ISSN: 2456-1452
Maths 2024; SP-9(5): 31-35

www.mathsjournal.com
Received: 23-06-2024
Accepted: 30-07-2024

Ume Kulsum
Ph.D. Scholar, Division of Agricultural Statistics, Sher-e-Kashmir University of Agricultural Sciences and Technology, Kashmir, India

Imran Khan
Associate Professor, Division of Agricultural Statistics, Sher-e-Kashmir University of Agricultural Sciences and Technology, Kashmir, India

Aqib Gul
Ph.D. Scholar, Division of Agricultural Statistics, Sher-e-Kashmir University of Agricultural Sciences and Technology, Kashmir, India

SA Mir
Professor, Division of Agricultural Statistics, Sher-e-Kashmir University of Agricultural Sciences and Technology, Kashmir, India

MS Pukhta
Professor, Division of Agricultural Statistics, Sher-e-Kashmir University of Agricultural Sciences and Technology, Kashmir, India

Arif Bashir
Ph.D. Scholar, Division of Agricultural Statistics Sher-e-Kashmir University of Agricultural Sciences and Technology Kashmir India

Ume Salma
MSc Plant Physiology, Division of Basic Sciences, Sher-e-Kashmir University of Agricultural Sciences and Technology Kashmir India

Corresponding Author:
Ume Kulsum
Ph.D. Scholar, Division of Agricultural Statistics, Sher-e-Kashmir University of Agricultural Sciences and Technology, Kashmir, India

Enhancing regression models: A comparative investigation of classical and ridge regression in the presence of multicollinearity in high density apple data

Ume Kulsum, Imran Khan, Aqib Gul, SA Mir, MS Pukhta, Arif Bashir and Ume Salma

DOI: https://doi.org/10.22271/maths.2024.v9.i5Sa.1804

Abstract
The application of statistical principles and methods is necessary for effective practice in resolving the different problems that arise in many branches of agricultural activity. To find out one such suitable estimation procedures for Gala species of high density apple, the present investigation has been carried out. The study revealed that ridge regression produces coefficients with minimum mean square error and produces regression coefficients which predict better than ordinary least squares when predictor variables are highly correlated. The value of ridge constant (θ) was estimated by three different methods, *viz*., ridge trace technique, method given by Hoerl, Kennard and Baldwin and cross-validation method. Based on the minimum value of mean square error, ridge trace method was selected to calculate the optimum value of ridge constant.

Keywords: Multiple linear regression, multicollinearity, regression estimates, ridge regression

Introduction
A ridge regression is an ordinary least squares estimation with a constraint on the sum of the squared coefficients (Hoerl & Kennard, 1970a) [7]. The ridge regression technique is extensively used to tackle the problem of multicollinearity between the X_i predictor variables (Hoerl & Kennard, 1970b; Marquardt, 1970) [8, 15]. It modifies the least-squares method of estimating the regression coefficients by adding a bias component (Hoerl, Kennard, & Baldwin, 1975) [9]. Serious multicollinearity often leads to wide confidence intervals for the regression coefficients that are too wide to be practical (Myers, 1990) [18]. Ridge regression technique estimates the biased β_i estimates purposely, but even so, will provide a much tighter confidence interval in which the true β_i value resides, though it is biased (Hoerl & Kennard, 1988) [11]. By using ridge regression procedure, the confidence interval for the biased estimator is tighter; as a result, the probability of biased estimate of regression coefficient (β_i biased), as it is closer to the actual population β_i parameter than the unbiased estimate of regression coefficient (β_i unbiased), is greater (Myers & Montgomery, 2002) [19]. Hence, the logic of ridge regression is simple: if a biased estimator can provide a more accurate estimate than can an unbiased one, yet still include the true β_i, it should be used (Hoerl & Kennard, 1976) [10]. Generally, the value of θ lies between 0 and 1 ($0 \leq \theta \leq 1$) (Marquardt, 1970) [15]. The population ridge regression equation, as proposed by Hoerl and Kennard in 1970 [7], is known as the ridge-regression method, and the estimator obtained by this method is known as the ridge estimator (Hoerl & Kennard, 1970b) [8]. The ridge regression procedure is intended to control "ill-conditioned" situations where near dependencies between various columns in X cause the X′X matrix to be close to singular, giving rise to unstable parameter estimates, typically with large standard errors (Lawless & Wang, 1976) [13]. Ridge regression procedure adds specific additional information to the problem to remove the ill-conditioning (Hoerl & Kennard, 1970a) [7].

