International Journal of
Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: 10.31695/IJASRE.2019.33217

Volume 5, Issue 5
May - 2019

Investigating the Effects of Multicollinearity on the Model Parameters of Ordinary Least Squares Estimator

Nusirat Funmilayo Gatta[1] and Banjoko Alabi Waheed[2]

[1,2]Faculty of Physical Sciences, Department of Statistics,

University of Ilorin, Ilorin, Kwara State

Nigeria

ABSTRACT

This study investigated the effects of multicollinearity on the model parameters of the ordinary least squares regression model. The aim was to examine the impacts of multicollinearity on the efficiency of classical Ordinary least squares (OLS). Data were simulated from a multivariate normal distribution with mean zero and variance-covariance matrix at various sample sizes 25, 50, 100, 200, 500 and 1000. To assess the asymptotic efficiency and consistency of the regression models in the presence of multicollinearity, the evaluation criteria used were the Variance, Absolute bias, Mean Square Error (MSE) and Mean Square Error of Prediction (MSEP). Results from the analysis revealed that the OLS is not efficient given the large MSE, MSEP, and Absolute bias.

Keywords: Ordinary least squares, Multicollinearity, Mean Square Error, Absolute Bias, Mean Square Error of Prediction

1. INTRODUCTION

The problem that often arises when the assumption of OLS "that the independent variables are independent of each other in the regression model in the sense that they do not move together in the same pattern" is disregarded is Known as multicollinearity problem. The word multicollinearity refers to a situation where there are perfect or near linear relationships among some (or all) of independent variables in the model. Exact multicollinearity occurs when there is a perfectly linear relationship among the explanatory variables and violates the assumption that data matrix X is full rank. In the case of perfect multicollinearity, the correlation coefficient of these variables is equal to unity and inverse of (X^1X) does not exist since determinant of X^1X is zero, making OLS estimate impossible. Recall the Gauss- Markov theorem which states that among all linear unbiased estimators, the least squares estimator has the smallest variance. consider a regression model that has two explanatory variables and a constant. The variance of parameter estimate is

$$\left.\begin{aligned} V(\beta_k) &= \frac{\sigma^2}{(1-r_{12}^2)\sum_{i=1}^{n}(X_{i-\bar{X}})} \\ &= \frac{\sigma^2}{(1-r_{12}^2)S} \end{aligned}\right\} \qquad (1)$$

Where r_{12} is the correlation between X_1 and X_2. If the two variables are perfectly correlated, then the variance is inestimable. The case of an exact linear relationship among the predictors is a serious failure of the assumptions of the model not of the data. When predictors are highly correlated but not perfectly correlated that is the column of X is close to linearly independent then variables are suffering from severe multicollinearity problem though the regression model retains all its assumed properties. The most common implication of severe multicollinearity is that individual parameter estimates will not be precise and the method of OLS breaks down. (Bruce 2008). Considering the regression model with two explanatory variables,

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad (2)$$

$X = (1 \quad X_1 \quad X_2)$

and the variance of β is:

$$V(\beta) = \frac{\sigma^2}{n}\begin{bmatrix}1 & \rho \\ \rho & 1\end{bmatrix}^{-1} \qquad (3)$$

The correlation ρ indicates collinearity between X_1 and X_2 , as this correlation approaches 1 the matrix X becomes singular and variance of a coefficient estimate $\sigma^2(1-\rho^2)^{-1}$ approaches infinity. If the predictors are not dependence that is the correlation coefficient for these variables are zeros, the Eigenvalues of the data matrix X are equal to one and matrix X is of full rank such variables are called orthogonal or uncorrelated variables. On the other hand, if the variables are nonorthogonal (correlated), at least one of the Eigenvalue will be close to zero. (see El-Dereny and Rashwan (2011))

Freund and Litell (2000), Batterham *et al* (1997), Wax (1992), Cavell *et al* (1998), Mofenson *et al* (1999), Elmstahl *et al* (1997), Parkin *et al* (2002) and Kinta *et al* (2002) indicated that collinearity leads to imprecise estimate of parameters, increases the estimate of standard error of coefficients, causing wider confidence interval and increasing the chance to reject the significance of the test statistic.

