JSPM
Page No. :

Assignment NO : 1

Title -
Assignment based on linear regression

Problem Definition :
The following table shows the results of recently conducted study on the correlation of the no. of hours spent dealing with the risk of developing acute backache. Find the equation of the best fit line for this data

Prerequisite
Basic of Python, data mining algorithm

Software Requirement
Anaconda with Python 3.7

Hardware Requirement
P4, 2GB Ram, 500 GB HDD

Theory concepts

Linear regression is used for finding linear relationship between target and one or more predictors there are two types of linear regression. simple and multiple

Least Squares regression line :-
Linear regression finds the straight line, called the least square regression line or LSRL that best represent observation in bivariate data set. Suppose y is a dependent variable and x is independent variable. the population regression line is :-

$$y = B_0 + B_1 x$$

How to Define a Regression line :
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 equation :-

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

$$b_1 = r^+ (S_y / S_x)$$

$$b_0 = y - b_1 + x$$

Standard Errors
The standard error about the Regression line 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 . prediction are likely to be

How to use :- The Regression Equation :
once you have the Regression equation, using it is a snap choose a value for the independant variable (x) . perform the computation and you have an estimated value (y) the dependant variable

In our example the independent variable is the student score on the aptitude test . the dependent variable is the students statistics grade . if a student made an 80 on the aptitute test . the estimated statistics grade (y) would be

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

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

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

Residuals :

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

Residual = observed value - predicted value

$e = y - \hat{y}$

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

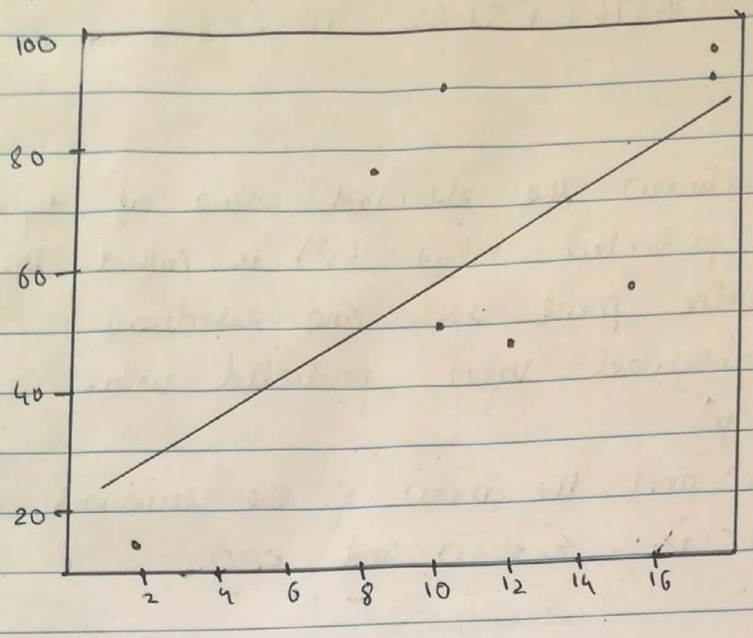
Algorithm

1) import the Required package
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 Regression object
6) Take input from user
7) Calculate the value of y
8) Draw scatter plot

Important function used for linear Regression :

1) coef : it is used to calculate ~~intercept~~ Slope in ML model
2) Intercept : it is used to calculate intercept in ML model
3) Fit : shows the relationship between two variable
4) Score : it display accuracy score of model

Scatter plot generated after implementation of code :-



Conclusion :

Thus we learn that how to find the trend of data using x as independent variable & y is dependent variable by linear regression