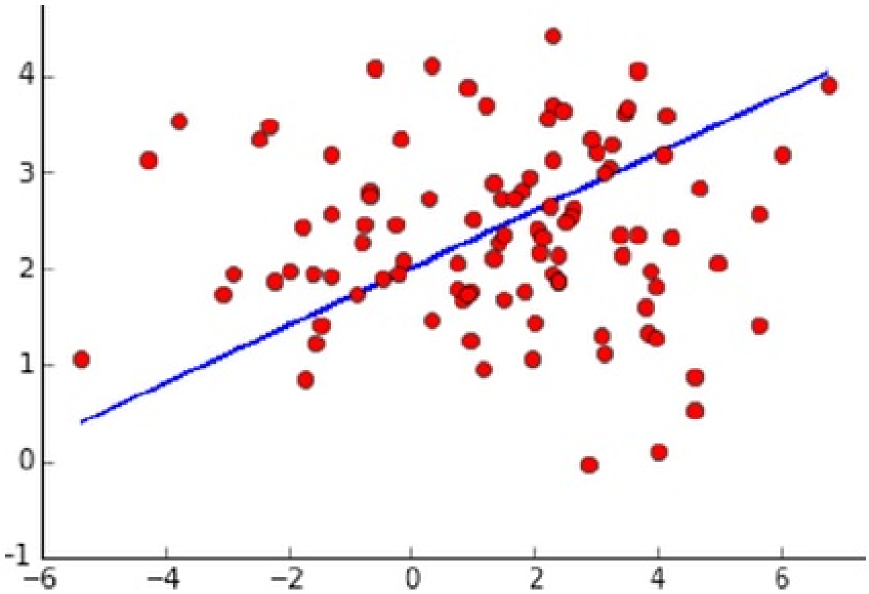
Understanding the concept of simple linear regression

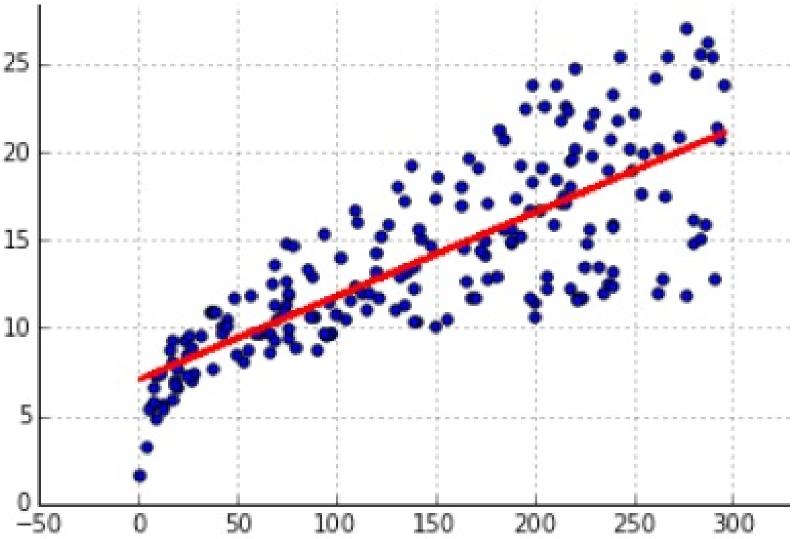
towardsdatascience.com/understanding-the-concept-of-simple-linear-regression-a572087c253

In simple words linear regression is predicting the value of a variable Y(dependent variable) based on some variable X(independent variable) provided there is a linear relationship between X and Y.

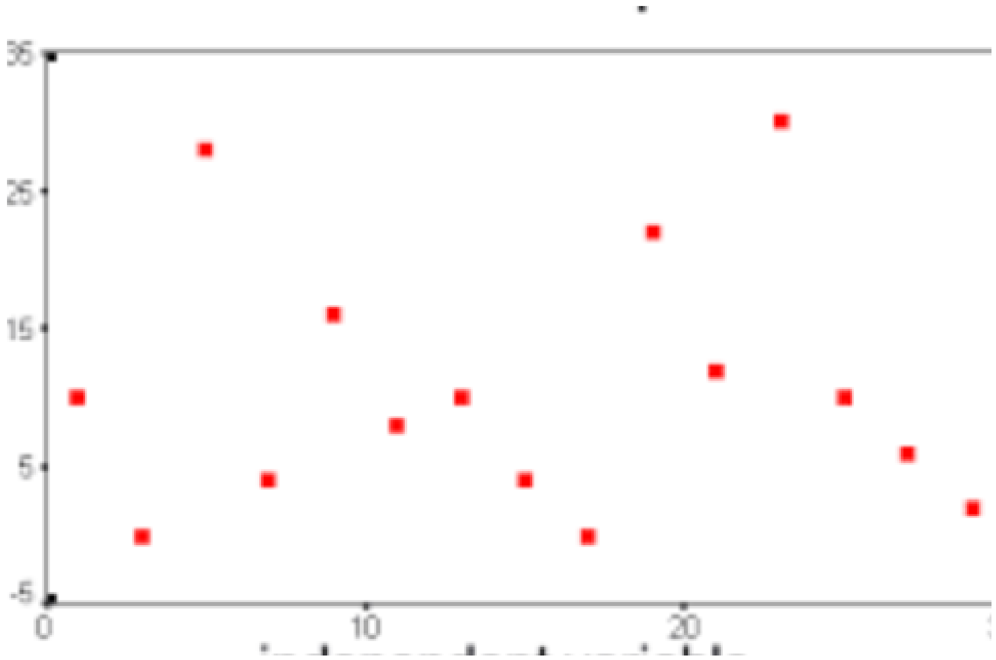
This linear relationship between the 2 variables can be represented by a straight line (called***regression line***).



Now to determine if there is a linear relationship between 2 variables we can simply plot a scatter plot of variable Y with variable X .If the plotted points are randomly scattered that it can be inferred that the variables are not related.

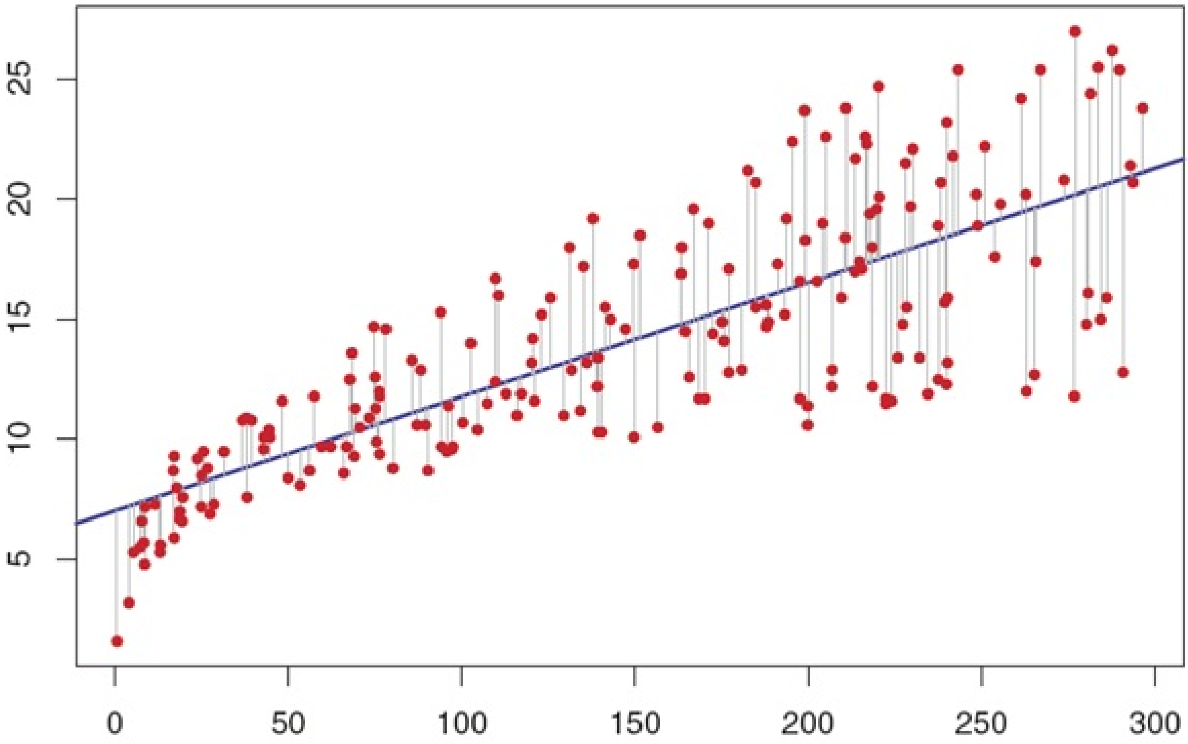


There is a linear relationship between the variables.

