KASU JOURNAL OF MATHEMATICAL SCIENCES (KJMS) VOL. 3, ISSUE 1, JUNE 2022
IISSN 2734-3839 (Print), ISSN 2735-962X (E-Copy)
http:/www.journal.kasu.edu.ng/index.php/kjms

KASU JOURNAL OF MATHEMATICAL SCIENCES

Penalization Techniques for Remedying Multicollinearity in Multiple Regression Model

[1*]Pascalis Kadaro Matthew, [1]Felix Adanaá Chama, [2]Nathan S. Agog

[1]Department of Mathematical Sciences, Taraba State University, Jalingo - Nigeria
[2]Department of Mathematical Sciences, Kaduna State University, Kaduna – Nigeria

***Corresponding Author:** E-mail: pkadaro@yahoo.com

Abstract
Multicollinearity is another important issue in multiple regression. Multicollinearity refers to a situation in which two or more predictor variables are highly linearly correlated, i.e. a linear relationship exists between two or more predictor variables. Penalization is one of the best technique for remedying multicollinearity in a model. In this paper, we used a data (with high possibility of multicollinearity) obtained from previous literature to assess the performance of four penalized regression methods, namely; LASSO, Ridge, Elastic net and SCAD. It shows that, all the penalized regression methods considered have lower mean square error (MSE), than the ordinary least squares regression. It also shows that, penalization techniques can make the predictive accuracy of a model better, by lowering the variability in the measures of a regression coefficients, which shrinks the estimates towards zero.

Keywords: Mean Square Error; Multicollinearity; Multiple Regression; Penalization Technique.

1. Introduction
Multicollinearity is a problem in multiple regression that develops when one or more of the independent variables is highly correlated with one or more of the other independent variables, i.e. if one independent variable is a perfect linear combination of the other independent variables. This happens, for example, when two predictor variables X_1 and X_2 satisfy the equation

$$X_2 = a + bX_1 \qquad (1)$$

for two real numbers a and b. The inclusion of both X_1 and X_2 in our model is problematic for estimation. Intuitively, a problem arises because the inclusion of both X_1 and X_2 adds no more information to the model than, the inclusion of just one of them. Effectively, we are asking the regression model to estimate an additional parameter, but we are not supplying it with any additional information. If X_1 is regressed on X_2 and the resulting $R^2 = 1.0$, then the matrix of inter-correlations between X_1 and X_2 is singular and there exists no unique solution for the regression coefficients.

Other consequences of multicollinearity are that some predictor variables are not statistically significant, but the model is overall significant, and that the usual interpretation of coefficient estimates fails in the presence of multicollinearity and there is high variability of parameter estimators, because the estimated variance-covariance matrix has large diagonal entries. On the other hand, suppose multicollinearity is detected and the predictor variables that cause multicollinearity are identified. As discussed by Ryan [7], multicollinearity may not be a problem if the goal is to use the linear regression model for prediction. However multicollinearity is a problem if we use the linear regression model for description or control. The term multicollinearity is use to describe situations where (i) there is a perfect linear relationship between the independent variables and (ii) there is a nearly perfect linear relationship between the independent variables (r_{12}^2 close to one). In practice, scholars almost never face perfect multicollinearity. However, they often encounter near-perfect multicollinearity, although the standard errors are technically correct and will have minimum variance. With near-perfect multicollinearity, the variance will be very large. This means that, the independent variables are not providing much needed independent information in the model and so the coefficients are not estimated with a lot of certainty.

1.1 Detecting Near-Perfect Multicollinearity

(i)A classic sign of near-perfect multicollinearity is when you have a high R^2 but none of the variables show significant effects.

(ii) If there is high pairwise correlations among the independent variables, there may likely be near-perfect multicollinearity.

(iii)Each of the X's can be regressed on all of the other X's to see if there are any strong linear dependencies. These are sometimes referred to as auxiliary regressions. If any of these models R^2 is greater than the main model R^2, then you may have a problem.

(iv)Variance Inflation Factors (VIF) can also be used to detect multicollinearity. These are measures of the how much the variance of the coefficients is inflated by multicollinearity. Variance Inflation Factors (VIF) is defined as

$$VIF = \frac{1}{1-R^2} \tag{2}$$

Where R^2 is the multiple coefficient of determination in a regression of the X_i on all other explanatory variables. As a rule of thumb if the $VIF > 10$, it indicates high collinearity.

