African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 5, Issue 2, 2022 (pp. 61-71)
www.abjournals.org

COMPARISON OF THE EFFECTS OF RIDGE BIASING CONSTANT IN REMEDYING MULTICOLLINEARITY ON GAMMA AND EXPONENTIALLY DISTRIBUTED DATA

Nwankwo Chike H. and Nnaji Peace O.

Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria

Corresponding author: peaceo_online@yahoo.com

Cite this article:

Nwankwo C.H., Nnaji P.O. (2022), Comparison of the Effects of Ridge Biasing Constant in Remedying Multicollinearity on Gamma and Exponentially Distributed Data. African Journal of Mathematics and Statistics Studies 5(2), 61-71. DOI: 10.52589/AJMSS-EY0EE7OV.

Manuscript History

Received: 24 May 2022

Accepted: 19 May 2022

Published: 30 June 2022

ABSTRACT: *Ridge regression as a solution to multicollinearity depends on the value of k, the ridge biasing constant. Since no optimum value can be found for k, as k is generally bounded between 0 and 1, i.e. $0\leq k \leq 1$ and it varies from one application to another. This has posed a major limitation of ridge regression in that ordinary inference procedures are not applicable and exact distributional properties are not known; and the choice of the biasing constant, k, is a judgmental one. This work examined the effect of ridge biasing constant, k, on different sample sizes using data combination from gamma and exponential distributions when multicollinearity exists. The sample sizes of 140, 100, 80, 50, 30, 20 and 10 and ridge constants, k=0.01, 0.02, ..., 0.1 respectively were used in the study. The Anderson Darling Test was used to check for the distribution of the data which were found to follow gamma and exponential distributions. The findings lay credence to how the ridge regression drastically remedies the effect of multicollinearity among independent variables. The study also revealed that the VIF consistently decreased as the ridge constant increased. While the ridge regression has a slight effect on the R-squared, sample sizes were found not to have any significant change or pattern on the VIFs. Since the study has shown that the VIF reduced drastically as the ridge constant increases, it is recommended to use a VIF that reduces multicollinearity to an acceptable minimum while maximizing the R-squared. This study recommends a ridge constant of 0.1 as all multicollinearity issues have been remedied at more than 90% if not completely. The study recommends using a large sample size to help stabilize the R^2 values while remedying multicollinearity.*

KEYWORDS: Ridge Biasing Constant, Remedying Multicollinearity, Gamma and Exponentially Distribution

Article DOI: 10.52589/AJMSS-EY0EE7OV
DOI URL: https://doi.org/10.52589/AJMSS-EY0EE7OV

INTRODUCTION

Multiple regression analysis is one of the most widely used statistical procedures with its applicability to varied types of data and problems, ease of interpretation and robustness to violations of the underlying assumptions. Predictive accuracy is calibrated by the magnitude of the R^2 and the statistical significance of the overall model (Mason &Perreault, 1991).

In the application of multiple regression, problems may arise when two or more predictor variables are correlated. Overall prediction is not affected, but the interpretation of conclusions based on the size of the regression coefficients, their standard errors, or the associated t-tests may be misleading because of the potentially confounding effects of collinearity among the independent variables. This situation is known as multicollinearity (Mason & Perreault, 1991).

Multicollinearity is a situation in which two or more predictor variables in a multiple regression model are highly correlated. Here, the estimates of the coefficient in the multiple regression any change erratically in response to small changes in the model or data. A multiple regression model with correlated predictors can indicate how well the entire bundle of predictors predict the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others.

Though no precise definition of collinearity has been firmly established in the literature, collinearity is generally agreed to be present if there is a significant linear relationship among some of the predictor variables in the Data (Mason and Perreault, 1991).

Multicollinearity has several potentially undesirable consequences: parameter estimates that fluctuate dramatically with negligible changes in the sample, parameter estimates with signs that are wrong in terms of theoretical consideration; theoretically, important variables with insignificant coefficients, and the inability to determine the relative importance of collinear variables. All of these consequences are symptoms of the same fundamental problem: near collinearities inflate the variance of the regression coefficients (Stewart 1987).

Pieces of literature provide numerous suggestions, ranging from simple rules of thumb to complex indices for diagnosing the presence of collinearity. Several of the most widely used procedures are examining the correlation matrix of the predictor variables, computing the coefficient of determination, R_k^2, of each X_k regressed on the remaining predictor variables, and measures based on the eigenstructure of the data matrix X, including variance inflation factor (VIF), trace of $(X^I X)^{-1}$, the condition number, ridge regression, etc. A common rule of thumb suggests that collinearity is a problem if any of the R_k^2 exceeds the R^2 of the overall model (Mason and Perreault, 1991).

The most common estimator for β is the ordinary least squares estimator, $\hat{\beta} = (X'X)^{-1}X'Y$, which is an unbiased estimator. But in the presence of multicollinearity ordinary least squares, estimators could become very unstable due to their very large variance, which leads to poor prediction. One of the popular methods for handling this problem is ridge regression estimation. (Dorugade and Kashid, 2010).

Ridge regression is the modification of the least-squares method that allows biased estimators of the regression coefficient. These biased estimators are preferred over the least-squares estimator because they have a larger probability of being close to the true parameter values

Article DOI: 10.52589/AJMSS-EY0EE7OV
DOI URL: https://doi.org/10.52589/AJMSS-EY0EE7OV

with smaller mean squared error (MSE) of regression coefficients, (Dorugade and Kashid, 2010).

Ridge regression was developed by Hoerl and Kennard in 1970. When multicollinearity exists, the matrix X′X, where X consists of the original regressors becomes nearly singular. Since the variance of $\hat{\beta}= \sigma^2(X'X)^{-1}$ and the diagonal elements of $(X'X)^{-1}$ become quite large, so the variance of $\hat{\beta}$ is to be large. This leads to an unstable estimate of β when the ordinary least squares is used (Al-Hassan, 2010).

In ridge regression, a standardized matrix X is used and a small constant, k, which is known as the ridge biasing constant is added to the diagonal elements of X′X. The addition of the small positive number, k where $k \geq 0$ to the diagonal elements of X′X causes X′X to be non-singular. Therefore, the ridge estimator is given as $\hat{\beta}_R = (X'X + kI)^{-1}X'Y$. (Dorugade and Kashid, 2010)

It is obvious that when k = 0, OLS estimators are recovered. As k increases, the ridge regression estimators are biased but more precise than OLS estimators, hence they will be closer to the true parameters.

The ridge regression estimator does not provide a unique solution to the problem of multicollinearity but provides a family of solutions. These solutions depend on the value of k, the ridge biasing constant. No optimum value can be found for k since k is generally bounded below by 0 and above by 1, i.e. $0\leq k \leq 1$ and it varies from one application to another.

Neter, Wasserman and Kutner (1983) posit that a major limitation of ridge regression is that ordinary inference procedures are not applicable and exact distributional properties are not known; and the choice of the biasing constant, k, is a judgmental one. While formal methods have been developed for making this choice, these methods have their own limitations

It was shown by Hoerl and Kennard (1970) that if a small enough k value for which the mean squared error is less than the mean squared error of the ordinary least squares is chosen, the procedure of the ridge regression is successful and $\hat{\beta}_R$ becomes stable. The main interest lies in finding a value of the ridge biasing constant, k, such that the reduction in the variance is attained with an accompanying increase in the stability of the regression coefficients. This study seeks to determine the effect of the ridge biasing constant, k for remedying multicollinearity when analyzing multiple regression data combinations from specifically Gamma and Exponential distributions.

REVIEW OF RELATED LITERATURE

The history of multicollinearity dates to the paper by Frisch in 1934 who introduced the concept to describe a situation where the variables dealt with are subject to two or more relations. Hoerl and Kennard, in 1970, introduced ridge regression to handle the problem of multicollinearity. At this stage, the main interest lies in finding a value of the ridge parameter which is the ridge biasing constant, k, such that the reduction in the variance term of the slope parameter is greater than the increase in its squared bias. The authors proved that there is a nonzero value of such ridge biasing constant for which the mean squared error (MSE) for the slope parameter using ridge regression is smaller than the variance of the ordinary least squares (OLS) estimator of the respective parameter (Muniz et al, 2012)

Article DOI: 10.52589/AJMSS-EY0EE7OV
DOI URL: https://doi.org/10.52589/AJMSS-EY0EE7OV

Vedide et al (2014), stated that a multiple regression model has got the standard assumptions and if the data cannot satisfy these assumptions some problems which have some serious undesired effects on the parameter estimates arise. One of the problems is called multicollinearity which means that there is a nearly perfect linear relationship between explanatory variables used in a multiple regression model. This undesirable problem is generally solved by using methods such as ridge regression which gives biased parameter estimates. Ridge regression shrinks the Ordinary Least Squares estimation vector of regression coefficients towards the origin, allowing a bias but providing a smaller variance. They observed, however, that the choice of the biasing constant k in ridge regression is another serious issue.

Yahya and Olaifa (2014), investigated the techniques of the ridge regression model as an alternative to the classical ordinary least square (OLS) method in the presence of correlated predictors. They observed that one of the basic steps for fitting efficient ridge regression models requires that the predictor variables be scaled to unit lengths or to have zero means and unit standard deviations prior to parameters' estimations. This was meant to achieve stable and efficient estimates of the parameters in the presence of multicollinearity in the data. Their work, therefore, examined the impacts of scaled collinear predictor variables on ridge regression estimators. Various results from simulation studies underscored the practical importance of scaling the predictor variables while fitting ridge regression models. They employed real-life data set on import activities in the French economy to validate the results from the simulation studies.

Nagai (2014) said that there are several model selection criteria for selecting the ridge parameter in multivariate ridge regression, e.g., the Cp criterion and the modified Cp (MCp) criterion and proposed the generalized Cp (GCp) criterion, which includes Cp and MCp criteria as special cases. The GCp criterion is specified by a non-negative parameter λ, which is referred to as the penalty parameter. He attempted to select an optimal penalty parameter such that the Predictive Mean Square Error (PMSE) of the predictor of ridge regression after optimizing the ridge parameter is minimized. Through numerical experiments, he verified that the proposed optimization methods exhibit better performance than conventional optimization methods, i.e., optimizing only the ridge parameter by minimizing the Cp or MCp criterion.

Fitrianto et al (2014) conducted some simulation studies to compare the performance of the ridge regression estimator and the OLS. Simulation studies of several methods for estimating the ridge parameters. The performance of each ridge estimator depends on the standard deviation (s) and the correlations between explanatory variables (g2). For s = 0.1, the HK estimator and HSL estimator have smaller MSE than the OLS estimator for all levels of correlations. However, the OLS estimator is reasonably better than the NHSL estimator for all levels of correlations for this given value of HK estimator might be recommended to be used to estimate the ridge parameter k. The study recommended further investigation of ridge estimators in future in order to make any definite statement.

Hanan Duzan et al (2015) investigated the problem of using Ordinarily Least Squares (OLS) estimators in the presence of multicollinearity in regression analysis. An alternative to OLS is ridge regression, which is believed to be superior to least-squares regression in the presence of multicollinearity. The robustness of this method was investigated and a comparison was made with the least-squares method via simulation studies. Their results have shown that the system stabilizes in a region of k, in which k is a positive quantity less than one and whose values

Article DOI: 10.52589/AJMSS-EY0EE7OV
DOI URL: https://doi.org/10.52589/AJMSS-EY0EE7OV

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 5, Issue 2, 2022 (pp. 61-71)
www.abjournals.org

depend on the degree of correlation between the independent variables. The results illustrate that *k* is a non-linear function of the correlation between the independent variables (r_{12}).

Pereira et al (2015) used lasso and ridge approaches for predicting corporate bankruptcy since they deal well with multicollinearity and display the ideal properties to minimize the numerical instability that may occur due to overfitting. The models were employed in a dataset of 2032 non-bankrupt firms and 401 bankrupt firms belonging to the hospitality industry, over the period 2010-2012. The results of the study showed that the lasso and ridge models tend to favour the category of the dependent variable that appears with heavier weight in the training set when compared to the stepwise methods implemented in SPSS.

Efendi & Effrihan (2017) discussed the simulation process for evaluating the characteristic of Bayesian Ridge regression parameter estimates using several simulation settings based on a variety of collinearity levels and sample sizes. The results of the study show that the Bayesian method gives better performance for relatively small sample sizes, and for other settings, the method does perform relatively similarly to the likelihood method.

