Regression Report

By: Kirtan

Introduction:

Regression is a statistical model to estimate the relation between dependent and one or more independent variables. In Regression we try to find free parameters by training on input parameters and output parameters.

Eq: y = mx + c + e

In the regression we have an error function and we try to minimize that error function

Error = (y - mx - c) ** 2

Some general type of regression:

1.) Linear Regression
2.) Multiple Linear Regression
3.) Polynomial Regression

1.) Linear Regression:

In this type of regression we have only one dimension of input and our function is a simple linear equation.

General Equation : Y = mX + c + E

(Where Y is output, m and c are free parameters, X is input and E is error)

2.) Multiple Linear Regression:

In this type of Regression we have multiple input dimensions and our function is a simple linear equation.

General Equation: H = { x ↦ <w,x> + b , w ∈ R**d , b ∈ R }

(w and b are free parameters, x is a input parameter)

3.) Polynomial Regression:

In this type of Regression we have one more than one input dimension and the function is polynomial of degree n (n >= 1).

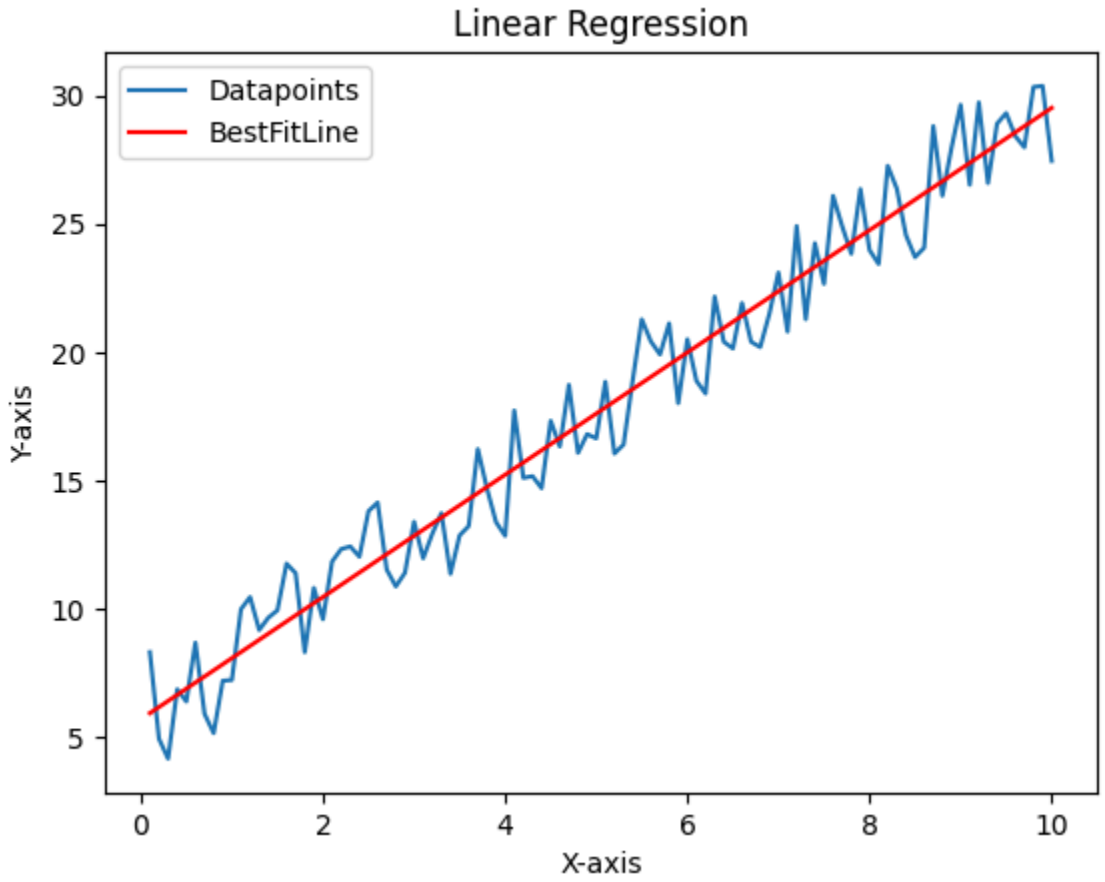
General Equation:

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

(where a0, a1,... are free parameters and x is input parameter)

Analysis of Regression:
@DataSet-1 is a Linear function.

In Machine Learning we always talk about a function , not about the function. So when I tried to fit a linear function in the dataset It satisfied the value and it gave me 0.96 R Square value. From this we can infer that it is a one of a function which satisfies this dataset.

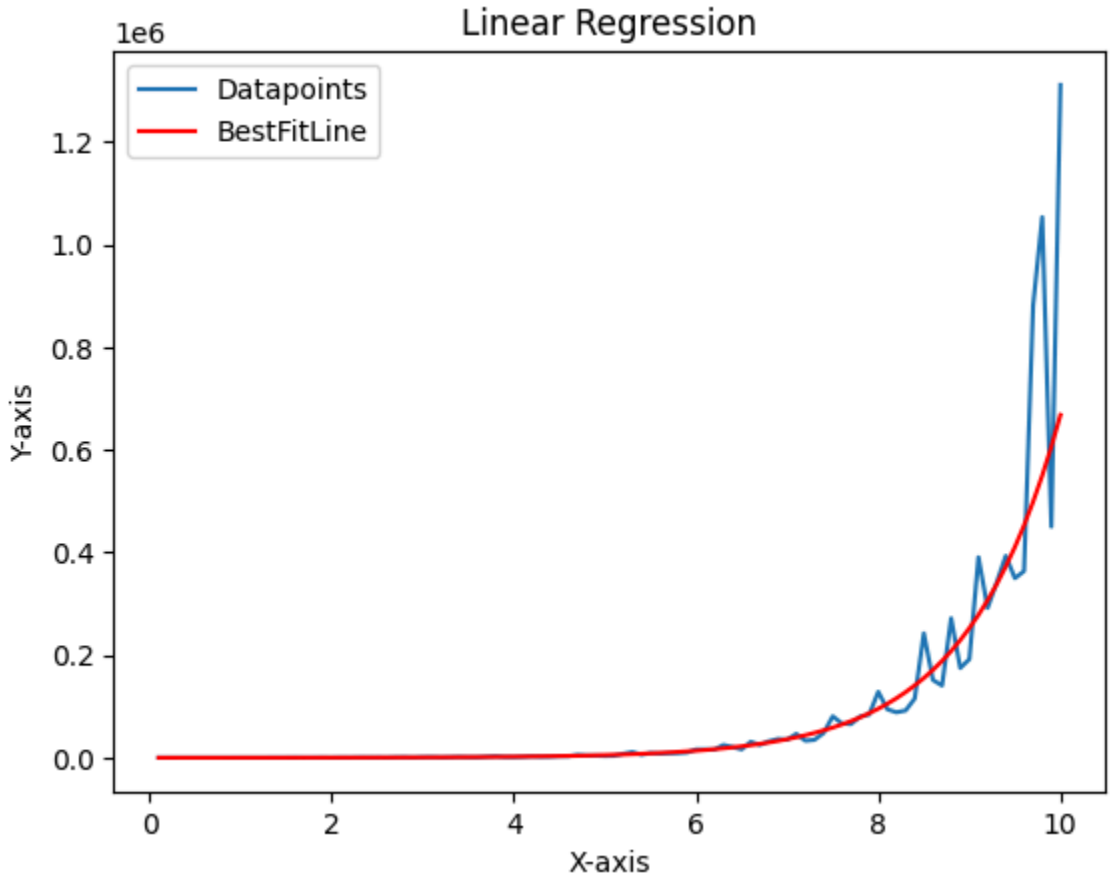


The value of W for this dataset is: [[5.68078713], [2.38406007]].
The value came out from the standard library:

Code	Mean Square Error	Root Mean Square Error	Absolute Mean Error	R Square
My Code	2.0785254017773265	1.4417091945941547	1.2805559784291146	0.9579571905586358
Scikit-Learn	2.0785254017773274	1.4417091945941551	1.2805559784291147	0.9579571905586357

@DataSet-2 is an exponential function.

In this dataset I applied linear regression but it gave me an R square value as 0.34 that was not good enough to say that this is a linear model. But, when I applied x square still i got 0.31 R square value then I tried some polynomial transformation till x to the power 8 then I got 0.88 R square value. From this I got that this can be exponential because it is a high power model. So I tried the power x model which gave me 0.91 R square value. But the prediction was far enough from the original value. Reason is that the e^x is going to infinity when x is around 100 hence, i took a log(y) then i applied the function and I got 0.99 R square value. Which is good enough to say that this is an exponential function.

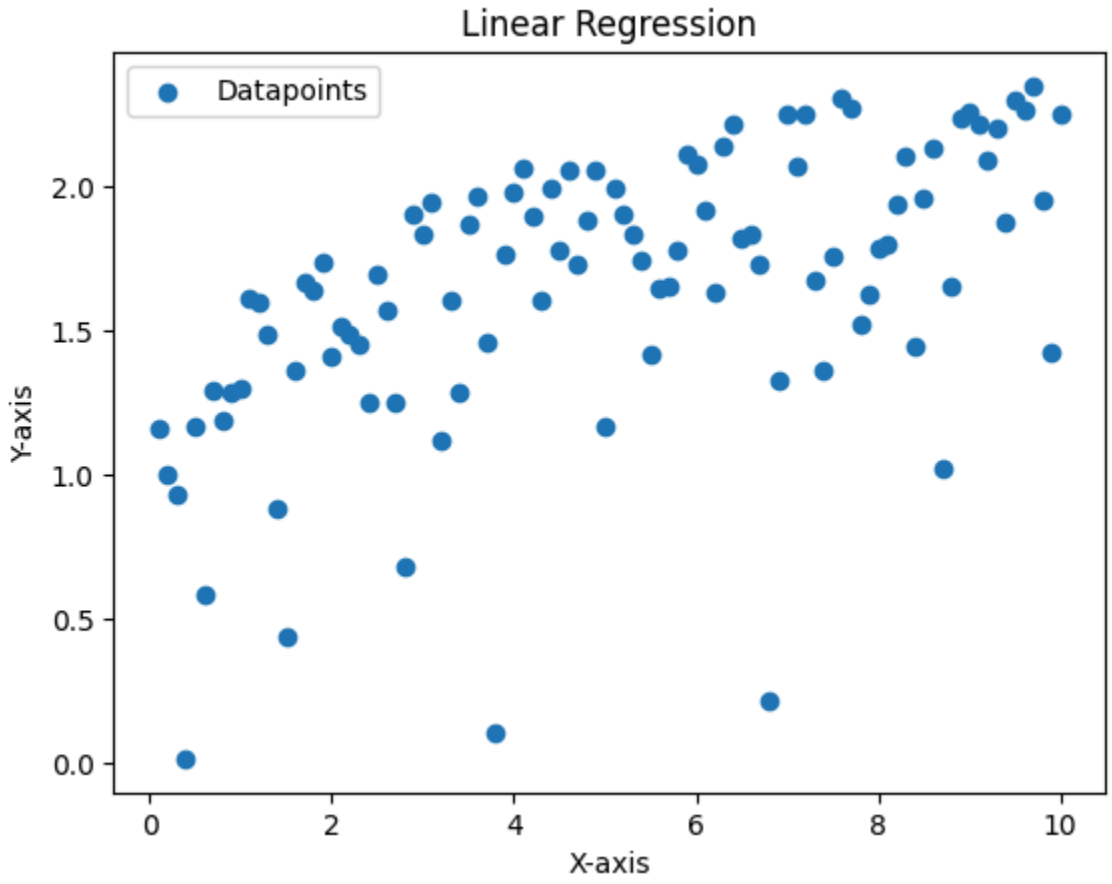


The value of W for this dataset is: [[3.68212267], [0.97299745]].
The value came out from the standard library:

Code	Mean Square Error	Root Mean Square Error	Absolute Mean Error	R Square
My Code	0.07643342704351962	0.2764659600086774	0.23498835289025688	0.9904038522690993
Scikit-Learn	0.07643342704351971	0.27646596000867757	0.23498835289025732	0.9904038522690994

@DataSet-3 [Regression is not applicable].

In this dataset, Regression can not be applied because this dataset contains too many outliers. In regression we square mean error loss function. When we have many outliers it increases the mean error which is prone to the wrong prediction and we are unable to fit the data in a function. When we plot this data set then we realize that the comment is true because this dataset contains too many outliers. I have tried 5 function transformations but it was not fitting to any one but still we can't say anything because there may be any one function which fits this data. But that data will lead to the wrong prediction for the testing data. By the help of above points we can say that Regression can not be fitted into this dataset.



In this we do optimisation by taking the function of degree equal to length of the dataset but it will not be generalized. So we prefer not to use regression for this dataset.

@DataSet-4 [Multiple Linear Regression is applicable in this dataset].

In this dataset I tried for Multiple Linear regression and I got the 0.98 R square value that is good enough to comment on a function and the function is Multiple linear equation.

Model	Mean Square Error	Root Mean Square Error	Absolute Mean Error	R Square
My Code	34.620480829243554	5.883917133104744	5.155505630377769	0.9841749058943147
Scikit-Learn	34.220480829243560	5.732546985620144	5.155505630378744 5	0.9841749058943147

By comparing with the standard library we can confirm that my values are correct.