Ex. No.: 5 Write a Program for Simple Linear Regression

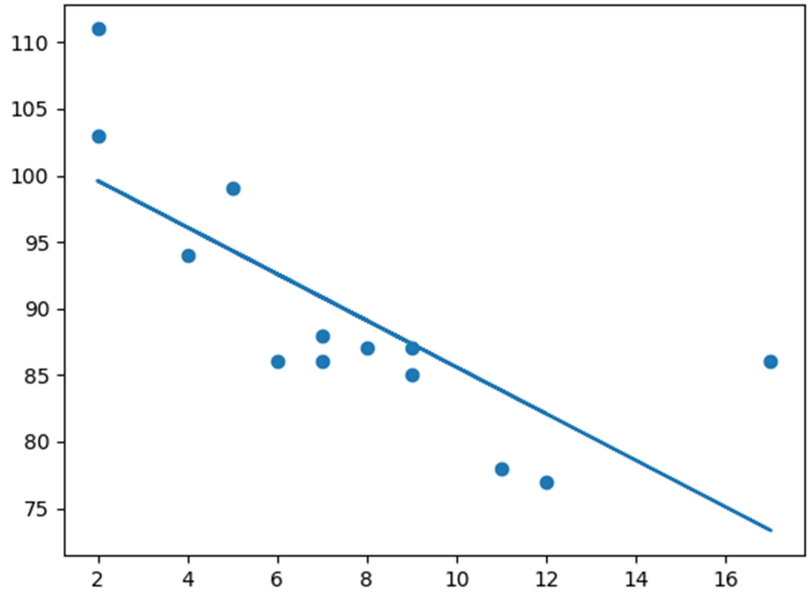
AIM:

To write a Python program to calculate the simple linear regression by using the given series and csv files.

(i) Program (By using the given series):

```
import matplotlib.pyplot as plt
from scipy import stats
x = [5,7,8,7,2,17,2,9,4,11,12,9,6]
y = [99,86,87,88,111,86,103,87,94,78,77,85,86]
slope, intercept, r, p, std_err = stats.linregress(x, y)
defmyfunc(x):
return slope * x + intercept
mymodel = list(map(myfunc, x))
plt.plot(x, mymodel)
plt.scatter(x, y)
plt.show()
```

Output:



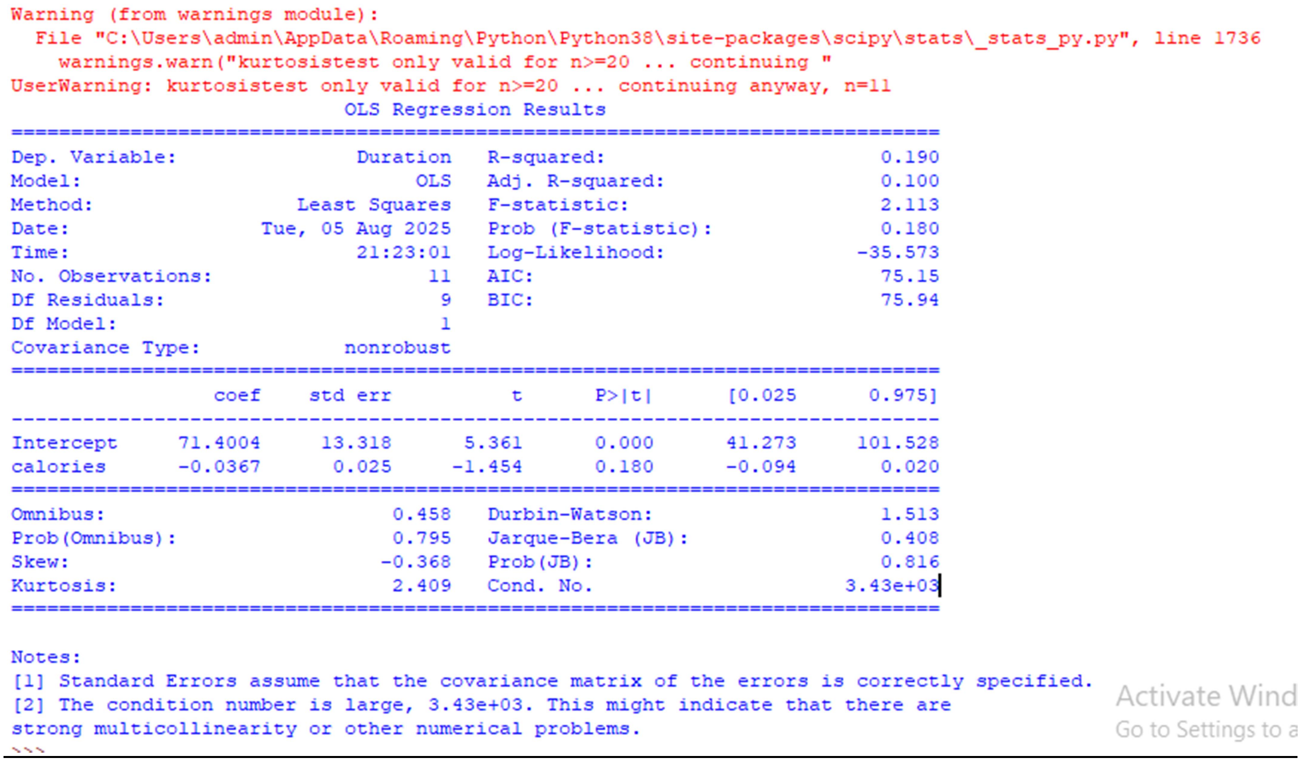
.

