Pertanika J. Sci. & Technol. 28 (1): 193 - 210 (2020)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Efficient Model Selection of Collector Efficiency in Solar Dryer using Hybrid of LASSO and Robust Regression

Anam Javaid[1,2]*, Mohd. Tahir Ismail[1] and Majid Khan Majahar Ali[1]

[1]*School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia*
[2]*Department of Statistics, The Women University Multan, Pakistan*

ABSTRACT

There are many variables involved in the real life problem so it is difficult to choose an efficient model out of all possible models relating to analytical factors. Interaction terms affecting the model also need to be addressed because of its vital role in the actual dataset. The current study focused on efficient model selection for collector efficiency of solar dryer. For this purpose, collector efficiency of solar dryer was used as a dependent variable with time, inlet temperature, collector average temperature and solar radiation as independent variables. Hybrid of the least absolute shrinkage and selection operator (LASSO) and robust regression were proposed for the identification of efficient model selection. The comparison was made with the ordinary least square (OLS) after performing a multicollinearity and coefficient test and with a ridge regression analysis. The final selected model was obtained using eight selection criteria (8SC). To forecast the efficient model, the mean absolute percentage error (MAPE) was used. As compared to other methods, the proposed method provides a more efficient model with minimum MAPE.

Keywords: model selection, ordinary least square, robust regression, selection criteria, sparse regression

INTRODUCTION

Food insecurity is considered to be a major problem in the agricultural sector (Ahmed et al., 2017). It is therefore necessary to produce more food due to the food insecurity problems (Rockstrom et al., 2009). There are many stages of crop management, such as nutrient supply, water, crop production environment, in the process of seeding to harvesting (Yan, 2011). Drying

ARTICLE INFO

Article history:
Received: 28 January 2019
Accepted: 04 November 2019
Published: 13 January 2020

E-mail addresses:
anamjavaid0786@yahoo.com (Anam Javaid)
m.tahir@usm.my (Mohd. Tahir Ismail)
majidkhanmajaharali@usm.my (Majid Khan Majahar Ali)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

is one of the key processes in agriculture or aquaculture. Air drying is the most commonly used dehydration operation in food and chemical industries (Ali et al., 2014). The global population has been growing over the years, and there are many variables with interactions needed to solve food insecurity in agriculture or aquaculture, so the main problem is to find the key variables amongst them so that the complexity of the model can be reduced and the model can be used to predict the supply demand for food using a more efficient model (Taylor & Adelman, 2003).

Many models only emphasise a single term without considering terms of interaction (Chen, 2012). Seaweed is one of the most common products used in agriculture and aquaculture. It is currently used mostly in food manufacturing, medical and manufacturing industries. Seaweed is regarded as a potential source for renewable energy. It also can be transformed into energy such as gas and biofuel oil. Dissa et al. (2011) claimed that carrageenan was a major cause of seaweed extraction. Ali et al. (2015) conducted the study to find out that carrageenan was also used in food and non-food products for humans, cosmetics, animal foods, meat binder and they developed a mathematical model for drying method and for smoothing drying rate. Many models have been proposed using analytical or empirical solutions in the aquacultural field (Neitsch et al., 2011). Different simple techniques such as ordinary least square (OLS) and other simple methods such as analysis of variance, principle component analysis was used to address problems related to agriculture and aquaculture, but these simple techniques have many constraints (Zuur et al., 2009).

Multicollinearity is one of the problems, particularly in the case of large-scale data analysis. Rischbeck et al. (2016) used multiple linear regression models for midly drought-stressed field trials that were impacted by multicollinearity problems. In the case of multicollinearity, OLS estimates have large variances and covariances that make it difficult to calculate precisely estimates (Gujarati, 2004). The OLS model was also used to predict grain yield (Montesinos-Lopez et al., 2017). Linear regression was also used for the development of the dengue forecast model **(**Guo et al., 2017). The regression analysis was used to explore the interaction between climate, water and agriculture by adding linear and quadratic terms (Mendelsohn & Dinar, 2003). Linear regression was used to estimate the structural economic model to increase productivity in agriculture (Pender et al., 2004). Some research has been performed on multicollinearity, as Giacalone et al. (2018) launched the regularizaton methods (L norm) for the compaction of the multicollinearity problem. Other work was done by Wouldiams et al. (2012) as they used the analysis of variance to examine the yield-related factors.

Variable selection is another issue with regression analysis as OLS does not cope with variable selection. Sparse regressions are implemented for this type of problem by adding a penalty term. This penalty term is introduced in the function of minimisation so it is necessary to work with sparse regression analysis for variable selection. Xu and Ying (2010)

used median regression with least absolute shrinkage and selection operator (LASSO) type penalty regression to select variables afterwards Zhao et al. (2012) investigated wavelet-based LASSO methodology to regress function scalars. They also explored its asymptotic convergence as well as its finite-sample performance by using both simulation and actual data examples. Zhang et al. (2016) implemented LASSO, adaptive LASSO, adaptive LASSO II, multitask LASSO, reweighted LASSO on quantitative trait loci analysis that offered helpful insights in the research of human cancer. Zou (2006) used adaptive weights to penalize distinct coefficients in terms of the absolute value of magnitude of coefficients (L1 penalty). In order to understand the contribution of individual observations and robustness outcome for evaluated values of the model parameters (Jang & Anderson-Cook, 2017) examined the influence plot of LASSO.