Materials and Methods

Secondary data on the Gala RedLum variety of High-Density Plantation (HDP) apples was obtained from the Division of Agricultural Statistics, Faculty of Horticulture, Sher-e-Kashmir University of Agricultural Sciences and Technology, Kashmir. The regressors and regressand which were taken into account were: trunk cross sectional area (X_1), tree height (X_2), canopy area (X_3), canopy volume (X_4), number of primary branches (X_5), average fruit number (X_6), number of flower buds (X_7) and tree yield (Y) respectively.

Detection of multicollinearity

The multicollinearity in the data was inspected by examination of the correlation matrix for independent variables, Variance Inflation Factor (VIF), eigenvalues, and Condition Indices (CI) (Montgomery *et al*., 2012; Kutner, Nachtsheim, & Neter, 2004) [17, 12]. Multicollinearity leads to problems in regression analysis by inflating the variances of the parameter estimates (Mason & Perreault, 1991) [16]. VIF quantifies the severity of multicollinearity by measuring how much the variance of a regression coefficient is inflated due to collinearity with other predictors (O'Brien, 2007) [20]. The condition index was used to detect multicollinearity, as recommended by Belsley, Kuh, and Welsch (1980) [1], where values exceeding 30 are indicative of significant collinearity.

Methods for selecting Ridge Parameter (θ)

The ridge estimator for model $\hat{\beta}$= (X'X)$^{-1}$X'Y, is given by

$$\hat{\boldsymbol{\beta}}^{r} = (\boldsymbol{X'X} + \theta \boldsymbol{I})^{-1} \boldsymbol{X'Y}$$

where θ is the constant and is known as the ridge estimator or biasing parameter or tuning parameter or penalty parameter, I is the identity matrix (Hoerl & Kennard, 1970a) [7] in which the diagonal and the off-diagonal values are 1 and 0, respectively. The value of penalty parameter is considered known or 'given'. In practice, the value of θ is not known and the experimenter needs to make an informed decision on its value (Hoerl & Kennard, 1970b) [8]. Several methods are available in literature for selecting appropriate value of θ,*viz*:

1. **Ridge Trace:** The ridge trace procedure was first suggested by A.E. Hoerl in 1962 and is discussed extensively in two seminal papers by Hoerl and Kennard (1970a, b). According to Hoerl and Kennard (1970b) [8], a reasonable value of θ can be estimated by Ridge Trace inspection. The Ridge Trace is a plot of the elements of $\hat{\beta}_r$ versus θ , for values of θ between 0 and 1 (Obenchain, 1975) [21]. The objective is to inspect the trace (curve) and identify a small value of θ at which the ridge regression estimators stabilize (Hastie, Tibshirani, & Friedman, 2009) [3].
2. **Hoerl, Kennard and Baldwin:** Hoerl, Kennard, and Baldwin (1975) [9] suggested another method to select for K (in our case θ), where the selection of θ is based on the following criterion:

$$K = p\hat{\sigma}^2 / \hat{\beta}'\hat{\beta}$$

where, p = no. of parameters to be estimated and $\hat{\beta}$ = vector of least squares estimator (Hoerl, Kennard, & Baldwin, 1975; Vinod, 1978) [9, 26].

3. **Cross-validation:** Cross-validation is a widely-used method for selecting the penalty parameter in ridge regression to ensure good prediction performance (Stone, 1974; Golub, Heath, & Wahba, 1979) [24, 4]. This technique splits the dataset into two groups: a training set and a test set (Picard & Cook, 1984) [22]. The model is built on the training set and evaluated on the test set. This process is repeated across various splits of the data to minimize prediction error (Shao, 1993; Friedman, Hastie, & Tibshirani, 2010) [22, 3]. The optimal θ is selected based on its performance across these test sets, as described by Cule, Vineis, and De Iorio (2011) [2].

The whole ridge regression procedure was performed in R software (version R 4.1.0) using various packages, such as "lmridge" (Imdad *et al*., 2018) [25] and "glmnet" (Friedman *et al*., 2010) [3]. R's extensive functionality and open-source nature make it ideal for regression modeling, as suggested by Venables & Ripley (2002) [27].

Results and Discussion

Multicollinearity diagnosis has shown that the multicollinearity issue in this study is very severe.

