

PAPER • OPEN ACCESS

Graphical user interface (GUI) for the least absolute shrinkage and selection operator (LASSO) regression

To cite this article: G A Dito *et al* 2019 *IOP Conf. Ser.: Earth Environ. Sci.* **299** 012031

View the [article online](#) for updates and enhancements.

Graphical user interface (GUI) for the least absolute shrinkage and selection operator (LASSO) regression

G A Dito¹, A Safitri¹, F M Afendi¹, R Anisa¹, A Salim² and B Sartono^{1*}

¹Department of Statistics, IPB University, Bogor 16680, Indonesia

²Department of Mathematics and Statistics, La Trobe University, Melbourne, 00115M, Australia

*E-mail: bagusco@gmail.com

Abstract. In Data Science, we usually encounter High-dimensional data. In this situation, the Classical Regression method usually cannot perform well because it is impossible to include all covariates in the model since the number of a parameter to be estimated is larger than the sample size. Least absolute shrinkage and selection operator (Lasso) method is one of the methods which can deal with this problem. Lasso regression perform the selection of covariates so that only the most influential covariates are used in the model. Unfortunately, most of Lasso method should be performed in CLI Software which is difficult to use for the general user. For this reason, we develop a web application by using Shiny to perform the Lasso method based on GUI which is easier to use. It allows users to analyze high-dimensional data without using programming language. The paper contains an implementation of Lasso Regression using web application on olive pomade oil data.

1. Introduction

Growth of data is significantly increased during this recent year. One of the problems arises is the dataset has more columns than its row, this problem is called high-dimensional data. This type of data unable to analyze by a classical Regression model. The reason is that the classical model unable to estimate a parameter which its number is larger than data. It is possible to solve this issue by doing a selection of predictor variables. Classical methods, such as forward selection, may give decent result but not efficient in implementation time. Lasso method is widely preferable than forwarding selection as it is able to perform the faster implementation. Lasso was coined and popularized by [1]. Since the first publication, the development of the model is rapidly increasing. One major concern is to find an efficient algorithm for computing the lasso method. A Notable result was achieved by [2], which find an efficient algorithm which outperforms its predecessor. This algorithm is implemented on a package in R Software, called glmnet. R is a software-based Command Line Interface(CLI) which means it uses programming languages. It is uneasy for general users to use glmnet because they require to learn the R language. For this reason, it is necessary to implement glmnet to become user-friendly. Software-based Graphical User Interface (GUI) mostly likable for general users because it only uses point-click features on a computer. This study is about the implementation of software-based GUI which is based on the glmnet package. This software was made by using a shiny package in R. Real dataset is used to illustrate steps for implementing this software.

Materials



Table 1. Definition of materials for the software.

No.	Material	Definition
1.	R 3.5.1	R is a language and environment for statistical computing and graphics. R is free and open source software. R 3.5.1 is R version that was released in August 2018.
2.	R Studio	R Studio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, and tools for plotting, history, debugging and workspace management.
3.	Shiny Packages	Shiny is an open-source R package that provides an elegant and powerful Web framework for building Web applications using R. Shiny helps turn your analyses into interactive Web applications without requiring HTML, CSS, or JavaScript [3].
4.	Glmnet Packages	Glmnet is a package that fits a generalized linear model via penalized maximum likelihood. The regularization path is computed for the lasso or elastic net penalty at a grid of values for the regularization parameter lambda. The algorithm is extremely fast and can exploit sparsity in the input matrix X. It fits linear, logistic and multinomial, Poisson, and Cox regression models. A variety of predictions can be made from the fitted models. It can also fit multi-response linear regression [4].
5.	Data	Data that can be analyzed using the GUI is the data with CSV extension. In this paper using olive pomade oil data. The data contains 18 observation or experimental runs, 31 potential influencing factors with 2 level, and the reaction yield of the whole process as the response variable [5].

2. Method

**Figure 1.** GUI development process.

At the design stage, the layout, features, and outputs of GUI-based software which can analyze any possible problems are planned. At the implementation stage, GUI-based software is designed using Shiny package. At the trial stage, the high dimensional data is analyzed using GUI-based software. At the publication stage, GUI-based software can be accessed by the public with uploading on the internet.

Table 2. Potential influencing factors.

No. Factors			Unity	Level (-1)	Level (+1)
1.	X1	Soap quality		1	2
2.	X2	Hydrolysis acid		H ₂ SO ₄	H ₃ PO ₄
3.	X3	Acid concentration	%	75	85
4.	X4	Hydrolysis temp.	°C	75	85
5.	X5	Hydrolysis time	Min	60	90
6.	X6	Fatty acids separation		Without NaCl	With NaCl
7.	X7	Rinse temp.	°C	70	95
8.	X8	Number of rinses		2	3
9.	X9	Esterification temp.	°C	115	125
10.	X10	Esterification time	H	3	5
11.	X11	Catalyst addition		After 15 min	At the start
12.	X12	Catalyst amount	%	0.2	0.4
13.	X13	MeOH amount	ml	55	75
14.	X14	MeOH quality		Absolute	HPLC
15.	X15	MeOH addition rate	ml/min	0.3	0.5
16.	X16	Stirring speed 1	Tr/min	200	400
17.	X17	Controlled Atmosphere		Air	Nitrogen
18.	X18	Amidation temp.	°C	120	130
19.	X19	Amidation time	H	10	12
20.	X20	Sodium methanoate amount	%	0.4	0.5
21.	X21	Molar ratio amine/ester		1.2	1.6
22.	X22	Methyl ester addition		At the start	After 15 min
23.	X23	Stirring speed 2	tr/min	200	400
24.	X24	Molar ratio SO ₃ /ester		1.8	2
25.	X25	Sulfation temp.	°C	5	15
26.	X26	Sulfation time.	H	2	3
27.	X27	Amide addition rate	g/min	0.667	0.333
28.	X28	Alkali reagent		NaOH	KOH
29.	X29	Alkali concentration	%	20	30
30.	X30	Neutralization temp.	°C	5	15
31.	X31	Neutralization time	Min	30	60

GUI-based software can perform Lasso regression. Lasso Regression is a regression method that minimizes the least squares and has an additional penalty. The lasso shrinks the coefficient estimates

towards zero, $\hat{\beta}_{\lambda}^L$, minimize the quantity.

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| = RSS + \lambda \sum_{j=1}^p |\beta_j|$$

The lasso uses an ℓ_1 (ell 1) penalty. The ℓ_1 norm of coefficient vector β is given by $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$. When the tuning parameter $\lambda = 0$, then the lasso simply gives the least squares fit, and when λ becomes sufficiently large, the lasso gives the null model in which all coefficient estimates to be exactly equal to zero. Lasso can be called as a method of selecting variables when λ is sufficiently large. Therefore, depending on the value of λ , the lasso can produce a model involving any number of variables or all variables [1].

GUI-based software can analyze the high dimensional data using Lasso Regression by performing these procedures:

- 1) Open website <https://grivz.shinyapps.io/PKLM/>
- 2) Upload data

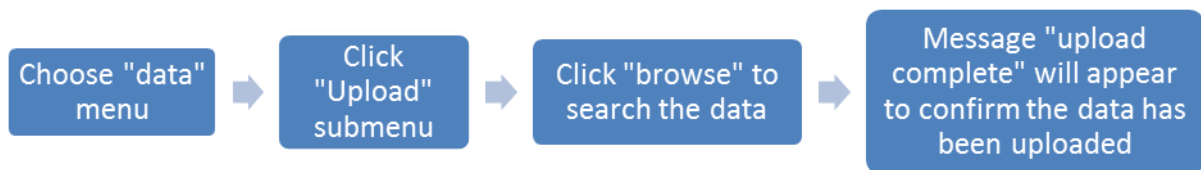


Figure 2. Procedure to upload the data.

- 3) Statistics descriptive analysis is an optional stage on "data" menu :
 - a. Choose the "dataset" menu to view the data.
 - b. Choose "summary" menu to know information about summary of the data such as number of observations, mean, standard deviation, Quartile 1 (Q1), Median, Quartile 3, (Q3) Interquartile Range (IQR), minimum observation, maximum observation, number of missing observation and number of outlying observations.
 - c. Choose the "graph" menu to visualize the data or shows graphics output such as scatterplot, boxplot, and histogram. The graph will appear when the response variable and the predictor variable have been selected.
- 4) Lasso Regression analysis.

