Linear Regression

We have a function (f(x)) for which given (x) predict (y). That is f(x) = y. And for regression, y is continuous.

Simplest type of function used in Linear Regression is Linear Function. That is:

$$f(x) = mx + b$$

And:

$$mx + b = y$$

$Note: mx + b$

$is\ a\ linear\ function\ which\ represents\ the\ equation\ of\ a\ straight\ line.\ But\ f(x)\ can\ comprise\ of\ any\ function.$

There are two types of Regression:

1. **Simple Regression.**
2. **Multiple Regression.**

1. **<u>Simple Regression:</u>**

Where 'x' comprises of a single feature.

2. **<u>Multiple Regression</u>**

Where 'x' comprises of multiple feature.

1. **Straight Line**

The geometrical representation of an equation of the form:

$$Ax + By + C = 0$$

Where A, B, C are constants. When B is not zero, i.e. the equation contains:

When B is not zero , i.e. the equation contains a term in y, the equation of the straight line can be solved for y ,

$$y = \left(-\frac{A}{B}\right)x + \left(-\frac{C}{B}\right)$$

which is the form of:

$$y = a + bx$$

The above form can also be written as:

$$y = mx + c$$

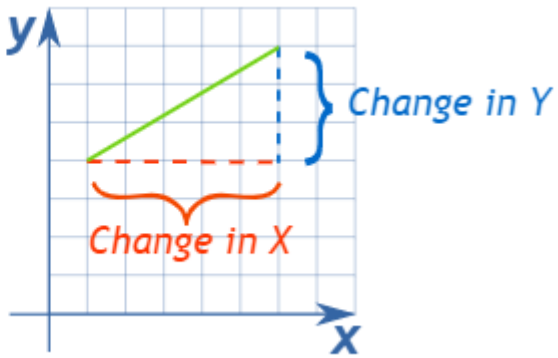
2. **Slope of Straight Line**

In the y = a + bx , the coefficient of 'x' on the right.

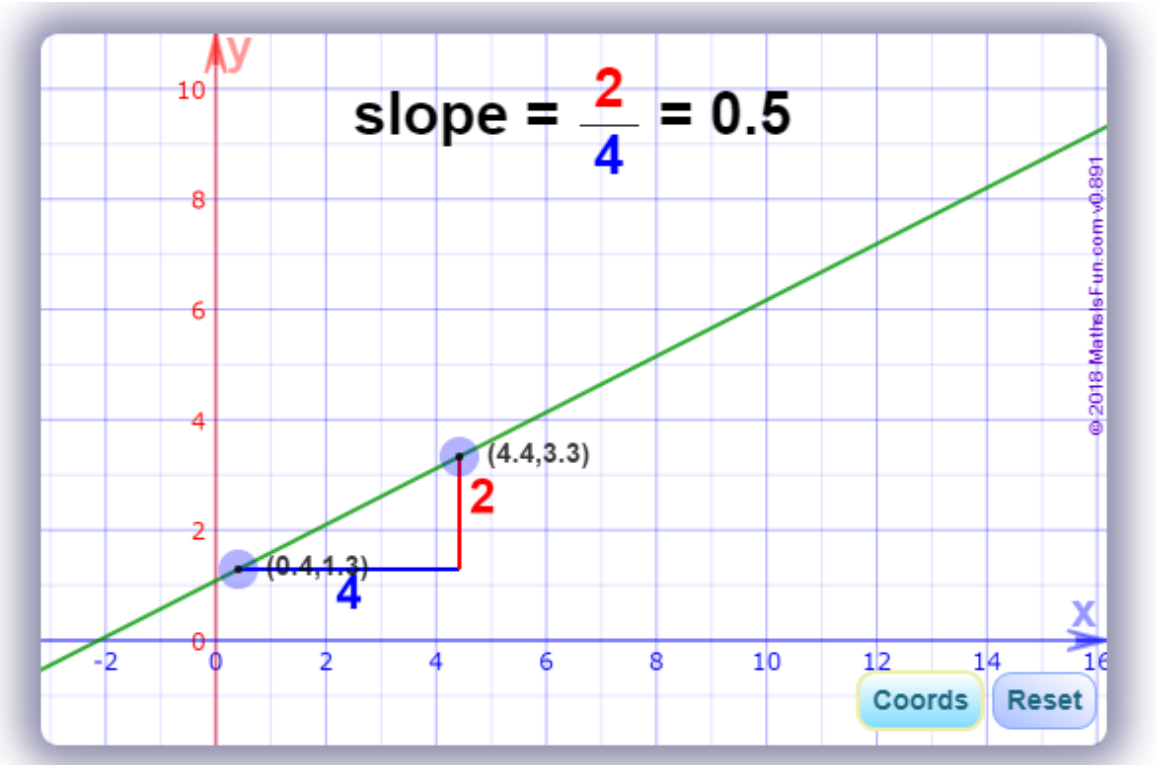
$$b = '\ slope\ of\ the\ straight\ line'$$

The slope may be positive, negative or zero.

$$\text{Slope} = \frac{\text{Change in Y}}{\text{Change in X}}$$



a) Slope is positive:

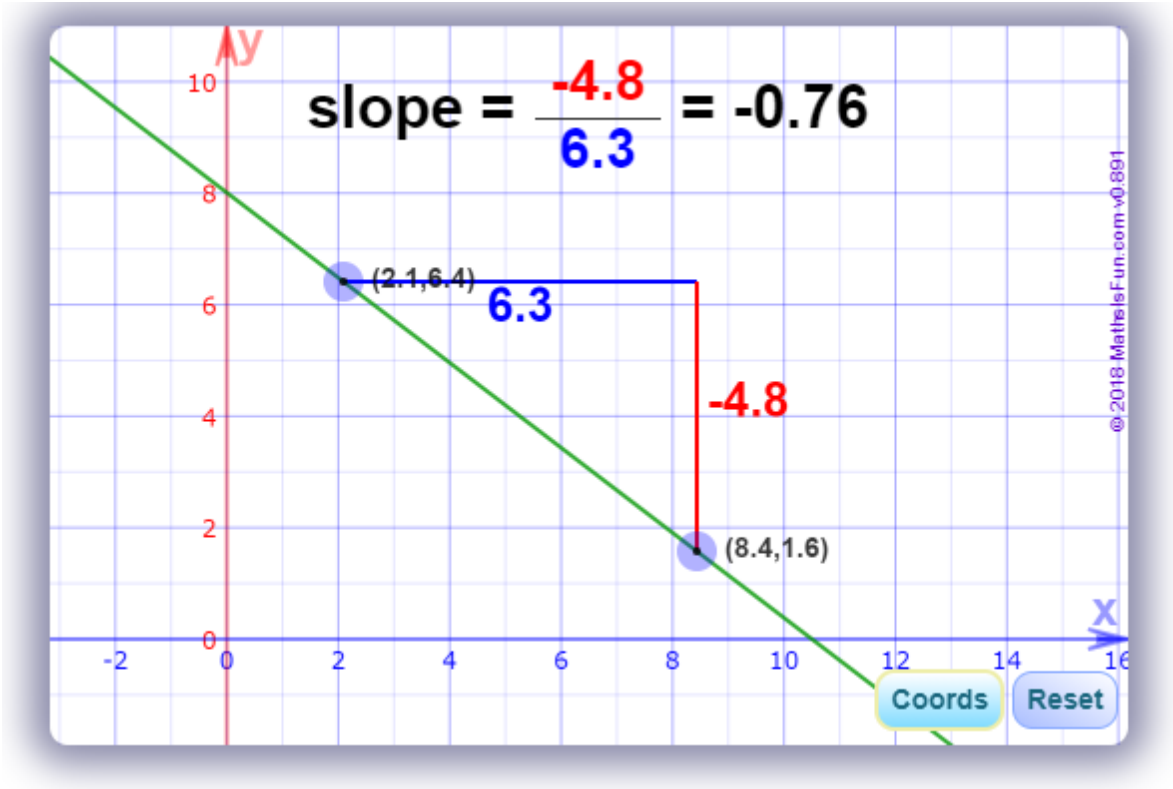


$$slope = \frac{Change\ in\ Y}{Change\ in\ X}$$

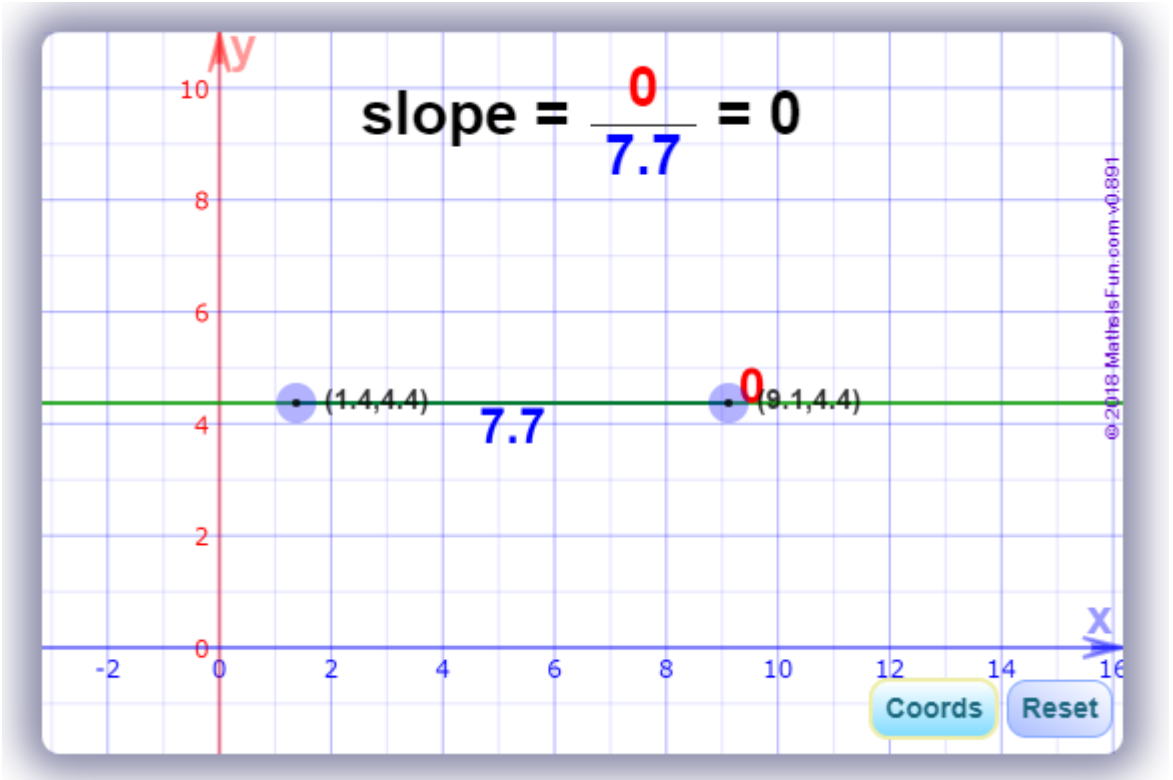
$$= \frac{3.3 - 1.3}{4.4 - 0.4}$$

$$= \frac{2}{4} = 0.5$$

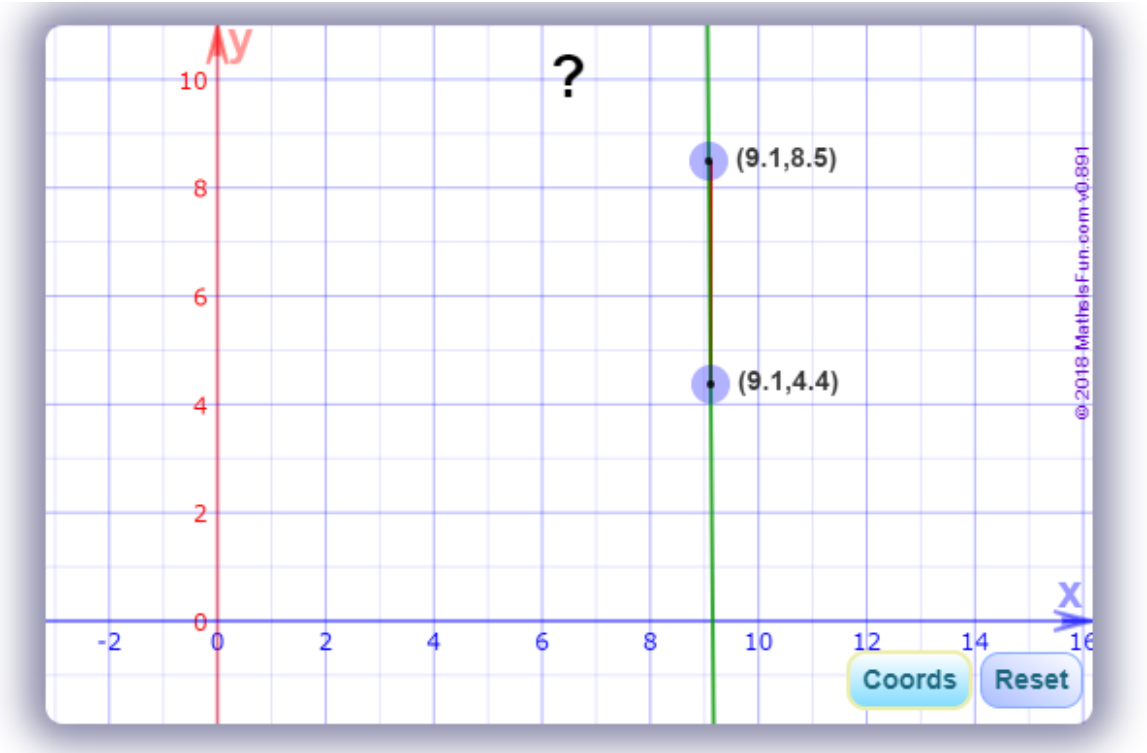
b) Slope is negative:



c) Slope is zero



d) Slope is Infinity:



Let there be an equation of a straight line:

$$5x + 2y - 6 = 0$$

$$=> 5x + 2y = 6$$

$$=> 2y = 6 - 5x$$

$$=> y = 3 - \frac{5x}{2}$$

We know the equation:

$$y = mx + c$$

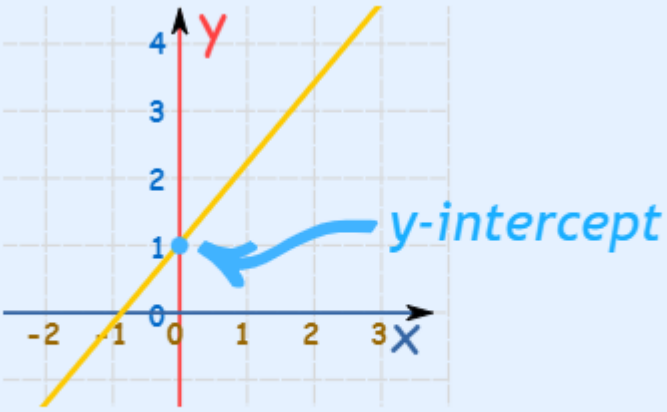
$$therefore\ , y = -\frac{5x}{2} + 3$$

$$therefore, mx = -\frac{5x}{2}\ and\ c = 3$$

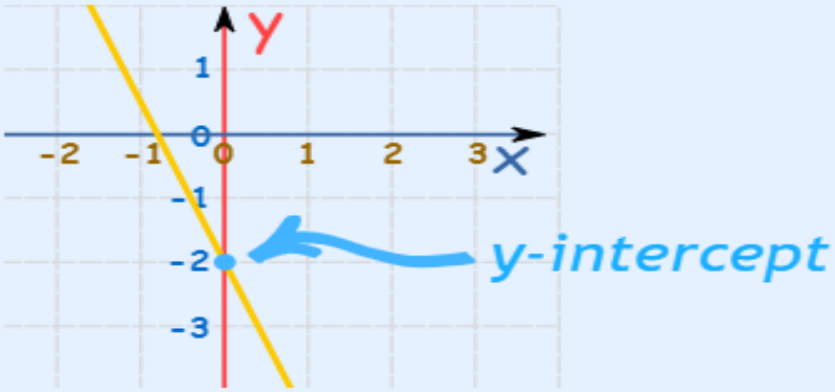
$$Here, c\ is\ called\ the\ y - intercept.$$