The correlation matrix of seven regressors is presented in the Table 1. It may be observed from the correlation matrix that correlation between X_1 and X_3, X_4, X_5, X_6 and X_7 is significant and is as high as 0.90 indicating the presence of multicollinearity. The correlation coefficient between X_2 and X_4 is 0.333, which is the lowest correlation coefficient value among the regressors. Values greater than 0.80 are considered as highly correlated and indicate the presence of multicollinearity and the same can be seen by scatter plot given in Fig 1.

Table 1: Correlation matrix for independent variables

Variable	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	1	0.338	0.990	0.994	0.955	0.903	0.990
X_2	0.338	1	0.335	0.333	0.349	0.357	0.336
X_3	0.990	0.335	1	0.991	0.969	0.926	0.990
X_4	0.994	0.333	0.991	1	0.955	0.899	0.992
X_5	0.955	0.349	0.969	0.955	1	0.898	0.940
X_6	0.903	0.357	0.926	0.899	0.898	1	0.910
X_7	0.990	0.336	0.990	0.992	0.940	0.910	1

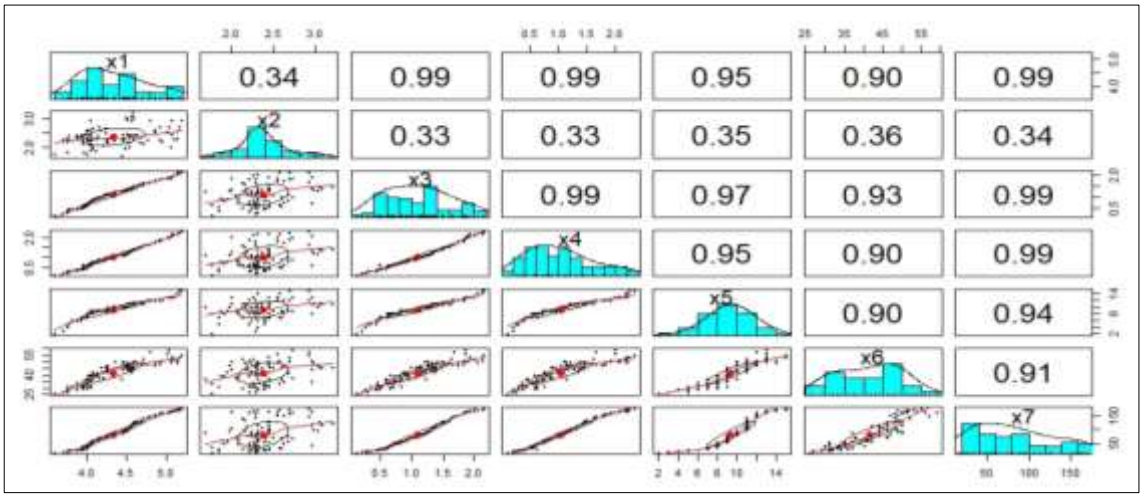

Fig 1: Scatter plot

Estimation of VIF's in Table 2 for independent variables show that X_1, X_3, X_4, X_5 and X_7 have VIF value greater than 10 and were responsible for causing multicollinearity.

Small eigenvalues shown in Table 2 with values 5.60×10^{-5}, 6.20×10^{-4}, etc is a sign of presence of multicollinearity. The values greater than 30 for CI are considered to be the indication for multicollinearity. The highest CI value of 5464.28 gives us the idea that the multicollinearity is severe. The problem of multicollinearity is tackled by ridge regression technique without removing any variable.

Table 2: Variance inflation factor (VIF), eigenvalues and condition indices (CI) for independent variables

Variables	VIF	Eigenvalues	CI
X_1	103.17	3.06×10^{-1}	1.00
X_2	1.33	1.23×10^{-2}	24.87
X_3	162.60	9.14×10^{-3}	33.48
X_4	60.90	5.17×10^{-3}	59.19
X_5	32.23	4.40×10^{-3}	69.54
X_6	7.76	6.20×10^{-4}	493.54
X_7	123.08	5.60×10^{-5}	5464.28

The package *"olsrr"* given by Aravind Hebbali (2017) [6] in R was used for multicollinearity diagnosis using following commands. Three different methods were used to determine the optimum value of θ; Ridge Trace method given by Hoerl and Kennard (1970) [7], method given by Hoerl, Kennard and Baldwin (1975) [9] and Cross-Validation method. All the three methods were run in R software using different packages, *viz. "lmridge"* and *"glmnet"*. One thousand admissible values of θ were used to produce the ridge trace. Ridge trace plot for the present investigation has been shown in Fig. 2. It can be seen from the plot how the coefficients begin to stabilize as θ increases from 0 up to 1. At $\theta = 0.0002$ the coefficients stabilize thus giving 0.0002 as the optimum value of θ. The values for θ obtained by other two methods, Hoerl, Kennard and Baldwin and cross-validation were 0.00033 and 0.00025, respectively.

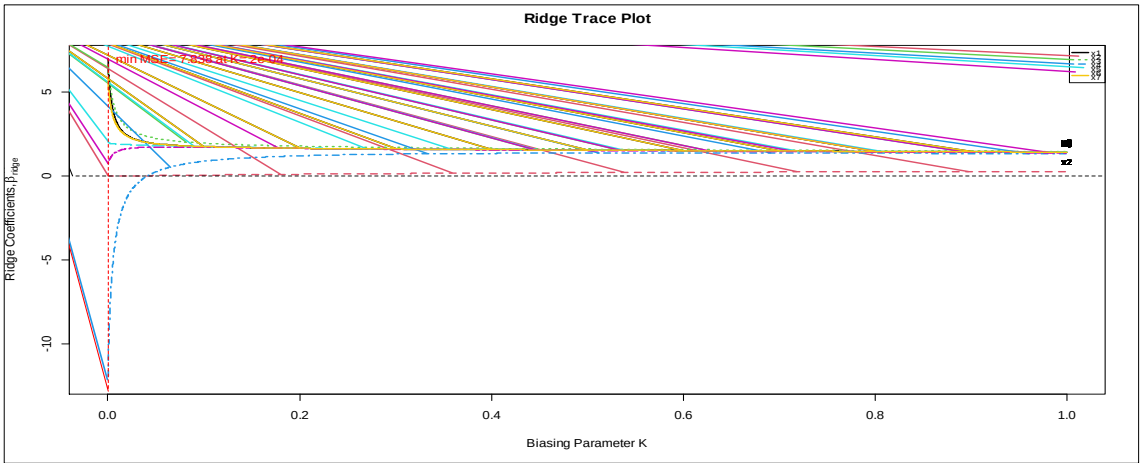

Fig 2: Ridge trace plot

Ridge estimates of regression coefficients for three different values of ridge constant (θ) were calculated. Estimates at value θ = 0, has given the least squares estimates. Table 3 presents the complete overview of the comparison made between the ridge estimates calculated by different methods.

As ridge regression technique leads to the shrinkage of the ordinary least squares estimates, there was considerable shrinkage in ridge estimates of $\hat{\beta}_1$, $\hat{\beta}_3$, $\hat{\beta}_5$ and $\hat{\beta}_7$ for all values of θ. The value of $\hat{\beta}_1$ decreased from 0.712 in OLS to 0.687 in ridge trace, method. Similarly, decrease in the value of regression estimates of $\hat{\beta}_3$ from 0.662 in OLS to 0.649 in ridge trace, method. Same pattern was followed by $\hat{\beta}_5$ and $\hat{\beta}_7$. Performance of ridge estimates of regression coefficients was found better on the basis of mean square error criterion, as for present investigation the mean square error of ridge estimates under ridge trace, Hoerl Kennard and Baldwin and cross-validation methods were 7.838, 0.081 and 7.860 respectively which were considerably less than mean square error value 8.099 obtained from ordinary least squares method.

