Multiple Linear Regression

Heaven Klair

1/30/2022

Encoding categorical data

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

Splitting the dataset into the Training set and Test set

```
# 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)
```

Fitting Multiple Linear Regression to the Training set

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

```
##
## Call:
## lm(formula = Profit ~ R.D.Spend + Administration + Marketing.Spend +
##     State, 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           2.376e+02  4.127e+03   0.058    0.954
## State3           1.213e+02  3.751e+03   0.032    0.974
## ---
## 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
```

The important columns to look at here are the last two: P-value and the Significance level because these columns tell us about the statistical significance of the independent variable onto the dependent variable. This means it tells us if each of the independent variable has a significant impact on on the dependent variable.

Lower the p-value is, more statistically significant the independent variable is going to be.

In the last column, the first two rows has stars on the side. This means that there will be highly statistical significance of the independent vairable onto the dependent variable.

Predict the test set Results

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

Look at the original values of the profit in the data set and compare them to these, we will see that both of them are quiet similar.

Building the optimal Model using Backward Elimination

Steps for Backward Elimination: 1. Select a significance level to stay in the model (e.g SL = 0.05) 2. Fit the model with all possible predictors 3. Consider the predictor with the highest P-value. If P > SL, go to step 4, otherwise go to FINISH 4. Remove the predictor 5. Fit the model without this variable. (and Repeat the step 3)

FINSIH: your model is ready

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

# Changing the data to "dataset" is not necessary, but we do it because we would like to use all the da
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           2.407e+02  3.339e+03   0.072    0.943
## State3           4.189e+01  3.256e+03   0.013    0.990
## ---
## 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
```

We can notice that State Florida and New York has very high P-values, above 90%, so lets remove them first.

```
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
```

```
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
```

```
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
```