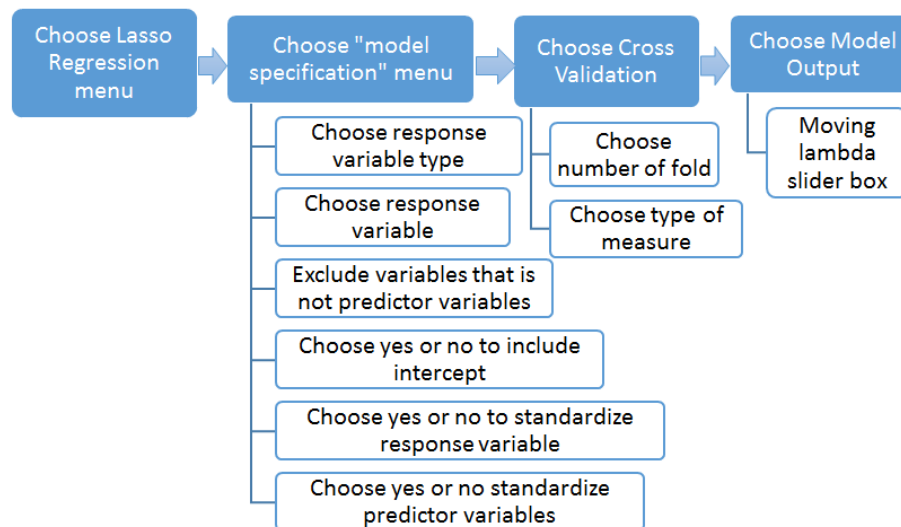


Figure 3. Lasso regression analysis procedure.

Response variable type consists of quantitative response, binary response, count response, multiclass response, and multi-quantitative response. The quantitative response is numerical, measurable, and continuous variable. The binary response is a categorical variable which has only two categories. The count response is a countable variable. The multiclass response is a categorical variable which has more than two categories. The multi-quantitative response is the quantitative response which has more than one variable. "Cross-Validation" menu to show the cross-validation plot. It used for finding optimal lambda value to perform LASSO Estimation. The vertical line shows optimal lambda value and the number of chosen variables. The fold is a number between 1 until 100 and type of measure consists of mean squared error, mean absolute error, deviance for binary or count response variable, the class for the multiclass response variable, and area under the curve for the binomial response variable. The default number of folds is 10, while the default type of measure is Mean Squared Error. "Model Output" menu to show about the result of Lasso Regression. "Model Output" menu displays lambda slider box and bar chart. The Default lambda slider box shows optimal lambda (Lasso Parameter). Lambda values can change by moving lambda slider to right or left. A higher value of lambda which can be applied by moving lambda slider to right cause fewer variables are selected. The bar chart shows the coefficients of the selected variables. Changing lambda value can influence the number and coefficient of selected variables, which show on the bar chart. The bar chart has two colors. First, the dark blue color represents a positive value of coefficients and light blue color represents a negative value of coefficients.

3. Results and Discussion

This section describes the implementation of Lasso Regression using GUI-based software. The final software designed can be viewed in Figure 4. The software layout consists of 4 main menus which are Home, Data, Lasso Regression, and Developers. The main menus which are used in for implementation purpose are Data and Lasso Regression.

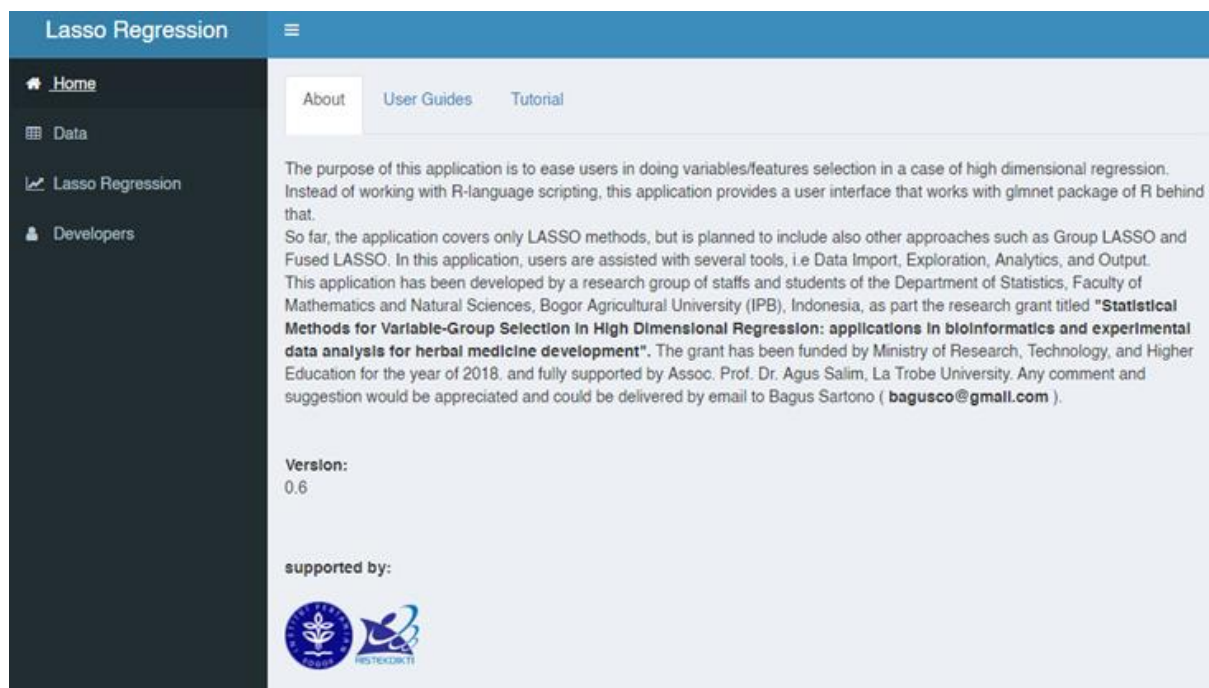


Figure 4. The result of GUI-based Software design.

The implementation uses the dataset which has more column (variable) than its row. The data is about Chemical experiments which involve olive pomace oil converted into sulfated fatty amides. In this experiment, the reaction yield can exceed 100% against the pure fatty amides (more than one mole of sulfate can be linked to one mole of amide), but in some cases, this yield is lower than 50% due to poor choices in experimental conditions [5]. The identification of factors (variables) which have influences on the reaction yield is a primary goal for the data. This identification can be achieved by performing Lasso Regression with GUI-based software. The following steps are performed to analyze the data:

1) Input data

Data can be input by choosing "data menu" and then click "browse". Choose the file from the computer directory.

