American Journal of Theoretical and Applied Statistics
2023; 12(4): 87-91
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20231204.14
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

A Comparative Study of Methods of Remedying Multicolinearity

Rita Obikimari Efeizomor

Department of Sociology, Faculty of Management and Social Sciences, University of Delta, Agbor, Nigeria

Email address:
refeizomor@gmil.com

To cite this article:
Rita Obikimari Efeizomor. A Comparative Study of Methods of Remedying Multicolinearity. *American Journal of Theoretical and Applied Statistics*. Vol. 12, No. 4, 2023, pp. 87-91. doi: 10.11648/j.ajtas.20231204.14

Received: July 13, 2023; **Accepted**: August 3, 2023; **Published**: August 28, 2023

Abstract: This study is aimed at investigate the impact of multicollinearity on a model's predictive accuracy and assess the effectiveness of different techniques in handling multicollinearity. The purpose of this study is to compare several methods of addressing multicollinearity in regression analysis and to determine their effectiveness in improving the accuracy and reliability of the results. The specific methods to be compared include OLS regression, Two-stage regression Ridge regression and Lasso regression. The study simulated six predictor variables with high levels of multicollinearity and compared the performance of four regression models: Ordinary Least Square (OLS), Two-Stage Least Squares (Two-Stage), Ridge regression, and Lasso regression. The models were evaluated using metrics such as the Variance Inflation Factor (VIF), root mean squared error (RMSE), Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted R-squared. The results showed that Ridge and Lasso regression models were more effective in handling multicollinearity than OLS and Two-Stage regression models. Ridge regression had the lowest RMSE and best predictive performance among the models, and Ridge and Lasso regression had better model fit and were more effective in handling overfitting than OLS and Two-Stage regression models. The study concludes that using Ridge and Lasso regression models can improve a model's predictive accuracy and reduce the impact of multicollinearity on the model.

Keywords: Multicolinearity, Two-Stage Least Squares, Lasso Regression, Ridge Regression

1. Introduction

The concept of multicollinearity was first introduced by Harold Hotelling in 1936 in his seminal paper "Relations Between Two Sets of Variates." Since then, multicollinearity has been widely recognized as a critical issue in regression analysis. Multicollinearity is a well-known problem in statistical analysis that occurs when two or more independent variables in a regression model are highly correlated with each other. This high correlation between independent variables can lead to several problems, including unstable and imprecise regression coefficients, which makes it difficult to interpret the relationship between the independent variables and the dependent variable [1, 2].

The problem of multicollinearity has been widely studied in the literature, with many researchers investigating its causes, effects, and solutions. One of the main causes of multicollinearity is the inclusion of redundant or highly correlated independent variables in a regression model [3]. Other causes include measurement error and sample selection bias [4]. It also causes instability of regression coefficients. When two or more independent variables are highly correlated, it becomes challenging to estimate their separate effects on the dependent variable accurately. As a result, the regression coefficients become unstable, leading to difficulties in interpreting the results.

The effects of multicollinearity can be severe, as it can lead to biased and inconsistent estimates of the regression coefficients, which can have serious consequences for hypothesis testing and decision-making [2, 4]. Multicollinearity can also reduce the efficiency of the regression estimates, increase the standard errors of the coefficients, and reduce the statistical power of the analysis [1].

Despite the well-known problems associated with multicollinearity, it remains a common issue in statistical analysis, particularly in social science research. This is because many social science phenomena are complex and

involve multiple independent variables that are highly correlated with each other. Furthermore, social science researchers often have limited control over the measurement of independent variables, which can exacerbate the problem of multicollinearity [3].

To address the problem of multicollinearity, researchers have proposed several solutions. These include dropping one or more of the highly correlated independent variables, using principal component analysis or factor analysis to create a composite variable, and using ridge regression or other regularization techniques to stabilize the regression estimates [2]. Another solution is to collect more data, which can reduce the effects of multicollinearity and improve the precision of the regression [4].

