In [27]:

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

executed in 15ms, finished 09:45:26 2021-11-26

In [2]:

```
data = pd.read_csv("50_Startups.csv")
data.head()
```

executed in 79ms, finished 09:24:30 2021-11-26

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

In [3]:

```
data.info()
```

executed in 43ms, finished 09:24:50 2021-11-26

```
<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 [5]:

```
data.isna().sum()
```

executed in 30ms, finished 09:25:12 2021-11-26

Out[5]:

```
R&D Spend          0
Administration     0
Marketing Spend    0
State              0
Profit             0
dtype: int64
```

In [6]:

```
data.corr()
```

executed in 34ms, finished 09:25:39 2021-11-26

Out[6]:

	R&D Spend	Administration	Marketing Spend	Profit
R&D Spend	1.000000	0.241955	0.724248	0.972900
Administration	0.241955	1.000000	-0.032154	0.200717
Marketing Spend	0.724248	-0.032154	1.000000	0.747766
Profit	0.972900	0.200717	0.747766	1.000000

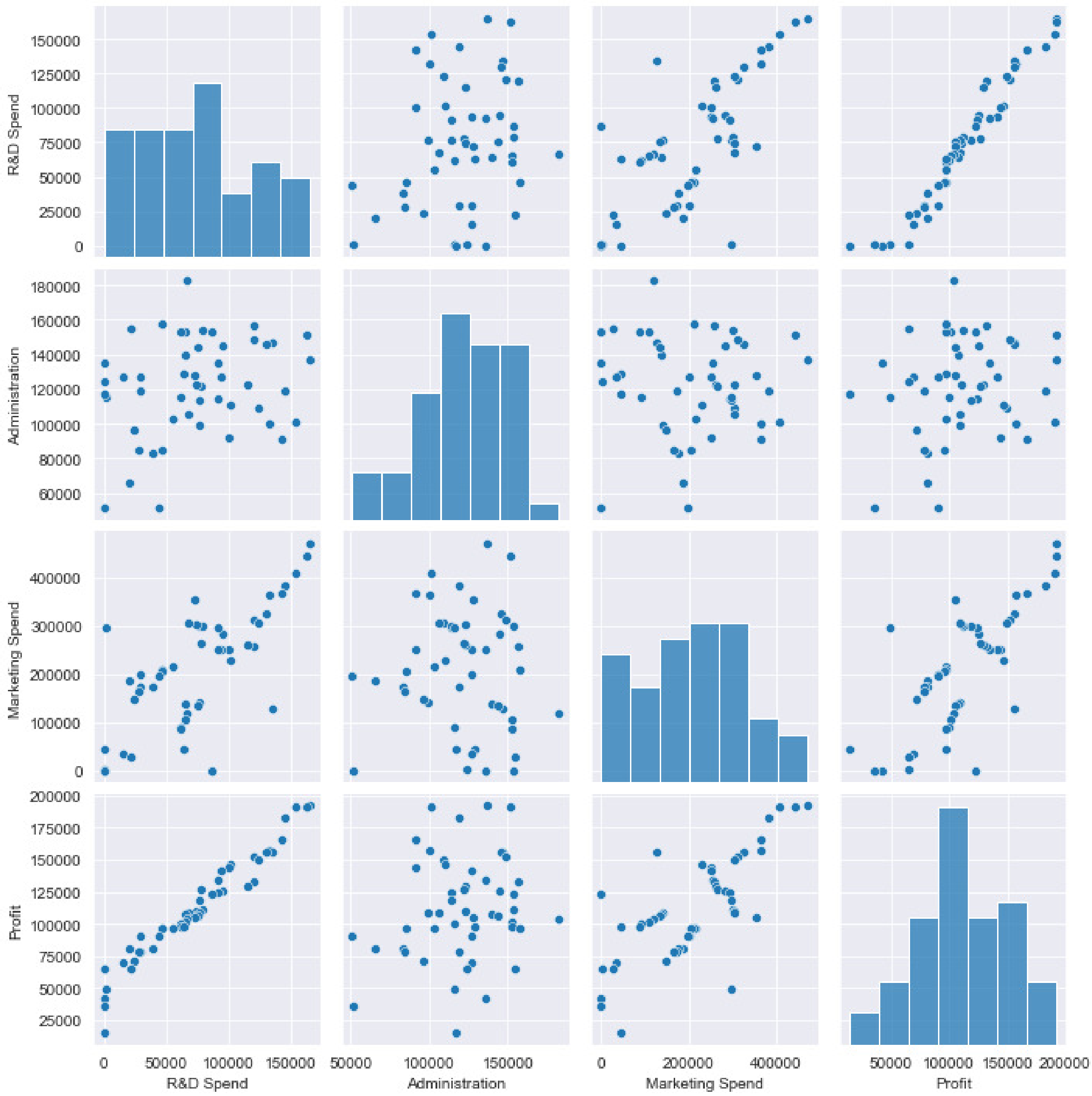
In [7]:

```
sns.set_style(style='darkgrid')
sns.pairplot(data)
```

executed in 5.33s, finished 09:26:19 2021-11-26

Out[7]:

```
<seaborn.axisgrid.PairGrid at 0x29acda0e610>
```



In [12]:

```
corrMatrix = data.corr()
```

executed in 22ms, finished 09:39:52 2021-11-26

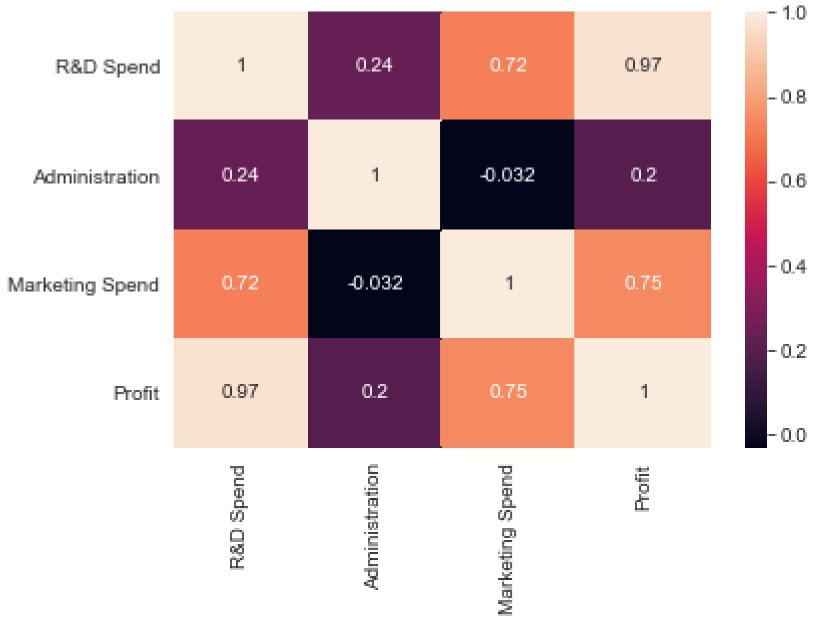
In [13]:

```
sns.heatmap(corrMatrix, annot=True)
```

executed in 577ms, finished 09:40:08 2021-11-26

Out[13]:

```
<AxesSubplot:>
```



	R&D Spend	Administration	Marketing Spend	Profit
R&D Spend	1	0.24	0.72	0.97
Administration	0.24	1	-0.032	0.2
Marketing Spend	0.72	-0.032	1	0.75
Profit	0.97	0.2	0.75	1

In [14]:

```
data = pd.get_dummies(data, columns=['State'])
```

executed in 17ms, finished 09:40:37 2021-11-26

In [15]:

```
X = data[['R&D Spend','Administration', 'Marketing Spend', 'State_California', 'State_Flori
Y = data[['Profit']]
```

executed in 23ms, finished 09:41:18 2021-11-26

In [19]:

```
model = sm.OLS(Y, X).fit()
predictions = model.predict(X)
```

executed in 22ms, finished 09:42:37 2021-11-26

In [20]:

```
model.summary()
```

executed in 48ms, finished 09:42:54 2021-11-26

Out[20]:

OLS Regression Results

```
       Dep. Variable:           Profit          R-squared:     0.951
              Model:               OLS      Adj. R-squared:     0.945
             Method:     Least Squares         F-statistic:     169.9
               Date:  Fri, 26 Nov 2021  Prob (F-statistic):  1.34e-27
               Time:          09:42:54      Log-Likelihood:   -525.38
   No. Observations:                50                 AIC:     1063.
       Df Residuals:                44                 BIC:     1074.
           Df Model:                 5
    Covariance Type:         nonrobust

                          coef     std err        t   P>|t|     [0.025     0.975]
        R&D Spend       0.8060       0.046   17.369   0.000      0.712      0.900
   Administration      -0.0270       0.052   -0.517   0.608     -0.132      0.078
  Marketing Spend       0.0270       0.017    1.574   0.123     -0.008      0.062
 State_California    5.013e+04    6884.820    7.281   0.000   3.62e+04     6.4e+04
    State_Florida    5.032e+04    7251.767    6.940   0.000   3.57e+04   6.49e+04
   State_New York    5.008e+04    6952.587    7.204   0.000   3.61e+04   6.41e+04

        Omnibus:  14.782     Durbin-Watson:      1.283
  Prob(Omnibus):   0.001  Jarque-Bera (JB):     21.266
           Skew:  -0.948          Prob(JB):   2.41e-05
       Kurtosis:   5.572          Cond. No.   2.45e+06
```

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.45e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

In [21]:

```
infl = model.get_influence()
```

executed in 22ms, finished 09:43:20 2021-11-26

In [22]:

```
summ_data = infl.summary_frame()
```

executed in 96ms, finished 09:43:40 2021-11-26

In [23]:

```
summ_data.sort_values('cooks_d', ascending=False)
```

executed in 100ms, finished 09:44:03 2021-11-26

Out[23]:

	dfb_R&D Spend	dfb_Administration	dfb_Marketing Spend	dfb_State_California	dfb_State_Florida	dfb_St
49	0.578956	-0.114232	0.080954	-0.566028	-0.246221	-
48	-0.112734	0.701599	0.418630	-0.783828	-0.801849	-(
45	-0.212843	0.091394	-0.189969	0.095382	0.140857	(
14	-0.221204	-0.257240	0.142195	0.267421	0.086725	(
36	-0.379353	0.189523	0.218405	-0.107545	0.053174	-(
38	-0.189819	-0.313449	0.109261	0.320201	0.309091	(
15	-0.208289	0.066627	0.071114	-0.002577	0.007587	-(
46	0.434369	-0.142646	-0.364064	0.106828	0.034265	(
19	0.252210	0.039342	-0.342025	0.009492	0.035168	(
27	0.271462	-0.146112	-0.339679	0.169919	0.186504	(
2	0.197811	-0.174765	-0.013702	0.080109	0.147293	(
3	0.110000	-0.049701	0.073783	-0.035947	-0.056370	
43	-0.090858	0.058906	-0.085052	0.024294	0.044997	(
10	0.186811	-0.137400	-0.159583	0.116978	0.216711	
12	0.069420	0.010085	-0.048411	-0.017490	0.098185	-(
34	-0.196836	0.242310	0.173390	-0.138400	-0.212207	-(
11	0.152695	-0.197500	-0.063320	0.226735	0.131358	(
16	-0.055403	0.029857	0.116725	0.037230	-0.065807	-(
4	-0.153851	0.163418	0.047737	-0.104342	-0.152925	-(
5	-0.081405	0.105665	-0.033055	-0.046292	-0.031296	-(
21	0.134008	-0.163215	-0.156986	0.162458	0.165701	(
35	-0.049889	-0.102221	0.034443	0.099727	0.095288	(
13	0.005568	0.054245	0.040526	-0.007538	-0.074145	-(
9	-0.087676	0.064819	0.009105	-0.073333	-0.016881	-(
26	-0.047078	-0.040934	0.086363	0.018763	-0.038037	(
24	-0.077455	0.089257	0.087987	-0.087996	-0.089711	-(
47	0.046774	-0.046937	0.031389	-0.029301	-0.001347	(
17	0.035255	-0.075381	-0.061166	0.081487	0.083661	(
7	-0.045130	-0.041183	0.001538	0.056695	0.017003	(
25	0.012139	0.046367	-0.028528	0.012690	-0.030258	-(
1	0.020841	0.046268	0.058280	-0.049608	-0.085419	-(
22	0.057093	-0.025665	-0.065599	0.028443	-0.010225	
18	-0.003027	0.019777	-0.017913	-0.008655	-0.050050	-(
6	-0.095187	0.004676	0.082074	-0.021299	-0.002912	(

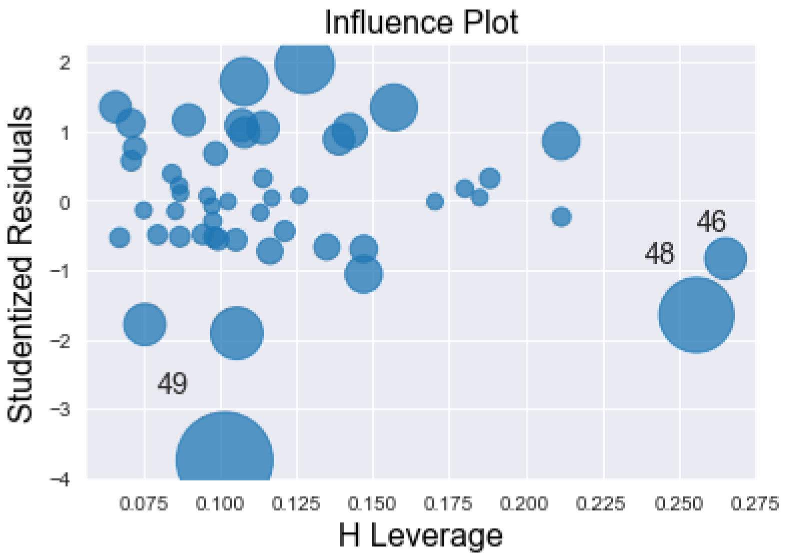
	dfb_R&D Spend	dfb_Administration	dfb_Marketing Spend	dfb_State_California	dfb_State_Florida	dfb_St
40	-0.063137	0.020142	0.044268	0.021070	-0.010930	-(
41	-0.022659	-0.051009	-0.006525	0.059567	0.087625	(
39	0.011576	0.050925	-0.005414	-0.074335	-0.047479	-(
28	-0.004892	0.054512	-0.014043	-0.041421	-0.021855	-(
20	-0.018937	0.001269	0.034093	0.009204	-0.011262	-(
32	-0.029851	0.005201	0.042128	-0.025767	-0.013770	-(
23	0.017345	0.004924	-0.019954	-0.003001	-0.014564	-(
33	0.005720	0.010764	-0.000214	-0.012256	-0.024085	-(
42	-0.010791	-0.008202	0.004902	0.018523	0.009901	(
44	-0.005257	0.013394	-0.004732	-0.001883	-0.006265	-(
29	-0.001210	-0.011523	0.007045	0.007881	0.006612	(
8	0.000992	0.008446	0.005091	-0.010483	-0.010897	-(
37	0.001322	-0.016087	-0.000850	0.018397	0.013923	(
30	0.004911	-0.003183	-0.008237	0.004519	0.008048	(
0	-0.000680	-0.001084	-0.002943	0.002627	0.003023	(
31	-0.000298	-0.002100	0.001610	0.001322	0.001054	(

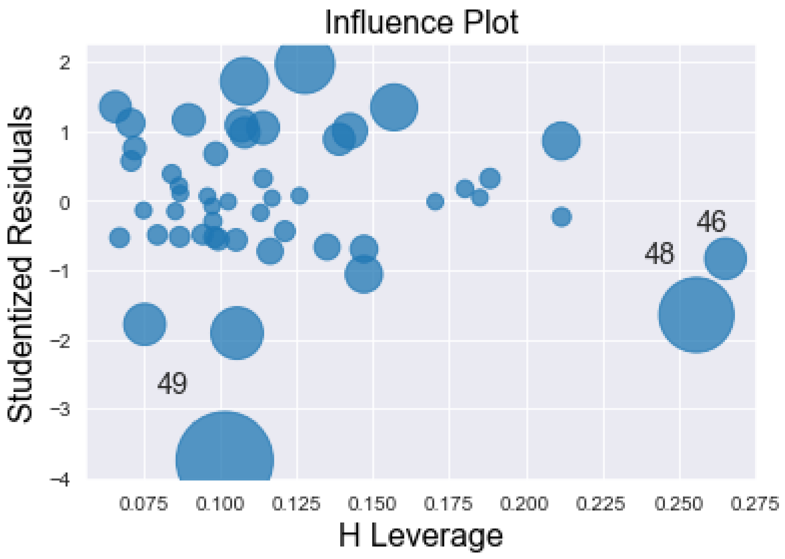
In [24]:

```
infl.plot_influence()
```

executed in 933ms, finished 09:44:22 2021-11-26

Out[24]:





In [25]:

```
vif = pd.DataFrame()
```

executed in 7ms, finished 09:44:41 2021-11-26

In [28]:

```
vif["VIF Factor"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
```

executed in 22ms, finished 09:45:32 2021-11-26

In [29]:

```
vif["features"] = X.columns
```

executed in 15ms, finished 09:45:49 2021-11-26

In [30]:

```
vif.round(1)
```

executed in 35ms, finished 09:46:03 2021-11-26

Out[30]:

	VIF Factor	features
0	2.5	R&D Spend
1	1.2	Administration
2	2.4	Marketing Spend
3	9.0	State_California
4	9.4	State_Florida
5	9.2	State_New York

In [32]:

```
new_X = data[['R&D Spend', 'Marketing Spend', 'State_California', 'State_Florida', 'State_N
```

executed in 19ms, finished 09:46:43 2021-11-26

In [33]:

```
new_model = sm.OLS(Y, new_X).fit()
new_predictions = new_model.predict(new_X)
```

executed in 20ms, finished 09:46:57 2021-11-26

In [34]:

```
new_model.summary()
```

executed in 63ms, finished 09:47:07 2021-11-26

Out[34]:

```
OLS Regression Results

     Dep. Variable:            Profit       R-squared:     0.950
             Model:               OLS   Adj. R-squared:     0.946
            Method:     Least Squares      F-statistic:     215.8
              Date:  Fri, 26 Nov 2021 Prob (F-statistic):  9.72e-29
              Time:          09:47:07   Log-Likelihood:   -525.53
  No. Observations:                50              AIC:     1061.
      Df Residuals:                45              BIC:     1071.
          Df Model:                 4
   Covariance Type:         nonrobust

