Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 3 (2021) 278–281

Journal of the Nigerian Society of Physical Sciences

Ridge Estimation's Effectiveness for Multiple Linear Regression with Multicollinearity: An Investigation Using Monte-Carlo Simulations

O.G. Obadina[a], A. F. Adedotun[b,*], O. A. Odusanya[c]

[a]*Department of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye, Ogun State, Nigeria*
[b]*Department of Mathematics, Covenant University Ota, Ogun State, Nigeria*
[c]*Department of Mathematics, D.S Adegbenro (ICT) Polytechnic, Itori, Ogun State, Nigeria*

Abstract

The goal of this research is to compare multiple linear regression coefficient estimation technique with multicollinearity. In order to quantify the effectiveness of estimations by the mean of average mean square error, the ordinary least squares technique (OLS), modified ridge regression method (MRR), and generalized Liu-Kejian method (LKM) are compared with the Average Mean Square Error (AMSE). For this study, the simulation scenarios are 3 and 5 independent variables with zero mean normally distributed random error of variance 1, 5, and 10, three correlation coefficient levels; i.e., low (0.2), medium (0.5), and high (0.8) are determined for independent variables, and all combinations are performed with sample sizes 15, 55, and 95 by Monte Carlo simulation technique for 1,000 times in total. As the sample size rises, the AMSE decreased. The MRR and LKM both outperformed the OLS. At random error of variance 10, the MRR is the most suitable for all circumstances.

DOI:10.46481/jnsps.2021.304

Keywords: Monte-Carlo, Multicollinearity, Regression Model, Ridge Estimation, Simulations

Article History :
Received: 16 July 2021
Received in revised form: 30 September 2021
Accepted for publication: 05 October 2021
Published: 29 November 2021

Communicated by: T. Latunde

1. Introduction

Multiple linear regression (MLR), a widely used and well-known statistical technique, is now applied in a variety of fields [1,2]. This method is a statistical strategy that predicts the values of a response by combining many predictors (independent variables). MLR's purpose is to identify the optimal model for describing the linear relationship between predictor and response variables. After obtaining the best subsets of predictors, MLR's main objective is to estimate coefficients, find the most fitting estimates, and minimize errors. The least squared approach has long been used as a tool for estimating. It is a common and acceptable approach. However, this technique has a multicollinearity constraint, which is a major MLR roadblock.

Literatures on ridge regression are concerned with the problem of finding a better substitute to the least square estimator. Common methods in dealing with multicollinearity include but not limited to ridge regression [3]. Estimation procedures are obtained using some specific assumptions such as the random vector ε should be independent and identically distributed random variables. But when these assumptions are violated, these methods do not yield the desirable results and leads to problems

*Corresponding author tel. no: +2348055711272
Email address: adedayo.adedotun@covenantuniversity.edu.ng (A. F. Adedotun)

such as heteroscedasticity and autocorrelation [4]. Li, & Yang [5,7] suggested Jackknifed Modified Ridge Estimator (JMRE) and it was shown that it superior to other models.

Giacalone *et al*, [7] define multicollinearity as a condition were regressor variables in multiple linear regression model is almost linearly dependent. This condition causes the variance of least squares estimator tends to be large and the estimator becomes unstable. Hence, this condition will result in a reduced explanation of the result of the regression model and ridge regression is used to address these difficulties.

Ref. [7] introduce the L_{pmin} method to determine and address the multicollinearity problem. The major advantage of the approach is that it produces more efficient estimates of the model's parameters than the ordinary least squares method. Ref. [8] proposed a new collinearity diagnostic test. Monte Carlo simulation study conducted to compare the existing and proposed tests. It is based on coefficient of determination and adjusted coefficient of determination on auxiliary regression of regressors while Ref. [9] examined estimators which are resistant to the combined problems of multicollinearity and non-normal disturbance distributions. Can the ridge estimator and some robust estimation technique be combined to produce a robust ridge regression estimator?

An algorithm that uses the α-level estimation method to evaluate the parameters of the ridge fuzzy regression model was proposed by Ref. [10]. Parameter bias, Type I and Type II error, and variance inflation factor (VIF) values produced by multiple regressions with two, four, and six predictors under various multicollinearity circumstances were examined [11]. Multicollinearity is not linked to Type I error, but it does increase Type II error, according to the findings. Multicollinearity appears to enhance the variability in parameter bias while resulting in overall parameter underestimate. VIF is also increased by collinearity. Increasing the number of predictors, on the other hand, interacts with multicollinearity in all diagnostics to exacerbate difficulties.