1.2 Remedying Multicollinearity

(i) One method is to select a collection of predictor variables that are minimally correlated with each other. This avoids over-fitting the regression model and can be normally done with statistical software. However information from other predictor variables is often lost, since there is no clear way of selecting a collection of predictor variables that forms the best subset.

Omitting predictor variables may result in potential loss of information.

(ii) Another method is to include interaction terms into the model to account for high linear correlation among the predictor variables. There are several problems with this approach. One is

that the form of interaction is not unique and must be carefully determined. The other is that the model is much more complex and has too many terms which reduce the degrees of freedom of the inference of the response, and hence reduces the power for predicting and estimating the response. (iv) The method of penalized least squares (PLS), which is equivalent to penalized maximum likelihood, helps to deal with the issue of multicollinearity by putting constraints on the values of the estimated parameters.

2. Literature Review

Penalization is one of the methods of handling the problem of multicollinearity. Various penalization methods have been proposed to handle multicollinearity, beginning with ridge penalty (Hoerl & Kennard, [6]). It estimates the regression coefficients through l_1-norm penalty. It is well-known that ridge regression shrinks the coefficients of correlated predictor variables toward each other, allowing them to borrow strength from each other (Friedman et.al [5]). The least absolute shrinkage and selection operator (LASSO) was proposed by Tibshirani [8], to estimate the regression coefficients through l_1-norm penalty. Zou and Hastie [11], proposed the elastic net penalty which is based on a combined penalty of LASSO and ridge regression penalties in order to overcome the drawbacks of using the LASSO and ridge regression on their own. Usually, in high dimensional data the explanatory variables are correlated. If there is a group of highly correlated variables, the LASSO will randomly select only one variable from this group and drop the rest whereas elastic net will select the whole group of the highly correlated explanatory variables (Zou & Hastie, [11]). Analogously, Bondell and Reich [1] proposed a penalty called OSCAR to encourage selection of a group of highly correlated explanatory variables. Elastic net often performs better than LASSO in terms of prediction error when there is correlation among variables, also OSCAR has a comparable performance similar to elastic net (Zeng & Xie, [10]). Tutz and Ulbricht [9] proposed correlation-based penalty to deal with grouping effects. This penalty just makes variable shrinkage rather than variable selection. Elastic net penalty lacks consistent variable selection (oracle property), even though it outperforms LASSO. Zou and Zhang [12] proposed adaptive elastic net to handle grouping effects and enjoying oracle property simultaneously. El-Anbari and Mkhadri [3] explained through experimental studies that elastic net seems to be slightly less reliable if the correlation between explanatory variables is not so extreme (i.e. $\rho \leq 0.95$)

3. Material and Methods

The data used in this research comes from a study conducted by Efron *et al* [2]; whereby 442 diabetic patients were measured on 10 baseline variables to get a prediction model that measure a disease progression one year after baseline. The 10 baseline variables include Age, Sex, Body Mass Index (BMI), Blood Pressure (BP), and six other blood serum measurements. The data analysis was performed using R package *glmnet* which utilizes the capabilities of fast cyclical coordinated descent (CCD) algorithm (Friedman et al., [5]).

3.1. Penalized Regression

Consider a standard multiple linear regression model

$$\mathrm{y} = \mathrm{X}\beta + \varepsilon \tag{3}$$

Let $y = (y_1, y_2, ..., y_n)^T$ be the response vector and $\mathrm{X} = [|\mathrm{X}_1|, ..., |\mathrm{X}_p|]$ be the model matrix, where $\mathrm{X}_i = (x_{1i}, x_{2i}, ..., x_{nj})^T$, and $j = 1, ..., p$ are predictors. β_0 is the intercept, $\beta = (\beta_1, ..., \beta_p)$ is a column vector that contains the regression coefficients and ε is a vector of error terms assuming normal distribution $\varepsilon \sim N(0, \sigma_e^2)$ .For models where $n > p$, the values of the unknown parameters β_0 and β can be uniquely estimated by minimizing the residual sum of squares,

$$RSS = (y - X\beta)^T (y - X\beta)$$
$$\text{Subject to } Pen(\beta) \le t \tag{4}$$

Where $Pen(\beta)$ (specific penalty) is a function of $\beta = (\beta_1, ..., \beta_p)^T$ and t is a tuning parameter. This constrained optimization problem can be solved with the equivalent Lagrangian formulation which minimizes

$$PLS = RSS + Penalty = (y - X\beta)^T (y - X\beta) + \lambda Pen(\beta) \tag{5}$$

Where λ is a tuning parameter and controls the strength of shrinkage. For example, when λ = 0, no penalty is applied and we have the ordinary least squares regression. When λ gets larger, more weight is given to the penalty term. After a location and scale transformation, we can assume the response is centred and the predictors are standardized,

3.1.1. Ridge Regression

Hoerl and Kennard [6] introduced ridge regression. It is also called l_2 penalized regression. The ridge estimator is defined as

$$\beta_{RIDGE} = (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta \tag{6}$$

Where λ is a non-negative regularization parameter.

3.1.2. LASSO Regression

The lasso penalty by Tibshirani [8], regularizes the linear regression coefficients through a l_1 penalized least squares procedure i.e. $P(\lambda, \beta) = \lambda \|\beta\|_{l_1}$

The LASSO regression is defined by

$$\beta_{LASSO} = (y - X\beta)^T (y - X\beta) + \lambda \sum_{i=1}^{p} |\beta_i| \tag{7}$$

The resulting regression problem is non-linear in *y* and results in a convex optimization problem. The regularization parameter λ controls the amount of shrinkage and needs to be tuned or chosen based on some prior results.

3.1.3. Elastic Net Regression

The elastic net method introduced by (Zou and Hastie,[11]) is based on a compromise between the LASSO and ridge regression penalties

$$\beta_{ELASTICNET} = (y - X\beta)^T (y - X\beta) + \sum_{i=1}^{p} [\tfrac{1}{2}(1-\alpha)\beta_i^2 + \alpha|\beta_i|] \quad (8)$$

where$0 \leq \alpha \leq 1$ is a penalty weight. The EN with $\alpha = 1$ is identical to the lasso, whereas it turns out to be ridge regression with $\alpha = 0$ (Friedman *et al*., [5]). Setting α close to1makes the EN to behave similar to the LASSO, but eliminates problematic behaviour caused by high correlations. When α increases from 0 to1, for a given λ the sparsity of the minimization (i.e., the number of coefficients equal to zero) increases monotonically from 0 to the sparsity of the LASSO estimation. The elastic net can select more variables than observations.

3.1.4. SCAD Regression

Smoothly Clipped Absolute Deviation Penalty (SCAD) of Fan & Li [4]. Instead of using an ℓ_1 -penalty, they minimize:

$$\hat{\beta}_{SCAD} = \frac{1}{2}\|Y - X\beta\| + \sum_{j=1}^{p} P_\lambda\left(|\beta_j|\right) \quad (9)$$

Where $P_\lambda'(\theta) = \lambda\left(I(\theta \leq \lambda)\right) + \frac{(a\lambda - \theta)}{(a-1)} I(\theta > \lambda)$ for some a > 2 and θ> 0. This penalty is chosen for its good model selection properties.

3.2. Fitting and analyzing models

The whole path of results (in λ) for the Bridge, LASSO, Elastic Net and SCAD models were calculated using the path wise cyclical coordinate descent (CCD) algorithms –in glmnet in R. We used 10-fold cross validation (CV) within glmnet to entirely search for the optimal λ. A regularized profile plot of the coefficient paths for the four methods was also shown. Predictive accuracy was also assessed using the mean squared error (MSE).

4. Results and discussions

In previous sections, we stated that penalization is one of the technique for remedying multicollinearity, we also, described the four penalized linear regression methods considered in this article; while in this section, analysis was conducted using numerical data obtained from previous literature to investigate their individual performances.

4.1 Ordinary least squares regression

Table 4.1: Results of Ordinary Least Squares

Variables	Estimate	Std.Error	t-Value	Pr(>|t|)	VIF
INTERCEPT	-0.0000	0.0334	-0.0000	1.0000	0.0000
AGE	-0.0062	0.0369	-0.1700	0.8670	1.2170
SEX	-0.1481	0.0378	-3.9200	0.0001	1.2780
BMI	0.3211	0.0411	7.8100	0.0000	1.5090
BP	0.2004	0.0404	4.9600	0.0000	1.4590
TC	-0.4893	0.2574	-1.9000	0.0579	59.2030
LDL	0.2945	0.2094	1.4100	0.1604	39.1930
HDL	0.0624	0.1313	0.4800	0.6347	15.4020
TCH	0.1094	0.0990	1.1000	0.2735	8.8910
LTG	0.4641	0.1062	4.3700	0.0000	10.0760
GLU	0.0418	0.0408	1.0200	0.3060	1.4850

Residual standard error: 0.7025 on 431 degrees of freedom

Multiple R-squared: 0.5177, Adjusted R-squared: 0.5066

F-statistic: 46.27 on 10 and 431 DF, p-value :< 2.2e-16.