Table 3: Regression coefficients, R^2, MSE, AIC and BIC at different values of θ

Ridge constant(θ)	OLS	Ridge Trace			Hoerl, Kennard and Baldwin			Cross-Validation		
	0.000	0.0002	0.00033	0.00025	0.0002	0.00033	0.00025	0.0002	0.00033	0.00025
$\hat{\beta}_1$	0.712	0.687	0.689	0.688	0.688	0.689	0.688	0.687	0.699	0.696
$\hat{\beta}_2$	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003	-0.003
$\hat{\beta}_3$	0.662	0.647	0.649	0.648	0.648	0.650	0.647	0.656	0.651	0.655
$\hat{\beta}_4$	-1.245	-1.191	-1.691	-1.227	-1.184	-1.159	-1.197	-1.195	-1.186	-1.218
$\hat{\beta}_5$	0.207	0.204	0.206	0.205	0.204	0.206	0.205	0.204	0.205	0.206
$\hat{\beta}_6$	0.069*	0.073*	0.073*	0.072*	0.074*	0.075*	0.073*	0.072*	0.073*	0.071*
$\hat{\beta}_7$	0.595	0.575	0.577	0.576	0.576	0.578	0.577	0.583	0.587	0.585
R^2	0.987	0.986	0.984	0.987	0.984	0.985	0.986	0.985	0.984	0.987
MSE	8.099	7.838	7.986	7.840	8.067	8.081	8.070	7.858	7.986	7.860
AIC	-412.521	-412.602	-412.193	-412.463	-412.602	-412.500	-412.563	-412.600	-412.478	-412.510
BIC	49.242	48.865	48.893	48.869	48.750	48.780	48.770	48.870	49.450	49.010

Significant codes: '*' 0.05

Package *"lmridge"* given by Imdad Ullah (2018) [25] in R was used for running ridge trace and Hoerl, Kennard and Baldwin method. The different commands used are.

The different packages used for cross-validation method in R include, "caret" given by Max Kuhn (2020), and "glmnet" given by Jerome Friedman (2009) [5].

Fig. 3 indicates how the VIF values had decreased after adopting the ridge regression technique. From the VIF trace plot it is observed how some variables with earlier values greater than 100 start to decrease as we adopt the ridge regression technique by adding some bias.

