**Lab 2**

**To implement linear regression**

**Linear regression** is a basic predictive analytics technique that uses historical data to predict an output variable. It is popular for predictive modelling because it is easily understood. Linear regression models have many real-world applications in an array of industries such as economics (e.g. predicting growth), business (e.g. predicting product sales, employee performance), social science (e.g. predicting political leanings from gender or race), healthcare (e.g. predicting blood pressure levels from weight, disease onset from biological factors), and more.

The basic idea is that if we can fit a linear regression model to observed data, we can then use the model to predict any future values. There are two kinds of variables in a linear regression model:

* The **input** or **predictor variable** is the variable(s) that help predict the value of the output variable. It is commonly referred to as ***X***.
* The **output variable** is the variable that we want to predict. It is commonly referred to as ***Y***.

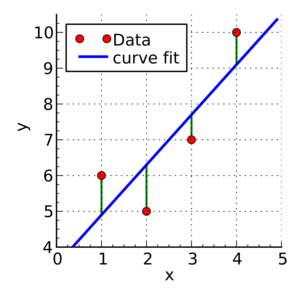
To estimate *Y* using linear regression, we assume the equation:

***Yₑ = α + β X***

where *Y*ₑ is the estimated or predicted value of *Y* based on our linear equation.

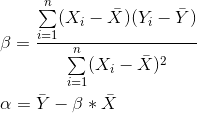
Our goal is to find statistically significant values of the **parameters *α*** and ***β*** that minimise the difference between *Y* and *Y*ₑ. If we are able to determine the optimum values of these two parameters, then we will have the **line of best fit** that we can use to predict the values of *Y*, given the value of *X*. How do we estimate *α* and *β*? We can use a method called ordinary least squares.

**Ordinary Least Squares**



Green lines show the difference between actual values Y and estimate values *Y*ₑ

The objective of the least squares method is to find values of *α* and *β* that minimise the sum of the squared difference between *Y* and *Y*ₑ. We will not go through the derivation here, but using calculus we can show that the values of the unknown parameters are as follows:



where *X̄* is the mean of *X* values and *Ȳ* is the mean of *Y* values.

If you are familiar with statistics, you may recognise *β* as simply   
*Cov(X, Y) / Var(X).*

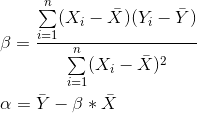
**Steps:**

1. Generate data with nonzero mean and standard deviation (X). Also, create random data by multiplying a constant and adding residual which is an actual data (Y).
2. Calculate the mean of X and Y

Mean\_X = X1+X2+…X\_N/ N

Mean\_Y = Y1+Y2+…Y\_N/ N

1. Calculate α and β using



1. Compute the predicted output y

***Yₑ = α + β X***

1. Plot regression against actual data