Evaluation Of Sailing Boat Performance Based on Ridge Regression and Mathematical Model Optimization

Zhang Li *

Beijing Polytechnic College Beijing, China

* Corresponding Author Email: zhli929@126.com

Abstract. The traditional linear regression model can analyze and determine the coefficient between variables, reflecting the interpretation degree of variables in the equation. However, if the model is too complex, the training data is too few or too extreme, overfitting will occur, and the prediction effect of multicollinearity data is not good. The regularized ridge regression model can solve this problem. In this paper, Ridge regression analysis is carried out on the sailing performance data. Firstly, the data are preprocessed, and then the root-mean-square error RMSE=10.68 is obtained through the ridge regression model evaluation. The model fitting effect is good, and the regression coefficient corresponding to the variable Fr is the largest, indicating that the Froude number is the most obvious to the performance of the yacht. The variable beam draft ratio of B/Dr Is the least important to the yacht's performance. The conclusion of this paper can predict the performance of yachts more accurately, and provide corresponding improvement suggestions for enterprises.

Keywords: Ridge regression, mathematical model; correlation coefficient, sailing performance.

1. Introduction

Regression analysis is a statistical analysis method based on observation data to establish the appropriate dependence between dependent variables and independent variables[1].Regression analysis can analyze the inherent law of data, and can be used for forecasting and control, etc. Through the analysis and transformation of practical problems, the corresponding mathematical model is established, and the actual measured data is substituted into the model for parameter solving. Finally, whether the model can fit the actual measurement data well is evaluated. The coefficient between variables is determined by regression analysis, which reflects the explanatory degree of variables in the equation. The so-called multiple regression refers to the regression model containing two or more independent variables x and one dependent variable Y. In practical problem solving, a certain problem is generally associated with multiple factors, so the influence of each factor on the problem should be considered comprehensively. Finally, the optimal combination of these independent variables can be used to predict or estimate dependent variables, which will be more in line with the reality. In order to better find the linear relationship between dependent and independent variables, it is necessary to establish a regression model, as shown below

Where is the dependent variable, is the independent variable, is the regression coefficient. However, this kind of method also has its shortcomings. In the process of using regression model, if the model is too complicated, the training data is too little or too extreme, and the sample features are too many, overfitting will occur, and its prediction effect on multicollinearity data is not good. The so-called overfitting is to make the hypothesis overly complicated in order to get a consistent hypothesis, Specifically, it means that the model has been trained, and indeed the rendering effect is ideal in the training set, but the opposite is true in the test set. Multicollinearity refers to the linear relationship between the sample observations of several explanatory variables or all explanatory variables in the linear regression model[2]. The multicollinearity problem in regression model will bring a series of consequences to the estimation of the model. If there is complete multicollinearity between explanatory variables, then the model parameters cannot be estimated and the variance of the parameter estimates will be infinite, which will completely invalidate the ordinary least square estimation of the regression model.

Based on this shortcoming of linear regression, ridge regression method will be adopted in this paper, which can effectively avoid overfitting in the process of regression, so as to make the model better fitting effect and better prediction effect [3].

This paper mainly studies the dynamic performance of sailing boats. By establishing Ridge regression model, six performance indexes such as Froude number and beam draft ratio are analyzed, which can predict sailing performance and provide a good reference for sailing manufacturing.

2. Data

The DELFT data set is used to predict the hydrodynamic performance and speed of sailing ships. The specific data is from the website "www.kaggle.com/datasets/heitornunes/ya

cht-hydrodynamics-dataset", a total of 7 column data, the six independent variables and one dependent variable. The prediction of residual drag during the initial design stage is of great value for evaluating the drag coefficient of a ship. The basic inputs include basic hull size and ship speed. The Delft dataset consists of 308 comprehensive experiments, which were carried out in the Delft Marine Fluid Mechanics Laboratory. These experiments included 22 different hull forms from the form of the mother hull.

This paper will predict the performance of the yacht based on these independent variables. The variation involves hull geometry and Froude number:

(1) Longitudinal position of buoyancy center [LC], three-dimensional. It is the position of the center of buoyancy for lifting the buoyancy of the salvage buoy, the buoyancy in the wreck, and other lifting forces used, with the head (tail) column as the starting point and calculated along the longitudinal direction.