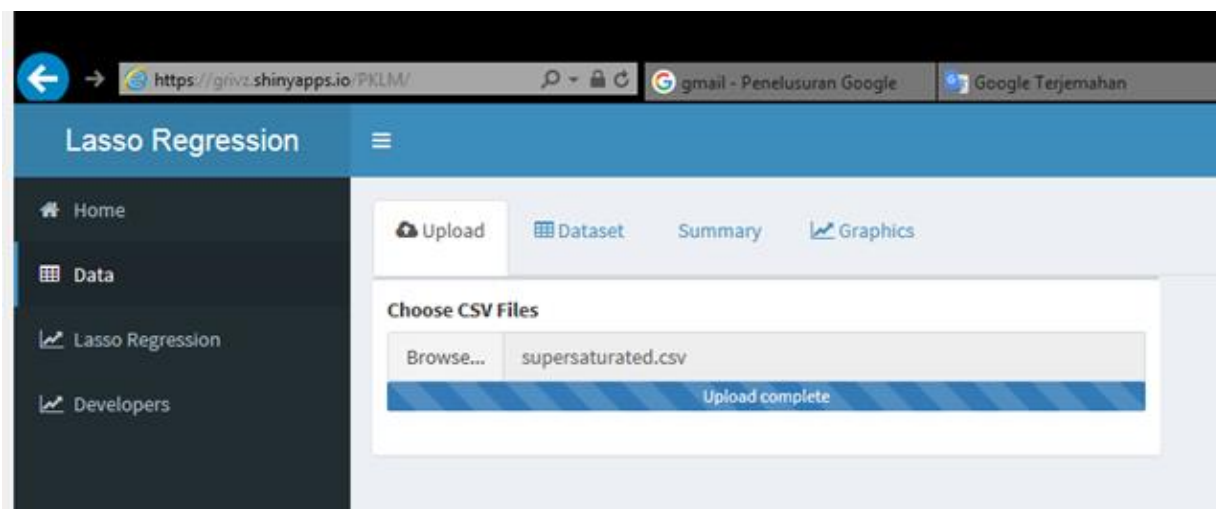


Figure 5. Uploading data.

Lasso Regression

Home

Data

Lasso Regression

Developers

Upload

Dataset

Summary

Graphics

Show 10 entries

Search:

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
1	-1	1	-1	1	1	1	-1	-1	-1	1	1	1	1	1	-1	1	1	1	-1	
2	1	-1	-1	1	-1	1	1	1	-1	-1	-1	1	1	1	1	1	-1	1	1	
3	-1	1	-1	-1	1	-1	1	1	1	-1	-1	-1	1	1	1	1	1	-1	1	
4	1	1	-1	-1	1	-1	-1	1	-1	1	1	1	-1	-1	-1	1	1	1	1	
5	1	-1	1	1	-1	-1	1	-1	-1	1	-1	1	1	1	-1	-1	-1	1	1	
6	1	-1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	1	1	1	-1	-1	-1	
7	-1	1	-1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	1	1	1	-1	-1	
8	-1	-1	1	-1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	1	1	1	-1	
9	-1	-1	-1	1	-1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	1	1	1	
10	1	-1	-1	-1	-1	1	-1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	1	

<

Showing 1 to 10 of 18 entries

Previous

1

2

Next

Figure 6. Data in tabular form.

2) Data exploration

GUI-based software can be used to explore data. In the "data" menu, dataset panel is used to explore data in tabular form, and the Summary panel is used to perform descriptive statistics and identify the number of missing and outlier observations. The Last Panel is a graphics panel which can be used to visualize data. Graphics panel performs scatterplot, box plot and bars chart simultaneously.

Variable	N	Mean	Sd	Q1	Median	Q3	IQR	Min	Max	N.missing	N.outlier
1 X1	18	0	1.03	-1	0	1	2	-1	1	0	0
2 X2	18	0	1.03	-1	0	1	2	-1	1	0	0
3 X3	18	0	1.03	-1	0	1	2	-1	1	0	0
4 X4	18	0	1.03	-1	0	1	2	-1	1	0	0
5 X5	18	0	1.03	-1	0	1	2	-1	1	0	0
6 X6	18	0	1.03	-1	0	1	2	-1	1	0	0
7 X7	18	0	1.03	-1	0	1	2	-1	1	0	0
8 X8	18	0	1.03	-1	0	1	2	-1	1	0	0
9 X9	18	0	1.03	-1	0	1	2	-1	1	0	0
10 X10	18	0	1.03	-1	0	1	2	-1	1	0	0

Figure 7. Summary data.