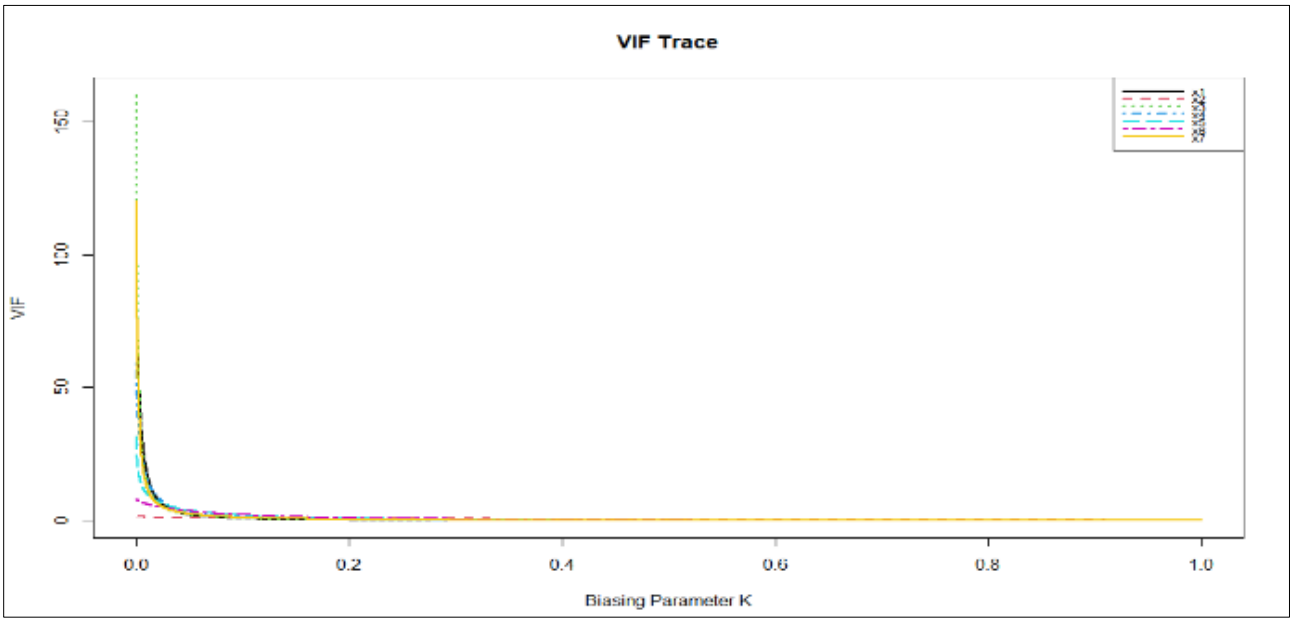

Fig 3: VIF trace plot

Summary and Conclusion

Three different methods were used to check the multicollinearity in the data, *viz.* examination of correlation matrix, VIF, CI and eigenvalue inspection. It was observed from the correlation matrix that explanatory variables have high and significant correlation among them with values as high as 0.99. Small eigenvalue of 5.60×10^{-5}, high VIF and CI value of 162.60 and 5464.28 respectively were also observed. These all measures of multicollinearity indicated that the problem of multicollinearity was serious for the present study. To find out the optimum value of ridge constant, three different methods were employed, *viz*. ridge trace method, method given by Hoerl, Kennard and Baldwin and cross-validation method. It was concluded that OLS estimator is dominated by all methods of in terms of MSE. Optimum value of ridge constant θ was found at 0.0002 by ridge trace method.

On the basis of performance of ridge estimators in presence of multicollinearity, it is concluded that there was severe multicollinearity in our data as a result the ordinary least square estimators, though estimable had a large variance and thus were no longer BLUE. It was seen that ridge regression is more than a last resort attempt to salvage least square linear regression in the case of multicollinearity in predictor variables. It is to be considered a major linear regression technique that proves its usefulness when collinearity is problematic. From the MSE point of view, it is not surprising that the use of traditional multiple linear regression suffers from multicollinearity problems and clearly shows that ridge regression performs best when input data are multicollineared.

References

1. Belsley DA, Kuh E, Welsch RE. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons; c1980.
2. Cule E, Vineis P, De Iorio M. Ridge regression in genetic epidemiology. Stat Appl Genet Mol Biol., 2011, 10(1).
3. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1-22.
4. Golub GH, Heath M, Wahba G. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics. 1979;21(2):215-223.
5. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer; c2009.
6. Hebbali A, Hebbali MA. Package 'olsrr'; c2017.
7. Hoerl AE, Kennard RW. Ridge Regression: Applications to Nonorthogonal Problems. Technometrics. 1970;12(1):69-82.
8. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. 1970;12(1):55-67.
9. Hoerl AE, Kennard RW, Baldwin KF. Ridge regression: Some simulations. Commun Stat. 1975;4(2):105-123.
10. Hoerl AE, Kennard RW. Ridge Regression Iterative Estimation of the Biasing Parameter. Commun Stat Theory Methods. 1976;5(1):77-88.
11. Hoerl AE, Kennard RW. Ridge Regression. In: Encyclopedia of Statistical Sciences. John Wiley & Sons; c1988. p. 129-34.
12. Kutner MH, Nachtsheim CJ, Neter J. Applied Linear Regression Models. McGraw-Hill; c2004.
13. Lawless JF, Wang P. A Simulation Study of Ridge and Other Regression Estimators. Commun Stat Theory Methods. 1976;5(4):307-323.
14. Lawrence A, Ridout M. Ridge regression for the linear mixed model. J Appl Stat. 2010;37(12):2149-2164.
15. Marquardt DW. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics. 1970;12(3):591-612.
16. Mason CH, Perreault WD Jr. Collinearity, power, and interpretation of multiple regression analysis. J Mark Res. 1991;28(3):268-280.
17. Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. John Wiley & Sons; c2012.
18. Myers RH. Classical and Modern Regression with Applications. PWS-Kent Publishing; c1990.
19. Myers RH, Montgomery DC. Response Surface Methodology: Process and Product Optimization Using Designed Experiments. John Wiley & Sons; c2002.
20. O'Brien RM. A caution regarding rules of thumb for variance inflation factors. Qual Quant. 2007;41(5):673-690.
21. Obenchain RL. Ridge Analysis Following a Preliminary Test of the Shrunken Hypothesis. Technometrics. 1975;17(4):431-441.
22. Picard RR, Cook RD. Cross-validation of regression models. J Am Stat Assoc. 1984;79(387):575-583.
23. Shao J. Linear model selection by cross-validation. J Am Stat Assoc. 1993;88(422):486-494.
24. Stone M. Cross-validatory choice and assessment of statistical predictions. J R Stat Soc Ser B Methodol. 1974;36(2):111-147.
25. Ullah MI, Aslam M, Altaf S. lmridge: A Comprehensive R Package for Ridge Regression. R J. 2018;10(2):326.
26. Vinod HD. A survey of ridge regression and related techniques for improvements over ordinary least squares. Rev Econ Stat. 1978;60(1):121-131.
27. Venables WN, Ripley BD. Modern Applied Statistics with S. Springer; c2002.