International Journal of Computational Research in Science and Technology
Volume 1 Issue 1
Sep-Oct 2025, PP 33-38

A Comparative Analysis of The Performance of Some Penalized Regression Techniques in The Presence of Multicollinearity

Nwuzor, Ozoemena[1], A.U .UDOM[2]

[1] Dept. of Industrial Mathematics and Statistics, Faculty of Science, Ebonyi State University

[2] Dept. of Statistics, Faculty of Physical Sciences, University of Nigeria, Nsukka

ozoemena.nwuzor@ebsu.edu.ng, akaninyene.udom@unn.edu.ng

Abstract. Multicollinearity is a common issue faced by statisticians and machine learning practitioners when building predictive models. This study compared the performance of some penalized regression techniques (Lasso, Ridge and Elastic Net) in the presence of multicollinearity using real-life datasets. The comparison was carried out in terms of the accuracy, precision, and recall scores of each technique using the root mean square error, residual sum of squares and R-square. The outcome of this study using the real-life dataset on 442 diabetes patients measured on 10 baseline predictor variables and one measure of disease progression showed that the Ridge regression performed better than Lasso and Elastic net. Comparing the results of the methods, It was observed that Ridge model performed better in comparing the performance of some penalized regression models with multicollinearity.

Index Terms: Multicollinearity.

I. Introduction

Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly related. Multicollinearity is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy (Funda, 2015). Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model (Adams, 1991). Multicollinearity can bring about wider confidence intervals that produce smaller reliable probabilities in terms of the effect of independent variables in a model. Multicollinearity occurs in totality when there are high correlations between two or more predictor variables in which one predictor variable can be used to predict others (Moses, 2018). This creates redundant information, skewing the results in a regression model. Examples of correlated predictor variables (also called multicollinear predictors) are: a person's height and weight, age and sales price of a car, or years of education and annual income. In order to deal with these challenges, variables selection and shrinkage estimation are becoming important. The traditional approach of automatic selection (such as forward selection backward elimination and stepwise selection) and best subset selection are computationally expensive and may not necessarily produce the best model (Breiman, 1996).

Multicollinearity problem can be solved using penalized least squares (PLS) method by putting some constraints on the values of the parameters estimated (Fan & Li, 2001). Penalized regression techniques are methods that keep all the predictor variables in the model but constrain the regression coefficients by shrinking them towards zero. If the amount of shrinkage is large enough, these methods can also perform variable selection by shrinking some coefficients to zero. A penalized regression technique gives a sequence of models, each associated with specified values for one or more tuning parameters (Funda, 2015). Examples of penalized regression techniques are:

- Least Absolute Shrinkage Selection Operator (LASSO)
- Ridge
- Elastic Net

LASSO

The LASSO is a penalized regression technique used for estimating linearmodels that involve minimizing sum of square errors with an L1-norm penalty.

Ridge regression technique

In ridge regression, a penalty is applied on the coefficients, so that they are shrunk towards zero.This is also having the effect of reducing the variance and hence, the prediction error.

Elastic Net

The elastic net regression is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods.The benefit of using this technique is that it allows a balance of both penalties, which can result in better performance than a model with either one or the other penalty on some problems. The Variance Inflation Factor (VIF) was used to evaluate the presence of multicollinearity in the dataset used in this study.

The VIF is the measure of the amount of multicollinearity ina set of multiple regression variables. V IF = 1/(1-R^2) where R^2 is the tolerance .Values of VIF exceeding 10 are constantly regarded as VIF with multicollinearity but for V IF (β) > 15 indicates that the multicollinearity is high.

The aim of this study is to compare the performance of some penalized regression techniques (Lasso, Elastic net and Ridge) in the presence of multicollinearity.