An extended conventional semi-parametric partial linear regression model was introduced [12]. The effectiveness of the proposed method is then illustrated through two numerical examples including a simulation study. They also compared with some common fuzzy multiple regression models with fuzzy predictors and fuzzy responses. In the method of ridge regression, a constant bias ridge k was added to $X'X$ matrix. This work illustrates the use of Restricted Ridge Regression method in disabling the multicollinearity in regression model. The method was developed by using the prior information of the parameter β [13].

2. Materials and Methods

Ridge regression is suited to deal with the problem of multicollinearity, especially when the predictors are highly correlated. In 1970, [14] proposed the ridge regression estimator, which included a scalar multiplication, the product of a positive real number and the identity matrix, within the inverse component of the least square estimator. This yielded more accurate ridge parameter estimations than least square estimates, and its variance and mean square errors are frequently lower than least square estimates. The following is a comparison of the three approaches for estimating multiple linear regression coefficients: Least Squares Method (OLS), Modified Ridge Regression Method (MRR), and Liu Kejian Method (LKM):

2.1. Ordinary Least Square Method (OLS)

The best linear unbiased estimator is a method for estimating multiple linear regression coefficients that is unbiased and has the least variation of estimation (BLUE). The estimated value is expressed as

$$\widehat{\beta}_{OLS} = (X'X)^{-1}X'Y \tag{1}$$

where X is the $n \times p$ predictors matrix, Y is the $n \times 1$ observation vector, $\widehat{\beta}_{OLS}$ is the vector of coefficients estimates. The mean square error of $\widehat{\beta}_{OLS}$ is $\sigma^2 tr\,(X'X)^{-1}$.

2.2. Modified Ridge Regression (MRR

The ridge estimation for multiple linear regression coefficients is

$$\widehat{\beta}_{Ridge} = (\boldsymbol{X}'\boldsymbol{X} + kI_n)^{-1}\boldsymbol{X}'\boldsymbol{Y}, \qquad k > 0, \tag{2}$$

where $\widehat{\beta}_{Ridge}$ is $p \times 1$ ridge estimator, k is a positive real number also known as a constant bias ridge, I_n is an identity matrix of size n. The approximation of the linear regression coefficients is more accurate and closer to the real values when historical data is used with the ridge regression approach. The modified ridge regression (MRR) approach is as follows:

$$\widehat{\beta}_{MRR} = \left(\boldsymbol{X}'\boldsymbol{X} + kI_n\right)^{-1}(\boldsymbol{X}'\boldsymbol{Y} + k\boldsymbol{J}) \tag{3}$$

where $\boldsymbol{J}$ is $p \times 1$ historical observation vector, $J = (\sum_{i=1}^{p} \frac{\widehat{\beta}_{OLS}}{p})\mathbf{1}$, $\mathbf{1}$ is $p \times 1$ vector of ones where every element is equal to one. From equation (3), $\widehat{\beta}_{MRR} = \widehat{\beta}_{OLS}$ when $k = 0$

The estimation of k is considered using the following two cases:

1. σ^2 is known,

$$\hat{k} = \begin{cases} \frac{p\sigma^2}{\left(\widehat{\beta}_{OLS}-J\right)'\left(\widehat{\beta}_{OLS}-J\right)-\sigma^2 tr(X'X)^{-1}}, \\ if\ \left(\widehat{\beta}_{OLS}-J\right)'\left(\widehat{\beta}_{OLS}-J\right)-\sigma^2 tr(X'X)^{-1} > 0 \\ \frac{p\sigma^2}{\left(\widehat{\beta}_{OLS}-J\right)'\left(\widehat{\beta}_{OLS}-J\right)},\ otherwise \end{cases}$$

2. σ^2 is unknown,

$$\hat{k} = \begin{cases} \frac{p\widehat{\sigma}^2}{\left(\widehat{\beta}_{OLS}-J\right)'\left(\widehat{\beta}_{OLS}-J\right)-\widehat{\sigma}^2 tr(X'X)^{-1}}, \\ if\ \left(\widehat{\beta}_{OLS}-J\right)'\left(\widehat{\beta}_{OLS}-J\right)-\widehat{\sigma}^2 tr(X'X)^{-1} > 0 \\ \frac{p\widehat{\sigma}^2}{\left(\widehat{\beta}_{OLS}-J\right)'\left(\widehat{\beta}_{OLS}-J\right)},\ otherwise \end{cases}$$

where $\widehat{\sigma}^2 = \frac{\left(Y-X\widehat{\beta}_{OLS}\right)'\left(Y-X\widehat{\beta}_{OLS}\right)}{n-p}$ is an unbiased estimator of σ^2.

The mean square of $\widehat{\beta}_{MRR}$ is $\widehat{\sigma}^2 tr\left(\left(\boldsymbol{X}'\boldsymbol{X} + \hat{k}I_n\right)^{-1}(X'X)\left(\boldsymbol{X}'\boldsymbol{X} + \hat{k}I_n\right)^{-1}\right) + \hat{k}^2\left(\widehat{\beta}_{OLS} - J\right)\left(\boldsymbol{X}'\boldsymbol{X} + \hat{k}I_n\right)^{-2}\left(\widehat{\beta}_{OLS} - J\right)$

2.3. Generalized Liu Kejian Method (LKM)

In the situation of a multiple-relationship between the independent variables, this is a method for estimating the multiple linear regression coefficient. The advantages of the Ridge Regression approach and the Stein method are combined. The Generalized Kejian Method is the name of this method, and the form of the multiple linear regression coefficient estimator is

$$\widehat{\beta}_{LKM} = (X'X + I_n)^{-1}\left(X'Y + d\widehat{\beta}_{OLS}\right),\ 0 < d < 1 \tag{4}$$

when $d = 1, \widehat{\beta}_{LKM} = \widehat{\beta}_{OLS}$ and

$$\begin{aligned}\widehat{\beta}_{LKM} &= (X'X + I_n)^{-1}\left(X'Y + D\widehat{\beta}_{OLS}\right)\\ &= (X'X + I_n)^{-1}\left(X'Y + D)\widehat{\beta}_{OLS}\right)\\ &= \left(I_n - (X'X + I_n)^{-2}(I_n - D)\right)\widehat{\beta}_{OLS}\\ &= \left(I_n - (X'X + I_n)^{-2}(I_n - D)^2\right)\widehat{\beta}_{OLS}\end{aligned} \tag{5}$$

where $D = diag\,(d_1, d_2, \ldots, d_p), 0 < d_i < 1, i = 1, 2, \ldots, p$ and the estimates of d_i is

$$\hat{d}_i = 1 - \frac{\widehat{\sigma}\,(\lambda_i + 1)}{\sqrt{\lambda_i \widehat{\beta}^2_{OLS_i} + \widehat{\sigma}^2}}$$

The mean square error of $\widehat{\beta}_{LKM}$ is $(I_n - \Delta^2)\,(X'X)^{-1}(I_n - \Delta^2)\sigma^2 + \Delta^2\beta\beta'\Delta^2$, where $\Delta = (X'X + I_n)^{-1}(I_n - D)$

2.4. Monte Carlo Simulation

A Monte Carlo simulation scenario with three and five independent variables was developed by [8], the properties were zero mean normally distributed random error of variance 1, 5, and 10, and three correlation coefficients levels; i.e., low (0.2), medium (0.5), and high (0.8) are determined for independent variables, and all combinations are performed with sample sizes 15, 55, and 95 by Monte Carlo simulation. The steps for carrying out the simulation are:

1. The random error (ε) is simulated as $\varepsilon \sim N\,(0, \sigma^2_\varepsilon I_n)$, where $\sigma^2_\varepsilon = 1, 5, 10$.
2. An observation matrix, X, is simulated from $X \sim N_n(0, I_N)$ with different levels of polynomial relations such that $\rho = 0.2,\ 0.5,\ 0.8$.
3. Generate response values of Y from the model with multiple linear regression coefficient β.
4. Multiple linear regression coefficients are estimated for all methods. The step in 1-3 are repeated 1,000 times in each scenario.
5. Then calculate the mean of the mean square error of multiple linear regression, $AMSE = \frac{1}{1000}\sum_{i=1}^{1000} MSE$), the method with lowest AMSE is selected as the best method for the scenario involved.