2. MATERIALS AND METHODS

This study makes use of multiple regression model where n sample observations of a dependent variable Y, explanatory variable X and the relationship between X and Y are observed. The project commences with the case of a K regressors that is, k= 1, 2 …K. and this is written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Where β_i (i=0, 1, 2 …k) are the regression coefficient and ε is the error term. Thus, the linear equation can be written in matrix form as:

$$Y = X\beta + \varepsilon$$

$$y = \begin{bmatrix}y_1 \\ y_2 \\ . \\ . \\ y_n\end{bmatrix}, \quad X = \begin{bmatrix}1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ . & . & \cdots & . \\ 1 & x_{n1} & \cdots & x_{nk}\end{bmatrix}, \quad \beta = \begin{bmatrix}\beta_1 \\ \beta_2 \\ . \\ . \\ \beta_k\end{bmatrix}, \quad \epsilon = \begin{bmatrix}\epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n\end{bmatrix}$$

Where Y is a vector of $n \times 1$ observations of the dependent variable, X is an $n \times (k+1)$ matrix of independent variables, β is a $(k+1) \times 1$ vector of unknown parameters, and ε is an $n \times 1$ vector of errors $\varepsilon \sim N(0, \sigma^2 I)$.

3. ORDINARY LEAST SQUARES (OLS) METHOD

The least square estimators $\hat{\beta}_i$ of β_i $(i = 0, 1, 2 \ldots k)$ are the ones that minimize the sum of squares

SSE= $\sum_{i=1}^{n}(y - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2$

SSE= $\left(y - x\hat{\beta}\right)^1 \left(y - x\hat{\beta}\right)$

Differentiate SSE concerning the parameters and equate to zero to minimize SSE that is,

$$\frac{dSSE}{d\beta} = -2\, x^1 y\beta + 2\, x^1\, x\, \beta$$

$$-2\, x^1 y\beta + 2\, x^1 x\, \beta = 0$$

$$\hat{\beta}_{ols} = (x^1 x)\, x^1 y$$

These estimates are unbiased so that the expected values of the estimates are the population value that is

$$E(\beta) = \beta$$

The variance-covariance matrix of the parameter is

DOI: 10.31695/IJASRE.2019.33217

$V(\beta) = \sigma^2(X^1X)^{-1}$

Absolute bias $= \frac{\sum_{i=1}^{k}|\beta_{OLS} - \beta|}{k}$

MSE= $\frac{\sum_{i=1}^{k}(\beta_i - \beta)^2}{k}$

MSEP = $\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}$

4. Simulation Study

In this research work, five explanatory variables values were generated with the same sample size n**.** The sample sizes are 25, 50, 100, 200, 500 and 1000 each of this sample sizes were generated over 1000 iteration in order to inspect the effect of the estimators with respect to this sample sizes. Data were generated using equation (3.1) above but in this case, we make use of five predictors and the regression parameters β_0, β_1, β_2, β_3, β_4 and β_5 are set to be 20, 40, 30, 50, 80 and 55 respectively. The variables are generated using multivariate normal distribution and make use of two correlation structures, these are:

$$\begin{pmatrix} 1 & 0.7 & 0.7 & 0.7 & 0.8 \\ 0.7 & 1 & 0.8 & 0.8 & 0.93 \\ 0.7 & 0.8 & 1 & 0.93 & 0.93 \\ 0.7 & 0.8 & 0.93 & 1 & 0.95 \\ 0.8 & 0.93 & 0.93 & 0.95 & 1 \end{pmatrix}$$

The error term was generated using a normal distribution with mean zero and variance ten (10), and the relationship below is used to generate the regress and variable

$$y = 20 + 40X_1 + 30X_2 + 50X_3 + 80X_4 + 55X_5 + \epsilon_i$$

The purpose of all these simulation studies is observe the performance of OLS estimator through the absolute bias, MSE and MSEP in order to examine the predictive ability of the estimator.

5. RESULTS

Considering positive high correlation structure that is, r1=r2=r3=0.7; r4=r5=r6=0.8; r7=r8=r9=0.93; r10=0.95 the VIF for the simulated data set are as follows:

Table 1: VARIANCE INFLATION FACTOR (VIF) OF THE VARIABLES

SAMPLE SIZE	X1	X2	X3	X4	X5
n=25	5.747766	41.27515	16.20155	56.92161	275.6749
n=50	6.210466	41.5237	10.504	41.00265	211.9157
n=100	9.433733	56.20102	13.50942	60.03766	328.5111
n=200	5.978252	30.85184	9.088591	38.95834	174.0173
n=500	5.762636	33.79607	11.28923	41.32586	197.7226
n=1000	6.219904	37.52424	11.57944	44.70781	209.2167

From the above table, it can be seen that the variance inflation factor of the variables are more than ten (10) when the correlation between the explanatory variables was very high with different sample sizes except for X1 variable. Then, it is clearly shown that multicollinearity problem exists. Using the method of OLS analyze the simulated data, the following results is obtained:

DOI: 10.31695/IJASRE.2019.33217

Table 2: Summary of the estimate of coefficient and variance using OLS Method at simulated sample size 25-1000

SAMPLE SIZE	OLS ESTIMATE		
		Coefficient	Variance
$n = 25$	β_0 =20	20.09694	5.30926
	β_1 =40	40.05853	34.64088
	β_2 =30	30.95643	197.7324
	β_3 =50	49.88632	59.07554
	β_4 =80	80.58534	217.3549
	β_5 =55	53.57662	1104.867
$n = 50$	β_0 =20	19.98742	2.315203
	β_1 =40	40.04765	16.37619
	β_2 =30	30.01891	85.73758
	β_3 =50	50.1007	26.559
	β_4 =80	79.86278	98.69653
	β_5 =55	54.99462	490.1038
$n = 100$	β_0 =20	20.03455	1.058805
	β_1 =40	40.02255	7.35934
	β_2 =30	30.08981	41.44856
	β_3 =50	50.16944	13.20461
	β_4 =80	80.02442	46.61918
	β_5 =55	54.68129	233.1581
$n = 200$	β_0 =20	20.00226	0.531346
	β_1 =40	40.02303	3.377697
	β_2 =30	30.18862	18.5691
	β_3 =50	50.12513	6.283709
	β_4 =80	80.03548	21.53696
	β_5 =55	54.67519	108.1894
$n = 500$	β_0 =20	20.00056	0.204958
	β_1 =40	39.92333	1.37536
	β_2 =30	29.8807	7.152052
	β_3 =50	49.92647	2.434517
	β_4 =80	79.89837	8.592135
	β_5 =55	55.34286	42.07386
$n = 1000$	β_0 =20	20.00803	0.088906
	β_1 =40	39.99234	0.685176
	β_2 =30	30.00141	3.816272
	β_3 =50	49.99058	1.201453
	β_4 =80	79.9891	4.683265
	β_5 =55	55.03201	22.83449

DOI: 10.31695/IJASRE.2019.33217

The OLS method produced close estimates to the true value across the sample sizes. Taking a look at the variances the OLS method has large variance. Increasing the sample sizes improve the variance of the OLS estimates.

Table 3: Absolute bias of the estimators at various sample sizes

SAMPLE SIZES	ORDINARY LEAST SQUARES
$n = 25$	10.22923
$n = 50$	6.853374
$n = 100$	4.701329
$n = 200$	3.267761
$n = 500$	2.026656
$n = 1000$	1.478323

Table 3 present the Absolute bias of the estimator which is used to measure the consistency of the estimators, it was discovered that the OLS has large absolute bias.

Table 4: Mean square error of the estimators

SAMPLE SIZES	OLS MSE
$n = 25$	270.1117
$n = 50$	119.8501
$n = 100$	57.10774
$n = 200$	26.41471
$n = 500$	10.32074
$n = 1000$	5.546268

Table 4 presents the Mean Square Error MSE which is used to measure the efficiency of the estimator to the true values, it was observed that the OLS has large MSE. While at large sample size OLS estimator tends to produce lower MSE.

Table 5: Mean square error of prediction the estimators

SAMPLE SIZE	ORDINARY LEAST SQUARES MSEP
$n = 25$	75.7604
$n = 50$	88.4472
$n = 100$	94.25558
$n = 200$	96.79781
$n = 500$	98.70971
$n = 1000$	99.26722

Considering the predictive ability, the MSEP of OLS produced large MSEP.

6. CONCLUSION

Based on the results from the analysis of the simulated data, it was observed that the OLS perform very poor with the simulated data and thus not to be considered when posed with multicollinearity issue. However, it may be considered if the data sample size is considerably large as the collinearity issues fizzle out at large sample.

REFERENCES

1. Batterham AM, Tolfrey K and George KP. (1997): Nevill's explanation of Kleiber's 0.75 mass exponent: an artifact of collinearity problems in least squares models? *Journal of Applied Physiology.; 82: 693–697.*

2. Cavell, A. C, Lydiate, D. J. and Parkin, I. A. P. (1998): Collinearity between a 30- centimorgan segment of Arabidopsis thaliana chromosome 4 and duplicated regions within the Brassica napus genome. Genome.; 41: 62–69.

3. El-Dereny, M.and Rashwan, N. I.(2011): Solving Multicollinearity Problem using Ridge Regression Models, *Int. Journal. Contemp. Math. Sciences, Vol. 6, 2011, no. 12, 585 – 600.*

DOI: 10.31695/IJASRE.2019.33217

4. Freund, R., and Littell, R. C. (2000): SAS System for Regression. 3rd ed. SAS Inst., Inc. Cary, NC.

5. Parkin, I. A. P., Lydiate, D. J. and Trick, M. (2002): Assessing the level of collinearity between Arabidopsis thaliana and Brassica napus for A. thaliana chromosome 5. Genome. 2002; 45:356– 366.

DOI: 10.31695/IJASRE.2019.33217