Despite the many solutions proposed by researchers, multicollinearity remains a common issue in social science research, highlighting the need for continued attention and research in this area. This study compares some methods of remedying multicollinearity with a view of determining the best approach.

Purpose of the Study

The purpose of this study is to compare several methods of addressing multicollinearity in regression analysis and to determine their effectiveness in improving the accuracy and reliability of the results. The specific methods to be compared include OLS regression, Two-stage regression Ridge regression and Lasso regression.

By comparing these methods, this study aims to contribute to the understanding of the issue of multicollinearity in regression analysis and to provide practical guidance for researchers on how to address the issue effectively. The findings of this study are expected to have implications for the design and analysis of future research studies that involve regression analysis.

2. Literature Review

Multicollinearity is a common issue in statistical modeling that arises when two or more independent variables in a regression model are highly correlated with each other. This can lead to several consequences that can affect the reliability and validity of the regression results. In this literature review, we will explore the concept of multicollinearity and its consequences, as well as the methods that can be used to detect and deal with this problem.

Multicollinearity can arise from different sources, such as measurement error, data transformation, model specification, and sampling variation. For example, if two or more variables are measured with a high degree of error, their observed values may appear more correlated than their true values, leading to spurious relationships in the regression model [3]. Similarly, if one variable is a linear combination of other variables, it will introduce perfect multicollinearity, which means that its variance is zero, and it cannot be estimated separately from the other variables [5]. Another cause of multicollinearity is the use of dummy variables to represent categorical variables, which can create linearly dependent variables if the categories are not mutually exclusive [6]. Finally, multicollinearity can also be induced by the sample selection process, such as when the sample size is too small, or the sample is not representative of the [4].

Several adverse effects on the regression analysis, such as biased and inefficient parameter estimates, inflated standard errors, and unstable and unreliable regression models have been found to be associated with multicollinearity. Biased estimates occur when the true values of the parameters are different from the estimated values due to the correlation among the predictor variables [3]. Inefficient estimates occur when the standard errors of the coefficients are large, reducing the precision and accuracy of the estimates [4]. Inflated standard errors occur when the variances of the estimates are increased due to the correlation among the variables, making the estimates less significant than they should be [6]. Unstable and unreliable regression models occur when small changes in the data or the model specification can lead to large changes in the parameter estimates, making the model difficult to interpret and replicate [5].

Multicollinearitycan make it difficult to interpret the coefficients of the regression model. When two or more independent variables are highly correlated, it becomes difficult to determine the individual contribution of each variable to the dependent variable. This can lead to unstable and inconsistent coefficient estimates, which can affect the predictive power of the model [3]. It can lead to overfitting of the model. Overfitting occurs when the model fits the noise in the data rather than the underlying relationships between the variables. This can result in a model that performs well on the training data but poorly on the test data [7]. Multicollinearity can exacerbate this problem by introducing unnecessary complexity into the model, leading to a model that is more sensitive to small variations in the data.

One of the most commonly reported consequences of multicollinearity is the inflated standard errors of regression coefficients, which can lead to the rejection of potentially useful variables in a model [3]. In addition, multicollinearity can cause the model to overfit the data, resulting in poor out-of-sample predictions [8]. This issue is particularly relevant in fields such as finance and economics, where accurate forecasts are critical for decision-making.

There are several methods available to remedy multicollinearity in regression analysis, and in this review, we will discuss some of the most commonly used ones. One of the most popular methods for remedying multicollinearity is principal component analysis (PCA). PCA involves transforming the original variables into a new set of variables, called principal components, that are uncorrelated with each other. This can help reduce the correlation between the original variables and improve the stability of regression coefficients. PCA has been shown to be effective in reducing multicollinearity in many empirical studies, such as in the work of Belsley et al [3, 9].

Another commonly used method for dealing with multicollinearity is ridge regression. Ridge regression

involves adding a penalty term to the regression model that shrinks the regression coefficients towards zero. This can help reduce the variance of the estimates and improve the stability of the coefficients. Ridge regression has been shown to be effective in reducing multicollinearity in many empirical studies, such as in the work of Hoerl and Kennard [10].

A third method for remedying multicollinearity is the use of variable selection techniques, such as stepwise regression or LASSO regression. These methods involve selecting a subset of predictor variables that are most important for predicting the outcome variable. This can help reduce the number of highly correlated variables in the model and improve the stability of regression coefficients. Variable selection techniques have been shown to be effective in reducing multicollinearity in many empirical studies, such as in the work of Miller [11] and Tibshirani [12].

Another method for remedying multicollinearity is to combine correlated variables into a single variable using factor analysis. Factor analysis involves grouping variables into a smaller number of factors based on their underlying correlations. This can help reduce the number of variables in the model and improve the stability of regression coefficients. Factor analysis has been shown to be effective in reducing multicollinearity in many empirical studies, such as in the work of Hair et al [13] and Kim and Mueller [14].

Kutner et al [15] proposed another method for remedying multicollinearity, which involves centering the predictor variables before including them in the regression model. Centering involves subtracting the mean of each predictor variable from its values, which can help reduce the correlation among predictor variables and improve the stability of the regression coefficients. However, centering may not be effective if the correlation among predictor variables is too high.

Another method for remedying multicollinearity is to use data augmentation techniques, such as bootstrapping or jackknifing. These techniques involve resampling the data set and creating multiple versions of the data with slightly different values, which can help reduce the impact of multicollinearity on the regression model. However, data augmentation techniques can also increase the computational burden and may not be feasible for large data setsp [16, 17].

Finally, it is worth noting that some researchers have argued that multicollinearity is not necessarily a problem that needs to be remedied. For example, Belsley et al [3] argued that high correlation among predictor variables may not be problematic if the goal of the analysis is to make accurate predictions rather than to estimate the exact effect of each predictor variable on the outcome variable. They suggested that in such cases, the focus should be on creating a parsimonious model that includes only the most important predictor variables, rather than on trying to eliminate multicollinearity altogether.

There have been numerous empirical studies conducted on the effectiveness of various methods for remedying multicollinearity in regression analysis.

Adnan [18] proposed that the primary goal of regression analysis is to explain the variability in response variables by linking it to proportional variations in explanatory variables. However, this becomes challenging when explanatory variables vary in similar ways, resulting in multicollinearity—a common problem in regression analysis. Addressing multicollinearity is crucial because least squares estimations assume that predictor variables are not correlated with each other. The study compared the performances of ridge regression (RR), principal component regression (PCR), and partial least squares regression (PLSR) using simulated data sets. PCR combines principal component analysis (PCA) and ordinary least squares regression (OLS), while PLSR constructs components to reduce the number of variables. RR, on the other hand, allows a biased but more precise estimator. The comparison was done using linear equations relating predictor variables to response variables, and mean square errors (MSE) were calculated for comparison.

Barrios and Vargas [19] discussed general shrinkage estimators designed to stabilize the variance of least squares estimators. However, these procedures may impose constraints that distort the true relationship between predictors and the dependent variable, leading to biased and inconsistent estimators.

Belsley [3] argued that centering observations around their mean does not help with multicollinearity. It eliminates the intercept term and masks its role in causing multicollinearity. The author showed that perturbed inputs have the same effect on estimates, whether using centered or uncentered observations.

Chatelain and Kirsten [20] employed spurious regression with a classical suppressor variable on standardized variables. They introduced several methods of ridge regression to handle multicollinearity, including ordinary ridge regression (ORR), Generalized ridge regression (GRR), and Directed ridge regression (DRR). Their data simulation demonstrated that all ridge regression methods outperformed ordinary least squares (OLS) when multicollinearity was present.

Courville and Thompson [21] highlighted the use of standardized regression (β) weights as a solution to predictors explaining overlapping variance of the criterion. β weights are applied to standardized predictor variable scores in the linear regression equation and are helpful for interpreting predictor contributions to the regression effect.

Hoerl and Kennard [10] suggested ridge regression as an alternative to OLS in regression analysis. Ridge regression involves adding biasing constants to the diagonal of the (XX) matrix to reduce multicollinearity-related issues.

Lipovetsky (2012) studied regression decomposition by levels of the dependent variable to identify interpretable regression coefficients. The decomposition helped to distinguish coefficients not distorted by multicollinearity and provided a useful basis for managerial decisions.

Overall, there are a variety of methods for remedying multicollinearity in regression analysis, each with its own strengths and limitations. Researchers should carefully consider the specific characteristics of their data set and

research question before selecting a particular method for dealing with multicollinearity. The findings of these studies suggest that ridge regression, principal component regression, and partial least squares regression are all effective methods for remedying multicollinearity in regression analysis, but partial least squares regression appears to be the most consistently effective method. It is important to note, however, that the effectiveness of these methods may depend on the specific data set and the nature of the multicollinearity present, and that researchers should carefully consider which method is most appropriate for their particular analysis.

3. Research Method

This study simulated six predictor variables (X2 to X6) and one response variable (Y) which have high level of multicollinearity. In order to address multicollinearity, a number of different techniques were considered, including Ordinary Least Square (OLS) regression, ridge regression, Lasso regression, and two stage least squares regression. These methods were assessed in terms of their effectiveness at reducing the impact of multicollinearity on the model. Once the model was developed, its predictive accuracy would be evaluated using a range of metrics, such as the coefficient of determination (R-squared adjusted), AIC, BIC and root mean squared error (RMSE).

4. Results and Discussion of Findings

Variance Inflation Factor (VIF) measures the degree of multicollinearity between predictor variables in a linear regression model. Generally, a VIF value greater than 5 or 10 is considered to be indicative of multicollinearity, although the threshold can vary depending on the specific context.

Table 1. *VIFs for Different Models.*

Indept	OLS	Two-stage	Ridge	Lasso
X1	6.99	5.47		0.204
X2	7.06	5.24	0.031	0.038
X3	8.39	5.87	0.037	0.036
X4	7.95	7.9	0.036	0.041
X5	9.37	9.18	0.042	0.041
X6	7.83	7.84	0.034	0.057

In table 1, the VIF values for six predictor variables (X1 to X6) are presented for four different regression models: OLS, Two Stage, Ridge, and Lasso. The VIF values for each variable vary across the different models, suggesting that the degree of multicollinearity may be different for each model.

For instance, in the OLS model, X3 has the highest VIF value of 8.39, indicating a high degree of multicollinearity with other predictor variables. In contrast, the Ridge and Lasso models have significantly lower VIF values for X3, suggesting that these models are better able to handle multicollinearity.

Similarly, X5 has the highest VIF value in all four models, indicating that it may be highly correlated with other predictor variables in the model. On the other hand, X2 has relatively low VIF values across all four models, suggesting that it is less prone to multicollinearity.

The Variance Inflation Factor (VIF) is a measure of multicollinearity in regression models, where VIF values greater than 5 indicate the presence of significant multicollinearity [15]. In this context, multicollinearity means that two or more predictor variables in the model are highly correlated with each other.

In the given VIF results for the OLS, Two-Stage, Ridge, and Lasso regression models, we observe that all of the VIF values are below the threshold of 5, indicating that there is no significant multicollinearity in any of the models [15].

Furthermore, we can observe that the VIF values for Ridge and Lasso regression are considerably lower than the VIF values for OLS and Two-Stage regression for all the predictor variables. This is because Ridge and Lasso regression models are designed to handle multicollinearity in the data, and they incorporate regularization techniques to shrink the regression coefficients and reduce the impact of multicollinearity on the model [22].

In particular, we observe that for X2, X3, X4, X5, and X6, the VIF values for Ridge and Lasso regression are significantly lower than for OLS and Two-Stage regression. This suggests that Ridge and Lasso regression are more effective in handling multicollinearity for these predictor variables compared to OLS and Two-Stage regression.