(ii) Program (By using the csv file):

```
import pandas as pd
import statsmodels.formula.api as smf
df=pd.read_csv("Book.csv", header=0, sep=",")
model=smf.ols('Duration ~ calories', data=df)
```

```
results = model.fit()
print(results.summary())
```

Output:



```
Warning (from warnings module):
  File "C:\Users\admin\AppData\Roaming\Python\Python38\site-packages\scipy\stats\_stats_py.py", line 1736
    warnings.warn("kurtosistest only valid for n>=20 ... continuing "
UserWarning: kurtosistest only valid for n>=20 ... continuing anyway, n=11
                            OLS Regression Results
==============================================================================
Dep. Variable:               Duration   R-squared:                       0.190
Model:                            OLS   Adj. R-squared:                  0.100
Method:                 Least Squares   F-statistic:                     2.113
Date:                Tue, 05 Aug 2025   Prob (F-statistic):              0.180
Time:                        21:23:01   Log-Likelihood:                -35.573
No. Observations:                  11   AIC:                             75.15
Df Residuals:                       9   BIC:                             75.94
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     71.4004     13.318      5.361      0.000      41.273     101.528
calories      -0.0367      0.025     -1.454      0.180      -0.094       0.020
==============================================================================
Omnibus:                        0.458   Durbin-Watson:                   1.513
Prob(Omnibus):                  0.795   Jarque-Bera (JB):                0.408
Skew:                          -0.368   Prob(JB):                        0.816
Kurtosis:                       2.409   Cond. No.                     3.43e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.43e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The linear regression function can be rewritten mathematically as:
Duration =-0.0367 * calories + 71.4004 [Y=aX+b]. Here P-value is 0.180. Which is greater than 0.05 and also R-Squared value is 0.190. Hence we conclude that there is a no significant relationship between the Duration and calories. Hence the regression line does not fit for the data.

Regression: The term regression is used to find the relationship between variables. In Machine Learning and in statistical modeling, that relationship is used to predict the outcome of events.

Least Square Method: Linear regression uses the least square method. The concept is to draw a line through all the plotted data points. The line is positioned in a way that it minimizes the distance to all of the data points. The distance is called residuals or errors.

Regression Table: The output from linear regression can be summarized in a regression table. The content of the table includes:

1. Information about the model.
2. Coefficients of the linear regression function.
3. Regression statistics.
4. Statistics of the coefficients from the linear regression function.

5. Other information that we will not cover in this module.
[smf.ols: statsmodels formula and Ordinary Least Squares]
Dep.Variable = Dependent Variable
Model = ols
Date and Time: shows the date and time the output was calculated.
Coef = Coefficient. It is the output of the linear regression function.
std err = Standard error
t = t is the t-value of the coefficients.
[0.025 0.975] represents the confidence interval of the coefficients.

P-value:P>t is called the P-value. The P-value is a statistical measure that helps determine the significance of results. If P-value <= 0.05, it indicates that there is a significant relationship between the two variables. If P-value > 0.05, it indicates that there is no significant relationship between the two variables.

R-Squared and Adjusted R-Squared: R-Squared and Adjusted R-Squared describes how well the linear regression model fits the data points. The value of R-Squared is always between 0 to 1 (0% to 100%). A **high R-Squared value** means that many data points are close to the linear regression function line. A **low R-Squared value** means that the regression function line does not fit the data well.

What is Duration if a calorie is: 56, 45, 65.

Program:

```
def Duration(calories):
   return (-0.0367 * calories + 71.4004)
print(Duration(56))
print(Duration(45))
print(Duration(65))
```

Output:

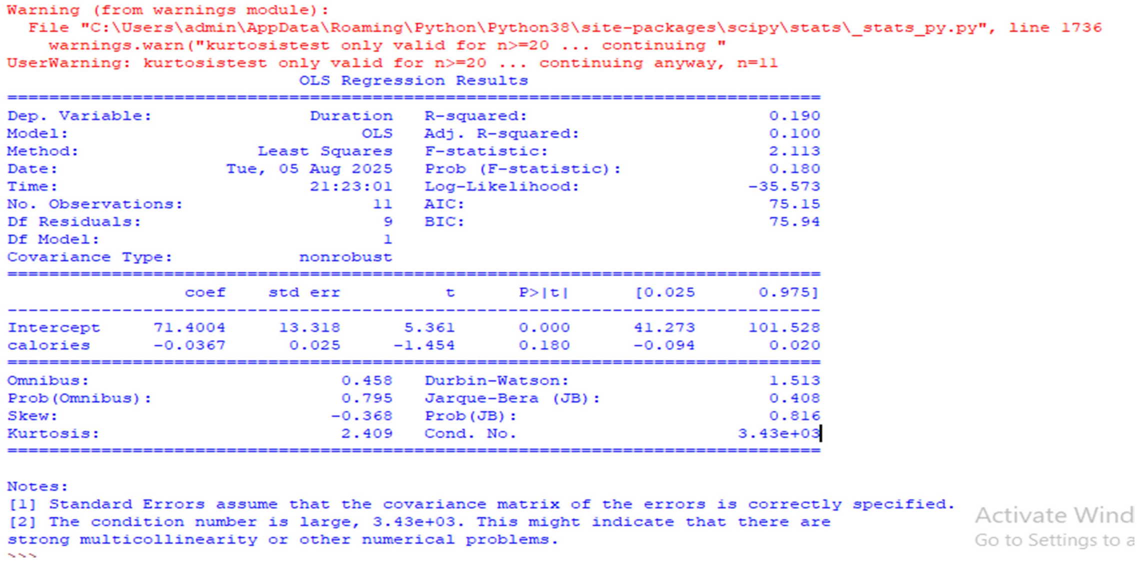
```
69.3452
69.7489
69.01490000000001
```

(iii) Program (By using the csv file):

