IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 8, No. 4, August, 2021, pp. 547-556
Journal homepage: http://iieta.org/journals/mmep

Accurate and Hybrid Regularization - Robust Regression Model in Handling Multicollinearity and Outlier Using 8SC for Big Data

Mukhtar[1,2*], Majid Khan Bin Majahar Ali[1], Anam Javaid[1], Mohd. Tahir Ismail[1], Ahmad Fudholi[3,4]

[1] School of Mathematical Sciences, Universiti Sains Malaysia, Penang 11800, Malaysia
[2] CEFORY (Local Food Innovation) Universitas Sultan Ageng Tirtayasa, Serang, Banten 42124, Indonesia
[3] Solar Energy Research Institute, National University of Malaysia (UKM), Bangi 43600, Selangor, Malaysia
[4] Research Centre for Electrical Power and Mechatronics, Indonesian Institute of Sciences (LIPI), Bandung 40135, Indonesia

Corresponding Author Email: mukhtar@untirta.ac.id

https://doi.org/10.18280/mmep.080407

Received: 23 March 2021
Accepted: 11 June 2021

Keywords:
variable selection, regularization regression, robust regression, model selection, 8 selection criteria

ABSTRACT

Regressions have been continuously received great attention. However, there are still open issues in regression, and two of the issues is regression with multicollinearity and outlier. Regularization (Ridge, Lasso, and Elastic Net) techniques implement a means to control regression coefficients. The methods can decrease the variance and reduce our sample error for tackle multicollinearity. In robust regression, it is a form of regression method designed to overcome outliers. Robust regression is an important method for analyzing data that are infected with outliers. The data have been interacted on the second order interaction. The data contained 435 different independent interaction variables. The primary focus of this paper is to analyze and compare the impact of three different variable selection techniques regularization regression algorithms for the data seaweed drying. After that, it will be analyzed through robust regression (Tukey Bi-Square, Hampel, and Huber). As the result, the Lasso-Hampel was better than others with the MAE (4.09641), RMSE (5.275992), MAPE (7.9962), SSE (182491.2), R-square (0.6514791), and R-square Adjusted (0.649279). The method of Lasso-Hampel is able to be relied on investigation of the accuracy in big data obtained from regularization and robust regression.

1. INTRODUCTION

Regression methods are algorithms of supervised learning, which are important both Machine Learning and Statistics Learning. The regression methods have been known for a long time because they are many new developments. The regression methods are extending these algorithms significantly [1].

The regression methods are frequently used to calculate an algorithm to forecast future responses. They are aims to investigate relationship between dependent variable (Y) and the independent variables (X) [2].

The regression analyses are often applied most sciences. The regression methods are ones of the main tasks in Machine Learning and Statistics Learning. The regression methods have been successfully applied in many fields such as agriculture and biology for this case using data seaweed drying.

Seaweed should get attention from the Malaysia Government because it has many advantages including lots of nutrients and short growth of only 45 days per cycle. The seaweed is widely cultivated in Sabah because of the environmental and geographical factors which support it. Sabah is very favorable compared to the Malaysian peninsula [3]. The seaweed as an agricultural sector plays an important role in providing a source of food and protein in Malaysia [4].

The abundant supply of seaweed in Malaysia offers promising opportunities to produce and extract such as fucoidan, alginate, agar, and carrageenan. The seaweed is used in various ingredients such as in foods, pharmaceuticals, nutraceuticals, medicals, and other industries.

Seaweed contains beneficial bioactive compounds such as carrageenan powder, agar, or alginate. Seaweed is of great commercial importance as a stabilizer, thickener, gelling agent, and emulsifier. The Malaysia Agro-Policy has developed seaweed as high-value and valuable commodity that makes seaweed an important industry. Malaysia has great potential to become a significant seaweed supplier in the country, provided Malaysia has fully developed and utilized existing resources [5].

Assessment and comparison of the performance of the available methods are thus important to select the best method with the seaweed drying data and determine when their performance is optimal. Here, we evaluate the relative performance regularization regressions (Ridge, Lasso, and Elastic Net) for selecting variables (to choose the most significant variable from their perspective) and will be analyzed with robust regression (Bi-Square, Hampel, and Huber) models.

The methods comprise Ridge, Lasso, and Elastic net regression [6-17].

Regularization regressions (Ridge, Lasso, and Elastic Net) are applied as a variable selection to select the most significant variables with their perspective. They provide methods for controlling the regression coefficient, which is able to decrease the variance and decrease the sample error to solve the

multicollinearity issue. They are applied in various fields of scientific disciplines [18].

Improvement in both Statistics Learning and Machine Learning method – driven by big data in various disciplines scientific – offers opportunities and challenges for agriculture data analysis (especially the seaweed drying data). Today, in the era of big data, variable selections are a fundamental task in the area of both Statistics Learning and Machine Learning.

In general, the process of variable selection aims to select which are important variables. For example, in regression, it is very useful to select and maintain variables with predictable capabilities.

The aims of variable selection usually are:

(i) To improve predictive model capabilities;

(ii) To avoid multicollinearity problems;

(iii) To provide a more comprehensive understanding of the prediction model by reducing ineffective and unnecessary variables [19].

Both Statistics Learning and Machine Learning aim to build a model that presents the best dataset, these methods involve the task of variable selections. In this paper, a dataset containing 1924 observations will use to study the effect of more 29 different independent variables on the one dependent variable. Then the data will be interacted with in the second interaction. The data contain the effect of 435 different interaction independent variables on the one dependent variable. The more detailed tables for each variable interaction are attached in the Appendix A.

In recent years, agricultural data has increased exponentially with the adoption of automated data collection tools and systems. Data generated from agricultural precision tools has been one of the most significant contributions to this improvement. Due to the fast growth of data, regularization regression (Ridge, Lasso, and Elastic Net) will help to find useful and meaningful in big data, especially in agriculture [20].

In this study, it is to analyse seaweed data with several variables including hourly solar radiation, temperature, humidity, and moisture content.

Big data technology in agriculture has increased adoption rates in precision agriculture and is expected to become more prevalent in the coming years. It is used in the precision agricultural in several aspects of crop production, such as accuracy, agriculture (weather forecasting, yield monitoring, soil conditioning), decision-making tool and in enhancing zones of food security. The big data repositories essential knowledge which can be applied to the scientific data, or to give knowledge on interdisciplinary decisions such as economics, politics, or often recently, 'artificial intelligence of farming' to enhance food security and potency of agriculture [21].

Regressions continue to get significant appreciation and attention. However, in regressions have still open problems such as multicollinearity and outlier.

The first issue in regression is multicollinearity. Multicollinearity is two or more independent variables with high correlation. It is a common problem which is often encountered in regression methods. It will reduce the accuracy of parameter evaluation in the regression methods [22].

Regularization regressions are applied as a variable selection to select the most significant variables with their perspective. They provide methods for controlling the regression coefficient, which is able to decrease the variance and decrease the sample error to solve the multicollinearity issue. They are applied in various fields of scientific disciplines [18]. So, regularization regression is a regression analysis designed to handle multicollinearity. In this paper, we will use three types of regularization regressions such as Lasso, Ridge, and Elastic Net.

The methods comprise Ridge, Lasso, and Elastic net regression [6-17]. So, an important property of regularization regressions is respect to multicollinearity in the database (big data).

The second issue in regression is outliers. Outliers are suspicious because they are much larger or much smaller than most of the observations [23, 24]. Outliers are objections that differ significantly from the remaining data. The outliers are also referred to as anomalies, abnormalities, and discordances [25]. The outliers are common in big data and can create severe regression problems. They can lead to model misspecification, inaccurate analysis results and make all evaluation methods meaningless.

So, an important property of robust regressions is method with respect to outliers in big data. Robust regressions are required where the estimated values are not much influenced by much smaller or much larger observations. So, robust regression is a regression method which is designed to address outliers.

Robust regression is an important method for analyzing data which are contaminated outliers [24, 26, 27]. Because ordinary least square (OLS) can be very sensitive to outliers. Robust regressions are applied to detect outliers and provide results that are resistant to the presence of outliers. In this paper, we will use three types of robust regression M-Estimation such as Bi-Square, Hampel, and Huber.

The methods comprise Tukey Bi-Square, Hampel, and Huber regression [28-32].

To assess models, we need a model selection. Model selection was also made by different researchers, Abdullah et al. [33] used eight selection criteria (8SC) to obtain the best model among all possible models. Similarly, Javaid et al. used in model selection problem [15, 34-36].

Several authors have reviewed 8 Selection Criteria (8SC), but our study is different from their paper. Javaid et al. have made a study of 8 Selection Criteria. They only conducted research on small data. They did not present visualization in comparing models [15, 36-38].

The primary focus of this paper is to analyze and compare the impact of three different variable selection techniques regression regularization algorithms (Lasso, Elastic Net, and Ridge) for the data seaweed drying. After that, it will be analyzed through robust regression (Tukey Bi-Square, Hampel, and Huber) and to compare the impact of three different regression algorithms for forecast the efficient model, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE)., comparing three regularization and robust regression algorithms – in terms of the best model eight selection criteria (8SC).

2. MATERIALS AND METHODS

2.1 Regularization regression

2.1.1 Lasso

Linear regression equations $\{(x_i, y_i)\}_{i=1}^{N}$ with N samples and independent variables are $p - dimensional$ and $y_i \in \mathbb{R}$

is dependent variable. The aim is to forecast the dependent variable from the independent variables. Forecast and find independent variables significant play an essential role in regression [39]. The equation assumes:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \quad (1)$$

β_0 and β_i are unknown parameters and ε_i is a residual term for $= 1, \dots, p$. The Eq. 2 is a requirement to constrain. For *Lasso* regression or $\ell_1 - regularize\ regression$,

$$\min_{\beta_0,\beta} \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \quad \text{subject to } \|\beta\|_1 \le t \quad (2)$$

2.1.2 Ridge

The ridge constrain is $\sum_{j=1}^{p} \beta_j^2 \le t$ for a positive value t. For *ridge* regression or $\ell_2 - regularize\ regression$ [39].

$$\min_{\beta_0,\beta} \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \quad \text{subject to } \sum_{J=1}^{P} \beta_J^2 \le t \quad (3)$$

2.1.3 Elastic net regression

Lasso and ridge regression could be stated with L_q. Both q=1 and q=2 are corresponding to lasso and ridge respectively. Eq. (4) can be solved by calculating of type L_q.

$$arg\min_{\beta}\{\boldsymbol{Y^T Y} - 2\beta^{\boldsymbol{T}}\boldsymbol{X^T Y} + \beta^{\boldsymbol{T}}\boldsymbol{X^T X}\beta\} \quad \text{subject to } \sum_{j=1}^{k}\left|\beta_j\right|^q \le t \quad (4)$$

Researchers recommend taking $1 < q < 2$, to choose a compromise between lasso and ridge [40]. The elastic net regression evolves combining between Lasso and Ridge [41]. The elastic net formulation was defined by Zou and Hastie [16] as:

$$\sum_{j=1}^{k}\left((1-\alpha)\beta_j^2 + \alpha\left|\beta_j\right|\right) \le d^2, \alpha \in [0,1] \quad (5)$$

The elastic net is then used as a penalizing term to obtain the elastic net estimate:

$$\hat{\beta}_{Elastic\ net} = arg\min_{\beta}\left\{\boldsymbol{Y^T Y} - 2\beta^{\boldsymbol{T}}\boldsymbol{X^T Y} + \beta^{\boldsymbol{T}}\boldsymbol{X^T X}\beta + \lambda\sum_{j=1}^{k}\left((1-\alpha)\beta_j^2 + \alpha\left|\beta_j\right|\right)\right\} \quad (6)$$

From the Eq. (7), selecting parameter q is not necessary. We require to select an α value between $0 < \alpha < 1$. Ridge and Lasso regression could be stated with α. Both $\alpha = 0$ and $\alpha = 1$ are corresponding to ridge and lasso respectively. The elastic net regression evolves combining between Lasso and Ridge.

The elastic net is a method of regularization regression that provides between ridge and lasso [42]. The advantage of the elastic net is achieving stability concerning random sampling [43].

2.2 Robust regression

The M-estimation is general method in robust regression. The M in M-estimation is "Maximum likelihood". The aim of M-estimation is minimizing error (residual) [44].

The first function regression method, suppose we have a data set of size n such that:

$$\begin{gathered} y_i = x_i^T\beta + e_i \\ e_i = y_i - \hat{y}_i = y_i - \boldsymbol{x_i^T}\beta \\ e_i(\beta) = y_i - x_i^T\beta \end{gathered} \quad (7)$$

M-estimator attempt to minimize the sum of a chosen function $\rho(\cdot)$ which is acting on the residual. Formally defined, M-estimators are given by:

$$\hat{\beta}_M = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(e_i(\beta)\right) \quad (8)$$

The above form with ρ function is the $\rho -$ type M-estimation. Suppose σ is known and let the residuals for some estimate β be $e_i = y_i - \beta^T x_i$ [45]. Then the regression M-estimate of β is the value that minimizes the objective function:

$$\sum_{i=1}^{n} \rho\left\{\frac{e_i(\beta)}{\sigma}\right\} \quad (9)$$

The σ should be estimated robustly. M-estimator of scale $\tilde{\sigma}_M$ is found by solution of the equation:

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{e_i}{\sigma}\right) = \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{y_i - \beta^T x_i}{\sigma}\right) = k \quad (10)$$

When β is the $p \times 1$ parameter vector, then $\psi -$ type function could be yielding as:

$$\sum_i \psi(e_i)\frac{\partial e_i}{\partial \beta_i}, \text{ for } j = 1,2, \dots, p \quad (11)$$

where the derivative function $\psi(e) = \frac{\partial \rho(e)}{\partial(e)}$ is the influence function. Then the weight function could be defined as below:

$$w(e) = \frac{\psi(e)}{e} \quad (12)$$

The $\psi(e)$-type function becomes:

$$\sum_i w(e_i)e_i\frac{\partial e_i}{\partial \beta_i} = 0, \text{ for } j = 1,2, \dots, p \quad (13)$$

And the object becomes to obtain the following iterated re-weighted least square problem:

$$min\sum_i w\left(e_i^{(k-1)}\right)e_i^2 \quad (14)$$

where, k indicates the iterate number [46].

