Import libraries

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.formula.api as smf
import statsmodels.api as sm
from statsmodels.graphics.regressionplots import influence_plot
```

Import dataset

In [2]:

```
data_50 = pd.read_csv('50_Startups.csv')
data_50
```

Out[2]:

	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94
5	131876.90	99814.71	362861.36	New York	156991.12
6	134615.46	147198.87	127716.82	California	156122.51
7	130298.13	145530.06	323876.68	Florida	155752.60
8	120542.52	148718.95	311613.29	New York	152211.77
9	123334.88	108679.17	304981.62	California	149759.96
10	101913.08	110594.11	229160.95	Florida	146121.95
11	100671.96	91790.61	249744.55	California	144259.40
12	93863.75	127320.38	249839.44	Florida	141585.52
13	91992.39	135495.07	252664.93	California	134307.35
14	119943.24	156547.42	256512.92	Florida	132602.65
15	114523.61	122616.84	261776.23	New York	129917.04
16	78013.11	121597.55	264346.06	California	126992.93
17	94657.16	145077.58	282574.31	New York	125370.37
18	91749.16	114175.79	294919.57	Florida	124266.90
19	86419.70	153514.11	0.00	New York	122776.86
20	76253.86	113867.30	298664.47	California	118474.03
21	78389.47	153773.43	299737.29	New York	111313.02
22	73994.56	122782.75	303319.26	Florida	110352.25
23	67532.53	105751.03	304768.73	Florida	108733.99
24	77044.01	99281.34	140574.81	New York	108552.04
25	64664.71	139553.16	137962.62	California	107404.34
26	75328.87	144135.98	134050.07	Florida	105733.54
27	72107.60	127864.55	353183.81	New York	105008.31
28	66051.52	182645.56	118148.20	Florida	103282.38
29	65605.48	153032.06	107138.38	New York	101004.64
30	61994.48	115641.28	91131.24	Florida	99937.59
31	61136.38	152701.92	88218.23	New York	97483.56
32	63408.86	129219.61	46085.25	California	97427.84
33	55493.95	103057.49	214634.81	Florida	96778.92

	R&D Spend	Administration	Marketing Spend	State	Profit
34	46426.07	157693.92	210797.67	California	96712.80
35	46014.02	85047.44	205517.64	New York	96479.51
36	28663.76	127056.21	201126.82	Florida	90708.19
37	44069.95	51283.14	197029.42	California	89949.14
38	20229.59	65947.93	185265.10	New York	81229.06
39	38558.51	82982.09	174999.30	California	81005.76
40	28754.33	118546.05	172795.67	California	78239.91
41	27892.92	84710.77	164470.71	Florida	77798.83
42	23640.93	96189.63	148001.11	California	71498.49
43	15505.73	127382.30	35534.17	New York	69758.98
44	22177.74	154806.14	28334.72	California	65200.33
45	1000.23	124153.04	1903.93	New York	64926.08
46	1315.46	115816.21	297114.46	Florida	49490.75
47	0.00	135426.92	0.00	California	42559.73
48	542.05	51743.15	0.00	New York	35673.41
49	0.00	116983.80	45173.06	California	14681.40

Exploratory data analysis

In [3]:

```
data_50.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   R&D Spend        50 non-null     float64
 1   Administration   50 non-null     float64
 2   Marketing Spend  50 non-null     float64
 3   State            50 non-null     object 
 4   Profit           50 non-null     float64
dtypes: float64(4), object(1)
memory usage: 2.1+ KB
```

In [4]:

```
data_1 = data_50.rename({'R&D Spend':'RDS','Administration':'ADMS','Marketing Spend':'MKTS'
data_1
```

Out[4]:

	RDS	ADMS	MKTS	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94
5	131876.90	99814.71	362861.36	New York	156991.12
6	134615.46	147198.87	127716.82	California	156122.51
7	130298.13	145530.06	323876.68	Florida	155752.60
8	120542.52	148718.95	311613.29	New York	152211.77
9	123334.88	108679.17	304981.62	California	149759.96
10	101913.08	110594.11	229160.95	Florida	146121.95
11	100671.96	91790.61	249744.55	California	144259.40
12	93863.75	127320.38	249839.44	Florida	141585.52
13	91992.39	135495.07	252664.93	California	134307.35
14	119943.24	156547.42	256512.92	Florida	132602.65
15	114523.61	122616.84	261776.23	New York	129917.04
16	78013.11	121597.55	264346.06	California	126992.93
17	94657.16	145077.58	282574.31	New York	125370.37
18	91749.16	114175.79	294919.57	Florida	124266.90
19	86419.70	153514.11	0.00	New York	122776.86
20	76253.86	113867.30	298664.47	California	118474.03
21	78389.47	153773.43	299737.29	New York	111313.02
22	73994.56	122782.75	303319.26	Florida	110352.25
23	67532.53	105751.03	304768.73	Florida	108733.99
24	77044.01	99281.34	140574.81	New York	108552.04
25	64664.71	139553.16	137962.62	California	107404.34
26	75328.87	144135.98	134050.07	Florida	105733.54
27	72107.60	127864.55	353183.81	New York	105008.31
28	66051.52	182645.56	118148.20	Florida	103282.38
29	65605.48	153032.06	107138.38	New York	101004.64
30	61994.48	115641.28	91131.24	Florida	99937.59
31	61136.38	152701.92	88218.23	New York	97483.56
32	63408.86	129219.61	46085.25	California	97427.84
33	55493.95	103057.49	214634.81	Florida	96778.92

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

In [5]:

```
data_1.shape
```

Out[5]:

```
(50, 5)
```

In [6]:

```
data_1.describe()
```

Out[6]:

	RDS	ADMS	MKTS	Profit
count	50.000000	50.000000	50.000000	50.000000
mean	73721.615600	121344.639600	211025.097800	112012.639200
std	45902.256482	28017.802755	122290.310726	40306.180338
min	0.000000	51283.140000	0.000000	14681.400000
25%	39936.370000	103730.875000	129300.132500	90138.902500
50%	73051.080000	122699.795000	212716.240000	107978.190000
75%	101602.800000	144842.180000	299469.085000	139765.977500
max	165349.200000	182645.560000	471784.100000	192261.830000

Correlation Analysis

In [7]:

```
data_1.corr()
```

Out[7]:

	RDS	ADMS	MKTS	Profit
RDS	1.000000	0.241955	0.724248	0.972900
ADMS	0.241955	1.000000	-0.032154	0.200717
MKTS	0.724248	-0.032154	1.000000	0.747766
Profit	0.972900	0.200717	0.747766	1.000000

sns.set_style(style='whitegrid') sns.pairplot(data_1)

Model Building

In [8]:

```
model =smf.ols("Profit~RDS+ADMS+MKTS",data = data_1).fit()
```

Model Testing

Finding Coefficient parameters

In [9]:

```
model.params
```

Out[9]:

```
Intercept    50122.192990
RDS              0.805715
ADMS            -0.026816
MKTS             0.027228
dtype: float64
```

Finding Tvalues and Pvalues

In [10]:

```
model.tvalues, np.round(model.pvalues,5)
```

Out[10]:

```
(Intercept     7.626218
 RDS          17.846374
 ADMS         -0.525507
 MKTS          1.655077
 dtype: float64,
 Intercept    0.00000
 RDS          0.00000
 ADMS         0.60176
 MKTS         0.10472
 dtype: float64)
```

Finding rsquared values

In [11]:

```
model.rsquared , model.rsquared_adj # Model accuracy is 94.75%
```

Out[11]:

```
(0.9507459940683246, 0.9475337762901719)
```

Build SLR and MLR models for insignificant variables 'ADMS' and 'MKTS'

Also find their tvalues and pvalues

In [12]:

```
slr_a = smf.ols("Profit~ADMS",data = data_1).fit()
```

In [13]:

```
slr_a.tvalues, slr_a.pvalues # ADMS has in-significant pvalue
```

Out[13]:

```
(Intercept    3.040044
 ADMS         1.419493
 dtype: float64,
 Intercept    0.003824
 ADMS         0.162217
 dtype: float64)
```

In [14]:

```
slr_a = smf.ols("Profit~MKTS",data = data_1).fit()
slr_a.tvalues, slr_a.pvalues  # ADMS has in-significant pvalue
```

Out[14]:

```
(Intercept    7.808356
 MKTS         7.802657
 dtype: float64,
 Intercept    4.294735e-10
 MKTS         4.381073e-10
 dtype: float64)
```

In [15]:

```
mlr_am = smf.ols("Profit~RDS+ADMS+MKTS",data = data_1).fit()
mlr_am.tvalues, mlr_am.pvalues # variable have siganificant pvalues
```

Out[15]:

```
(Intercept     7.626218
 RDS          17.846374
 ADMS         -0.525507
 MKTS          1.655077
 dtype: float64,
 Intercept    1.057379e-09
 RDS          2.634968e-22
 ADMS         6.017551e-01
 MKTS         1.047168e-01
 dtype: float64)
```

Model validation

In [16]:

```
# Two Techniques: 1. colinearity check & 2. Residual Analysis
# 1) Colinearity problem check
# Caluculate VIF = 1/(1-Rsqure) for all independent variables

rsq_r = smf.ols("RDS~ADMS+MKTS",data = data_1).fit().rsquared
vif_r=1/(1-rsq_r)

rsq_a = smf.ols("ADMS~RDS+MKTS",data=data_1).fit().rsquared
vif_a = 1/(1-rsq_a)

rsq_m = smf.ols("MKTS~RDS+ADMS",data=data_1).fit().rsquared
vif_m = 1/(1-rsq_m)
```

Putting the values in Dataframe format

In [17]:

```
d1={'Variables':['RDS','ADMS','MKTS'],'Vif':[vif_r,vif_a,vif_m]}
vif_df = pd.DataFrame(d1)
vif_df
```

Out[17]:

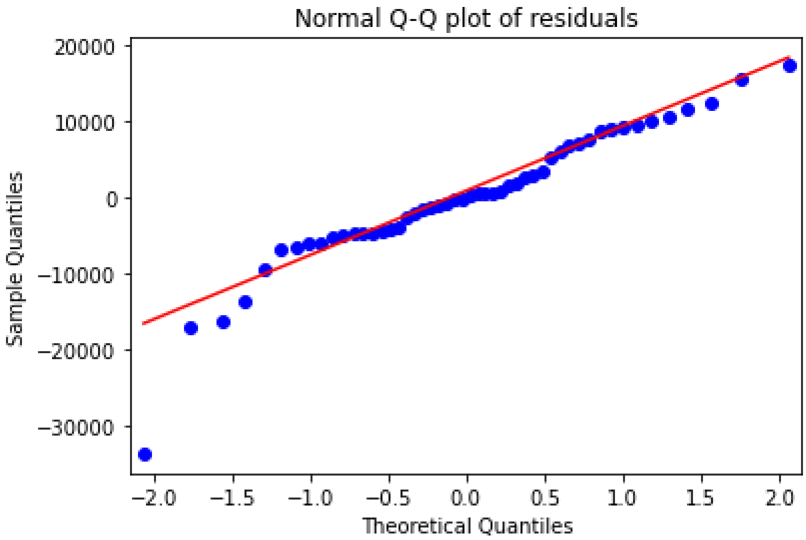
	Variables	Vif
0	RDS	2.468903
1	ADMS	1.175091
2	MKTS	2.326773

Residual Analysis

In [18]:

```
# Test for Normality of Residuals (Q-Q plot) using residual model (model.residual)

sm.qqplot(model.resid, line='q')
plt.title("Normal Q-Q plot of residuals")
plt.show()
```



In [19]:

```
list(np.where(model.resid<-30000))
```

Out[19]:

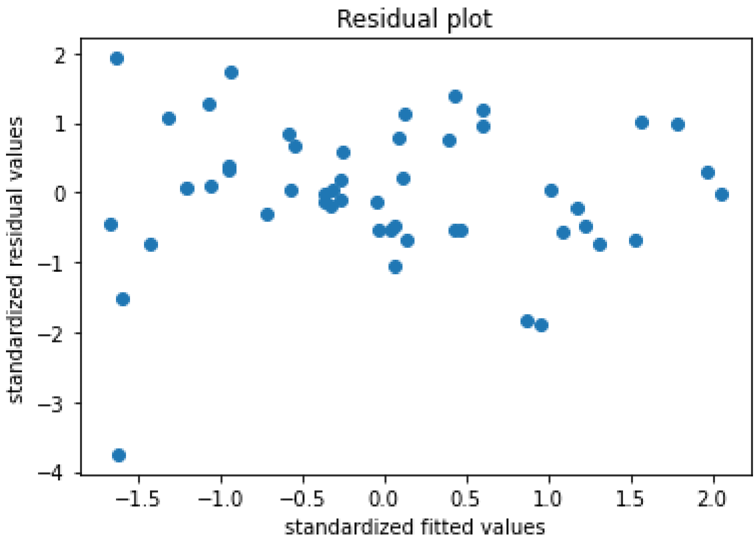
```
[array([49], dtype=int64)]
```

Test for Homoscedasticity or Heteroscadasticity

In [20]:

```
# (plotting model's standerdized fitted values vs standarized residual values)

def standard_values(vals) : return (vals-vals.mean())/vals.std()  # User defined z = (x- mu
plt.scatter(standard_values(model.fittedvalues),standard_values(model.resid))
plt.title('Residual plot')
plt.xlabel('standardized fitted values')
plt.ylabel('standardized residual values')
plt.show()
```



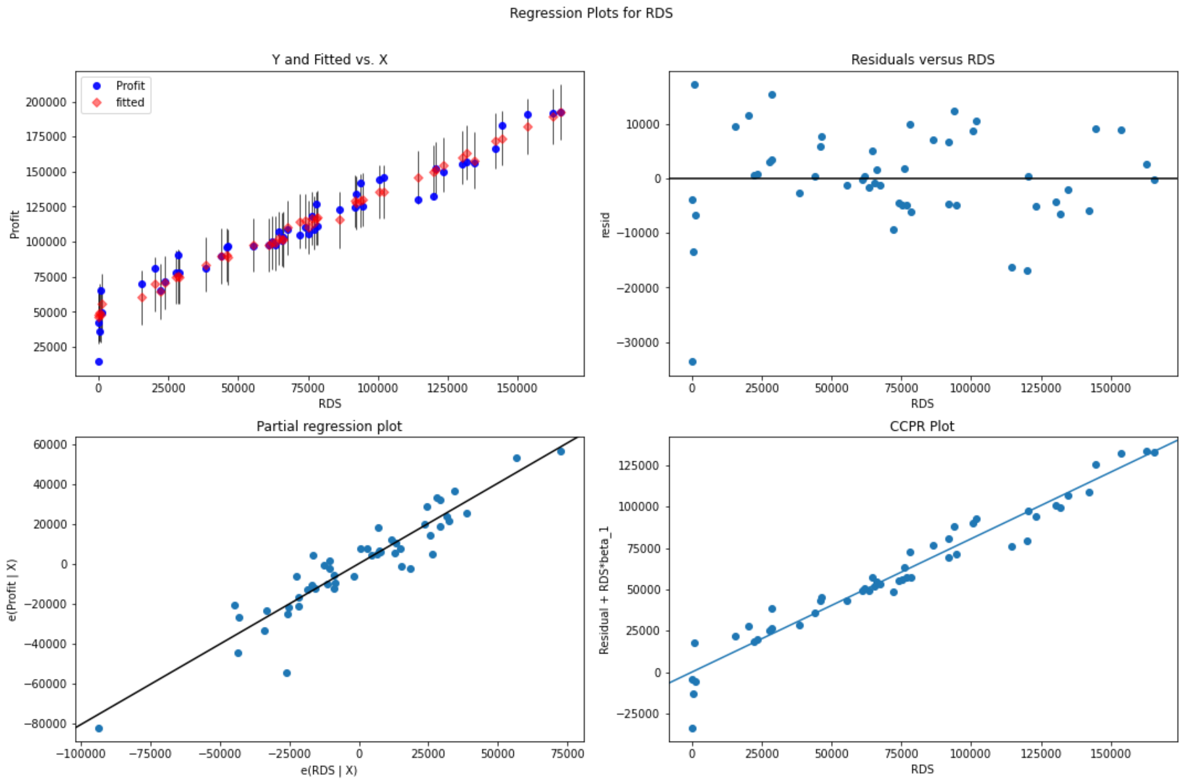
In [21]:

```
# Test for errors or Residuals Vs Regressors or independent 'x' variables or predictors
# using Residual Regression Plots code graphics.plot_regress_exog(model,'x',fig) # exog = x
```

In [22]:

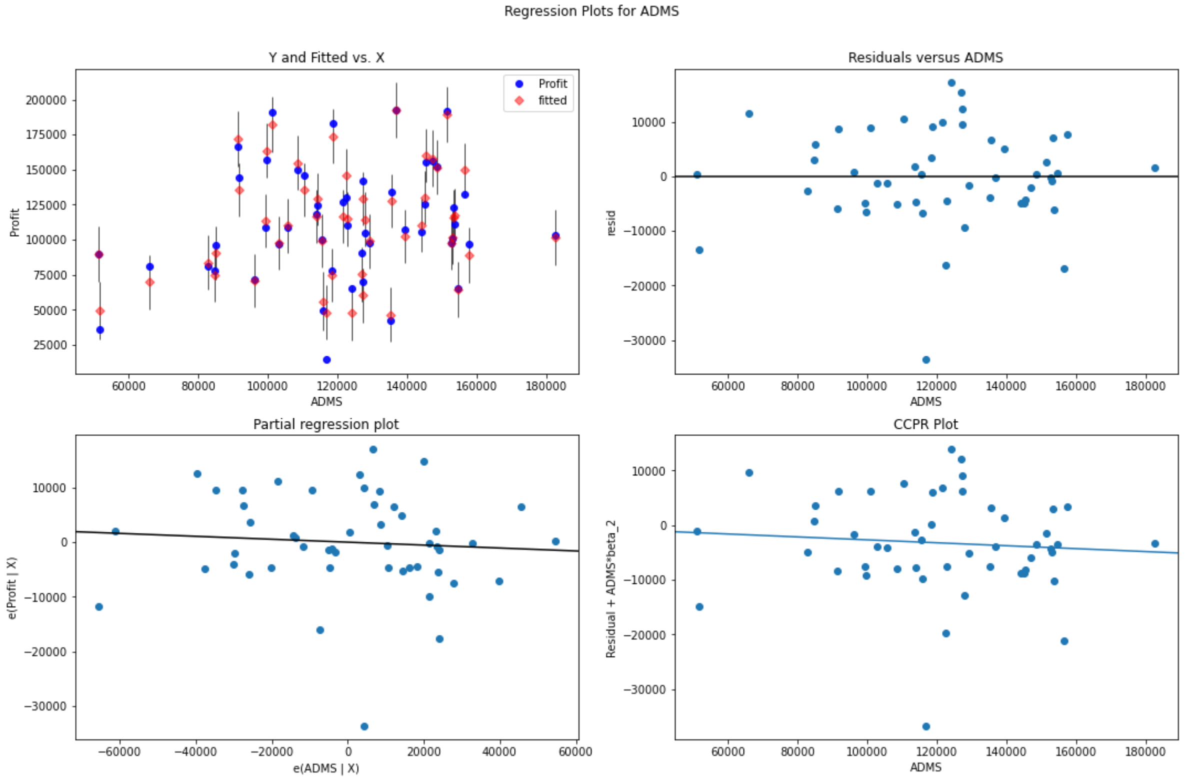
```
# Test for errors or Residuals Vs Regressors or independent 'x' variables or predictors
# using Residual Regression Plots code graphics.plot_regress_exog(model,'x',fig) # e
#exog = x-variable & endog = y-variable

fig=plt.figure(figsize=(15,10))
sm.graphics.plot_regress_exog(model,'RDS',fig=fig)
plt.show()
```



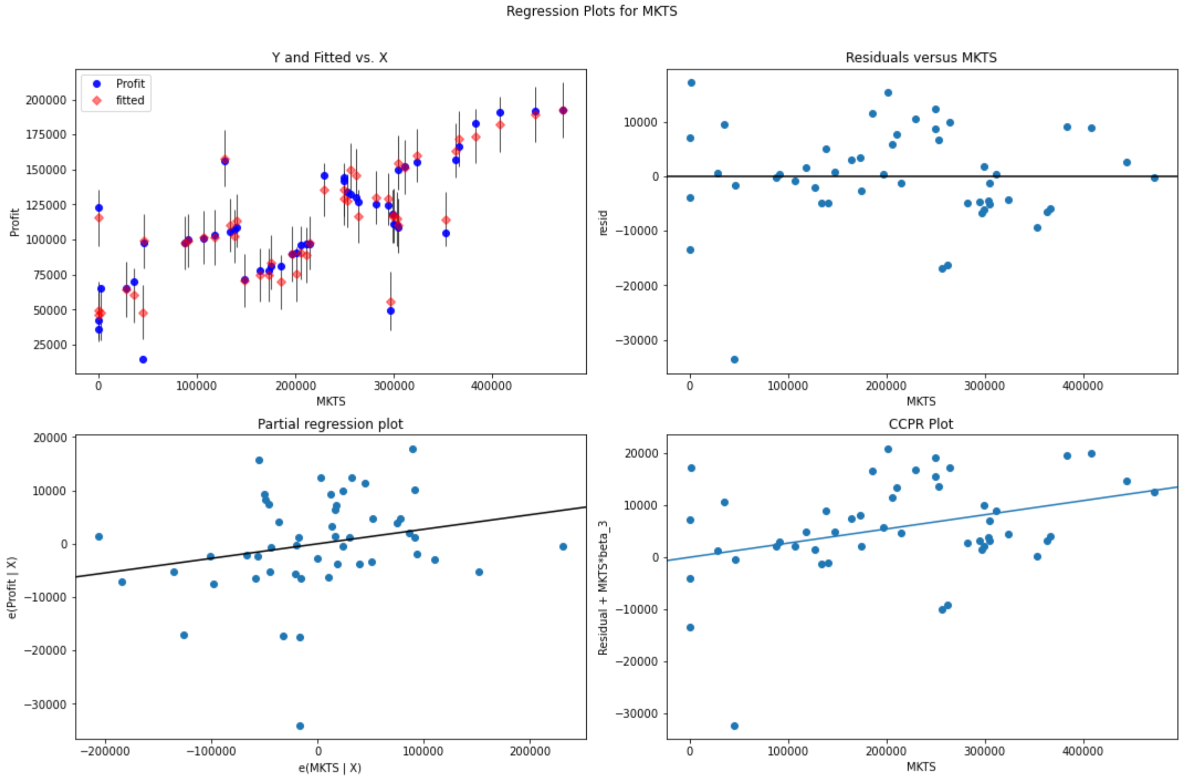
In [23]:

```
fig=plt.figure(figsize=(15,10))
sm.graphics.plot_regress_exog(model,'ADMS',fig=fig)
plt.show()
```



In [24]:

```
fig=plt.figure(figsize=(15,10))
sm.graphics.plot_regress_exog(model,'MKTS',fig=fig)
plt.show()
```



In [25]:

```
# Model Deletion Diagnostatics (checking outliers or Influencers)
# Two Techniques : 1. Cook's Distance & 2, Leverage value
# 1. Cook's Distance : If Cook's distance > 1, then it's on outlier
# Get influencers using cook's distance.
(c,_) = model.get_influence().cooks_distance
c
```

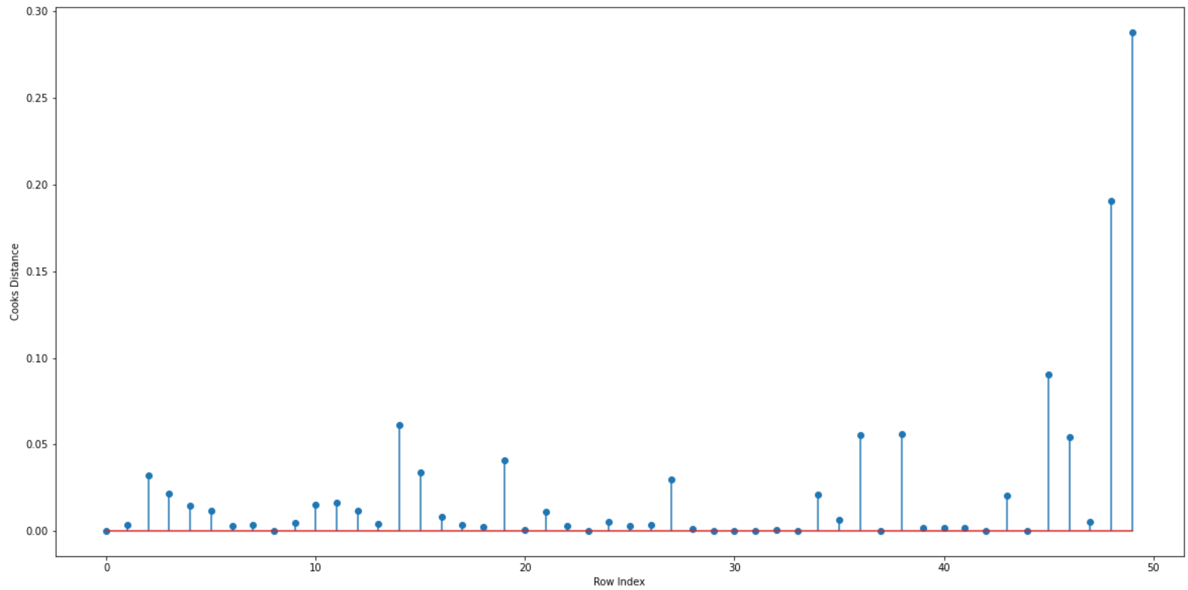
Out[25]:

```
array([3.21825244e-05, 3.27591036e-03, 3.23842699e-02, 2.17206555e-02,
       1.44833032e-02, 1.17158463e-02, 2.91766303e-03, 3.56513444e-03,
       4.04303948e-05, 4.86758017e-03, 1.51064757e-02, 1.63564959e-02,
       1.15516625e-02, 4.01422811e-03, 6.12934253e-02, 3.40013448e-02,
       8.33556413e-03, 3.30534399e-03, 2.16819303e-03, 4.07440577e-02,
       4.25137222e-04, 1.09844352e-02, 2.91768000e-03, 2.76030254e-04,
       5.04643588e-03, 3.00074623e-03, 3.41957068e-03, 2.98396413e-02,
       1.31590664e-03, 1.25992620e-04, 4.18505125e-05, 9.27434786e-06,
       7.08656521e-04, 1.28122674e-04, 2.09815032e-02, 6.69508674e-03,
       5.55314705e-02, 6.55050578e-05, 5.61547311e-02, 1.54279607e-03,
       1.84850929e-03, 1.97578066e-03, 1.36089280e-04, 2.05553171e-02,
       1.23156041e-04, 9.03234206e-02, 5.45303387e-02, 5.33885616e-03,
       1.90527441e-01, 2.88082293e-01])
```

plot the influencers using the stem plot

In [26]:

```
fig=plt.figure(figsize=(20,10))
plt.stem(np.arange(len(data_1)),np.round(c,5))
plt.xlabel('Row Index')
plt.ylabel('Cooks Distance')
plt.show()
```



In [27]:

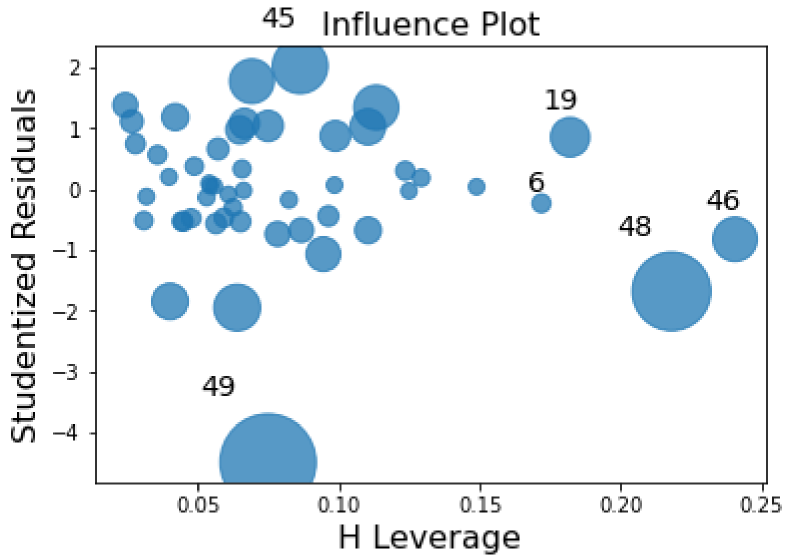
```
# Index and value of influencer where C>0.5
np.argmax(c), np.max(c)
```

Out[27]:

```
(49, 0.28808229275432634)
```

In [28]:

```
# 2. Leverage value using High Influence points : points beyond Leverage_cutoff value are i
influence_plot(model)
plt.show()
```



In [29]:

```
# Leverage Cuttoff Value = 3*(k+1)/n ; k = no.of features/columns & n = no.of datapoints
k=data_1.shape[1]
n=data_1.shape[0]
leverage_cuttoff = (3*(k+1))/n
leverage_cuttoff
```

Out[29]:

```
0.36
```

In [30]:

```
data_1[data_1.index.isin([49])]
```

Out[30]:

	RDS	ADMS	MKTS	State	Profit
49	0.0	116983.8	45173.06	California	14681.4

In [31]:

```
# Improving the Model
# Discard the data points which are influencers and reassign the row number (reset_index(dr

data_2 = data_1.drop(data_1.index[[49]],axis=0).reset_index(drop=True)
data_2
```

Out[31]:

	RDS	ADMS	MKTS	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94
5	131876.90	99814.71	362861.36	New York	156991.12
6	134615.46	147198.87	127716.82	California	156122.51
7	130298.13	145530.06	323876.68	Florida	155752.60
8	120542.52	148718.95	311613.29	New York	152211.77
9	123334.88	108679.17	304981.62	California	149759.96

Model Deletion Diagnostics and Final Model

In [32]:

```
while np.max(c)>0.5 :
    model=smf.ols("Profit~RDS+ADMS+MKTS",data=data_2).fit()
    (c,_)=model.get_influence().cook_distance
    c
    np.argmax(c) , np.max(c)
    data_2=data_2.drop(data_2.index[[np.argmax(c)]],axis=0).reset_index(drop=True)
    data_2
else:
    final_model = smf.ols("Profit~RDS+ADMS+MKTS",data=data_2).fit()
    final_model.rsquared, final_model.aic
    print("Thus model accuracy is improved to",final_model.rsquared)
```