```
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats
import statsmodels.formula.api as smf
df=pd.read_csv("Book.csv", header=0, sep=",")
x=df["Duration"]
y=df["calories"]
slope, intercept, r, p, std_err = stats.linregress(x, y)
def myfunc(x):
  return slope * x + intercept
mymodel = list(map(myfunc, x))
print(mymodel)
model=smf.ols('Duration ~ calories', data=df)
results = model.fit()
print(results.summary())
plt.plot(x, mymodel)
plt.scatter(x, y)
plt.ylim(ymin=0,ymax=700)
plt.xlim(xmin=0,xmax=80)
plt.xlabel("Duration")
plt.ylabel("calories")
plt.show()
```

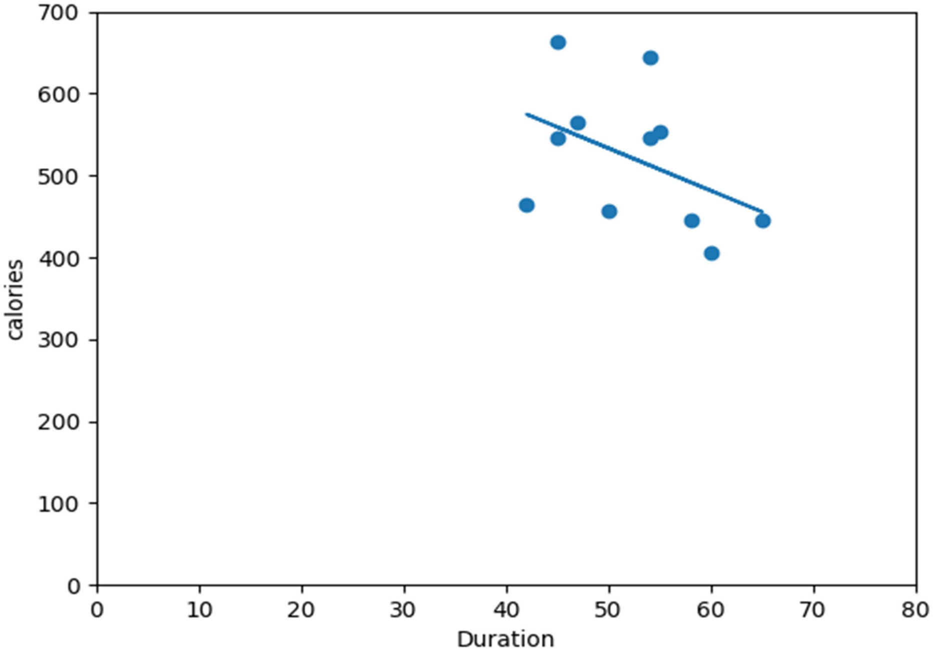
Output:

```
[481.3214412495563, 533.1407525736599, 559.0504082357118,
512.4130280440185, 507.2310969116081, 491.68530351437704,
455.4117855875045, 559.0504082357118, 574.5962016329429,
548.6865459708911, 512.4130280440185]
```



```
Warning (from warnings module):
  File "C:\Users\admin\AppData\Roaming\Python\Python38\site-packages\scipy\stats\_stats_py.py", line 1736
    warnings.warn("kurtosistest only valid for n>=20 ... continuing "
UserWarning: kurtosistest only valid for n>=20 ... continuing anyway, n=11
                            OLS Regression Results
==============================================================================
Dep. Variable:               Duration   R-squared:                       0.190
Model:                            OLS   Adj. R-squared:                  0.100
Method:                 Least Squares   F-statistic:                     2.113
Date:                Tue, 05 Aug 2025   Prob (F-statistic):              0.180
Time:                        21:23:01   Log-Likelihood:                -35.573
No. Observations:                  11   AIC:                             75.15
Df Residuals:                       9   BIC:                             75.94
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     71.4004     13.318      5.361      0.000      41.273     101.528
calories      -0.0367      0.025     -1.454      0.180      -0.094       0.020
==============================================================================
Omnibus:                        0.458   Durbin-Watson:                   1.513
Prob(Omnibus):                  0.795   Jarque-Bera (JB):                0.408
Skew:                          -0.368   Prob(JB):                        0.816
Kurtosis:                       2.409   Cond. No.                     3.43e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.43e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
>>>
```


700
600
500
400
300
200
100
0
calories
0
10
20
30
40
50
60
70
80
Duration