Name : Sagar Patil
Roll No : 8061
BE Computer

Assignment No. 1

1.1 Title

Assignment based on Linear Regression.

1.2 Problem Definition:

The following table shows the results of a recently conducted study on the correlation of the number of hours spent driving with the risk of developing acute backache. Find the equation of the best fit line for this data.

1.3 Prerequisite:

Basic of Python, Data Mining Algorithm

1.4 Software Requirements:

Anaconda with Python 3.7

1.5 Hardware Requirement:

PIV, 2GB RAM, 500 GB HDD, Lenovo A13-4089Model.

1.6 Learning Objectives:

Learn Linear Regression for given different Dataset

1.7 Outcomes:

After completion of this assignment students are able to understand the How to find the correlation between to Two variable, How to Calculate Accuracy of the Linear Model and how to plot graph using matplotlib.

1.8 Theory Concepts:

1.8.1 Linear Regression

Regression analysis is used in stats to find trends in data. For example, you might guess that there's a connection between how much you eat and how much you weight; regression analysis can help you quantify that.

What is Linear Regression?

In a cause and effect relationship, the **independent variable** is the cause, and the **dependent variable** is the effect. **Least squares linear regression** is a method for predicting the value of a dependent variable *Y*, based on the value of an independent variable *X*.

Prerequisites for Regression

Simple linear regression is appropriate when the following conditions are satisfied.

- The dependent variable *Y* has a linear relationship to the independent variable *X*. To check this, make sure that the XY scatterplot is linear and that the residual plot shows a random pattern. For each value of X, the probability distribution of Y has the same standard deviation σ.
- When this condition is satisfied, the variability of the residuals will be relatively constant across all values of X, which is easily checked in a residual plot.
- For any given value of X,
    - The Y values are independent, as indicated by a random pattern on the residual plot.
    - The Y values are roughly normally distributed. A little skewness is ok if the sample size is large. A histogram or a dotplot will show the shape of the distribution.

The Least Squares Regression Line

Linear regression finds the straight line, called the **least squares regression line** or LSRL, that best represents observations in a bivariate data set. Suppose *Y* is a dependent variable, and *X* is an independent variable. The population regression line is:

$$Y = B_0 + B_1X$$

where B_0 is a constant, B_1 is the regression coefficient, X is the value of the independent variable, and Y is the value of the dependent variable.

Given a random sample of observations, the population regression line is estimated by:

$$\hat{y} = b_0 + b_1x$$

where b_0 is a constant, b_1 is the regression coefficient, x is the value of the independent variable, and $\hat{y}$ is the *predicted*value of the dependent variable.

How to Define a Regression Line

Normally, you will use a computational tool - a software package (e.g., Excel) or a graphing calculator - to find b_0 and b_1. You enter the *X* and *Y* values into your program or calculator, and the tool solves for each parameter.

In the unlikely event that you find yourself on a desert island without a computer or a graphing calculator, you can solve for b_0 and b_1 "by hand". Here are the equations.

$$b_1 = \Sigma [(x_i - x)(y_i - y)] / \Sigma [(x_i - x)^2]$$

$$b_1 = r * (s_y / s_x)$$

$$b_0 = y - b_1 * x$$

where b_0 is the constant in the regression equation, b_1 is the regression coefficient, r is the correlation between x and y, x_i is the *X* value of observation *i*, y_i is the *Y* value of observation *i*, x is the mean of *X*, y is the mean of *Y*, s_x is the standard deviation of *X*, and s_y is the standard deviation of *Y*.

Properties of the Regression Line

When the regression parameters (b_0 and b_1) are defined as described above, the regression line has the following properties.

- The line minimizes the sum of squared differences between observed values (the *y* values) and predicted values (the ŷ values computed from the regression equation).
- The regression line passes through the mean of the *X* values (x) and through the mean of the *Y* values (y).
- The regression constant (b_0) is equal to the y intercept of the regression line.
- The regression coefficient (b_1) is the average change in the dependent variable (*Y*) for a 1-unit change in the independent variable (*X*). It is the slope of the regression line.

The least squares regression line is the only straight line that has all of these properties.

The Coefficient of Determination

The **coefficient of determination** (denoted by R^2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

- The coefficient of determination ranges from 0 to 1.
- An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.
- An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.10 means that 10 percent of the variance in *Y* is predictable from *X*; an R^2 of 0.20 means that 20 percent is predictable; and so on.

The formula for computing the coefficient of determination for a linear regression model with one independent variable is given below.

Coefficient of determination. The coefficient of determination (R^2) for a linear regression model with one independent variable is:

$$R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y) \}^2$$

where N is the number of observations used to fit the model, Σ is the summation symbol, x_i is the x value for observation i, x is the mean x value, y_i is the y value for observation i, y is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

If you know the linear correlation (r) between two variables, then the coefficient of determination (R^2) is easily computed using the following formula: $R^2 = r^2$.

Standard Error

The **standard error** about the regression line (often denoted by SE) is a measure of the average amount that the regression equation over- or under-predicts. The higher the coefficient of determination, the lower the standard error; and the more accurate predictions are likely to be.

Example

Last year, five randomly selected students took a math aptitude test before they began their statistics course. The Statistics Department has three questions.

- What linear regression equation best predicts statistics performance, based on math aptitude scores?
- If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?
- How well does the regression equation fit the data?

How to Find the Regression Equation

In the table below, the x_i column shows scores on the aptitude test. Similarly, the y_i column shows statistics grades. The last two columns show deviations scores - the difference between the student's score and the average score on each test. The last two rows show sums and mean scores that we will use to conduct the regression analysis.