```
Thus model accuracy is improved to 0.9613162435129847
```

In [33]:

```
final_model.rsquared
```

Out[33]:

```
0.9613162435129847
```

In [34]:

```
data_2
```

Out[34]:

	RDS	ADMS	MKTS	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94
5	131876.90	99814.71	362861.36	New York	156991.12
6	134615.46	147198.87	127716.82	California	156122.51
7	130298.13	145530.06	323876.68	Florida	155752.60
8	120542.52	148718.95	311613.29	New York	152211.77
9	123334.88	108679.17	304981.62	California	149759.96
10	101913.08	110594.11	229160.95	Florida	146121.95
11	100671.96	91790.61	249744.55	California	144259.40
12	93863.75	127320.38	249839.44	Florida	141585.52
13	91992.39	135495.07	252664.93	California	134307.35
14	119943.24	156547.42	256512.92	Florida	132602.65
15	114523.61	122616.84	261776.23	New York	129917.04
16	78013.11	121597.55	264346.06	California	126992.93
17	94657.16	145077.58	282574.31	New York	125370.37
18	91749.16	114175.79	294919.57	Florida	124266.90
19	86419.70	153514.11	0.00	New York	122776.86
20	76253.86	113867.30	298664.47	California	118474.03
21	78389.47	153773.43	299737.29	New York	111313.02
22	73994.56	122782.75	303319.26	Florida	110352.25
23	67532.53	105751.03	304768.73	Florida	108733.99
24	77044.01	99281.34	140574.81	New York	108552.04
25	64664.71	139553.16	137962.62	California	107404.34
26	75328.87	144135.98	134050.07	Florida	105733.54
27	72107.60	127864.55	353183.81	New York	105008.31
28	66051.52	182645.56	118148.20	Florida	103282.38
29	65605.48	153032.06	107138.38	New York	101004.64
30	61994.48	115641.28	91131.24	Florida	99937.59
31	61136.38	152701.92	88218.23	New York	97483.56
32	63408.86	129219.61	46085.25	California	97427.84
33	55493.95	103057.49	214634.81	Florida	96778.92

	RDS	ADMS	MKTS	State	Profit
34	46426.07	157693.92	210797.67	California	96712.80
35	46014.02	85047.44	205517.64	New York	96479.51
36	28663.76	127056.21	201126.82	Florida	90708.19
37	44069.95	51283.14	197029.42	California	89949.14
38	20229.59	65947.93	185265.10	New York	81229.06
39	38558.51	82982.09	174999.30	California	81005.76
40	28754.33	118546.05	172795.67	California	78239.91
41	27892.92	84710.77	164470.71	Florida	77798.83
42	23640.93	96189.63	148001.11	California	71498.49
43	15505.73	127382.30	35534.17	New York	69758.98
44	22177.74	154806.14	28334.72	California	65200.33
45	1000.23	124153.04	1903.93	New York	64926.08
46	1315.46	115816.21	297114.46	Florida	49490.75
47	0.00	135426.92	0.00	California	42559.73
48	542.05	51743.15	0.00	New York	35673.41

Model Predictions

In [35]:

```
# say New data for Prediction is
new_data = pd.DataFrame({'RDS':70000,'ADMS':90000,'MKTS':140000},index=[0])
new_data
```

Out[35]:

	RDS	ADMS	MKTS
0	70000	90000	140000

Manual prediction of price

In [36]:

```
final_model.predict(new_data)
```

Out[36]:

```
0    108727.154753
dtype: float64
```

In [37]:

```
# Automatic prediction of price with 90.02% accuracy
pred_y = final_model.predict(data_2)
pred_y
```

```
32    100325.741335
33     98962.303136
34     90552.307809
35     91709.288672
36     77080.554255
37     90722.503244
38     71433.021956
39     85147.375646
40     76625.510303
41     76492.145175
42     72492.394974
43     62592.049718
44     67025.731107
45     50457.297206
46     58338.443625
47     49375.776655
48     51658.096812
dtype: float64
```

In [38]:

```
# table containg R^2 value for each prepared model
d2={'Prep_Models':['Model','Final_Model'],'Rsquared':[model.rsquared,final_model.rsquared]}
table = pd.DataFrame(d2)
table
```

Out[38]:

	Prep_Models	Rsquared
0	Model	0.950746
1	Final_Model	0.961316

In [ ]: