Subject Name: **Machine Learning**

Subject Code: **MCA-4014**

Subject Topic: **Mathematics for Machine Learning : Linear Regression & Least Square Regression**

Abhishek Dwivedi

Assistant Professor

Department of Computer Application

UIET, CSJM University, Kanpur

What is Linear Regression?

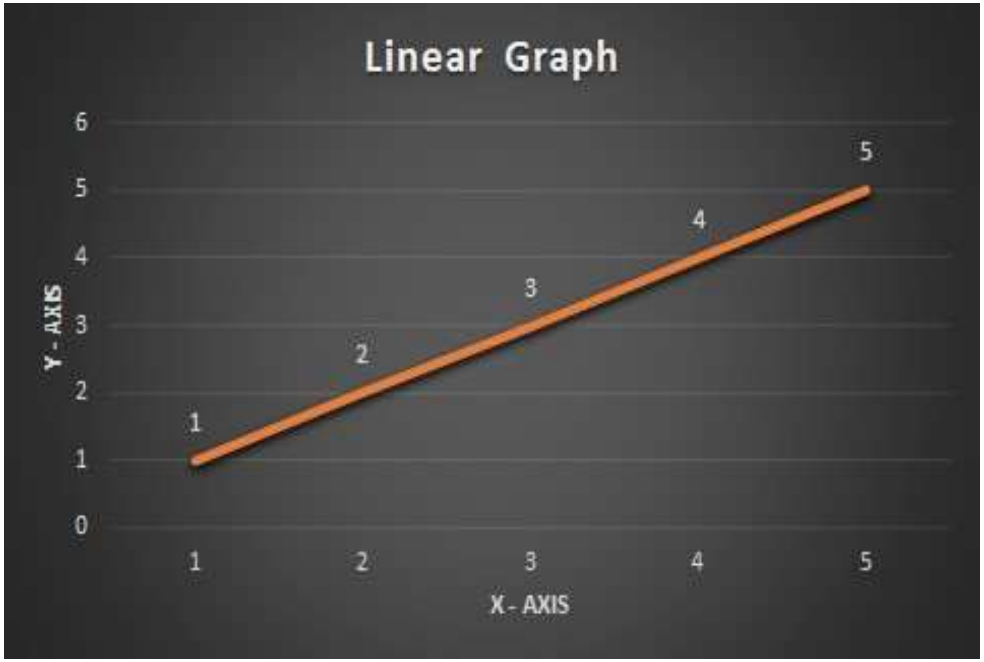
Linear Regression is a predictive algorithm which provides a Linear relationship between **Prediction** *(Call it* **'Y'***) and* ***Input*** *(Call is* **'X'***).*

As we know from the basic maths that if we plot an **'X','Y'** graph, a linear relationship will always come up with a straight line. For example, if we plot the graph of these values

(Input) X = 1,2,3,4,5

(Prediction) Y = 1,2,3,4,5

It will be a perfectly straight line



Linear Straight Line graph

- **Equation of Straight Line from 2 Points**

The equation of a straight line is written using the y = mx + b, where m is the slope (Gradient) and b is y-intercept (where the line crosses the Y axis).

Once we get the equation of a straight line from 2 points in space in y = mx + b format, we can use the same equation to predict the points at different values of x which result in a straight line.

In this formula, m is the slope and b is y-intercept.

Note: Linear regression is a way to predict the 'Y' values for unknown values of Input 'X' like 1.5, 0.4, 3.6, 5.7 and even for -1, -5, 10 etc.

Linear Regression with Real World Example

Let’s take a real world example to demonstrate the usage of linear regression and usage of Least Square Method to reduce the errors.

a real world example of the price of agricultural products and how it varies based on the location its sold. The price will be low when bought directly from farmers and high when brought from the downtown area.

Given this dataset, we can predict the price of the product in intermediate locations.

Note: When a dataset is used for predictions, it’s also called as Training Data Set

Agricultural Product and its price at point of sale

In this example, if we consider Input 'X — Axis' as Sale Location and 'Y — Axis' as Price (think of any currency you're familiar with), we can plot the graph as

Graph: Agricultural Product and its price at point of sale

Problem Statement

Note: Given this dataset, predict the price of agricultural product, if it's sold in intermediate locations between farmers house and city downtown

Training DataSet:

The dataset provided above can be considered as Training DataSet for the problem statement stated above, If we consider these inputs as Training Data for the model, we can use that model to predict the price at locations between

- Farmers home — Village
- Village — Town
- Town — City
- City — City Downtown

Our aim is to come with a straight line which minimizes the error between training data and our prediction model when we draw the line using the equation of straight line.

