Regressão Linear Múltipla

Contents

Informações iniciais

Há 50 empresas com gastos em:

- R&D Spend
- Administration
- Marketing Spend
- State

assim como o lucro (`profit`), variável dependente.

Importando os dados

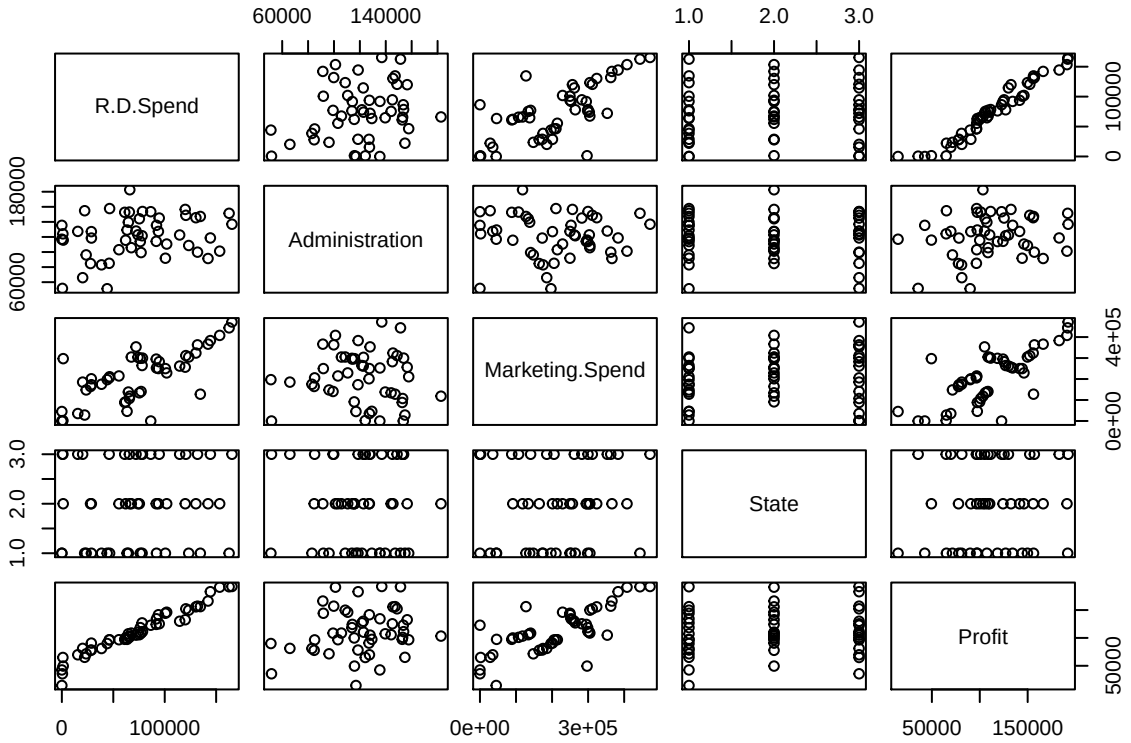
```
dataset = read.csv('https://pastebin.com/raw/UaFmFF4j')
head(dataset)
```

```
##   R.D.Spend Administration Marketing.Spend      State   Profit
## 1  165349.2      136897.80        471784.1   New York 192261.8
## 2  162597.7      151377.59        443898.5 California 191792.1
## 3  153441.5      101145.55        407934.5    Florida 191050.4
## 4  144372.4      118671.85        383199.6   New York 182902.0
## 5  142107.3       91391.77        366168.4    Florida 166187.9
## 6  131876.9       99814.71        362861.4   New York 156991.1
```

```
str(dataset)
```

```
## 'data.frame':    50 obs. of  5 variables:
##  $ R.D.Spend      : num  165349 162598 153442 144372 142107 ...
##  $ Administration : num  136898 151378 101146 118672 91392 ...
##  $ Marketing.Spend: num  471784 443899 407935 383200 366168 ...
##  $ State          : Factor w/ 3 levels "California","Florida",..: 3 1 2 3 2 3 1 2 3 1 ...
##  $ Profit         : num  192262 191792 191050 182902 166188 ...
```

```r
plot(dataset)
```



Codificar a variável categórica `State`

```r
dataset$State = factor(dataset$State,
                       levels = c('New York', 'California', 'Florida'), labels = c(1, 2, 3))
```

Separando os dados em `training_set` e `test_set`

```r
library(caTools)
set.seed(123)
split = sample.split(dataset$Profit, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE) # 80% dos dados serão usados para TREINO
test_set = subset(dataset, split == FALSE) # 20% dos dados serão usados para TESTE
```

Criar o modelo

Criar o modelo dos dados de `training_set` sendo `profit` a variável dependente e todas as outras são independentes

```r
regressor = lm(formula = Profit ~ .,
               data = training_set)
summary(regressor)
```

```
##
## Call:
## lm(formula = Profit ~ ., data = training_set)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33128  -4865      5   6098  18065
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.965e+04  7.637e+03   6.501 1.94e-07 ***
## R.D.Spend        7.986e-01  5.604e-02  14.251 6.70e-16 ***
## Administration  -2.942e-02  5.828e-02  -0.505    0.617
## Marketing.Spend  3.268e-02  2.127e-02   1.537    0.134
## State2           1.213e+02  3.751e+03   0.032    0.974
## State3           2.376e+02  4.127e+03   0.058    0.954
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9908 on 34 degrees of freedom
## Multiple R-squared:  0.9499, Adjusted R-squared:  0.9425
## F-statistic:   129 on 5 and 34 DF,  p-value: < 2.2e-16
```

Predição dos lucros de `test_set` com o modelo de `trainig_set`

```
y_pred = predict(regressor, newdata = test_set)
head(y_pred)
```

```
##        4        5        8       11       16       20
## 173981.1 172655.6 160250.0 135513.9 146059.4 114151.0
```