The presence of outliers is also a major problem in the dataset. The removal of the outlier is not always a good option for analysis, so robust methods are necessary in order to detect and remove outlier as (Gad & Qura, 2016) have reviewed a wide variety of robust outliers methods. Midi et al. (2011) proposed some practical lower bound (LB) and upper bound (UB) for high leverage collinearity influential measure (HLCIM) that was an essential measure for the detection of multicollinearity degree. Ridge regression is also used in cases of multicollinearity but is considered to be affected in the presence of outliers (Shariff & Ferdaos, 2017). Gusnanto and Pawitan (2015) had compared various methods including sparse regression in case of number of variables were greater than number of observations ($p > n$) and had preferred sparse methods for high multicolinearity.

The method named two-step robust weighted least squares (TSRWLS) method was studied in Midi et al. (2014). Beath (2018) worked on robustness method for linear models but the disadvantage was that it could only deal with group logistics, not binary logistics, as binary logistics could not exactly fit observation. It had been evident, that LASSO was mostly applied to medical fields and gene data since in this area, there was a large number of variables to deal with, but little research was done in relation to LASSO as a wavelet-based LASSO technique was done by Zhao et al. (2012). Gusnanto and Pawitan (2015) compared ridge, cauchy, LASSO, mixture of Normals and adaptive LASSO on near infrared (NIR) instruments. Similarly, in terms of robust or ridge regression analysis, not much research has been performed on agriculture. Many types of estimators were used in robust regression analysis as Susanti et al. (2014) presented maximum likelihood type estimators (*M* estimators), modified *M* estimates (*MM*) and estimators of scale (*S*) estimates on maize production data while mostly researchers preferred *M* estimates as Sinova and Van Aelst (2018) showed advantages for Tukey bisquare-based *M* estimates by comparing them with the hampel loss function for fuzzy number value calculation. Shariff and Ferdaos (2017) provided a robust ridge method of regression model for multi-collinearity and outlier problems. Model selection was also made by different reserachers, as Abdullah et al. (2011)

used eight selection criteria (8SC) to obtain the best model among all possible models. Similarly Zainodin et al. (2011) used 8SC in model selection problem.

It is clear that the work was done separately on LASSO and robust, but there is no such model that uses 8SC to combine Robust and LASSO. In this study, therefore, this gap is addressed in the development of a robust and LASSO models and the objective is to select the best model by using 8SC that can be used to efficient prediction. In this study, collector efficiency factors are observed for solar drier using a hybrid model of LASSO and Huber M estimator, and comparisons are made with OLS and ridge regression analysis.

MATERIAL AND METHODS

OLS, Tikhonov regularisation (Ridge), LASSO and robust regression would be used for dataset assessment. The flow chart used in this research can be found in Figure 1.

The following phases were performed for the application of the flow chart referred to in Figure 1.

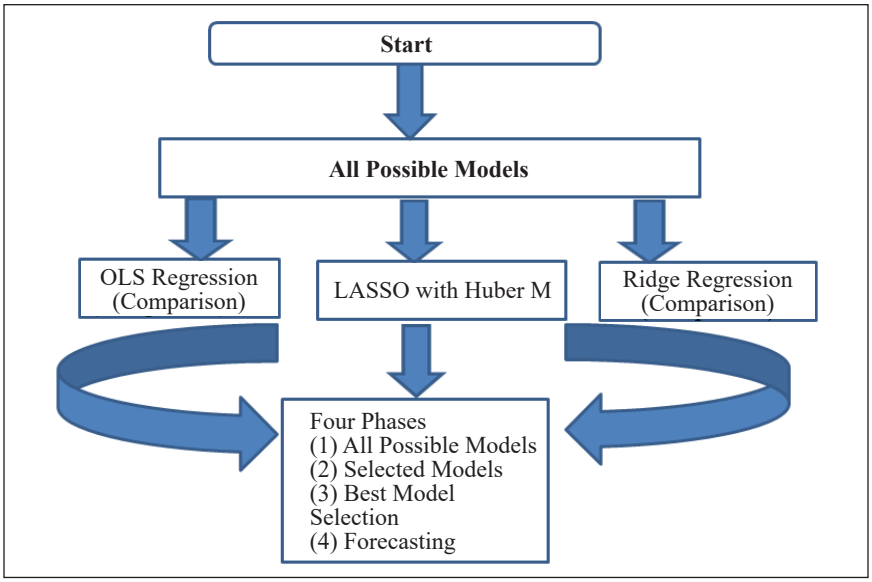

Figure 1. Flow Chart for best selected model

Phase 1– All Possible Models

According to Khuneswari et al. (2008), all possible models are the prerequisite for determination of the best model and can be derived by using Equation 1

$$N = \sum_{j=1}^{k} j \binom{k}{j} \quad (1)$$

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

These all possible models would be used by OLS, LASSO and ridge after that the procedure would be moved on to the next phase.

No observation was missing in the dataset. Thus, approximately 16.67% of data reserved for the mean absolute percentage error (MAPE) would be used to predict the best model in Phase 3 later.

Phase 2- Selected Variables

For this phase, two tests were performed for OLS, *i.e.* the multicollinearity test and the coefficient test. After performing these two tests, selected models would be obtained in the OLS regression analysis while the significant variables would be selected for the ridge and LASSO regression because LASSO is a sparse regression to perform an automatic selection of variables. The coefficients were compared to the 0.05 level of significance for ridge regression.

Multicollinearity Test. Multicollinearity occurs in the case of a correlation of independent variables and is considered to be a problem in multiple regression analysis as a problem arises in the model validity of the investigation (Gujarati, 2004).

