Concrete Definition of Linear Regression

By Sunil Ghimire - “**Be Unique, Be Identifiable, Be You**”

Regression is the method which measures the average relationship between two or more continuous variables in term of the response variable and feature variables. In other words, regression analysis is to know the nature of the relationship between two or more variables to use for predicting the most likely value of dependent variables for a given value of independent variables. Linear regression is a mostly used regression algorithm.

For a more concrete understanding, let’s say there is a high correlation between day temperature and sales of tea and coffee. Then the salesman might wish to know the temperature for the next day to decide for the stock of tea and coffee. This can be done with the help of regression.

The variable, whose value is estimated, predicted, or influenced is called a dependent variable. And the variable which is used for prediction or is known is called an independent variable. It is also called explanatory, regressor, or predictor variable.

LINEAR REGRESSION

Linear Regression is a supervised method that tries to find a relation between a continuous set of variables from any given dataset. So, the problem statement that the algorithm tries to solve linearly is to best fit a line/plane/hyperplane (as the dimension goes on increasing) for any given set of data.

This algorithm use statistics on the training data to find the best fit linear or straight-line relationship between the input variables (X) and output variable (y). The simple equation of the Linear Regression model can be written as:

```
Y=mX+c; Here m and c are calculated on the training
```

In the above equation, m is the scale factor or coefficient, c being the bias coefficient, Y is the dependent variable and X is the independent variable. Once the coefficient m and c are known, this equation can be used to predict the output value Y when input X is provided.

Mathematically, coefficients m and c can be calculated as:

```
m = sum((X(i) - mean(X)) * (Y(i) - mean(Y))) / sum( (X(i) - mean(X))^2 )
```

```
c = mean(Y) - m * mean(X)
```

As you can see, the red point is very near the regression line; its error of prediction is small. By contrast, the yellow point is much higher than the regression line and therefore its error of prediction is large. The best-fitting line is the line that minimizes the sum of the squared errors of prediction.

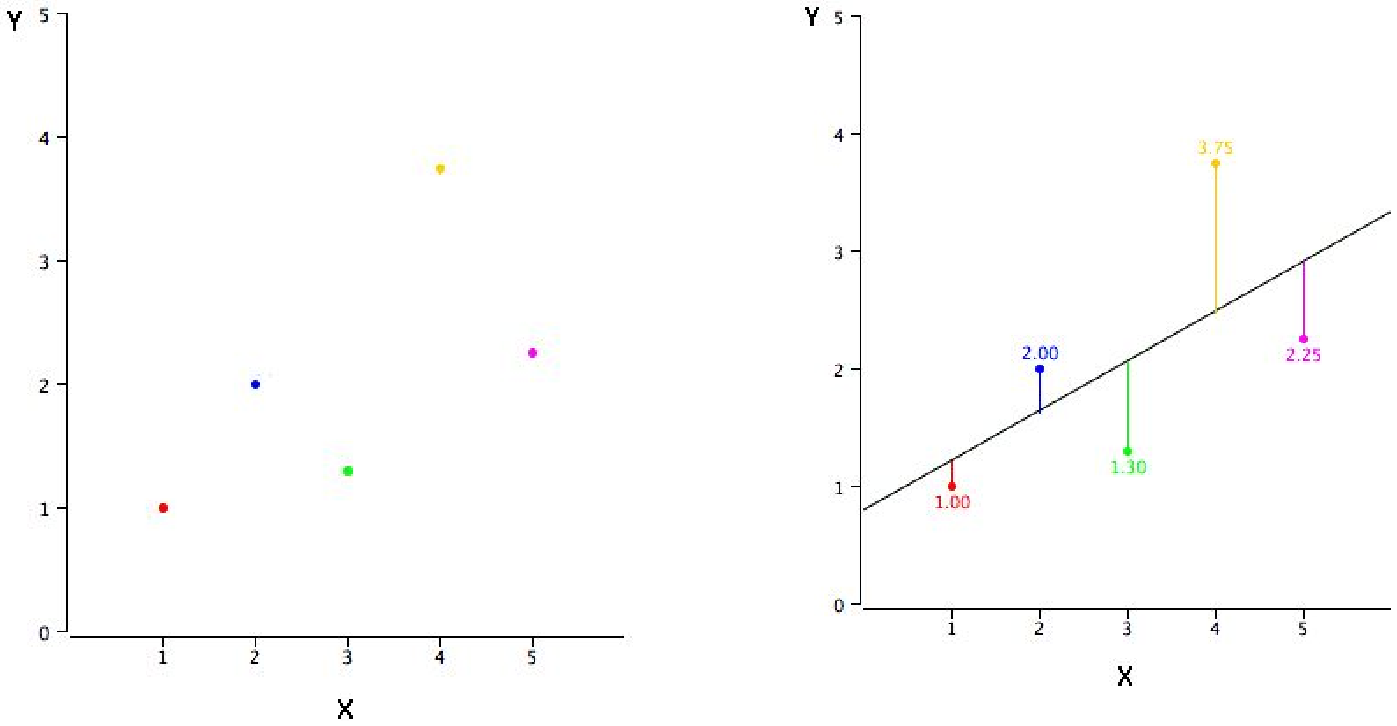

Figure 02: Scatter plot of example data.

In the above figure, the black line consists of the predictions, the points are the actual data, and the vertical lines between the points and the black line represent errors of prediction.

CONCLUSION

We need to able to measure how good our model is (accuracy). There are many methods to achieve this but we would implement Root mean squared error and coefficient of Determination (R^2 Score).

a. Try Model with Different error metric for Linear Regression like Mean Absolute Error, Root mean squared error
b. Try algorithm with large data set, imbalanced & balanced dataset so that you can have all flavors of Regression.

Note: **For the scratch implementation of linear regression**, feel free to connect me any time and wherever you like without any hesitation.