Further, the M robust regression was applied to address the outliers through M–bi square, M–Hampel, and M–Huber. For more detail, we applied in Table 1-Formulas for Robust Regression M-estimation.

2.3 Validation models

The metric evaluations are needed to evaluate the appropriateness of a model. They become very important to analyze whether the model is adequate. The metrics including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Sum of Square Error (SSE), and R-squared.

Table 1. Formulas for Robust Regression M-estimation

Methods	Objective Function	Weight Function
Bi-Square	$\rho_B = \begin{cases} \frac{k^2}{6}\left\{1-\left[1-\left(\frac{e}{k}\right)^2\right]^3\right\} \; for\; \lvert e\rvert \le k \\ \frac{k^2}{6} \; for\; \lvert e\rvert > k \end{cases}$	$w_B = \begin{cases} \left[1-\left(\frac{e}{k}\right)^2\right]^2 \; for\; \lvert e\rvert \le k \\ 0 \; for \lvert e\rvert > k \end{cases}$
Huber	$\rho_{Hu} = \begin{cases} \frac{1}{2}e^2 \; for\; \lvert e\rvert \le k \\ k\lvert e\rvert - \frac{1}{2}k^2 \; or\; \lvert e\rvert > k \end{cases}$	$w_{Hu} = \begin{cases} 1 \; for\; \lvert e\rvert \le k \\ \frac{k}{\lvert e\rvert} \; for\; \lvert e\rvert < k \end{cases}$
Hampel	$\rho_{Ha} = \begin{cases} \frac{e^2}{2}, \quad 0 < \lvert e\rvert < a \\ a\lvert e\rvert - \frac{e^2}{2}, \quad b < \lvert e\rvert \le c \\ \frac{-a}{2(c-b)}(c-e)^2 + \frac{a}{2}(b+c-a), b < \lvert e\rvert \le c \end{cases}$	

Table 2. Formulas for validation methods

Validation	Formulation	Reference
Mean Absolute Error (MAE)	$MAE = \sum_{i=1}^{n}\left\lvert\frac{Y-\hat{Y}_i}{\hat{Y}_i}\right\rvert$	[47]
Mean Square Error (MSE)	$MSE = \sum_{i=1}^{n}\left(\frac{Y-\hat{Y}_i}{\hat{Y}_i}\right)^2$	[48]
Mean Absolute Percentage Error (MAPE)	$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left\lvert\frac{Y-\hat{Y}_i}{\hat{Y}_i}\right\rvert$	[49]
Sum of Square Error (SSE)	$SSE = \sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$	[50]
Sum of Squared Total (SST)	$SST = \sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2$	[50]
R-squared	$R^2 = \frac{SSR}{SST} = \frac{SST-SSE}{SST} = 1-\frac{SSE}{SST}$	[50]

The metric evaluations are used to measure the accuracy of the regression model in forecasting the dependent variable within the acceptable rage of accuracies. The formula of the metrics is shown in Table 2.

2.4 Selection models

2.4.1 Phase 1 – all possible model

$$N = \sum_{j=1}^{k} j\left(C_j^k\right)$$

2nd Order

where, N is the number of all Possible models, k is total number of independent variables and $j = 1,2,\dots,k$.

A dataset containing 1924 observations will use to study the effect of 29 different independent variables on the one dependent variable. Then the data will be interacted with in the second interaction. The data contain the effect of 435 different interaction independent variables on the one dependent variable. For more detail, the second order inteactions as depicted in Appendix A.

2.4.2 Phase 2 – selected models

In this paper, we will analyze regression model amongst Lasso, Ridge, and Elastic Net. From these regression model, we will take variable importance. We will take subset of top highest 30 influential variables from each technique and will apply three robust regression algorithms (Bi-Square Tukey, Hampel, and Huber).

2.4.3 Phase 3 – the best model

Regularization such as Lasso, Ridge, and Elastic Net, serve as variable important to take top highest 30 influential variables. After the subset of the influential variable will be followed by robust regression Bi-Square, Hampel, and Hubber to determine a regression model.

The next step was to get the best model after a list of selected models was obtained. 8SC were defined for this purpose by Zainodin et al. [34]. 8SC formula can be displayed as shown in Table 3. By using mentioned formulas in Table 3, Akaike information criterion (AIC), RICE, Final Prediction Error (FPE), SCHWARZ (SBC), Generalized Cross Validation (GCV), Sigma square (SGMASQ), SHIBATA, and Hannan-Quinn (HQ) information on the basis of minimum value obtained from all mentioned criteria.

Table 3. Formula used for 8SC

No	Methods	Formulation	Reference
1.	AIC	$\left(\frac{sse}{n}\right)e^{\frac{2(k+1)}{n}}$	[51]
2.	RICE	$\frac{\left(\frac{sse}{n}\right)}{\left[1-\left(\frac{2(k+1)}{n}\right)\right]}$	[52]
3.	Final prediction error (FPE)	$\left(\frac{(SSE)^2}{n}\right)\frac{n+(k+1)}{n-(k+1)}$	[53]
4.	Schwarz	$\left(\frac{sse}{n}\right)n^{\left(\frac{k+1}{n}\right)}$	[54]
5.	GCV	$\frac{\left(\frac{sse}{n}\right)}{\left[1-\left(\frac{k+1}{n}\right)\right]^2}$	[55]
6.	SGMASQ	$\frac{\left(\frac{sse}{n}\right)}{\left[1-\left(\frac{k+1}{n}\right)\right]}$	[55]
7.	SHIBATA	$\left(\frac{sse}{n}\right)\left(\frac{n+2(k+1)}{n}\right)$	[56]
8.	HQ	$\left(\frac{sse}{n}\right)\ln n^{\frac{2(k+1)}{n}}$	[57]

where, n is total number of observations, k+1 is estimated parameters numbers (including constant), and SSE is sum of square error.

2.4.4 Phase 4 – goodness fit

The goodness of fit test was performed on the final models selected in phase 3 to check the efficiency of the selected model. Residual data would be gathered by taking into account the difference in real and expected value for the best model in Phase 3 used MAE, RMSE, and MAPE.

3. RESULTS AND DISCUSSION

3.1 Data

The data was collected from time period of 8.00 am until 5.00 pm starting on 08/04/2017 to 12/04/2017. That is almost four days data. The original data was for each second and then it was converted in hour for data analysis. The variables taken are data contain hourly solar radiation, temperature, humidity, and moisture content. The detailed factor of modelling is shown in Table 4.

In this paper, a dataset containing 1924 observations will be used to study the effect of more 29 different independent variables on the one dependent variable. Significance of interaction terms had also been observed in this study. So, T1*T2 represents the interaction between T1 and T2. Another example H1*PY represents the interaction between H1 and PY. The data contain the effect of 435 different interaction independent variables on the one dependent variable. The more detailed tables for each variable interaction are attached in the Appendix A.

We require to select an α value between 0<α<1. Ridge and Lasso regression could be stated with α. Both α=0 and α=1 are corresponding to ridge and lasso respectively. The elastic net regression evolves combining between Lasso and Ridge.

The elastic net regression, this method evolves in the case of combining Lasso and Ridge. We require to choose an α value between $0 < \alpha < 1$ because the elastic net regression formulation is $\lambda \sum_{j=1}^{k} \left((1-\alpha)\beta_j^2 + \alpha|\beta_j| \right) \leq d^2$, $\alpha \in [0,1]$.

Table 4. Factors of modelling

Symbols	Factors	Definitions
Y	Dependent	Moisture
H1	Independent	Relative Humidity Ambient
H5	Independent	Relative Humidity Chamber
PY	Independent	Solar Radiation
T1	Independent	Temperature (℃) ambient
T2, T3, T4	Independent	Temperature (℃) before enter solar collector
T5	Independent	Temperature (℃)in front of down v-Groove (Solar Collector)
T6, T8	Independent	Temperature (℃) in front of up v-Groove (Solar Collector)
T7, T14, T15, T16, T21, T22	Independent	Temperature (℃) Solar Collector
T8, T9, T10, T11, T12	Independent	Temperature (℃) behind inside chamber
T13, T17, T18, T19, T23	Independent	Temperature (℃) Infront of (Inside Chamber)
T20, T23, T24, T25, T28	Independent	Temperature (℃) from solar collector to chamber

Table 5. Results alpha criteria for elastic net

No	alpha	Mean Square Error-Minimum	Lambda Minimum
1	0	47.2	0.925
2	0.1	27.2	0.00925
3	0.2	26.5	0.00462
4	0.3	25.8	0.00308
5	0.4	25.5	0.00231
6	0.5	24.9	0.00185
7	0.6	25.2	0.00154
8	0.7	24.8	0.00132
9	0.8	24.4	0.00116
10	0.9	24.0	0.00103
11	1	24.2	0.000925

Table 5 shows the estimates on the various value alpha to choose the minimum. The minimum value alpha obtained in step 9 and 10. We select alpha between 0.8 and 0.9 to 0.85 (interpolation) and lambda values between 0.00116 and 0.00103 to 0.001095. It shows that the model that minimized Mean Square Error (MSE) used an alpha of 0.85 and lambda of 0.001095 with the minimum MSE. From Table 5, we will convert to Figure 1. In order to, the selection of alpha value shows the minimized of MSE.

The Figure 1 shows mean square error (MSE) is widely used in model evaluations. Variations of MSE with alpha are portrayed. The Figure 1 depicts alpha (0.85) for minimum MSE (24.1).

Table 6 shows that the values of α=0, α=0.85, α=1 are corresponding to ridge, elastic net, and lasso respectively. The elastic net is a method of regularization regression that provides between ridge and lasso.

In this study, different methods for variable selections are ridge, elastic net and lasso which were performed. The variable selection is the most significant variables with their perspective. Variable selections only provide the rank of highest important variables, which means that techniques didn't have no rules in selecting the suitable range of variable important [58]. Hence, we choose the 30 highest variable importance. The variable important is shown in Table 7.

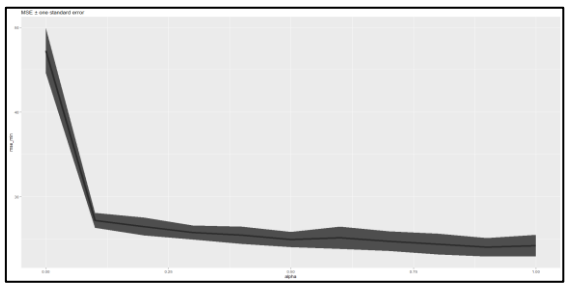

Figure 1. Minimized MSE for elastic net regression

Table 6. Choosing values alpha (α)

No	alpha	Methods
1	0	Ridge
2	0.85	Elastic Net
3	1	Lasso

Table 7. The 30 highest of variable importance

No	Methods	Variable Importance
1	Lasso	T14, T25, T22, T1*T23, T16, H1, T1*T6, PY, T9*H5, T1*T7, T4*T23, T7*T13, T7*T10, T7*T8, T2*T7, T12*H5, T29, T8, H5, T6, T3, T1, T4, T13, T2, T7, T23
2	Elastic Net	T14, T22, T25, T1*T23, H1, T1*T6, T9*H5, T17*H5, T7*T13, T2*T23, T4*T23, T1*T7, T1*T2, T6*T8, T7*T8, T2*T7, T3*T17, T8*T19, T12*H5, T8, H5, T6, T3, T4, T1, T13, T2, T7, and T23
3	Ridge	T1, T9, T6, T2, T17, T5, T23, T22, T14, T21, T28, T27, T11, T3, T1*T6, T1*T2, T7*T9, T1*T9, T26, T8, T19, H5, T15, T16, T10, T13, T4, T29, T12, and T7

Table 8. Results for the validation methods

ML	Robust Regression	MAE	MSE	MAPE	Sum Square of Error	R-square	R-square Adjusted
Ridge	Bi-Square	5.50670320	87.6181	10.7123	167701.12	0.6797253	0.674623
	Hampel	5.46121692	50.451657	10.41787	96564.47	0.8155816	0.812643
	Huber	5.42828203	50.754504	10.29807	97144.12	0.8144746	0.811519
Elastic Net	Bi-Square	5.87494596	127.19119	9.188008	243443.92	0.5350719	0.527665
	Hampel	5.494431217	48.762572	10.26748	93331.56	0.8217558	0.818916
	Huber	5.41403189	49.728886	9.966258	95181.09	0.8182236	0.815328
Lasso	Bi-Square	5.518140568	83.476643	9.124401	159774.3	0.6948638	0.690002
	Hampel	5.473598081	48.411285	9.17489	92659.21	0.8230399	0.820221
	Huber	5.395837661	49.351041	9.864292	94457.89	0.8196048	0.816731

It shows subset of 30 variable important that are taken by each technique. Three regression algorithms are applied for this purpose. i.e., Ridge, Elastic Net, and Lasso. They show the final result that was obtained by each variable important ranking technique. All the variable importance was ranked according to their importance score computed by their respective techniques. The more detailed tables for the highest 30 important variables are attached in the Appendix B. In Figure B-1 the 30 highest important variables for Ridge Regression, while in Figure B-2 the 30 highest important variables for Elastic Net Regression, and Figure B-3 the 30 highest important variables for Lasso Regression, respectively.

In order to measure the prediction accuracy, predicted responses with the actual responses are compared of each regression-based model in terms of the validation methods described in Table 8.

Predefined performances measures for Ridge, Elastic Net, and Lasso sets of data are given in Table 8. All performance measures (MAE, MSE, MAPE, Sum Square of Error, R-square, and R-square Adjusted) indicate that significantly better results were obtained by Lasso-Hampel in comparison to others. Considering Mean Absolute Error (MAE) values for Lasso-Hampel (5.473598081), MSE (48.411285), MAPE (9.17489), Sum Square of Error (92659.21), R-square (0.8230399), and R-square Adjusted (0.820221), respectively.

As can be seen in Table 8, In the context of validation (MAE, MSE, MAPE, and Sum Square of Error), the Lasso-Hamper also exhibited the lowest error data which provides the most relevant data of the result. It can be assumed that the method of Lasso-Hamper is able to be relied on investigation of the accuracy in big data obtained from regularization and robust regression.

According to Figure 2, the forecast generated by individual models show that Lasso – Hampel method leads to more accurate forecasts than the other models, because the forecasts by Lasso – Hampel method follow the pattern of actual data better than the other forecast by models used in this study. Based on the accuracy of the MAE, MSE, MAPE, Sum Square of Error, R-square, and R-square Adjusted, the obtained result is proved by Figure 2.

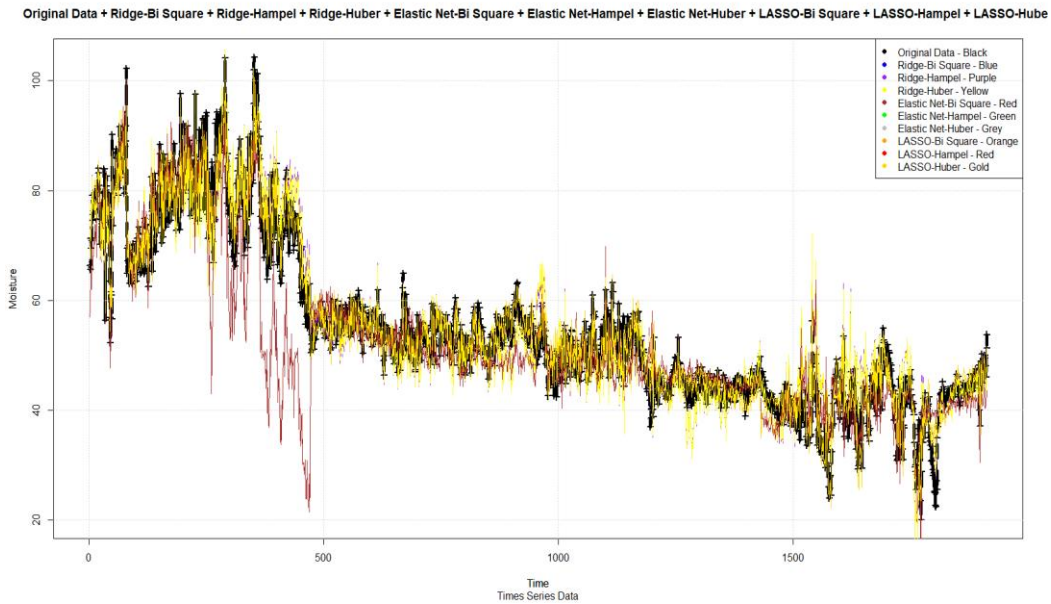

Figure 2. Accuracy measure for three regularization regressions and three robust regressions

Table 9. Results for 8 Selection Criteria for Ridge, Elastic Net, and Lasso

ML	Robust regression	AIC	GCV	HQ	RICE	SCHWARZ	SGMASQ	SHIBATA	FPE
RIDGE	Bi-Square	90.5028	90.5268	93.55051	90.55134	99.02611	89.06059	90.45633	90.50305
	Hampel	52.11269	52.12651	53.8676	52.14064	57.02053	51.28225	52.08593	52.11284
	Huber	52.42551	52.43941	54.19095	52.45363	57.36281	51.59008	52.39859	52.42566
ELASTIC NET	Bi-Square	131.3787	131.4136	135.8029	131.4492	143.7517	129.2852	131.3113	131.3791
	Hampel	50.36799	50.38135	52.06415	50.39501	55.11152	49.56535	50.34213	50.36814
	Huber	51.36612	51.3797	53.09589	51.39368	56.20365	50.54758	51.33975	51.36627
LASSO	Bi-Square	86.22497	86.24784	89.12862	86.27122	94.34541	84.85093	86.18069	86.22521
	Hampel	50.00514	50.01841	51.68908	50.03197	54.7145	49.20828	49.97947	50.00528
	Huber	50.97584	50.98936	52.69246	51.00318	55.77661	50.16351	50.94966	50.97598

All possible models have 9 models such as Regularization (Ridge, Lasso, and Elastic Net) and Robust Regression (Tukey – Bi Square, Hampel, and Hubber). The minimum value for 8SC were found for model Lasso-Hampel meaning that subset the highest 30 variable important from Lasso and continue with Hampel Regression. The results obtained from 8SC are observed in Table 9.

4. CONCLUSIONS

The results show that Lasso-Hampel model provides the best model as compared to other existing methods used in this study. The selection of efficient model needs to deal with all possible models with the second interaction terms. The proposed hybrid (Lasso-Hampel) model is found to be better in terms of MAE, MSE, and MAPE value in comparison to other existing methods. Therefore, the proposed hybrid model Lasso-Hampel can therefore be used for the efficient selection of the model including the interaction terms in it. For future work, each of 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 highest variable important were selected.

ACKNOWLEDGMENT

The author would like to thank the University of Sultan Ageng Tirtayasa Banten Indonesia and University Sains Malaysia for the support in this research project.

REFERENCES

[1] Emmert-Streib, F., Dehmer, M. (2019). High-dimensional LASSO-based computational regression models: regularization, shrinkage, and selection. Machine Learning and Knowledge Extraction, 1(1): 359-383. https://doi.org/10.3390/make1010021

[2] van der Kooij, A.J., Meulman, J.J., Heiser, W.J. (2006). Local minima in categorical multiple regression. Computational Statistics & Data Analysis, 50(2): 446-462. https://doi.org/10.1016/j.csda.2004.08.009

[3] M Ali, M.K., Ruslan, M.H., Muthuvalu, M.S., Wong, J., Sulaiman, J., Yasir, S.M. (2014). Mathematical modelling for the drying method and smoothing drying rate using cubic spline for seaweed Kappaphycus Striatum variety Durian in a solar dryer. AIP Conference Proceedings, 1602(1): 113-120. https://doi.org/10.1063/1.4882475

[4] Nurhanna, A., Othman, M. (2017). Multi-class support vector machine application in the field of agriculture and poultry: A review. Malaysian Journal of Mathematical Sciences, 11: 35-52.

[5] Ali, M.K.M., Fudholi, A., Muthuvalu, M., Sulaiman, J., Yasir, S.M. (2017). Implications of drying temperature and humidity on the drying kinetics of seaweed. AIP Conference Proceedings, 1905(1): 050004. https://doi.org/10.1063/1.5012223

[6] Yahya, W.B., Olaifa, J.B. (2014). A note on ridge regression modeling techniques. Electronic Journal of Applied Statistical Analysis, 7(2): 343-361. https://doi.org/10.1285/i20705948v7n2p343

[7] Dorugade, A., Kashid, D. (2010). Alternative method for choosing ridge parameter for regression. Applied Mathematical Sciences, 4(9): 447-456. https://doi.org/10.1016/j.jaubas.2013.03.005

[8] Kibria, B., Lukman, A.F. (2020). A new ridge-type estimator for the linear regression model: Simulations and applications. Scientifica, 2021: 9758378. https://doi.org/10.1155/2020/9758378

[9] Tinungki, G. (2019). Orthogonal iteration process of determining K value on estimator of Jackknife ridge regression parameter. Journal of Physics: Conference Series, 1341(9): 092001. https://doi.org/10.1088/1742-6596/1341/9/092001

[10] Duzan, H., Shariff, N.S.B.M. (2015). Ridge regression for solving the multicollinearity problem: Review of methods and models. Journal of Applied Sciences, 15(3): 392-404. https://doi.org/10.3923/jas.2015.392.404

[11] de Vlaming, R., Groenen, P.J. (2015). The current and future use of ridge regression for prediction in quantitative genetics. BioMed Research International, 2015: 143712. https://doi.org/10.1155/2015/143712

[12] Dorugade, A.V. (2014). New ridge parameters for ridge regression. Journal of the Association of Arab Universities for Basic and Applied Sciences, 15: 94-99. https://doi.org/10.1016/j.jaubas.2013.03.005

[13] McDonald, G.C. (2009). Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics, 1(1): 93-100. https://doi.org/10.1002/wics.14

[14] Chen, C., Chen, S., Chen, L., Zhu, Y. (2017). Method for solving LASSO problem based on multidimensional weight. Advances in Artificial Intelligence, 2017: 1736389. https://doi.org/10.1155/2017/1736389

[15] Javaid, A., Ismail, M., Ali, M.K.M. (2020). Efficient model selection of collector efficiency in solar dryer using hybrid of LASSO and robust regression. Pertanika Journal of Science & Technology, 28(1): 193-210.

[16] 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. https://doi.org/10.1111/j.1467-

9868.2005.00503.x
[17] Algamal, Z.Y., Lee, M.H. (2015). Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. Computers in Biology and Medicine, 67: 136-145. https://doi.org/10.1016/j.compbiomed.2015.10.008
[18] Ahrens, A., Hansen, C.B., Schaffer, M.E. (2020). Lassopack: Model selection and prediction with regularized regression in Stata. The Stata Journal, 20(1): 176-235. https://doi.org/10.1177/1536867X20909697
[19] Guyon, I., Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3: 1157-1182. https://doi.org/10.1162/153244303322753616
[20] Cebeci, Z., Yildiz, F. (2015). Comparison of k-means and fuzzy c-means algorithms on different cluster structures. Agrárinformatika/Journal of Agricultural Informatics, 6(3): 13-23. https://doi.org/10.17700/jai.2015.6.3.196
[21] Christensen, A., Srinivasan, V., Hart, J.C., Marshall-Colon, A. (2018). Use of computational modeling combined with advanced visualization to develop strategies for the design of crop ideotypes to address food security. Nutrition Reviews, 76(5): 332-347. https://doi.org/10.1093/nutrit/nux076
[22] Tamura, R., Kobayashi, K., Takano, Y., Miyashiro, R., Nakata, K., Matsui, T. (2017). Best subset selection for eliminating multicollinearity. Journal of the Operations Research Society of Japan, 60(3): 321-336. https://doi.org/10.15807/jorsj.60.321
[23] Cousineau, D., Chartier, S. (2010). Outliers detection and treatment: A review. International Journal of Psychological Research, 3(1): 58-67. https://doi.org/10.21500/20112084.844
[24] Begashaw, G.B., Yohannes, Y.B. (2020). Review of outlier detection and identifying using robust regression model. International Journal of Systems Science and Applied Mathematics, 5(1): 4-11. https://doi.org/10.11648/j.ijssam.20200501.12
[25] Salgado, C.M., Azevedo, C., Proença, H., Vieira, S.M. (2016). Noise versus outliers. Secondary Analysis of Electronic Health Records, pp. 163-183. https://doi.org/10.1007/978-3-319-43742-2_14
[26] Alma, Ö.G. (2011). Comparison of robust regression methods in linear regression. Int. J. Contemp. Math. Sciences, 6(9): 409-421.
[27] Fox, J., Weisberg, S. (2018). Visualizing fit and lack of fit in complex regression models with predictor effect plots and partial residuals. Journal of Statistical Software, 87(1): 1-27. https://doi.org/10.18637/jss.v087.i09
[28] Nahar, J., Purwani, S. (2017). Application of robust M-estimator regression in handling data outliers. 4th ICRIEMS Proceedings, pp. 53-60.
[29] Riani, M., Cerioli, A., Atkinson, A.C., Perrotta, D. (2014). Monitoring robust regression. Electronic Journal of Statistics, 8(1): 646-677. https://doi.org/10.1214/14-EJS897
[30] Yu, C., Yao, W. (2017). Robust linear regression: A review and comparison. Communications in Statistics-Simulation and Computation, 46(8): 6261-6282. https://doi.org/10.1080/03610918.2016.1202271
[31] Muthukrishnan, R., Radha, M. (2010). M-estimators in regression models. Journal of Mathematics Research, 2(4): 23-27. https://doi.org/10.5539/jmr.v2n4p23
[32] Alamin, M., Xu, H., Mollah, M., Haque, N. (2020). Robustification of linear regression and its application in genome-wide association studies. Frontiers in Genetics, 11: 549. https://doi.org/10.3389/fgene.2020.00549
[33] Abdullah, N., Jubok, Z.H., Ahmed, A. (2011). Improved stem volume estimation using p-value approach in polynomial regression models. Research Journal of Forestry, 5(2): 50-65. https://doi.org/10.3923/rjf.2011.50.65
[34] Zainodin, H., Noraini, A., Yap, S. (2011). An alternative multicollinearity approach in solving multiple regression problem. Trends in Applied Sciences Research, 6(11): 1241-1255. https://doi.org/10.3923/tasr.2011.1241.1255
[35] Zainodin, H., Khuneswari, G. (2010). Model-building approach in multiple binary logit model for coronary heart disease. Malaysian Journal of Mathematical Sciences, 4(1): 107-133.
[36] Ali, M.K.M., Fudholi, A., Muthuvalu, M.S., Sulaiman, J., Yasir, S.M. (2017). Implications of drying temperature and humidity on the drying kinetics of seaweed. In AIP Conference Proceedings, 1905(1): 050004. https://doi.org/10.1063/1.5012223
[37] Sulaiman, J., Ali, M., Tuah, P., Yasir, S., Lee, W. (2017). Productivity cost model in 308 ross chicken poultry systems: Case study of contract farming in rural development cooperative. Malaysian Journal of Mathematical Sciences, 11: 17-33.
[38] Lim, H.Y., Fam, P.S., Javaid, A., Ali, M., Khan, M. (2020). Ridge regression as efficient model selection and forecasting of fish drying using v-groove hybrid solar drier. Pertanika Journal of Science & Technology, 28(4): 1179-1202. https://doi.org/10.47836/pjst.28.4.04
[39] Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer, 27(2): 83-85. https://doi.org/10.1007/BF02985802
[40] Griva, I., Nash, S.G., Sofer, A. (2009). Linear and nonlinear optimization (Vol. 108). Siam.
[41] Zou. H., Hastie. T., Tibshirani, R. (2006). Journal of Computational and Graphical Statistics. 15(2): 265-285. https://doi.org/10.1198/106186006X113430
[42] Hans, C. (2011). Elastic net regression modeling with the orthant normal prior. Journal of the American Statistical Association, 106(496): 1383-1393. https://doi.org/10.1198/jasa.2011.tm09241
[43] De Mol, C., De Vito, E., Rosasco, L. (2009). Elastic-net regularization in learning theory. Journal of Complexity, 25(2): 201-230. https://doi.org/10.1016/j.jco.2009.01.002
[44] Nadia, H., Mohammad, A.A. (2013). Model of robust regression with parametric and nonparametric methods. Mathematical Theory and Modeling, 3(5): 27-39.
[45] Ding, J. (2015). An evaluation of some robust estimators of regression coefficients. Faculty of Graduate Studies and Research, University of Regina.
[46] Fox, J., Weisberg, S. (2018). An R Companion to Applied Regression. Sage Publications.
[47] Chai, T., Draxler, R.R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3): 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014
[48] Rougier, J. (2016). Ensemble averaging and mean squared error. Journal of Climate, 29(24): 8865-8870.

https://doi.org/10.1175/JCLI-D-16-0012.1
[49] Kim, S., Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. International Journal of Forecasting, 32(3): 669-679. http://dx.doi.org/10.1016/j.ijforecast.2015.12.003
[50] Mo, Z. (2014). An empirical evaluation of OLS hedonic pricing regression on Singapore private housing market. Master thesis, Department of Real Estate and Construction management. Centre of Finance and Banking. https://doi.org/10.13140/RG.2.2.24071.24484
[51] Akaike, H. (1969). Fitting autoregressive models for prediction. Annals of the Institute of Statistical Mathematics, 21(1): 243-247. https://doi.org/10.1007/BF02532251
[52] Rice, J. (1984). Bandwidth choice for nonparametric regression. Annals of Statistics, 12(4): 1215-1230. https://doi.org/10.1214/aos/1176346788
[53] Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6): 716-723. https://doi.org/10.1109/TAC.1974.1100705
[54] Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6(2): 461-464. https://doi.org/10.1214/aos/1176344136
[55] Golub, G.H., Heath, M., Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics, 21(2): 215-223. https://doi.org/10.2307/1268518
[56] Shibata, R. (1981). An optimal selection of regression variables. Biometrika, 68(1): 45-54. https://doi.org/10.1093/biomet/68.1.45
[57] Hannan, E.J., Quinn, B.G. (1979). The determination of the order of an autoregression. Journal of the Royal Statistical Society: Series B(Methodological): 41(2): 190-195. https://doi.org/10.1111/j.2517-6161.1979.tb01072.x
[58] Gómez-Verdejo, V., Parrado-Hernández, E., Tohka, J. (2019). Sign-consistency based variable importance for machine learning in brain imaging. Neuroinformatics, 17(4): 593-609. https://doi.org/10.1007/s12021-019-9415-3

APPENDIX

Appendix A

data.frame(T1,T2,T3,T4,T5,T6,T7,T8,T9,T10,T11,T12,T13,T14,T15,T16,T17,T19,T21,T22,T23,T25,H5,PY,T1*T2,T1*T3,T1*T4,T1*T5,T1*T6,T1*T7,T1*T8,T1*T9,T1*T10,T1*T11,T1*T12,T1*T13,T1*T14,T1*T15,T26,T27,T28,T29,H1,T1*T16,T1*T17,T1*T19,T1*T21,T1*T22,T1*T23,T1*T25,T1*T26,T1*T27,T1*T28,T1*T29,T2*T3,T2*T4,T2*T5,T2*T6,T2*T7,T2*T8,T2*T9,T2*T10,T2*T11,T2*T12,T2*T13,T2*T14,T2*T15,T2*T16,T2*T17,T3*T5,T3*T6,T1*H1,T1*H5,T1*PY,T2*T19,T2*T21,T2*T22,T2*T23,T2*T25,T2*T26,T2*T27,T2*T28,T2*T29,T2*H1,T2*H5,T2*PY,T3*T4,T3*T7,T3*T8,T3*T9,T3*T10,T3*T11,T3*T12,T3*T13,T3*T14,T3*T15,T3*T16,T3*T17,T3*T19,T3*T21,T3*T22,T3*T23,T3*T25,T3*T26,T3*T27,T3*T28,T3*T29,T3*H1,T3*H5,T3*PY,T4*T5,T4*T6,T4*T7,T4*T8,T4*T9,T4*T10,T4*T11,T4*T12,T4*T13,T4*T14,T4*T15,T4*T16,T4*T17,T4*T19,T4*T21,T4*T22,T4*T23,T4*T25,T4*T26,T4*T27,T4*T28,T4*T29,T4*H1,T4*H5,T4*PY,T5*T6,T5*T7,T5*T8,T5*T9,T5*T10,T5*T11,T5*T12,T5*T13,T5*T14,T5*T15,T5*T16,T5*T17,T5*T19,T5*T21,T5*T22,T5*T23,T5*T25,T5*T26,T5*T27,T5*T28,T5*T29,T5*H1,T5*H5,T5*PY,T6*T7,T6*T8,T6*T9,T6*T10,T6*T11,T6*T12,T6*T13,T6*T14,T6*T15,T6*T16,T6*T17,T6*T19,T6*T21,T6*T22,T6*T23,T6*T25,T6*T26,T6*T27,T6*T28,T6*T29,T6*H1,T6*H5,T6*PY,T7*T8,T7*T9,T7*T10,T7*T11,T7*T12,T7*T13,T7*T14,T7*T15,T7*T16,T7*T17,T7*T19,T7*T21,T7*T22,T7*T23,T7*T25,T7*T26,T7*T27,T7*T28,T7*T29,T7*H1,T7*H5,T7*PY,T8*T9,T8*T10,T8*T11,T8*T12,T8*T13,T8*T14,T8*T15,T8*T16,T8*T17,T8*T19,T8*T21,T8*T22,T8*T23,T8*T25,T8*T26,T8*T27,T8*T28,T8*T29,T8*H1,T8*H5,T8*PY,T9*T10,T9*T11,T9*T12,T9*T13,T9*T14,T9*T15,T9*T16,T9*T17,T9*T19,T9*T21,T9*T22,T9*T23,T9*T25,T9*T26,T9*T27,T9*T28,T9*T29,T9*H1,T9*H5,T9*PY,T10*T11,T10*T12,T10*T13,T10*T14,T10*T15,T10*T16,T10*T17,T10*T19,T10*T21,T10*T22,T10*T23,T10*T25,T10*T2,T10*T27,T10*T28,T10*T29,T10*H1,T10*H5,T10*PY,T11*T12,T11*T13,T11*T14,T11*T15,T11*T16,T11*T17,T11*T19,T11*T21,T11*T22,T11*T23,T11*T25,T11*T26,T11*T27,T11*T28,T11*T29,T11*H1,T11*H5,T11*PY,T12*T13,T12*T14,T12*T15,T12*T16,T12*T17,T12*T19,T12*T21,T12*T22,T12*T23,T12*T25,T12*T26,T12*T27,T12*T28,T12*T29,T12*H1,T12*H5,T12*PY,T13*T14,T13*T15,T13*T16,T13*T17,T13*T19,T13*T21,T13*T22,T13*T23,T13*T25,T13*T26,T13*T27,T13*T28,T13*T29,T13*H1,T13*H5,T13*PY,T14*T15,T14*T16,T14*T17,T14*T19,T14*T21,T14*T22,T14*T23,T14*T25,T14*T26,T14*T27,T14*T28,T14*T29,T14*H1,T14*H5,T14*PY,T15*T16,T15*T17,T15*T19,T15*T21,T15*T22,T15*T23,T15*T25,T15*T26,T15*T27,T15*T28,T15*T29,T15*H1,T15*H5,T15*PY,T16*T17,T16*T19,T16*T21,T16*T22,T16*T23,T16*T25,T16*T26,T16*T27,T16*T28,T16*T29,T16*H1,T16*H5,T16*PY,T17*T19,T17*T21,T17*T22,T17*T23,T17*T25,T17*T26,T17*T27,T17*T28,T17*T29,T17*H1,T17*H5,T17*PY,T19*T21,T19*T22,T19*T23,T19*T25,T19*T26,T19*T27,T19*T28,T19*T29,T19*H1,T19*H5,T19*PY,T21*T22,T21*T23,T21*T25,T21*T26,T21*T27,T21*T28,T21*T29,T21*H1,T21*H5,T21*PY,T22*T23,T22*T25,T22*T26,T22*T27,T22*T28,T22*T29,T22*H1,T22*H5,T22*PY,T23*T25,T23*T26,T23*T27,T23*T28,T23*T29,T23*H1,T23*H5,T23*PY,T25*T26,T25*T27,T25*T28,T25*T29,T25*H1,T25*H5,T25*PY,T26*T27,T26*T28,T26*T29,T26*H1,T26*H5,T26*PY,T27*T28,T27*T29,T27*H1,T27*H5,T27*PY,T28*T29,T28*H1,T28*H5,T28*PY,T29*H1,T29*H5,T29*PY,H1*H5,H1*PY,H5*PY)

Appendix B

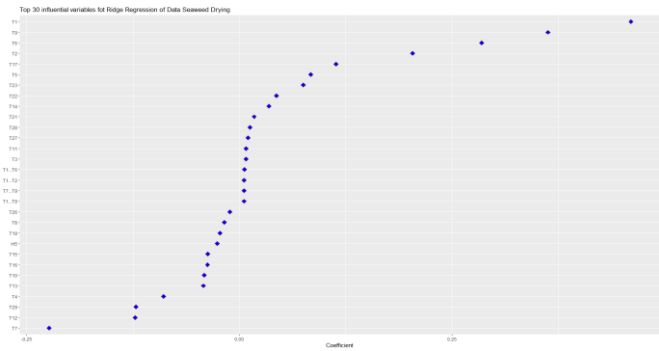

Figure B-1. The 30 highest important variables for Ridge Regression

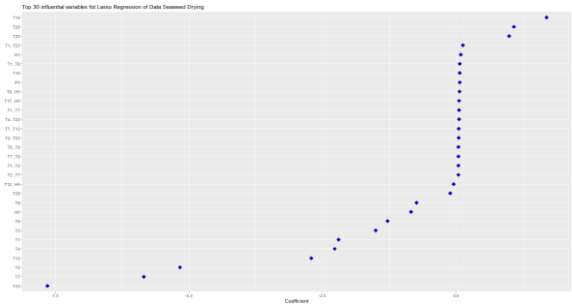

Figure B-2. The 30 highest important variables for Elastic Net Regression

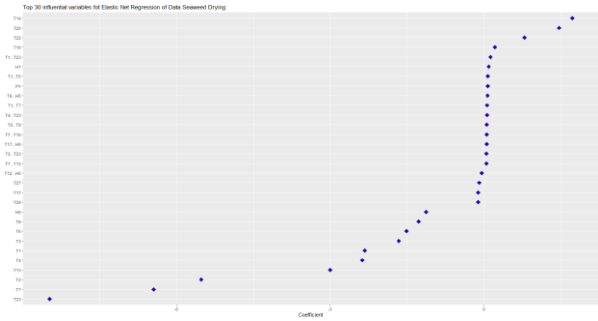

Figure B-3. The 30 highest important variables for Lasso Regression