The following steps were taken to address the problem of multicollinearity.

i. In the first step, the correlation coefficient is calculated for all variables in each model and the verification is performed between independent variables with a high value (coefficient > 0.95).
ii. Following this, most common high correlation coefficient variable was removed and the correlation coefficient recalculated.
iii. Steps (i) and (ii) are repeated until there is no variable left with a high multicollinearity problem, if any, the variable with a lower value of the absolute correlation coefficient with the dependent variable is removed.
iv. The correlation coefficient between the dependent variable and the entire multicollinearity source was checked for the existence or non-existence of multicollinearity between the dependent and the other variables.

Coefficient Test. According to Ramanathan (2002), the coefficient test is considered to be a test for each independent variable whether or not it differs significantly from zero. *i.e.* it can be tested under the following hyporthesis.

H_0: $\beta_j = 0$

H_1: $\beta_j \neq 0$

Where is the coefficient of variable in the model for $j = 1,2, ..., k$ and the t test would be performed at a 5% level of significance for the test. The best model would undergo a fitness test that includes a normality test and a randomness test on residuals for the model (Abdullah et al., 2011).

For LASSO regression analysis, robust regression was conducted as a coefficient test on each model. Tibshirani (1996) first introduced LASSO that could select coefficients β to minimise (Equation 2).

$$(y - X\beta)(y - X\beta)' + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$= (y - X\beta)(y - X\beta)' \quad \text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| \leq s \qquad (2)$$

Where s and λ are considered to be non-negative regularisation parameters. LASSO used the $L1$ norm that explains the coordinates vertices and the edge polotype where some coordinate values are zero. A solution for LASSO is commonly found on polotype vertex or polotype edges. LASSO can therefore be called a variable selection method where coefficient shrinkage to zero can eliminate variables from the model. Hoerl and Kennard (1970) introduced ridge regression with a bias parameter $b(d)$ obtained by using all variables as follows (Equation 3).

$$b(d) = (X'X + dI)^{-1}X'y \qquad (3)$$

Where d is the bias parameter. If $d = 0$, the ridge parameter is equal to the OLS parameter.

There are many types of estimators available in the case of robust regression analysis, but the most common types are the M estimators where Huber, Hampel and bisquares were mostly used. Stuart (2011) defined the typical tuning constant for Huber as a= 1.345 for 95% relative efficiency, For Hampel the typical tunning constants are $a = 2$, $b = 4$ and $c = 8$ and for Tukey's Bisquare the typical tuning constant is a= 4.685 results in 95% relative efficiency with the weight functions defined in Table 1

Phase 3 - The Best Model

Once the selected models have been obtained, the best model can be obtained among selected models. Ali et al. (2017) stated the 8SC that could be used to choose the best model from the list of selected models. For best model selection, 8SC would be used in this research. The formulae are outlined in Table 2.

Table 1
Weight function used for different Regression methods

	Objective Function $\rho(u)$	Score function	Weight Function $w(u) = \frac{\psi(u)}{u}$
a) **Least Squares**	$\frac{1}{2}u^2 \quad -\infty \le u \le \infty$	u	1
b) Huber M ***a>0***	$\begin{cases} \frac{1}{2}u^2 & if \ \lvert u\rvert < a \\ a\lvert u\rvert - \frac{1}{2}a^2 & if \ \lvert u\rvert \ge a \end{cases}$	$\begin{cases} u & if \ \lvert u\rvert < a \\ a \ sign \ u & if \ \lvert u\rvert \ge a \end{cases}$	$\begin{cases} 1 & if \ \lvert u\rvert < a \\ \frac{a}{\lvert u\rvert} & if \ \lvert u\rvert \ge a \end{cases}$
c) Hampel M ***a,b,c>0***	$\begin{cases} \frac{1}{2}u^2 & if \quad \lvert u\rvert < a \\ a\lvert u\rvert - \frac{1}{2}a^2 & if \ a \le \lvert u\rvert < b \\ a\frac{c\lvert u\rvert - \frac{1}{2}u^2}{c-b} - \frac{7a^2}{6} & if \ b \le \lvert u\rvert \le c \\ a(b+c-a) & Otherwise \end{cases}$	$\begin{cases} u & if \quad \lvert u\rvert < a \\ a \ sign \ u & if \ a \le \lvert u\rvert < b \\ a\frac{c \, sign \, u - u}{c-b} & if \ b \le \lvert u\rvert \le c \\ 0 & Otherwise \end{cases}$	$\begin{cases} 1 & if \quad \lvert u\rvert < a \\ \frac{a}{\lvert u\rvert} & if \ a \le \lvert u\rvert < b \\ 0 & Otherwise \end{cases}$
d) Tukey Bisquare M ***a>0***	$\begin{cases} \frac{a^2}{6}\left(1-\left(1-\frac{u}{a}\right)^2\right)^3 & if \ \lvert u\rvert \le a \\ \frac{1}{6}a^2 & if \ \lvert u\rvert > a \end{cases}$	$\begin{cases} u\left(1-\left(\frac{u}{a}\right)^2\right)^2 & if \ \lvert u\rvert < a \\ 0 & if \ \lvert u\rvert > a \end{cases}$	$\begin{cases} \left(1-\left(\frac{u}{a}\right)^2\right)^2 & if \ \lvert u\rvert \le a \\ 0 & if \ \lvert u\rvert > a \end{cases}$