$$y - intercept\ is\ the\ point\ where\ the\ straight\ line$$

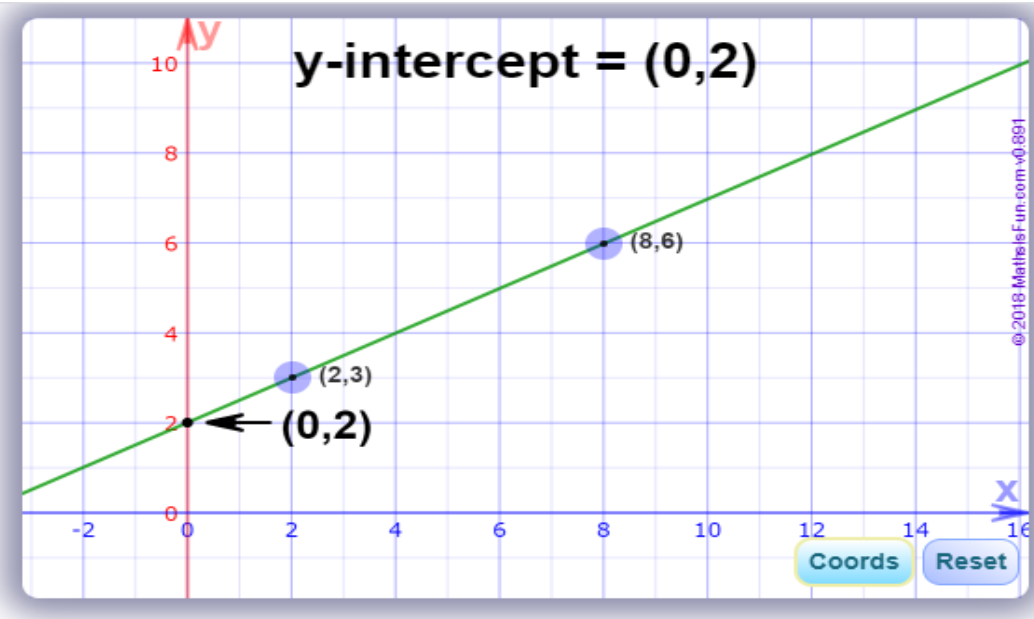
$$touches(intercept)\ the\ y\ axis\ ,where\ x = 0.$$



In the above diagram the line crosses the y axis at $\mathbf{y} = \mathbf{1}$



Here the line crosses the y axis at $\mathbf{y} = -\mathbf{2}$



Some useful information about straight lines:

a) If a straight line $y = mx + c$ *or* $y = a + bx$ passes through the point (x_0, y_0) *the coordinate must satisfy the equation i.e.* $y_0 = a + bx_0$ *or* $y_0 = mx_0 + c$

b) If a straight line passes through the points (x_1, y_1) *and* (x_2, y_2), *its slope is* $b = \frac{y_1 - y_2}{x_1 - x_2}$. If a straight line passes through the point (x_0, y_0) *and has slope* b, Its equation is:

$$y - y_0 = b(x - x_0)$$

Parabola

Parabola is the geometrical representation of an equation of the form:

$$y = a + bx + cx^2$$

Where a, b, c are constants (the term in x^2 *must be* present, i.e. $c \neq 0$). The parabola is a special type of curve.

Following equations represent parabolas:

1. $y = 3 - 2x + 7x^2$
2. $y = 6 - 4x^2$
3. $y = 0.5x^2$
4. $y = a + bx + cx(x - 1)$
   $=> a + bx + cx^2 - cx$
   $=> a + bx - cx + cx^2$
   $=> a + (b - c)x + cx^2$

Method of Least Square

Method of Least Squares is a device for finding the equation of a specified type of curve, which best fits, a given set of observations.

The method depends upon the *Principle of Least Squares*, which suggests that for the "best-fitting" curve, the sum of the squares of differences between the observed and the corresponding estimated values should be the minimum possible.

Suppose we are given 'n' pairs of observation $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$**, and it is required to fit a straight line to these data. The general equation of a straight line** $y = a + bx$ **is taken, where** $a \; and \; b \; are \; constants$.

Any values for $'a' \; and \; 'b'$ **would give a straight line, and once these values are obtained, an estimate of** $'y'$**can be had by substituting the values of** $'x'$.

That is to say, the estimated values of $'y'$ when $x = x_1, x_2, \ldots\ldots, x_n$ would be :

$a + bx_1, a + bx_2, \ldots\ldots, a + bx_n$ respectively.

In order that the equation $y = a + bx$ gives a good representation of the relationship between $x \; and \; y$, it is desirable that the estimated values,

$a + bx_1, a + bx_2, \ldots\ldots, a + bx_n$

are on whole, close enough to the corresponding observed values : $y_1, y_2, \dots, y_n$.

For the best fitting straight line ,therefore, our problem is only to choose such values of $'a' and\ 'b'$ for the equation : $y = a + bx$ which will provide estimates of 'y' as close as possible to the observed values . This can be done in different ways.

However according to the 'Principle of Least Square', the "best-fitting" equation is interpreted as that which minimises the sum of square of difference/ sum of square of errors (SSD/SSE):

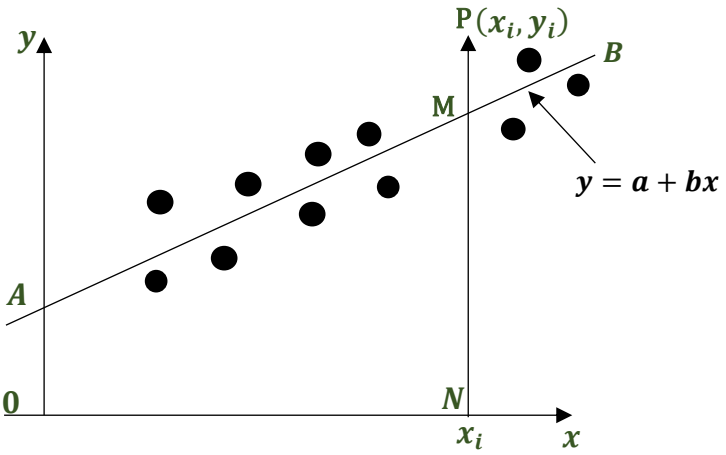
$$\sum_{i=1}^{n} (y_i - a - bx_i)^2$$

i.e.

$$(y_1 - a - bx_1)^2 + (y_2 - a - bx_2)^2 + \cdots + (y_n - a - bx_n)^2$$

Principle of Least Squares in Straight Line

x -(1)	$observed\ y$ -(2)	$Estimated\ y = a + bx$ -(3)	Difference (2)-(3)	(Difference)^2
x_1	y_1	$a + bx_1$	$y_1 - a - bx_1$	$(y_1 - a - bx_1)^2$
x_2	y_2	$a + bx_2$	$y_2 - a - bx_2$	$(y_2 - a - bx_2)^2$
....	....	....	....	....
x_n	y_n	$a + bx_n$	$y_n - a - bx_n$	$(y_n - a - bx_n)^2$
$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} a + bx_i$	$\sum_{i=1}^{n} y_i - a - bx_i$	$\sum_{i=1}^{n} (y_i - a - bx_i)^2$



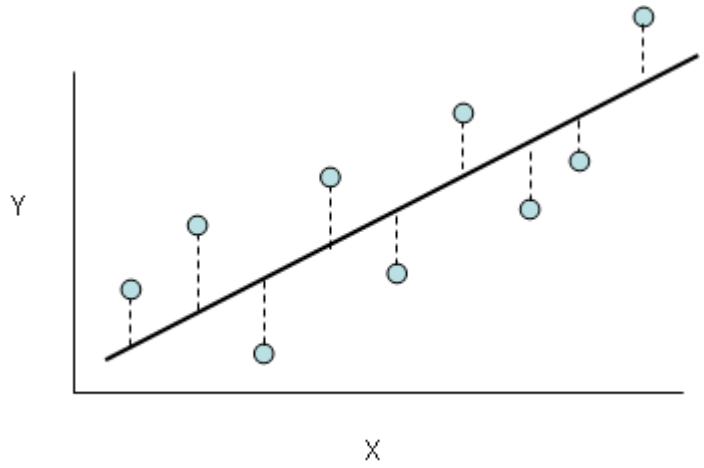
(*Geometrical Interpretation*)

Fig: *Method of Least Squares*

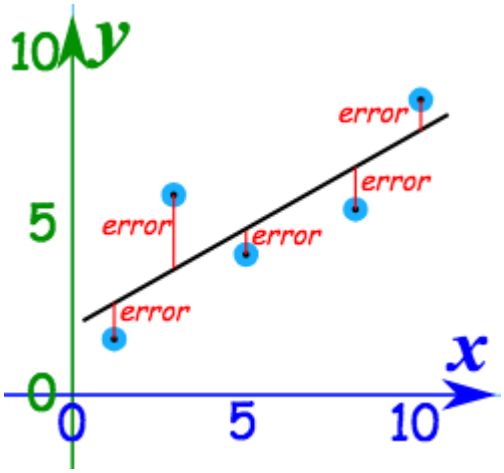
In the geometric sense, the problem of finding the best fitting straight line as follows:

If the pairs of observations: $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$

are plotted as points on a graph paper and all the possible straight lines are drawn on it, the straight line will be considered the "best fitting" for which the "*sum of the squares of vertical distances*" PM between the plotted points P and the line AB is the least.