From table 4.1 above, variables *TC, LDL, HDL, TCH and LTG* all have Variance Inflation Factors (VIF) greater than 10. This indicate that, there may likely be a problem of multicollinearity in the data.

4.2. Results on Elastic net regression

The (Fig. 1) below gives the relationship between $ln\,\lambda$ and MSE. The integers at the top show the number of non-zero estimators for the model. The left line gives the smallest MSE with eight variables in the model, and the right line gives the smallest Standard Error (SE) with only seven variables in the model. The Elastic Net was calculated based on an optimal value of $\alpha = 0.16$.

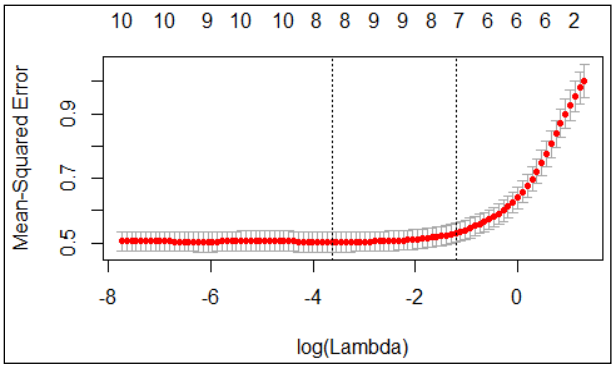

Fig. 1: MSE Plot and the Number of Variables in the Model as A Function of Log (λ) for the 10-Fold Cross Validation.

4.3. Results on LASSO regression

Figure 2 illustrates the relationship between ln λ and MSE. The integers at the top show the number of non-zero estimators for the model. The left line gives the smallest MSE with seven variables in the model, and the right line gives the smallest Standard Error (SE) with only four variables in the model. The LASSO regression model was obtained based on an optimal value of $\lambda = 0.0129$.

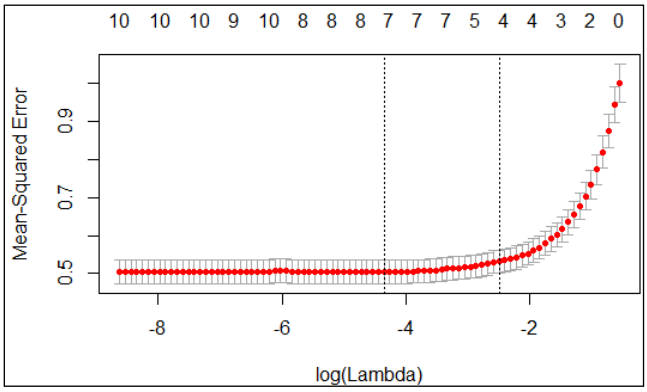

Fig. 2: MSE Plot and the Number of Variables in the Model as A Function of Log (λ) for the 10-Fold Cross Validation.

4.4 Results on Ridge Regression

Figure 3 illustrates the relationship between ln λ and MSE. The integers at the top show the number of non-zero estimators for the model. The ridge regression only does the variables shrinkage but not the variable selection. The Ridge regression model was obtained based on an optimal value of $\lambda = 0.01$.

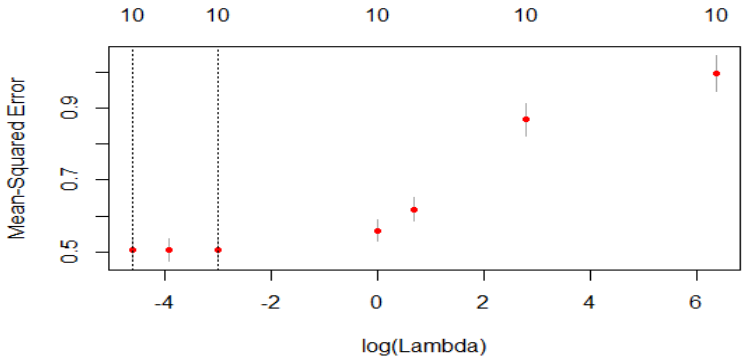

Fig. 3: MSE Plot and the Number of Variables in the Model as A Function of Log (λ) for the 10-Fold Cross Validation.