Table 1. The best method of all scenarios from 1000-Monte Carlo simulations

Predictors	σ	ρ	n=15	n=55	n=90
3	1	0.3	MRR	MRR	MRR
3	1	0.6	MRR	MRR	MRR
3	1	0.9	MRR	MRR	MRR
3	5	0.3	MRR	MRR	MRR
3	5	0.6	MRR	MRR	MRR
3	5	0.9	MRR	MRR	MRR
3	10	0.3	LKM	MRR	MRR
3	10	0.6	LKM	LKM	MRR
3	10	0.9	LKM	MRR	MRR
5	1	0.3	MRR	MRR	MRR
5	1	0.6	MRR	MRR	MRR
5	1	0.9	MRR	MRR	MRR
5	5	0.3	LKM	MRR	MRR
5	5	0.6	MRR	MRR	MRR
5	5	0.9	MRR	MRR	MRR
5	10	0.3	LKM	MRR	MRR
5	10	0.6	LKM	LKM	MRR
5	10	0.9	LKM	LKM	LKM

3. Conclusion

The optimum multicollinearity MLR coefficients estimation approach for each simulation circumstance is shown in Table 1. Clearly, the OLS is not suited in every application. MRR and LKM are the best approaches for determining MLR coefficients when multicollinearity exists. MRR is appropriate for all sample sizes and data with low predictor correlation degrees and small to moderate error variance. The Generalized Liu Kejian Method is well suited to small datasets with a high degree of predictor correlation and a large error variance. LKM outperforms MRR as the number of predictors grows, however the more predictors, the greater the risk of multicollinearity.

References

[1] K. K. Adesanya, A. I. Taiwo, A. F. Adedotun & T. O. Olatayo "Modeling Continuous Non-Linear Data with Lagged Fractional Polynomial Regression", Asian Journal of Applied Sciences **06** (2018) 315.

[2] G. Ciulla, & A. D'Amico "Building energy performance forecasting: A multiple linear regression approach", Applied Energy **253** (2019) 113500.

[3] H. Yang, & X. Chang "A new two-parameter estimator in linear regression", Communications in Statistics - Theory and Methods **39** (2010) 923 doi: 10.1080/0361092090280791.

[4] B. M. G. Kibria "Performance of some new ridge regression estimators", Communications in Statistics - Simulation and Computation, **32** (2003) 419. doi: 10.1081/SAC-120017499.

[5] Y. Li, & H. Yang "On the performance of the jackknifed modified ridge estimator in the linear regression model with correlated or heteroscedastic errors", Communications in Statistics - Theory and Methods **40** (2011) 2695. doi: 10.1080/03610926.2010.491589

[6] D. C. Montgomery, E. A. Peck E & G. G. Vining "Introduction to Linear Regression Analysis", (United States: John Wiley & Sons) 2001.

[7] M. Giacalone, D. Panarello, & R. Mattera "Multicollinearity in regression: an efficiency comparison between L p-norm and least squares estimators", Quality & Quantity **52** (2018) 1831.

[8] M. I Ullah, M. Aslam, S. Altaf, & M. Ahmed "Some new diagnostics of multicollinearity in linear regression model", Sains Malaysiana **48** (2019) 2051.

[9] A. F. Lukman, K. Ayinde, & A. S. Ajiboye "Monte Carlo study of some classification-based ridge parameter estimators", Journal of Modern Applied Statistical Methods **16** (2017) 24.

[10] S. H. Choi, H. Y. Jung, & H. Kim, "Ridge fuzzy regression model", International Journal of Fuzzy Systems **21** (2019) 2077.

[11] M. R. Lavery, P. Acharya, S. A. Sivo, & L. Xu, "Number of predictors and multicollinearity: What are their effects on error and bias in regression?", Communications in Statistics-Simulation and Computation **48** (2019) 27.

[12] M. G. Akbari, & G. Hesamian "A partial-robust-ridge-based regression model with fuzzy predictors-responses", Journal of Computational and Applied Mathematics **351** (2019) 290.

[13] F. A. O Rumere, S. M. Soemartojo & Y. Widyaningsih, "Restricted Ridge Regression estimator as a parameter estimation in multiple linear regression model for multicollinearity case", Journal of Physics: Conference Series **24** (2021) 1725.

[14] A. E. Hoerl & R. W. Kennard "Ridge Regression: Biased Estimation for Nonorthogonal Problems", American Society for Quality **12** (1970) 44.