How can I use regsubsets() in R for model selection?

Authored by

stats writer

June 23, 2024

Regsubsets() is a function in the R programming language that allows users to perform model selection by systematically evaluating different combinations of variables in a given dataset. This function uses the concept of best subsets, where it generates all possible models containing a specified number of variables and then selects the best model based on a chosen criterion, such as adjusted R-squared or Bayesian Information Criterion (BIC). By using regsubsets(), users can efficiently compare and select the most appropriate model for their data, leading to more effective and accurate statistical analyses.

Use regsubsets() in R for Model Selection

You can use the regsubsets() function from the leaps package in R to find the subset of predictor variables that produces the best regression model.

The following example shows how to use this function in practice.

Example: Using regsubsets() for Model Selection in R

For this example we'll use the built-in dataset in R, which contains measurements on 11 different attributes for 32 different cars.

#view first six rows of mtcars datasethead(mtcars)

mpg cyl disp hp drat wt qsec vs am gear carb
Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4
Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4
Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1

Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1
Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2
Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

Suppose we would like to fit a regression model using hp as the response variable and the following potential predictor variables:

mpgwtdratqsec

We can use the regsubsets() function from the leaps package to perform an exhaustive search to find the best regression model:

```
library(leaps)

#find best regression model
bestSubsets <- regsubsets(hp ~ mpg + wt + drat + qsec, data=mtcars)

#view results
summary(bestSubsets)

Subset selection object
Call: regsubsets.formula(hp ~ mpg + wt + drat + qsec,
```

data = mtcars)
4 Variables (and intercept)
Forced in Forced out
mpg FALSE FALSE
wt FALSE FALSE
drat FALSE FALSE
qsec FALSE FALSE
1 subsets of each size up to 4
Selection Algorithm: exhaustive
mpg wt drat qsec
1 (1) "*" " " " " " " "
2 (1) " " "*" " " "*"
3 (1) "*" "*" " " "*"
4 (1) "*" "*" "*" "*"

The stars (*) at the bottom of the output indicate which predictor variables belong in the best regression model for each possible model with a different number of predictor variables.

Here is how to interpret the output:

For a model with only one predictor variable, the best regression model is produced by using mpg as the

predictor variable.

For a model with two predictor variables, the best regression model is produced by using wt and qsec as the predictor variables.

For a model with three predictor variables, the best regression model is produced by using mpg, wt and qsec as the predictor variables.

For a model with four predictor variables, the best regression model is produced by using mpg, wt, drat and qsec as the predictor variables.

Note that you can also extract the following metrics for each model:

rsq :The for each modelRSS: The for each modeladjr2: The for each modelcp: for each modelbic: The for each model

#view adjusted R-squared value of each model summary(bestSubsets)$adjr2

0.5891853 0.7828169 0.7858829 0.7787005

From the output we can see:

The adjusted R-squared value for the model with mpg as the predictor variable is 0.589.The adjusted R-squared value for the model with wt and qsec as the predictor variables is 0.783.The adjusted R-squared value for the model with mpg, wt and qsec as the predictor variables is 0.786.The adjusted R-squared value for the model with mpg, wt, drat and qsec as the predictor variables is 0.779.

These values give us an idea of how well the set of predictor variables are able to predict the value of the response variable, adjusted for the number of predictor variables in the model.