Table 2
Formula used for eight selection criteria

Selection criteria	Formula	Reference
AIC	$\left(\frac{SSE}{n}\right)(e)^{2(k+1)/n}$	Akaike, 1969
RICE	$\left(\frac{SSE}{n}\right)\left[1-\left(\frac{2(k+1)}{n}\right)\right]^{-1}$	Rice, 1984
FPE	$\left(\frac{SSE^2}{n}\right)\frac{n+(k+1)}{n-(k+1)}$	Akaike, 1974
SCHWARZ	$\left(\frac{SSE}{n}\right)n^{(k+1)/n}$	Schwarz, 1978
GCV	$\left(\frac{SSE}{n}\right)\left[1-\left(\frac{k+1}{n}\right)\right]^{-2}$	Golub et al., 1979
SGMASQ	$\left(\frac{SSE}{n}\right)\left[1-\left(\frac{k+1}{n}\right)\right]^{-1}$	Ramanathan, 2002
HQ	$\left(\frac{SSE}{n}\right)(ln\, n)^{2(k+1)/n}$	Hannan and Quinn, 1979
SHIBATA	$\left(\frac{SSE}{n}\right)\frac{n+2(k+1)}{n}$	Shibata, 1981

where
n = total number of observations
$k+1$ = estimated parameters numbers (including constant)
SSE = sum of square error

By using formula in Table 2, Akaike information criterion *(AIC*), RICE, Final prediction error (FPE), SCHWARZ(SBC), generalized cross validation (GCV), sigma square(SGMASQ), Hannan-Quinn information criterion (HQ) and SHIBATA were calculated for the purpose of efficient model selection.

Phase 4 - Goodness of Fit

Gujarati (2004) had described some assumptions concerning the least square estimators, such as that there should be no ideal multicollinearity and that the model should be completely identified. Ramanathan (2002) stated that the goodness of the fitness test ensured that the model fitted well into the data. In this phase, 16.67% of Phase 1 datasets were used for the calculation of the MAPE value, in order to determine model efficiency. Residual data would be gathered by taking into account the difference in real and expected value for the best model in Phase 3. Ali et al. (2017) used the MAPE Formula as in Equation 4.

$$MAPE = \frac{100}{N}\left(\frac{\sum_{i=1}^{j}|A_i - E_i|}{A_i}\right) \qquad i=1,2,\ldots,j \qquad (4)$$

Where

A = actual value of dependent variable (y)

E = expected value ($\hat{y}$)

N = number of observations points

A non-parametric test such as a randomness test would be performed to check the random pattern of observations. For normality assumptions, the Sharpio wilk test and the Kolmogrov smirnov test would be used with the sporting documents of the scatter plot, the histogram and the box plot of the residues obtained from the efficient selected models.

RESULT AND DISCUSSION

Data Collection and Procedure

The information used in this research was drawn from Sabah. In Sabah, solar dryer is used for the drying method and various variables influence on the effectiveness of the collector. In this research, four factors, such as time, inlet temperature, collector average temperature and solar radiation, were taken as independent variables while the collector efficiency was maintained as dependent variables. For analysis purposes, 66 observations were taken. Data were collected for every second and then converted into hour to analyse the behaviour of different variables during the given time frame. Dataset was collected for four days in which solar radiation was at the peak during this time period, from 8:00 a.m. to 5:00 p.m. The purpose of this study was to monitor each factor behaviour on collector efficiency of solar dryer in which Y was used to indicate the efficiency of the collector as a dependant variable whereas x_1, x_2, x_3 and x_4 represented independent variables such as time, inlet temperature, collector average temperature and solar radiation, respectively, where x_{12} represented the interaction between x_1 and x_2 and was used to observe the combined behaviour of x_1 and x_2 on the collector effectiveness. Inlet temperature was observed to be between 27.9°C and 58.3°C in the complete data procedure, while the collector average temperature was found to be between 33.0°C and 87.7°C. Solar radiation was observed to be between 104.3 W/m² and 819.8 W/m² at 8:00 a.m. to 5:00 p.m. in four days.

For four independent variables, 32 possible models were available until the third order interaction term. All possible models consisting of four independent variables could be observed as shown in Table 3.

All possible models were calculated as indicated in Table 3 and, following a multicolinearity test and a coefficient test, a list of selected models was obtained. The list of selected models was achieved by the sum of square of error (SSE) and the number of variables left in the selected model *(k)* can be seen in Table 4.

From Table 4, the original model demonstrated that Model *M32* had to go through two phases. The multicollinearity test was applied to this original model and, as a result,

Table 3
All possible models

No of variables	single	Interact			Total
		1st Order	**2nd Order**	**3rd Order**	
1	4	-	-	-	**4**
2	6	6	-	-	**12**
3	4	4	4	-	**12**
4	1	1	1	1	**4**
Total Models	15	5	5	1	**32**
Model ID	***M1-M15***	***M16-M26***	***M27-M31***	***M32***	

Table 4
Selected models by using ordinary least square method

Sr. NO	Selected models using OLS	*k*	SSE
1	*M1.0.0=M5.0.1*	1	2810.25
2	*M2.0.2*	1	3021.1
3	*M3.0.0= M6.0.1*	1	2090
4	*M4.0.0 = M7.0.1*	1	1210.8
5	*M8.0.0=M19.1.0*	2	1547.1
6	*M9.0.0=M12.0.1=M25.4.0*	2	948.53
7	*M10.0.0=M13.0.1=M14.0.1=M15.0.2=M21.1.0*	2	824.86
8	*M11.0.0*	3	1361.23
9	*M16.0.1*	2	2588.28
10	*M17.0.1*	2	1566.93
11	*M18.0.1*	2	915.71
12	*M20.1.0*	2	1341.24
13	*M22.2.0=M27.2.1*	4	990.71
14	*M23.1.1*	4	650.97
15	*M24.2.0*	4	742.16
16	*M26.5.1*	4	585.31
17	*M28.2.1*	4	650.97
18	*M29.2.0*	5	562.59
19	*M30.4.1*	2	619.59
20	*M31.8.4*	3	544.92
21	*M32.7.2*	6	444.07

seven variables were removed from the model. So it became as *M32.7.0.* Coefficient test was performed and two variables were removed from the model so that the best selected model obtained as *M32.7.2*. With all significant variables in the model, the resulting model was now free of multicollineraity. The best model selected using the formula for the 8SC set out in Table 2 can be found in Table 5.

