How can I use stepAIC in R for feature selection?

Authored by

stats writer

June 23, 2024

StepAIC, or Stepwise Akaike Information Criterion, is a feature selection method that can be used in R to identify the most important variables for a particular model. It works by systematically adding or removing variables from a model, based on their impact on the Akaike Information Criterion (AIC) value. This helps to streamline the model and improve its overall performance by selecting only the most relevant variables. To use StepAIC in R, you can start by importing the "MASS" package, which contains the necessary functions. Then, you can specify your model and use the "stepAIC" function to automatically perform the stepwise feature selection process. This approach can be particularly useful when dealing with large datasets with numerous variables, as it helps to reduce the complexity of the model and improve its interpretability.

Use stepAIC in R for Feature Selection

The Akaike information criterion (AIC) is a metric that is used to quantify how well a model fits a dataset.

It is calculated as:

AIC = 2K - 2*ln*(L)

where:

K: The number of model parameters. The default value of K is 2, so a model with just one predictor variable will have a K value of 2+1 = 3.*ln*(L): The log-likelihood of the model. Most statistical software can automatically calculate this value for you.

The AIC is designed to find the model that explains the most variation in the data, while penalizing for models

that use an excessive number of parameters.

You can use the stepAIC() function from the MASS package in R to iteratively add and remove predictor variables from a regression model until you find the set of predictor variables (or "features") that produces the model with the lowest AIC value.

This function uses the following basic syntax:

stepAIC(object, direction, ...)

where:

object: The name of a fitted modeldirection: The type of stepwise search to use ("backward", "forward", or "both")

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

Example: Using stepAIC() for Feature 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

library(MASS)

#fit initial multiple linear regression model
model <- lm(hp ~ mpg + wt + drat + qsec, data=mtcars)

#use both forward and backward selection to find model with lowest AIC
stepAIC(model, direction="both")

Start: AIC=226.88

hp ~ mpg + wt + drat + qsec

Df Sum of Sq RSS AIC

- drat 1 94.9 28183 224.98

- mpg 1 1519.4 29608 226.56

none 28088 226.88

- wt 1 3861.9 31950 229.00

- qsec 1 28102.2 56190 247.06

Step: AIC=224.98

hp ~ mpg + wt + qsec

Df Sum of Sq RSS AIC

- mpg 1 1424.5 29608 224.56

none 28183 224.98

+ drat 1 94.9 28088 226.88

- wt 1 3797.9 31981 227.03

- qsec 1 29625.1 57808 245.97

Step: AIC=224.56

hp ~ wt + qsec

Df Sum of Sq RSS AIC

none 29608 224.56

+ mpg 1 1425 28183 224.98

+ drat 1 0 29608 226.56

- wt 1 43026 72633 251.28

- qsec 1 52881 82489 255.35

Call:

lm(formula = hp ~ wt + qsec, data = mtcars)

Coefficients:

(Intercept) wt qsec

441.26 38.67 -23.47

Here is how to interpret the output:

(1) First, we start by fitting a regression model with all four predictor variables. This model has an AIC value of 226.88.

(2) Next, stepAIC determines that removing drat as a predictor variable will further reduce the AIC value to 224.98.

(3) Next, stepAIC model determines that removing mpg as a predictor variable will further reduce the AIC value to 224.56.

(4) Lastly, stepAIC determines that there is no way to further reduce the AIC value by adding or removing any variables.

Thus, the final model is:

hp = 441.26 + 38.67(wt) - 23.47(qsec)

This model has an AIC value of 224.56.