                        coef    std err       t   P>|t|     [0.025     0.975]
        R&D Spend     0.7967      0.042  18.771   0.000      0.711      0.882
  Marketing Spend     0.0298      0.016   1.842   0.072     -0.003      0.062
 State_California  4.696e+04   3119.471  15.053   0.000    4.07e+04   5.32e+04
    State_Florida   4.71e+04   3670.129  12.833   0.000    3.97e+04   5.45e+04
   State_New York  4.694e+04   3342.591  14.043   0.000    4.02e+04   5.37e+04

       Omnibus:  14.640    Durbin-Watson:     1.257
 Prob(Omnibus):   0.001 Jarque-Bera (JB):    21.037
          Skew:  -0.938         Prob(JB):  2.70e-05
      Kurtosis:   5.565         Cond. No.  9.45e+05


Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.45e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [35]:

```
new_df = data.drop(data.index[[49,48]])
```

executed in 11ms, finished 09:47:55 2021-11-26

In [37]:

```
new_X = new_df[['R&D Spend', 'Marketing Spend', 'State_California', 'State_Florida', 'State
new_Y = new_df[['Profit']]
```

executed in 10ms, finished 09:48:34 2021-11-26

In [38]:

```
final_model = sm.OLS(new_Y, new_X).fit()
predictions = final_model.predict(new_X)
```

executed in 23ms, finished 09:48:52 2021-11-26

In [39]:

```
final_model.summary()
```

executed in 58ms, finished 09:49:03 2021-11-26

Out[39]:

```
OLS Regression Results

    Dep. Variable:            Profit          R-squared:     0.961
            Model:               OLS      Adj. R-squared:     0.958
           Method:     Least Squares         F-statistic:     265.9
             Date:  Fri, 26 Nov 2021   Prob (F-statistic):  1.02e-29
             Time:          09:49:03      Log-Likelihood:   -494.30
 No. Observations:                48                 AIC:     998.6
     Df Residuals:                43                 BIC:     1008.
         Df Model:                 4
  Covariance Type:         nonrobust

                        coef    std err        t   P>|t|    [0.025     0.975]
        R&D Spend     0.7692      0.035   22.072   0.000     0.699      0.840
  Marketing Spend     0.0251      0.013    1.908   0.063    -0.001      0.052
 State_California  5.183e+04   2710.866   19.120   0.000   4.64e+04   5.73e+04
    State_Florida  5.046e+04   3078.590   16.391   0.000   4.43e+04   5.67e+04
   State_New York   5.09e+04   2936.767   17.333   0.000    4.5e+04   5.68e+04

        Omnibus:  0.133      Durbin-Watson:     1.645
  Prob(Omnibus):  0.936   Jarque-Bera (JB):     0.304
           Skew:  0.097           Prob(JB):     0.859
       Kurtosis:  2.661           Cond. No.  1.02e+06


Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.02e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [40]:

```
X_sqrt = np.sqrt(new_df[['R&D Spend', 'Marketing Spend', 'State_California', 'State_Florida
```

executed in 19ms, finished 09:49:31 2021-11-26

In [42]:

```
model3 = sm.OLS(new_Y, X_sqrt).fit()
predictions3 = model3.predict(X_sqrt)
```

executed in 16ms, finished 09:49:57 2021-11-26

In [43]:

```
model3.summary()
```

executed in 43ms, finished 09:49:59 2021-11-26

Out[43]:

OLS Regression Results

Dep. Variable:	Profit	**R-squared:**	0.887
Model:	OLS	**Adj. R-squared:**	0.877
Method:	Least Squares	**F-statistic:**	84.44
Date:	Fri, 26 Nov 2021	**Prob (F-statistic):**	8.67e-20
Time:	09:49:59	**Log-Likelihood:**	-519.91
No. Observations:	48	**AIC:**	1050.
Df Residuals:	43	**BIC:**	1059.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
R&D Spend	340.5455	25.777	13.211	0.000	288.560	392.531
Marketing Spend	20.0497	15.481	1.295	0.202	-11.170	51.270
State_California	1.836e+04	6267.224	2.930	0.005	5724.219	3.1e+04
State_Florida	1.692e+04	7013.669	2.413	0.020	2779.320	3.11e+04
State_New York	1.908e+04	6591.247	2.894	0.006	5782.772	3.24e+04

Omnibus:	7.588	**Durbin-Watson:**	0.777
Prob(Omnibus):	0.023	**Jarque-Bera (JB):**	7.161
Skew:	0.941	**Prob(JB):**	0.0279
Kurtosis:	3.197	**Cond. No.**	3.04e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.04e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

In [44]:

```
Y_sqrt = np.sqrt(new_df['Profit'])
```

executed in 18ms, finished 09:50:17 2021-11-26

In [45]:

```
model4 = sm.OLS(Y_sqrt, new_X).fit()
predictions4 = model4.predict(new_X)
```

executed in 13ms, finished 09:50:30 2021-11-26

In [46]:

```
model4.summary()
```

executed in 45ms, finished 09:50:42 2021-11-26

Out[46]:

```
OLS Regression Results

   Dep. Variable:          Profit         R-squared:     0.954
           Model:             OLS    Adj. R-squared:     0.950
          Method:   Least Squares       F-statistic:     223.3
            Date: Fri, 26 Nov 2021 Prob (F-statistic): 3.68e-28
            Time:        09:50:42    Log-Likelihood:   -185.87
No. Observations:              48               AIC:     381.7
    Df Residuals:              43               BIC:     391.1
        Df Model:               4
 Covariance Type:       nonrobust

                       coef   std err       t   P>|t|    [0.025    0.975]
      R&D Spend      0.0012  5.64e-05  20.622   0.000     0.001     0.001
Marketing Spend   2.473e-05  2.13e-05   1.159   0.253 -1.83e-05  6.78e-05
State_California   241.0032     4.390  54.894   0.000   232.149   249.857
  State_Florida    240.7325     4.986  48.283   0.000   230.678   250.787
 State_New York    240.9886     4.756  50.669   0.000   231.397   250.580

      Omnibus:  4.530     Durbin-Watson:    1.406
Prob(Omnibus):  0.104  Jarque-Bera (JB):    3.371
         Skew: -0.532          Prob(JB):    0.185
     Kurtosis:  3.745         Cond. No. 1.02e+06
```

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.02e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

In [47]:

```
model5 = sm.OLS(Y_sqrt, X_sqrt).fit()
predictions5 = model5.predict(X_sqrt)
```

executed in 14ms, finished 09:51:00 2021-11-26

In [48]:

```
model5.summary()
```

executed in 54ms, finished 09:51:11 2021-11-26

Out[48]:

```
OLS Regression Results

      Dep. Variable:           Profit          R-squared:     0.929
              Model:              OLS      Adj. R-squared:     0.923
             Method:    Least Squares         F-statistic:     141.7
               Date:  Fri, 26 Nov 2021  Prob (F-statistic):  3.64e-24
               Time:         09:51:11      Log-Likelihood:   -196.16
   No. Observations:               48                 AIC:     402.3
       Df Residuals:               43                 BIC:     411.7
           Df Model:                4
    Covariance Type:        nonrobust

                         coef  std err       t   P>|t|   [0.025   0.975]
       R&D Spend       0.5271    0.030  17.371   0.000    0.466    0.588
 Marketing Spend       0.0231    0.018   1.270   0.211   -0.014    0.060
State_California     187.8689    7.377  25.465   0.000  172.991  202.747
   State_Florida     187.0162    8.256  22.652   0.000  170.366  203.666
  State_New York     189.8076    7.759  24.463   0.000  174.160  205.455

      Omnibus:  7.976      Durbin-Watson:     1.243
Prob(Omnibus):  0.019   Jarque-Bera (JB):     7.007
         Skew:  0.870           Prob(JB):    0.0301
     Kurtosis:  3.692           Cond. No.  3.04e+03


Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.04e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```