Table 5
Eight selection criteria for OLS selected models

Selected models from OLS	AIC	FPE	GCV	HQ	RICE	SCHWARZ	SGMASQ	SHIBATA
M1.0.0=M5.0.1	54.94	54.95	55.02	56.52	55.10	59.11	53.0	54.81
M2.0.2	59.07	59.07	59.15	60.764	59.237	63.55	57.001	58.92
M3.0.0= M6.0.1	40.86	40.86	40.9	42.0	40.98	43.96	39.43	40.76
M4.0.0 = M7.0.1	23.67	23.67	23.70	24.35	23.74	25.46	22.84	23.61
M8.0.0=M19.1.0	31.37	31.37	31.46	32.72	31.57	35.00	29.75	31.19
M9.0.0=M12.0.1=M25.4.0	19.23	19.23	19.293	20.06	19.35	21.45	18.24	19.12
M10.0.0=M13.0.1=M14.0.1=M15.0.2=M21.1.0	16.7	16.72	16.77	17.44	16.83	18.66	15.86	16.63
M11.0.0	27.60	27.60	27.68	28.79	27.78	30.79	26.17	27.44
M16.0.1	52.48	52.48	52.64	54.7	52.82	58.5	49.77	52.19
M17.0.1	31.7	31.77	31.87	33.14	31.9	35.44	30.1	31.59
M18.0.1	18.56	18.57	18.62	19.3	18.68	20.7	17.60	18.46
M20.1.0	27.19	27.19	27.28	28.3	27.3	30.34	25.79	27.04
M22.2.0=M27.2.1	21.60	21.6	21.79	23.18	22.01	25.92	19.81	21.28
M23.1.1	14.19	14.202	14.32	15.23	14.4	17.03	13.01	13.98
M24.2.0	16.18	16.19	16.32	17.36	16.49	19.42	14.84	15.94
M26.5.1	12.76	12.77	12.87	13.69	13.00	15.31	11.70	12.57
M28.2.1	14.19	14.20	14.32	15.23	14.46	17.03	13.01	13.98
M29.2.0	12.7	12.73	12.88	13.84	13.08	15.83	11.48	12.46
M30.4.1	12.56	12.56	12.60	13.10	12.64	14.01	11.91	12.49
M31.8.4	11.45	11.46	11.52	12.12	11.59	13.25	10.68	11.34
M32.7.2	10.41	10.42	10.60	11.49	10.83	13.44	9.251	10.12

From Table 5, it is clear that the minimum selection criterion value for model *M32.7.2* was obtained from all the best possible models by using formulae for each selection criterion, thus the final selected model *M32.7.2* can be observed with its coefficient values in Equation 5 using R software.

$$M32.7.2 = Y = 32.12 + 1.58x_1 - 0.0548x_4 - 0.0274x_{12} - 0.002\ x_{14} + 0.005\ x_{23} + 0.00000058x_{1234} \quad (5)$$

For the purpose of analysis, coefficients were observed in the final selection of models for each variable. After multicollineraity and coefficient testing of 15 variables, there were 6 variables left in the model including interaction terms in it. From the selected model in Equation 4, the significant variables were time, solar radiation, interaction of time and inlet temperature, interaction of time and solar radiation, interaction of inlet temperature and collector average temperature, interaction of time, inlet temperature, collector average temperature and solar radiation. The interaction terms between variables could be seen to be crucial for model selection so that we could not ignore them. From the coefficient, as the time increased, dryer collector efficiencies were increased by 1.58 units. Similar to the increase in solar radiation, there would be a 0.05 unit decrease in collector dryer efficiency, as solar radiation would mostly be effective from 11: 00 a.m. to 1: 00 p.m. Time and inlet temperature interactions would cause collector efficiency to decrease by 0.0274 units as inlet temperature can not be controlled. Solar radiation would cause collector efficiency to decrease by 0.002 units. Interaction between the inlet temperature and the average collector temperature had a positive effect on the collector's efficiency.

For this selected model, the MAPE value was calculated using a specific formula as defined in Equation 4. SSE with the number of variables left in the model *(k)* was used to obtain MAPE value. The MAPE value for the dataset was found to be MAPE = 29.2198 .

The MAPE value is not so high that the selected model can be used for forecasting. The standardised residuals for the selected model were calculated after MAPE calculation. The standardised residual for this final selected model can be viewed as in Figure 2.

In Figure 2, the pattern for an effective chosen model is random and suggests a good fit for a linear model. Outliers can be seen outside the 2 sigma boundaries. The randomness test and the normality test were also conducted for proof. Stuart (2011) explained that the performance of the least square estimators was not good in the case of outliers or in the case of deviations from normal assumptions, so that Ridge regression was considered as an alternative method in the case of highly collinear predictors. The ridge regression estimates are biased, but the mean square error of the ridge estimator is smaller than the OLS estimators of Hoerl and Kennard (1970). Since the data used in this study also has multicollinearity problem, for comparison purposes ridge regression is performed on all