Hence dotted points are predicted y and the straight line is the actual y, Therefore the distance between actual y and predicted y is the difference / error .



Such error or difference is known as Sum of Squares of Difference (SSD)/Sum of Squares of Error (SSE) i.e.

$$\sum_{i=1}^{n}(y_i - a - bx_i)^2$$

Obtained from Least Square Method.

Now we will do partial derivative of the above obtained difference:

$$SSE = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$

Partial derivative of 'a' :

$$\frac{\partial SSE}{\partial a} = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$

$$=> \frac{\partial}{\partial a}\left(\sum_{i=1}^{n}(y_i - a - bx_i)^2\right)$$

By applying power rule$\left(\frac{\partial}{\partial x}(x^n) = nx^{n-1} \right)$:

$$=> \sum_{i=1}^{n} \left\{ 2(y_i - a - bx_i)^{2-1} \times \frac{\partial}{\partial a}(y_i - a - bx_i) \right\}$$

We know while doing partial derivation except 'a' everything will be considered as constants:

$$=> \sum_{i=1}^{n} \{2(y_i - a - bx_i) \times (0 - 1 - 0)\}$$

$$=> \sum_{i=1}^{n} \{2(y_i - a - bx_i) \times (-1)\}$$

$$=> 2\sum_{i=1}^{n} (y_i - a - bx_i) \times (-1) \cdots eqn(i)$$

Partial derivative of 'b':

$$\frac{\partial SSE}{\partial b} = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

$$=> \frac{\partial}{\partial b} \left(\sum_{i=1}^{n} (y_i - a - bx_i)^2 \right)$$

By applying power rule$\left(\frac{\partial}{\partial x}(x^n) = nx^{n-1} \right)$:

$$=> \sum_{i=1}^{n} \left\{ 2(y_i - a - bx_i)^{2-1} \times \frac{\partial}{\partial b}(y_i - a - bx_i) \right\}$$

We know while doing partial derivation except 'b' everything will be considered as constants:

$$=> \sum_{i=1}^{n} \{2(y_i - a - bx_i) \times \left(0 - 0 - (b\frac{\partial}{\partial b}(x_i) + (x_i)\frac{\partial}{\partial b}(b)\right)\}$$

$$=> \sum_{i=1}^{n} \{2\left(y_i - a - bx_i\right) \times (0 - 0 - (0 + (x_i) \times 1)\}$$

$$=> \sum_{i=1}^{n} \left\{2\left(y_i - a - bx_i\right) \times (-x_i)\right\}$$

$$=> 2\sum_{i=1}^{n} (y_i - a - bx_i) \times (-x_i) \cdots eqn(ii)$$

Now, $eqn(i)$ and $eqn(ii)$ is a linear equation i.e. equals to zero('0')

$$2\sum_{i=1}^{n} (y_i - a - bx_i) \times (-1) = 0 \ldots . eqn(iii)$$

$$2\sum_{i=1}^{n} \left(y_i - a - bx_i\right) \times (-x_i) = 0 \cdots eqn(iv)$$

eqn(iv) can be written as:

$$\mathbf{2}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\left(\boldsymbol{y_i}-\boldsymbol{a}-\boldsymbol{bx_i}\right)\times\left((\mathbf{-1})\times(\boldsymbol{x_i})\right)=\mathbf{0}$$

Now we will make all the constants i.e. -1 and 2 to right side and divide with zero in eqn(iii) and eqn(iv):

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_i}-\boldsymbol{a}-\boldsymbol{bx_i})=\mathbf{0}\ldots.\boldsymbol{eqn}(\boldsymbol{v})$$

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_i}-\boldsymbol{a}-\boldsymbol{bx_i})\times(\boldsymbol{x_i})=\mathbf{0}\cdots\boldsymbol{eqn}(\boldsymbol{vi})$$

Now we multiply summation(Σ) in eqn(v) and x_i and summation(Σ):

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_i}-\boldsymbol{a}-\boldsymbol{bx_i})=\mathbf{0}\ldots.\boldsymbol{eqn}(\boldsymbol{v})$$

$$=>\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{y_i}-\boldsymbol{a}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\mathbf{1}_{\boldsymbol{i}}-\boldsymbol{b}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{x_i}=\mathbf{0}$$

We know,

$$\sum_{i=1}^{n} 1_i = 1_1 + 1_2 + \cdots + 1_n = n$$

Therefore,

$$=> \sum_{i=1}^{n} y_i - an - b\sum_{i=1}^{n} x_i = 0$$

$$=> \sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i \ldots . eqn(vii)$$

Similarly eqn(vi):

$$\sum_{i=1}^{n} (y_i - a - bx_i) \times (x_i) = 0 \cdots eqn(vi)$$

$$=> \sum_{i=1}^{n} (y_i x_i - ax_i - bx_i^2) = 0$$

$$=> \sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0$$

$$=> \sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \ldots eqn(viii)$$

Now eqn(vii) and eqn(viii)

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i \ldots . eqn(vii)$$

$$\sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \ldots eqn(viii)$$

is considered the best-fit equations where least squared values can be fitted with minimum no. of errors hence these two equations is known as best fit of a straight line.

Hence equations for best fit of a straight line are:

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i \ldots . eqn(i)$$

$$\sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \ldots eqn(ii)$$

Best fit on Curve/Curve fit though Method of Least Square:

We know equation of parabola:

$$y = a + bx + cx^2$$

Principle of Least Squares in Curve

x -(1)	$observed\ y$ -(2)	$Estimated$ $y = a + bx + cx^2$ -(3)	Difference (2)-(3)	(Difference)^2
x_1	y_1	$a + bx_1 + cx_1^2$	$y_1 - a - bx_1 - cx_1^2$	$(y_1 - a - bx_1 - cx_1^2)^2$
x_2	y_2	$a + bx_2 + cx_2^2$	$y_2 - a - bx_2 - cx_2^2$	$(y_2 - a - bx_2 - cx_2^2)^2$
....	....	....	....	....
x_n	y_n	$a + bx_n + cx_n^2$	$y_n - a - bx_n - cx_n^2$	$(y_n - a - bx_n - cx_n^2)^2$
$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} a + bx_i + cx_n^2$	$\sum_{i=1}^{n} y_i - a - bx_i - cx_n^2$	$\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2)^2$

Therefore the sum of square of difference/ sum of square of errors (SSD/SSE) obtained from Least Square Method:

$$\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2)^2$$

$$= (y_1 - a - bx_1 - cx_1^2)^2 + (y_2 - a - bx_2 - cx_2^2)^2 + \cdots + (y_n - a - bx_n - cx_n^2)^2$$

Now we will do partial derivative of the above-obtained difference:

$$SSE = \sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2)^2$$

1)

$$\frac{\partial SSE}{\partial a} = \sum_{i=1}^{n} \left(y_i - a - bx_i - cx_i^2\right)^2$$

$$\Rightarrow \frac{\partial SSE}{\partial a} = \frac{\partial}{\partial a} \sum_{i=1}^{n} \left\{\left(y_i - a - bx_i - cx_i^2\right)^2\right\}$$