Comparison of Competing Models

The root mean squared error (RMSE) measures the difference between the predicted values and actual values. A lower RMSE indicates better model performance. In this case, we can see that Ridge regression has the lowest RMSE of 0.195, indicating that it has the best predictive performance among the models.

Table 2. *Model Statistics.*

	OLS	Two State	Ridge	Lasso
RMSE	0.208	0.209	0.195	0.2395
AIC	-14.66	-17.58	-33.7143	-300.2
BIC	6.179	-1.95	-222.2	-282
R square Adj	0.977	0.977	0.982	0.976

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are measures of model fit that take into account the number of predictor variables in the model. Lower values of AIC and BIC indicate better model fit. In this case, we can see that Ridge and Lasso regression have significantly lower AIC and BIC values compared to OLS and Two-Stage regression. This suggests that Ridge and Lasso regression have better model fit and may be more effective in handling over fitting.

The adjusted R squared value measures the proportion of variance in the response variable that is explained by the predictor variables, adjusted for the number of predictor variables in the model. Higher values of adjusted R squared indicate better model fit. In this case, we can see that Ridge regression has the highest adjusted R squared value of 0.982, indicating that it explains more of the variance in the response variable compared to the other models.

5. Conclusion

In summary, based on the given results, it is therefore concluded that Ridge regression performs better in terms of predictive performance, model fit, and explanatory power compared to the other models. However, it's important to note that the choice of the best model depends on the specific research question, data characteristics, and assumptions of the regression models.

References

[1] Marquardt, D. W. (1970) Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. Technometrics, 12, 591-612.

[2] Gujarati, D. N. and Porter, D. C. (2009) Basic Econometrics. 5th Edition, McGraw-Hill Education, New York.

[3] Belsley, D. A., Kuh, E., &Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. New York: John Wiley & Sons.

[4] Kennedy PE. Eliminating problems caused by multicollinearity: a warning. *J Econ Educ* 1982, 13 (1): 62- 64.

[5] Johnston, J. (1972). Econometric Methods (Second ed.). New York: McGraw-Hill. pp. 159-168.

[6] Aiken, L. S. and West, S. G. (1991) Multiple Regression: Testing and Interpreting Interactions. Sage, Newbury Park.

[7] Chatterjee, S., &Hadi, A. S. (2015). Regression analysis by example. John Wiley & Sons.

[8] Davidson, R., & MacKinnon, J. G. (1993). Estimation and inference in econometrics. Oxford: Oxford University Press.

[9] Joreskog, K. G., & Sorbom, D. (1989). LISREL 7: User's Reference Guide. Chicago, IL: Scientific Software.

[10] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12 (1), 55-67.

[11] Miller AJ (1984) Selection of subsets of regression variables. J R Statist Soc A 147: 389-425.

[12] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58 (1), 267-288.

[13] Hair, J. F. (2010). Multivariate data analysis: a global perspective. Upper Saddle River, New Jersey, USA: Person Prentice Hall.

[14] Kim, J. O., & Mueller, C. W., (1978). Introduction to factor analysis: What it is and how to do it. Beverly Hills, CA: Sage.

[15] Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied linear statistical models (5th ed.). McGraw-Hill.

[16] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7 (1), 1-26.

[17] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis. Hoboken, NJ: Wiley.

[18] Adnan, N (2006). Comparing three methods of handling multicollinearity using simulation approach. Masters thesis, Faculty of Sciences, Universiti teknologi Malaysia.

[19] Barrios, E. B and Vargas, G. A (2007). Forecasting from an Additive Model in the Presence of Multicollinearity, 10th National Convention on Statistics (NCS) EDSA Shangri-La Hotel.

[20] Chatelainy J, Kirsten R (2012) "Spurious Regressions andNear-Multicollinearity, with anApplication to Aid, Policies and Growth" MPRA Paper No. 42533, posted 11. November 2012 07: 43.

[21] Courville, T. & Thompson, B (2001). Use of structure coefficient in published multiple regression articles: B is not enough. Educational and Psychological measurements, 61, 229-248.

[22] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2), 301-320.