Video 138: R lab - Subset selection

Arnab Chakraborty
Scribe-Arghya sarkar

October 2022

1 Introduction

As we discussed earlier subset selection method is not popular because of more useful methods like, Ridge regression or LASSO. But we still demonstrate that in R.

2 R code

For these we will need ’leaps’ package and we will be using ”state” data (about average life expectancy is various states of USA).

```
#Installing ’leap’ package
> install.packages("leaps")
> library(leaps)

#Preparing the data
> data(state)
> statedata=data.frame(state.x77,row.names = state.abb)
> names(statedata)
[1] "Population" "Income"     "Illiteracy" "Life.Exp"   "Murder"     "HS.Grad"
[7] "Frost"      "Area"
```

At this point we print first few rows which will give some idea about the data.

```
> head(statedata)
   Population Income Illiteracy Life.Exp Murder HS.Grad Frost   Area
AL       3615   3624        2.1    69.05   15.1    41.3    20  50708
AK        365   6315        1.5    69.31   11.3    66.7   152 566432
AZ       2212   4530        1.8    70.55    7.8    58.1    15 113417
AR       2110   3378        1.9    70.66   10.1    39.9    65  51945
CA      21198   5114        1.1    71.71   10.3    62.6    20 156361
CO       2541   4884        0.7    72.06    6.8    63.9   166 103766
> dim(statedata)    #size of data
[1] 50  8
```

The function we will use is ”regsubsets”. And its usage is similar to ’lm’ function.

```
> b=regsubsets(Life.Exp~.,data=statedata)
> rs=summary(b)
> rs
Subset selection object
Call: regsubsets.formula(Life.Exp ~ ., data = statedata)
7 Variables  (and intercept)
           Forced in Forced out
Population     FALSE      FALSE
Income         FALSE      FALSE
Illiteracy     FALSE      FALSE
Murder         FALSE      FALSE
HS.Grad        FALSE      FALSE
Frost          FALSE      FALSE
Area           FALSE      FALSE
1 subsets of each size up to 7
Selection Algorithm: exhaustive
         Population Income Illiteracy Murder HS.Grad Frost Area
1  ( 1 ) " "        " "    " "        "*"    " "     " "   " " 
2  ( 1 ) " "        " "    " "        "*"    "*"     " "   " " 
3  ( 1 ) " "        " "    " "        "*"    "*"     "*"   " " 
4  ( 1 ) "*"        " "    " "        "*"    "*"     "*"   " " 
5  ( 1 ) "*"        "*"    " "        "*"    "*"     "*"   " " 
6  ( 1 ) "*"        "*"    "*"        "*"    "*"     "*"   " " 
7  ( 1 ) "*"        "*"    "*"        "*"    "*"     "*"   "*" 
```

This is very difficult to interpret so we do the following.

```
> rs$which
  (Intercept) Population Income Illiteracy Murder HS.Grad Frost  Area
1        TRUE      FALSE  FALSE      FALSE   TRUE   FALSE FALSE FALSE
2        TRUE      FALSE  FALSE      FALSE   TRUE    TRUE FALSE FALSE
3        TRUE      FALSE  FALSE      FALSE   TRUE    TRUE  TRUE FALSE
4        TRUE       TRUE  FALSE      FALSE   TRUE    TRUE  TRUE FALSE
5        TRUE       TRUE   TRUE      FALSE   TRUE    TRUE  TRUE FALSE
6        TRUE       TRUE   TRUE       TRUE   TRUE    TRUE  TRUE FALSE
7        TRUE       TRUE   TRUE       TRUE   TRUE    TRUE  TRUE  TRUE
```

3 Interpretation

The way this function works is : it finds the 1 sized subset of variables which has best fitting model among all subset of variable of size 1. Then it finds the 2 sized subset of variables which has best fitting model among all subset of variable of size 2. And so on. And finally it finds the 7 sized subset of variables which has best fitting model among all subset of variable of size 7. Since there is only 7 total variables it takes all of them. Here TRUE means that the specific variable is present in the optimal set of variable and FALSE means that the specific variable is not present in the optimal set of variable. In our case :

- Among all set of size 1 of variables it takes only 'Murder' variable.
- Among all set of size 2 of variables it takes only 'Murder' and 'HS grad' variable.
- Among all set of size 3 of variables it takes only 'Murder' and 'HS grad' and 'Frost' variable.
- Among all set of size 4 of variables it takes only 'Murder' and 'HS grad' and 'Frost' and 'population' variable.
- Among all set of size 5 of variables it takes only 'Murder' and 'HS grad' and 'Frost' and 'population' and 'Income' variable.
- Among all set of size 6 of variables it takes only 'Murder' and 'HS grad' and 'Frost' and 'population' and 'Income' and 'Illiteracy' variable.
- For set of 7 variable the answer is trivially all variables.
- Note that here as we increase no of variables in the subset we keep on adding new variable in the existing set. Although this is not generally true but this phenomenon happens quite often.

Remark

These R commands uses BIC smartly to find which model is better than others. But here choice of criterion function does not matter since this function returns separate best models among sets of same sizes and since different model selection criteria such as AIC, BIC, CIC, DIC, ... differ only in how models of different sizes are compared, the results do not depend on the choice of cost-complexity tradeoff.
For more details see : https://cran.r-project.org/web/packages/leaps/leaps.pdf