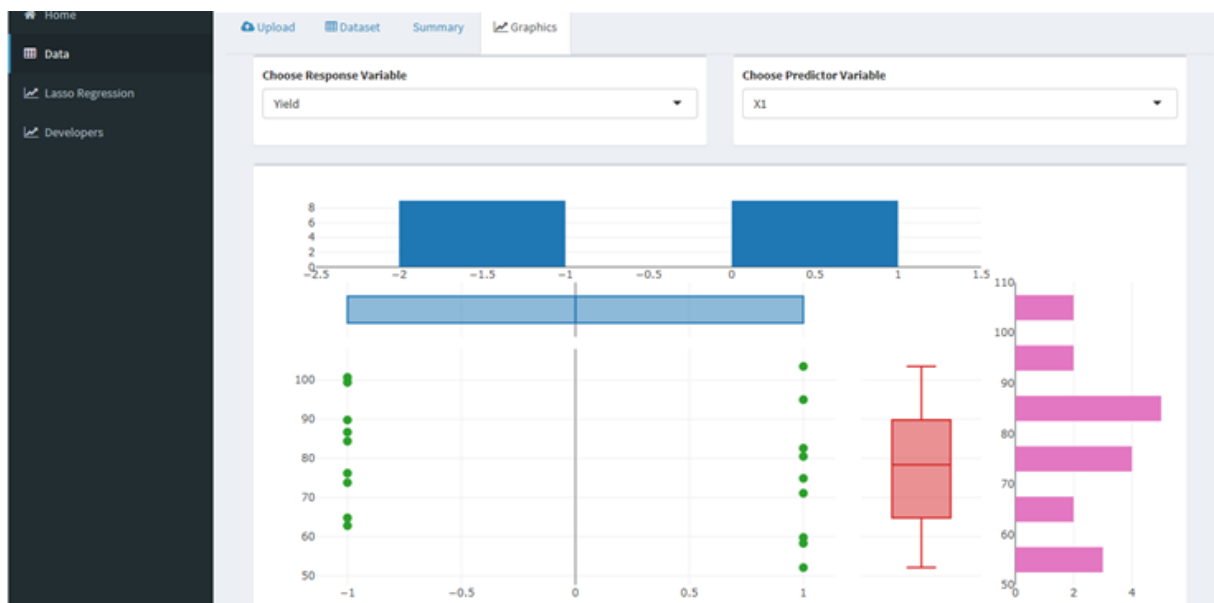


Figure 8. Data visualization.

3) Lasso regression

“Lasso Regression” menu has three panels. The first panel is the model specification panel. In this panel, choose a quantitative type of response because the reaction yield is numeric. Then, in the second panel, the cross-validation plot is used to identify optimal lasso parameter. The dashed line shows the number of selected predictors. Optimal lasso parameter can be chosen based on minimum MSE or a minimum confidence interval of MSE. Finally, the third panel shows a bar chart of selected predictors which its values show an estimated coefficient of predictors. Based on figure 11, there are 14 factors which have a strong influence on reaction yield.

Figure 9. Model specification.

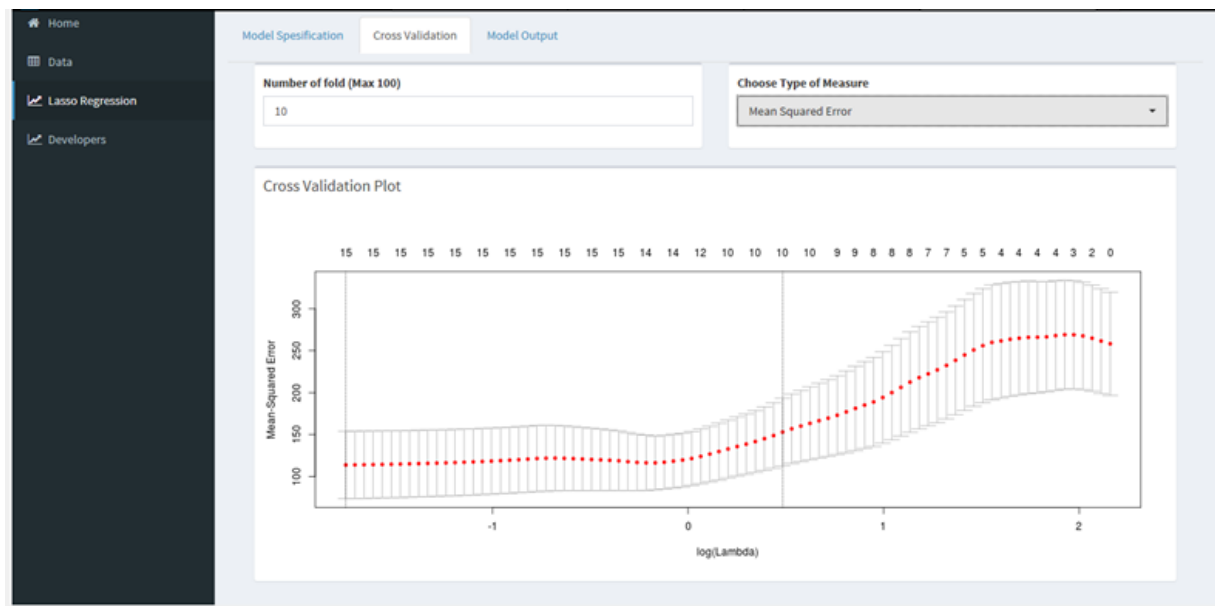


Figure 10. Cross-validation.

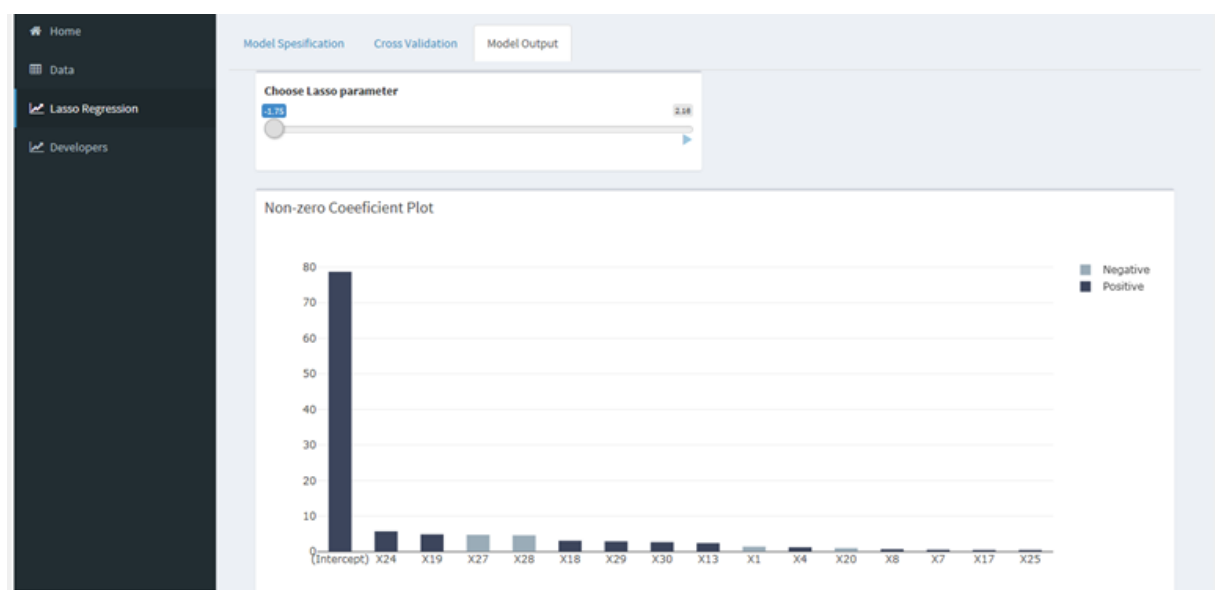


Figure 11. D Model output.

4. Conclusion

GUI-based Software is very intuitive to use for general users. Lasso Regression analysis becomes more user-friendly to implement rather than using CLI-based Software such as R. Besides performing Lasso Regression analysis, the software also can be used for visualization and exploration of the data. This software still under development, for future release more model for analyzing high-dimensional data will be added.

References

- [1] Tibshirani R 1996 *Journal of the Royal Statistical Society, Series B* **58** 267.
- [2] Friedman J, Hastie T and Tibshirani R 2010 *Journal of Statistical Software* **33** 1.
- [3] Gunuganti A 2018 *Application development framework for r/shiny* online:

- [4] <https://www.pharmasug.org/proceedings/2018/AD/PharmaSUG-2018-AD24.pdf>
- [5] Hastie T and Qian J 2014 *Glmnet vignette* online:
- [6] https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html
- [7] Rais F, Kamoun A, Chaabouni M, Claeys-Bruno M, Phan-Than-Luu R and Sergent M 2009 *Chemometrics and Intelligent Laboratory Systems* **99** 71.