By applying power rule$\left(\frac{\partial}{\partial x} (x^n) = nx^{n-1} \right)$:

$$\Rightarrow \sum_{i=1}^{n} \left\{2\left(y_i - a - bx_i - cx_i^2\right)^{2-1} \times \frac{\partial}{\partial a}\left(y_i - a - bx_i - cx_i^2\right)\right\}$$

$$\Rightarrow \sum_{i=1}^{n} \left\{2\left(y_i - a - bx_i - cx_i^2\right) \times (0 - 1 - 0 - 0)\right\}$$

$$\Rightarrow \sum_{i=1}^{n} \left\{2\left(y_i - a - bx_i - cx_i^2\right) \times (-1)\right\}$$

$$\Rightarrow 2\sum_{i=1}^{n} \left(y_i - a - bx_i - cx_i^2\right) \times (-1) \ldots . eqn(i)$$

2)

$$\frac{\partial SSE}{\partial b} = \sum_{i=1}^{n} \left(y_i - a - bx_i - cx_i^2\right)^2$$

$$=> \frac{\partial SSE}{\partial b} = \frac{\partial}{\partial b} \sum_{i=1}^{n} \left\{\left(y_i - a - bx_i - cx_i^2\right)^2\right\}$$

By applying power rule$\left(\frac{\partial}{\partial x}(x^n) = nx^{n-1} \right)$:

$$=> \sum_{i=1}^{n} \left\{2\left(y_i - a - bx_i - cx_i^2\right)^{2-1} \times \frac{\partial}{\partial b}\left(y_i - a - bx_i - cx_i^2\right)\right\}$$

$$=> \sum_{i=1}^{n} \left\{2\left(y_i - a - bx_i - cx_i^2\right) \times \left(0 - 0 - (b\frac{\partial}{\partial b}x_i + x_i\frac{\partial}{\partial b}b) - 0\right)\right\}$$

$$=> \sum_{i=1}^{n} \left\{2\left(y_i - a - bx_i - cx_i^2\right) \times (0 - 0 - (0 + x_i \times 1) - 0)\right\}$$

$$=> \sum_{i=1}^{n} \left\{2\left(y_i - a - bx_i - cx_i^2\right) \times (-x_i)\right\}$$

$$=> 2\sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2) \times (-x_i) \ldots\ldots eqn(ii)$$

iii)

$$\frac{\partial SSE}{\partial c} = \sum_{i=1}^{n} \left(y_i - a - bx_i - cx_i^2\right)^2$$

$$\Rightarrow \frac{\partial SSE}{\partial c} = \frac{\partial}{\partial c} \sum_{i=1}^{n} \left\{\left(y_i - a - bx_i - cx_i^2\right)^2\right\}$$

By applying power rule$\left(\frac{\partial}{\partial x}(x^n) = nx^{n-1} \right)$:

$$\Rightarrow \sum_{i=1}^{n} \left\{2\left(y_i - a - bx_i - cx_i^2\right) \frac{\partial}{\partial c}\left(y_i - a - bx_i - cx_i^2\right)\right\}$$

$$\Rightarrow \sum_{i=1}^{n} \left\{2\left(y_i - a - bx_i - cx_i^2\right) \times \left(0 - 0 - 0 - (c\frac{\partial}{\partial c}x_i^2 + x_i^2 \frac{\partial}{\partial c}c)\right)\right\}$$

$$\Rightarrow \sum_{i=1}^{n} \left\{2\left(y_i - a - bx_i - cx_i^2\right) \times \left(0 - 0 - 0 - (0 + x_i^2 \times 1)\right)\right\}$$

$$\Rightarrow \sum_{i=1}^{n} \left\{2\left(y_i - a - bx_i - cx_i^2\right) \times \left(-x_i^2\right)\right\}$$

$$\Rightarrow 2\sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2) \times \left(-x_i^2\right) \ldots . eqn(iii)$$

Now we get three equations:

$$2\sum_{i=1}^{n}\left(y_i - a - bx_i - cx_i^2\right) \times (-1) \ldots . eqn(i)$$

$$2\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2) \times (-x_i) \ldots . . eqn(ii)$$

$$2\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2) \times \left(-x_i^2\right) \ldots . eqn(iii)$$

Now , we putting eqns = 0 we get,

$$2\sum_{i=1}^{n}\left(y_i - a - bx_i - cx_i^2\right) \times (-1) = 0 \ldots . eqn(iv)$$

$$2\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2) \times (-x_i) = 0 \ldots . . eqn(v)$$

$$2\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2) \times \left(-x_i^2\right) = 0 \ldots . eqn(vi)$$

from eqn (iv) we get:

$$\mathbf{2}\sum_{i=1}^{n}\left(y_i - a - bx_i - cx_i^2\right) \times (-1) = 0 \ldots . eqn(iv)$$

$$=> \sum_{i=1}^{n}\left(y_i - a - bx_i - cx_i^2\right) = 0$$

$$=> \sum_{i=1}^{n}(y_i) = \sum_{i=1}^{n} a + bx_i + cx_i^2 \ldots . eqn(vii)$$

Similarly from eqn(v),

$$\mathbf{2}\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2) \times (-1) \times (x_i) = 0 \ldots \ldots eqn(v)$$

$$=> \sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2) \times (x_i) = 0$$

$$=> \sum_{i=1}^{n}(y_i x_i - ax_i - bx_i^2 - cx_i^3) = 0$$

$$=> \sum_{i=1}^{n}(y_i x_i) = \sum_{i=1}^{n} ax_i + bx_i^2 + cx_i^3 \ldots eqn(viii)$$

Similarly from eqn(vi),

$$2\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2) \times (-x_i^2) = 0 \ldots . eqn(vi)$$

$$=> \ 2\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2) \times (-1) \times (x_i^2) = 0$$

$$=> \ \sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2) \times (x_i^2) = 0$$

$$=> \ \sum_{i=1}^{n}(y_i x_i^2 - ax_i^2 - bx_i^3 - cx_i^4) = 0$$

$$=> \ \sum_{i=1}^{n}(y_i x_i^2) = \sum_{i=1}^{n} ax_i^2 + bx_i^3 + cx_i^4 \ldots eqn(ix)$$

Therefore we get three equations through which we get best fit in a curve:

$$\sum_{i=1}^{n}(y_i) = \sum_{i=1}^{n} a + bx_i + cx_i^2 \ldots . eqn(i)$$

$$\sum_{i=1}^{n}(y_i x_i) = \sum_{i=1}^{n} ax_i + bx_i^2 + cx_i^3 \ldots eqn(ii)$$

$$\sum_{i=1}^{n}(y_i x_i^2) = \sum_{i=1}^{n} ax_i^2 + bx_i^3 + cx_i^4 \ldots eqn(iii)$$

Through the above three equations we achieve best fit in a curve i.e. less error in experimental values very nearest to the curve also known as curve fitting or curve fit obtained from least square method.

Linear regression:

There are two types of Linear Regression:

1. Simple Linear Regression.
2. Multiple Linear Regression.

Simple Linear Regression:

Simple linear regression is used to estimate the relationship between two quantitative variables.

$$y = a + bx$$

Or,

$$y = mx + c$$

Or,

$$y = \beta_0 + \beta_1 x + \epsilon$$

$'\epsilon'$ Stands for 'epsilon' is the error while constructing a linear regression model, where:

β_0 *or* a *or* c *is the* $'y - intercept'$

β_1 *or* m *or* b *is the* $'slope'$

What we got in best fit while calculation for a straight line from partial derivation:

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i \ldots . eqn(i)$$

$$\sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \ldots eqn(ii)$$

can be written as :

$$\sum_{i=1}^{n} y_i = \beta_0 n + \beta_1\sum_{i=1}^{n} x_i \ldots . eqn(i)$$

$$\sum_{i=1}^{n} y_i x_i = \beta_0\sum_{i=1}^{n} x_i + \beta_1\sum_{i=1}^{n} x_i^2 \ldots eqn(ii)$$

Therefore, multiplying 'n' in equation (ii) and $\sum_{i=1}^{n} x_i$ in eqn(i) we get:

$$\sum_{i=1}^{n} x_i y_i = n \times \beta_0\sum_{i=1}^{n} x_i + \beta_1 (\sum_{i=1}^{n} x_i \times \sum_{i=n}^{n} x_i) .. eqn(iii)$$

$$n\sum_{i=1}^{n} x_i y_i = n \times \beta_0 \sum_{i=1}^{n} x_i + n \times \beta_1 \sum_{i=1}^{n} x_i^2 \ \ldots . eqn(iv)$$

Subtracting from eqn(iv) from eqn(iii) we get:

$$\sum_{i=1}^{n} x_i y_i = n \times \beta_0 \sum_{i=1}^{n} x_i + \beta_1 (\sum_{i=1}^{n} x_i \times \sum_{i=n}^{n} x_i) \ .. eqn(iii)$$

-

$$n\sum_{i=1}^{n} x_i y_i = n \times \beta_0 \sum_{i=1}^{n} x_i + n \times \beta_1 \sum_{i=1}^{n} x_i^2 \ \ldots . eqn(iv)$$

$$\sum_{i=1}^{n} x_i y_i - n\sum_{i=1}^{n} x_i y_i$$

$$= 0 + \beta_1 (\sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} x_i) - \left(n \times \beta_1 \sum_{i=1}^{n} x_i^2 \right)$$

$$\sum_{i=1}^{n} x_i y_i - n\sum_{i=1}^{n} x_i y_i$$

$$= \beta_1\left(\sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} x_i\right) - \left(n \times \beta_1 \sum_{i=1}^{n} x_i^2\right)$$

Interchanging sides we get:

$$=> \left(n \times \beta_1 \sum_{i=1}^{n} x_i^2\right) - \beta_1\left(\sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} x_i\right)$$

$$= n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i y_i$$

$$=> \beta_1\left(n\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} x_i\right)$$

$$= n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i y_i$$

$$=> \beta_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i y_i}{\left(n\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} x_i\right)}$$

Hence, we got slope as:

$$\beta_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i y_i}{\left(n\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} x_i\right)}$$

Similarly,

Again taking the best fit of a straight line:

$$\sum_{i=1}^{n} y_i = \beta_0 n + \beta_1 \sum_{i=1}^{n} x_i \ldots . eqn(i)$$

$$=> \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i = \beta_0 n$$

$$=> \beta_0 = \frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n}$$

$$We\ got\ y - intercept\ as: -$$

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n}$$

Hence for 2-d problem:

$$y = \beta_0 + \beta_1 x$$

We got two learn parameter $\beta_0\ and\ \beta_1$ where:

$$\beta_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i y_i}{(n \sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} x_i)}$$

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n}$$

How Simple Linear Regression works:

Suppose we have 'x' data and 'y' data:

x	1	2	3	4	5
y	3	4	2	4	5

we know :

$$y = \beta_0 + \beta_1 x$$

and

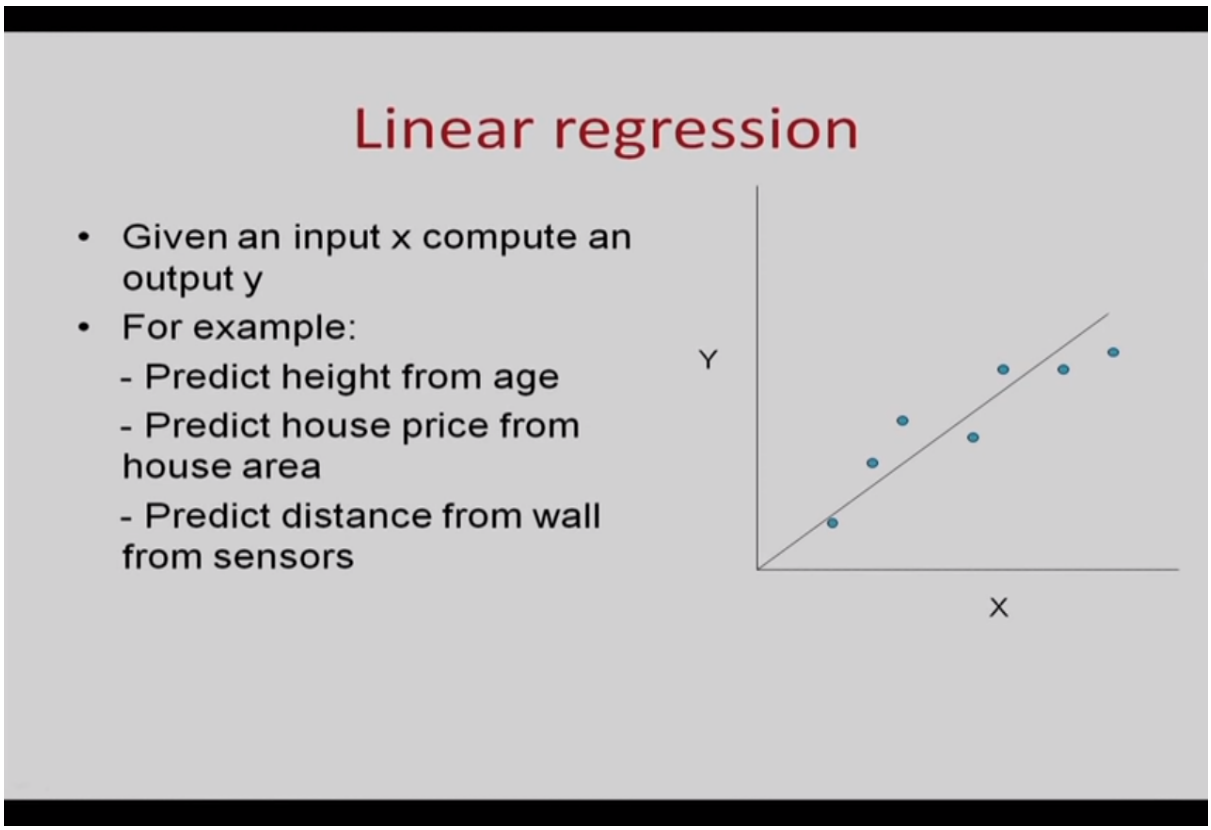
$$\beta_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i y_i}{(n\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} x_i)}$$

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n}$$

The linear regression function will take each value of 'x', calculate 'y', and plot them in the graph with a random error 'ε' making the equation:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

But we in practical sense , we do plotting in huge number of data size such as :



House Number	*Y*: Actual Selling Price	*X*: House Size (100s ft2)
1	89.5	20.0
2	79.9	14.8
3	83.1	20.5
4	56.9	12.5
5	66.6	18.0
6	82.5	14.3
7	126.3	27.5
8	79.3	16.5
9	119.9	24.3
10	87.6	20.2
11	112.6	22.0
12	120.8	.019
13	78.5	12.3
14	74.3	14.0
15	74.8	16.7
Averages	88.84	18.17

Sample 15 houses from the region.

Therefore 'Y' becomes:

$$Y = \beta_0 x^0 + \beta_1 x^1 + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon$$

When we have 'p' predictor variables.

The above equation is the equation of the population line i.e. the equation from which the examples are actually drawn.

Therefore expected value of 'y' given 'x' represented as:

$$(\mathrm{E}(Y|X)) =$$

$$\hat{\beta_0}x^0 + \hat{\beta_1}x^1 + \hat{\beta_2}x^2 + \cdots + \hat{\beta_p}x^p$$

^ is known as hat.

We need to find out $\hat{\beta_0}, \hat{\beta_1}, \hat{\beta_2}, .., \hat{\beta_n}$ so that sum of squared difference or sum of squared error is minimised through least squared line.

Now about error,

Error(ε)

We have to make assumption about the error:

$$d_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$$

$$d_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2$$

.

.

$$d_n = \beta_0 + \beta_1 x_n + \varepsilon_n$$

we assume that expected value of epsilon 'i':

$$E(\varepsilon_i) = 0$$

Residual error will have a Mean of Zero(Zero Conditional Mean),

And standard deviation of the error =

$$\sigma(\varepsilon_i) = \sigma(\varepsilon)$$

Errors $(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$are independent . And we can also assume that these errors are normally distributed.

This type of noise is known as Gaussian Noise or white noise.

We know sum of the squared errors:

$$\sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Also written as:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Or,

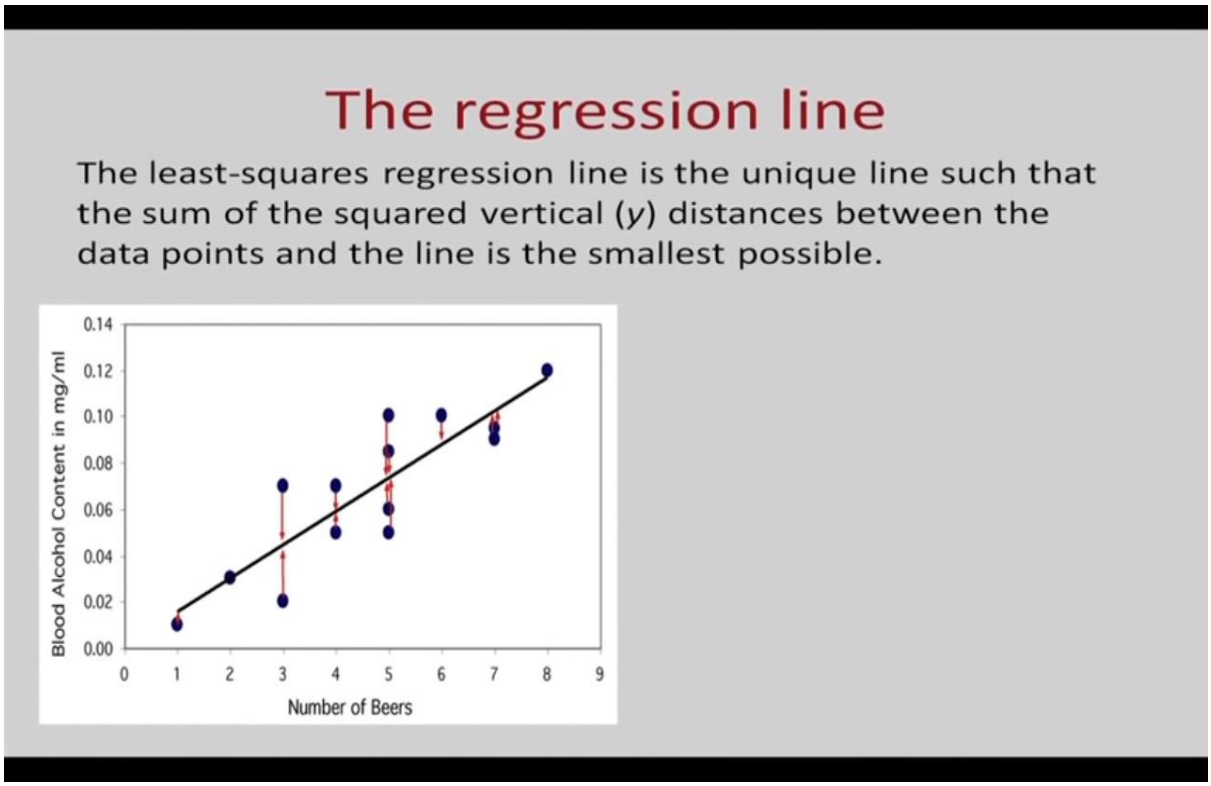
$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Sum of the squared errors for estimated value of 'y' on 'x':

$$\sum_{i=1}^{n}\left(y_i - (\hat{\beta_0} + \hat{\beta_1} x_i)\right)^2$$

Now , if $d_i = (x_i, y_i)$,the above equation can also be written as:

$$\sum_{d \epsilon D}\left(y_i - (\hat{\beta_0} + \hat{\beta_1} x_i)\right)^2$$



How do we "learn" parameters

- For the 2-*d* problem

$$Y = \beta_0 + \beta_1 X$$

- To find the values for the coefficients which minimize the objective function we take the partial derivates of the objective function (SSE) with respect to the coefficients. Set these to 0, and solve.

$$\beta_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} \qquad \beta_0 = \frac{\sum y - \beta_1 \sum x}{n}$$

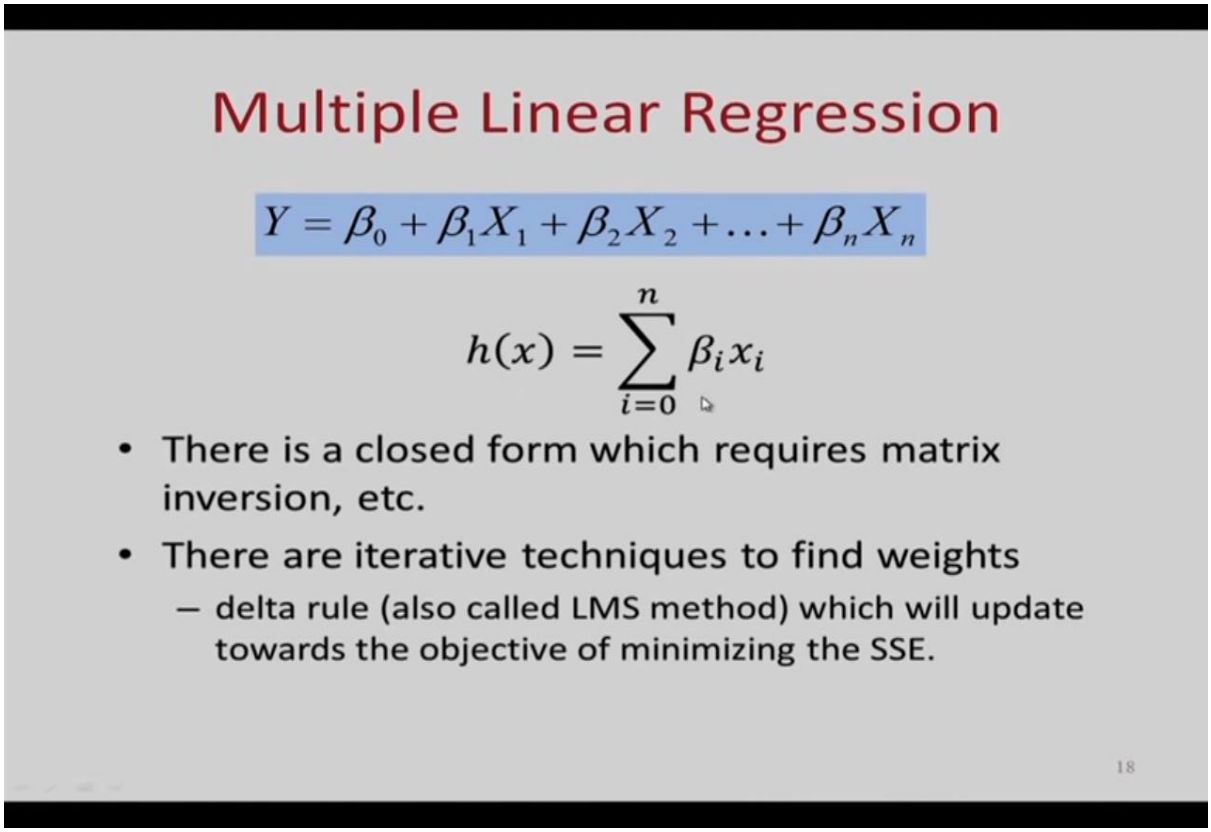
17

i.e.:

$$\boldsymbol{\beta_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i y_i}{(n\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} x_i)}}$$

$$\boldsymbol{\beta_0 = \frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n}}$$

Now coming to Multiple Linear Regression:



Alternative to matrix inversion etc. methods, the most famous rule called delta rule also called LMS (least minimum slope) method.

Linear Regression

$$h(x) = \sum_{i=0}^{n} \beta_i x_i$$

To learn the parameters θ (β_i) ?

- Make h(**x**) close to *y*, for the available *training examples*.
- Define a cost function $J(\theta)$

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h(x)^{(i)} - (y)^{(i)})^2$$

- Find θ that minimizes J(θ).

