

INSTITUTO TECNOLOGICO DE TIJUANA

CARRERA

INGENIERÍA EN SISTEMAS COMPUTACIONALES

MATERIA

MINERÍA DE DATOS

TAREA

PRÁCTICA #2, UNIDAD #3

FECHA ENTREGA

18/05/2022

ALUMNO(A)

HOWARD HERRERA ERWIN #18210716

PÉREZ LÓPEZ ALICIA GUADALUPE #18210514

DOCENTE

JOSE CHRISTIAN ROMERO HERNANDEZ

Importamos el archivo “csv” con el nombre “50_Startups.csv”, luego tenemos una columna de estados con 3 diferentes, para poder hacerlos numéricos usaremos “factor” el número 1 para New York, 2 para California, 3 para Florida.

```
> dataset <- read.csv(file.choose())
> dataset$State = factor(dataset$State,
+                        levels = c('New York', 'California', 'Florida'),
+                        labels = c(1,2,3))
> dataset
   R.D.Spend Administration Marketing.Spend State    Profit
1  165349.20      136897.80       471784.10     1 192261.83
2  162597.70      151377.59       443898.53     2 191792.06
3  153441.51      101145.55       407934.54     3 191050.39
4  144372.41      118671.85       383199.62     1 182901.99
5  142107.34       91391.77       366168.42     3 166187.94
6  131876.90       99814.71       362861.36     1 156991.12
7  134615.46      147198.87       127716.82     2 156122.51
8  130298.13      145530.06       323876.68     3 155752.60
9  120542.52      148718.95       311613.29     1 152211.77
10 123334.88      108679.17       304981.62     2 149759.96
11 101913.08      110594.11       229160.95     3 146121.95
12 100671.96       91790.61       249744.55     2 144259.40
13  93863.75      127320.38       249839.44     3 141585.52
14  91992.39      135495.07       252664.93     2 134307.35
15 119943.24      156547.42       256512.92     3 132602.65
16 114523.61      122616.84       261776.23     1 129917.04
17  78013.11      121597.55       264346.06     2 126992.93
18  94657.16      145077.58       282574.31     1 125370.37
19  91749.16      114175.79       294919.57     3 124266.90
```

Utilizaremos la librería “caTools” para la función aleatoria con “seed” y la dividiremos 0.8, y con el training_set probará si es verdadero o falso.

```
> 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)
```

Ahora tendremos un filtro para la regresión lineal múltiple con lm del modelo lineal tomando el perfil y todas las demás columnas de nuestro conjunto de entrenamiento. Y obtenemos nuestro resumen o resumen de los datos.

```
> 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

> |
```

Hacemos la predicción

```
> y_pred = predict(regressor, newdata = test_set)
> y_pred
       4        5        8       11       16       20       21       24
173981.09 172655.64 160250.02 135513.90 146059.36 114151.03 117081.62 110671.31
      31       32
 98975.29  96867.03
> |
```

Ahora vemos que el valor de P tenemos que elegir los más cercanos a 0 y el modelo nos dice que eliminemos las variables mayores a 0,05, que es el valor significativo.

Y estamos haciendo el modelo con todas las variables independientes.

```
> regressor = lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend
 + State,
+               data = dataset )
> summary(regressor)


Call:
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend +
    State, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33504  -4736     90   6672  17338

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.008e+04  6.953e+03   7.204 5.76e-09 ***
R.D.Spend        8.060e-01  4.641e-02  17.369  < 2e-16 ***
Administration  -2.700e-02  5.223e-02  -0.517    0.608
Marketing.Spend  2.698e-02  1.714e-02   1.574    0.123
State2           4.189e+01  3.256e+03   0.013    0.990
State3           2.407e+02  3.339e+03   0.072    0.943
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9439 on 44 degrees of freedom
Multiple R-squared:  0.9508,    Adjusted R-squared:  0.9452
F-statistic: 169.9 on 5 and 44 DF,  p-value: < 2.2e-16
```

Eliminamos el estado por su valor P.

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

Call:
lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend,
    data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33534  -4795     63   6606  17275

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.012e+04  6.572e+03   7.626 1.06e-09 ***
R.D.Spend        8.057e-01  4.515e-02  17.846  < 2e-16 ***
Administration  -2.682e-02  5.103e-02  -0.526    0.602
Marketing.Spend  2.723e-02  1.645e-02   1.655    0.105
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9232 on 46 degrees of freedom
Multiple R-squared:  0.9507,    Adjusted R-squared:  0.9475
F-statistic:   296 on 3 and 46 DF,  p-value: < 2.2e-16
```

Eliminamos el valor de la administración por el valor significativo de 0.05

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

Call:
lm(formula = Profit ~ R.D.Spend + Marketing.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-33645  -4632   -414   6484  17097

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.698e+04  2.690e+03  17.464   <2e-16 ***
R.D.Spend       7.966e-01  4.135e-02  19.266   <2e-16 ***
Marketing.Spend 2.991e-02  1.552e-02   1.927     0.06 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9161 on 47 degrees of freedom
Multiple R-squared:  0.9505,    Adjusted R-squared:  0.9483
F-statistic: 450.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

Aquí tenemos el gasto en marketing que está muy cerca de 0,05 pero el modelo nos dice que tenemos que eliminarlo.

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

Call:
lm(formula = Profit ~ R.D.Spend, data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-34351  -4626   -375   6249  17188

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.903e+04  2.538e+03   19.32   <2e-16 ***
R.D.Spend   8.543e-01  2.931e-02   29.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9416 on 48 degrees of freedom
Multiple R-squared:  0.9465,    Adjusted R-squared:  0.9454
F-statistic: 849.8 on 1 and 48 DF,  p-value: < 2.2e-16
```

Y con esto tenemos nuestro regresor.

