ASSIGNMENT NO. 2

AIM: Assignment on Linear Regression

PREREQUISITE: Python programming

THEORY:

Simple Linear Regression

When we have a single input attribute (x), and we want to use linear regression, this is called simple linear regression.

If we had multiple input attributes (e.g. x1, x2, x3, etc.) This would be called multiple linear regression. The procedure for linear regression is different and simpler than that for multiple linear regression, so it is a good place to start.

In this section we are going to create a simple linear regression model from our training data, then make predictions for our training data to get an idea of how well the model learned the relationship in the data.

With simple linear regression we want to model our data as follows:

$$y = B0 + B1 * x$$

This is a line where y is the output variable we want to predict, x is the input variable we know and B0 and B1 are coefficients that we need to estimate that move the line around.

Technically, B0 is called the intercept because it determines where the line intercepts the y-axis. In machine learning we can call this the bias, because it is added to offset all predictions that we make. The B1 term is called the slope because it defines the slope of the line or how x translates into a y value before we add our bias.

The goal is to find the best estimates for the coefficients to minimize the errors in predicting y from x.

Simple regression is great, because rather than having to search for values by trial and error or calculate them analytically using more advanced linear algebra, we can estimate them directly from our data.

We can start off by estimating the value for B1 as:

$$B1 = \frac{\sum (Xi - \bar{X}) * (Yi - \bar{Y})}{\sum (Xi - \bar{X})^2}$$

Where mean() is the average value for the variable in our dataset. The xi and yi refer to the fact that we need to repeat these calculations across all values in our dataset and i refers to the i'th value of x or y. We can calculate B0 using B1 and some statistics from our dataset, as follows:

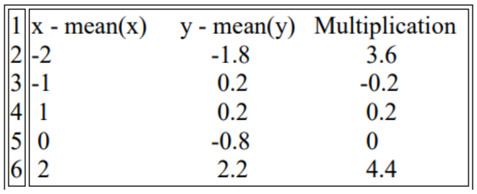
$$B0 = \bar{Y} - (B1 * \bar{X})$$

Not that bad right? We can calculate these right in our spreadsheet.

Estimating Slope (B1) Let's start with the top part of the equation, the numerator. First we need to calculate the mean value of x and y. The mean is calculated as: sum(x) / n Where n is the number of values (5 in this case). Let's calculate the mean value of our x and y variables:

$$\bar{X} = 3 \text{ , } \bar{Y} = 2.8$$

We now have the parts for calculating the numerator. All we need to do is multiple the error for each x with the error for each y and calculate the sum of these multiplications.



1	x - mean(x)	y - mean(y)	Multiplication
2	-2	-1.8	3.6
3	-1	0.2	-0.2
4	1	0.2	0.2
5	0	-0.8	0
6	2	2.2	4.4

Summing the final column we have calculated our numerator as 8.

Now we need to calculate the bottom part of the equation for calculating B1, or the denominator. This is calculated as the sum of the squared differences of each x value from the mean.

We have already calculated the difference of each x value from the mean, all we need to do is square each value and calculate the sum.

Calculating the sum of these squared values gives us up denominator of 10 Now we can calculate the value of our slope.

B1 = 8 / 10 so further B1 = 0.8

Estimating Intercept (B0) This is much easier as we already know the values of all of the terms involved. $B0 = \bar{Y} - (B1 * \bar{X})$

or

B0 = 2.8 – 0.8 * 3 , or further B0 = 0.4

Making Predictions

We now have the coefficients for our simple linear regression equation.

y = B0 + B1 * x

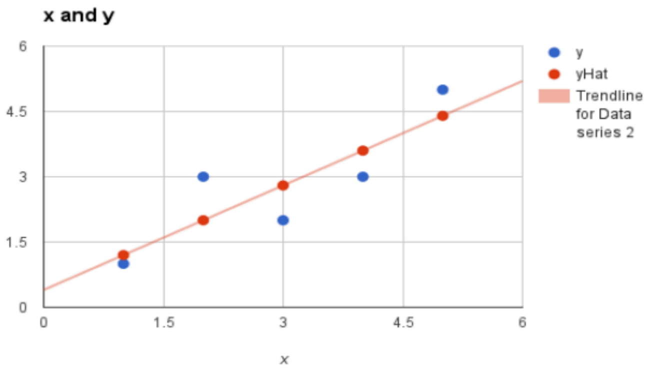
or

y = 0.4 + 0.8 * x

Let’s try out the model by making predictions for our training data.

1	x	y	predicted y
2	1	1	1.2
3	2	3	2
4	4	3	3.6
5	3	2	2.8
6	5	5	4.4

We can plot these predictions as a line with our data. This gives us a visual idea of how well the line models our data.



Simple Linear Regression Model

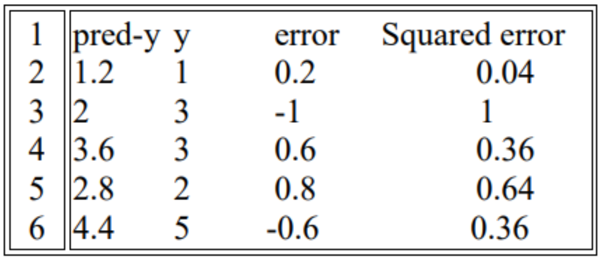
Estimating Error We can calculate a error for our predictions called the Root Mean Squared Error or RMSE.

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2.$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

Where sqrt() is the square root function, p is the predicted value and y is the actual value, i is the index for a specific instance, n is the number of predictions, because we must calculate the error across all predicted values.

First we must calculate the difference between each model prediction and the actual y values. We can easily calculate the square of each of these error values (error*error or error^2).



1	pred-y	y	error	Squared error
2	1.2	1	0.2	0.04
3	2	3	-1	1
4	3.6	3	0.6	0.36
5	2.8	2	0.8	0.64
6	4.4	5	-0.6	0.36

The sum of these errors is 2.4 units, dividing by n and taking the square root gives us: RMSE = 0.692 Or, each prediction is on average wrong by about 0.692 units.

REFERENCES:

1. Mitchell M., T., Machine Learning, McGraw Hill (1997) 1st Edition.

2. Alpaydin E., Introduction to Machine Learning, MIT Press (2014) 3rd Edition.

CONCLUSION:

Linear regression is a fundamental technique in statistical modeling and machine learning, used to predict a continuous outcome based on one or more input variables. In simple linear regression, the relationship between the input variable (x) and the output variable (y) is modeled as a straight line, characterized by two coefficients: the slope (B1) and the intercept (B0). By estimating these coefficients, we can make predictions and evaluate the model's accuracy using metrics like Root Mean Squared Error (RMSE). Linear regression provides an intuitive and powerful approach to understand and predict the relationship between variables, making it a key tool in data analysis and machine learning.