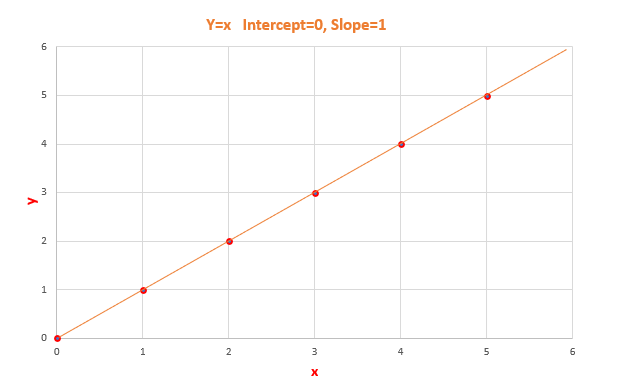
**Line of Best Fit in Linear Regression**

# Equation of Straight Line

# Y=mx+c

m→slope  
c →intercept



y=x [Slope=1, Intercept=0]

# Model Coefficient

Slope **m** and Intercept **c**are model coefficient/model parameters/regression coefficients.

## Slope →m

Slope basically says how steep the line is. The slope is calculated by a change in y divided by a change in x



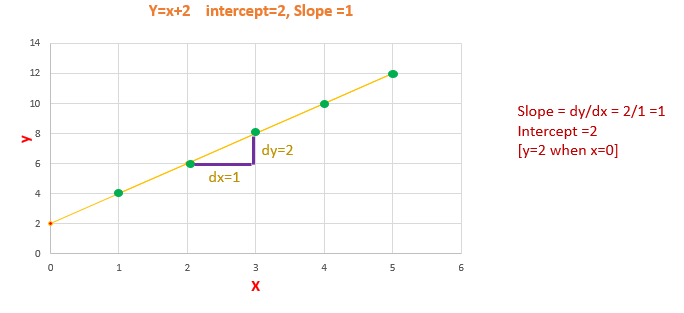
The slope will be negative if one increases and the other one decreases.  
The slope will be positive if x increases and y increases.

*The value of slope will range from -∞ to + ∞.*

[Since we didn’t normalize the value, the slope will depend on units. So, it can take any value from -∞ to + ∞]

## Intercept → c

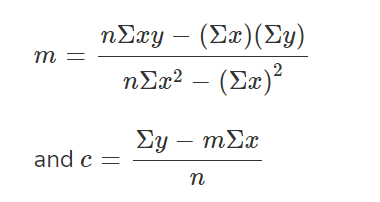
The value of y when x is 0.  
When the straight line passes through the origin intercept is 0.



Calculating Slope and Intercept

The slope will remain constant for a line. We can calculate the slope by taking any two points in the straight line, by using the formula dy/dx.

The fit for given data is of the type **y=mx+c,** m being slope of regression line and c being intercept of regression line.



# Line of Best Fit

The Linear Regression model have to find the line of best fit.

We know the equation of a line is y=mx+c. There are infinite m and c possibilities, which one to chose?

Out of all possible lines, how to find the best fit line?

The line of best fit is calculated by using the cost function — Least Sum of Squares of Errors.

The line of best fit will have the least sum of squares error.

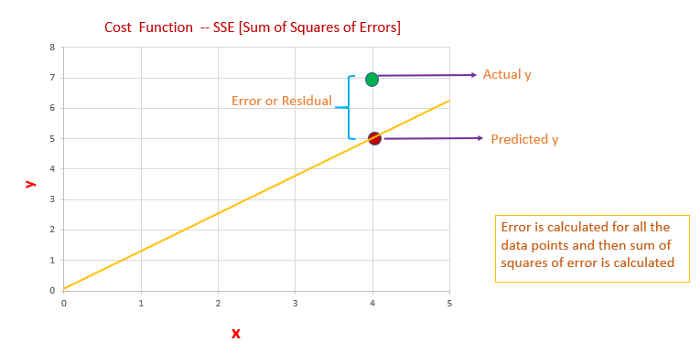
# Cost Function

The least Sum of Squares of Errors is used as the cost function for Linear Regression.

For all possible lines, calculate the sum of squares of errors. The line which has the least sum of squares of errors is the best fit line.

## Error/Residuals

Error is the difference between the actual value of y and the predicted value of y.



Error or Residual

1. We have to calculate error/residual for all data points
2. square the error/residuals.
3. Then we have to calculate the sum of squares of all the errors.
4. Out of all possible lines, the line which has the least sum of squares of errors is the line of best fit.

## The reason behind squaring the error/residuals

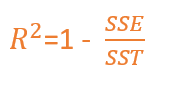
1. If we are not squaring the error, the negative and positive signs will cancel. We will end up with error=0
2. So we are interested only in the magnitude of the error. How much the actual value deviates from the predicted value.
3. So, why we didn’t consider the absolute value of error. Our motive is to find the least error. If the errors are squared, it will be easy to differentiate between the errors comparing to taking the absolute value of the error.
4. Easier to differentiate the errors, it will be easier to identify the least sum of squares of error.

Out of all possible lines, the linear regression model comes up with the best fit line with the least sum of squares of error. Slope and Intercept of the best fit line are the model coefficient.

Now we have to measure how good is our best fit line?

# Coefficient of Determination R² → R-squared

R-squared is one of the measures of goodness of the model. (best-fit line)



SSE →Sum of squares of Errors  
SST →Sum of Squares Total

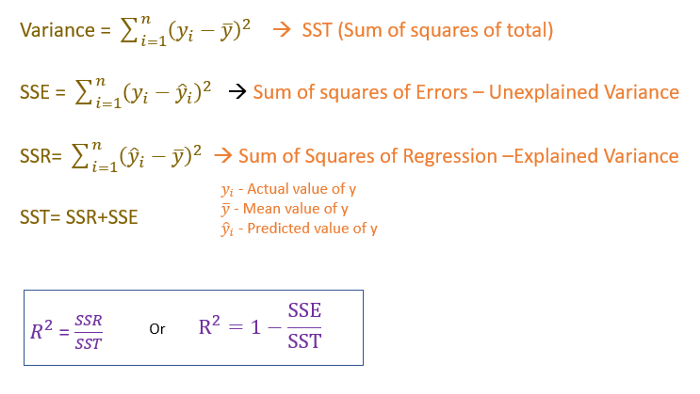
**What is the Total Error?**

Before building a linear regression model, we can say that the expected value of y is the mean/average value of y. The difference between the mean of y and the actual value of y is the **Total Error.  
Total Error is the Total variance.**Total Variance is the amount of variance present in the data.

After building a linear regression model, our model predicts the y value. The difference between the mean of y and the predicted y value is the Regression Error.  
Regression Error is the **Explained Variance**. Explained Variance means the amount of variance captured by the model.

Residual/Error is the difference between the actual y value and the predicted y value.  
Residual/Error is the **Unexplained Variance**.

Total Error = Residual Error + Regression Error



Coefficient of determination or R-squared measures how much variance in y is explained by the model.

The R-squared value ranges between 0 and 1

0 → being a bad model and 1 being good.

