Evaluating Regression Models Performance

Heaven Klair

5/4/2022

Regression Model Selection

After building the regression model on a dataset, we can evaluate its performance by looking at the adjusted R value produced by the model. We can change the regression formula by adding or subtracting the variables in the formula. Once we do that all the variables, we can compare the Adjusted R value of each formula to choose the best regression model with the respective variables.

One more thing we can do is follow the backward elimination method to choose the number of variables in the regression formula, and look at respective Adjusted R Squared value. This is all !!

This is specifically related to linear regression model

Interpreting coefficients in the linear regression model equation

Once we call summary of a linear function, we get a lot of coefficients and their values. In this file, we will address them and try to make sense of them.

Let us build a linear regression model as a example and call its summary.

```
dataset = read.csv('50_Startups.csv')
dataset$State = factor(dataset$State,
                       levels = c('New York', 'Florida', 'California'),
                       labels = c(1, 2, 3))

#install.packages('caTools')
library(caTools)
set.seed(123)
split = sample.split(dataset$Profit, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)

regressor = lm(formula = Profit ~ R.D.Spend  + Marketing.Spend,
               data = training_set)
summary(regressor)
```

```
##
## Call:
## lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = training_set)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33294  -4763   -354   6351  17693
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)     4.638e+04  3.019e+03  15.364   <2e-16 ***
## R.D.Spend       7.879e-01  4.916e-02  16.026   <2e-16 ***
## Marketing.Spend 3.538e-02  1.905e-02   1.857   0.0713 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9533 on 37 degrees of freedom
## Multiple R-squared:  0.9495, Adjusted R-squared:  0.9468
## F-statistic: 348.1 on 2 and 37 DF,  p-value: < 2.2e-16
```

We see that there are a lot of coefficients that we get from calling summary of the model. Let's interpret these coefficients.

First we see that we are trying to predict profit using the R.D. Spend of the company, and the Marketing Spend. The Estimate column tells us the values of the coefficients in the multiple linear equation

$$\text{profit} = b_1 \times \text{R.D. Spend} + b_2 \times \text{Marketing Spend}$$

So b_1 and b_2 are those coefficients (estimate values.) How do we interpret them? We check if the coefficients of the equation are positive or negative. If they are positive, then the dependent variable is correlated with the independent variable. That means if we change the value of R.D. Spend value, then profit will change. Furthermore, if we increase the value of R.D. Spend, profit will increase too.

If we increase the spend on Marketing, then the profit will increase too.

If the coefficients are negative, then the effects are opposite. Increasing the value of independent variable will decrease the dependent variable's value.

Now, let's talk about magnitude of the coefficients. In this example, we have $b_1 > b_2$. This is a tricky question because we can change the value of coefficient by a multiple of a constant. For example, Marketing spend is being calculated in dollars, what if we change it to cents, then the value of b_2 will get increased by a multiple of 100 and become greater than b_1.

A way to address the magnitude of the coefficients, one can address them along with their units for example "b_1" is in dollars and b_2 is in cents. Another thing we can do is state the units as "units". Here, we know that $b_1 > b_2$, so "R.D. spend has a greater impact on profits than Marketing spend per unit".

Further, what does the value of coefficients tells us?

The value of 0.79 for R.D Spend tells us that if we keep all other independent variables constant (not change them), and only change value of R.D spend, then the profit will increase by 0.79 cents per unit increase in R.D spend.

Similarly, if we keep all the other independent variables constant and only change the Marketing spend, then the profit will increase by 0.02 cents for every one unit increase in Marketing spend.