Equation of Straight Line ($y = mx + b$)

The maths allow us to get a straight line between any two (x,y) points in two dimensional graph. For this example, let's consider farmers home and price as starting point and city downtown as ending point.

The coordinates of the start and end points will be

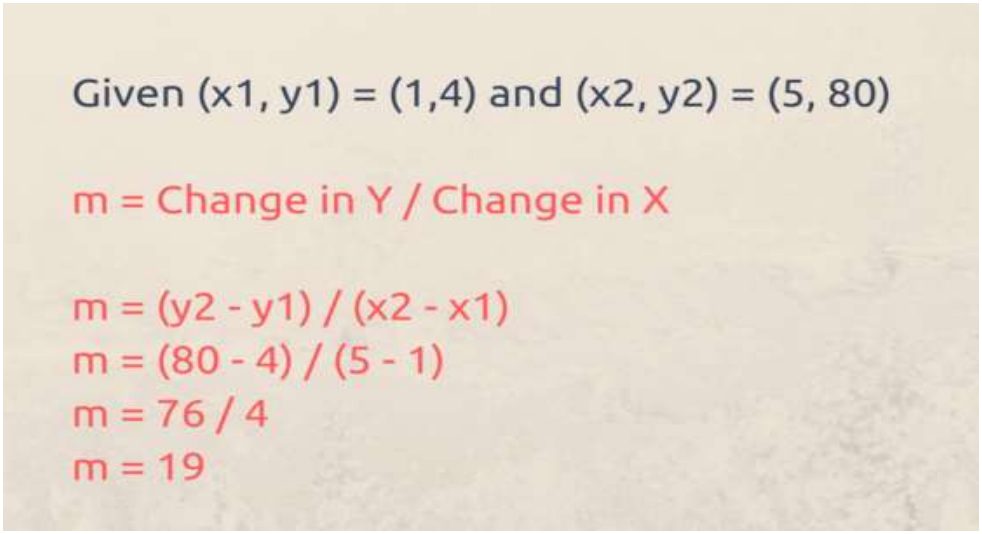
(x1,y1) = (1, 4)
(x2,y2) = (5, 80)

Note: where x represents the location and y represent the price.

The first step is to come up with a formula in the form of
y = mx +b where x is a known value and y is the predicted value.
To calculate the Prediction y for any Input value x we have two unknowns, the m = slope(Gradient) and b = y-intercept(also called bias)

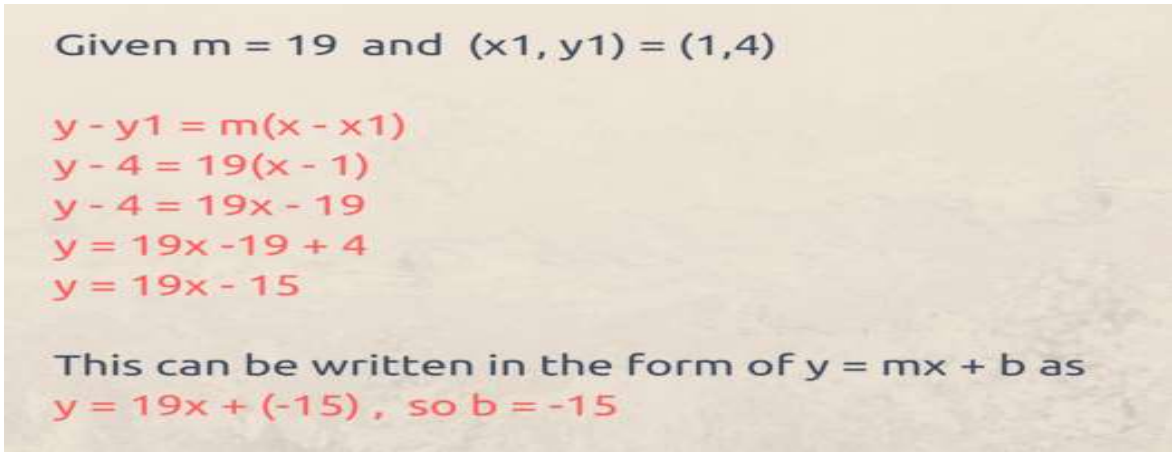
Slope (m = Change in y/ Change in x)

The slope of the line is calculated as the change in y divided by change in x, so the calculation will look like



Calculating m = Change in Y / Change in X

The y-intercept / bias shall be calculated using the formula y-y1 = m(x-x1)



Finding y = mx + b

Once we arrived at our formula, we can verify the same by substituting x for both starting and ending points which were used to calculate the formula as it should provide the same y value.

Given Formula	{ y = 19x + (-15) }
x1 => 1	{ y = 19 * 1 - 15 => 19 - 15 => 4 (i.e. y1)
x2 => 5	{ y = 19 * 5 - 15 => 95 - 15 => 80 (i.e. y2)

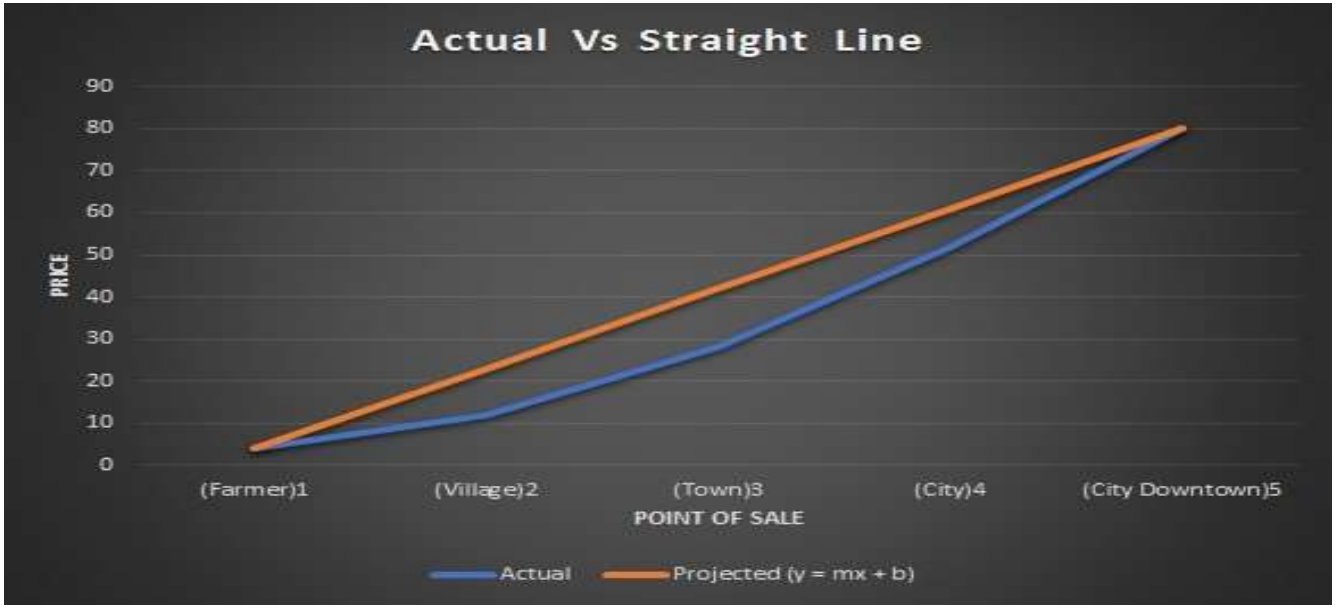
Verifying y = mx + b

Now we know that our formula is correct as we get the same y value by substituting the x value, but what about other values of x in between i.e 2,3,4 , let's find out

Given Formula	{ y = 19x + (-15) }
x => 2	{ y = 19 * 2 - 15 => 38 - 15 => 23
x => 3	{ y = 19 * 3 - 15 => 57 - 15 => 42
x => 4	{ y = 19 * 4 - 15 => 76 - 15 => 61

Predicting Y values for unknown X values

These values are different from what was actually there in the training set (understandably as original graph was not a straight line), and if we plot this(x,y) graph against the original graph, the straight line will be way off the original points in the graph of x=2,3, and 4.



Graph: Actual Line Vs Projected Straight Line

Note: Nevertheless, the first step is successful as we managed to predict the Y *for unknown values of* X

Minimizing the Error :

The error is defined as the difference of values between actual points and the points on the straight line). Ideally., we'd like to have a straight line where the error is minimized across all points.

Note: The are many mathematical ways to do the same and one of the methods is called Least Square Regression

Least Square Regression

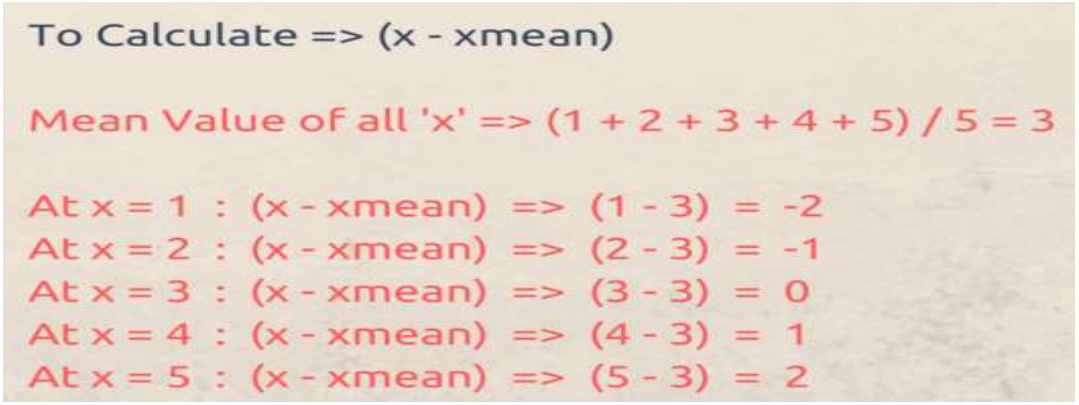
Least Square Regression is a method which minimizes the error in such a way that the sum of all square error is minimized. Here are the steps you use to calculate the Least square regression.

First, the formula for calculating m = slope is

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

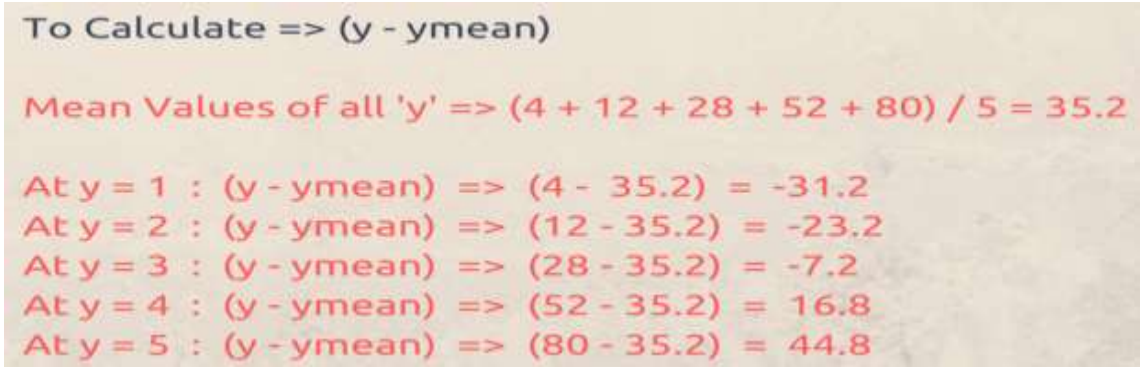
Calculating slope(m) for least square

So let’s calculate all the values required to come up with the slope(m), first start with calculating values with x



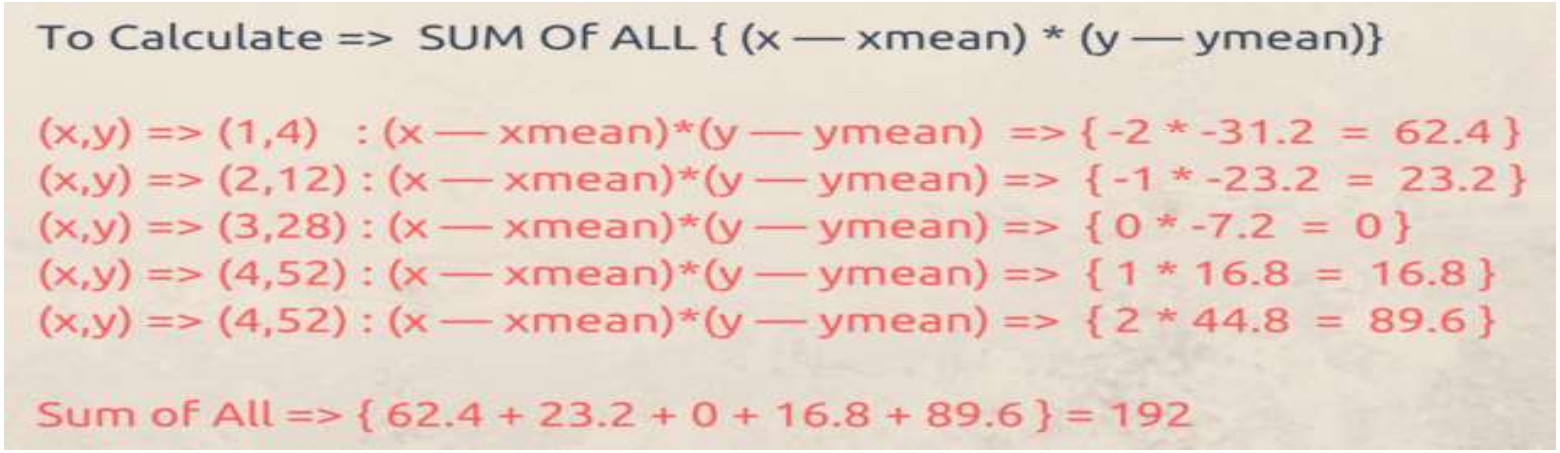
Calculating x — xmean for all X values

Now let’s calculate the values with y

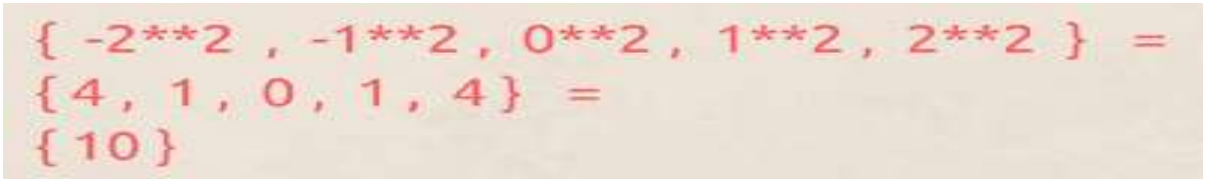


Calculating y— ymean for all Y values

The availability of these values allows us to calculate Sum of all (x — xmean)*(y — ymean)



Now let's calculate the denominator part of the equation which is Sum of $(x — xmean)^2$



So the overall calculation would be
m = 192/10 = 19.2

Calculation of y-Intercept

The y-intercept is calculated using the formula b = ymean — m * xmean

```
To Calculate  =>  b = ymean — m * xmean

b = 35.2 - 19.2 * 3
b = 35.2 - 57.6
b = -22.4
```

The overall formula can now be written in the form of y = mx + b as

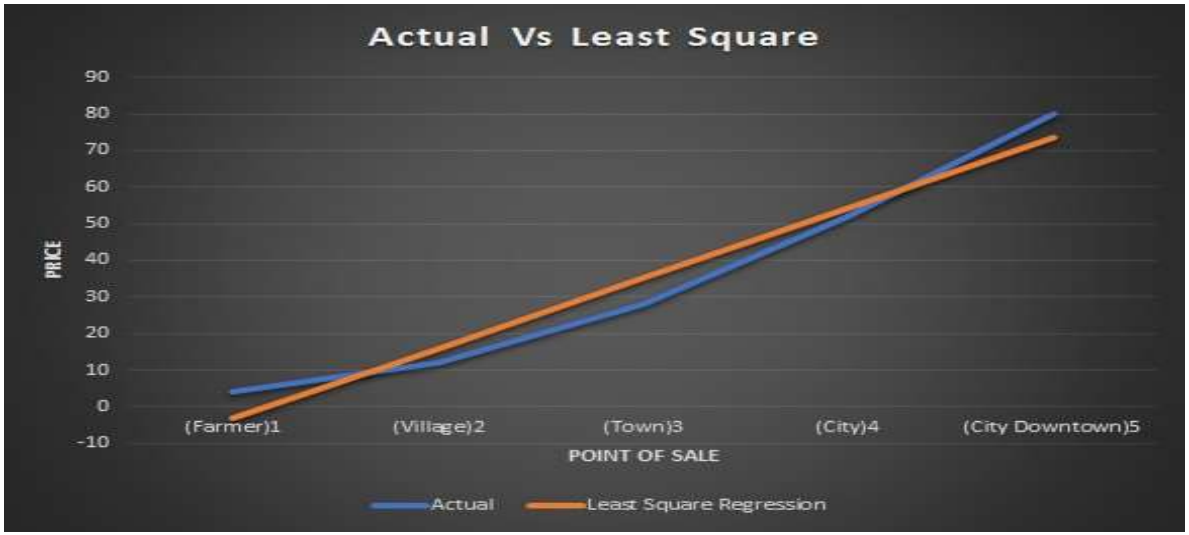
y = 19.2 x – 22.4

Using Least Square Regression on X,Y values

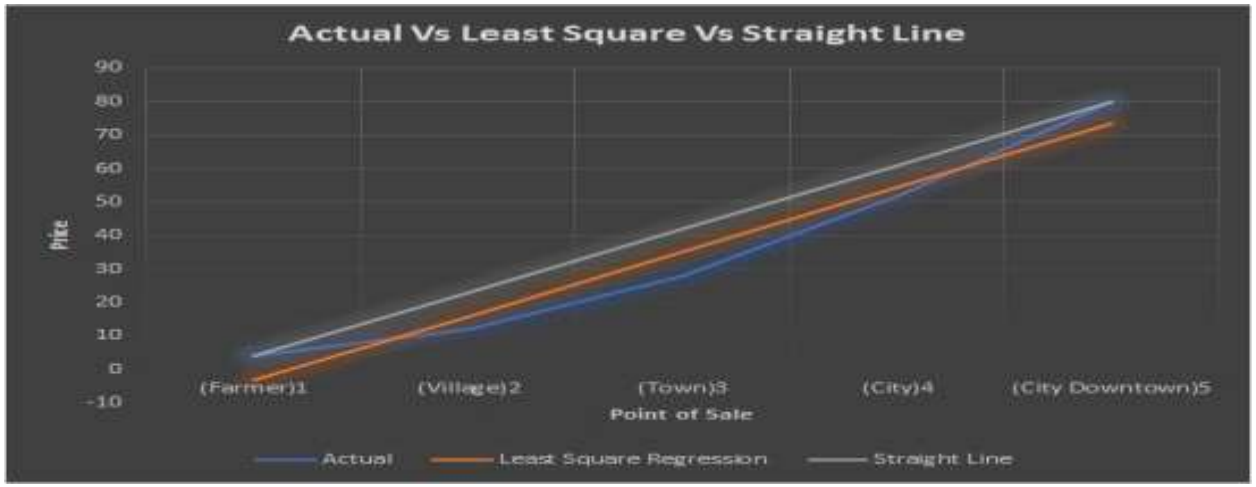
Let's see how the prediction y changes when we apply
y = 19.2 x -22.4 on all x values.

```
x => 1   { y = 19.2 * 1 - 22.4 } = -3.2
x => 2   { y = 19.2 * 2 - 22.4 } = 16
x => 3   { y = 19.2 * 3 - 22.4 } = 35.2
x => 4   { y = 19.2 * 4 - 22.4 } = 54.4
x => 5   { y = 19.2 * 5 - 22.4 } = 73.6
```

Let’s plot this particular straight line graph against the standard values.



As we can see that these values are nearer to the actual line as compared to direct straight line values between starting and end points. If we compare this with the straight line graph we visualize the difference



Why this method is called Least Square Regression ?

This method is intended to reduce the sum square of all error values. The lower the error, lesser the overall deviation from the original point. We can compare the same with the errors generated out of the straight line as well as with the Least Square Regression

Error in 'Y' With Straight Line Equation

At X-Value	Y-Value	Actual-Value	Error	Square-Error
1	4	4	0	0
2	12	23	-11	121
3	28	42	-14	196
4	52	61	-9	81
5	80	80	0	0

Sum of Square Error = { 0 + 121 + 196 + 81 + 0 } = 398

Error in 'Y' With Least Square Equation

At X-Value	Y-Value	Actual-Value	Error	Square-Error
1	4	-3.2	7.2	51.84
2	12	16	-4	16
3	28	35.2	-7.2	51.84
4	52	54.4	-2.4	5.76
5	80	73.6	6.4	40.96

Sum of Square Error = { 51.84 + 16 + 51.84 + 5.76 + 40.96 } = 166.4

So that Least Square Method provide better results than a plain straight line between two points calculation.