Student	x_i	y_i	(x_i-x)	(y_i-y)
1	95	85	17	8
2	85	95	7	18
3	80	70	2	-7
4	70	65	-8	-12
5	60	70	-18	-7
Sum	390	385		
Mean	78	77		

And for each student, we also need to compute the squares of the deviation scores (the last two columns in the table below).

Student	x_i	y_i	$(x_i-x)^2$	$(y_i-y)^2$
1	95	85	289	64
2	85	95	49	324
3	80	70	4	49
4	70	65	64	144
5	60	70	324	49
Sum	390	385	730	630
Mean	78	77		

And finally, for each student, we need to compute the product of the deviation scores.

Student	x_i	y_i	$(x_i-x)(y_i-y)$
1	95	85	136
2	85	95	126
3	80	70	-14
4	70	65	96
5	60	70	126
Sum	390	385	470
Mean	78	77	

The regression equation is a linear equation of the form: $\hat{y} = b_0 + b_1x$. To conduct a regression analysis, we need to solve for b_0 and b_1. Computations are shown below. Notice that all of our inputs for the regression analysis come from the above three tables.

First, we solve for the regression coefficient (b_1):

$$b_1 = \Sigma [(x_i - x)(y_i - y)] / \Sigma [(x_i - x)^2]$$

$$b_1 = 470/730$$

$$b_1 = 0.644$$

Once we know the value of the regression coefficient (b_1), we can solve for the regression slope (b_0):

$$b_0 = y - b_1 * x$$

$$b_0 = 77 - (0.644)(78)$$

$$b_0 = 26.768$$

Therefore, the regression equation is: $\hat{y} = 26.768 + 0.644x$.

How to Use the Regression Equation

Once you have the regression equation, using it is a snap. Choose a value for the independent variable (*x*), perform the computation, and you have an estimated value ($\hat{y}$) for the dependent variable.

In our example, the independent variable is the student's score on the aptitude test. The dependent variable is the student's statistics grade. If a student made an 80 on the aptitude test, the estimated statistics grade ($\hat{y}$) would be:

$$\hat{y} = b_0 + b_1 x$$

$$\hat{y} = 26.768 + 0.644x = 26.768 + 0.644 * 80$$

$$\hat{y} = 26.768 + 51.52 = 78.288$$

How to Find the Coefficient of Determination

Whenever you use a regression equation, you should ask how well the equation fits the data. One way to assess fit is to check the coefficient of determination, which can be computed from the following formula.

$$R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y) \}^2$$

where N is the number of observations used to fit the model, Σ is the summation symbol, x_i is the x value for observation i, x is the mean x value, y_i is the y value for observation i, y is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

Computations for the sample problem of this lesson are shown below. We begin by computing the standard deviation of x (σ_x):

$$\sigma_x = \text{sqrt} [\Sigma (x_i - x)^2 / N]$$

$$\sigma_x = \text{sqrt}(730/5) = \text{sqrt}(146) = 12.083$$

Next, we find the standard deviation of y, (σ_y):

$$\sigma_y = \text{sqrt} [\Sigma (y_i - y)^2 / N]$$

$$\sigma_y = \text{sqrt}(630/5) = \text{sqrt}(126) = 11.225$$

And finally, we compute the coefficient of determination (R^2):

$$R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y) \}^2$$

$$R^2 = [(1/5) * 470 / (12.083 * 11.225)]^2$$

$$R^2 = (94 / 135.632)^2 = (0.693)^2 = 0.48$$

A coefficient of determination equal to 0.48 indicates that about 48% of the variation in statistics gradescan be explained by the relationship to math aptitude scores. This would be considered a good fit to the data, in the sense that it would substantially improve an educator's ability to predict student performance in statistics class.

Residuals

The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the **residual** (e). Each data point has one residual.

$$\text{Residual} = \text{Observed value} - \text{Predicted value}$$
$$e = y - \hat{y}$$

Both the sum and the mean of the residuals are equal to zero. That is, $\Sigma\, e = 0$ and $e = 0$.

What is Best Fit Line- A line of best fit (or "trend" line) is a straight line that best represents the data on a scatter plot. This line may pass through some of the points, none of the points, or all of the points.

1.9 Given Dataset in Our Definition-

Number of hours spent driving (x)	Risk score on a scale of 0-100 (y)
10	95
9	80
2	10
15	50
10	45
16	98
11	38
16	93

1.10 Algorithm

1. Import the Required Packages
2. Read Given Dataset
3. Import the Linear Regression and Create object of it
4. Find the Accuracy of Model using Score Function
5. Predict the value using Regressor Object
6. Take input from user.
7. Calculate the value of y
8. Draw Scatter Plot

Conclusion :

Thus we learn that to how to find the trend of data using X as Independent Variable and Y is and Dependent Variable by using Linear Regression.

CODE:

```
#import the packges
import matplotlib.pyplot as plt
import pandas as pd
#Read Dataset
dataset=pd.read_csv("hours.csv")
#index read
x=dataset.iloc[:,:-1].values #slice all column
y=dataset.iloc[:,1].values #last Column

#import packages of LR
from sklearn.linear_model import LinearRegression
regressor=LinearRegression() #create object of LR

# Fit Function
regressor.fit(x,y)

#score Function
Accuracy=regressor.score(x,y)*100
print('Accuracy')
print(Accuracy)

#Predict Function
y_pred=regressor.predict([[10]])
print(y_pred)

#input from user
hours=int(input("Enter the no of hours"))

# Coefficient
# intercept
eq=regressor.coef_*hours+regressor.intercept_
print("Risk Score",eq[0])
```

```
plt.plot(x,y,'o')
plt.plot(x,regressor.predict(x));
plt.show()
```

OUTPUT: