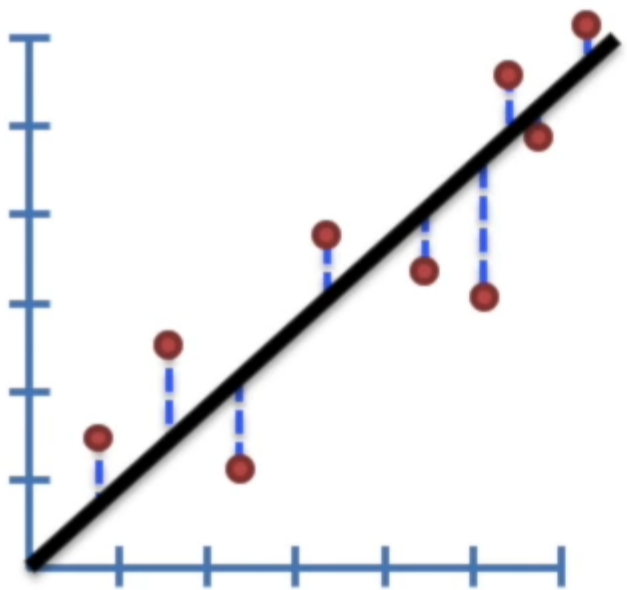
Linear regression

Supervised learning

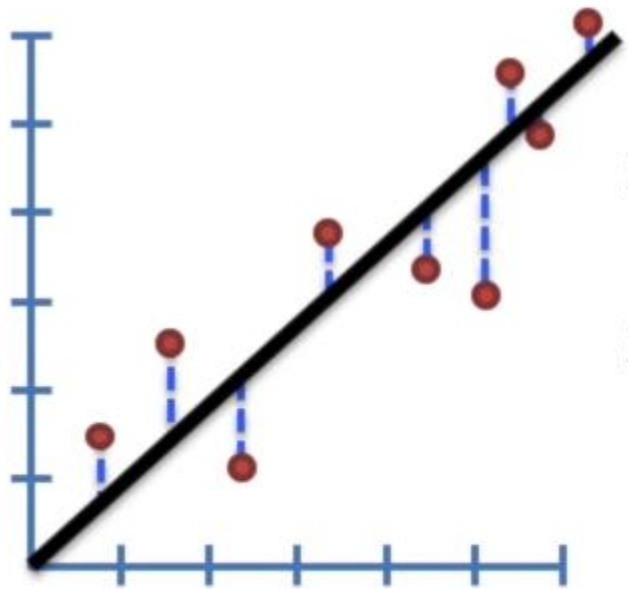
The Main Ideas!

1) Use least-squares to fit a a line to the data.



The Main Ideas!

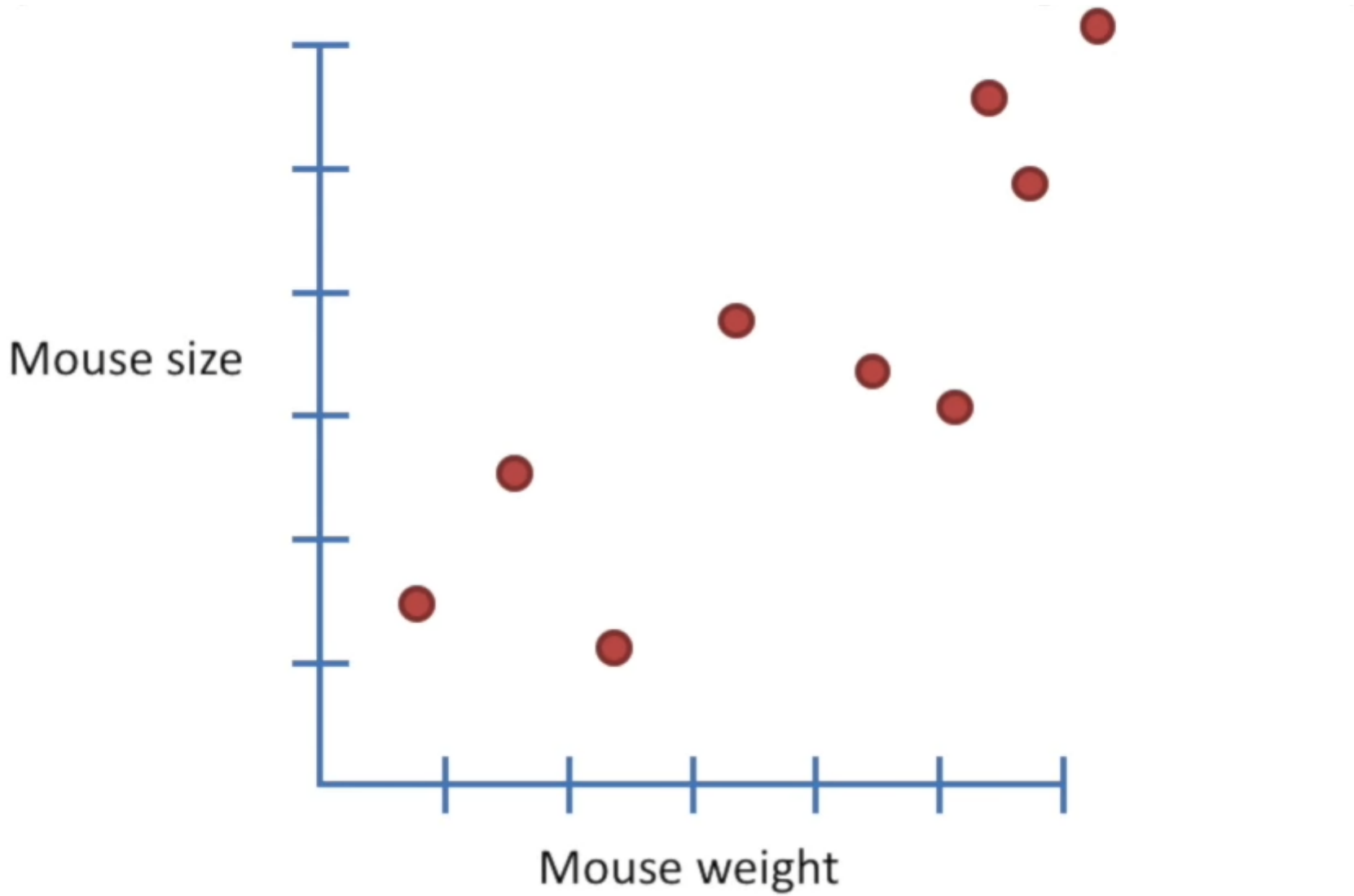
1) Use least-squares to fit a a line to the data.

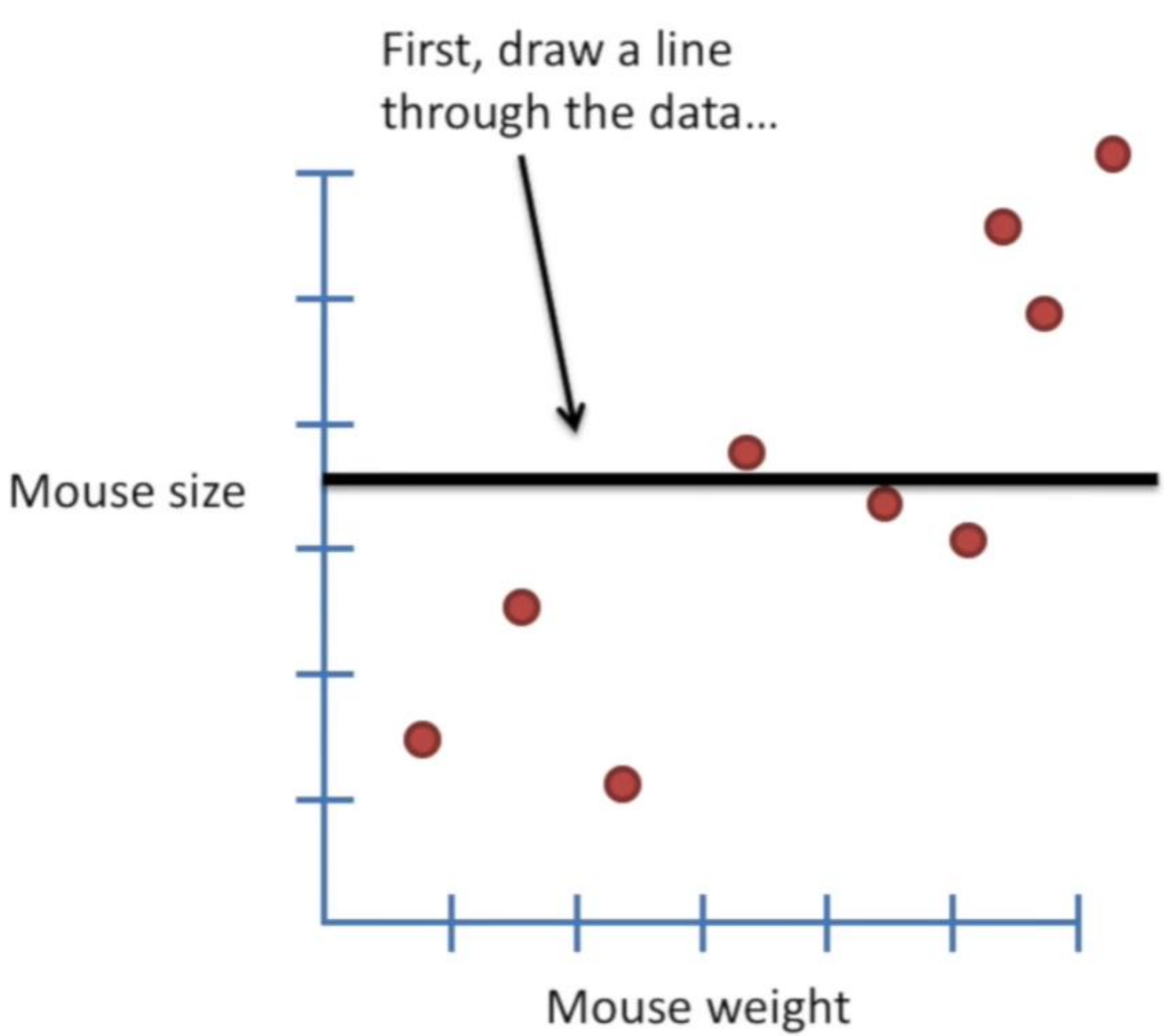


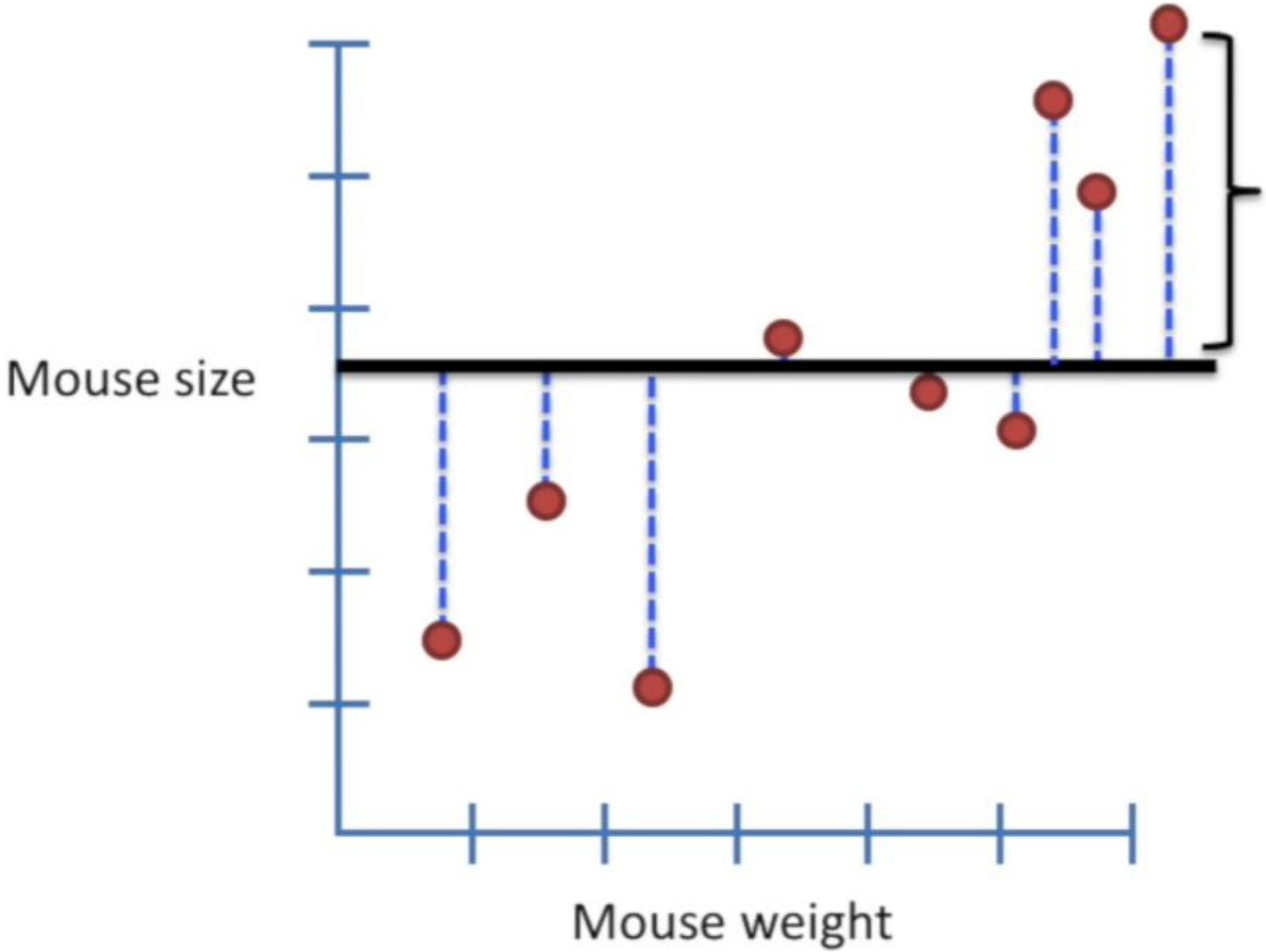
2) Calculate R^2

3) Calculate a p-value for R^2

Let’s do a quick review…

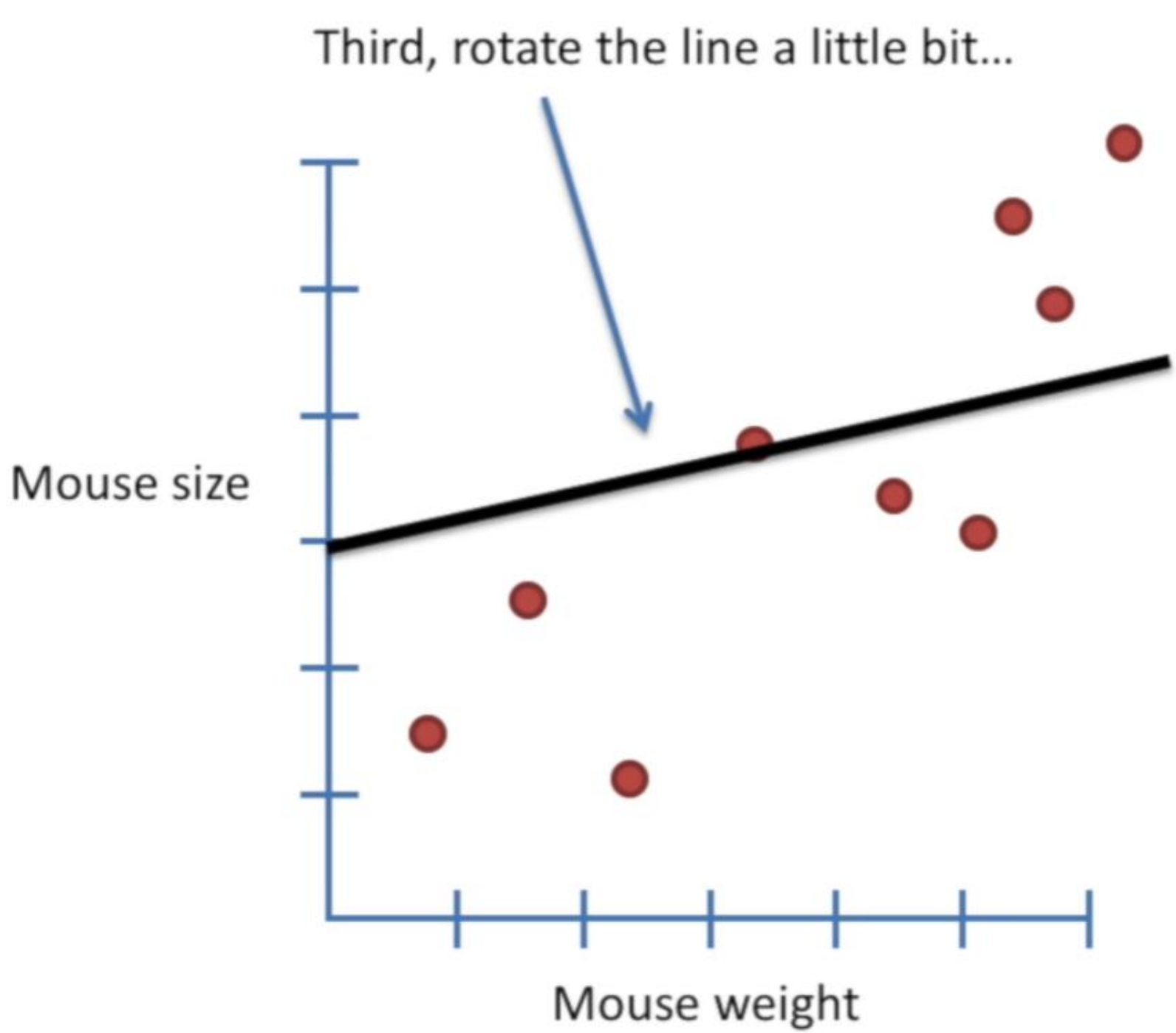


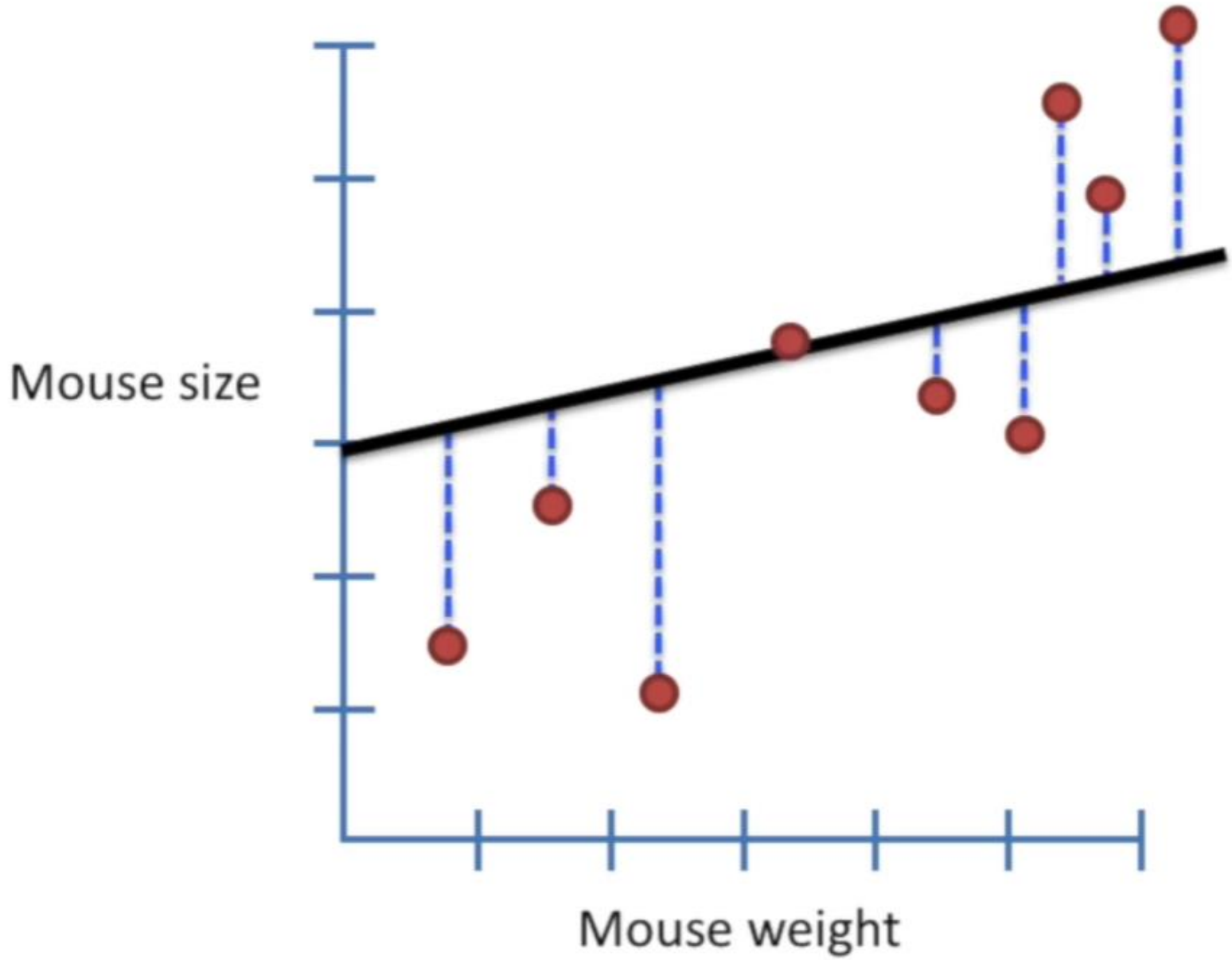

First, draw a line
through the data…
Mouse size
Mouse weight



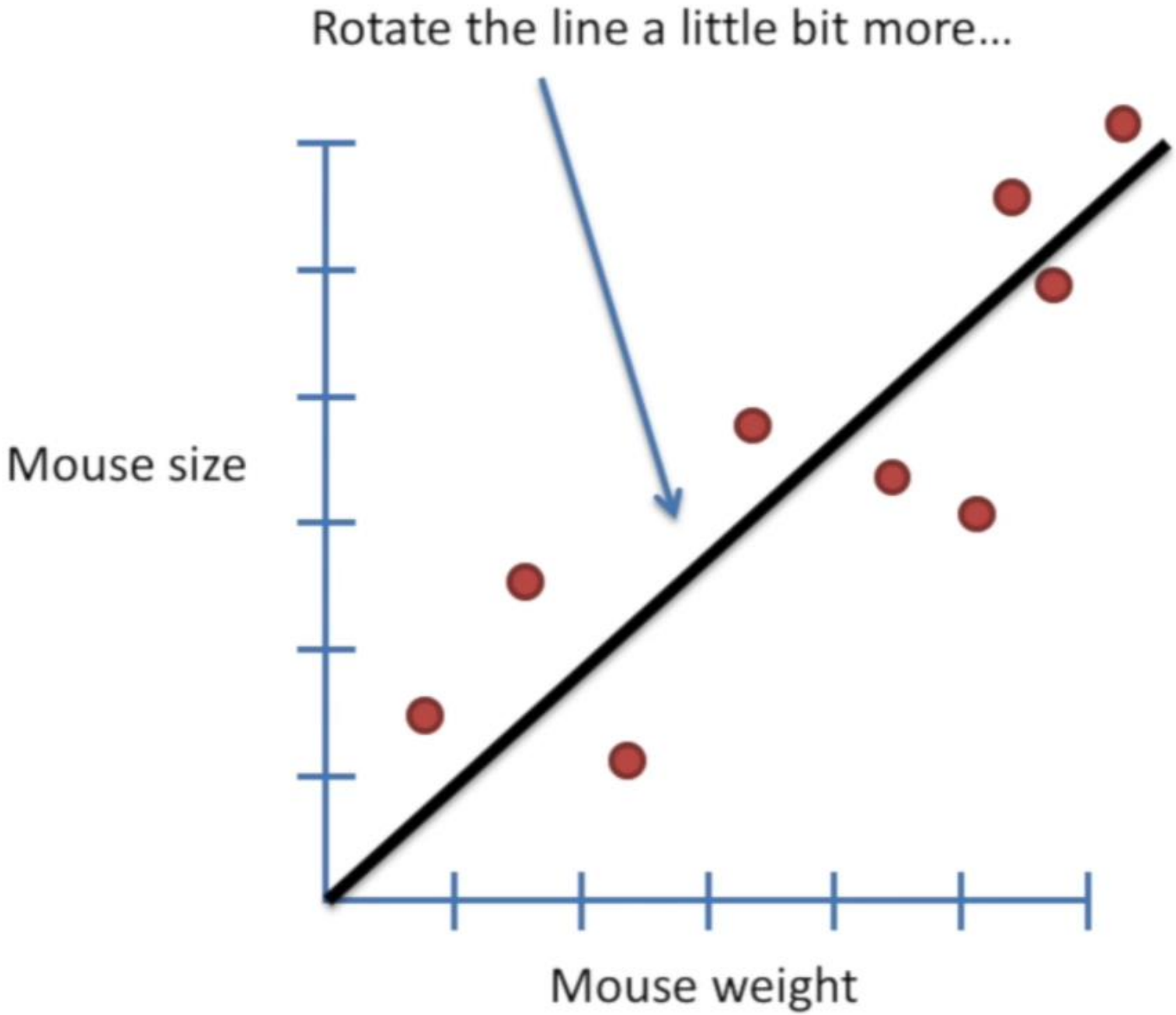
Second, measure the distance from the line to the data, square each distance, and then add them up.

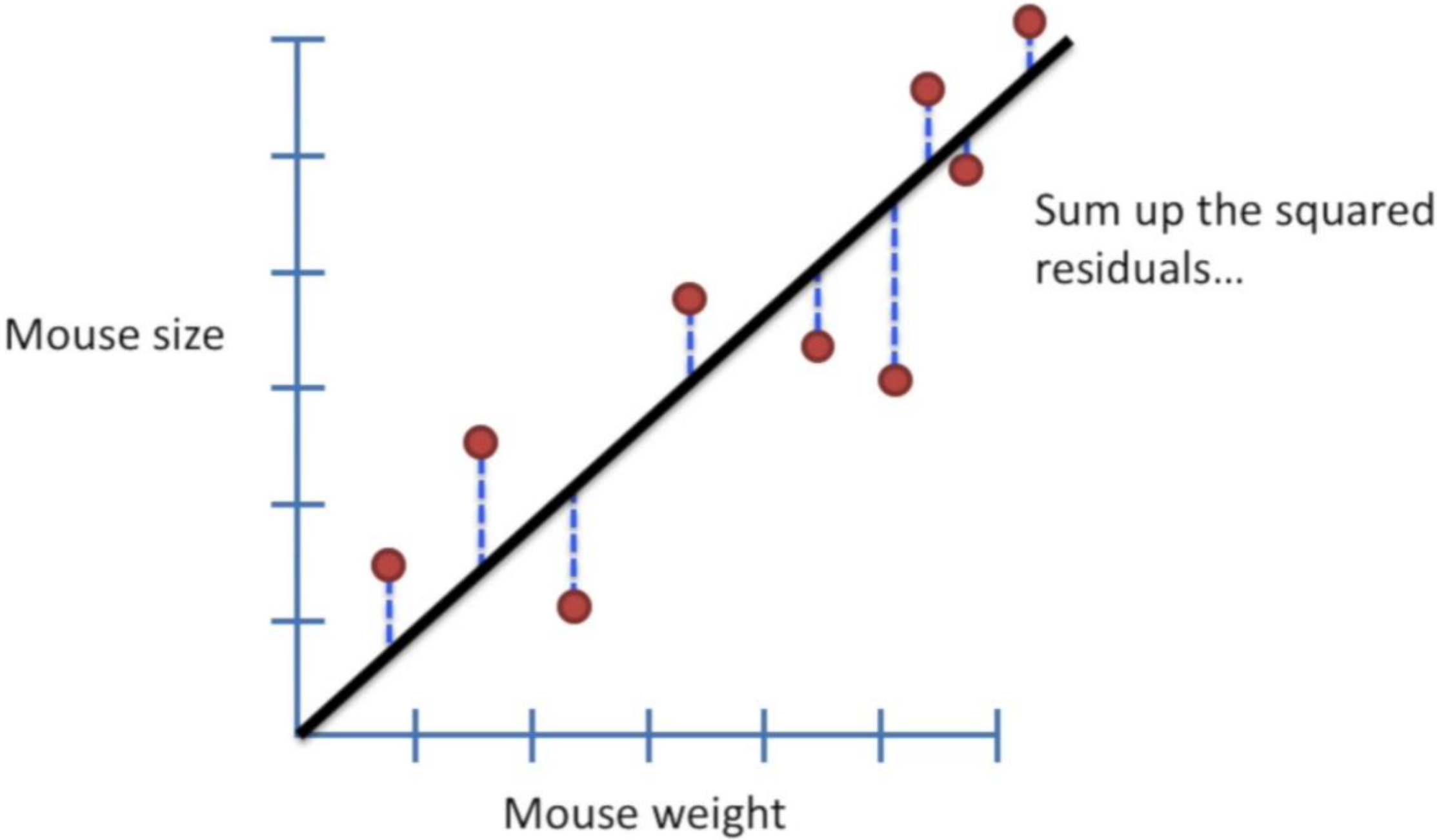
Terminology alert!
The distance from a line to a data point is called a "**residual**".

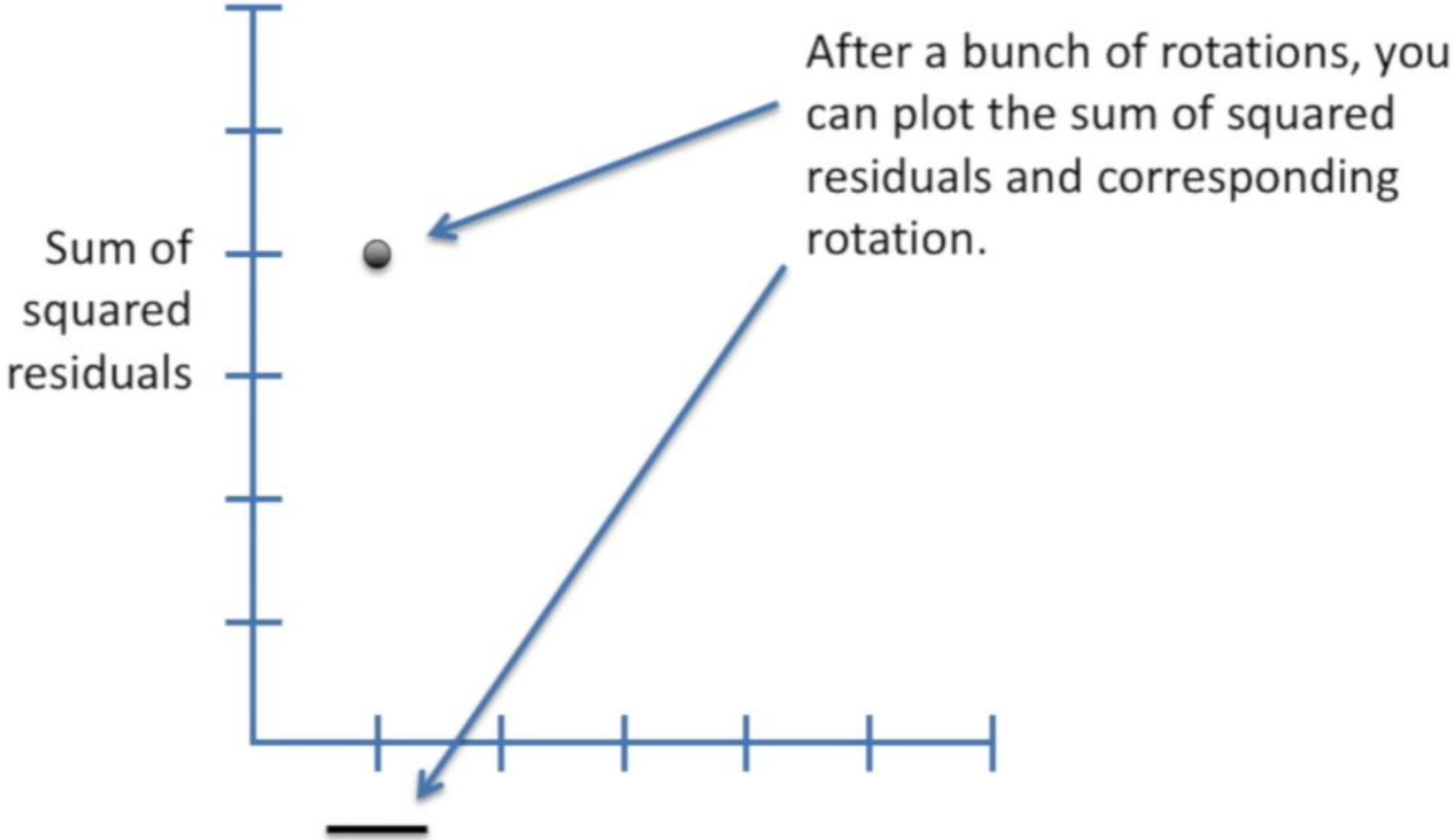

Third, rotate the line a little bit…
Mouse size
Mouse weight

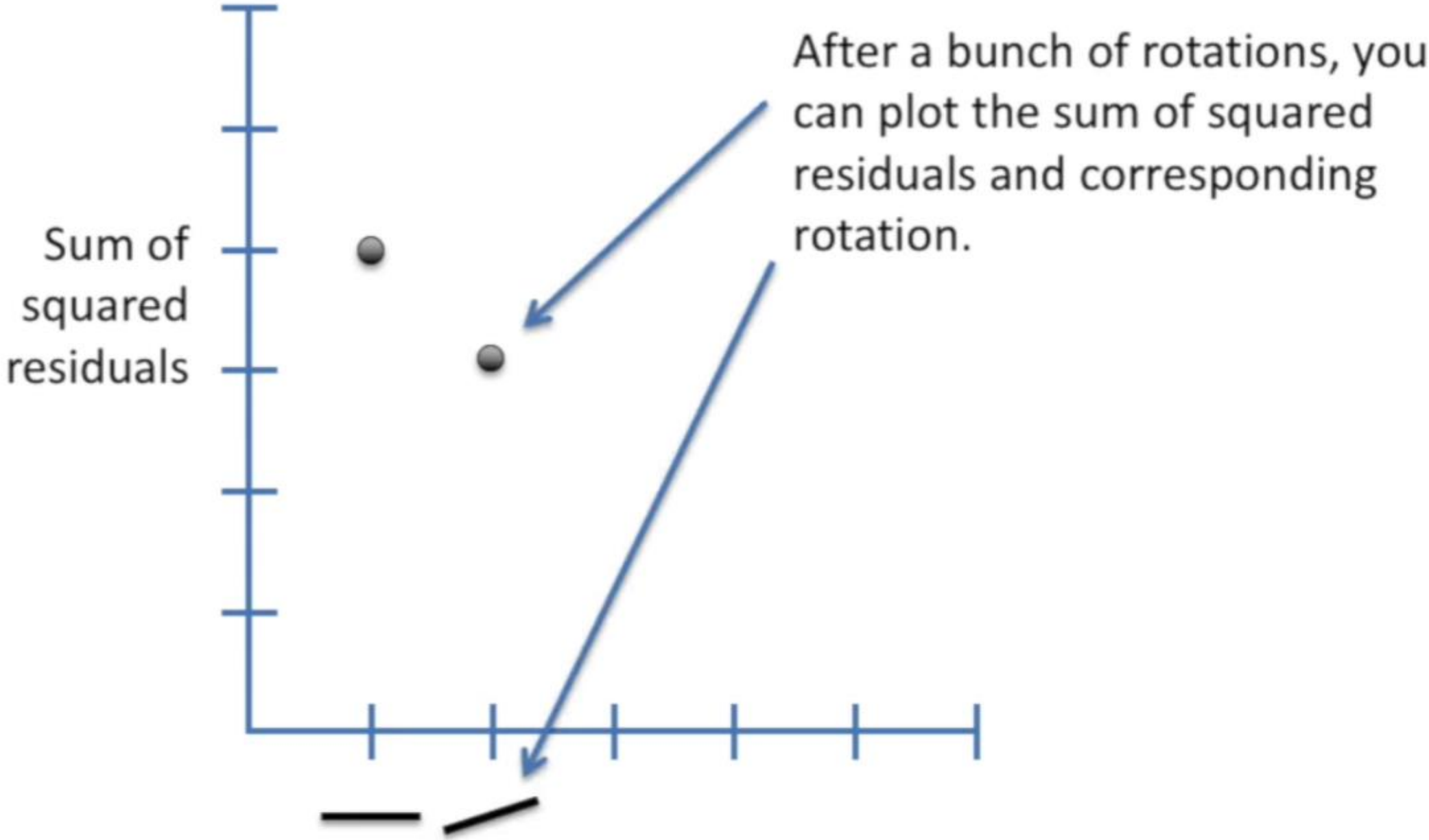


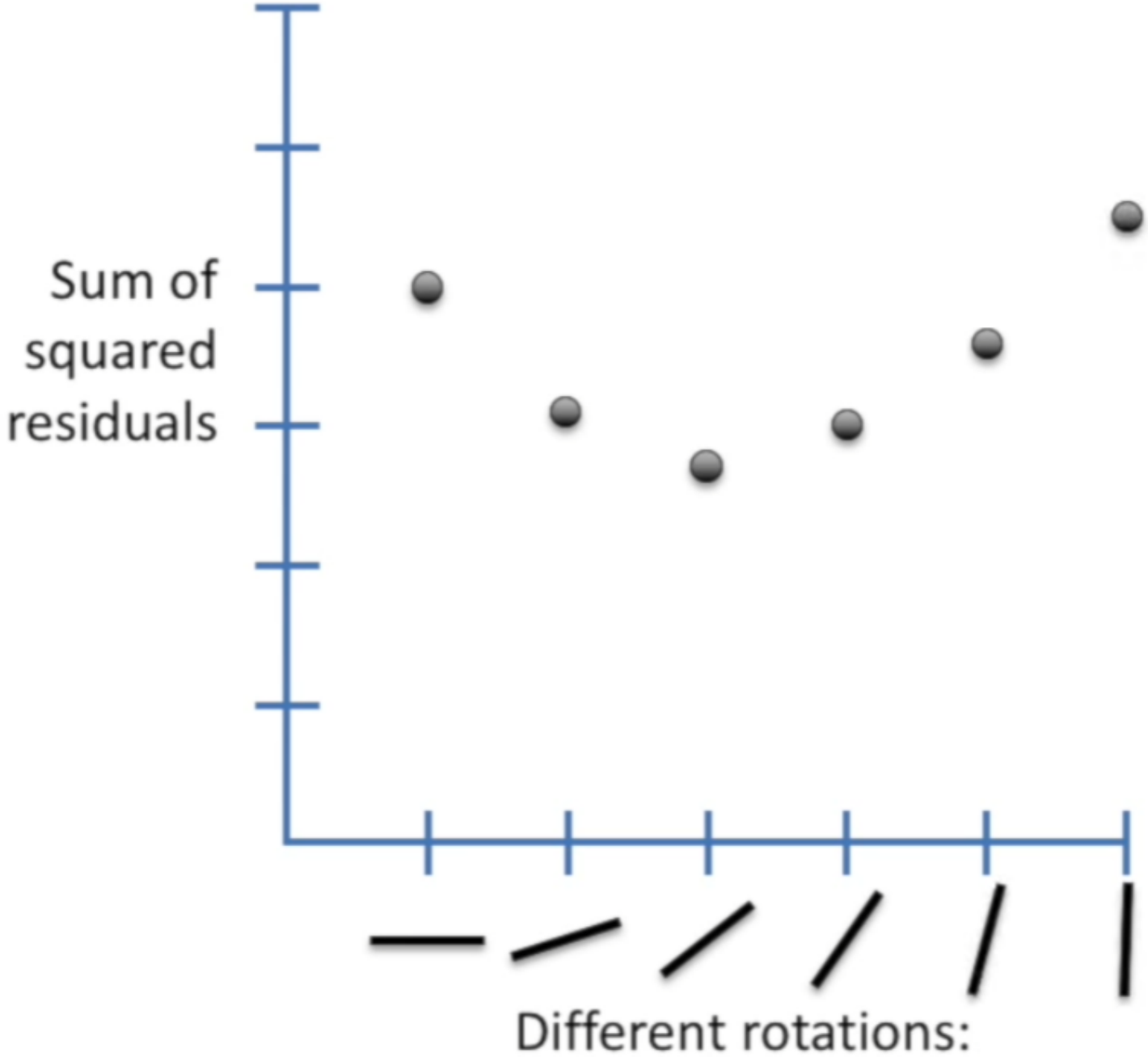
With the new line, measure the residuals, square them, and then sum up the squares.

