Linear Regression

Linear regression is used for finding linear relationship between target and one or more predictors. There are two types of linear regression- Simple and Multiple.
Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight.
The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the point to the regression line.

Real-Time Example

We have a dataset which contains information about relationship between 'number of hours studied' and 'marks obtained'. Many students have been observed and their hours of study and grade are recorded. This will be our training data. Goal is to design a model that can predict marks if given the number of hours studied. Using the training data, a regression line is obtained which will give minimum error. This linear equation is then used for any new data. That is, if we give number of hours studied by a student as an input, our model should predict their mark with minimum error.

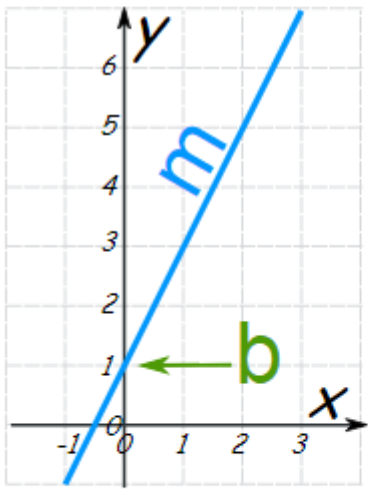
Y(Pred)= mx+c **c=intercept** **m=slope** **x= independent variable**

The equation of a straight line is usually written this way:

$$y = mx + b$$

"y = mx + c"

What does it stand for?



$$y = mx + b$$

Slope or Gradient

y when x=0 *(see Y Intercept)*

y = how far up

x = how far along

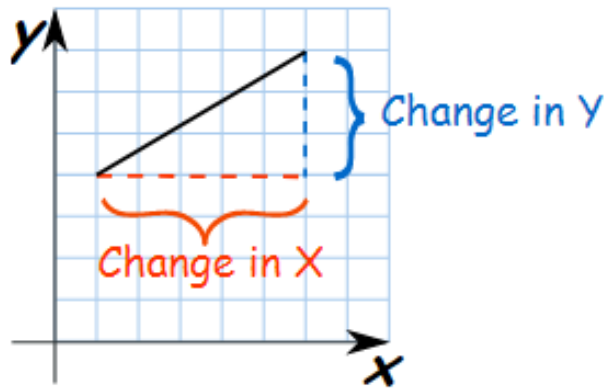
m = Slope or Gradient (how steep the line is)

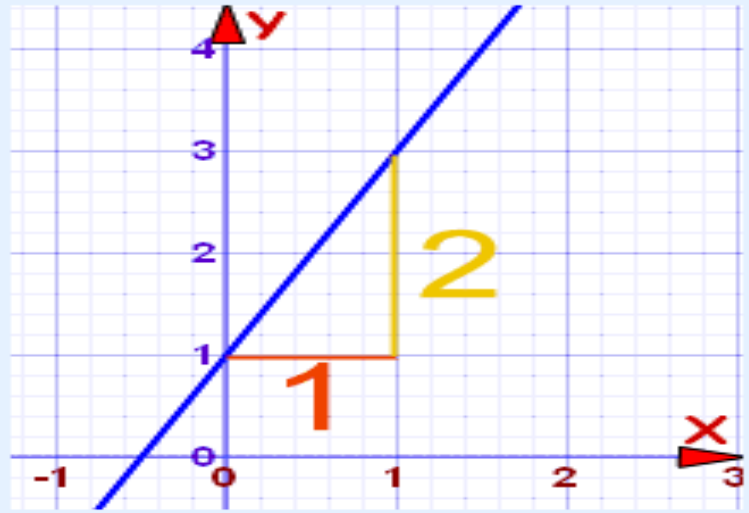
b = value of **y** when **x=0**

How do you find "m" and "b"?

- b is easy: just see where the line crosses the Y axis.
- m (the Slope) needs some calculation:

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



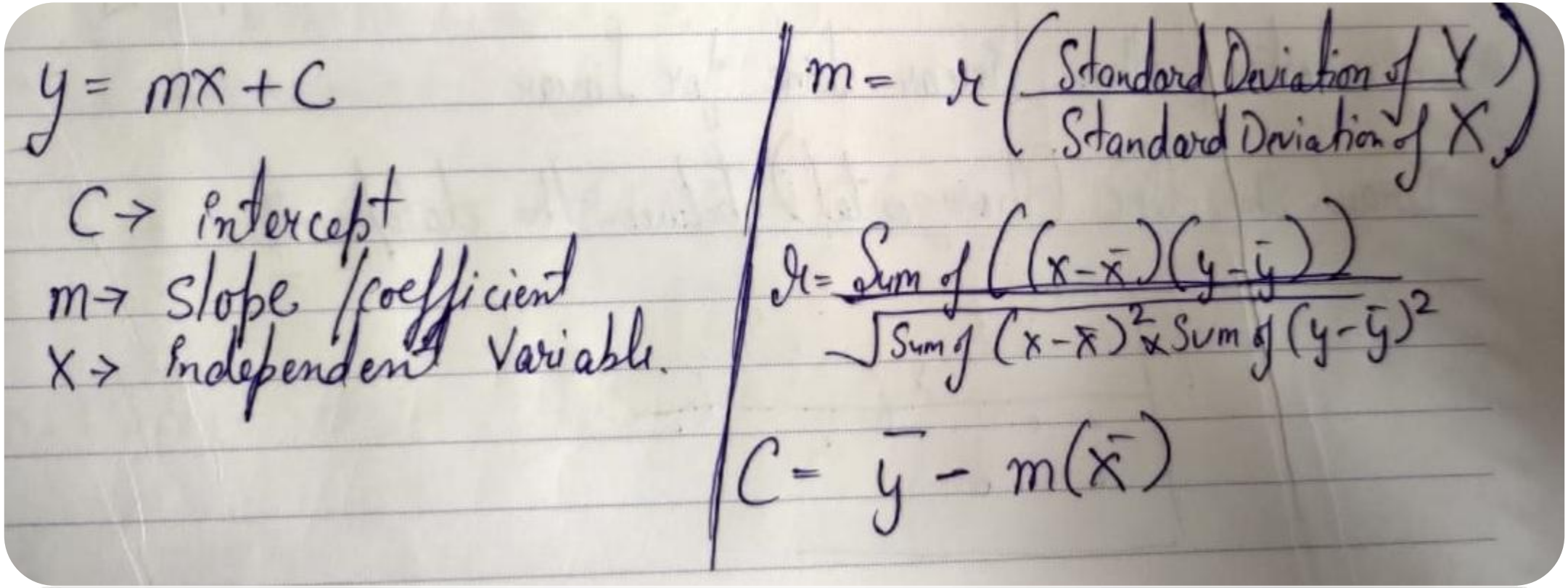


$$m = \frac{2}{1} = 2$$

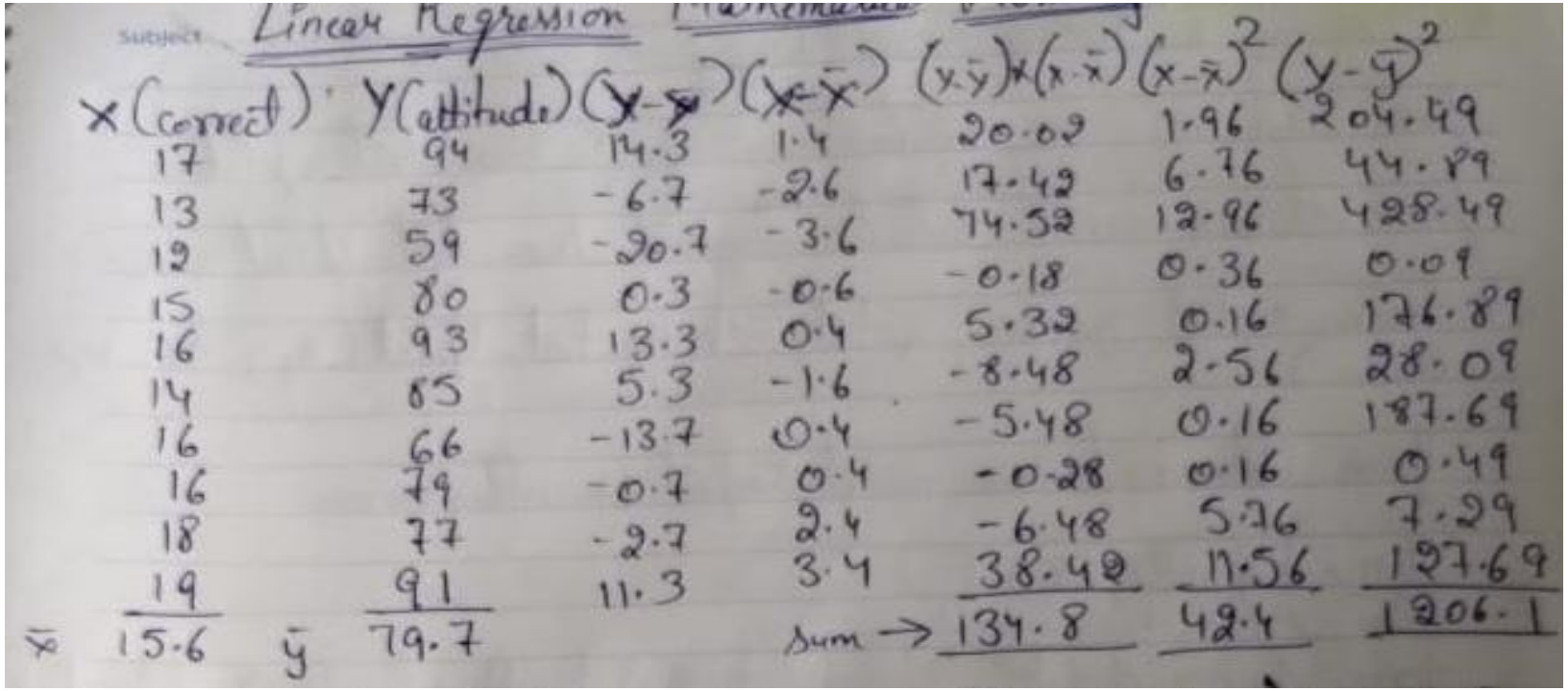
b = 1 *(value of* ***y*** *when x=0)*

So: y = 2x + 1

Mathematical Intuition



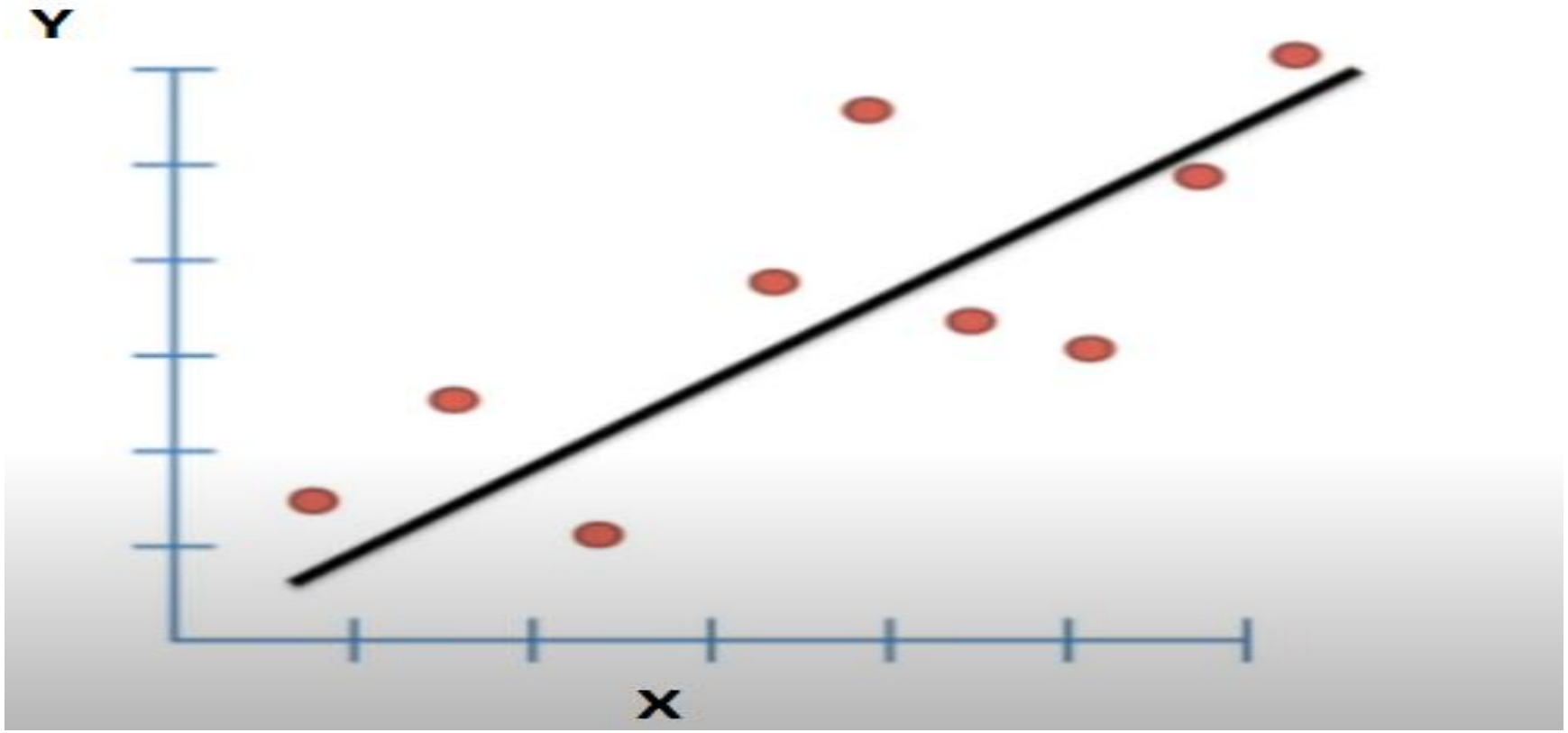
A Basic Linear Regression Exercise



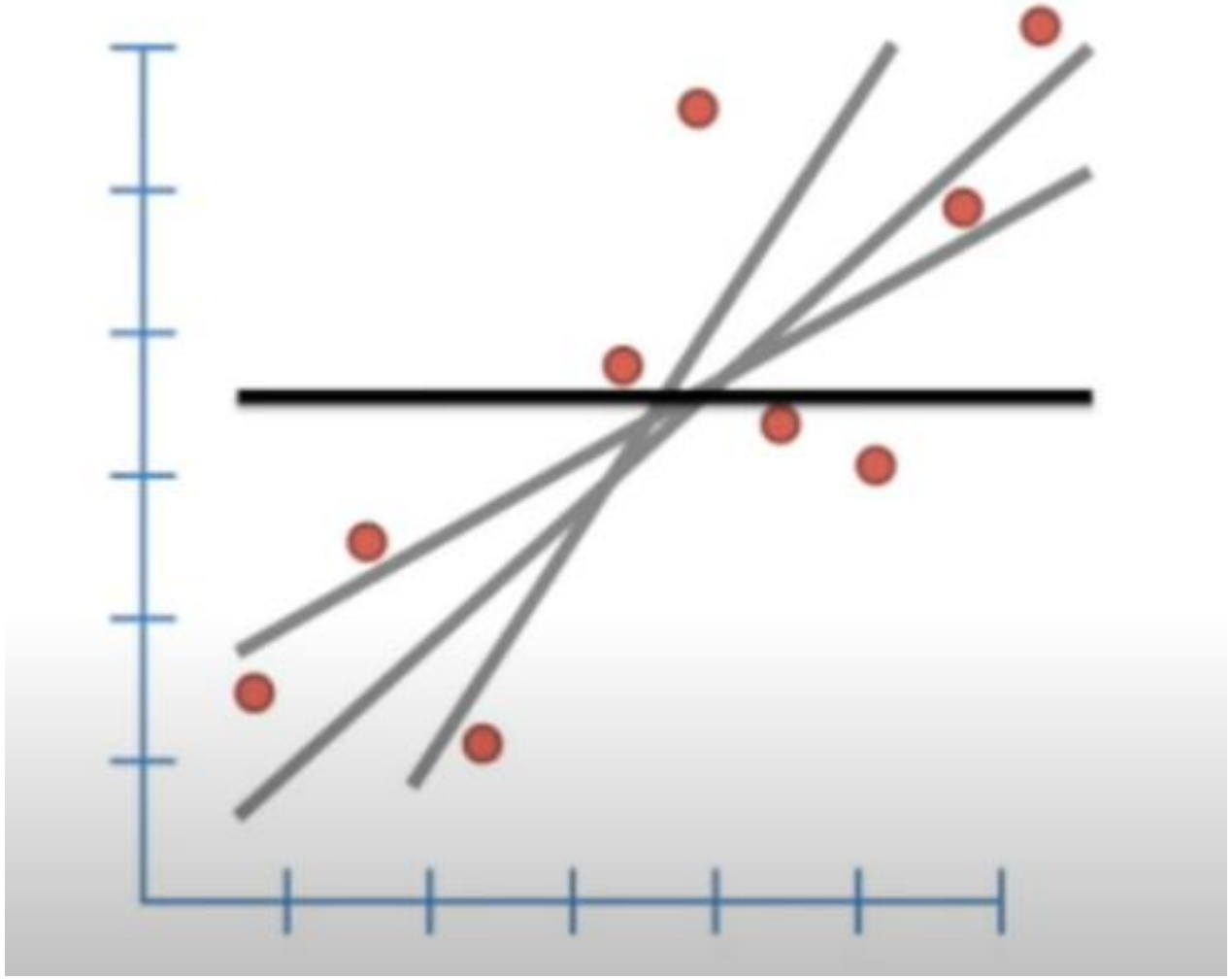
Linear Regression

x (correct)	y (attitude)	$(y-\bar{y})$	$(x-\bar{x})$	$(x-\bar{x})(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
17	94	14.3	1.4	20.02	1.96	204.49
13	73	-6.7	-2.6	17.42	6.76	44.89
12	59	-20.7	-3.6	74.52	12.96	428.49
15	80	0.3	-0.6	-0.18	0.36	0.09
16	93	13.3	0.4	5.32	0.16	176.89
14	85	5.3	-1.6	-8.48	2.56	28.09
16	66	-13.7	0.4	-5.48	0.16	187.69
16	79	-0.7	0.4	-0.28	0.16	0.49
18	77	-2.7	2.4	-6.48	5.76	7.29
19	91	11.3	3.4	38.42	11.56	127.69
$\bar{x}$ 15.6	$\bar{y}$ 79.7		Sum →	134.8	42.4	1206.1

How to find best fit Line?

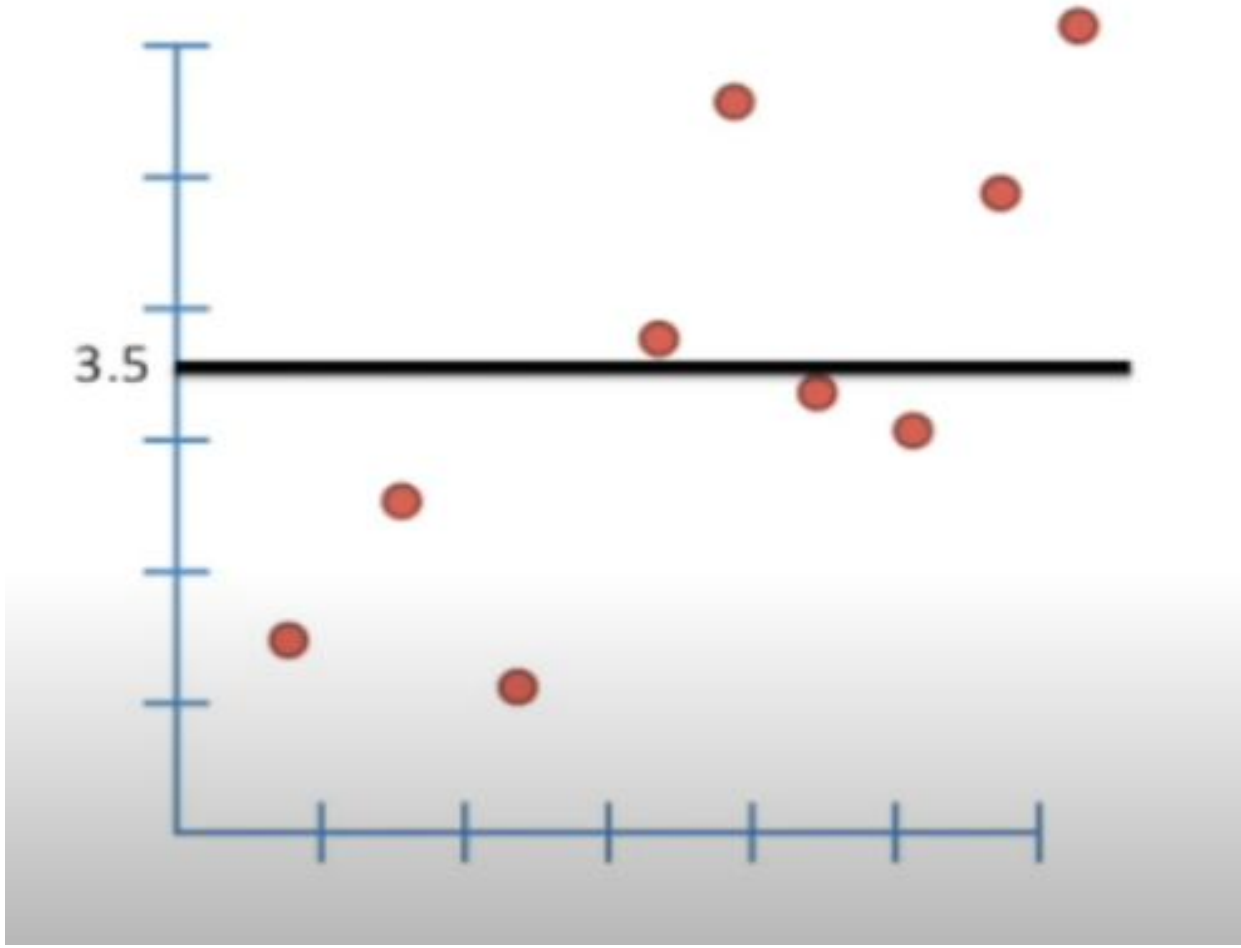


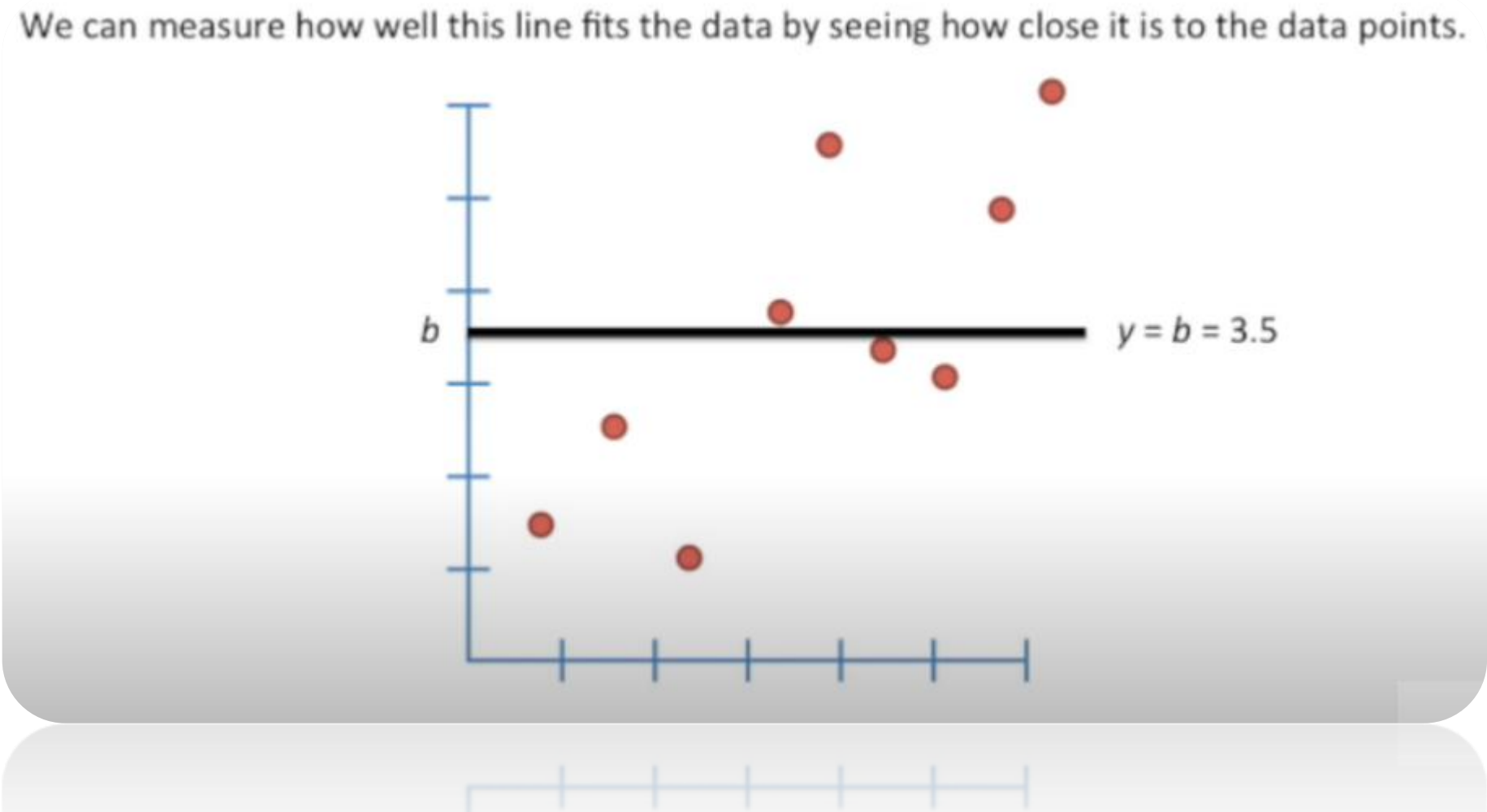
Let’s start with some datapoints that are plotted on the graph and also X axis and Y axis are also denoted. Make sure to remember the X axis and Y axis as you want to see in coming slides. A perfect line is drafted randomly but how?



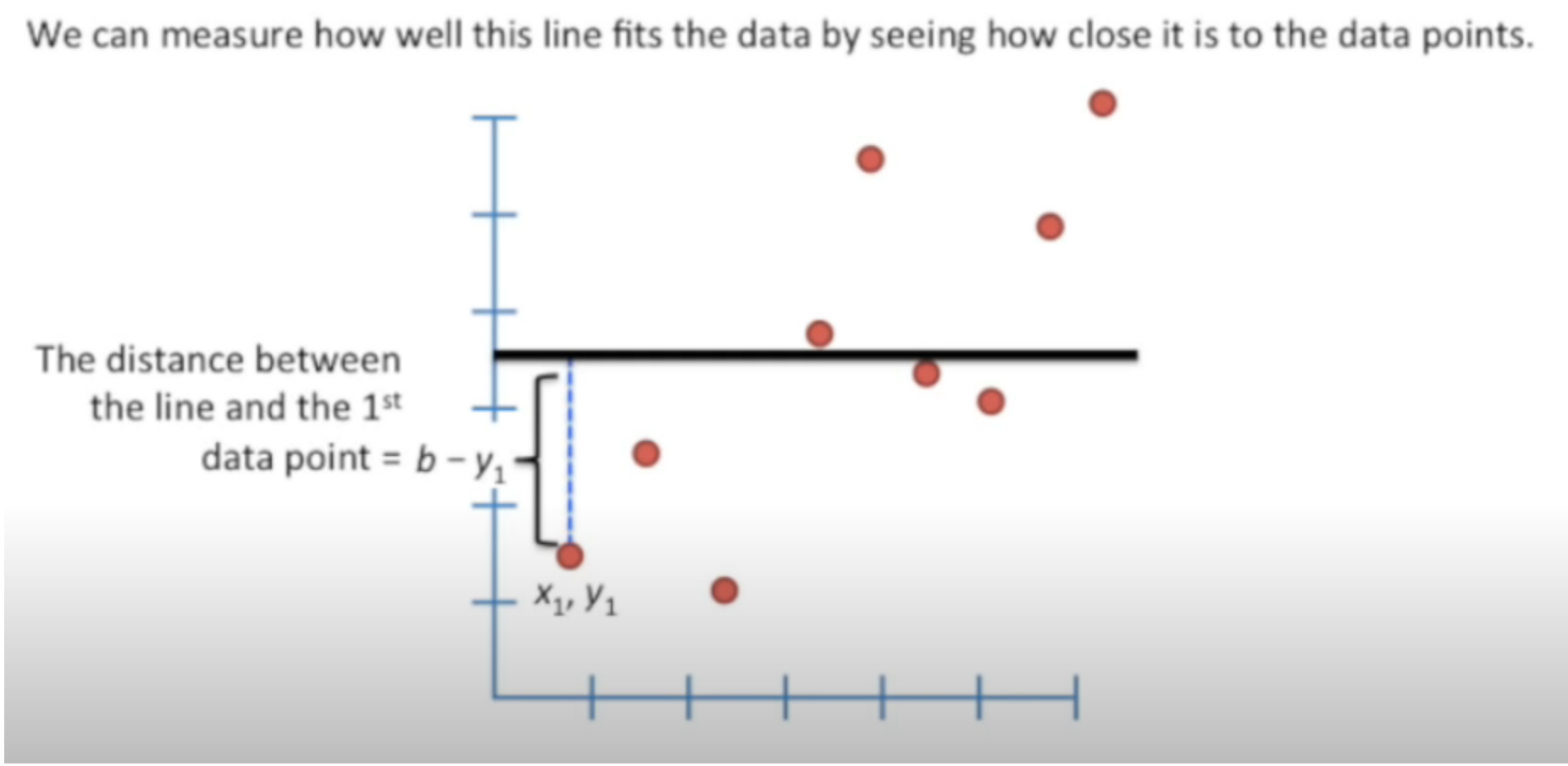
There are some randomly drawn Lines as well that justify the correctly fit line

We’ll start the exercise by inserting a line on the average of Y axis where in this exercise it is 3.5.

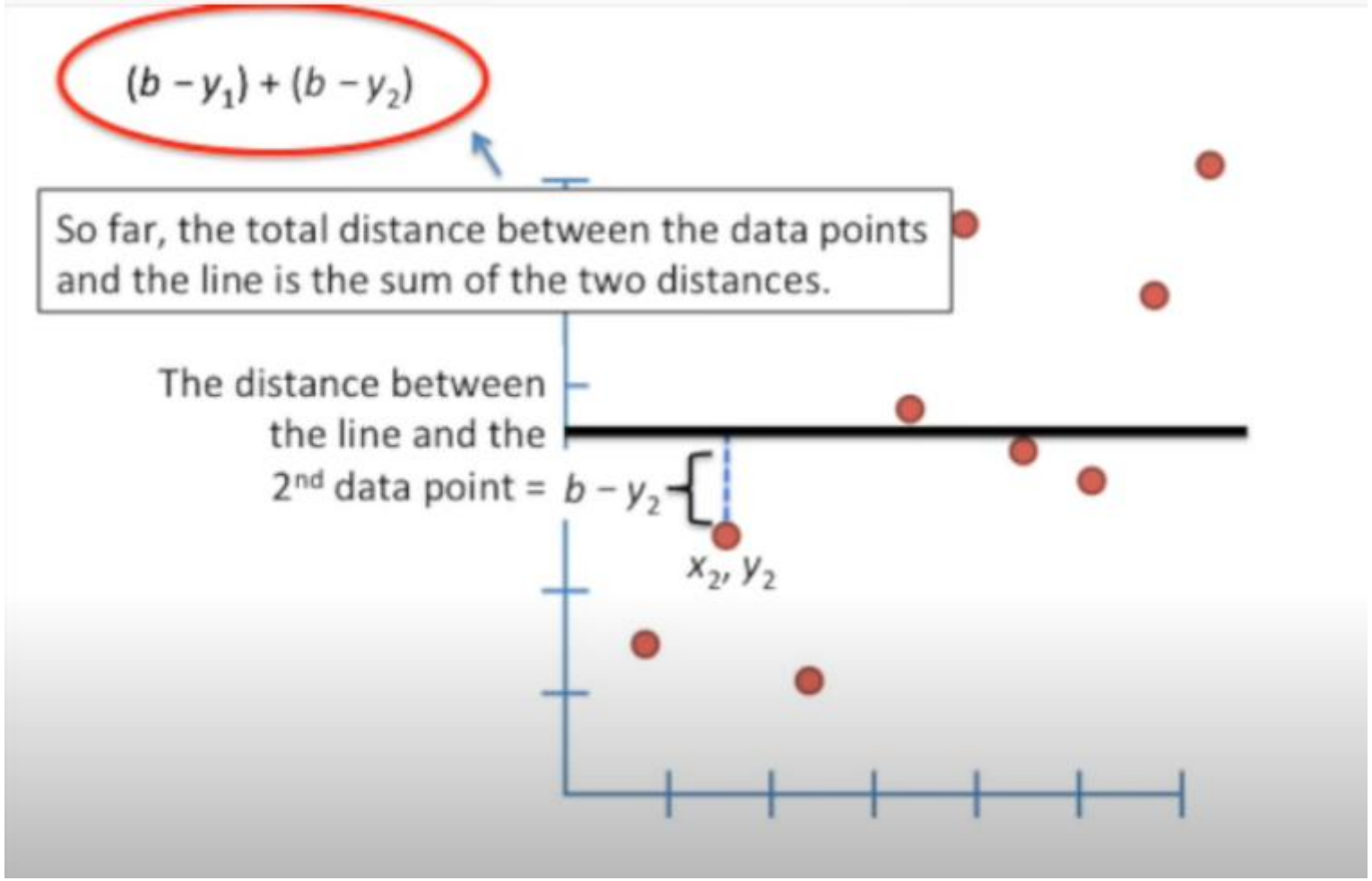




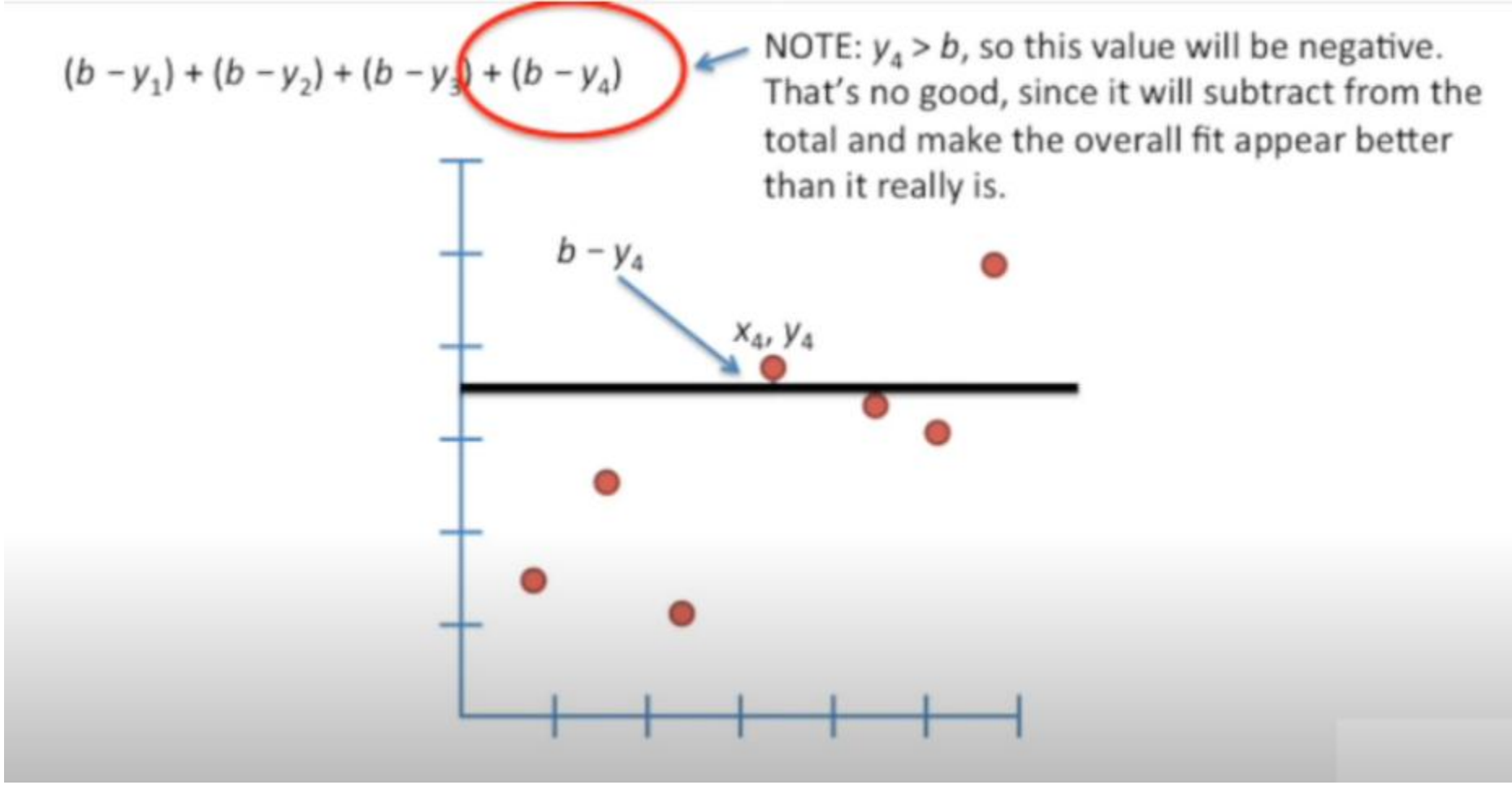
In this image we have named the line as **b** and **y** is just a axis because it is formed from **y-axis.**



From the above image the distance is calculated with the help of b-y1 equation and will be calculated for further datapoints



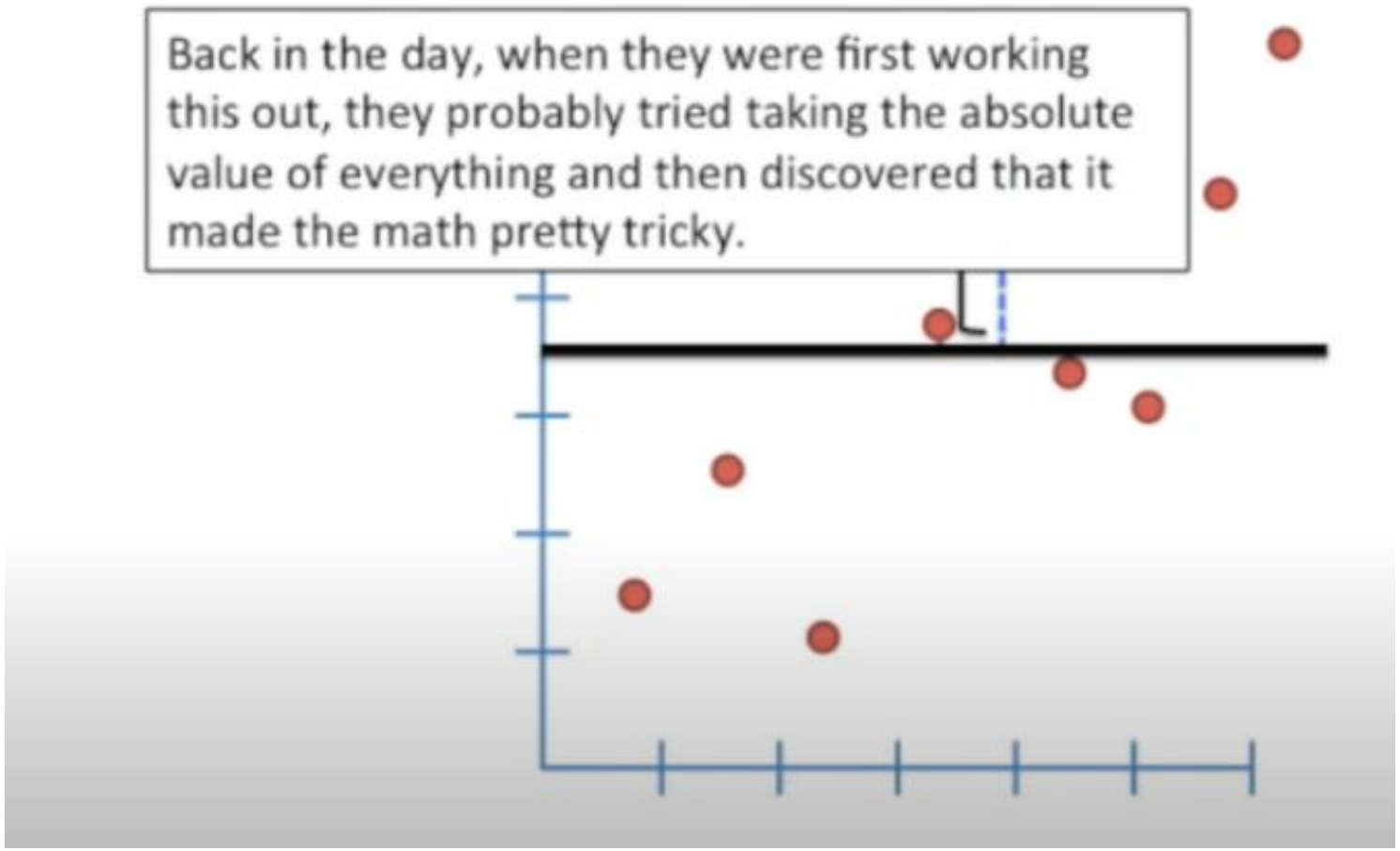
We will also adding the equations by subtracting the datapoints from different datapoints



The data points that are being calculated will be squared because there are some points that are in Negative. To eliminate the negative points we will square it.

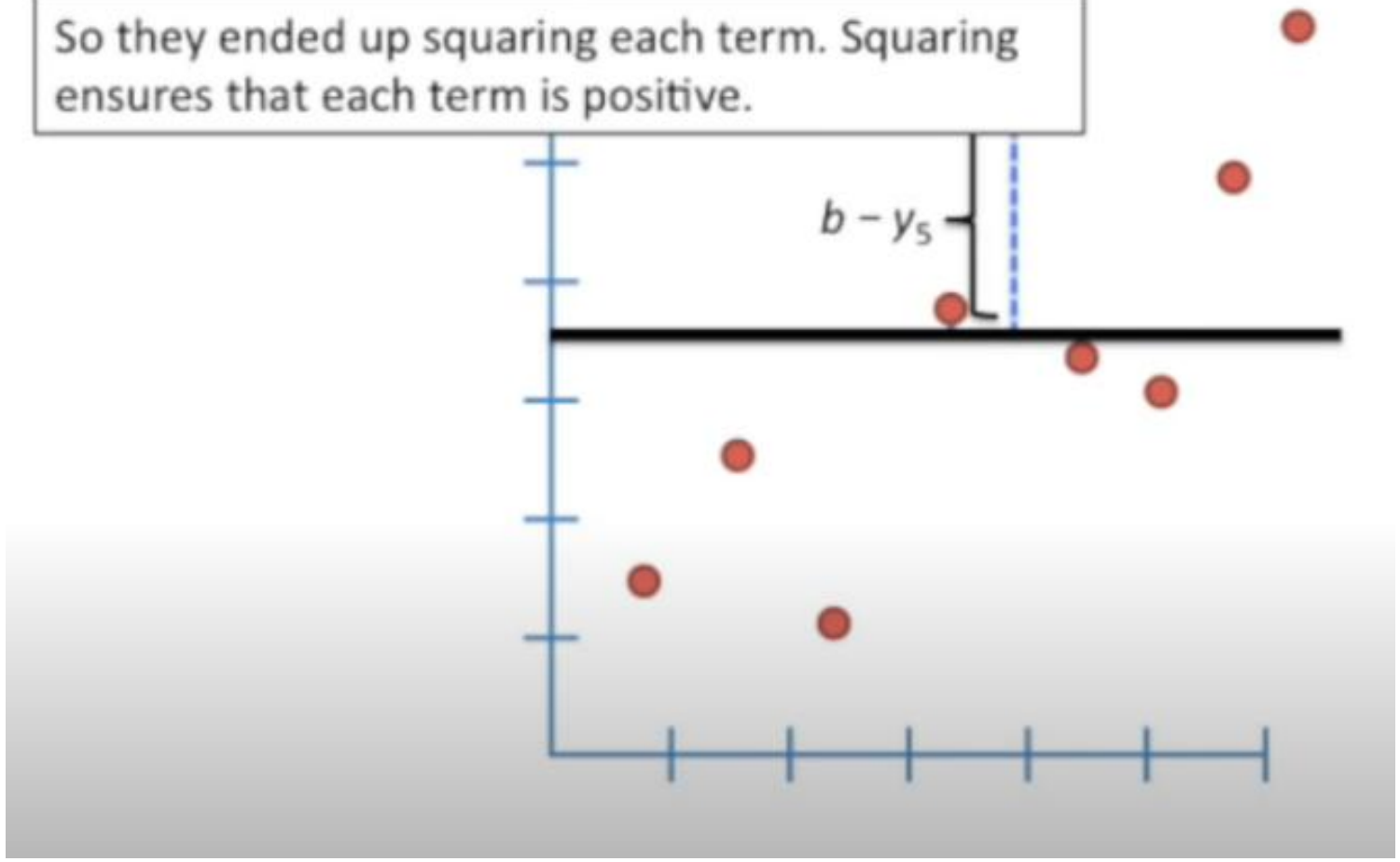
$$|(b - y_1)| + |(b - y_2)| + |(b - y_3)| + |(b - y_4)| + |(b - y_5)|$$

Back in the day, when they were first working this out, they probably tried taking the absolute value of everything and then discovered that it made the math pretty tricky.



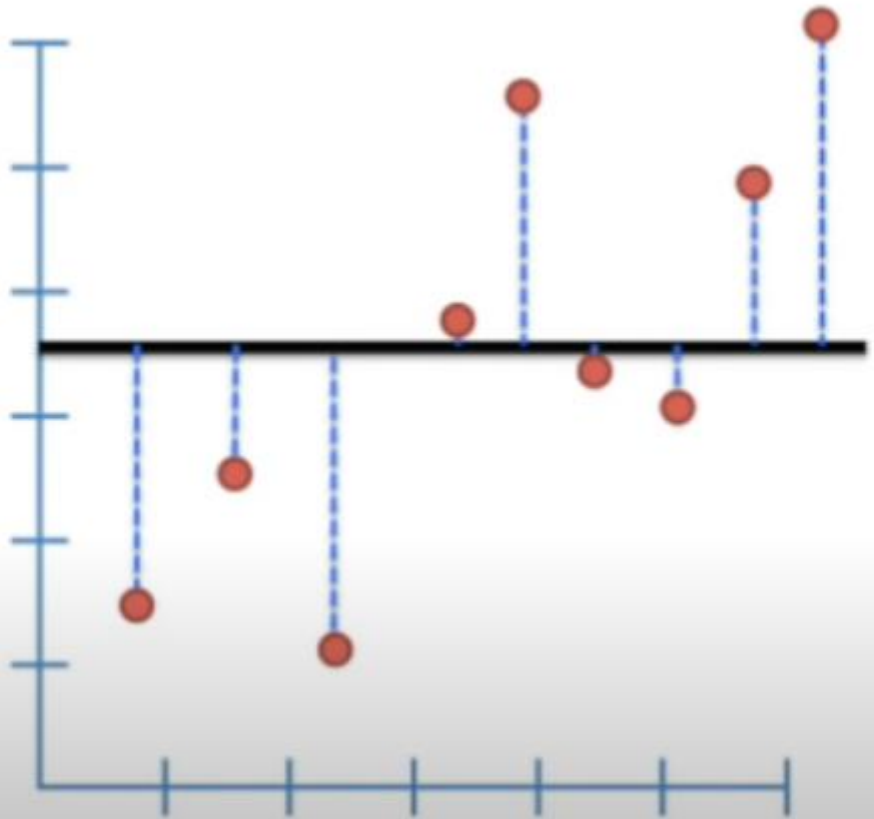
$$(b - y_1)^2 + (b - y_2)^2 + (b - y_3)^2 + (b - y_4)^2 + (b - y_5)^2$$

So they ended up squaring each term. Squaring ensures that each term is positive.



Understood the Idea?

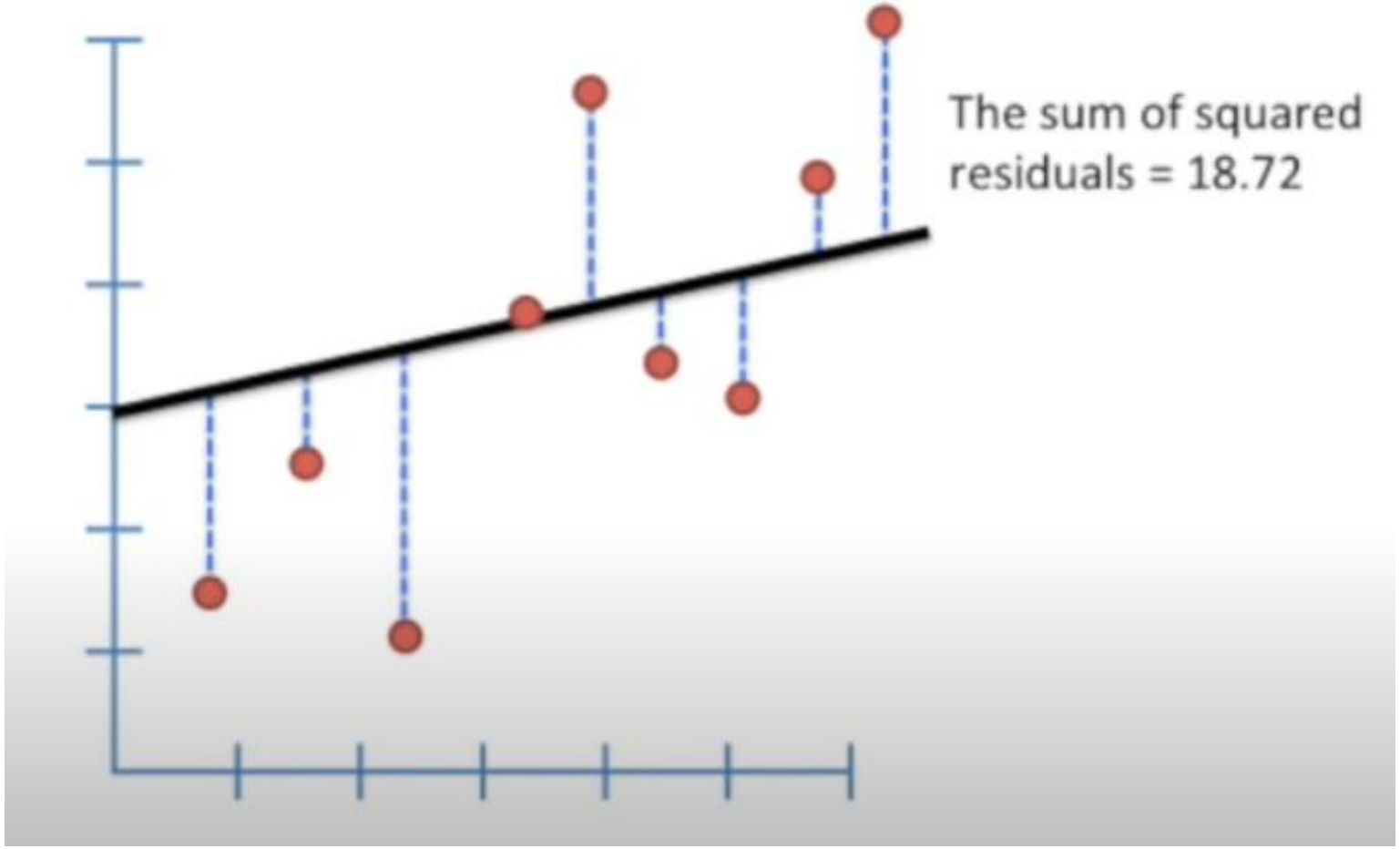
$$(b - y_1)^2 + (b - y_2)^2 + (b - y_3)^2 + (b - y_4)^2 + (b - y_5)^2 + (b - y_6)^2 + (b - y_7)^2 + (b - y_8)^2 + (b - y_9)^2$$



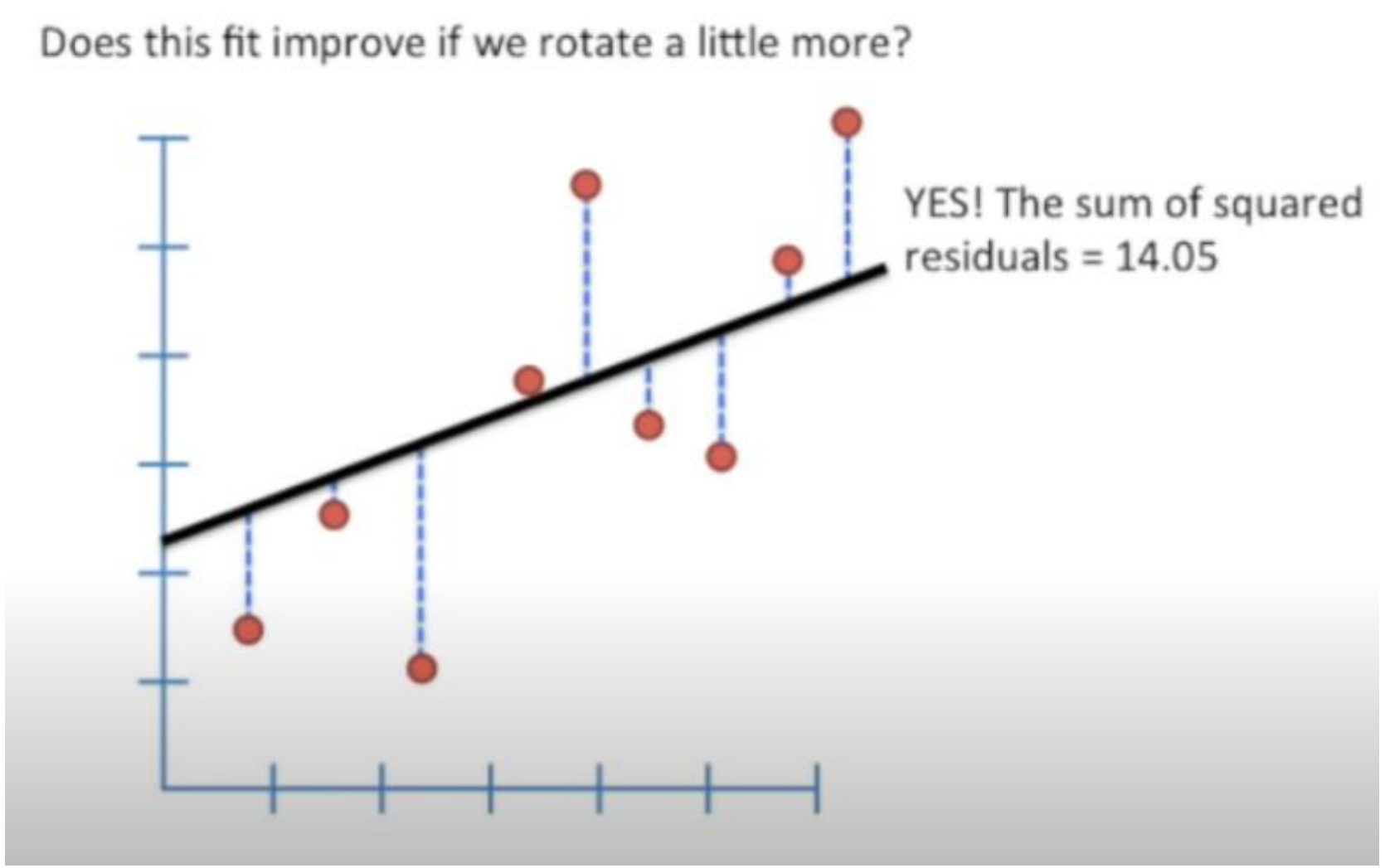
= 24.62

This is our measure of how well this line fits the data.

It's called the "sum of squared residuals, because the residuals are the differences between the real data and the line, and we are summing the square of these values.

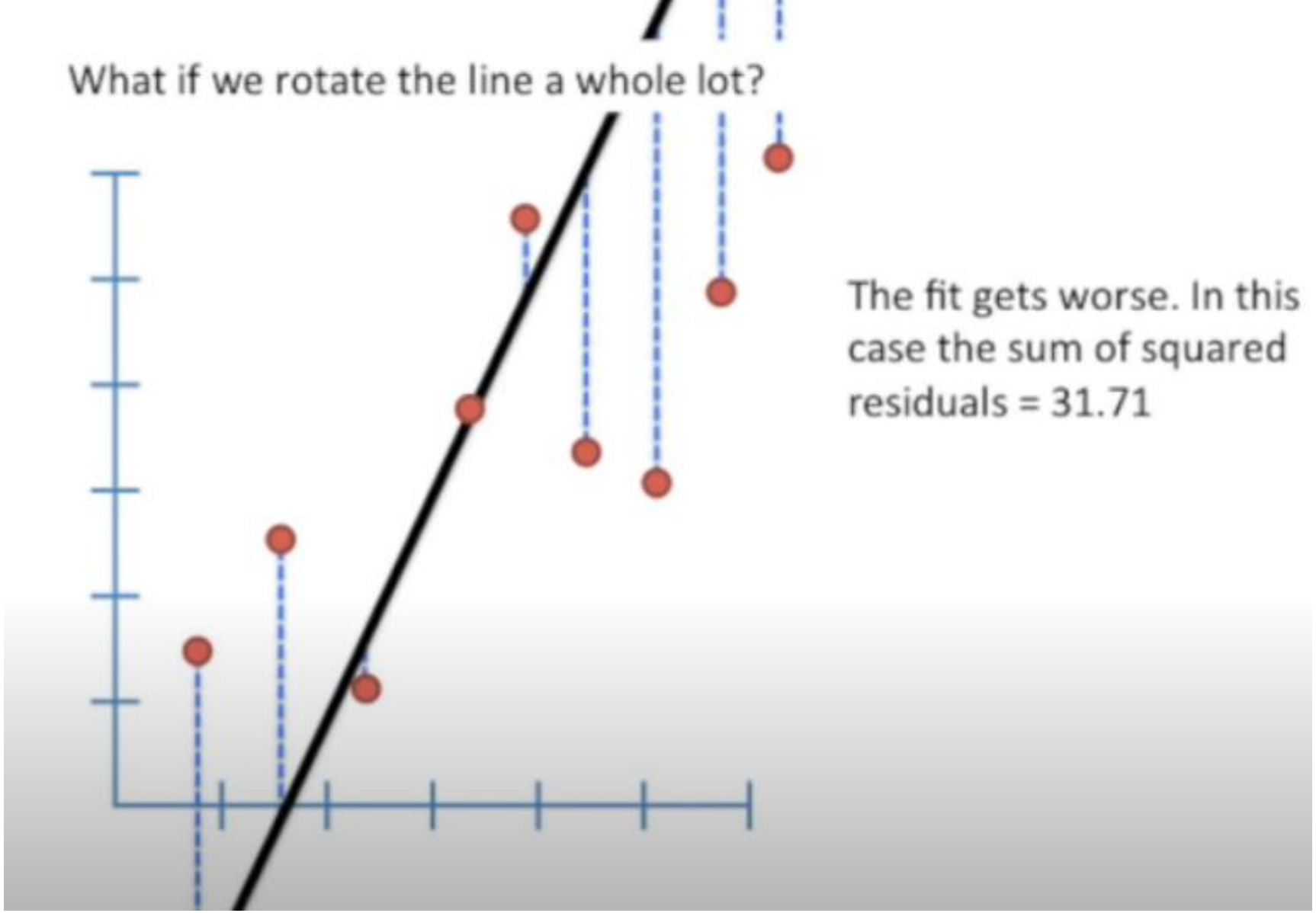


We will be repeating the same process but with this time the line is shifted a bit.

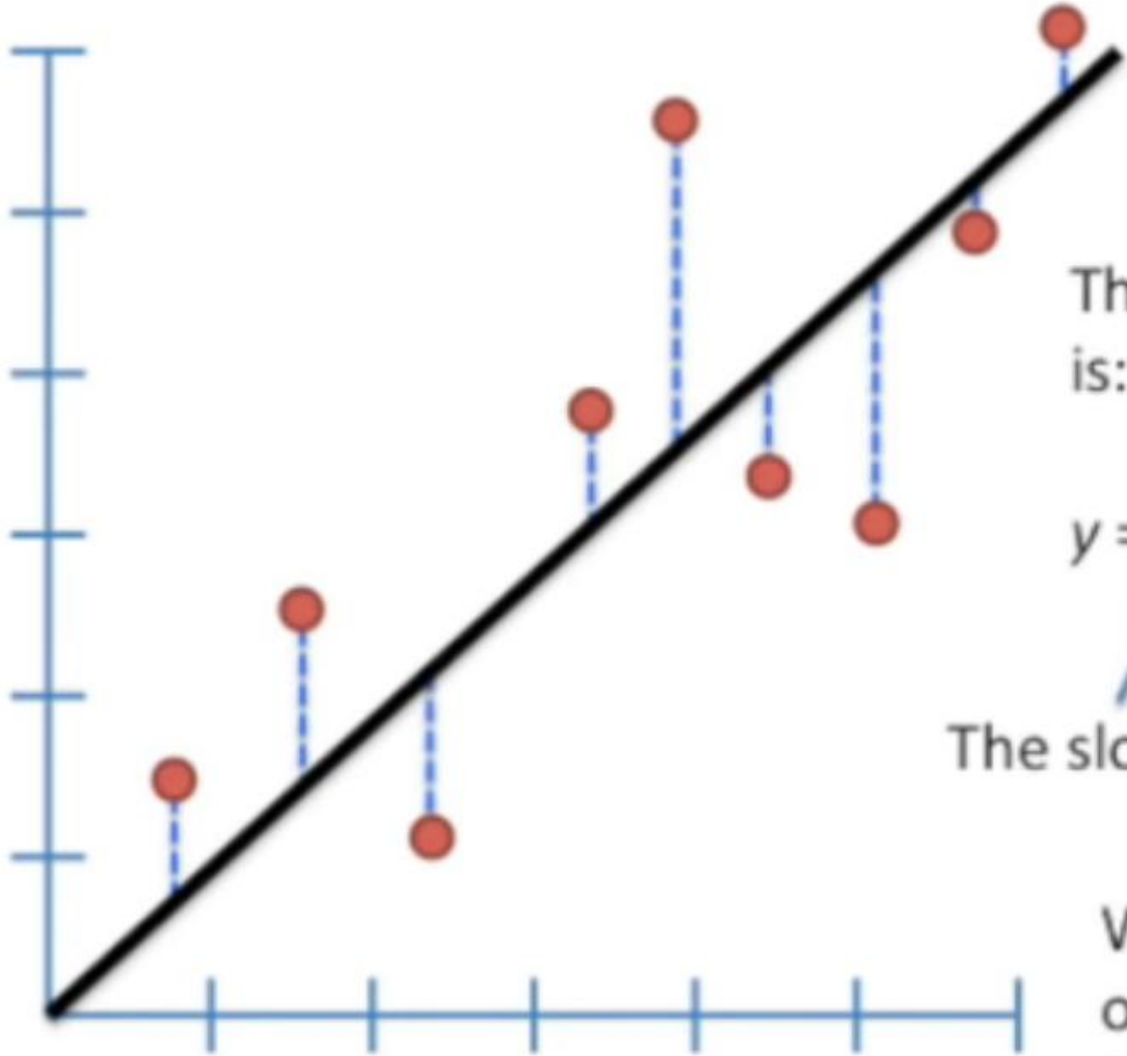


We will be shifting the line till we get a point the Sum of Squared residual is started to increase!

What if we rotate the line a whole lot?



The fit gets worse. In this case the sum of squared residuals = 31.71



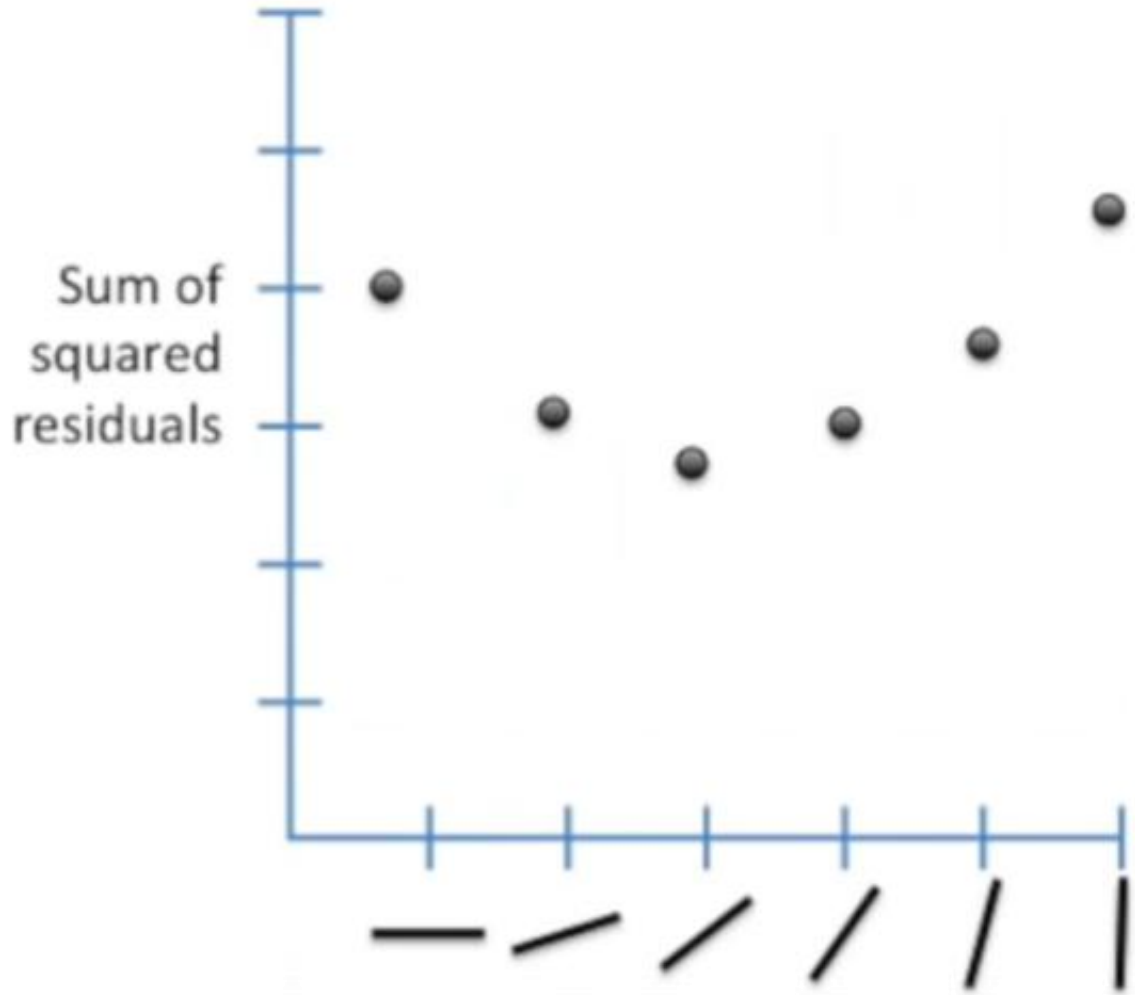
The generic line equation is:

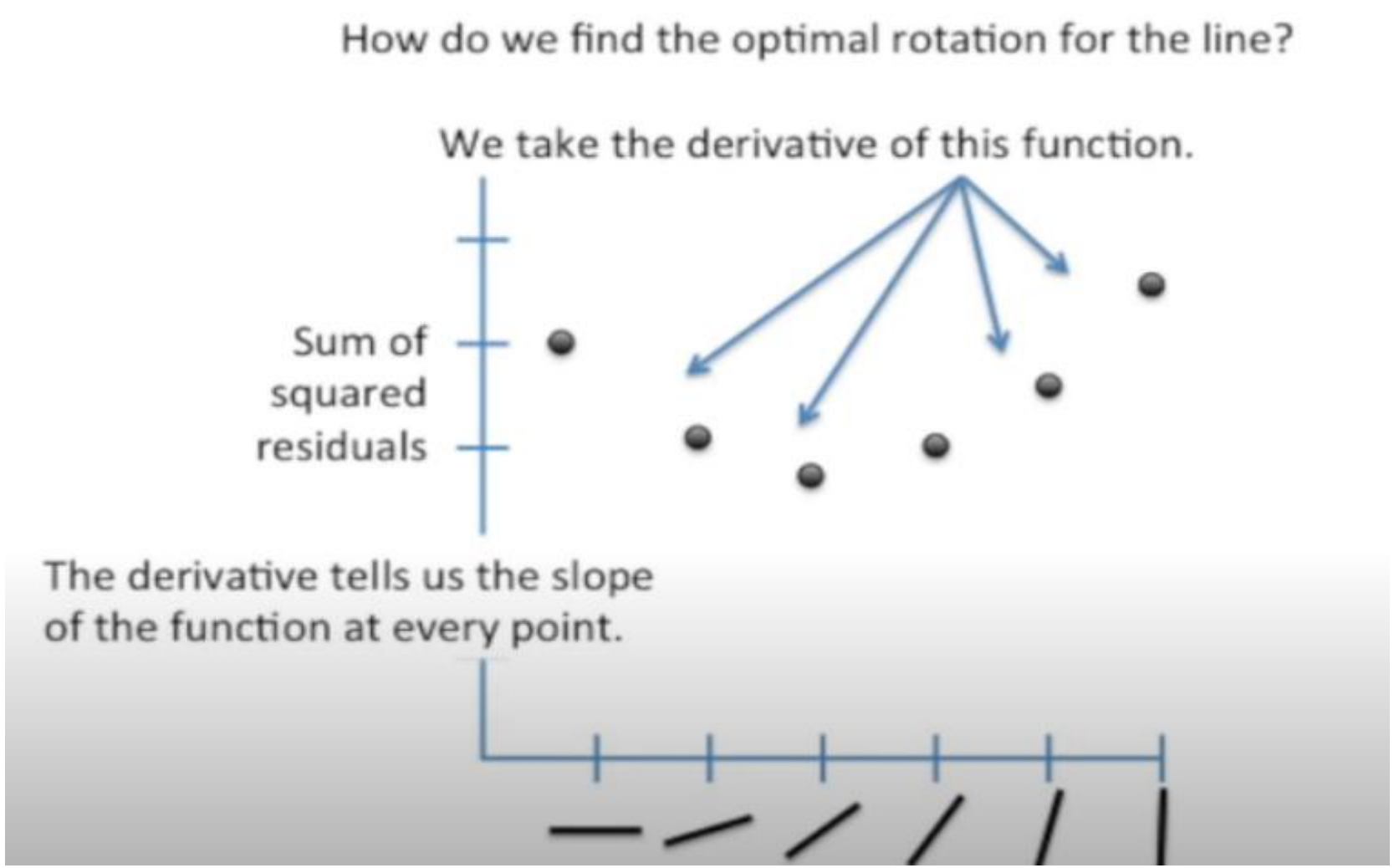
$y = a^*x + b$

The slope… …the intercept

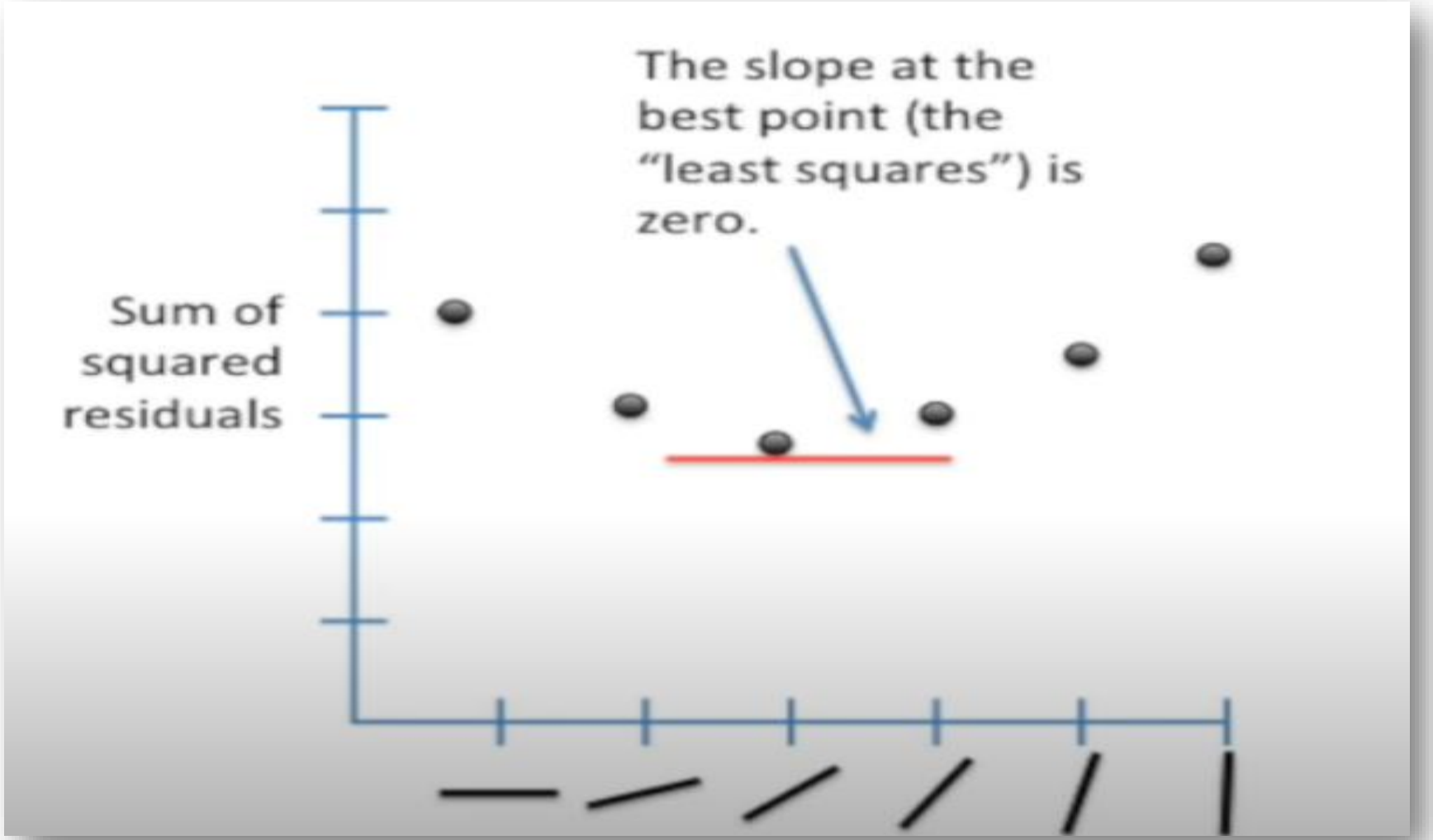
We want to find the optimal values for "*a*" and "*b*" so that we minimize the sum of squared residuals.

If we plotted the sum of squared residuals vs. each rotation, we'd get something like this…





The Derivative value is noting but it is the values that have extracted from Sum of Squared Residual.


The slope at the best point (the "least squares") is zero.
Sum of squared residuals