Daoud (2017) posits that when multicollinearity is present, the standard error of the coefficients will increase. An increased standard error means that the coefficients for some or all independent variables may be found to be significantly different from 0. In other words, by overinflating the standard errors, multicollinearity makes some variables statistically insignificant when they should be significant.

Herawati et al (2018) compared the performance of Ordinary Least Square (OLS), Least Absolute Shrinkage and Selection Operator (LASSO), Ridge Regression (RR) and Principal Component Regression (PCR) methods in handling severe multicollinearity among explanatory variables in multiple regression analysis using data simulation. A Monte Carlo experiment was carried out to select the best method, it was found that the simulated data contain severe multicollinearity among all explanatory variables ($\rho = 0.99$) with different sample sizes ($n = 25, 50, 75, 100, 200$) and different levels of explanatory variables ($p = 4, 6, 8, 10, 20$). The performances of the four methods are compared using Average Mean Square Errors (AMSE) and Akaike Information Criterion (AIC). The result shows that PCR has the lowest AMSE among other methods. It indicates that PCR is the most accurate regression coefficients estimator in each sample size and various levels of explanatory variables studied. PCR also performs as the best estimation model since it gives the lowest AIC values compare to OLS, RR, and LASSO.

Schreiber-Gregory (2018) reviewed and provided examples of the different ways in which multicollinearity can affect a research project, how to detect multicollinearity and how one can reduce it through Ridge Regression applications.

The data used for this study are secondary data retrieved from the Central bank of Nigeria and the Nigerian Bureau of Statistics databases. The data were macroeconomic data of Nigeria from the first quarter (Q_1)1986 to the fourth quarter (Q_4) 2020. The data include Government expenditure (GXP), Real Gross Domestic Product (RGDP), Nominal Gross Domestic Product (NGDP), Export, Import and international trade (Trade). Data were tested and confirmed to follow Gamma and exponential distributions which are being used for the purpose of this study.

Testing for Multicollinearity

Article DOI: 10.52589/AJMSS-EY0EE7OV
DOI URL: https://doi.org/10.52589/AJMSS-EY0EE7OV

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 5, Issue 2, 2022 (pp. 61-71)
www.abjournals.org

The presence of multicollinearity is tested using the Variance Inflation Factor. WonSuk et al (2010) defined the variance inflation factor as a measure of how much the variance of the estimated regression coefficient b_i is "inflated" by the existence of correlation among the predictor variables in the model. According to the author, a VIF of 1 means that there is no correlation between the *i*th predictor and the remaining predictor variables, and hence the variance of b_i is not inflated at all. The general rule of thumb is that VIFs exceeding 10 are signs of serious multicollinearity requiring correction. The test for multicollinearity is necessary to ascertain the independence of the predictor variables.

The variance inflation factor for the *i*th suspected predictor variable is given by:

$$VIF_i = \frac{1}{1-R_i^2}$$

Where R_i^2 is the R^2 (coefficient of determination) value obtained by regressing the ith predictor on the remaining predictors**.**

Multiple Regression (Ordinary Least Squares Approach)

Multiple Linear Regression Model is given as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \qquad . \qquad . \qquad . \qquad (1)$$

where:

β_0 = a constant term (the Y-intercept)

β_i = is the rate of change of Y with respect to X_i.

ε = error term.

Where, Y and X are given as:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{pmatrix} \qquad X = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots X_{1k} \\ 1 & X_{21} & X_{22} & \cdots X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots X_{nk} \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$ is the vector of regression parameters,

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$ is the vector of random errors.

Article DOI: 10.52589/AJMSS-EY0EE7OV
DOI URL: https://doi.org/10.52589/AJMSS-EY0EE7OV

Y is an n × 1 vector of observations on the dependent variable, X is an n × p matrix of

observations on independent variables, $\beta$is a p × 1 vector of unknown parameters which are called regression coefficients and $\varepsilon$is a × 1 vector of random errors which are assumed as ε_i =N~ (0, σ^2)

Thus, we can matrix-multiply X to β. The product $X\beta$ has dimension $n \times 1$, that is, it is an n-dimensional column vector.

