FCU ePaper

Feng Chia University Outstanding Academic Paper by Students

Regression variable selection using LASSO algorithm

藉由 LASSO 演算法選擇迴歸變數

Author(s): Wei-Lun Li

Class: 1nd year of Department of Statistics

Student ID: M0825429

Course: Statistical Computing

Instructor: Dr. Cathy W.S. Chen

Department: Institute of Statistic

Academic Year: Semester 2, 2019

Abstract

This study considers using the LASSO (least absolute shrinkage and selection operator) method to select important independent variables in linear regression models. The LASSO method is a supervised learning algorithm. By constraining the properties of the residuals, an interpretable regression model is developed, and the constrained parameters are set to be zero. This report discusses how to appropriately select important variables via a simulation study when there are too many independent variables or even the number of independent variables greater than the number of sample size. We use simulated data in 500 replications to show how the LASSO method selects important variables of the regression model. Finally, the results of simulation data are provided, which show that almost all important variables and half of the correct combination of variables can be accurately selected.

Keyword： LASSO; big data; linear regression; machine learning; variable selection.

摘要

在許多數據集中，包含的變量動輒上百甚至更多，這使得我們必須適當選取變量以及降維技術，使得我們可以在最大程度發揮模型的解釋能力。本研究考慮使用 LASSO（最小絕對收縮和選擇算子）方法在線性回歸模型中選擇重要的迴歸變數。 LASSO 模型是一種監督學習算法。通過約束殘差的屬性，提出可解釋的迴歸模型，並將約束參數設置為零。本報告討論了當自變量過多，甚至迴歸變量的數量大於樣本數量時，如何通過模擬研究適當選擇重要變量。我們重複 500 次中的模擬數據來顯示 LASSO 模型如何選擇迴歸模型的重要變量。最後，提供了模擬數據的結果，表明幾乎可以準確選擇所有重要變量和變量正確組合的一半。

關鍵字: LASSO、大數據，線性迴歸、機器學習、變數選取

Table of Contents

目錄

1 Introduction

With the advent of big data, data acquisition is more convenient than ever, but it has also created large-scale data. The big data contained in huge amounts of data is beyond the capacity of traditional software. Due to recent technological advancements, the convenience of data release, and the transparency of data, big data and machine learning have attracted attention. In many data sets, hundreds or more variables are included, which makes it necessary to select variables and dimensionality reduction techniques appropriately to maximize the explanatory power of the model.

Big data makes machine learning and statistics a popular subject, and it also makes the development of big data tools faster, easier to obtain and use. E-commerce giant Amazon uses big data to predict customer behavior, significantly reducing logistics and warehousing costs. U.S. restaurant review website Yelp also provides a recommendation system for consumers to increase consumption through big data.

Machine learning is an algorithm that automatically analyzes and obtains rules from data, and uses the rules to predict unknown data. Machine learning is divided into supervised learning, unsupervised learning and semi-supervised learning. LASSO (least absolute shrinkage and selection operator) method proposed by Tibshirani (1996) is an algorithm in supervised learning. The LASSO method is one of the methods of variable selection in the regression model. Many variable selection methods exist for regression model such as forward selection method, backward elimination method, stepwise method and C_p, but when there are a large number of variables, there are too many combinations of regression models which often takes up a lot of time and energy. In other words, when we face with sparsity and p >> n problems, it is impossible to use any of the above-mentioned methods to select the important variables. The LASSO method improves this problem and can effectively select model variables.

The LASSO method improves general regression model by forcing the parameters to return to 0 through the restriction, and effectively selects the model parameters to make the model simpler. This article mainly discusses how to use R package,"glmnet", to select variables based on the LASSO method and provides results through simulation data in 500 replications wherein almost all important variables can be correctly selected.

2. The LASSO Method

In the general setting, the linear model is as follows

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} \qquad i = 1,2,\ldots,n \ ,$$

where $\boldsymbol{Y} = (\ y_1, y_2, \ldots, y_n\ )^T$ is a response variable and $\boldsymbol{X}_1 = (\ \mathrm{x}_{11}, \mathrm{x}_{12}, \ldots, \mathrm{x}_{1n}\ )^T \ldots\ \boldsymbol{X}_p = (\ \mathrm{x}_{p1}, \mathrm{x}_{p2}, \ldots, \mathrm{x}_{pn}\ )^T$ are independent variables.

Regression variable selection is an important step for building a regression model because, in the process of selecting variables, we have a total of 2^p combinations. Therefore, we aim at how to choose important variables to interpret the model and find an appropriate subset of independent variables.

Regarding regression variable selection, the common methods are stepwise selection, forward selection, backward elimination, all subset methods (e.g. adjusted R squares) and C_p. When we deal with sparsity and p >> n problems, it is impossible to use any of the above-mentioned methods to select the important variables. There are insufficient degrees of freedom to estimate the full model when there are too many variables or even when these exceed the number of samples. LASSO method is a relatively new alternative which overcomes this shortcoming.

In the linear regression model, we usually use the least square method to estimate our parameter $\beta_0, \beta_1, \ldots, \beta_p$ using the value that minimize formula (1)

$$\mathrm{RSS} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{y_i} - \beta_0 - \sum_{\mathrm{j}=1}^{\mathrm{p}} \beta_\mathrm{j} \mathrm{X_{ij}} \right)^2 . \qquad \textbf{(1)}$$

James et al. (2013) uses the constraints of the LASSO method to minimize the number of coefficients and effectively select our model variables as follows

$$\sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{y_i} - \beta_0 - \sum_{\mathrm{j}=1}^{\mathrm{p}} \beta_\mathrm{j} \mathrm{X_{ij}} \right)^2 + \lambda \sum_{\mathrm{j}=1}^{\mathrm{p}} |\beta_\mathrm{j}| = \mathrm{RSS} + \lambda \sum_{\mathrm{j}=1}^{\mathrm{p}} |\beta_\mathrm{j}| \ . \qquad \textbf{(2)}$$

In formula (2), a penalty term λ is added to limit coeficcient estimates β_j towards zero, and when the turning parameter λ is large enough, the penalty term can force some parameter estimates to be exactly equal to zero.

3. Simulation

In this section, we consider to select regression variables and use simulation data to present the results when the variables are greater than the number of samples. Consider the sample size N=100, number of variables P=200, independent variables $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{200}$ from the normal distribution. There are 1.60694E+60 (2^{200}-1) combination of subset variables. We only allow 10 variables to be the important variables. We use the LASSO method to solve sparsity and p >> n problems.

Randomly select 10 variables as the important variables and let the parameter $\boldsymbol{\beta}$ of these 10 variables come from the uniform distribution from -5 to -1 and 1 to 5 while the parameters of the remaining variables are zero. Our response variable **Y** is equal to **X** times beta and plus an error term, where the error term comes from the T distribution with a degree of freedom of 5. Under the simulation setting, we use LASSO method to select our important variables in 500 replications.

Step1:

Generate data containing 200 variables **X** from the normal distribution, where the first column is fixed to 1 in order to allow β_0 to be presented. We use "rmvnorm" in package "mvtnorm" to generate data from the normal distribution, with the number of samples "n" equal to 100, the mean range between 0 and 200 and the method "svd".

```
library("mvtnorm")
n <- 100
x=matrix(0,100,201)
x[,1]=1
x[,2:201]=rmvnorm(n = n, mean = rep(0,200), method = "svd")
```

Table 1 shows the generated data **X** with sample size N equal to 100 and number of variables P equal to 200 plus a constant term. From this, randomly select 10 variables from 1 to 200 variables as the selected variables with beta ranging from -5 to -1 and 1 to 5.

Table 1. Generated data (X)

N				
1	1	X_1	X_2	$\ldots\ldots X_{200}$
2	1	X_1	X_2	$\ldots\ldots X_{200}$
.	1	X_1	X_2	$\ldots\ldots X_{200}$
.				
.				
100	1	X_1	X_2	$\ldots\ldots X_{200}$

To randomly select 10 important variables from the 200 variables, use "sample" and set "replace=F" so that parameters do not repeat. To include the constant term, set the dimension "p" to be equal to 201 and let "s" be the variable selected randomly.

Next, to determine the range of the parameters, set the range between -5 to -1 and 1 to 5. Here, 0.1 is used to divide the range, and then use "sample" to determine the parameters.

```
p <- 201
s <- c(1,sort(sample(2:201,10,replace = F)))
beta=rep(0,p)
pos=seq(1,5,by=0.1)
neg=-seq(1,5,by=0.1)
b=c(neg,pos)
beta[s]=sample(b,11,replace = T)
```

Step2:

Let Y be equal to X times beta plus an error term from the T distribution with 5 degrees of freedom, which is then fitted with the LASSO method to find the selected variable beta. Use the R package "glmnet" to help find the best variable selected. Separately record the number of times each variable has been selected, and the combined result of the variable.

```
library("glmnet")
y = x %*% beta + rt(n,df=5)
fit2=glmnet(x,y,alpha=1,pmax=10,standardized=T,intercept=T)
b=coef(fit2,s=0.001)
```

"fit2" is our first regression model, "alpha" set to 1 means the LASSO method is used, "pmax" means to find the number of important variables in the model which is set here to 10. In addition, "standardized=T" is used to standardize the data and "intercept=T" to include the constant term as our important variable. Finally, extract all the variables with "coef".

It is easier to record the number of variable selections separately, but relatively difficult to record the result of the combination. Therefore, when recording the results of each combination, convert the numbers presented in the results into "string" to extract the variables in sequence and consequently record them together.

Here loops pick out each variable first and then convert them to string form one by one before connecting them together which becomes the combined result. Under the definition of string space "d", we use loops to extract the selected variables "b@i", respectively, and use "paste" to link them together to obtain the combined result.

```
d=as.character()
for (v in 1:length(b@i)) {
  q=b@i[v]
  d=paste(d,q)
}
```

Step3:

Finally, repeat the above steps 500 times to calculate the number of times the variables have been selected and determine which group of variables has been selected the most. Compare the difference between the results of 500 simulations and the variables originally selected.

The ten true independent variables originally selected are (0, 9, 32, 33, 35, 59, 66, 84, 94, 106, 199). Comparing with the first simulation result of the LASSO method, the results are as follows.

Table 2. Result of variable selection based on one simulated dataset

True independent Variables	β_0	β_9	β_{32}	β_{33}	β_{35}	β_{59}	β_{66}	β_{94}	β_{106}	β_{199}	β_{84}
	-1.7	2.6	-1.9	2.8	-2.9	3.8	-3.3	3.7	-1.1	-2.4	1.4
One simulated dataset	β_0	β_9	β_{32}	β_{33}	β_{35}	β_{59}	β_{66}	β_{94}	β_{106}	β_{199}	β_{117}
	-1.4	1.14	-0.8	2.05	-1.1	2.2	-1.9	2.6	-0.3	-1.4	-1.3

Table 2 shows that the variables selected through the LASSO method are slightly different from those set at the beginning. Observe that only β_{84} from the true independent variables is not selected, and β_{117} is selected instead. It can be clearly

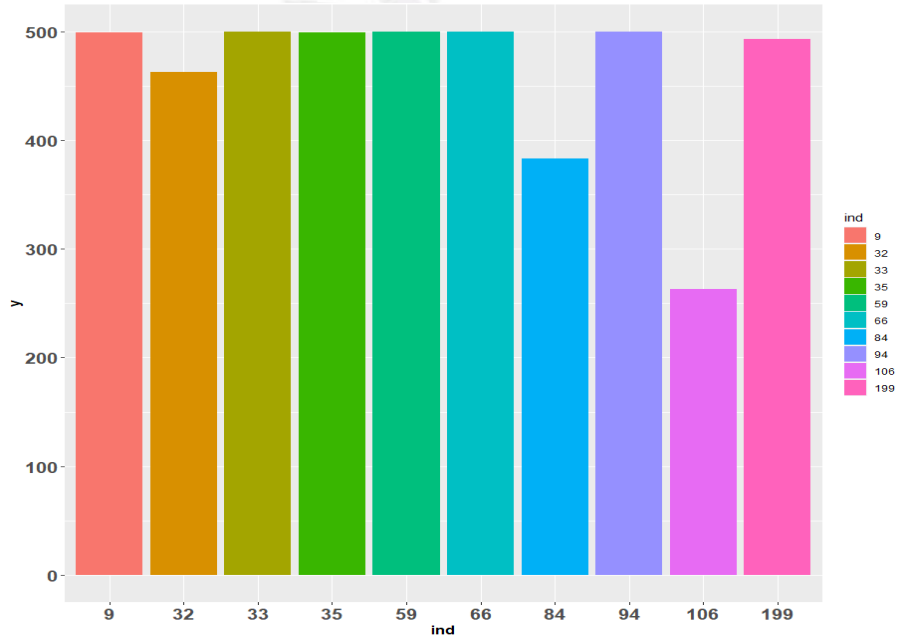
seen that when the variables selected by LASSO method are consistent with the true variables, the parameter estimates are close, and when the selected variables do not include the true variables, the error is obviously different. Table 3 shows that the most of correct rates for individual parameter are greater than 92% except for two variables, β_{84} and β_{106}, whose correct rates are 76.6% and 52.6%, respectively. Nevertheless, these two variables have been selected more than half the number of times.

Table 3. The proportional result of the number of times each variable is selected

	β_9	β_{32}	β_{33}	β_{35}	β_{59}
True Independent	99.8%	92.6%	100%	99.8%	100%
Variables Rate	β_{66}	β_{84}	β_{94}	β_{106}	β_{199}
	100%	76.6%	100%	52.6%	98.6%

Figure1 exhibits the frequency of each variable that have been selected by the LASSO method based on 500 replications. The X axis represents the real variable and the Y axis represents the number of times each variable is selected. Figure 1 shows that most of the important variables are correctly selected.

Figure 1. The number of times each variable has been selected



Finally, a comparison of the number of selected model variables for each group based on the simulation study. There are 233 combinations of results for 500

replications. Three highest frequencies are listed in Table 4. The highest relative frequency is 41%, which is the true model, the second and third best models are the subsets of the true model. It is interesting to note that the seond best model only excludes β_{106} for the true paramters and the third best model only lacks of β_{84}.

Table 4. The result of combination of selected variables in 500 times

True independent Variables 0 9 32 33 35 59 66 84 94 106 199		
Simulated dataset in 500 replications	Frequnecy	Relative Frequency
0 9 32 33 35 59 66 84 94 106 199	205	41%
0 9 32 33 35 59 66 84 94 199	22	4.4%
0 9 32 33 35 59 66 94 106 199	6	1.2%

The results show that the number of correct selection of our true important variables is the highest, so through the LASSO method it is indeed possible to select our important variables when the variables are greater than the number of samples.

4. Conclusion

We provide the simulation results of the LASSO method. When there are too many variables or even when these are greater than the number of samples, we compare the parameter estimates and we obtain effective variable selection. The results also clearly show that when the variables are greater than the number of samples, our important variables can be selected through the LASSO method. However, we only consider the situation where the variable is a numerical value. When the variable is categorical, the problem of whether an accurate variable selection can be obtained may be considered to extend the discussion in the future.

References

Braun, W. J., & Murdoch, D. J. (2016). *A first course in statistical programming with R*. Cambridge University Press.

Friedman J, Hastie T, Tibshirani R (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, **33**(1), 1–22. http://www.jstatsoft.org/v33/i01/.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Tibshirani, R. (1996). "Regression shrinkage and selection via the LASSO." *Journal of the Royal Statistical Society: Series B (Methodological),* 267-288.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.