CSE303 (Section 1)

[Spring 2022]

Lab Assignment Submission Report

Assignment Title: LAB 06

Submitted by:

Rizvee Hassan Prito

2019-3-60-041

1. Screenshots

Implementing Linear regression model:

Using Statsmodel OLS Method:

Taking input for file location/path from user:

```
Enter file location/path:

D:\Courses\CSE 303\literacy_rate.csv
```

Taking input for columns from user:

```
Enter columns:
Country
Newspapers
Radios
TVsets
literacyRate
```

Taking input for independent variable names (predictor variables) from user:

```
Enter independent variable names:
Newspapers
Radios
TVsets
```

Taking input for dependent variable names (response variables) from user:

```
Enter dependent variable names:
literacyRate
```

Dataset selected by user:

Given dataset:

	Country	Newspapers	Radios	TVsets	literacyRate
0	Czech Republic	280	266	228	0.98
1	Italy	142	230	201	0.93
2	Kenya	10	114	2	0.25
3	Norway	391	313	227	0.99
4	Panama	86	329	82	0.79
5	Philippines	17	42	11	0.72
6	Tunisia	21	49	16	0.32
7	USA	314	1695	472	0.99
8	Russia	333	430	185	0.99
9	Venezuela	91	182	89	0.82

Independent variables selected by user:

Independent Variables:

	Newspapers	Radios	TVsets
0	280	266	228
1	142	230	201
2	10	114	2
3	391	313	227
4	86	329	82
5	17	42	11
6	21	49	16
7	314	1695	472
8	333	430	185
9	91	182	89

Dependent variable selected by user:

Dependent Variable:

	literacyRate
0	0.98
1	0.93
2	0.25
3	0.99
4	0.79
5	0.72
6	0.32
7	0.99
8	0.99
9	0.82

Output for Statsmodel OLS Method:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.699
Model:                            OLS   Adj. R-squared:                  0.548
Method:                 Least Squares   F-statistic:                     4.640
Date:                Thu, 14 Apr 2022   Prob (F-statistic):            0.0526
Time:                        13:20:35   Log-Likelihood:                 5.1604
No. Observations:                  10   AIC:                            -2.321
Df Residuals:                       6   BIC:                            -1.110
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.5149      0.094      5.496      0.002       0.286       0.744
x1             0.0005      0.001      0.626      0.554      -0.002       0.003
x2            -0.0004      0.000     -1.076      0.323      -0.001       0.000
x3             0.0020      0.002      1.282      0.247      -0.002       0.006
==============================================================================
Omnibus:                        0.776   Durbin-Watson:                   1.822
Prob(Omnibus):                  0.679   Jarque-Bera (JB):                0.628
Skew:                          -0.265   Prob(JB):                        0.731
Kurtosis:                       1.893   Cond. No.                     1.02e+03
==============================================================================
```

Using Scikit-learn library:

Taking input for file location/path from user:

```
Enter file location/path:

D:\Courses\CSE 303\USA_Housing.csv
```

Taking input for columns from user:

```
Enter columns:

Avg. Area Income

Avg. Area House Age

Avg. Area Number of Rooms

Avg. Area Number of Bedrooms

Area Population

Price

Address
```

Taking input for independent variable names (predictor variables) from user:

```
Enter independent variable names:

Avg. Area Income

Avg. Area House Age

Avg. Area Number of Rooms

Avg. Area Number of Bedrooms

Area Population
```

Taking input for dependent variable name (response variables) from user:

```
Enter dependent variable names:

Price
```

Dataset selected by user:

```
Given dataset:
      Avg. Area Income  ...                                            Address
0          79545.45857  ...  208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1          79248.64245  ...  188 Johnson Views Suite 079\nLake Kathleen, CA...
2          61287.06718  ...  9127 Elizabeth Stravenue\nDanieltown, WI 06482...
3          63345.24005  ...                          USS Barnett\nFPO AP 44820
4          59982.19723  ...                        USNS Raymond\nFPO AE 09386
...                ...  ...                                               ...
4995       60567.94414  ...                USNS Williams\nFPO AP 30153-7653
4996       78491.27543  ...             PSC 9258, Box 8489\nAPO AA 42991-3352
4997       63390.68689  ...  4215 Tracy Garden Suite 076\nJoshualand, VA 01...
4998       68001.33124  ...                          USS Wallace\nFPO AE 73316
4999       65510.58180  ...  37778 George Ridges Apt. 509\nEast Holly, NV 2...

[5000 rows x 7 columns]
```

Independent variables selected by user:

```
Independent Variables:

      Avg. Area Income  ...  Area Population
0          79545.45857  ...      23086.80050
1          79248.64245  ...      40173.07217
2          61287.06718  ...      36882.15940
3          63345.24005  ...      34310.24283
4          59982.19723  ...      26354.10947
...                ...  ...              ...
4995       60567.94414  ...      22837.36103
4996       78491.27543  ...      25616.11549
4997       63390.68689  ...      33266.14549
4998       68001.33124  ...      42625.62016
4999       65510.58180  ...      46501.28380

[5000 rows x 5 columns]
```

Dependent variable selected by user:

```
             Price
0     1.059034e+06
1     1.505891e+06
2     1.058988e+06
3     1.260617e+06
4     6.309435e+05
...            ...
4995  1.060194e+06
4996  1.482618e+06
4997  1.030730e+06
4998  1.198657e+06
4999  1.298950e+06

[5000 rows x 1 columns]
```

Outputs:

```
Intercept: [-2646630.5310881]

Coefficients: [[2.16604083e+01 1.65809651e+05 1.20329408e+05 2.19309558e+03
  1.52858855e+01]]

Comparing actual values of Y and predicted values of Y:
          Actual     Predicted
0    8.942511e+05  9.713052e+05
1    9.329794e+05  9.547172e+05
2    9.207479e+05  9.075616e+05
3    6.918549e+05  4.932514e+05
4    7.327332e+05  7.178465e+05
..            ...           ...
995  7.549606e+05  8.391244e+05
996  1.205568e+06  1.142525e+06
997  6.682555e+05  4.986132e+05
998  1.398760e+06  1.235228e+06
999  1.277381e+06  1.156382e+06

[1000 rows x 2 columns]
MAE:  82657.94604671016
MSE:  10549721685.894304
RMSE:  287.5029496313214
R-squared:  0.9146454505152328
```

2. Learning outcomes:

From this lab, I have learned how to implement linear regression model using OLS function of Statsmodel library and using Scikit-learn library of pyhton. OLS stands for Ordinary Least Squares. We can find from the output from Statsmodel OLS method that the more R-squared and F-statistic values are higher the more the model fits good for the given dataset. By using the Scikit-learn learn library, we can find the intercept of regression line, value of R-squared and can compare actual values of y and predicted values of y etc. We can also find the coefficients of independent variables by using any of those libraries.