Multiple Linear Regression Model using R

Name: Abhishek K Singh

Polynomial Regression Output Considering All Independent variables

```
> summary(regressor1)

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

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 ***
State2           1.213e+02  3.751e+03   0.032    0.974    
State3           2.376e+02  4.127e+03   0.058    0.954    
Administration  -2.942e-02  5.828e-02  -0.505    0.617    
Marketing.Spend  3.268e-02  2.127e-02   1.537    0.134    
---
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
```

Conclusion: We can clearly observe from above model output that

Pr = 0.974 & 0.954 >> 0.05, therefore State variable is not significant to “Profit”

So, we can ignore “State1” & “State2”, from our model

Polynomial Regression Output Removing “State” Variable

```
> summary(regressor2)

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

Residuals:
   Min     1Q Median     3Q    Max 
-33117  -4858    -36   6020  17957 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      4.970e+04  7.120e+03   6.980 3.48e-08 ***
R.D.Spend        7.983e-01  5.356e-02  14.905  < 2e-16 ***
Administration  -2.895e-02  5.603e-02  -0.517    0.609    
Marketing.Spend  3.283e-02  1.987e-02   1.652    0.107    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9629 on 36 degrees of freedom
Multiple R-squared:  0.9499,	Adjusted R-squared:  0.9457 
F-statistic: 227.6 on 3 and 36 DF,  p-value: < 2.2e-16
```

Conclusion: We can clearly observe from above model output that

Pr = 0.609 >> 0.05, therefore “Administration” variable is not significant to “Profit”

So, we can ignore “Administration” variable from our model

Polynomial Regression Output Removing "Administration" Variable

```
> summary(regressor3)

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

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
```

Conclusion: We can clearly observe from above model output that

Pr = 0.07 ≈ 0.05, therefore "Marketing.Spend" variable is may or may not be significant to "Profit"

So, lets see what happens if we ignore "Marketing.Spend" variable from our model

Polynomial Regression Output Removing "Marketing.Spend" Variable

```
> summary(regressor4)

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

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
```

Conclusion: We can clearly observe from above model output that

"Adjusted R-squared" decreased from "0.9468" in the previous model to "0.9434" in this model & "Adjusted R-squared" should be close to 1 for Modeling Parameters to have stronger relationship

So, 3rd model was better than 4th model

Prediction Comparison of different model on Test data

S.N.	R.D.Spend	Marketing.Spend	Profit	Predict1	Predict2	Predict3	Predict4
4	144372.4	383199.62	182902	173981	174095	173687	172648
5	142107.3	366168.42	166188	172656	172517	171300	170708
8	130298.1	323876.68	155753	160250	160135	160499	160596
11	101913.1	229160.95	146122	135514	135378	134783	136288
16	114523.6	261776.23	129917	146059	146167	145873	147087
20	86419.7	0	122777	114151	114244	114468	123020
21	76253.86	298664.47	118474	117082	117082	117025	114315
24	67532.53	304768.73	108734	110671	110555	110370	106847
31	61994.48	91131.24	99937.6	98975.3	98834.3	98447.4	102104
32	61136.38	88218.23	97483.6	96867	96980.7	97668.2	101369

Conclusion: We can clearly observe from the above prediction output that "Predict3" is Closest prediction to our actual prediction

So, Regression Model 3 is best model for this dataset