The Least Squares estimate of the multiple regression coefficient β is given as:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \quad . \quad . \quad . \quad (2)$$

Mean Square and Standard Error

MSE = $s^2 = \frac{SSE}{n-k}$

Analysis of Variance Table

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F- RATIO
REG	SSR=$\sum \quad (X_i - \underline{X})^2$	K-1	MSR=$\frac{SSR}{K-1}$	F_{cal}=MSR/MSE
ERROR	SSE=$\sum \quad (Y_i - \hat{Y})^2$	n-k	MSE=$\frac{SSE}{n-k}$	
TOTAL	SST	n-1		

There are two methods of concluding the ANOVA test, both of which produce the same result:

Firstly, the use of probability value (p-value). The null hypothesis is rejected if this probability value is less than or equal to the significance level (α).

Secondly, the observed value of F with the critical value of F may be compared using the F distribution tables. The critical value of F is a function of the degrees of freedom of the numerator and the denominator and the significance level (α). If $F \geq F_{critical,}$ the null hypothesis is rejected.

Coefficient of Multiple Determination (R^2)

The coefficient of multiple determination, R^2 is the percentage of the variance in the dependent variable explained uniquely or jointly by the independent variables. R^2 can also be interpreted as the proportionate reduction in error in estimating the dependent variable when knowing the independent variables. It is the proportion of change in the dependent variable that is attributable to the independent variables. The formula is given as:

Article DOI: 10.52589/AJMSS-EY0EE7OV
DOI URL: https://doi.org/10.52589/AJMSS-EY0EE7OV

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 5, Issue 2, 2022 (pp. 61-71)

$\mathbf{R^2} = \frac{SSR}{SST}$

SSR is the Sum of Squares Regression while SST is the Sum of Squares Total which is alternatively calculated from the ANOVA table as: SSR + SSE

But SSR = $\beta(X'Y) - (\frac{1}{n})Y'JY$

SST= $Y'Y - (\frac{1}{n})Y'JY$

Where J is an n x n square matrix with all elements 1.

Ridge Regression

The ridge regression estimator is obtained by solving the normal equations of least squares estimation. The normal equations are modified as:

$$(X'X + KI)\hat{\beta}_{ridge} = X'Y$$

$$\Rightarrow \hat{\beta}_{ridge} = (X'X + KI)^{-1}X'Y \quad . \quad . \quad . \quad (3)$$

As $K \to 0, \hat{\beta} \to \beta(OLS)$ and as $K \to 1, \hat{\beta} \to 0$

The OLSE is inappropriate to use in the sense that it has very high variance when multicollinearity is present in the data. On the other hand, a very small value of $\widehat{\beta}$ may tend to accept the null hypothesis H_0: β= 0 indicating that the corresponding variables are not relevant. The value of biasing parameter controls the amount of shrinkage in the estimates.

METHODOLOGY

The analysis started by testing the distribution of the macroeconomic data comprising Government expenditure (GXP), Real Gross Domestic Product (RGDP), Nominal Gross Domestic Product (NGDP), Export, Import and international trade (Trade) from the first quarter (Q_1) 1986 to the fourth quarter (Q_4) 2020 data using the Anderson Darling Test, and it was confirmed that the data followed gamma and exponential distributions.

The Multiple Linear Regression equation obtained from the data using the Ordinary Least Square (OLS) approach and the Variance Inflation Factor (VIF) resulting from the OLS estimates were calculated using the computer software SPSS.

The test of multicollinearity using the Variance Inflation Factor (VIF) obtained are as follows: NGDP=35.0, RGDP=40.7, Exp=6.7, Import=26.0, and trade=20.0. These indicate the presence of significant multicollinearity among the independent variables. The R^2value is 92.2%. The ridge regression is then fitted and the VIF recalculated. The interest is on the effect of ridge regression and sample size on the VIF. The sample size is then varied as 100, 80, 50, 30, 20 and 10 respectively for various values of the ridge constants of k=0.01, 0.02, … 0.1 respectively. Also, the ridge parameter (k) that gave the best ridge model was considered.