4.5 Results on SCAD Regression

Figure 4 illustrates the relationship between ln λ and MSE. The integers at the top show the number of non-zero estimators for the model. The left line gives the smallest MSE with seven variables in the model, and the right line gives the smallest Standard Error (SE) with only four variables in the model. The SCAD regression model was obtained based on an optimal value of $\lambda = 0.0167$.

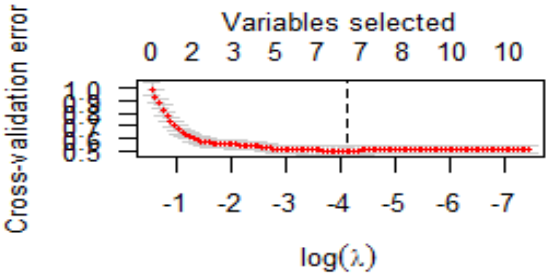

Fig. 4: MSE Plot and the Number of Variables in the Model as A Function of Log (λ) for the 10-Fold Cross Validation.

5. Main Results

MSE Comparison of OLS, LASSO, Elastic Net and Ridge and SCAD Regressions

Variable	OLS	LASSO	ELASTIC NET	RIDGE	SCAD
AGE	-0.0062	0.0000	0.0000	-0.00354	0.0000
SEX	-0.1481	-0.1211	-0.1342	-0.1426	-0.1380
BMI	0.3211	0.3225	0.3189	0.3200	0.3286
BP	0.2004	0.1830	0.1907	0.1964	0.1975
TC	-0.4893	-0.0630	-0.1010	-0.1537	-0.07432
LDL	0.2945	0.0000	0.0000	0.0292	0.0000
HDL	0.0624	-0.1379	-0.1078	-0.0826	-0.1472
TCH	0.1094	0.0000	0.0513	0.0730	0.0000
LTG	0.4641	0.3173	0.3151	0.3324	0.3318
GLU	0.0418	0.0333	0.0423	0.0454	0.02148
MSE	**0.5050**	**0.5040**	**0.5034**	**0.5046**	**0.5000**

5. Conclusion

In this paper, we used a data (with high possibility of multicollinearity) obtained from previous literature to assess the performance of four penalized regression methods, namely; LASSO, Ridge, Elastic net and SCAD. It shows that, all the penalized regression methods considered have lower mean square error (MSE), than the ordinary least squares regression. It also shows that, penalization techniques can make the predictive accuracy of a model better, by lowering the variability in the measures of a regression coefficients, which shrinks the estimates towards zero. This research is hope to provide assistance to researchers to ease their decision making as to which technique to be used when encountered with the problem of multicollinearity

References

[1] Bondell, H. D., & Reich, B. J. (2008). Simultaneous regression shrinkage, variable Selection, and supervised clustering of predictors with OSCAR. *Biometrics, 64*(1), 115-123.

[2] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R.(2004). Least Angle Regression. The *Annals of Statistics,*32(6), pp: 407-499.

[3] El Anbari, M., & Mkhadri, A. (2014). Penalized regression combining the l_1 norm and a correlation based penalty. *Sankhya B, 76*(1), pp: 82-102.

[4] Fan, J. and Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96 .1348-1360.

[5] Friedman, J., Hastie, T., Tibshirani, R., (2010)"Regularization paths for generalized linear models via coordinate descent", *Journal of Statistical Software*, 33, pp:1 – 22.

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

[7] Ryan, T. (2009). *Modern Regression Methods (Second Edition).* John Wiley & Sons. Hoboken, New Jersey, pp: 50-80.

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

[9] Tutz, G., & Ulbricht, J. (2009). Penalized regression with correlation-based penalty. *Statistics and Computing, 19*(3), pp:239-253.

[10] Zeng, L., & Xie, J. (2011). Group variable selection for data with dependent structures. *Journal of Statistical Computation and Simulation, 82*(1), pp:95-106.

[11] 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), pp:301-320.

[12] Zou, H. and Zhang, H. (2009). On the adaptive elastic-net with a diverging number of parameters. The *Annals of Statistics*, 12, pp:1149-1173.