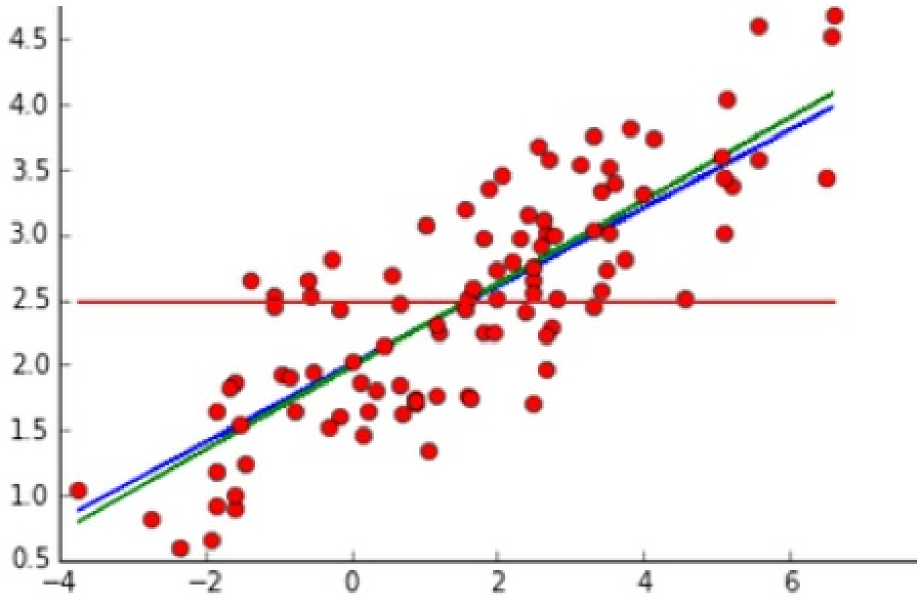


There is no linear relationship between the variables.

When regression line is drawn some points will lie on the regression line other points will lie in the close vicinity of it. This is because our regression line is a **probabilistic model** and our prediction is approximate. So there will be some errors/deviations from actual/observed value of variable Y.



But when the linear relationship exist between X and Y we can plot more than one line through these points. Now how do we know which one is the best fit?



To help us choose the best line we use the concept of “least squares”.

Least Squares

Y=b0 + b1X+e

This the mathematical representation for the regression line where

Y-Dependant variable.

X-Independent variable.

b0 –intercept of the regression line.

b1-slope of the regression line.

e- error/deviation from actual/observed value of variable Y.

Suppose we fit n points of the form (x1,y1) ,(x2,y2)…..(xn,yn)to the above regression line then

Where *ei* is the difference between *i*th observed response value and the *i*th response value that is predicted by our regression line.

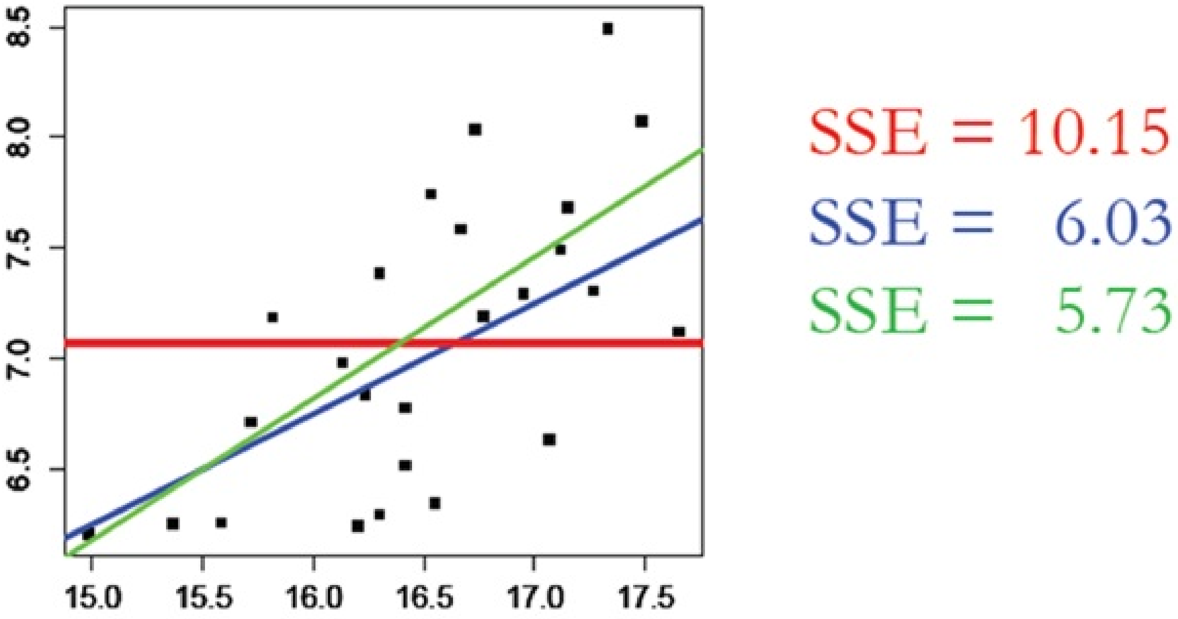
$$e_i = (Y_i - b_0 - b_1 X_i\ )$$

Our aim here is to minimize this error so that we can get the best possible regression line.

Now this error *ei* can be positive or negative but we are only interested in the magnitude of the error and not in its sign. Hence we square the errors and minimize the *sum of squared errors(SSE).*

$$SSE = \sum_{i=1}^{n} e_i^2$$

$$= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$



(In the above graph the green line is the best fit.)

How do we minimize the *sum of squared errors(SSE)?*

Remember that b1 and b0 are still unknown to us.

In the least square approach we minimize *sum of squared errors(SSE)* by choosing the value of b1 and b0 to be (not diving into math of it)

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x}.$$