Ya tenemos nuestra predicción

```
> y_pred = predict(regressor, newdata = test_set)
> y_pred
       4        5        8       11       16       20       21       24
172369.0 170434.0 160345.5 136096.4 146869.4 122860.5 114175.9 106725.4
      31       32
101994.2 101261.2
```

Esta función lo que hace es tomar los parámetros del conjunto de entrenamiento que son nuestros datos y SL o el valor significativo de 0.05 y lo que hace es un ciclo que comienza en 1 y numvars que toma la cantidad de datos en el conjunto de entrenamiento, y con esto obtiene el regresor con el modelo lineal entre la ganancia y las otras columnas de nuestra x que será el conjunto de entrenamiento, a continuación, establece maxVar con el coeficiente máximo que se extraerá del resumen del regresor y pasa a la condición si este valor es mayor que el SL o valor significativo de 0. 05 asigna a la variable J la columna que tiene ese valor y en x que es el dataset la quitará y restará numVars por 1 hasta que se quede con las variables que no superen el valor significativo y devolverá el resumen con las columnas con el valor significativo menor a 0.05.

```
> backwardElimination <- function(x, sl) {
+   numVars = length(x)
+   for (i in c(1:numVars)){
+     regressor = lm(formula = Profit ~ ., data = x)
+     maxVar = max(coef(summary(regressor))[c(2:numVars), "Pr(>|t|)"])
+     if (maxVar > sl){
+       j = which(coef(summary(regressor))[c(2:numVars), "Pr(>|t|)"] == maxVar)
+       x = x[, -j]
+     }
+     numVars = numVars - 1
+   }
+   return(summary(regressor))
+ }
> SL = 0.05
```

```
> training_set
   R.D.Spend Administration Marketing.Spend State    Profit
1  165349.20       136897.80        471784.10     1 192261.83
2  162597.70       151377.59        443898.53     2 191792.06
3  153441.51       101145.55        407934.54     3 191050.39
6  131876.90        99814.71        362861.36     1 156991.12
7  134615.46       147198.87        127716.82     2 156122.51
9  120542.52       148718.95        311613.29     1 152211.77
10 123334.88       108679.17        304981.62     2 149759.96
12 100671.96        91790.61        249744.55     2 144259.40
13  93863.75       127320.38        249839.44     3 141585.52
14  91992.39       135495.07        252664.93     2 134307.35
15 119943.24       156547.42        256512.92     3 132602.65
17  78013.11       121597.55        264346.06     2 126992.93
18  94657.16       145077.58        282574.31     1 125370.37
19  91749.16       114175.79        294919.57     3 124266.90
22  78389.47       153773.43        299737.29     1 111313.02
23  73994.56       122782.75        303319.26     3 110352.25
25  77044.01        99281.34        140574.81     1 108552.04
26  64664.71       139553.16        137962.62     2 107404.34
27  75328.87       144135.98        134050.07     3 105733.54
28  72107.60       127864.55        353183.81     1 105008.31
29  66051.52       182645.56        118148.20     3 103282.38
30  65605.48       153032.06        107138.38     1 101004.64
33  63408.86       129219.61         46085.25     2  97427.84

34  55493.95       103057.49        214634.81     3  96778.92
35  46426.07       157693.92        210797.67     2  96712.80
36  46014.02        85047.44        205517.64     1  96479.51
37  28663.76       127056.21        201126.82     3  90708.19
38  44069.95        51283.14        197029.42     2  89949.14
39  20229.59        65947.93        185265.10     1  81229.06
40  38558.51        82982.09        174999.30     2  81005.76
41  28754.33       118546.05        172795.67     2  78239.91
42  27892.92        84710.77        164470.71     3  77798.83
43  23640.93        96189.63        148001.11     2  71498.49
44  15505.73       127382.30         35534.17     1  69758.98
45  22177.74       154806.14         28334.72     2  65200.33
46   1000.23       124153.04          1903.93     1  64926.08
47   1315.46       115816.21        297114.46     3  49490.75
48      0.00       135426.92             0.00     2  42559.73
49    542.05        51743.15             0.00     1  35673.41
50      0.00       116983.80         45173.06     2  14681.40
```

```
> backwardElimination(training_set, SL)

Call:
lm(formula = Profit ~ ., data = x)

Residuals:
   Min     1Q Median     3Q    Max
-34334  -4894   -340   6752  17147

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.902e+04  2.748e+03   17.84   <2e-16 ***
R.D.Spend   8.563e-01  3.357e-02   25.51   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9836 on 38 degrees of freedom
Multiple R-squared:  0.9448,    Adjusted R-squared:  0.9434
F-statistic: 650.8 on 1 and 38 DF,  p-value: < 2.2e-16
```