β_3
Presence of multicollinearity	3.202 (3.416)	0.963 (0.242) *21.631	-2.632 (3.901) *21.900	-0.130 (3.885) *1.334
Absence of multicollinearity When X_1 removed	3.114 (3.416)	0.804 (0.059) *1.317		-0.130 (3.813) *1.317
Absence of multicollinearity When X_2 removed	1.614 (3.416)		0.059 (5.024) *1.334	12.388 (1.245) *1.334

Note: The value in the parenthesis is standard error of estimate of the coefficient of predictions in the model. And the * value is VIF value.

Discussion

When Farrah-Glauber test was used it shows the existing of multicollinearity, however when VIF method were employed, it showed that all the variables showed the presence of multicollinearity. The Farrah-Glauber test indicates a higher presence of multicollinearity than the Variance Inflation Function (VIF).

The model showed the presence of multicollinearity exist when all the independent variables were analysed together, however, when variable the second variable X_2 $(weight)$ was removed using the backward elimination method, it shows that none of VIF value was more than 10 and the adjusted R^2 was 0.909. Finally, it was shown that variable X_1 provided a more significant contribution to the model than variable X_2 given that contributed more significantly to the model

There was an introduction to the concept of multiple linear regression, multicollinearity and definition of the aims and objectives of this study. Furthermore, a general outlook to ordinary least squares regression (OLS), multicollinearity in regression analysis, its effect on the least square regression, methods of detecting multicollinearity and also reviews of relevant literature of published work done in relevant to this research topic was done. A discussion into detail on how to detect (using Farrah-Glauber test and VIF method) and control multicollinearity using variable selection method.

CONCLUSION

This current study focuses on how to use variable selection method to handling multicollinearity justified by the fact that the OLS estimates fail in the presence of multicollinearity. The method of Farrah-Glauber test and VIF method is used to test presence of multicollinearity on a given data set to see how to perform and made some conclusion. Variable selection method is also used to control multicollinearity, we can now conclude that when multicollinearity issue occurred the best way is to drop one of the variables that are collinear to each other.

REFERENCES

Aldrich, J. (2005). Fisler and regression. *Statistical Science, 20*(4), 401-417.

Akintunde, M. O., Olawale, A. O., Amusan, A. S., & Azeez, A. I. A. (2021). Comparing Two Classical Methods of Detecting Multicollinearity in Financial and Economic Time Series Data. *International Journal of Applied Mathematics and Theoretical Physics*, *10*(1), 62-67.

Armstrong, J. S. (2012). Illusions in regression analysis. *International Journal of Forecasting, 28*(3), 689.

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.

Becker, A. J., & Hoover, E. M. (1998). Population growth and economic development in low-income countries. Princeton: Princeton University Press, 610-619.

CIA World Facts Books. (2011). Vol. 5, No. 1, 89-95.

Cressie, N. (1996). Change of support and the modifiable area unit problem. *Geographical Systems, 3*, 169-180.

Freedom, D. A. (2005). Statistical models: Theory and practice. *Cambridge*: Cambridge University Press.

Fotheringham, A. S., & Wong, D. W. S. (2002, January 1). The modifiable area unit problem in multivariate statistical analysis. *Environment and Planning A, 23*(7), 1025-1044.

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). Applied linear regression models. 25(7), 25-31.

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.

Smith, J. B. (1997). Effects of eighth-grade transition programs on high school retention and experiences. *The Journal of Educational Research, 90*(3), 144-152.

Thirlwall, D. (2007). Growth and development with special reference to developing economies. University of Kent at Canterbury, 5th Edition, 111(8), 143-155.

Tofallis, C. (2009). Least squares percentage regression. *Journal of Modern Applied Statistical Methods*, 7, 526-534.

Waegeman, W., De Baets, B., & Boullart, L. (2009). ROC analysis in ordinal regression learning. *Pattern Recognition Letters, 29*, 1-9.

Yang, J. L. (2009). Human age estimation by metric learning for regression problems. 21(6), 74-82.