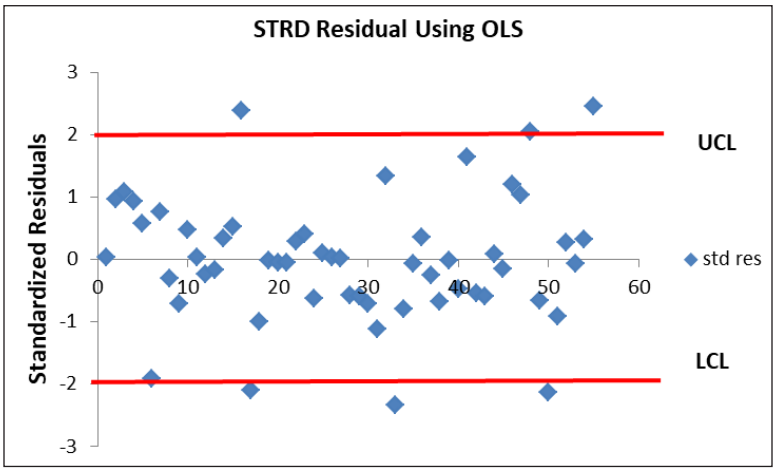

Figure 2. Standardized residual for OLS

possible models and the selected models are obtained. In these selected models, 8SC is performed in the same way as in OLS and the best model is selected. *M24* was observed as the best model with a minimum SSE value of 8SC. The coefficients are obtained using the library *glmnet* in R software (Equation 6).

$$M24 = 31.938 + 0.3144\,x_1 + 0.0604\,x_3 - 0.0218x_4 + 0.0022x_{13} + 0.00082x_{14} - 0.000032x_{34} \quad (6)$$

The best model in Equation 6 can be observed with the key variables. Since ridge regression and OLS contained all the variables in the model as they did not have the ability to select the model, so for the purpose of a sparse regression analysis, the modified LASSO was performed using the Huber M estimation method. LASSO was performed on all possible models for a sparse regression analysis. Significant variables were observed in LASSO with 17 models after grouping a model consisting of the same variables, while Huber M was used for efficient model selection. After the performance of Huber M, 12 models were left at a 0.05% significance level in the modified LASSO. 8SC for efficient model selection were performed on these 12 models and *M29.1.1* was observed to describe the efficient model with minimum SSE as in Equation 7.

With 239.48 SSE and significant variables for collector efficiency can be seen from the above selected models. Using R software (*glmnet* library), other models with their coefficients were selected the same way. *M29.1.1* represented that from model 29, one variable was removed in LASSO and one variable was removed using Huber M as non-significant variables. The resulting model notation became as *M29.1.1*.

Table 6
MAPE for final selected models of all methods

Selected Model	Technique used	Variables in the Model	SSE	MAPE
M32.7.2	OLS	$Y = \beta_0 + \beta_1 x_1 + \beta_4 x_4 + \beta_{12} x_{12} + \beta_{14} x_{14} + \beta_{23} x_{23} + \beta_{1234} x_{1234}$	444.07	29.21
M24	Ridge Regression	$Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \beta_{13} x_{13} + \beta_{14} x_{14} + \beta_{34} x_{34}$	740.801	33.89
M29.1.1	**LASSO with Huber M**	$\boldsymbol{Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \beta_{14} x_{14} + \beta_{134} x_{134}}$	**239.48**	**28.28**

From Table 6, it is clear that LASSO used the Huber M estimator to provide a good forecasting fit as compared to other methods because the MAPE value was smaller than the Huber M estimator compared to other methods. While the weight function for OLS is *1/n*, it means that all observations including outliers are given equal weight as well as ridge regression MAPE is high as compared to others because ridge regression is capable of dealing with multicollinearity but has an effect in the presence of outliers.

As a consequence, five variables remain in the effcient model chosen from LASSO with Huber M with 28.28% MAPE. There are 6 variables in the final model in OLS and ridge regression analysis, but MAPE is greater than the suggested method with a enormous difference in SSE.

By comparing all methods with OLS, it is evident that there is a 46.07% reduction in SSE for Huber M compared to OLS. For all methods used in the analysis, a standardised residual graph is noted as in Figure 3 and Figure 4.

The box plot for an effective model is also observed for all three methods used in the analysis

From box plots in Figure 5, 6 and 7, it is possible to see the outlier detection for each method. It is clear that OLS shows three observations as outlier but there is one observation as outlier for the ridge regression and for Huber's M method.

