ASSIGNMENT

Designs of Experiment– I (ELECTIVE)

REGRESSION ANALYSIS

TYBSC - SEM V

2021-22

TOPIC: Analyzing if profit earned by a startup depends upon the revenue spend on the start up for Research & development , administration and marketing.

Assignment By
Aanchal Yadav ***2015228*** **TYBSC STATISTICS**
Hrutuja Patkar ***2115205*** **TYBSC STATISTICS**

Estimated Regression Equation

$$\widehat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

FOUR Variable Model :

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \varepsilon$$

Where

Y = Dependent Variable (Profit)

X_1 = 1st independent variable (R&D spend)

X_2 = 2nd independent variable (Administration)

X_3 = 3rd independent variable (Marketing Speed)

	X_1	X_2	X_3	Y
Observation	R&D Spend	Administration	Marketing Spend	Profit
1	165349.2	136897.8	471784.1	192261.8
2	162597.7	151377.6	443898.5	191792.1
3	153441.5	101145.6	407934.5	191050.4
4	144372.4	118671.9	383199.6	182902
5	142107.3	91391.77	366168.4	166187.9
6	131876.9	99814.71	362861.4	156991.1
7	134615.5	147198.9	127716.8	156122.5
8	130298.1	145530.1	323876.7	155752.6
9	120542.5	148719	311613.3	152211.8
10	123334.9	108679.2	304981.6	149760
11	101913.1	110594.1	229161	146122
12	100672	91790.61	249744.6	144259.4
13	93863.75	127320.4	249839.4	141585.5
14	91992.39	135495.1	252664.9	134307.4
15	119943.2	156547.4	256512.9	132602.7
16	114523.6	122616.8	261776.2	129917
17	78013.11	121597.6	264346.1	126992.9
18	94657.16	145077.6	282574.3	125370.4
19	91749.16	114175.8	294919.6	124266.9
20	86419.7	153514.1	0	122776.9
21	76253.86	113867.3	298664.5	118474
22	78389.47	153773.4	299737.3	111313
23	73994.56	122782.8	303319.3	110352.3
24	67532.53	105751	304768.7	108734
25	77044.01	99281.34	140574.8	108552
26	64664.71	139553.2	137962.6	107404.3
27	75328.87	144136	134050.1	105733.5
28	72107.6	127864.6	353183.8	105008.3
29	66051.52	182645.6	118148.2	103282.4
30	65605.48	153032.1	107138.4	101004.6
31	61994.48	115641.3	91131.24	99937.59
32	61136.38	152701.9	88218.23	97483.56
33	63408.86	129219.6	46085.25	97427.84
34	55493.95	103057.5	214634.8	96778.92
35	46426.07	157693.9	210797.7	96712.8
36	46014.02	85047.44	205517.6	96479.51
37	28663.76	127056.2	201126.8	90708.19
38	44069.95	51283.14	197029.4	89949.14
39	20229.59	65947.93	185265.1	81229.06
40	38558.51	82982.09	174999.3	81005.76
41	28754.33	118546.1	172795.7	78239.91
42	27892.92	84710.77	164470.7	77798.83
43	23640.93	96189.63	148001.1	71498.49
44	15505.73	127382.3	35534.17	69758.98
45	22177.74	154806.1	28334.72	65200.33
46	1000.23	124153	1903.93	64926.08
47	1315.46	115816.2	297114.5	49490.75
48	0	135426.9	0	42559.73
49	542.05	51743.15	0	35673.41
50	0	116983.8	45173.06	14681.4

MULTIPLE REGRESSION OUTPUT

First, we perform F test.
F statistic is a test of significance for the entire regression.
At α = 0.05, this regression is statistically significant because p-value < 0.05.

ANOVA Table at 5% Loss

ANOVA					
	df	SS	MS	F	Significance F
Regression	3	75683964196	2.5228E+10	295.9781	0.00
Residual	46	3920856301	85236006.5		
Total	49	79604820497			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	50122.193	6572.3526	7.6262	0	36892.7333	63351.65	36892.73	63351.65
R&D Spend	0.8057	0.0451	17.8464	0	0.7148	0.8966	0.7148	0.8966
Administration	-0.0268	0.051	-0.5255	0.6018	-0.1295	0.0759	-0.1295	0.0759
Marketing Spend	0.0272	0.0165	1.6551	0.1047	-0.0059	0.0603	-0.0059	0.0603

Linear Regression Model = $\hat{Y}$ = 50122.193 +0.8057 X_1 -0.0268 X_2 + 0.0272 X_3

Descriptive statistics from the regression output

Regression Statistics	
Multiple R	0.975062046
R Square	0.950745994
Adjusted R Square	0.947533776
Standard Error	9232.334837
Observations	50

Coefficient of Determination
A measure of "explained variation." Result shows that about 95% of the total variation in Profit Gained (Y) is explained by the regression.
$r^2 = SSR/SST$

Standard Error of Estimate
A measure of "unexplained variation." Std. Error = SQRT(MSE) = SQRT(9232.3348)

95.07% variability in Y is explained by X_1 X_2 & X_3

Checking for assumptions of the regression

1. Multicollinearity

2. Homoscedasticity test:

3. Autocorrelation

1. Multicollinearity

Multicollinearity may be checked by correlation matrix:

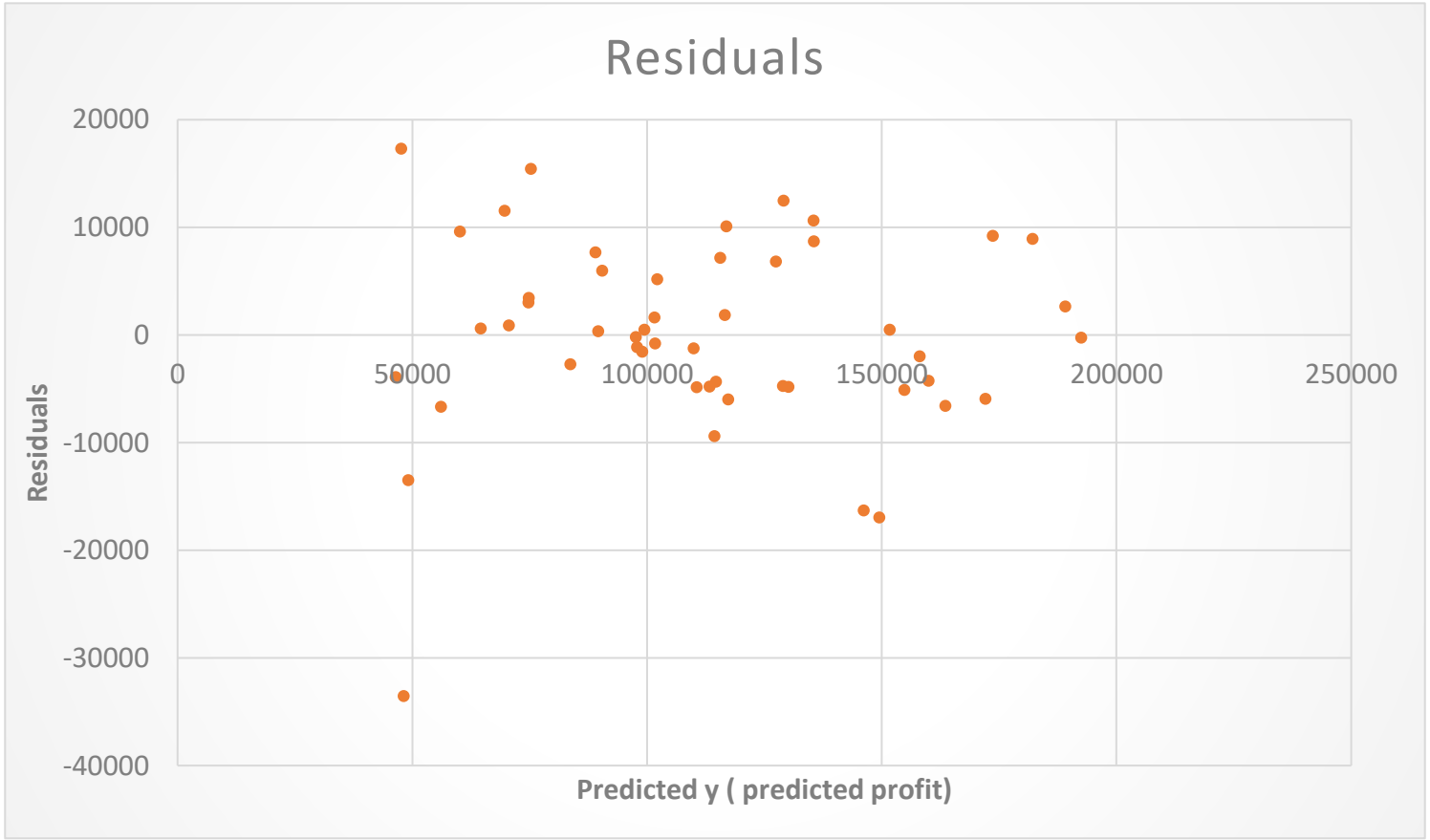
	R&D Spend	Administration	Marketing Spend	Profit
R&D Spend	**1**			
Administration	**0.241955**	**1**		
Marketing Spend	**0.724248**	**-0.03215**	**1**	
Profit	**0.9729**	**0.200717**	**0.747766**	**1**

When computing a matrix of Pearson's bivariate correlations among all independent variables, the magnitude of the correlation coefficients should be less than .80

As there are no correlation coefficients above than .80, there is no multicollinearity in the data

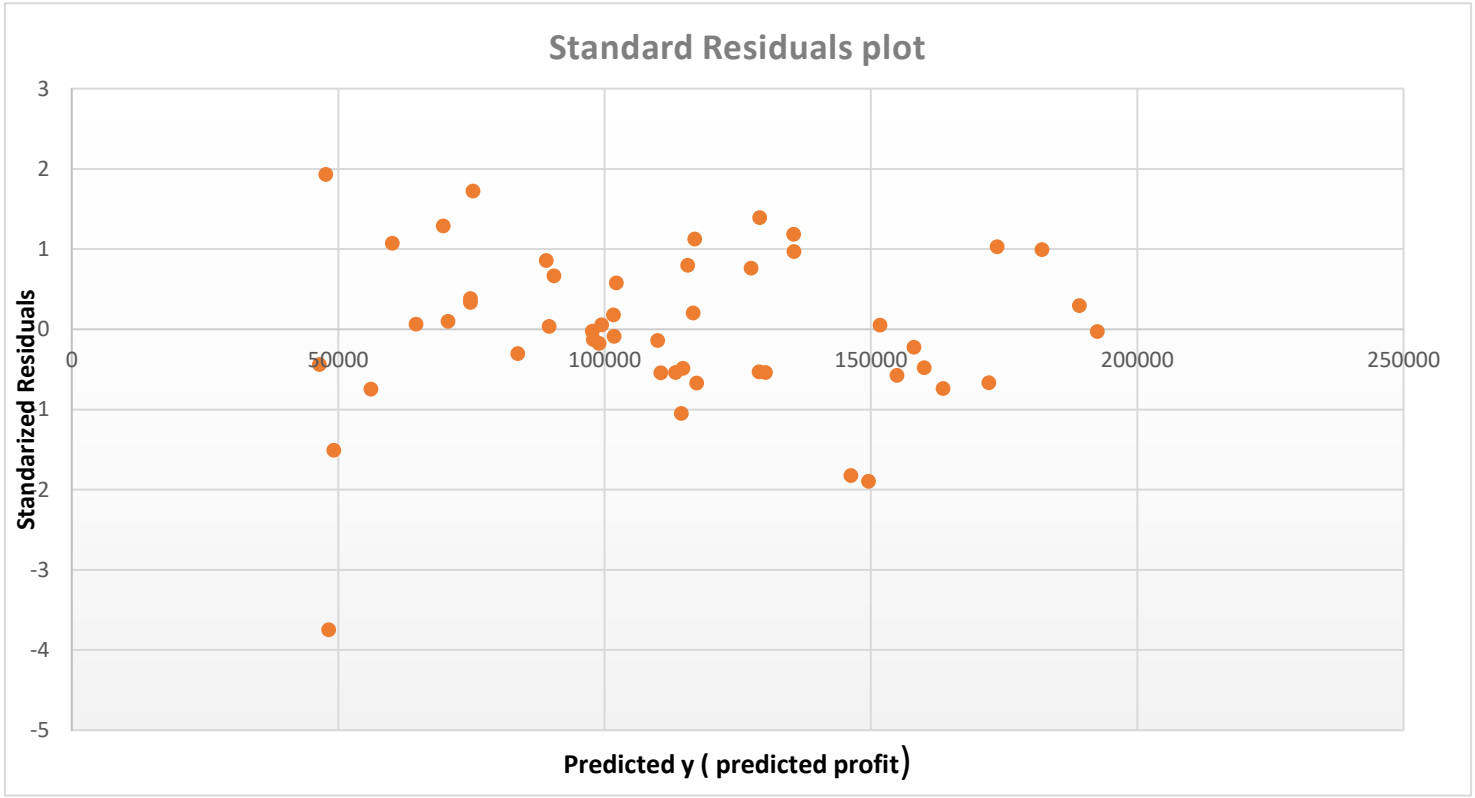
2. Homoscedasticity test:

A scatterplot of residuals versus predicted values is good way to check for homoscedasticity.
There should be no clear pattern in the distribution; if there appears a cone-shaped or any other shaped the data is *heteroscedastic*.



A data transformation scale is applied to see if we get any other results.

We applied log-transformation, residual square, residual square root and standardized residual transformation.



From all the plots, we found that the data follows **Homoscedasticity.**

3. Autocorrelation

One of the assumptions of linear regression is that there is no **autocorrelation** between the residuals.

For testing the assumption we perform *Durbin Watson Test* in which we test the autocorrelation existence within independent variable by observing ***d*** value after testing the data.

Step 1

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	*0.975062*							
R Square	*0.950746*							
Adjusted R Square	*0.947534*							
Standard Error	*9232.335*							
Observations	*50*							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	*3*	*7.57E+10*	*2.52E+10*	*295.9781*	*4.53E-30*			
Residual	*46*	*3.92E+09*	*85236007*					
Total	*49*	*7.96E+10*						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	*50122.19*	*6572.353*	*7.626218*	*1.06E-09*	*36892.73*	*63351.65*	*36892.73332*	*63351.65*
R&D Spend	*0.805715*	*0.045147*	*17.84637*	*2.63E-22*	*0.714838*	*0.896592*	*0.714838309*	*0.896592*
Administration	*-0.02682*	*0.051029*	*-0.52551*	*0.601755*	*-0.12953*	*0.0759*	*-0.129531575*	*0.0759*
Marketing Spend	*0.027228*	*0.016451*	*1.655077*	*0.104717*	*-0.00589*	*0.060343*	*-0.005886553*	*0.060343*

Test statistic for Durbin Watson test :

$$d = \sum_{t=2}^{T} (e_t - e_{t-1})^2 / \sum_{t=1}^{T} {e_t}^2$$

$d = 0.286785$

n = 50
K = 50
By following the multiple regression table, we get
The lower critical value : 1.245
The upper critical value : 1.491

Since,
our test statistic *(d = 0.286785)* does not lie in the range of *1.245- 1.491* we conclude that the data is ***Autocorrelated***.