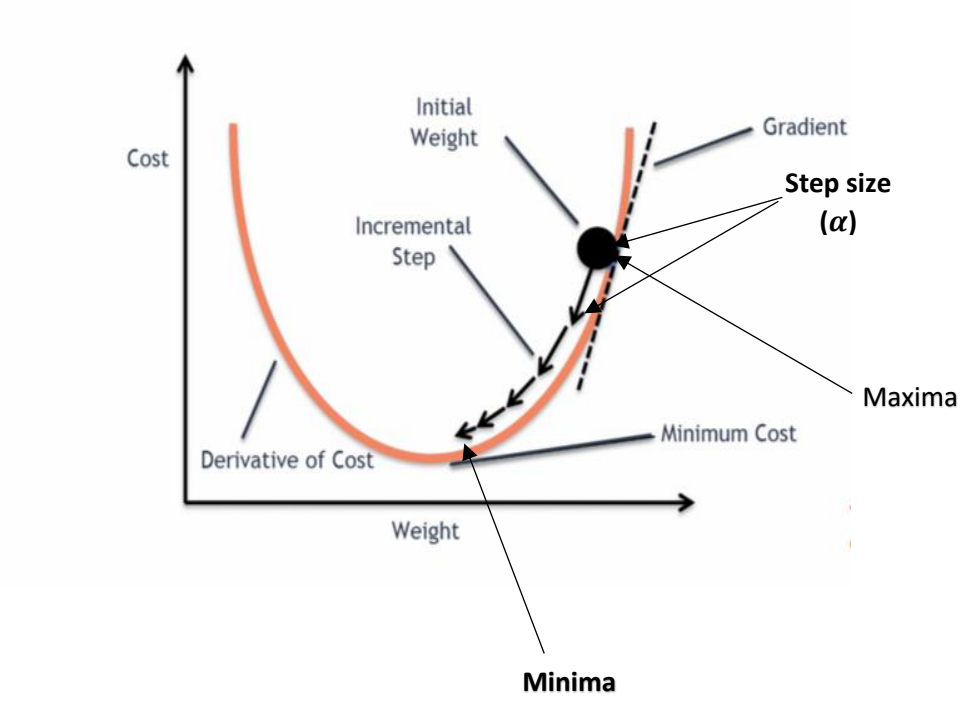
LMS Algorithm

- Start a search algorithm (e.g. gradient descent algorithm,) with initial guess of θ.
- Repeatedly update θ to make J(θ) smaller, until it converges to minima.

$$\beta_j = \beta_j - \alpha \frac{\partial}{\partial \beta_j} J(\theta)$$

- *J* is a convex quadratic function, so has a single global minima. gradient descent eventually converges at the global minima.
- At each iteration this algorithm takes a step in the direction of steepest descent(-ve direction of gradient).

α *is the step size, if* α *is small we take smaller steps and if* α *is large we take larger steps.*



$$\beta_j = \beta_j - \alpha \frac{\partial}{\partial \beta_j} J(\theta)$$

Where

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left(h(x)^i - y^i \right)^2$$

Therefore putting value of (θ) , if we have only one training example (x,y):

$$\beta_j = \beta_j - \alpha \frac{\partial}{\partial \theta_j} \times \frac{1}{2} \sum_{i=1}^{m} \left(h(x)^i - y^i \right)^2$$

$$=> \beta_j - \alpha \frac{\partial}{\partial \theta_j} \times \frac{1}{2} \sum_{i=1}^{m} \left(h(x)^i - y^i \right)^2$$

By applying power rule$\left(\frac{\partial}{\partial x} (x^n) = n x^{n-1} \right)$:

$$=> \beta_j - \alpha \times \frac{1}{2} \sum_{i=1}^{m} \{ 2 (h(x)^i - y^i)^{2-1} \times \frac{\partial}{\partial \theta_j} (h(x_i)^i - y^i) \}$$

$$=> \beta_j - \alpha \times \frac{1}{2} \sum_{i=1}^{m} \{ 2 (h(x)^i - y^i) \times \frac{\partial}{\partial \theta_j} (h(x_i)^i - y^i) \}$$

we know θ *is the change in the gradient for* i_{th} *value* , hence

$$h = \sum_{i=1}^{m} \theta_i$$

$$similarly \ we \ know,$$

$$h(x) = \sum_{i=0}^{m} \beta_i x_i$$

$$\boldsymbol{hence\ we\ also\ can\ write\ ,} \boldsymbol{h(x) = \sum_{i=0}^{m} \theta_i x_i}$$

Similarly,

$$\boldsymbol{h(x)^i = \sum_{i=0}^{m} \theta_i (x_i)^i}$$

$$\boldsymbol{=> \beta_j - \alpha \times [\frac{1}{2} \sum_{i=1}^{m} \{2(\, h(x)^i - y^i) \frac{\partial}{\partial \theta_j} \left(\sum_{i=0}^{m} \theta_i (x_i)^i - y^i \right)\}]}$$

Then for $\boldsymbol{\theta_j}$

$$\boldsymbol{=> \beta_j - \alpha \times [\frac{1}{2} \sum_{i=1}^{m} \{2(\, h(x)^i - y^i) \frac{\partial}{\partial \theta_j} \left(\sum_{j=0}^{m} \theta_j (x_j)^i - y^i \right)\}]}$$

$$\boldsymbol{=> \beta_j - \alpha \times [\frac{1}{2} \sum_{i=1}^{m} \{2(\, h(x)^i - y^i) \left(\sum_{j=0}^{m} (\theta_j \frac{\partial}{\partial \theta_j} (x_j)^i + (x_j)^i \frac{\partial}{\partial \theta_j} \theta_j - 0 \right)\}]}$$

$$\boldsymbol{=> \beta_j - \alpha \times [\frac{1}{2} \sum_{i=1}^{m} \{2(\, h(x)^i - y^i) \left(\sum_{j=0}^{m} (0 + (x_j)^i \times 1 - 0 \right)\}]}$$

$$=> \boldsymbol{\beta_j - \alpha} \times [\frac{\mathbf{1}}{\mathbf{2}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\{\mathbf{2}(\boldsymbol{h(x)^i - y^i})\left(\sum_{\boldsymbol{j=0}}^{\boldsymbol{m}}(\boldsymbol{x_j})^{\boldsymbol{i}}\right) + (\boldsymbol{h(x)^i - y^i})\left(\sum_{\boldsymbol{j=0}}^{\boldsymbol{m}}(\boldsymbol{x_j})^{\boldsymbol{i}}\right)\}]$$

$$=> \boldsymbol{\beta_j - \alpha} \times [\frac{\mathbf{1}}{\mathbf{2}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\{\mathbf{2}(\boldsymbol{h(x)^i - y^i})\left(\sum_{\boldsymbol{j=0}}^{\boldsymbol{m}}(\boldsymbol{x_j})^{\boldsymbol{i}}\right)\}]$$

$$=> \boldsymbol{\beta_j - \alpha} \times [\frac{\mathbf{2}}{\mathbf{2}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\{(\boldsymbol{h(x)^i - y^i})\left(\sum_{\boldsymbol{j=0}}^{\boldsymbol{m}}(\boldsymbol{x_j})^{\boldsymbol{i}}\right)\}]$$

$$=> \boldsymbol{\beta_j - \alpha} \times [\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\{(\boldsymbol{h(x)^i - y^i})\left(\sum_{\boldsymbol{j=0}}^{\boldsymbol{m}}(\boldsymbol{x_j})^{\boldsymbol{i}}\right)\}]$$

$$=> \boldsymbol{\beta_j + \alpha} \times [\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\{(\boldsymbol{y^i - h(x)^i})\left(\sum_{\boldsymbol{j=0}}^{\boldsymbol{m}}(\boldsymbol{x_j})^{\boldsymbol{i}}\right)\}]$$

If we remove Σ from the equations, we get:

$$\boldsymbol{\beta_j + \alpha \times \left(y^i - h(x)^i\right)\left(x_j\right)^i}$$

LMS Update Rule

- If you have only one training example (x, y)

$$\frac{\partial}{\partial\theta}J(\theta) = \frac{\partial}{\partial\theta_j}\frac{1}{2}(h(x)-y)^2$$

$$= 2.\frac{1}{2}(h(x)-y)\frac{\partial}{\partial\theta_j}(h(x)-y)$$

$$= (h(x)-y).\frac{\partial}{\partial\theta_j}\left(\sum_{i=0}^{n}\theta_i x_i - y\right)$$

$$= (h(x)-y)x_j$$

- For a single training example, this gives the update rule:

$$\beta_j = \beta_j + \alpha(y^{(i)} - h(x^i))x_j{}^{(i)}$$