Data Science

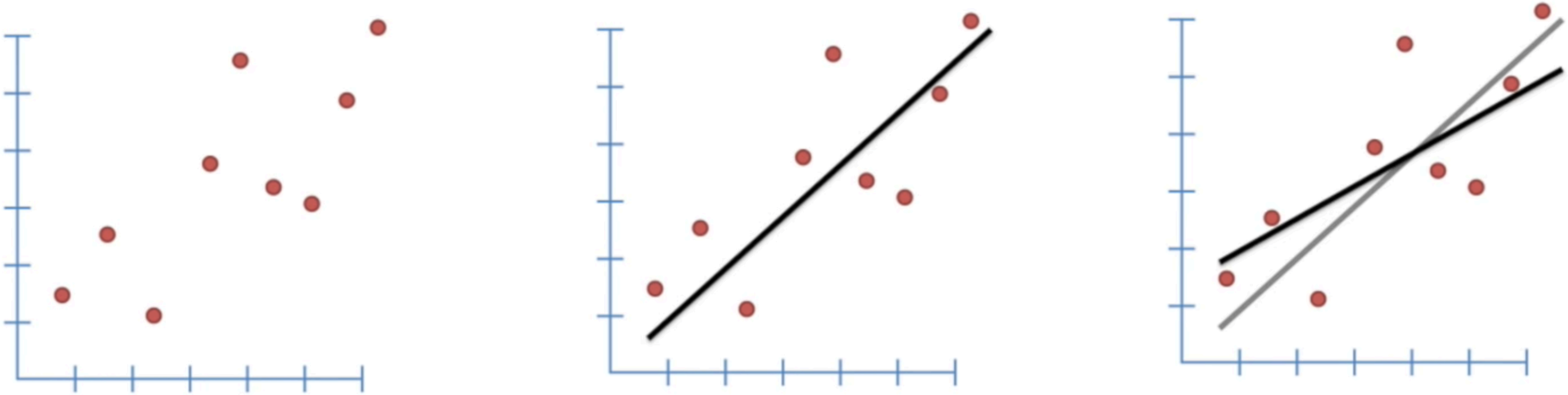
Intro to Linear Regression

10.31.22

Linear Regression

Fitting a line to data

Suppose you worked an experiment Retrieved data and drew a scatter plot. We usually like to add a line to our scatter plot to see what the ‘trend’ is. But we don’t often know if which trend line is the best fit for our data.



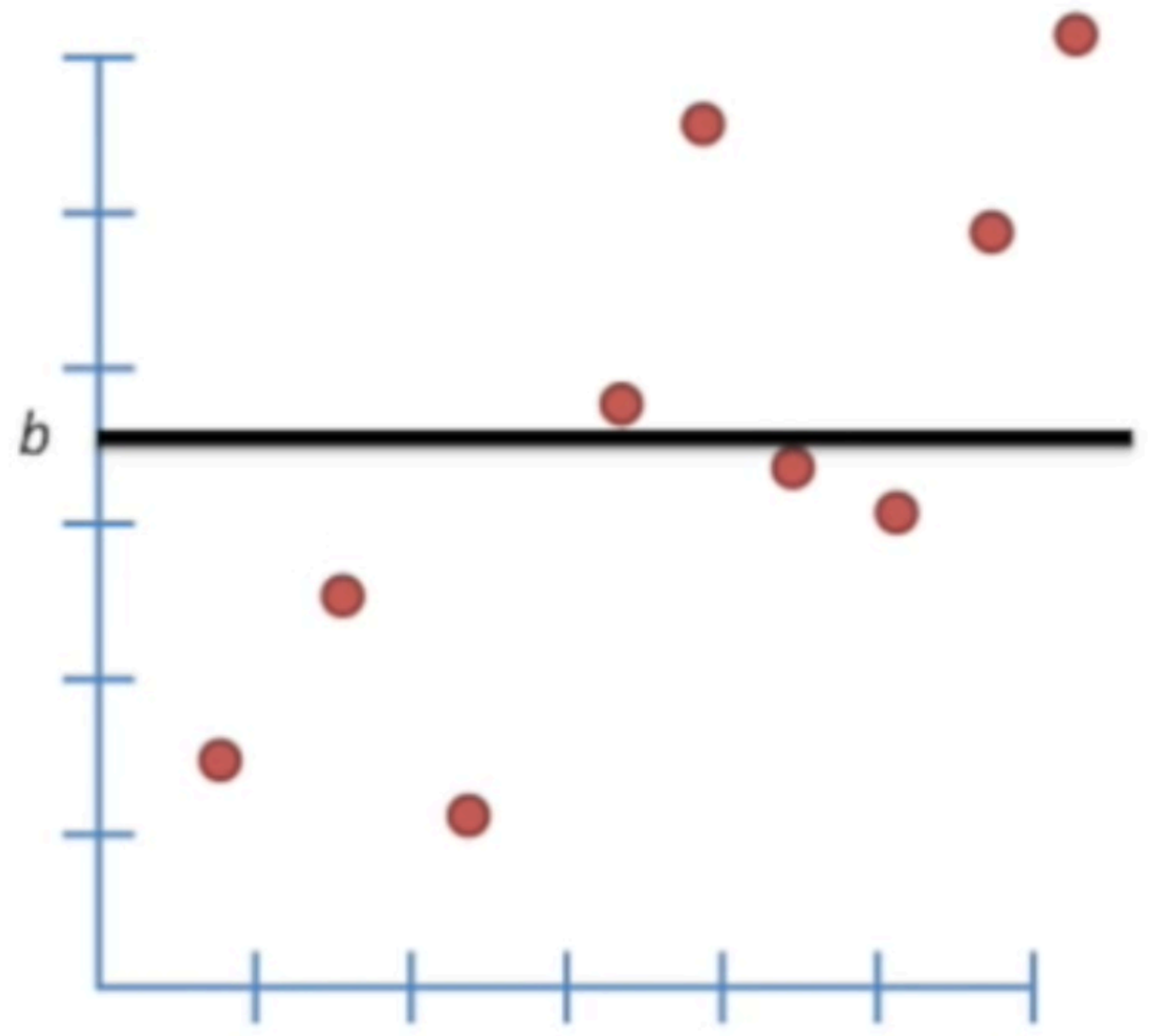
Linear regression

Fitting a line to data

Let’s start by talking about the horizontal line.

This line cuts our data through the average y-value which we will denote as b.

We can measure how well this line fits the data by measuring how close the line is to the data points.



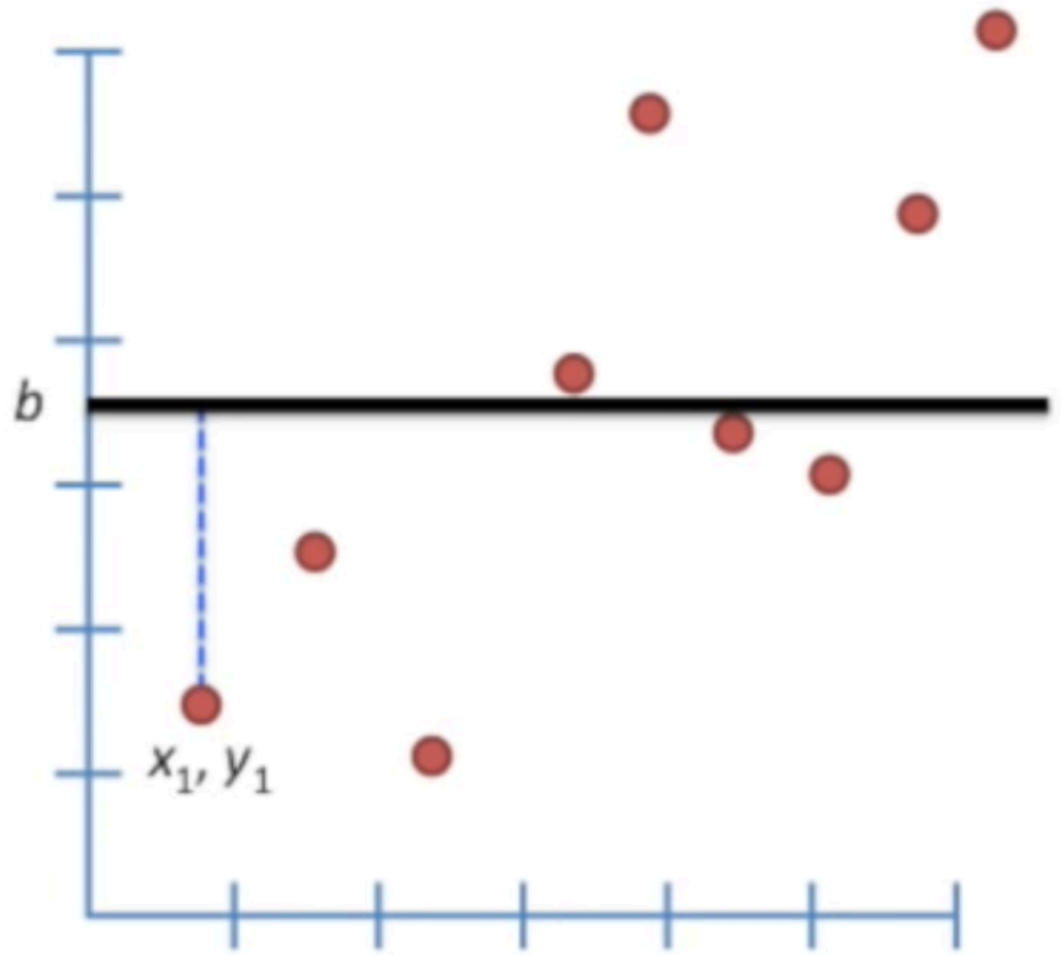
Linear Regression

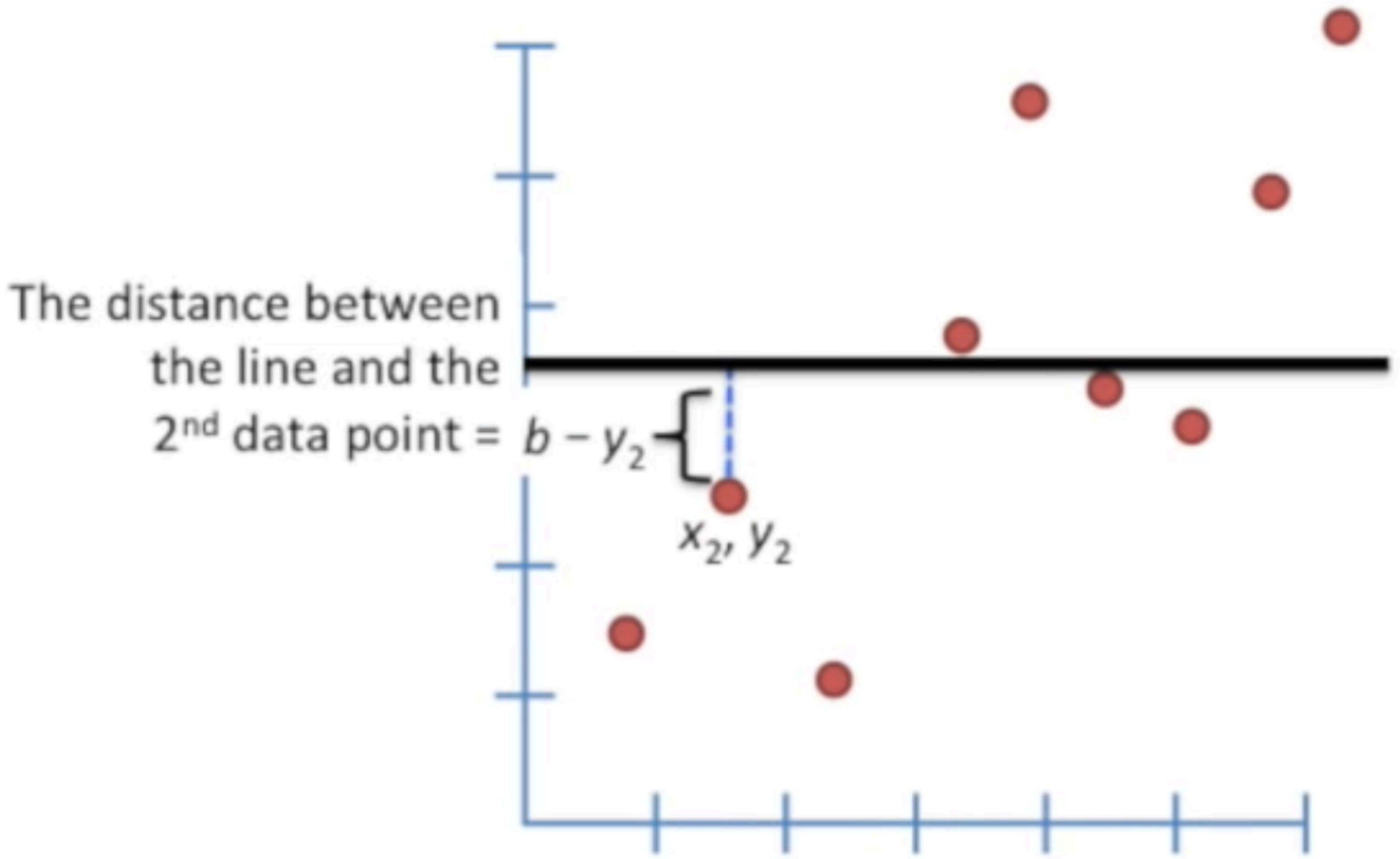
Fitting a line to data

Let’s begin by taking the left most data point on my data set, (x_1,y_1).

The y-distance between the line at y=b and (x_1,y_1) is $b-y_1$.

Similarly, the y-distance of the next point (x_2,y_2) is $b-y_2$.





The distance between the line and the 2nd data point = $b - y_2$
x_2, y_2

Linear Regression

Fitting a line to data

Continuing this way we get the total distance between the line at y=b and the data set is the sum of the distances between the data points and the line at y=b. Given by;

$(b - y_1) + (b - y_2) + \ldots + (b - y_n)$, where n is the number of data points in your given data set.

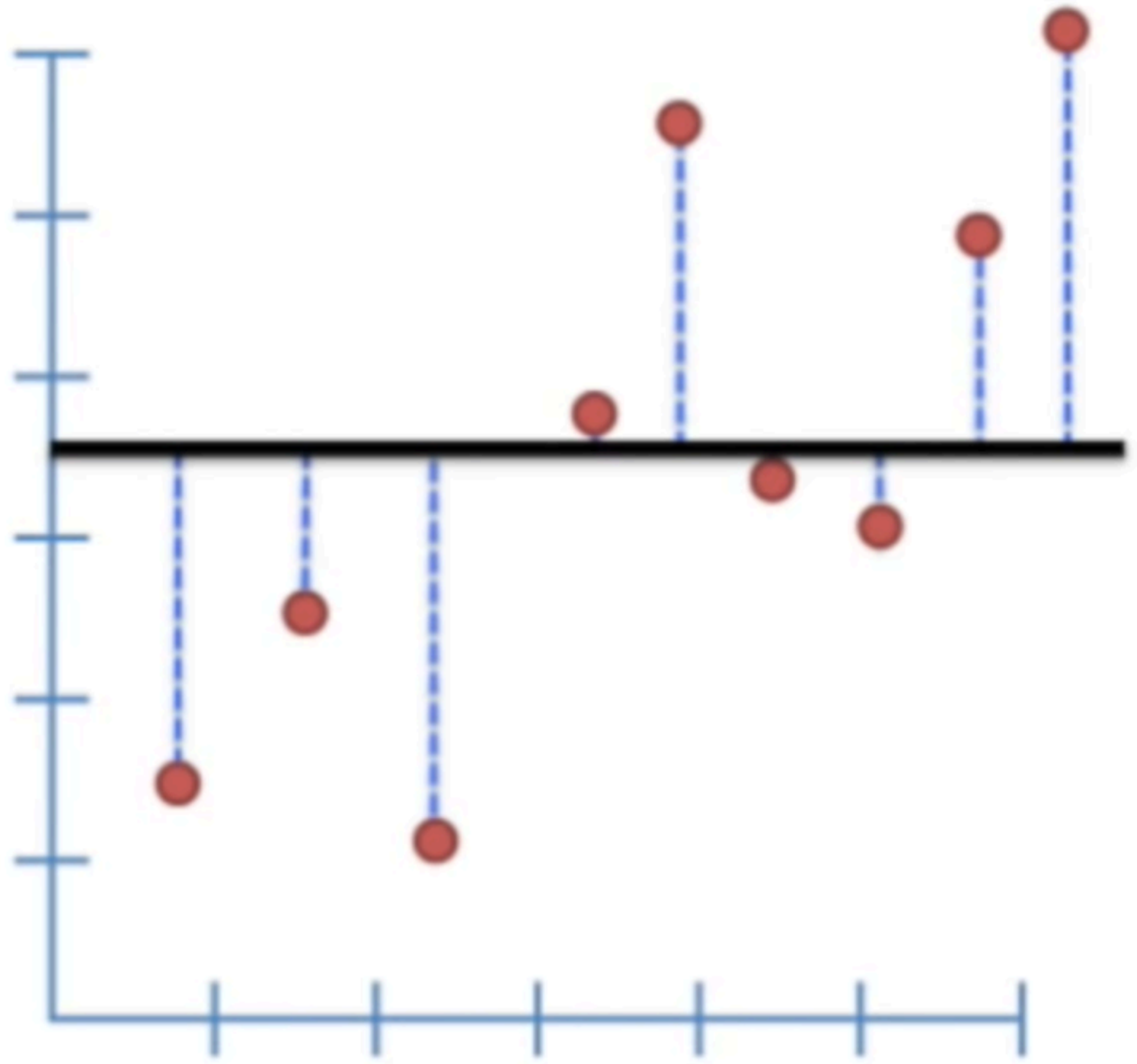
Note: that we have data points above the line at y=b. This means that for those points $y_i > b$ which means $b - y_i < 0$. That’s not good because it will subtract from the total and make the overall fit appear better than it really is.

Linear Regression

Fitting a line to data

To avoid such an issue, one can try to take the absolute value or they can square to make each value non-negative. Hence we arrive at the following:

$$S = (b - y_1)^2 + (b - y_2)^2 + \dots + (b - y_n)^2$$



This sum S is 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.

S will be the measure of how well our line fits the data.

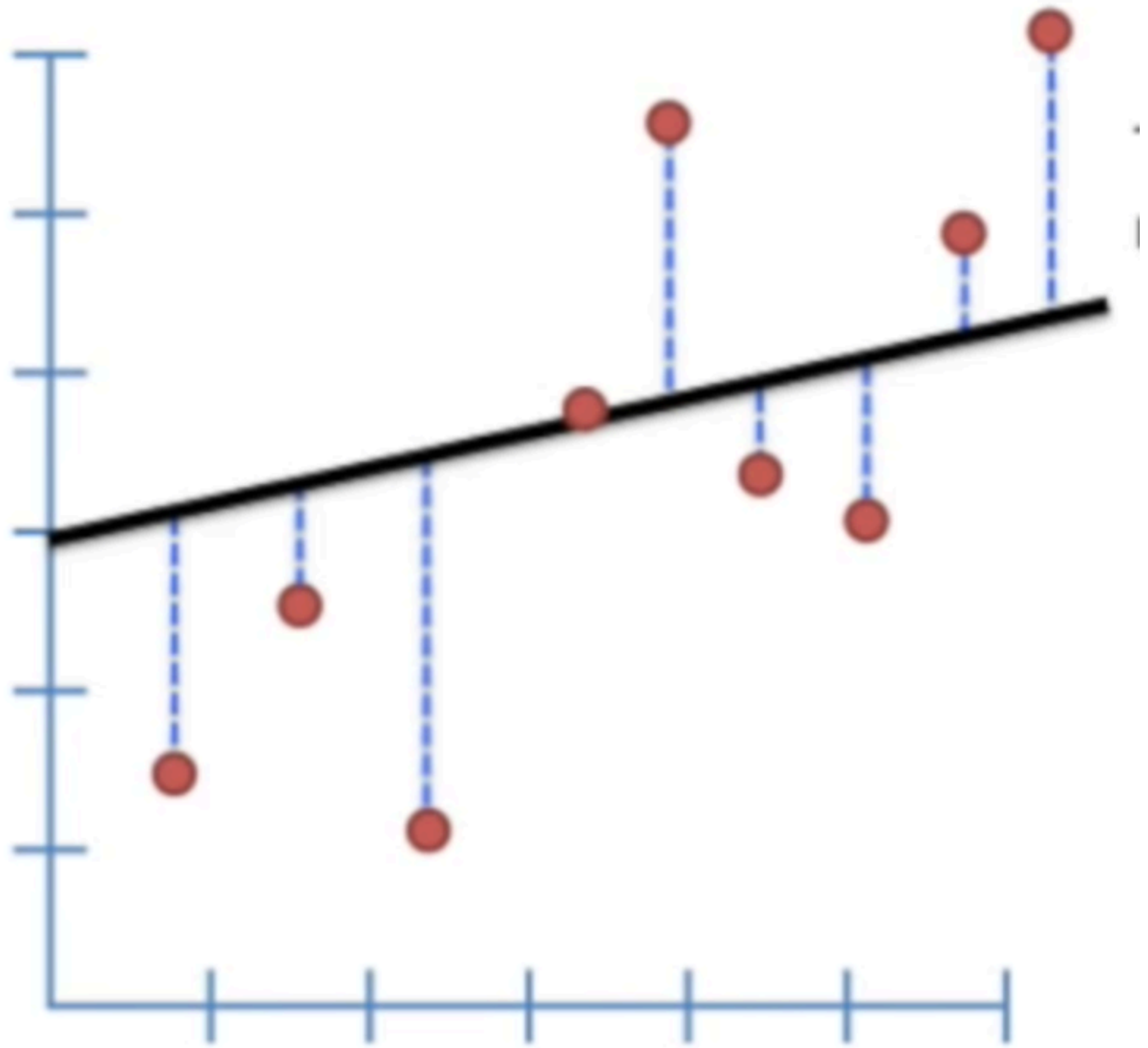
Linear Regression

Fitting a line to data

We expect S (the sum of squared residuals) to get smaller as our line becomes a better fit with the data points.

As we rotate the line, S decreases and as we continue to rotate the line further S will increases again.

This means that there is a good spot between our horizontal line at y=b and a line close to being a vertical line.



Linear Regression

Fitting a line to data

To find that best fitted line, we can start with a generic line equation given by $y = mx + b$, m is called the slope of the line and b is the y-intercept of the line.

We want to find the optimal m and b so that we can minimize S, the sum of squared residuals.

Given a line $y = mx + b$, notice that S is calculated as follows:

$S = ((a*x_1 + b) - y_1)^2 + ((a*x_2 + b) - y_2)^2 + \ldots + ((a*x_n + b) - y_n)^2$

The terms $a*x_i + b$ is just the value of the line at x_i and y_i is the observed value at x_i. Hence, all we are doing is calculating the distance between the line $y=mx+b$ and the observed value.

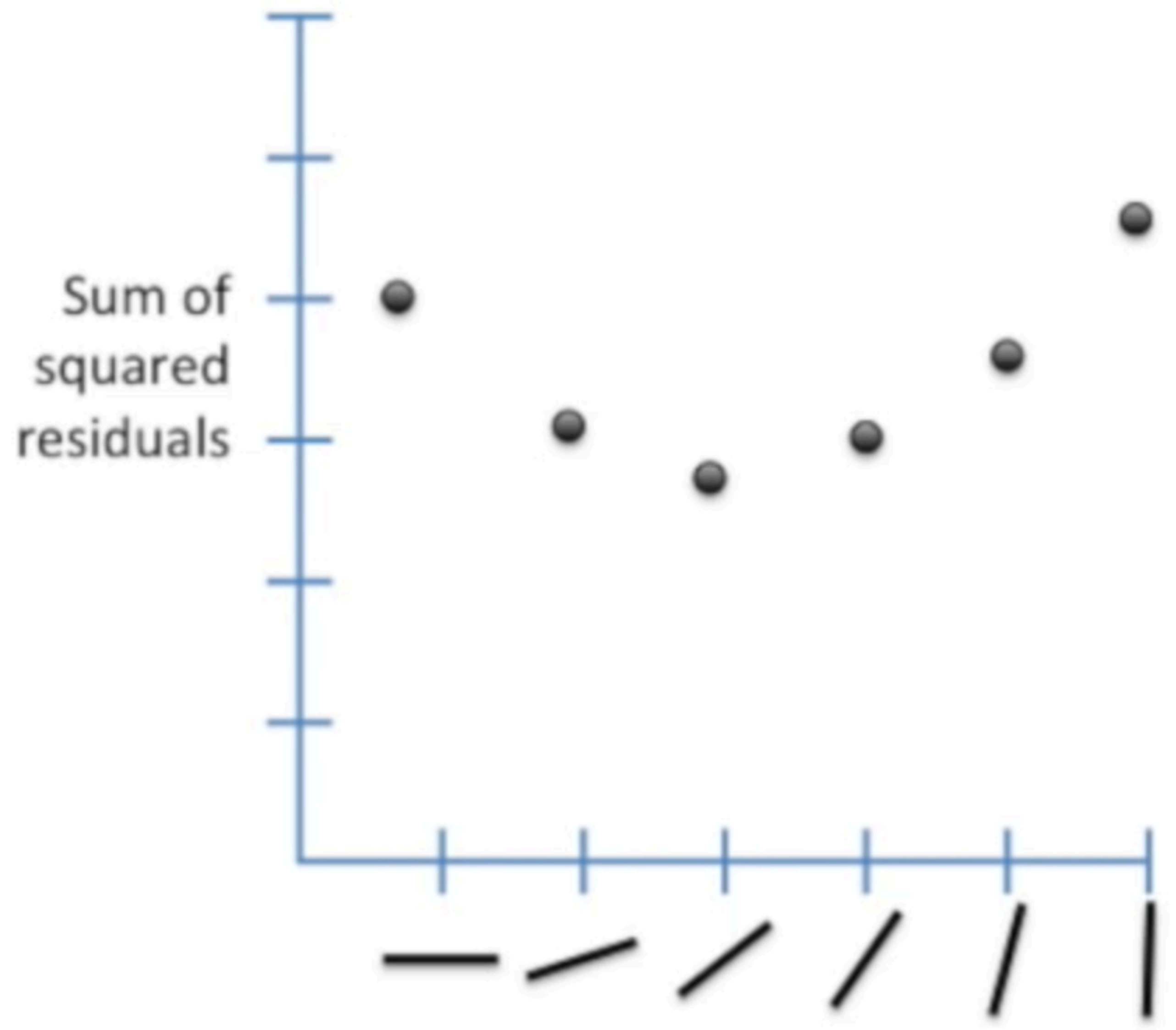
Linear Regression

Fitting a line to data

Since we want the line that will give us the smallest sum of squares, this method for finding the best values for ‘m’ and ‘b’ is called ‘least squares’. Hence, often times people refer to linear regression as least squares.

If we plotted the sum of squared residuals for each rotation we would get a graph like to your right.

How would we find the optimal rotation for the best fitted line?



Linear Regression

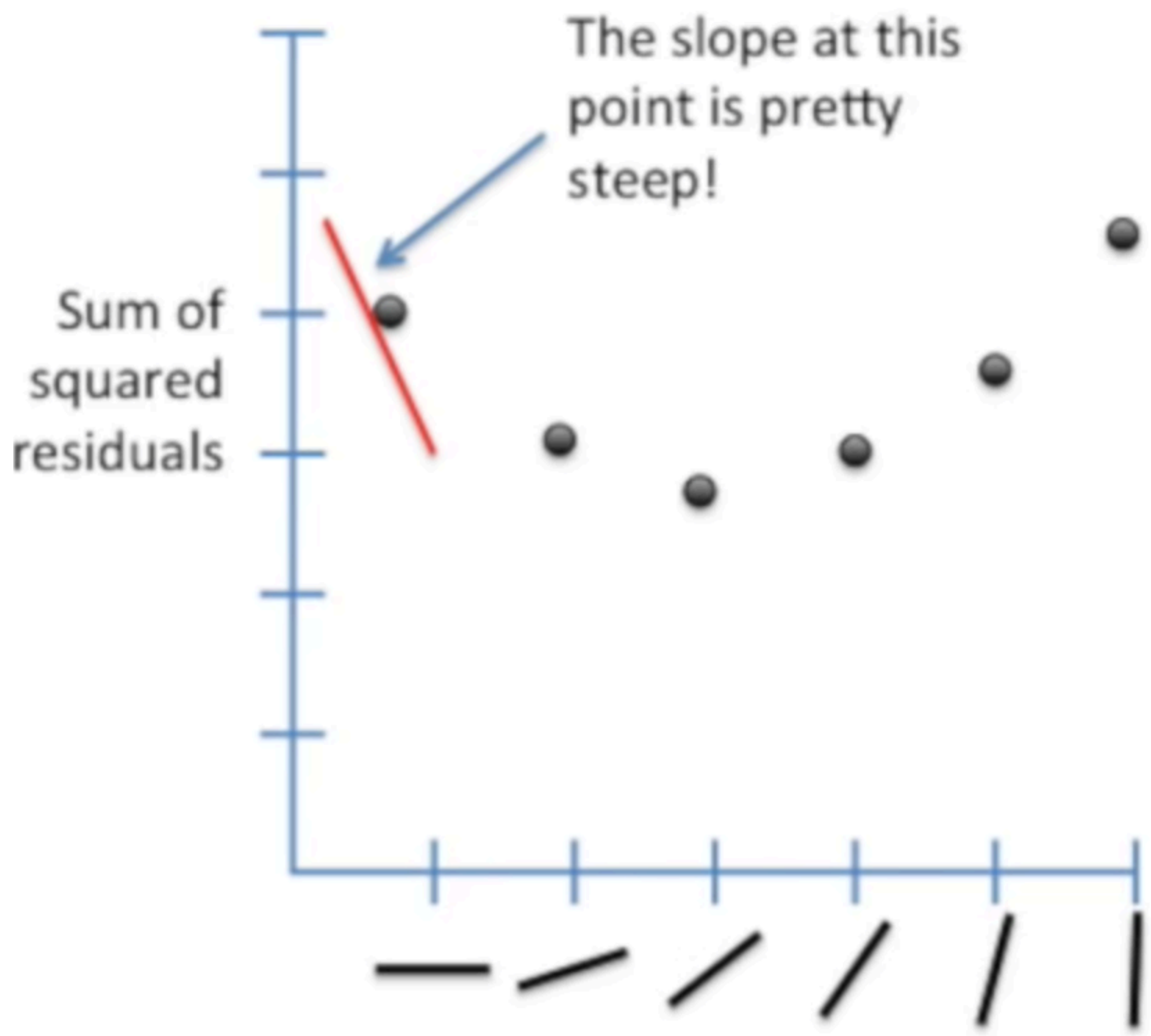
Fitting a line to data

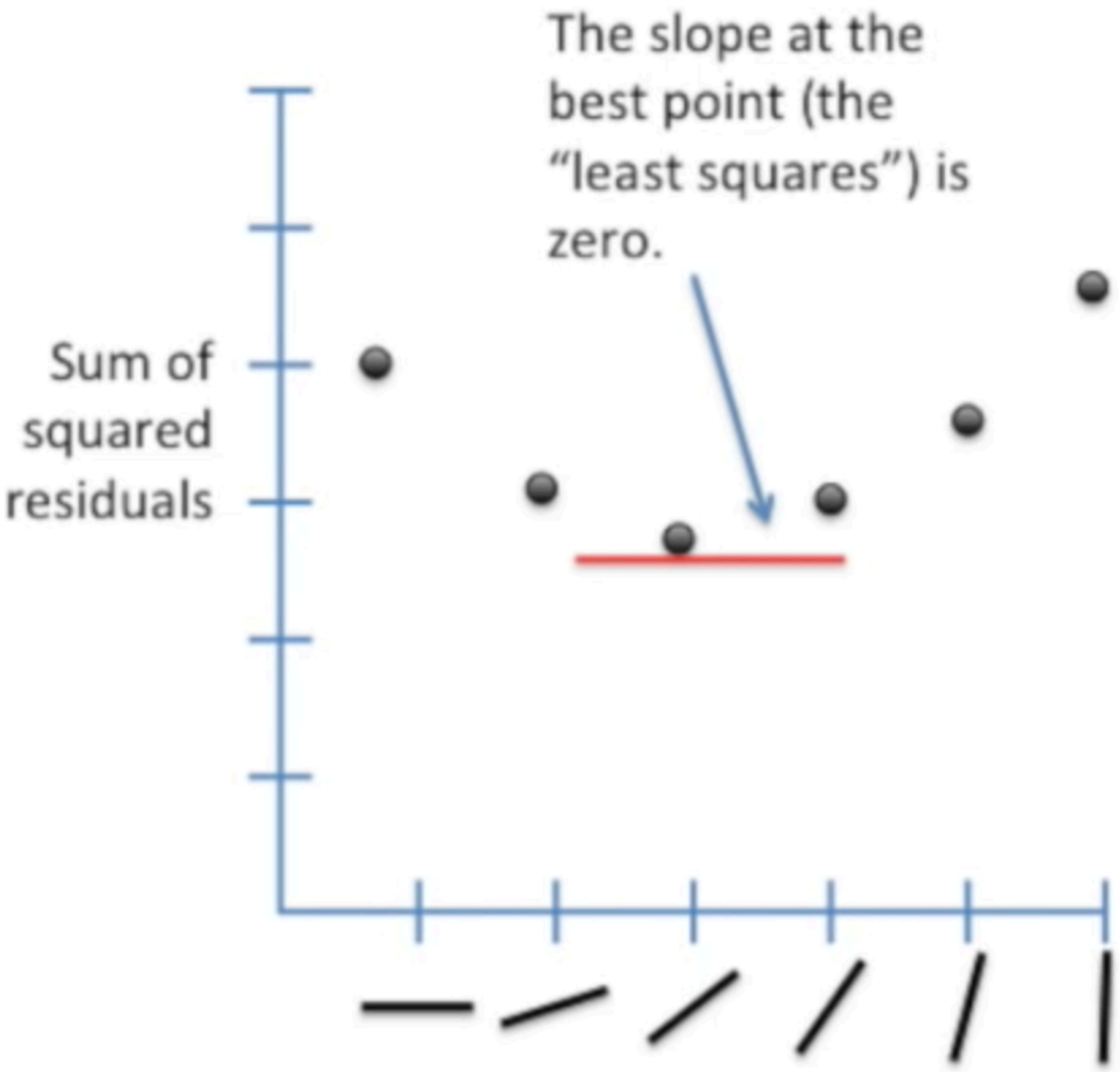
Well, we are looking for the minimum of S i.e. the minimum point on the graph shown in the previous slide. For those of you who took calculus, you would recall how to find the min and max given a function. You take the derivative of the function and set it equal to zero.

For those of you who didn’t take calculus, the derivative of a function tells you the slope of the tangent line to the function at every point. Hence in our example, the derivative should tell you the slope of the line at each point (see next slide).

Linear Regression

Fitting a line to data





Linear Regression

Fitting a line to data

Now, remember that the slope of the line gives you different rotation of the line. This means that for each x-value in our graph (in previous slide) we just have different values of ‘m’ and ‘b’ for y = mx+b.

Taking the derivatives of the slope and the intercepts tells us where the optimal values are for the best fit.

Note: no one ever solves this problem by hand! (Whew~ thats good news!) Since this is done by the computer you do not need to know how to take these derivatives. However, it is important to understand the concepts.

Linear Regression

Wrapping up key concepts

- We want to minimize the square of the distance between the observed values and the line.
- We do this by taking the derivative and finding where it is equal to 0.
- The final line minimizes the sum of squares (it gives the “least squares”) between it and the real data.

Linear Regression

Main ideas

- Use least-squares to fit a line to a data
- Calculate R^2.
- Calculate a p-value for R^2.

Linear Regression

R-squared

R, is what we call the correlation coefficient. We calculate R in the following way:

1. Calculate $\bar{x}$, the mean of all of the first coordinates of the data x_i.
2. Calculate $\bar{y}$, the mean of all of the second coordinates of the data y_i.
3. Calculate s_x the sample standard deviation of all of the first coordinates of the data x_i.
4. Calculate s_y the sample standard deviation of all of the second coordinates of the data y_i.
5. Use the formula $(z_x)_i = (x_i – \bar{x}) / s_x$ and calculate a standardized value for each x_i.
6. Use the formula $(z_y)_i = (y_i – \bar{y}) / s_y$ and calculate a standardized value for each y_i.
7. Multiply corresponding standardized values: $(z_x)_i (z_y)_i$
8. Add the products from the last step together.
9. Divide the sum from the previous step by *n* – 1, where *n* is the total number of points in our set of paired data. The result of all of this is the correlation coefficient *r*.

Linear Regression

R-squared

To see exactly how the value of *r* is obtained we look at an example. Again, it is important to note that for practical applications we would want to use our calculator or statistical software to calculate *r* for us.

We begin with a listing of paired data: (1, 1), (2, 3), (4, 5), (5,7). The mean of the *x* values, the mean of 1, 2, 4, and 5 is $\bar{x} = 3$. We also have that $\bar{y} = 4$. The standard deviation of the x values is $s_x = 1.83$ and $s_y = 2.58$.

The table in the next slide summarizes the other calculations needed for *r*. The sum of the products in the rightmost column is 2.969848. Since there are a total of four points and 4 – 1 = 3, we divide the sum of the products by 3. This gives us a correlation coefficient of $r = 2.969848/3 = 0.989949$.

Table for Example of Calculation of Correlation Coefficient

x	y	z_x	z_y	$z_x z_y$
1	1	-1.09544503	-1.161894958	1.272792057
2	3	-0.547722515	-0.387298319	0.212132009
4	5	0.547722515	0.387298319	0.212132009
5	7	1.09544503	1.161894958	1.272792057

Linear Regression

R-squared

Our correlation coefficient R, tells us correlation between x-variable and y-variable. For example suppose we have a sample of mice and we are comparing their weight and size. As expected the bigger our mouse, the more it weighs. There is a heavy correlation.

A positive correlation is one who's R = 1 or close to 1 and a negative correlation is one who's R = -1 or close to -1. When R = 0, that means we have no correlation.

So why do we care about R^2?

Linear Regression

R-squared

Well, R^2 is very similar to R but the interpretation is much easier.

It’s not very obvious that for $R = 0.7$ is twice as good a correlation as $R = 0.5$. However, $R^2 = 0.7$ is what it looks like and is 1.4 times as good as $R^2 = 0.5$.

It’s also very intuitive and easy to calculate.

Linear Regression

Lets look at an example:

Going back to our first example with a sample of mice. Suppose we have a graph that shows weight of the mouse and the height of the mouse.

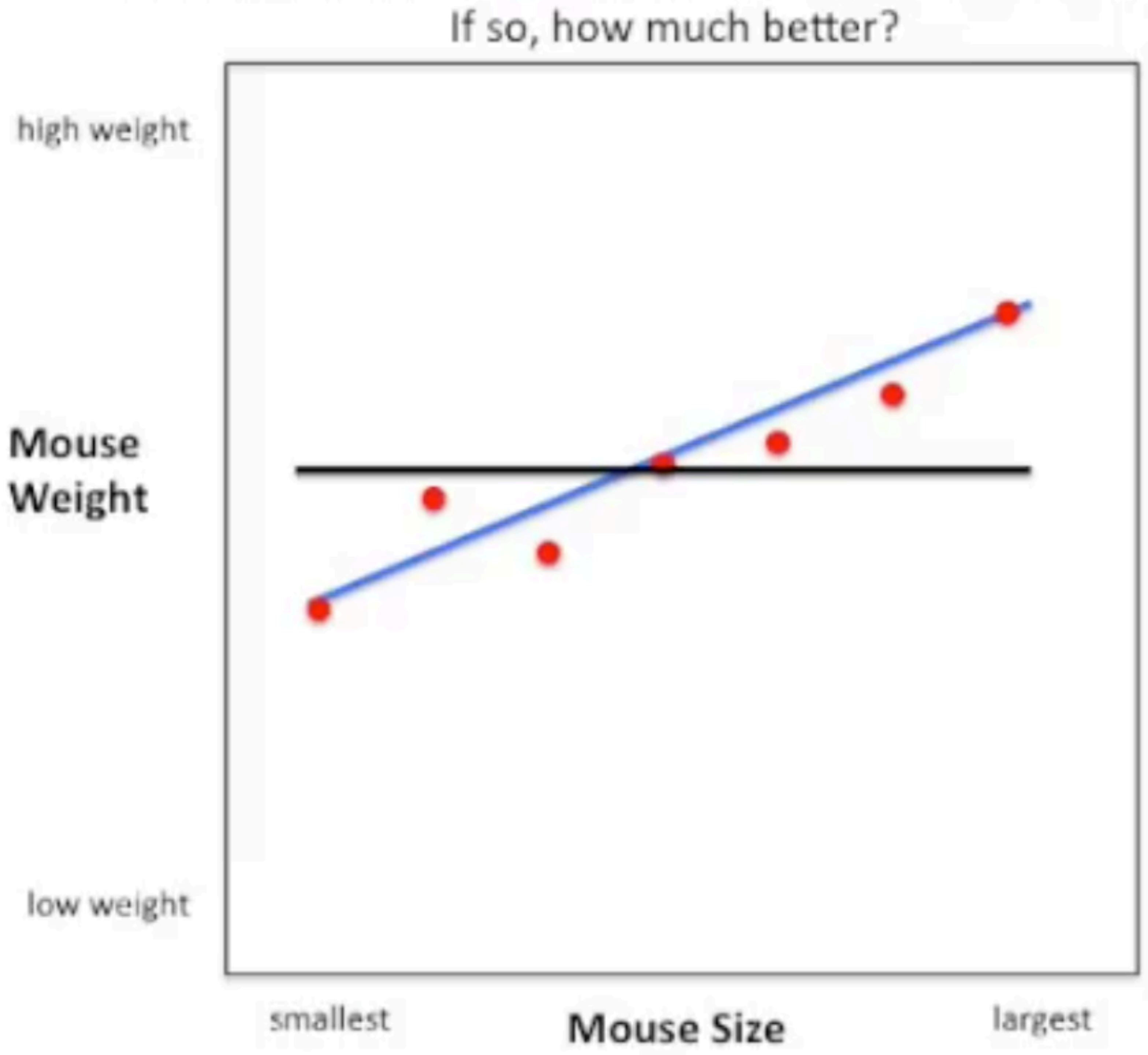
We have a line in blue that is best fitted and

a black line to show the line at y = ȳ.

it looks like the blue line is better than the black

line as far as fitting goes. We can quantify that

difference using R-squared.



Linear Regression

R-squared

R-square is given by the following formula:

$$R^2 = \mathrm{var}(\mathbf{mean}) - \mathrm{var}(\mathrm{line})/ \mathrm{var}(\mathbf{mean})$$

- var(**mean**) is just the variation around the mean. This is just the sum of the squared differences of the actual data values from the mean.

- var(line) is calculated in a similar to calculating the var(mean) it is the sum of the square differences of the actual data values from the new blue line.

This will in turn make $R^2 < 1$. This will give you a percentage value when multiplied with 100.