(2) Three-dimensional prismatic coefficient [PC], a general term for various dimensionless coefficients representing the geometric shape, area or bulk slimness of the underwater part of the hull.

(3) depth to length displacement ratio [L/D], one dimension. Is the ratio of perpendiculars length to profile depth, which is mainly related to hull strength. A large ratio is bad for hull strength.

(4) Beam draft ratio [B/D], dimensionless. The ratio is related to stability, roll period, wave endurance, rapidity and other factors. Large ratio, large hull width, good stability, but roll period is small, poor seaworthiness, navigation resistance increase.

(5) Width ratio of length beam ratio [L/B], one dimension. Is the ratio of perpendiculars length to profile width, which is related to speediness and course stability. The larger the ratio, the slimmer the hull, the better its speediness and course stability, but the less flexible handling in harbour and vice versa.

(6) Froude number, one-dimensional. Froude number, named after British ship designer W.Froude, Froude number is a very important similarity criterion number for simulating liquid flow with free surface.

This group of data is dimensionless data, the units between the data are unified, and different indicators can be compared with each other, so there is no need to standardize the data.

3. Data exploration

Data exploration is the basis of subsequent data mining in order to further discover various information among data after data collection and data sorting. In fact, the process of data exploration is to analyze the nature of data and describe the form and correlation between data through different drawing forms. This allows for better prediction during subsequent data modeling and analysis.

In the process of data exploration, it is necessary to make clear the specific situation and characteristics of the sample data, which variables are included and the relationship between them, and whether the sample data can meet the modeling requirements.

3.1. Correlation coefficient diagram

In the part of data exploration, firstly, the correlation between variables is studied. Correlation shows some linear relationship between variables, so for these independent variables, they can be regarded as multicollinearity. Correlation coefficient graphs are often used to show collinearity (see Figure 1).

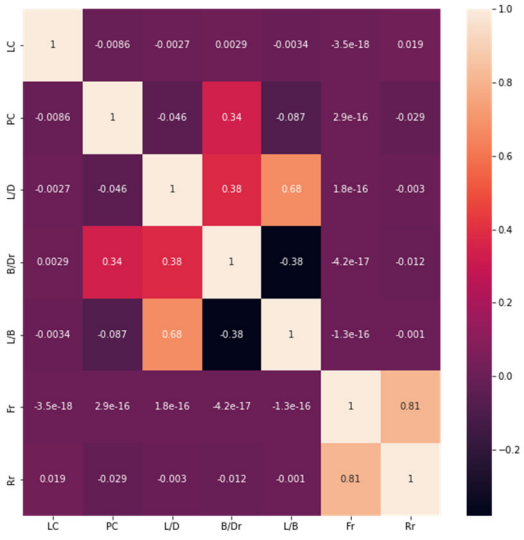


Fig. 1 Correlation coefficient diagram

Correlation coefficient graph is a statistical analysis method commonly used in empirical research. It is a data visualization technique, Linear relationships between two or more variables can be visually presented. Chart uses graphics and colors to transform quantitative data into different kinds of visual charts, so that users can analyze and discover their relationships more intuitively. Therefore, the correlation coefficient graph is an important part of statistical analysis, which can help us establish the relationship between variables and provide a basis for further empirical analysis. It can quickly capture the basic data between variables, such as average, median, degree of dispersion and degree of aggregation, and also find the correlation between variables, so as to point out the trend and law of data.

In addition, correlation coefficient diagrams can also be used to detect the presence of abnormal data. For example, the correlation between two variables may change suddenly, or a variable may produce abnormal observations, which usually indicate problems or possible errors in the data. Therefore, in research, correlation coefficient graph can effectively discover and solve data problems. In statistics, the most commonly used correlation coefficient plot is the scatter plot, and often to show the relationship between more variables. It allows users to understand the data more deeply, and then develop the corresponding strategy or plan. Correlation graphs can also be applied to various areas of data analysis, such as health, economy, society, etc. It can provide important basis for selection, model evaluation and decision making.

In conclusion, correlation coefficient graph is a practical data visualization tool, which can better understand the correlation among multiple variables and provide important support for data utilization. It is an indispensable analysis method in empirical research.

Specifically, it can be seen from the different color regions in Figure 1 that the degree of linear relationship between different variables represents different correlation coefficients. The light-colored region corresponds to a higher correlation coefficient, the purple region corresponds to a lower correlation coefficient, and the region close to black corresponds to a negative correlation coefficient. The horizontal and vertical axes correspond to 6 independent variables and 1 dependent variable respectively. The first 6 rows and 6 columns in the figure show the correlation between independent variables, while the last row and column show the correlation between independent variables and dependent variables.

For example, the correlation coefficient between variable L/D and variable L/B in the figure is 0.68, indicating strong collinearity between the two variables, while the correlation coefficient between variable B/Dr, variable PC and variable L/D is 0.34 and 0.38, indicating weak collinearity between them. Similarly, Fr and Rr are strongly correlated, while the rest are weakly correlated with Rr. Based on the multicollinearity between these variables, the ridge regression method adopted in this paper can better deal with these multicollinearity data sets.

3.2. Box Diagram

In order to describe the central position and distribution of a group or more groups of continuous data distribution, we often use box-like graphics to represent, called boxplot. For some distributions or data sets, more information is needed to describe the central tendency, understand its range of variation and dispersion, which can be better achieved by boxplot.

Different positions of the box represent different meanings and explain different problems. For example, a line in the middle of the box vividly represents the average level of sample data, also known as the median of data; The upper limit and lower limit of the box correspond to the upper and lower quartiles of the data respectively, that is, 50% of the data is in the box. The vertical length of the box gives a very vivid picture of how the data fluctuates. The maximum and minimum values in the data are distinguished by the lines above and below the box. Sometimes there are points that "pop up", which can be interpreted as "outliers". The so-called quartile is a group of data in accordance with the order from small to large, the group of data quartered number, called the quartile. It can be divided into the first quartile, the second quartile and the third quartile. According to the definition of quartile, the quartile corresponds to the 25%, 50% and 75% numbers in descending order. We know that there are two basic types of numbers in statistics, mode and average. The difference between the two can monitor distribution bias. If the average is greater than the mode, it is called positive bias (or right bias). Conversely, it is called negative skewness (or left skewness).

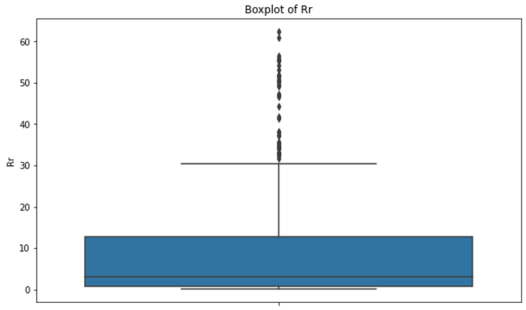


Fig. 2 Box diagram

In order to describe the central position and distribution of a group or more groups of continuous data distribution, we often use box-like graphics to represent, called boxplot. For some distributions or data sets, more information is needed to describe the central tendency, understand its range of variation and dispersion, which can be better achieved by boxplot.

Different positions of the box represent different meanings and explain different problems. For example, a line in the middle of the box vividly represents the average level of sample data, also known as the median of data; The upper limit and lower limit of the box correspond to the upper and lower quartiles of the data respectively, that is, 50% of the data is in the box. The vertical length of the box gives a very vivid picture of how the data fluctuates. The maximum and minimum values in the data are distinguished by the lines above and below the box. Sometimes there are points that "pop up", which can be interpreted as "outliers". The so-called quartile is a group of data in accordance with the order from small to large, the group of data quartered number, called the quartile. It can be divided into the first quartile, the second quartile and the third quartile. According to the definition of quartile, the quartile corresponds to the 25%, 50% and 75% numbers in descending order. We know that there are two basic types of numbers in statistics, mode and average. The difference

between the two can monitor distribution bias. If the average is greater than the mode, it is called positive bias (or right bias). Conversely, it is called negative skewness (or left skewness).

4. Method (Principle of Ridge Regression)

4.1. A.Underfitting and overfitting

In the process of fitting, we all hope to get the most appropriate fitting model[4-6]. In the process of data processing and analysis, if the assumptions are too idealized, the corresponding model will be too simple and cannot grasp the law of the data well. This situation is called underfitting. If you consider too carefully and set too many influence parameters, such a data model can predict the existing data well, but not the unknown data that may be generated in the future. This situation is called overfitting.(Figure 3)

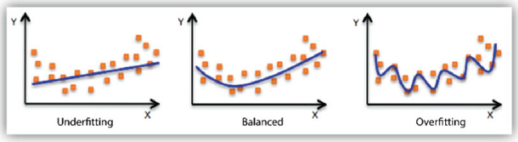


Fig. 3 Underfitting and overfitting

In order to get a more stable data state, we need to adjust the model to get the best performance on the training data. There are many methods to alleviate the overfitting phenomenon, of which there are three common ones. First, in terms of data, more training data can be obtained to train the model, because more samples will make the model learn more comprehensive and effective features, thus reducing the impact of noise. It can be said that this is the most effective way to alleviate the overfitting of the model. If it is difficult to obtain more training data, the existing training data can be reasonably expanded to increase the diversity of data. For example, image data can be expanded by means of image translation, rotation, scaling and cropping. In addition, early termination of training can also alleviate the overfitting phenomenon, because the initial weight of the network is generally small when initializing the network, but the longer the training time, the larger the weight of part of the network may be. If the training is stopped at an appropriate time, the ability of the network to fit data can be limited to a certain range. The second is to reduce the complexity of the model. When the data is small, the complex model is the main factor for overfitting. The neural network can reduce the number of network layers and neurons, and the decision tree can reduce the number of nodes by pruning, reducing the depth, and so on. In addition, ensemble learning can also be adopted to integrate multiple models together to reduce the risk of overfitting a single model. The third is to add regular term. When optimizing the objective function, adding regular term after the objective function is also a major way to alleviate the overfitting phenomenon. To alleviate the underfitting phenomenon, feature engineering, such as context feature and combination feature, can be used mainly from the data, because the factors that cause underfitting are mainly insufficient feature or weak correlation between feature and target. Increase the complexity of the model to increase the fitting ability of the model, such as increasing the number of neurons, network layers, high-order terms, etc. The regularization term can also be reduced, because the regularization term is to prevent overfitting, and when underfitting occurs, the regularization coefficient can be reduced.

4.2. B. Ridge regression principle

Ridge regression is an optimized least square method which can be used for biased estimation of collinear data[7-8]. Compared with the least square method, this method loses a small part of information and reduces the accuracy after fitting. The purpose is to make the regression coefficient more realistic and reliable, so that the abnormal data can be better fitted, which is also the key point that ridge regression is better than the least square method.

In the input/output description, the input independent variable X has at least one or more quantitative variables or binary variables, and the dependent variable Y is required to be a quantitative variable. The output is the result of model test goodness, For the error function, constraint conditions should be selected in accordance with the direction of gradient reduction, so as to make the calculated results more consistent with the information known in advance (for example, the general l-norm prior shows that it is a relatively simple initial problem, and after optimization and analysis, it is easier to get a solution with a smaller magnitude of parameter values, which can also be considered as a smooth solution of sparse parameters). For the discomfort of inverse problems, regularization also provides certain theoretical methods. For the results calculated by relying on data, not only exist, but also are unique. In this way, the influence of other noises will be reduced, so that overfitting is not easy to occur. The number of non-zero parameters of the model parameters is called the L0 regularization value, the sum of absolute values of each parameter is called the L1 regularization, and the square root value of the sum of the squares of each parameter is called the L2 regularization.

Ridge regression mainly solves the problem that there is a high degree of multicollinearity between data sets and when the number of predicted independent variables is greater than or equal to the number of variables to be observed, because such cases will lead to the existence of no solution to the coefficient of the regression model.

The general (matrix) form of regression analysis is as follows:

$$y = \sum_{j=1}^{p} \beta_j x_j + \beta_0 \tag{1}$$

The goal of using the least square method to solve the above regression problem is to minimize the following formula:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i)^2 \tag{2}$$

Ridge regression is to add a penalty term to the above minimization goal:

$$\hat{\beta}^{bridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i)^2 + \lambda \sum_{j=1}^{p} \beta_j{}^2 \right\} \tag{3}$$

Where $\beta_0 + \sum_{j=1}^{p} \beta_j x_i$ is the estimated value, y_i is the actual value, the actual value minus the sum of the squares of the deviation of the estimated value is an objective function that linear regression needs to train, then by minimizing this objective function, we can get the estimated value of β, followed by adding $\lambda \sum_{j=1}^{p} \beta_j{}^2$, it becomes the formula for calculating the coefficient of ridge regression, it is a regular term, and is L2 regular term.

In practical applications, the most common types of regularization are L1 and L2. In order to constantly update the general cost function, the so-called cost function refers to loss plus regular term, adding a regular term to the cost function will reduce the value of the weight matrix. In order to make the model simpler, it can be assumed that the neural network has a smaller weight matrix. This can greatly reduce the occurrence of overfitting phenomenon.

However, the regular term in L1 and L2 is different. In L2, what is added after linear regression is $\lambda \sum_{j=1}^{p} \beta_j{}^2$ that λ is the regularization parameter, it is a hyperparameter, optimizing its value will make the model perform better, L2 regularization is also called weight decay, because it forces the weight to decay to 0 (but never really equal to 0). And for L1, what we're adding after linear regression is

$\lambda\sum_{j=1}^{p}|\beta_j|$, here we're penalizing the absolute value of the weight. Unlike L2, the weight here may be reduced to 0. So, if we want to compress the model, this is very useful. In other cases, L2 regularization is usually preferred. In this paper, L2 regularization is used in the calculation of Zhongling regression model, which makes the coefficients of those insignificant variables approach 0, thus reducing the influence of such variables on dependent variable y, while L1 regularization makes the coefficients of those insignificant variables equal to 0, which becomes a feature screening process.

5. Results

5.1. Cross validation method

This paper divides the training set and the test set through the method of cross-validation. 75% of the samples are selected as the training set and 25% as the test set. If we do not divide the data and use all the data for modeling, we cannot evaluate the effect of the model. Therefore, we need to select a part of samples as the test set to evaluate the quality of the model. The so-called cross validation is mainly used in regression modeling.The so-called cross validation is mainly used in the modeling of regression analysis. For the known modeling data samples, the data is divided into two parts: one part has a large amount of data, the other part has a small amount of data, and the data samples with a large amount of data are modeled. The remaining samples are tested and predicted by the established model, and the prediction error value is calculated, and their sum of squares is further calculated. The process is followed until all data samples have been predicted, one and only one time. The basic idea of this method is to divide the sample data into training set and verification set, and use the training set to train the classifier. After the model is obtained, the verification set is used to test the model, and finally the performance indicators of the classifier are evaluated.

The purpose of cross-validation is to get a reliable and stable model. It can effectively evaluate the quality of the model, effectively select the model with the best performance on the data set, and effectively avoid overfitting and underfitting. The so-called underfitting means that the model cannot obtain the main information of the data set, and the representation on the training set and the test set is not good. Overfitting means that the model not only obtains the information of the data set, but also extracts the information of the noise data, so that the model has a very good performance in the training set but a very poor performance in the test set.

5.2. Parameter Adjustment

When adjusting the model parameters, we use the horizontal axis to represent the parameter lambda in the model, whose value increases gradually from to, to; The vertical axis represents the regression coefficient in the model. With the increasing of parameters, the regression coefficient of the top curve decreases sharply, while the regression coefficient corresponding to several other variables does not change significantly. Figure 3 can show how the regression coefficient changes with the change of parameters (Figure 4). Finally, the parameter alpha=10 of the Zhongling regression model is obtained.

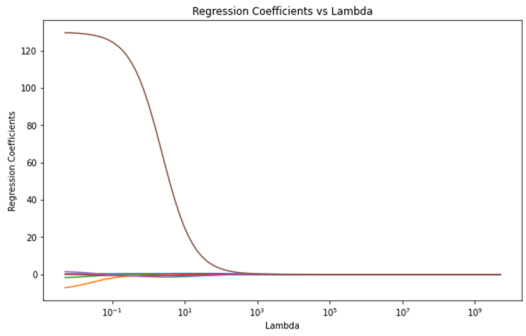


Fig. 4 Illustration of the parameter setting method.

5.3. Model Evaluation

In this paper, the Root Mean Square ErrOr (RMSE) method was used to evaluate the model. Root means square error (RMSE) is a commonly used index to measure the degree of fitting of quantitative prediction models in statistics [9]. It is formed from the square root of the average error, representing the difference between each sample point and its predicted value. It can measure the degree of fitting and accuracy of the data prediction model, that is, measure the degree of deviation between the predicted value of the model and the real value.

Root means square error can be used to represent the degree of discretization of a set of data, which reflects the degree of discretization of data [10]. The smaller it is, the more accurate it is, and can be used to measure the fit degree of the model. It can be used to measure the degree of fit of nonlinear regression models and can also be used to measure the degree of fit of time series models.

The formula for calculating the root mean square error is.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} \left\| y(i) - \widehat{y(i)} \right\|^2}{N}} \tag{4}$$

Where, $y(i)$is the actual value, $\widehat{y(i)}$is the predicted value, and N is the number of samples. In the calculation formula of root mean square error, the difference between the actual value and the predicted value of each sample is squared first to reduce the impact of positive and negative values, and the absolute value of the error can be converted to the square value, so as to better reflect the degree of fit of the model.

Root mean square error is a statistical index, which can help us to evaluate the degree of fit of the model, and can be used to measure the merits and demerits of the product prediction model, so as to provide us with a rational and scientific evaluation system.In the model established in this paper, RMSE=10.68.

5.4. Model Interpretation

Tab. 1 The coefficients of ridge model

Variable	LC	PC	L/D	B/Dr	L/B	Fr
Coefficient	β_1	β_2	β_3	β_4	β_5	β_6
Value	0.048	-1.1	-0.2	-0.004	-0.24	11.8

According to the above table, 6 independent variables and their regression coefficients including longitudinal position of buoyancy center, prism coefficient, length displacement ratio, beam draft ratio, long beam ratio and Froude number can be known. It can be seen that the coefficient of the last variable, namely the regression coefficient of Fr, is the largest, which is 11.8, indicating that this variable has the strongest influence on yacht performance. The regression coefficient of the fourth variable is the smallest, indicating that the variable has the weakest influence on the performance of the yacht. It is obvious that the first variable and the sixth variable are positively correlated with the

dependent variable y, while the other four variables are negatively correlated with the dependent variable y.

6. Conclusion

The regression coefficient corresponding to the variable Fr is the largest, indicating that the Froude number is the most obvious to the yacht's performance, while the variable beam draft ratio of B/Dr Is the least important to the yacht's performance. If the larger the residual drag is, the better, then in the actual application scenario, it is recommended to increase the values of x1 and x6 variables, so that the value of dependent variable y will also increase, which will be helpful to improve the performance of the yacht. Similarly, if the value of other four negatively correlated variables is reduced, the performance of the yacht will also be improved.

Ridge regression model is established in this paper. Compared with linear regression model, this model is more effective, has better model fitting and prediction, and can predict the performance of yachts more accurately. Then in the actual industrial scenario, if the shipbuilding company uses the ridge regression model established in this paper, it can effectively evaluate the performance of the yacht. At the same time, this paper finds out that the Froude number and beam draft ratio have a great impact on the performance of the yacht. If the enterprise improves the performance of these two indicators, it will play a great role in the improvement of the hull quality, so as to provide corresponding improvement suggestions for the enterprise.

References

[1] Arkes, Jeremy. Regression analysis: a practical introduction. Taylor & Francis, 2023.

[2] Shrestha, Noora. "Detecting multicollinearity in regression analysis." American Journal of Applied Mathematics and Statistics 8.2 (2020): 39-42.

[3] Ying, Xue. "An overview of overfitting and its solutions." Journal of physics: Conference series. Vol. 1168. IOP Publishing, 2019.

[4] Hawkins, Douglas M. "The problem of overfitting." Journal of chemical information and computer sciences 44.1 (2004): 1-12.

[5] Bartlett, Peter L., et al. "Benign overfitting in linear regression." Proceedings of the National Academy of Sciences 117.48 (2020): 30063-30070.

[6] Tsigler, Alexander, and Peter L. Bartlett. "Benign overfitting in ridge regression." arXiv preprint arXiv:2009.14286 (2020).

[7] Hastie, Trevor. "Ridge regularization: An essential concept in data science." Technometrics 62.4 (2020): 426-433.

[8] Arashi, Mohammad, AK Md Ehsanes Saleh, and BM Golam Kibria. Theory of ridge regression estimation with applications. John Wiley & Sons, 2019.

[9] Karunasingha, Dulakshi Santhusitha Kumari. "Root mean square error or mean absolute error? Use their ratio as well." Information Sciences 585 (2022): 609-629.

[10] Brassington, Gary. "Mean absolute error and root mean square error: which is the better metric for assessing model performance." EGU General Assembly Conference Abstracts. 2017.