Step 2

Residual output		
Observation	**Predicted Profit**	**Residuals**
1	192521.3	-259.423
2	189156.8	2635.292
3	182147.3	8903.111
4	173696.7	9205.29
5	172139.5	-5951.57
6	163580.8	-6589.66
7	158114.1	-1991.59
8	160021.4	-4268.76
9	151741.7	470.0703
10	154884.7	-5124.72
11	135509	10612.93
12	135573.7	8685.687
13	129138.1	12447.47
14	127488	6819.358
15	149548.6	-16946
16	146235.2	-16318.1
17	116915.4	10077.52
18	130192.4	-4822.08
19	129014.2	-4747.33
20	115635.2	7141.644
21	116639.7	1834.361
22	117319.5	-6006.43
23	114707	-4354.73
24	109996.6	-1262.63
25	113363	-4810.93
26	102237.7	5166.615
27	110600.6	-4867.04
28	114408.1	-9399.76
29	101660	1622.354
30	101795	-790.343
31	99452.37	485.2171
32	97687.86	-204.296
33	99001.33	-1573.49
34	97915.01	-1136.09
35	89039.27	7673.526
36	90511.6	5967.91
37	75286.17	15422.02
38	89619.54	329.6023
39	69697.43	11531.63
40	83729.01	-2723.25
41	74815.95	3423.956
42	74802.56	2996.274
43	70620.41	878.0782
44	60167.04	9591.94
45	64611.35	588.9751
46	47650.65	17275.43
47	56166.21	-6675.46
48	46490.59	-3930.86
49	49171.39	-13498
50	48215.13	-33533.7

Fitting a linear regression model to the selected data using
Backward Elimination Method.

Step 1:

Fit all 4-variable model: $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.975062046					
R Square	0.950745994					
Adjusted R Square	0.947533776					
Standard Error	9232.334837					
Observations	50					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	3	75683964196	2.5228E+10	295.9781	0.0000	
Residual	46	3920856301	85236006.5			
Total	49	79604820497				
	Coefficients	Standard Error	t Stat	P-value		
Intercept	50122.193	6572.3526	7.6262	0		
R&D Spend	0.8057	0.0451	17.8464	0		
Administration	-0.0268	0.051	-0.5255	0.6018		
Marketing Spend	0.0272	0.0165	1.6551	0.1047		

We eliminate X_2 (Administration) variable as it has the highest p value greater than 0.05(5% level of significance**)**

Step 2:

Eliminating X_2 variable,
We fit rest 3-variable model: $\beta_0 + \beta_1 X_1 + \beta_3 X_3$

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.9749104				
R Square	0.9504503				
Adjusted R Square	0.9483418				
Standard Error	9160.9658				
Observations	50				
ANOVA					
	df	SS	MS	F	Significance F
Regression	2	7.57E+10	3.78E+10	450.7713	0.0000
Residual	47	3.94E+09	83923295		
Total	49	7.96E+10			
	Coefficients	Standard Error	t Stat	P-value	
Intercept	46975.8642	2689.9329	17.4636	0.0000	
R&D Spend	0.7966	0.0413	19.2656	0.0000	
Marketing Spend	0.0299	0.0155	1.9271	0.0600	

We eliminate X_3 (Marketing Spend) variable as it has p-value greater than 0.05(5% level of significance**)**

Step 3:

Eliminating X_3 variable,

We fit rest 2-variable model: $\beta_0 + \beta_1 X_1$

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.9729				
R Square	0.946535				
Adjusted R Square	0.945421				
Standard Error	9416.349				
Observations	50				
ANOVA					
	df	SS	MS	F	Significance F
Regression	1	7.53E+10	7.53E+10	849.7889	0
Residual	48	4.26E+09	88667637		
Total	49	7.96E+10			
	Coefficients	Standard Error	t Stat	P-value	Lower 95%
Intercept	49032.9	2537.897	19.3203	0	43930.12
R&D Spend	0.8543	0.0293	29.1511	0	0.7954

The given model is significant as all the variables have p-value lesser than 0.05 (5% level of significance).

Fitting a linear regression model to the selected data using
Forward Selection Method.

Step 1:

We first find linear regression output of individual variable with Y

Model 1 : Linear regression output of X_1 variable with Y

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	**0.9729**	**0.9729**						
R Square	**0.946535**							
Adjusted R Square	**0.945421**							
Standard Error	**9416.349**							
Observations	**50**							
ANOVA								
	df	**SS**	**MS**	**F**	**Significance F**			
Regression	**1**	**7.53E+10**	**7.53E+10**	**849.7889**	**0**			
Residual	**48**	**4.26E+09**	**88667637**					
Total	**49**	**7.96E+10**						
	Coefficients	**Standard Error**	**t Stat**	**P-value**	**Lower 95%**	**Upper 95%**	**Lower 95.0%**	**Upper 95.0%**
Intercept	**49032.9**	**2537.897**	**19.3203**	**0**	**43930.12**	**54135.68**	**43930.12**	**54135.68**
R&D Spend	**0.8543**	**0.0293**	**29.1511**	**0**	**0.7954**	**0.9132**	**0.7954**	**0.9132**

Model 2 : Linear regression output of X_2 variable with Y

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	**0.200717**	**0.200717**						
R Square	**0.040287**							
Adjusted R Square	**0.020293**							
Standard Error	**39895.12**							
Observations	**50**							
ANOVA								
	df	**SS**	**MS**	**F**	**Significance F**			
Regression	**1**	**3.21E+09**	**3.21E+09**	**2.01496**	**0.162217**			
Residual	**48**	**7.64E+10**	**1.59E+09**					
Total	**49**	**7.96E+10**						
	Coefficients	**Standard Error**	**t Stat**	**P-value**	**Lower 95%**	**Upper 95%**	**Lower 95.0%**	**Upper 95.0%**
Intercept	**76974.47**	**25320.18**	**3.040044**	**0.003824**	**26064.83**	**127884.1**	**26064.83**	**127884.1**
Administration	**0.288749**	**0.203417**	**1.419493**	**0.162217**	**-0.12025**	**0.697747**	**-0.12025**	**0.697747**

Model 3 : Linear regression output of X_3 variable with Y

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.747766	0.747766						
R Square	0.559154							
Adjusted R Square	0.549969							
Standard Error	27039.13							
Observations	50							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	4.45E+10	4.45E+10	60.88145	0			
Residual	48	3.51E+10	7.31E+08					
Total	49	7.96E+10						
	Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	60003.55	7684.53	7.8084	0	44552.77	75454.33	44552.77	75454.33
Marketing Spend	0.2465	0.0316	7.8027	0	0.183	0.31	0.183	0.31

By considering Coefficient of determination and variables having *p* values (i.e < 0.05), we find Model 1 to be highly significant among all

Step 2:

Selecting X_1 variable, we find linear regression output of X_1 & X_2 variable with Y

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.973557							
R Square	0.947813							
Adjusted R Square	0.945592							
Standard Error	9401.609							
Observations	50							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	2	7.55E+10	3.77E+10	426.8032	7.29E-31			
Residual	47	4.15E+09	88390248					
Total	49	7.96E+10						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	54886.62	6016.718	9.1224	0	42782.54	66990.7	42782.54	66990.7
R&D Spend	0.8621	0.0302	28.5889	0	0.8015	0.9228	0.8015	0.9228
Administration	-0.053	0.0494	-1.0727	0.2889	-0.1524	0.0464	-0.1524	0.0464

Selecting X_1 variable, we find linear regression output of X_1 & X_3 variable with Y

Regression Statistics								
Multiple R	0.97491							
R Square	0.95045							
Adjusted R Square	0.948342							
Standard Error	9160.966							
Observations	50							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	2	7.57E+10	3.78E+10	450.7713	2.16E-31			
Residual	47	3.94E+09	83923295					
Total	49	7.96E+10						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	46975.86	2689.933	17.46358	3.50E-22	41564.42	52387.31	41564.42	52387.31
R&D Spend	0.796584	0.041348	19.26556	0.0000	0.713403	0.879765	0.713403	0.879765
Marketing Spend	0.029908	0.01552	1.927052	0.06003	-0.00131	0.06113	-0.00131	0.06113

By considering Coefficient of determination and variables having *p* values (i.e < 0.05), we find no Model to be significant

Therefore, by Forward selection and Backward Elimination method we get our significant linear regression model as:

$$\mathbf{Y = \beta_0 + \beta_1 X_1}$$

Which is given as:

$$\mathbf{Y = 49032.9 + 0.8543 X_1}$$

THANKYOU