Article DOI: 10.52589/AJMSS-EY0EE7OV
DOI URL: https://doi.org/10.52589/AJMSS-EY0EE7OV

Furthermore, the effect of ridge constant on various sample sizes and VIF was considered. Won Suk et al (2010) posits that a VIF of 1 means that there is no correlation between the *ith* predictor and the remaining predictor variables, hence the variance of b_i is not inflated at all. The general rule of thumb is that VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

R^2 and MSE are acceptable values for the model. Predictive accuracy is calibrated by the magnitude of the R^2 (Mason &Perreault, 1991).

The Least Squares solution maximizes R-squared, the largest value of R-squared occurs when *k* is zero, which is the OLS estimate (NCSS Statistical Software).

Table 1 - R^2 statistics (Actual R-squared)

K	n=140	n=100	n=80	n=50	n=30	n=20	n=10
OLS	0.9217	0.8802	0.8408	0.9540	0.9695	0.9861	0.9712
0.01	0.9191	0.874536	0.8339	0.9514	0.9687	0.9851	0.9697
0.02	0.9161	0.8688	0.8274	0.9481	0.9676	0.9837	0.9680
0.03	0.9132	0.8641	0.8221	0.9450	0.9663	0.9822	0.9665
0.04	0.9105	0.8598	0.8174	0.9422	0.9651	0.9808	0.9654
0.05	0.9078	0.8558	0.8129	0.9397	0.9640	0.9795	0.9644
0.06	0.9051	0.8520	0.8085	0.9373	0.9629	0.9784	0.9636
0.07	0.9025	0.8482	0.8041	0.9351	0.9619	0.9772	0.9629
0.08	0.8998	0.8444	0.7998	0.9330	0.9609	0.9762	0.9623
0.09	0.8972	0.8407	0.7955	0.9310	0.9600	0.9752	0.9617
0.1	0.8945	0.837011	0.7911	0.9291	0.9592	0.9743	0.9612

From the table, the R^2 for OLS is the highest, but gradually decreased as k increased. This indicates the ridge regression performed better for the data than the OLS estimate.

Ridge Constant and Model Statistics

The ridge regression effect on the R^2 was assessed using the different sample sizes. In the case of n = 140 using k = 0.01, 0.02, 0.03, …0.1. The result revealed that the ridge regression caused small changes in the R^2 values. As the ridge constant k, increases (from 0.01 to 0.1), the R^2 value for all sample sizes decreases. At k=0.01 the R^2 decreased by 0.28%. The R^2 also decreased by 0.61, 0.92 and 1.21% for k=0.02, 0.03 and 0.04 respectively. Overall, the R^2 decreased by a minimum of 0.08 at k=0.01 and a maximum of 5.90% at k=0.1 and n=100. This shows that as k increases, the R^2 decreases. The changes in the sample size did not have any noticeable change in the values of the R^2.

Table 2: Percentage change in R^2 for the selected Ridge Constants and Sample sizes

K	n=140	n=100	n=80	n=50	n=30	n=20	n=10
0.01	-0.28	-0.62	-0.82	-0.27	-0.08	-0.10	-0.15
0.02	-0.61	-1.29	-1.59	-0.62	-0.20	-0.25	-0.33
0.03	-0.92	-1.83	-2.22	-0.94	-0.33	-0.40	-0.48

Article DOI: 10.52589/AJMSS-EY0EE7OV
DOI URL: https://doi.org/10.52589/AJMSS-EY0EE7OV

0.04	-1.21	-2.31	-2.78	-1.23	-0.45	-0.54	-0.60
0.05	-1.51	-2.77	-3.32	-1.50	-0.57	-0.67	-0.69
0.06	-1.79	-3.20	-3.84	-1.75	-0.69	-0.79	-0.78
0.07	-2.08	-3.63	-4.35	-1.98	-0.79	-0.90	-0.85
0.08	-2.37	-4.06	-4.87	-2.20	-0.89	-1.01	-0.92
0.09	-2.66	-4.48	-5.39	-2.41	-0.98	-1.11	-0.98
0.1	-2.95	-4.89	-5.90	-2.61	-1.06	-1.20	-1.03

The result revealed that the ridge regression caused a small change in the R^2 values. As the ridge constant k increases from 0.01 to 0.1, the R^2 value for all sample sizes decreased. At k=0.01 the R^2 decreased by 0.28%. The R^2 also decreased by 0.61, 0.92 and 1.21% for k=0.02, 0.03 and 0.04 respectively. Overall, the R^2 decreased by a minimum of 0.08 at k=0.01 and a maximum of 5.90% at k=0.1 and n=100.

This shows that as k increases, the R^2 decreases. The changes in the sample size did not have any noticeable change in the actual values of the R^2.

CONCLUSION

This study presented the effect of the ridge biasing constant on multicollinear data drawn from gamma and exponential distributions.

The following are the findings of this study:

1. Overall, the ridge regression was found to have caused a slight change in the R^2 values of the models. It was found that as*k* increases, the R^2 slightly decreased by a minimum of 0.08% and a maximum of 5.9%.
2. Sample size was found not to have any noticeable change or pattern across all studied variables, ridge constant and R^2 values.
3. It was, however, observed that the R^2 was relatively higher for small sample sizes ($n \leq 30$) compared to the large sample sizes ($n \geq 30$).
4. At k= 0.1 all issues of multicollinearity were highly resolved, as the VIF values were reduced to values that are indicative of the absence of multicollinearity (VIF approximately equal to 1)

The study validates the effect of ridge regression biasing constant on model statistics (R^2). The study also revealed that the VIF consistently decreases as the ridge biasing constant increases. While the ridge regression has a slight effect on the R^2, sample sizes were found not to have any significant change or pattern on the VIFs.

Article DOI: 10.52589/AJMSS-EY0EE7OV
DOI URL: https://doi.org/10.52589/AJMSS-EY0EE7OV

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 5, Issue 2, 2022 (pp. 61-71) www.abjournals.org

REFERENCES

Achmad Efendi and Effrihan (2017). A Simulation Study on Bayesian Ridge Regression Models for Several Collinearity Levels. AIP Conference Proceedings 1913, 020031 (2017); Published Online: 05 December 2017.

Al-Hassan, Y. M (2010). Performance of a new ridge regression estimator. Journal of the Association of Arab Universities for Basic and Applied Sciences, Vol. 9: 23–26

Daoud, J. I (2017). Multicollinearity and Regression Analysis. IOP Conf. Series: Journal of Physics: Conf. Series 949 (2017) 012009

Dorugade, A. V., Kashid, D. N. (2010). An alternative method for choosing ridge parameters for regression. Applied Mathematical Science, 4(9), 447-456.

Fitrianto, A & Yik, L. C. (2014). Performance of Ridge Regression Estimator Methods on Small Sample Size by Varying Correlation Coefficients: A Simulation Study. Journal of Mathematics and Statistics 10 (1): 25-29.

Herawati, H., Nisa, K., Setiawan, E & Nusyirwan, T. (2018). Regularized Multiple Regression Methods to Deal with Severe Multicollinearity. International Journal of Statistics and Applications, 8(4): 167-172

Isamu Nagai (2014). Selection of model selection criteria in multivariate ridge regression. Hiroshima Math. J. 43(1): 73 - 106

Mason, C. H & Perreault, W. D (1991). Collinearity, Power, and Interpretation of Multiple Regression Analysis. Journal of Marketing Research, Vol. 28(3):268-280.

Neter John, Michael H. Kutner, William Wasserman (1983). Applied Regression Models. 2nd Edition.

Pereira, J. M., Bastoa, M & Ferreira da Silvab, A. (2015). 3rd GLOBAL Conference On Business, Economics, Management And Tourism, 26-28 November 2015, Rome, Italy

Stewart G. W. (1987) Collinearity and Least Squares Regression. Statistical Science, Vol. 2, No. 1, 68 - 100

Vedide Rezan Uslu, Erol Egrioglu, Eren Bas (2014). Finding Optimal Value for the Shrinkage Parameter in Ridge Regression via Particle Swarm Optimization. American Journal of Intelligent Systems 2014, 4(4): 142-147

Won Suk Hye, Hwang, Heungsun (2010) Regularized fuzzy clusterwise ridge regression in: Advances in Data Analysis and Classification, Vol. 4, No. 1, 2010, p. 35-51.

Article DOI: 10.52589/AJMSS-EY0EE7OV
DOI URL: https://doi.org/10.52589/AJMSS-EY0EE7OV