II. Methodology

2.1 Penalized Regression Techniques Model

In a linear model, the response variable Y is represented as a linear combination of the predictor variables, $X_1, \ldots, X_p$, plus random noise

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip+}\epsilon_i$$

where β_0, β_1, ..,β_p are regression parameters, ϵ is the random noise term, and N is the observation number . The general form of the shrinkage and regularization methods for linear models is

$$\beta = minarg\beta \sum_{i=0}^{N} (Y - (X\beta)^2), i = 1,2, \ldots, N$$

Subject to Pen $(\beta) \leq t$,

Where Pen(β) is a specific penalty function while t is a tuning parameter, N is the number of observations in the data set. ordinary least squares which minimizes the residual sum of squares can be written as:

$RSS(\beta) = \sum_{i=0}^{N}(\beta_0+\beta_1 X_{i1}+\ldots+\beta_p X_{ip})$

Lasso penalty function is written as:

β $Lasso=argmin$ $\beta \epsilon R(Y - X\beta)^T(Y - X\beta) + \lambda + \lambda$

Where λ is a nonnegative regularization parameter that controls the amount of shrinkage. $\lambda \sum |\beta|$ is called "L_1 penalty," . The estimates from the elastic net method are expressed by

$\widehat{\beta}$ = argmin$\beta(||Y - X\beta||^2 + (\lambda_2||\beta||^2 + (\lambda_1||\beta||.^1)$ In ridge regression, the OLS loss function is augmented in such a way that it does not only minimize the sum of square residuals but also penalize the size of parameter estimates, in order to shrink them towards zero:

$L_{ridge}(\hat{\beta}) = \sum_{i=1}^{n}(Y_i - x_{1i}(\hat{\beta}))^2 + \lambda \sum_{j=1}^{n}(\widehat{\beta})^2 = ||y - X\hat{\beta}||2 + \lambda||\hat{\beta}||^2$

III. Real-life Dataset Results

Table 1. The Variance Inflation Factor(VIF) values of the ten predictor variables of the real-life dataset.

Variable	VIF
AGE	1.2250
SEX	1.2976
BMI	1.6043
BP	1.5282
TC	62.9675
LDL	41.7728
HDL	15.4509
TCH	8.8167
LTG	10.3493
GLU	1.5127

Table 2 Model Evaluation using the test data

Model	RSS	RMSE
Ridge	864618.6	53.33048
Lasso	866256.7	53.38097
Elastic net	3745235	92.05098

IV. Real-life Dataset Results

This work assessed the performance of three penalized regression techniques (Lasso, Ridge and En) in dealing with multicollniearity in regression analysis. The findings suggest that Ridge models outperform Lasso and En models when using real-life datasets. This indicates that the performance of these techniques may vary depending on the characteristics of the dataset being analyzed. Based on findings from the research, It can be concluded that the performance of these techniques can vary depending on the dataset and the method of analysis.

References

1. Adams, J. (1991, August). A computer experiment to evaluate regression strategies. Paper presented at the Statistical Computing Section, American Statistical Association Meeting, New York, USA.Retrieved fromhttp://www.evolvinglogic.com/news/el_validation.html
2. Altelbany, S. (2021). Evaluation of ridge, elastic net and lasso regression methods in precedence of multicollinearity problem: A simulation study. Journal of Applied Economics and Business Studies, 5(1), 131-142.
3. Andre, N., Young, T. M., & Rials, T. (2006). Online monitoring of the buffer capacity of particle Board furnish by near-infrated spectroscopy. Applied Spectroscopy,60(10),1204-1209.
4. Ayers, K. L., & Cordell, H. J. (2010). SNP selection in genome-wide and candidate gene studies via penalized logistic regression. Genetic Epidemology. 34(4), 879-891.
5. Breiman, I. (1996). Heuristics of instability and stability in model selection. The Annals of Statistics, 24(6), 2350-2383.
6. Buhlmann, P., & Vandegeer, S. (2011). Statistics for high dimensional data: methods, theory and applications. Bern , Switzerland: Springer Science and Business Media.
7. Cardot, H., & Sarda, P. (2005). Estimation in generalized linear models for functional data. Journal of Multivariate Analysis, 92(1), 24-41.
8. Doreswamy, V., & Chanabasayya, M.V. (2013). Performance analysis of regularized linearregression models. International Journal of Computational Science and Information Technology, 1(4), 20-33.
9. Efron, B. Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression.The Annals of Statistics, 32(2), 407-499.

10. Fan, J. & Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(4), 1348-1360.
11. Friedman, J., Hastie, T., Hofling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization, Annals of Applied Statistics, 1(2), 302-322.
12. Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization path for generalized linear models via Coordinate decent. Journal of Statistical Software, 33(1), 1-22.
13. Funda, G.(2015). Penalized regression methods for linear models in SAS/STATA. Paper presented at the SAS Global Forum 2015 Conference. North Carolina, USA. Retrieved from http://support. sas. com/rnd/app/stat/papers/2015/PenalizedRegression_LinearModels.pdf
14. Gertheiss, J., & Tutz, G. (2009). Penalized regression with ordinal predictors. International Statistical Review, 77(3), 345-365.
15. Greenwood, C. J., Youssef, G. J., Letcher, P., Macdonald, J. A., Hagg, L. J., Sanson, A., & Olsson,C. A. (2020). A comparison of penalised regression methods for informing the selection of predictive markers. PloS One, 15(11), e0242730.
16. Heckman, N. E., & Ramsay, J. O. (2000). Penalized regression with model-based penalties. Canadian Journal of Statistics, 28(2), 241-258.
17. Jason, B. (2020). How to develop elastic net regression model in python. Retrieved from https://machinelearningmastery.com/elastic-net-regression-in-python/
18. Kassambara, A.(2018). Linear regression essentials in R-Articles. (1st ed.). Paris, France:Independent publishing platform.
19. Kooperberg, C., LeBlane, M., & Obenchain, V. (2010). Risk prediction using genome- wideassociation studies, Genetics Epidemiology, 34(7), 643 – 652.
20. Kutner, M.H.,Nachtsheim, C.J., Neter, J. & Li, W. (2004). Applied linear statistical models,(5th
21. ed.). New York, USA: McGraw Hill Inc.
22. Lee, S. (2015). A note on standardization in penalized regressions. Journal of the Korean Data and Information Science Society, 26(2), 505-516.
23. Matthew, P.K., & Yahaya, A. (2015). Performance analysis on lasso, elastic net and correlation Adjusted elastic net regression methods. International Journal of Advanced Statistics and Probability, 3(1), 93-99.
24. Moses, J. (2018). Comparison of the penalized regression techniques with classical least squares in minimizing the effect of multicollinearity(Master's dissertation).Ahmadu Bello University, Zaria, Nigeria.
25. Pan, W., Xie, B., & Shen, X. (2010). Incorporating predictor network in penalized regression with application to microarray data. Biometrics, 66(2), 474-484.
26. Tharmatranam, K.,Claeskens, G., Croux, C., & Salibian-Barrera, M.(2010). S-Estimation for Penalized Regression Splines. Journal of Computational & Graphical Statistics, 19(3), 609-625.
27. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of Royal Statistical Society, Series B, 58, 267-288.
28. Tutz, G., & Ulbricht, J. (2009).Penalized regression with correlation-based penalty. Statistics and Computing, 19(3), 239-253.
29. Waldron, L., Pintilie, M., Tsao, M. S., Shepherd, F. A., Huttenhower, C., & Jurisica, I. (2011). Optimized application of penalized regression methods to diverse genomic data. Bioinformatics, 27(24), 3399-3406.

30. Zhao, Z., & Zucknick, M. (2020). Structured penalized regression for drug sensitivity prediction. Journal of the Royal Statistical Society Series C: Applied Statistics, 69(3), 525-545.
31. Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101, 1418–1429.
32. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of Royal Statistical Society, Series B, 67(2), 301-320.