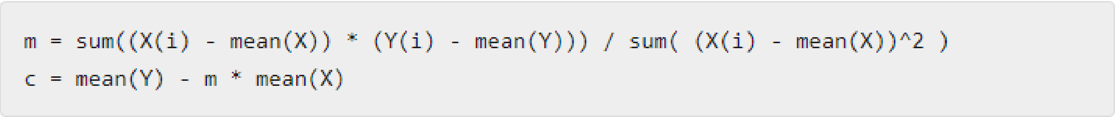
Linear Regression

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.

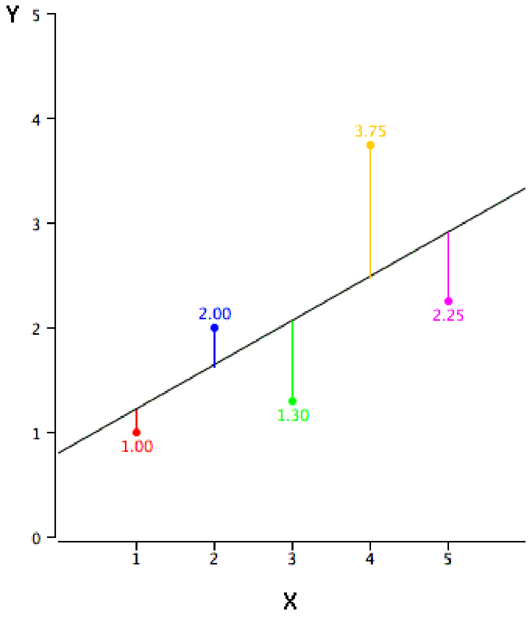
This algorithm use statistics on the training data to find the best fit linear or straight-line relationship between the input variables (X) and output variable (y). Simple equation of Linear Regression model can be written as:

Y=mX+c; Here m and c are calculated on 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.



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



Link: https://onlinestatbook.com/2/regression/intro.html

Linear regression from scratch:

Linear regression assumes a linear or straight-line relationship between the input variables (X) and the single output variable (y).

More specifically, that output (y) can be calculated from a linear combination of the input variables (X). When there is a single input variable, the method is referred to as a simple linear regression.

In simple linear regression we can use statistics on the training data to estimate the coefficients required by the model to make predictions on new data.

The line for a simple linear regression model can be written as**: y = b0 + b1*x + e**

b0 and b1 are known as the regression beta coefficients or parameters:

b0 is the intercept of the regression line; that is the predicted value when x = 0.
b1 is the slope of the regression line.
e is the error term (also known as the residual errors), the part of y that can be explained by the regression model

Note: Why do we use the above formula:
The mathematical **formula** of the linear regression can be written as y = **b0** + **b1***x + e, where: **b0** and **b1** are known as the regression beta coefficients or parameters: **b0** is the intercept of the regression line; that is the predicted value when x = 0 . **b1** is the slope of the regression line.

Once the coefficients are known, we can use this equation to estimate output values for y given new input examples of x.

It requires that you calculate statistical properties from the data such as mean, variance and covariance.

All the algebra has been taken care of and we are left with some arithmetic to implement to estimate the simple linear regression coefficients.

Briefly, we can estimate the coefficients as follows:
B1 = sum((x(i) - mean(x)) * (y(i) - mean(y))) / sum((x(i) - mean(x))^2)
B0 = mean(y) - B1 * mean(x)
where the i refers to the value of the ith value of the input x or output y