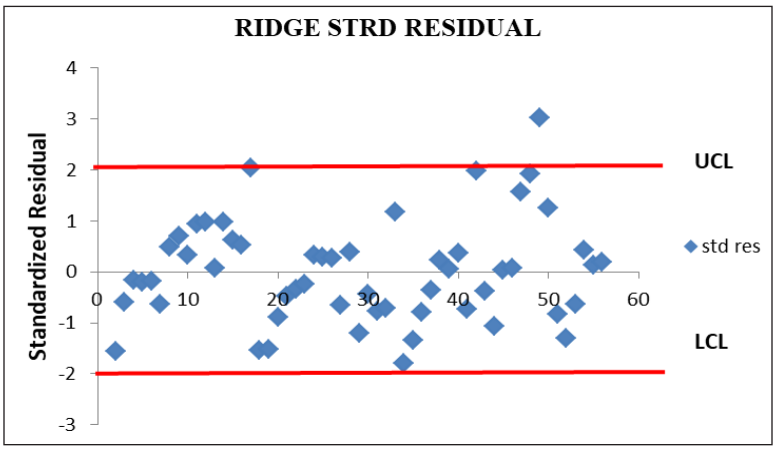

Figure 3. Standardized residual by using ridge regression

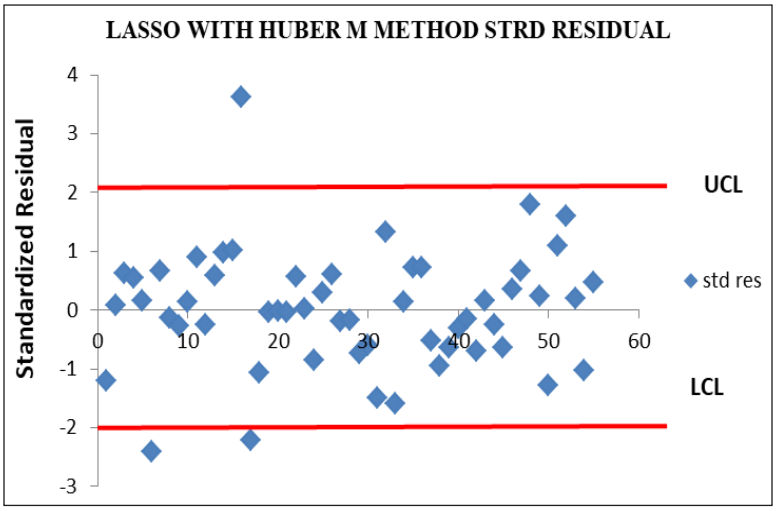

*Figure 4.*Standardized residual using Huber M after LASSO

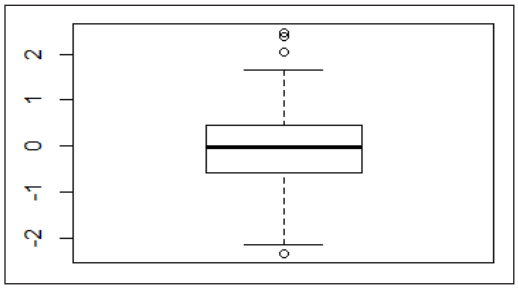

*Figure 5.*Box plot for OLS

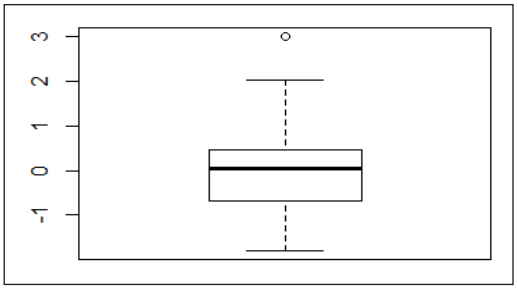

Figure 6. Box Plot for Ridge Regression

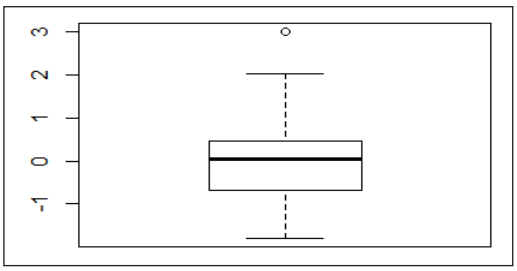

Figure 7. Box plot for Huber M method after LASSO

CONCLUSION

From the above results, it can be concluded that LASSO with the Huber M estimator provides the most efficient model compared to other methods with minimum MAPE. Thus, the best model for forecasting can be chosen by using the significant variables as time, collector average temperature, solar radiation, interaction of time and solar radiation,

interaction of time, collector average temperature and solar radiation. The model is ready to predict the collector efficiency of solar drier. By using 8SC for different types of data, this developed model may also be used in big data analyses.

ACKNOWLEDGEMENT

The authors would like to thank the Women University Multan Pakistan and University Sains Malaysia for the support in this research project.

REFERENCES

Abdullah, N., Jubok, Z., & Ahmed, A. (2011). Improved stem volume estimation using P-value approach in polynomial regression models. *Research Journal of Forestry*, *5*(2), 50- 65.

Ahmed, U. I., Ying, L., Bashir, M. K., Abid, M., & Zulfiqar, F. (2017). Status and determinants of small farming households' food security and role of market access in enhancing food security in rural Pakistan. *PLoSONE*, *12*(10), 1-15.

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, *21*(1), 243-247.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans Auto Control*, *19*, 716-723.

Ali, M. K. M., Sulaiman, J., Yasir, S. M., & Ruslan, M. (2015). The effectiveness of sauna technique on the drying period and kinetics of seaweed Kappaphycus alvarezii using solar drier. *Advances in Environmental and Agricultural Science*, *1*, 86-95.

Ali, M. K. M., 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.

Ali M. K. M., Fudholi, M., Muthuvalu, S., Sulaiman, J., & Yasir, S. M. (2017). Implications of drying temperature and humidity on the drying kinetics of seaweed. *AIP Conference Proceedings*, 1905(1), 1-7.

Beath, K. J. (2018). A mixture-based approach to robust analysis of generalised linear models. *Journal of Applied Statistics*, *45*(12), 2256-2268.

Chen, G. J. (2012). A simple way to deal with multicollinearity. *Journal of Applied Statistics*, *39*(9), 1893-1909.

Dissa, A. O., Bathiebo, D. J., Desmorieux, H., Coulibaly, O., & Koulidiati, J. (2011). Experimental characterisation and modelling of thin layer direct solar drying of Amelie and Brooks mangoes. *Energy*, *36*(5), 2517-252

Gad, A. M., & Qura, M. E. (2016). Regression estimation in the presence of outliers: A comparative study. *International Journal of Probability and Statistics*, *5*(3), 65-72

Giacalone, M., Panarello, D., & Mattera, R. (2018). Multicollinearity in regression: an efficiency comparison between L p-norm and least squares estimators. *Quality and Quantity*, *52*(4), 1831-1859.

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics, 21*, 215-223.

Gujarati, D. N. (2004). Basic econometrics (4th Ed.). New York, USA: The McGraw-Hill Companies.

Guo, P., Liu, T., Zhang, Q., Wang, L., Xiao, J., Zhang, Q., … & Ma, W. (2017). Developing a dengue forecast model using machine learning: A case study in China. *PLoS Neglected Tropical Diseases*, *11*(10), 1-22.

Gusnanto, A., & Pawitan, Y. (2015). Sparse alternatives to ridge regression: A random effects approach. *Journal of Applied Statistics*, *42*(1), 12-26.

Hannan, E. J., & Quinn, B. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*, 190-195.

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

Jang, D. H., & Anderson-Cook, C. M. (2017). Influence plots for LASSO. *Quality and Reliability Engineering International*, *33*(7), 1317-1326.

Khuneswari, G., Zainodin, H. J., Darmesah, G., & Sim, S. H. (2008). An alternative approach in getting a representative model in a multiple regression analysis. *The 3rd international Conference on Mathematics and Statistics (ICoMS-3), 23*(1), 47-64.

Mendelsohn, R., & Dinar, A. (2003). Climate, water, and agriculture. *Land Economics*, *79*(3), 328-341.

Midi, H., Bagheri, A., & Imon, A. H. M. R. (2011). A Monte Carlo simulation study on high leverage collinearity-enhancing observation and its effect on multicollinearity pattern. *Sains Malaysiana*, *40*(12), 1437-1447.

Midi, H., Rana, S., & Imon, A. H. M. R. (2014). Two-step robust estimator in heteroscedastic regression model in the presence of outliers. *Economic Computation and Economic Cybernetics Studies and Research*, *48*(3), 3-8.

Montesinos-Lopez, O. A., Montesinos-Lopez, A., Crossa, J., de los Campos, G., Alvarado, G., Suchismita, M., … & Burgueno, J. (2017). Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data. *Plant Methods*, *13*(1), 1-23.

Neitsch, S. L., Arnold, J. G., Kiniry, J. R., & Wouldiams, J. R. (2011). *Soil and water assessment tool theoretical documentation version 2009*. Berlin, Germany: Springer.

Pender, J., Nkonya, E., Jagger, P., Sserunkuuma, D., & Ssali, H. (2004). Strategies to increase agricultural productivity and reduce land degradation: evidence from Uganda. *Agricultural Economics*, *31*(2-3), 181-195.

Ramanathan, R. (2002). *Introductory econometrics with application* (5th Ed.). Ohio, USA: Thomson Learning.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics*, *12*, 1215-1230.

Rischbeck, P., Elsayed, S., Mistele, B., Barmeier, G., Heil, K., & Schmidhalter, U. (2016). Data fusion of spectral, thermal and canopy height parameters for improved yield prediction of drought stressed spring barley. *European Journal of Agronomy*, *78*, 44-59.

Rockström, J., Falkenmark, M., Karlberg, L., Hoff, H., Rost, S., & Gerten, D. (2009). Future water availability for global food production: the potential of green water for increasing resilience to global change. *Water Resources Research*, *45*(7), 1-16.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Shariff, N. S. M., & Ferdaos, N. A. (2017). An application of robust ridge regression model in the presence of outliers to real data problem. *Journal of Physics: Conference Series Paper*, *890*(1), 1-6.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, *68*(1), 45-54.

Sinova, B., & Van Aelst, S. (2018). Advantages of M-estimators of location for fuzzy numbers based on Tukey's biweight loss function. *International Journal of Approximate Reasoning*, *93*, 219-237.

Stuart, C. (2011). *Robust regression*. Durham, England: Durham University.

Susanti, Y., Pratiwi, H., Sulistijowati H., S., & Liana, T. (2014). M estimation, S estimation, and MM estimation in robust regression. *International Journal of Pure and Apllied Mathematics*, *91*(3), 349-360.

Taylor, J. E., & Adelman, I. (2003). Agricultural household models: Genesis, evolution and extensions. *Review of Economic Households*, *1*(1), 1-44.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, 58*(1), 267-288.

Wouldiams, R., Borges, L. F., Lacoste, M., Andersen, R., Nesbitt, H., & Johansen, C. (2012). On-farm evaluation of introduced maize varieties and their yield determining factors in East Timor. *Field Crops Research*, *137*, 170-177.

Xu, J., & Ying, Z. (2010). Simultaneous estimation and variable selection in median regression using lasso-type penalty. *Annals of the Institute of Statistical Mathematics*, *62*(3), 487-514.

Yan, E. D. (2011, March 28-29). Design of intelligent agriculture management information system based on IoT. In *4th International Conference on Intelligent Computation Technology and Automation* (pp. 1045-1049). Shenzhen, Guangdong, China.

Zainodin, H. J., Noraini, A., & Yap, S. J. (2011). An alternative multicollinearity approach in solving multiple regression problem. *Trends in Applied Sciences Research*, *6*(11), 1241-1255.

Zhang, K., Zhe, S., Cheng, C., Wei, Z., Chen, Z., Chen, H., … & Ye, J. (2016, August 13-17). Annealed sparsity via adaptive and dynamic shrinking. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1325-1334). San Francisco, California, USA.

Zhao, Y., Ogden, R. T., & Reiss, P. T., (2012). Wavelet-based LASSO in functional linear regression. *Journal of Computational and Graphical Statistics*, *21*(3), 600-617.

Zou, H., (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418-1429.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M., (2009). Limitations of linear regression applied on ecological data. In A. F. Zuur, E. N. Ieno, N. J.Walker, A. A. Saveliev & G. M. Smith (Eds.), *Mixed effects models and extensions in ecology with R* (pp. 11-33). New York, USA: Springer.