aaa

kuntal

2/15/2023

```
knitr::opts_chunk$set(echo = TRUE)
```

Multiple Linear Regression

Importing the dataset

```
dataset = read.csv('50_Startups.csv')
```

Encoding categorical data

```
dataset$State = factor(dataset$State,
                       levels=c('New York', 'California', 'Florida'),
                       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 )
```

Feature Scaling

training_set = scale(training_set)

test_set = scale(test_set)

Fitting Multiple Linear Regression to the Training set

can be also written as if you want selected independent columns and . for all columns

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

```
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
```

#Analizing the result of summery coffecient we see R.D.Spend has the maximum effect as P-value is minimum and has # "***" statistical significance # that gives the benefit that we can also use simple linear regression for faster process #like formula = Profit ~ R.D.Spend that will be ok and will give same prediction

Predicting the Test set results

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

#building optimum modal using backward elimination

#Backward Elimination #Step1 - select significance level (SL) to stay in the modal (eg SL =.05) #Step2 - fit the full modal with all possible predictor #Step3 - consider the predictor with the highest P-Value. if P > SL go to step 4 otherwise go to FIN (finish Modal is ready)
#Step4 - remove the predictor #Step5 - fit the modal without the variable* (Note* means rebuild the modal again so if 100 after remove its 99 you have the rebuild the modal with 99 variable) go back to Step 3 till (not P > SL)

#you can take training set also instead of whole dataset

```
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
```

#remove independent variable one by one where P-value > than .05 or 5% in Coefficients #first remove State has very high P-value 99% and 93% no statistical significance or impact on dependent variable Profit

```
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
```

#remove Administration has very high P-values %60 no statistical significance or impact on dependent variable Profit

```
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
```

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

```
##         4         5         8        11        16        20        21        24
```

```
## 173441.31 171127.62 160455.74 135011.91 146032.72 115816.42 116650.89 109886.19
##        31        32
##  99085.22  98314.55
```

#You can see R.D.Spend excellent impact on Profit # but now you can also see Marketing.Spend has some statistical significance #remove Marketing.Spend as it is > 5% following strictly the elimination rules #Note You may think about it to keep also

```
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_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
```

#Note: So final modal is either based on one independent variable R.D.Spend or # team with R.D.Spend +Marketing.Spend depending upon your choice # now need check adjusted R squared and R squared and # then choose based on adjusted R squared near to 1 for best modal

Note:Coefficient are correlated to between dependent and independent variable

these are measured per unit. since here we are talking in dollars

+ve mean increase and -ve means decrease

R.D.Spend per unit(dollar) increase the profit per unit (in dollar)

by 7.9 cent and similarly for marketing. Always treat it as Per Unit both variable

#Also Coefficient talk about the additional effect of every single variable # given that the other variable already in place. example here R.D.Spend already # in the modal and Marketing.Spend give additional effect

###Automated Code for Backward Elimination

```
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
dataset = dataset[, c(1,2,3,4,5)]
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
```