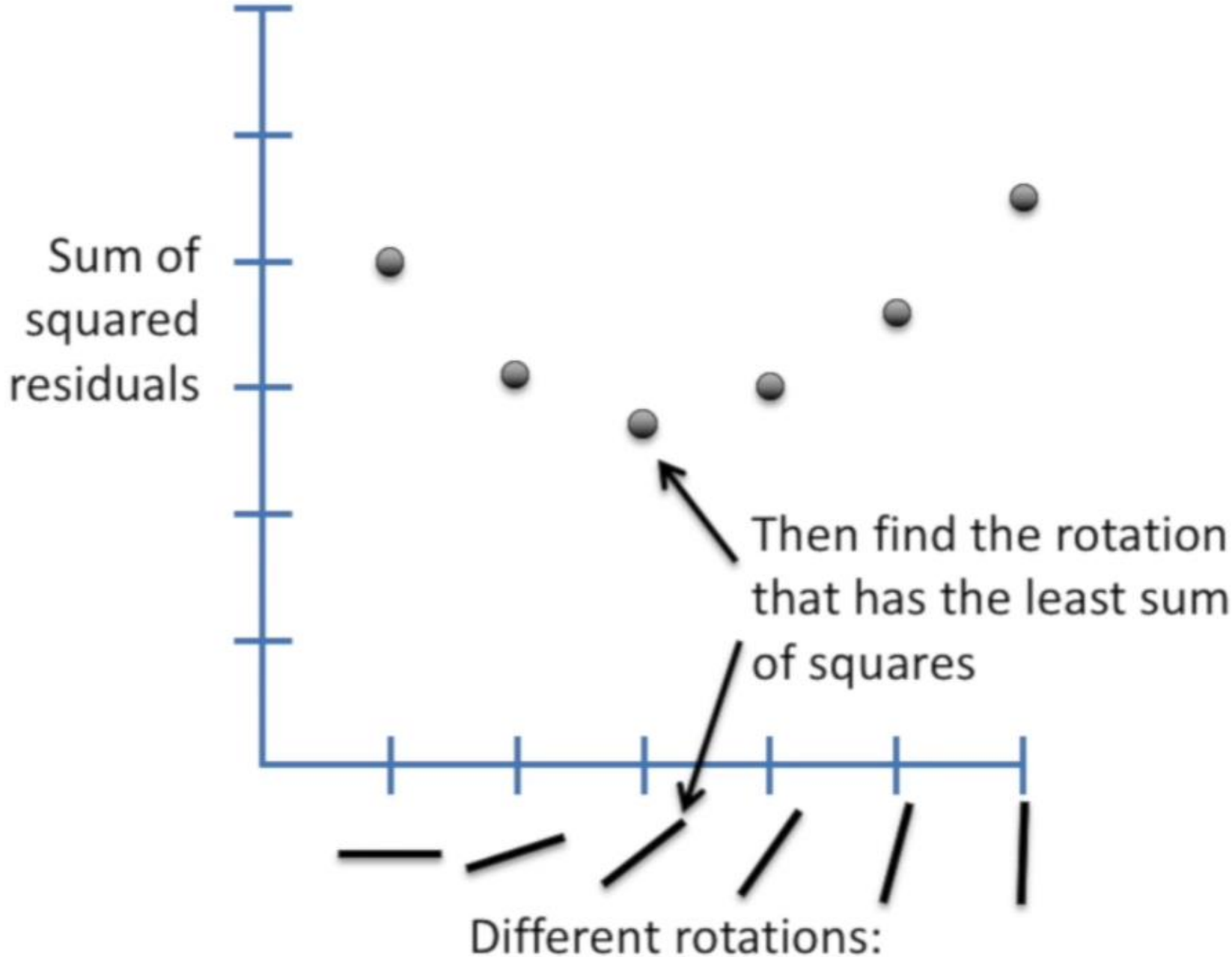

Rotate the line a little bit more…
Mouse size
Mouse weight

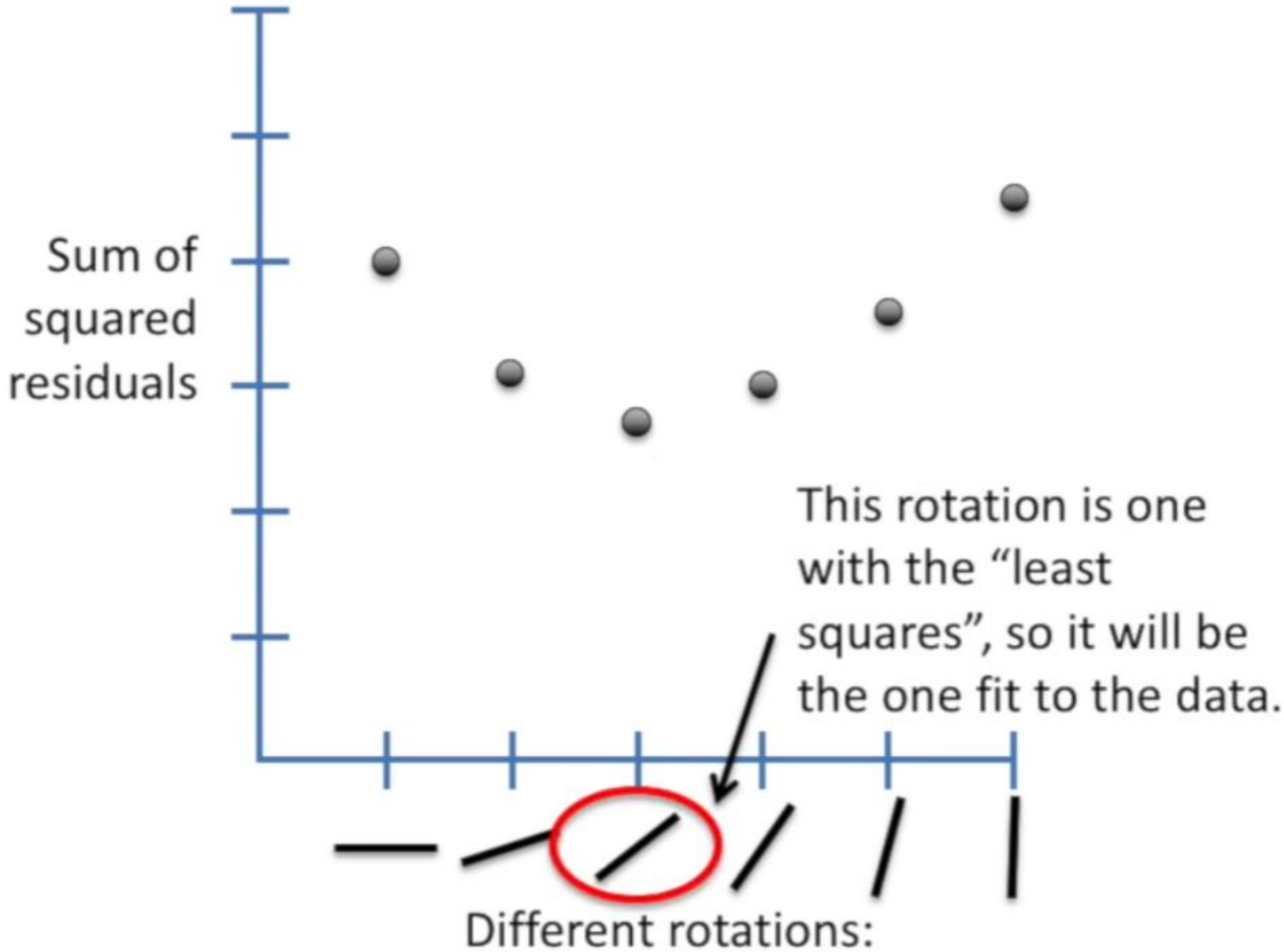

Mouse size
Sum up the squared residuals…
Mouse weight

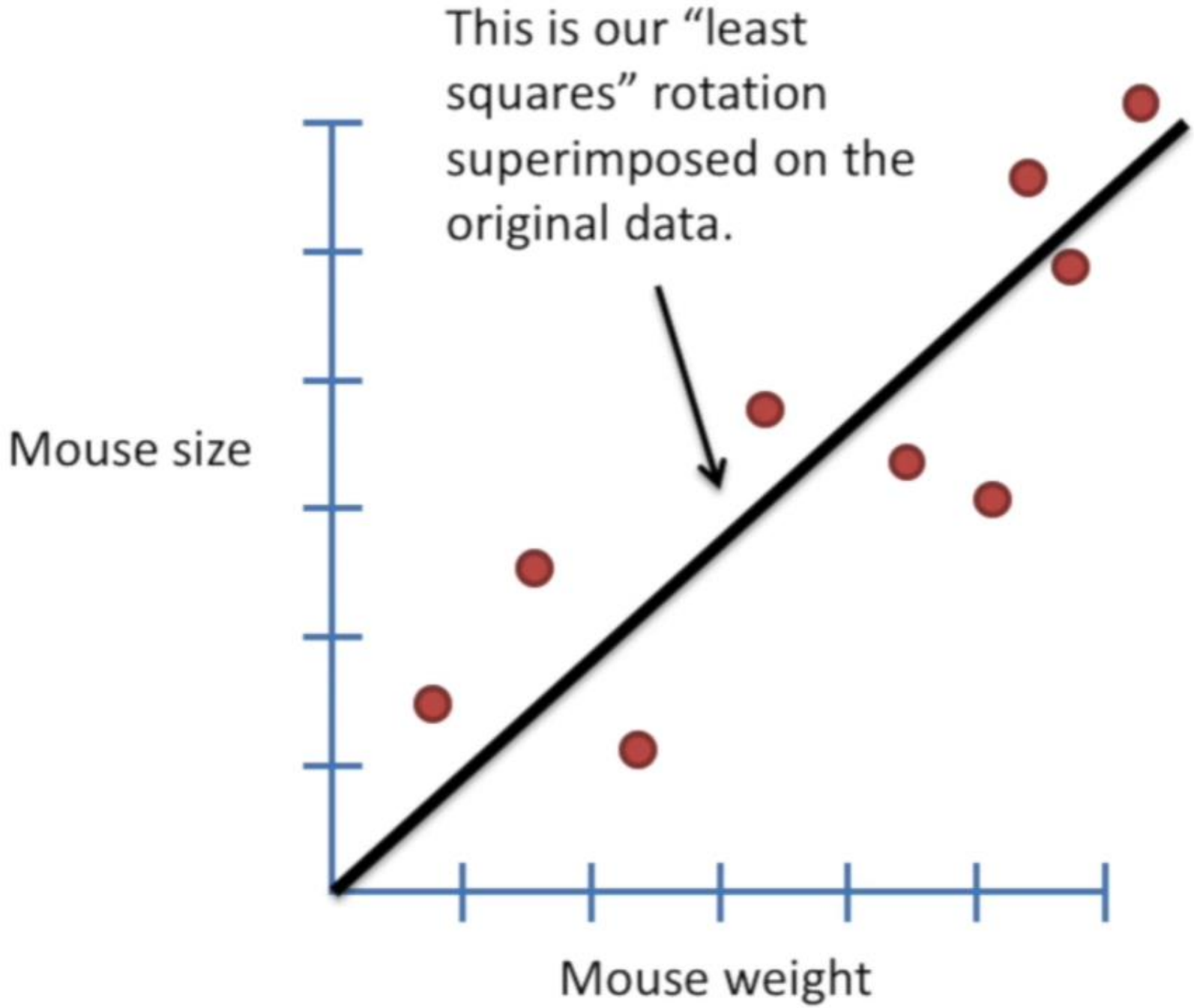

After a bunch of rotations, you can plot the sum of squared residuals and corresponding rotation.
Sum of squared residuals
Different rotations:

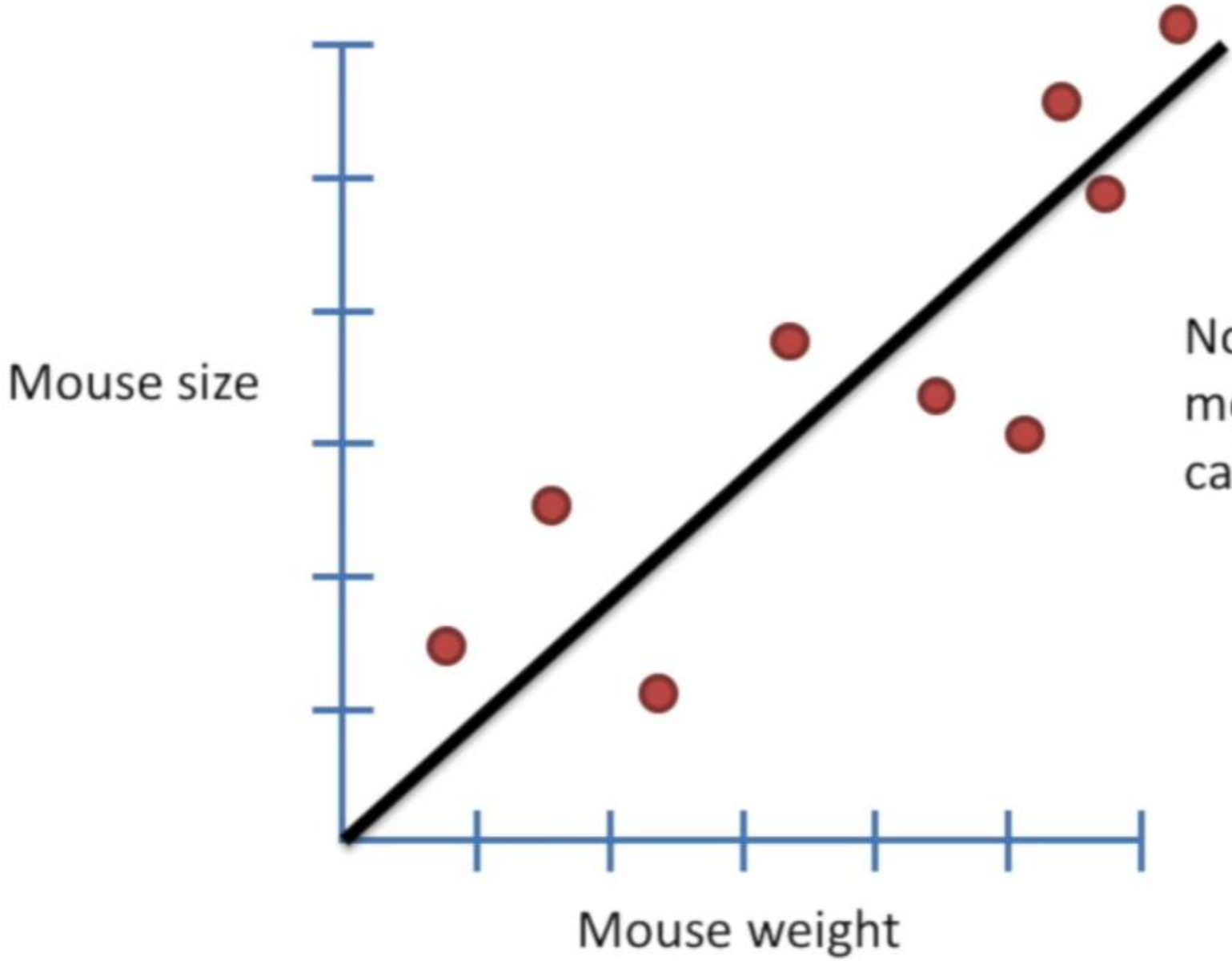

After a bunch of rotations, you can plot the sum of squared residuals and corresponding rotation.
Sum of squared residuals
Different rotations:


Sum of squared residuals
Different rotations:


Sum of squared residuals
Then find the rotation that has the least sum of squares
Different rotations:

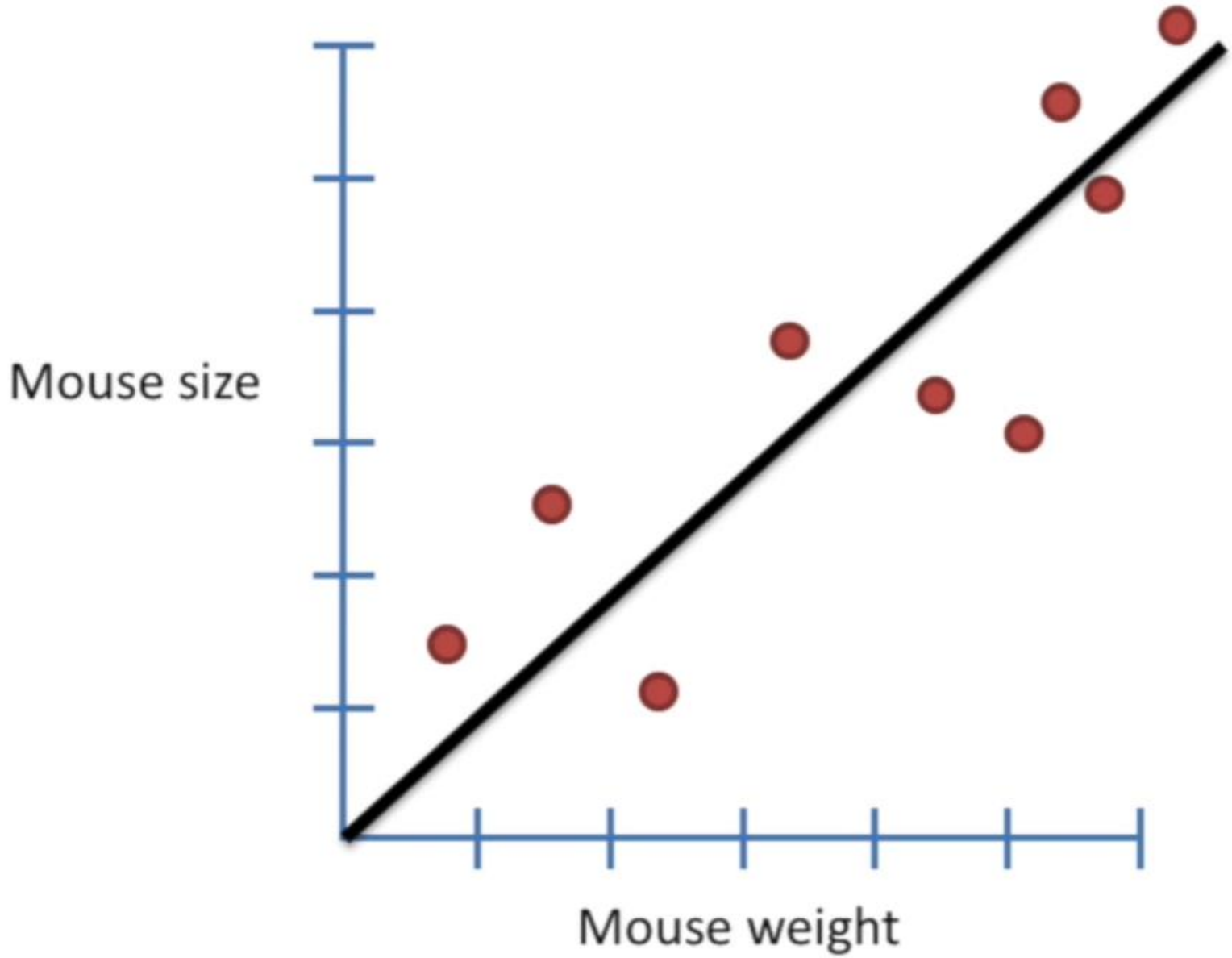

Sum of squared residuals
This rotation is one with the “least squares”, so it will be the one fit to the data.
Different rotations:


This is our “least squares” rotation superimposed on the original data.
Mouse size
Mouse weight

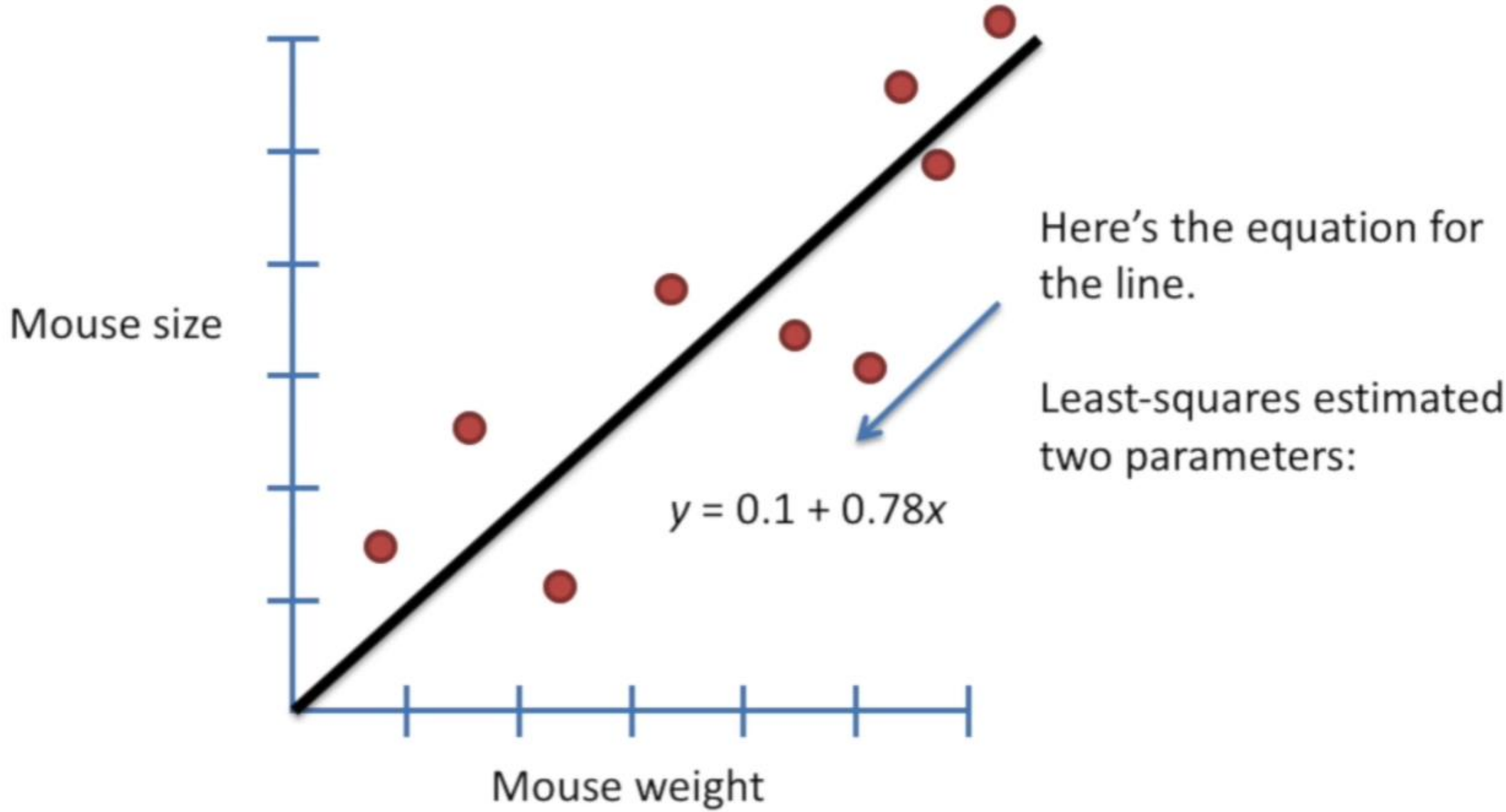


Now we know why the method for fitting a line is called “Least Squares”.

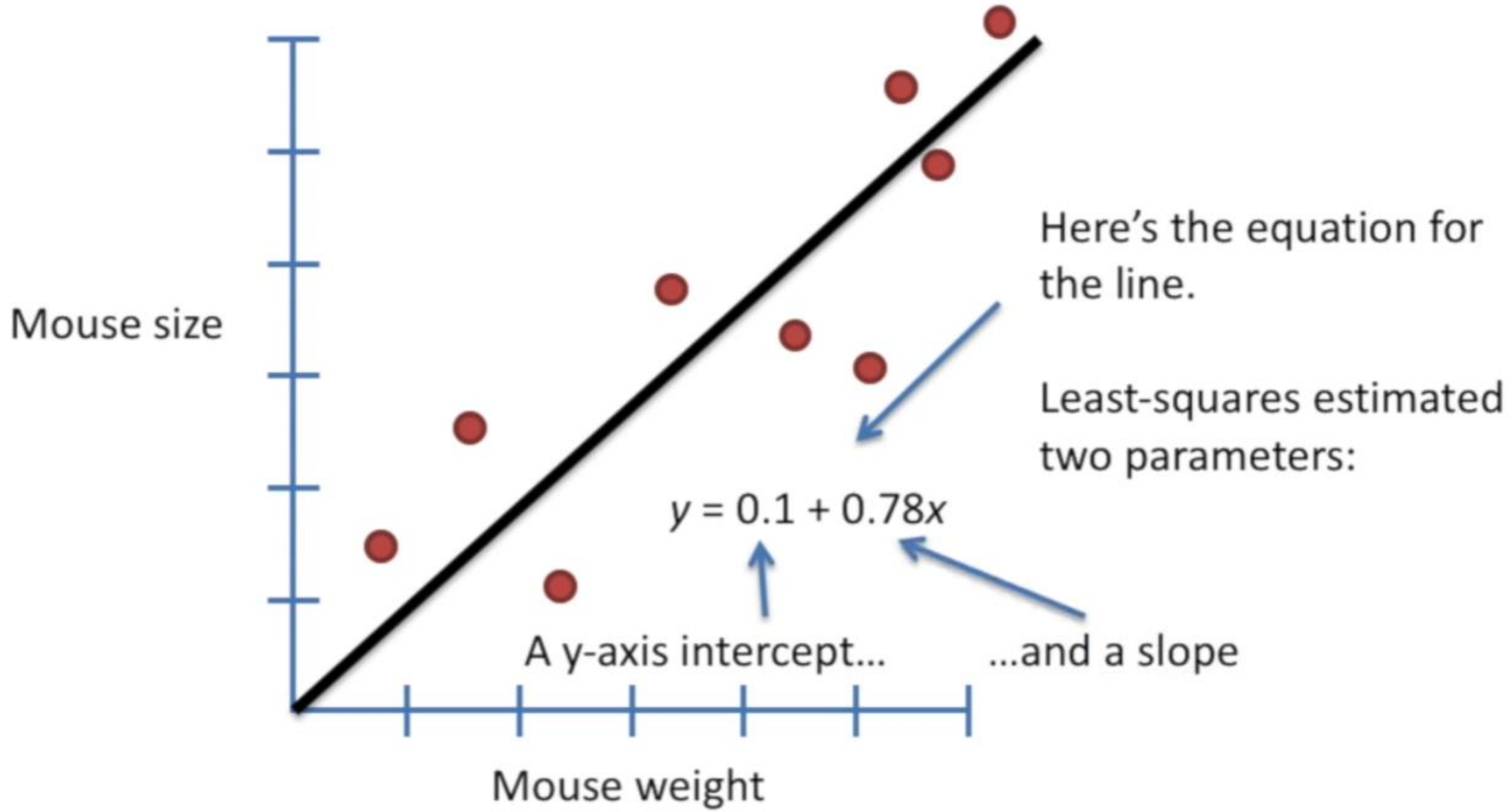
Now we have fit a line to the data! This is awesome!



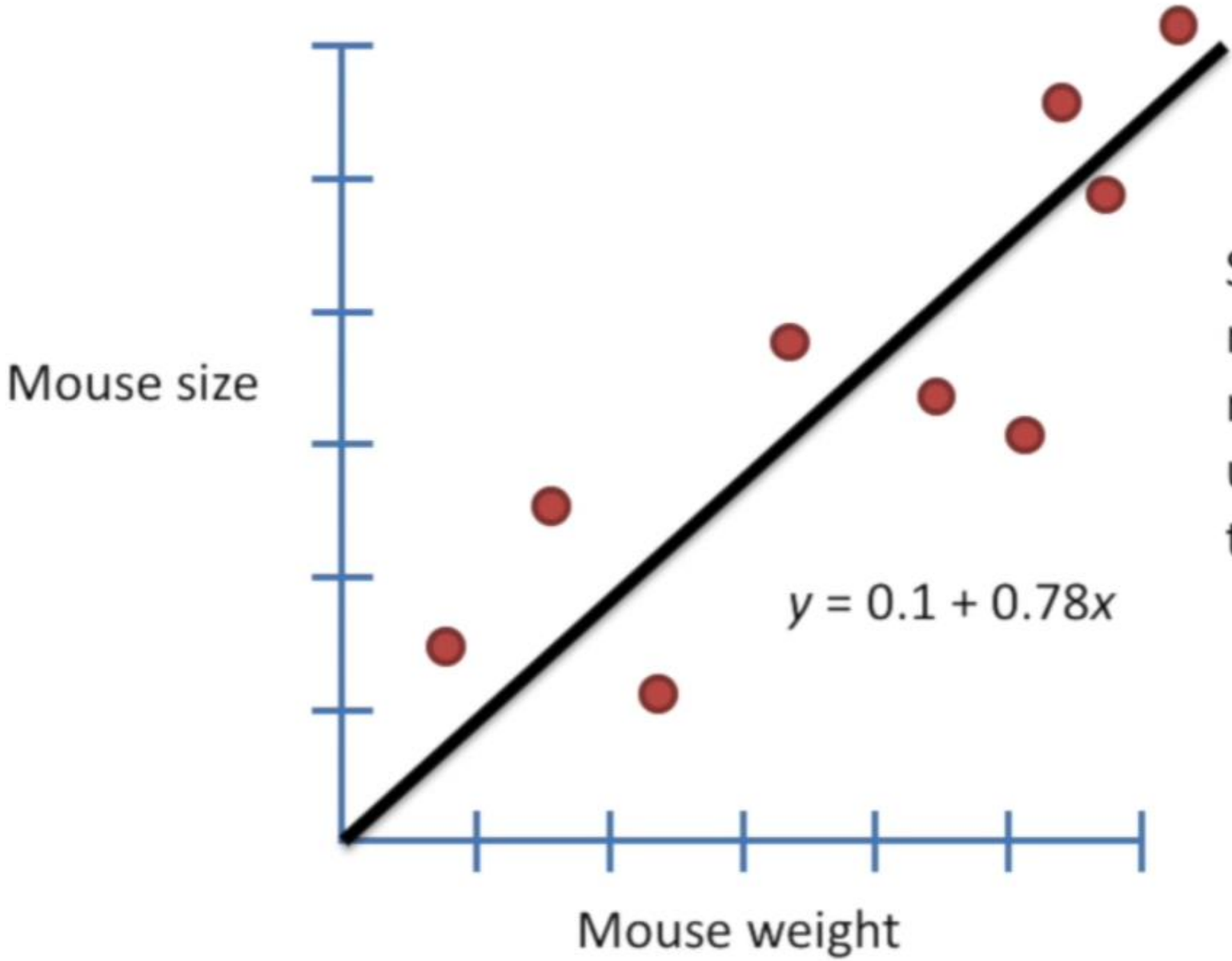
Now we have fit a line to the data! This is awesome!



Now we have fit a line to the data! This is awesome!

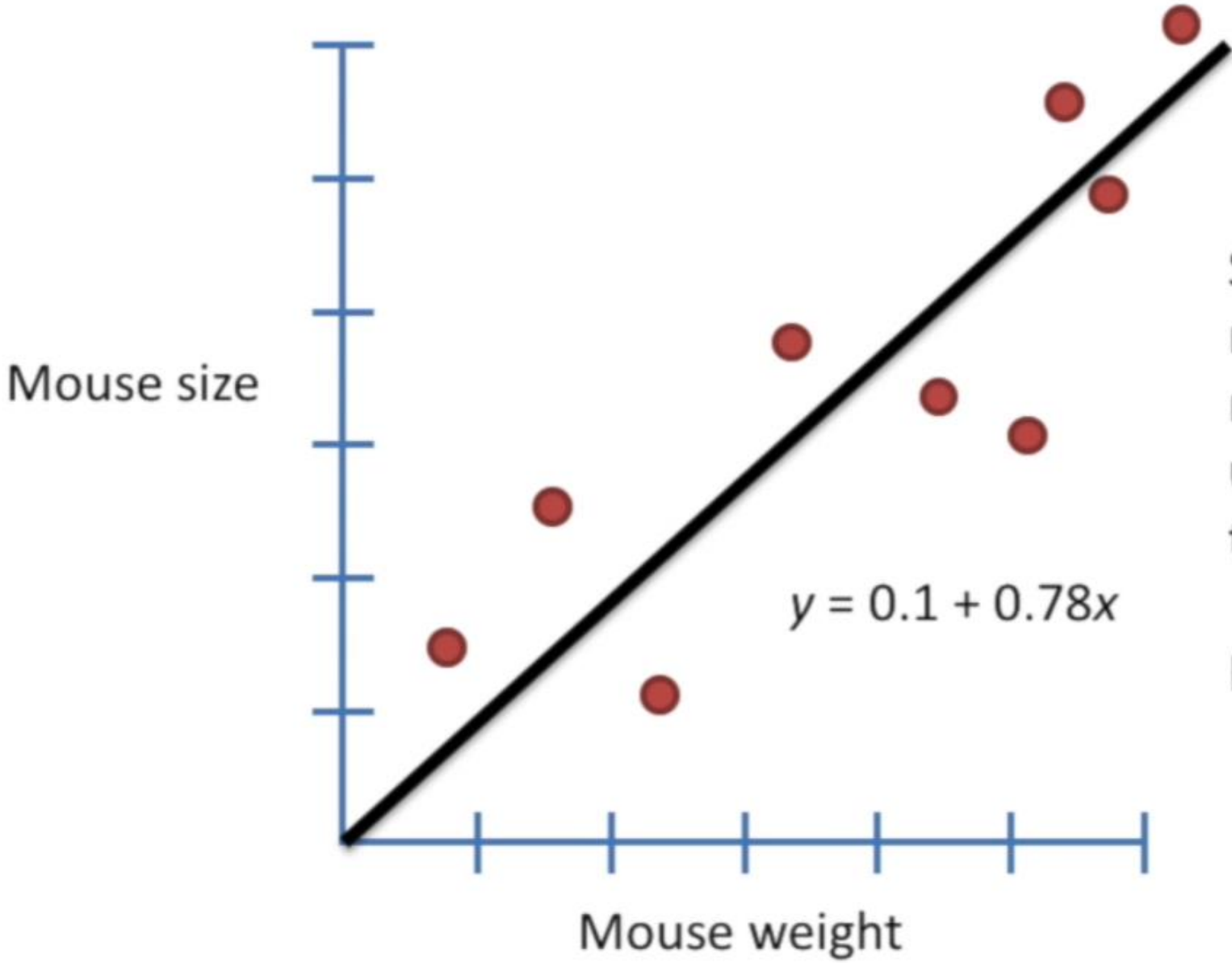


Now we have fit a line to the data! This is awesome!



Since the slope is not 0, it means that knowing a mouse's weight will help us make a guess about that mouse's size.

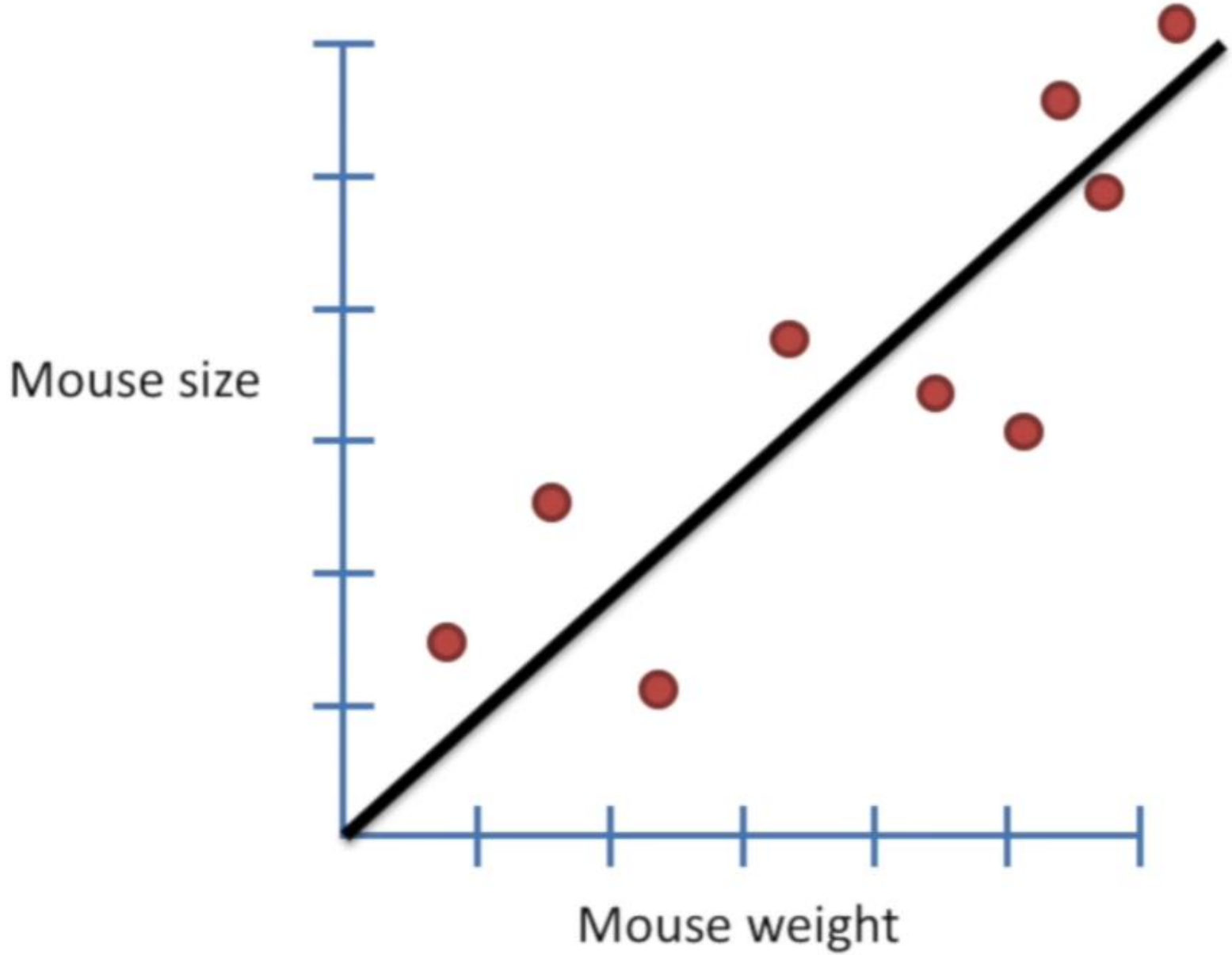
Now we have fit a line to the data! This is awesome!



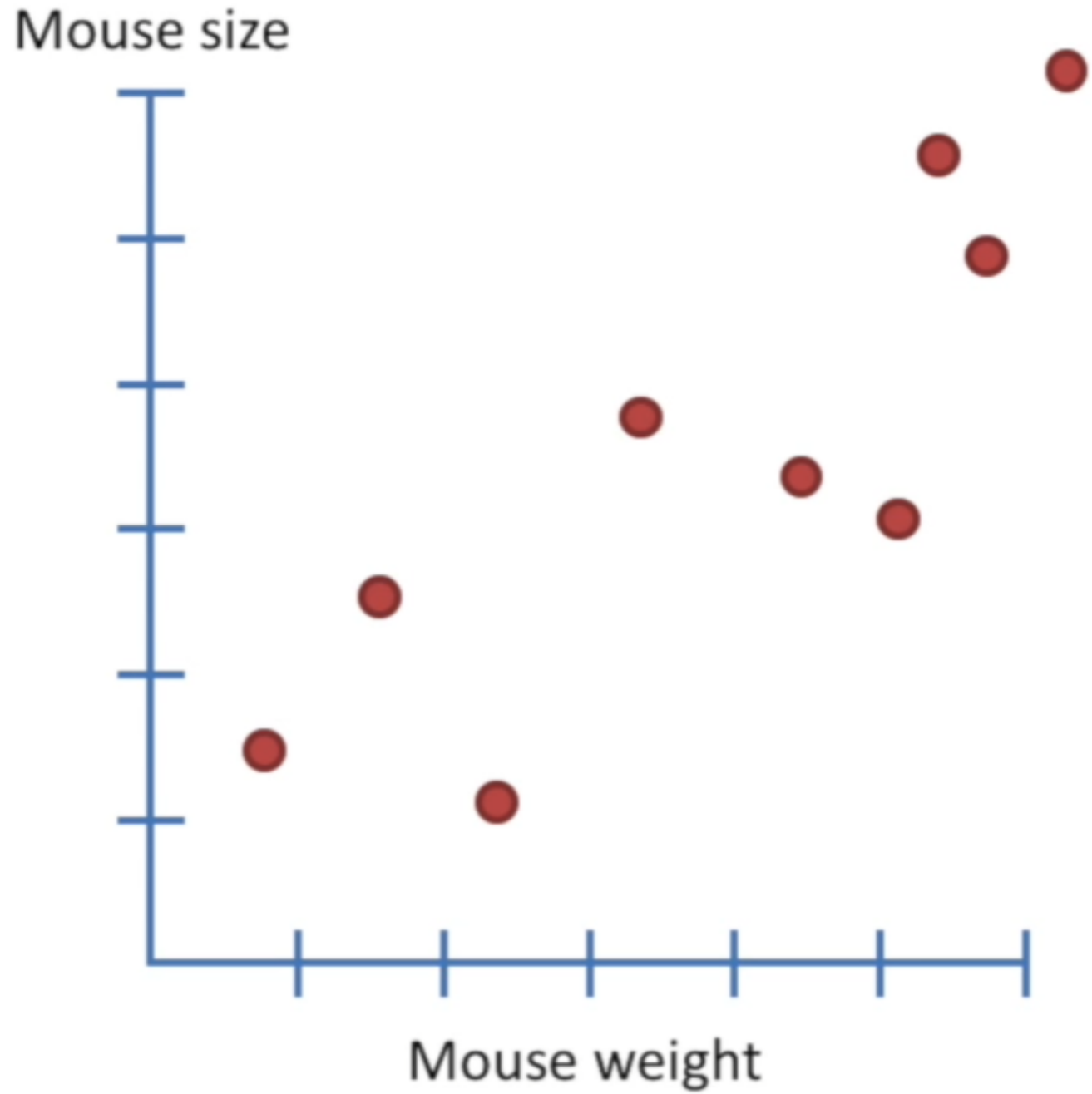
Since the slope is not 0, it means that knowing a mouse's weight will help us make a guess about that mouse's size.

How good is that guess?

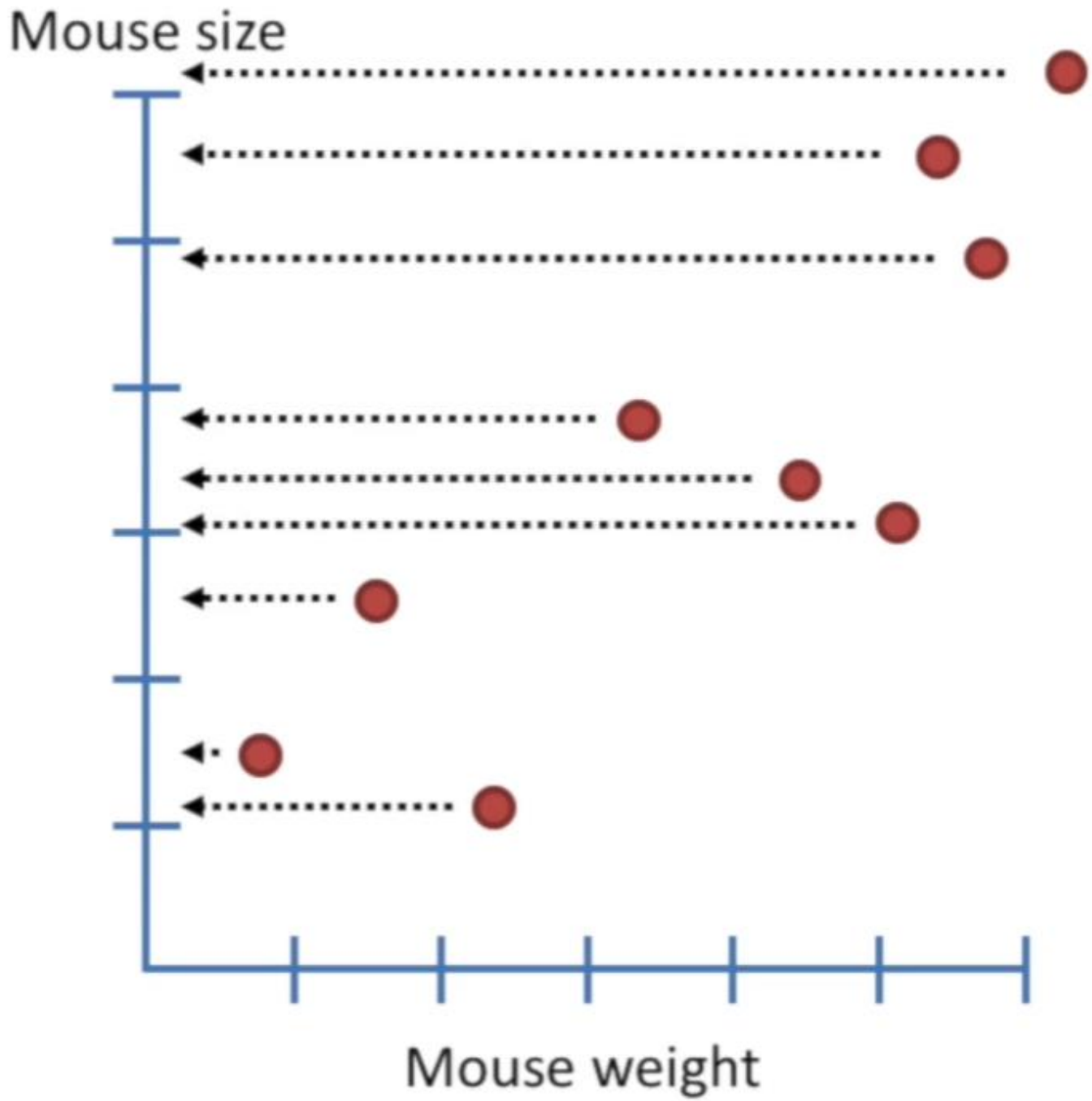
Calculating R^2 is the first step in determining how good that guess will be.



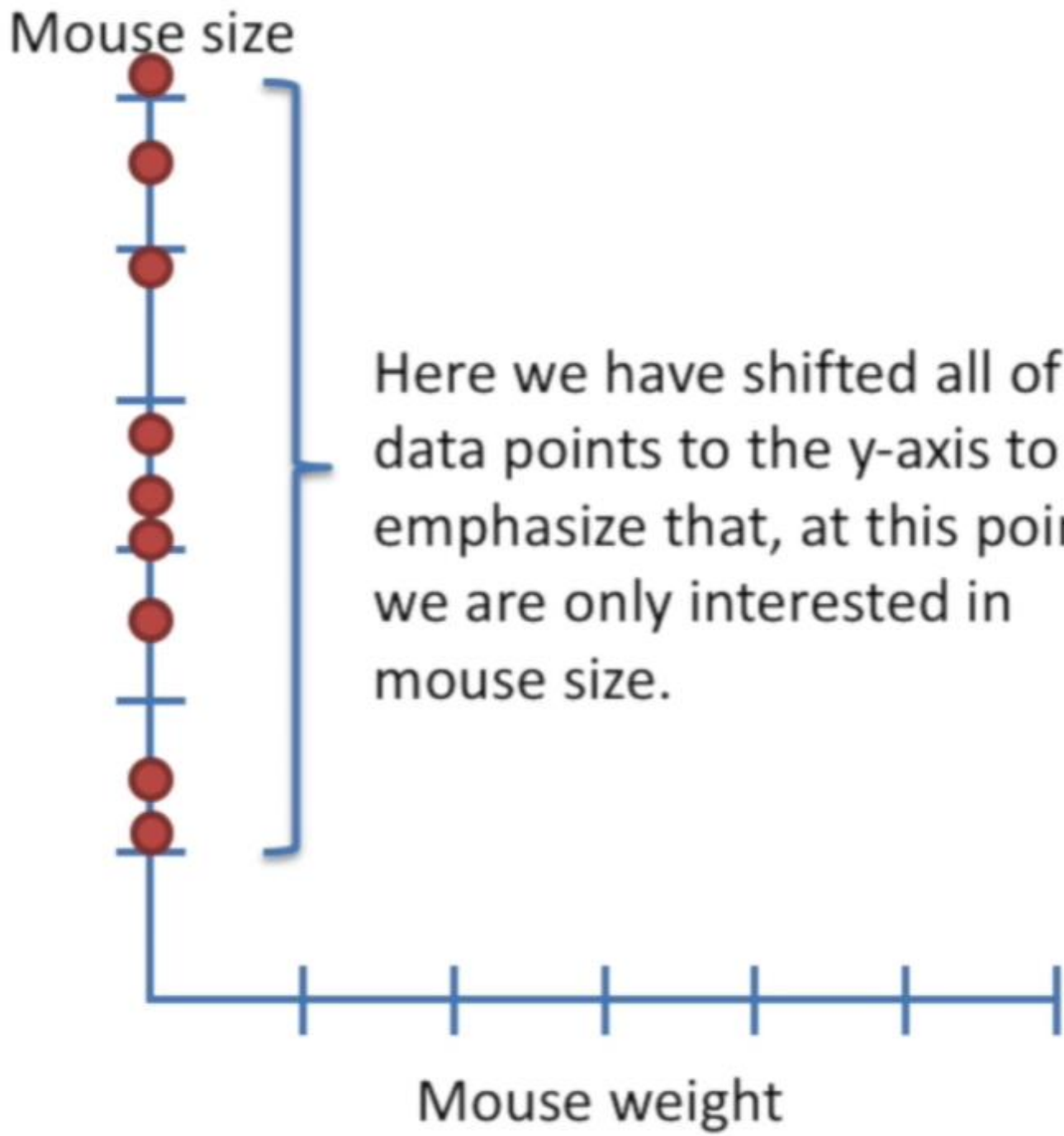
First, calculate the average mouse size.

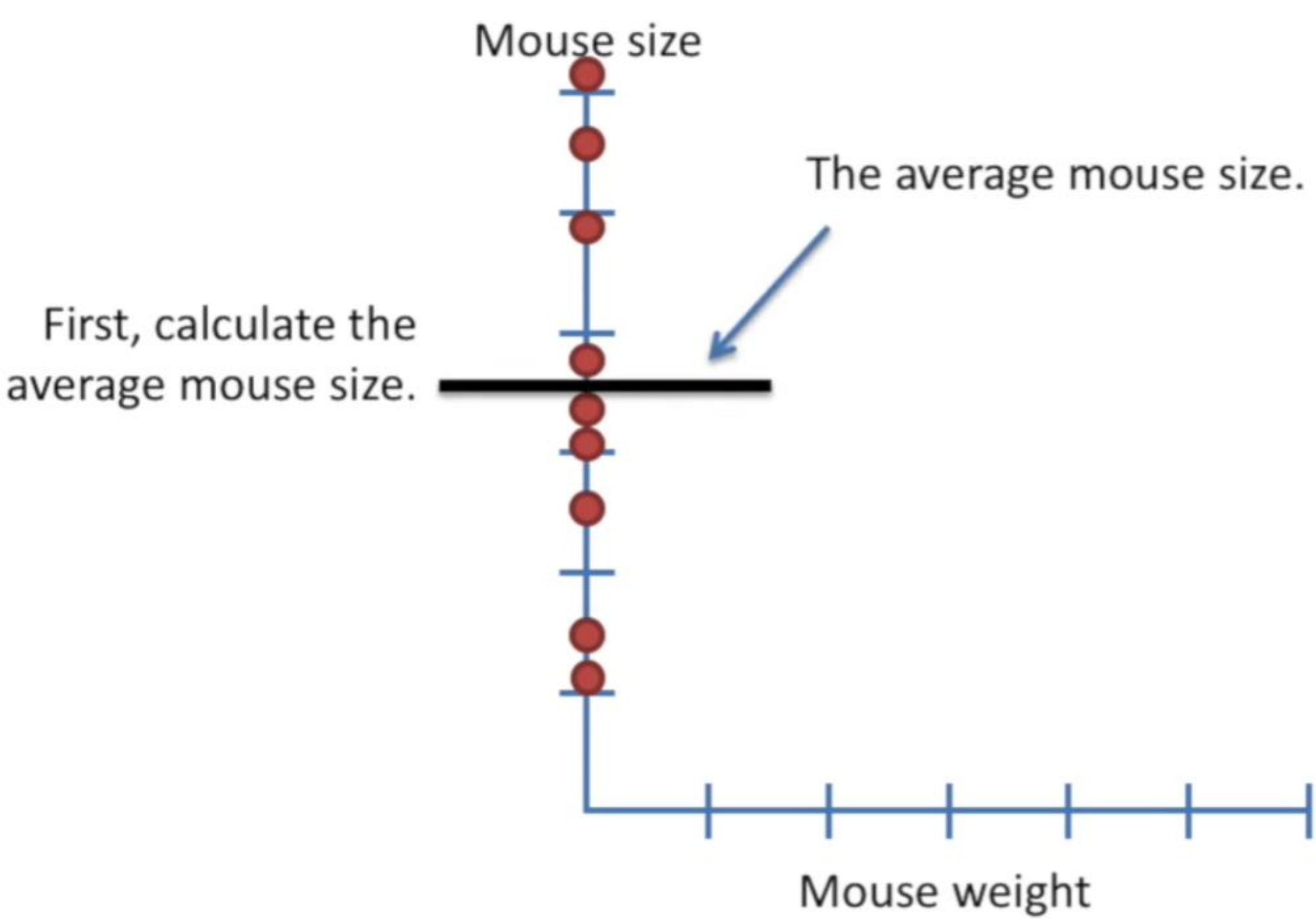


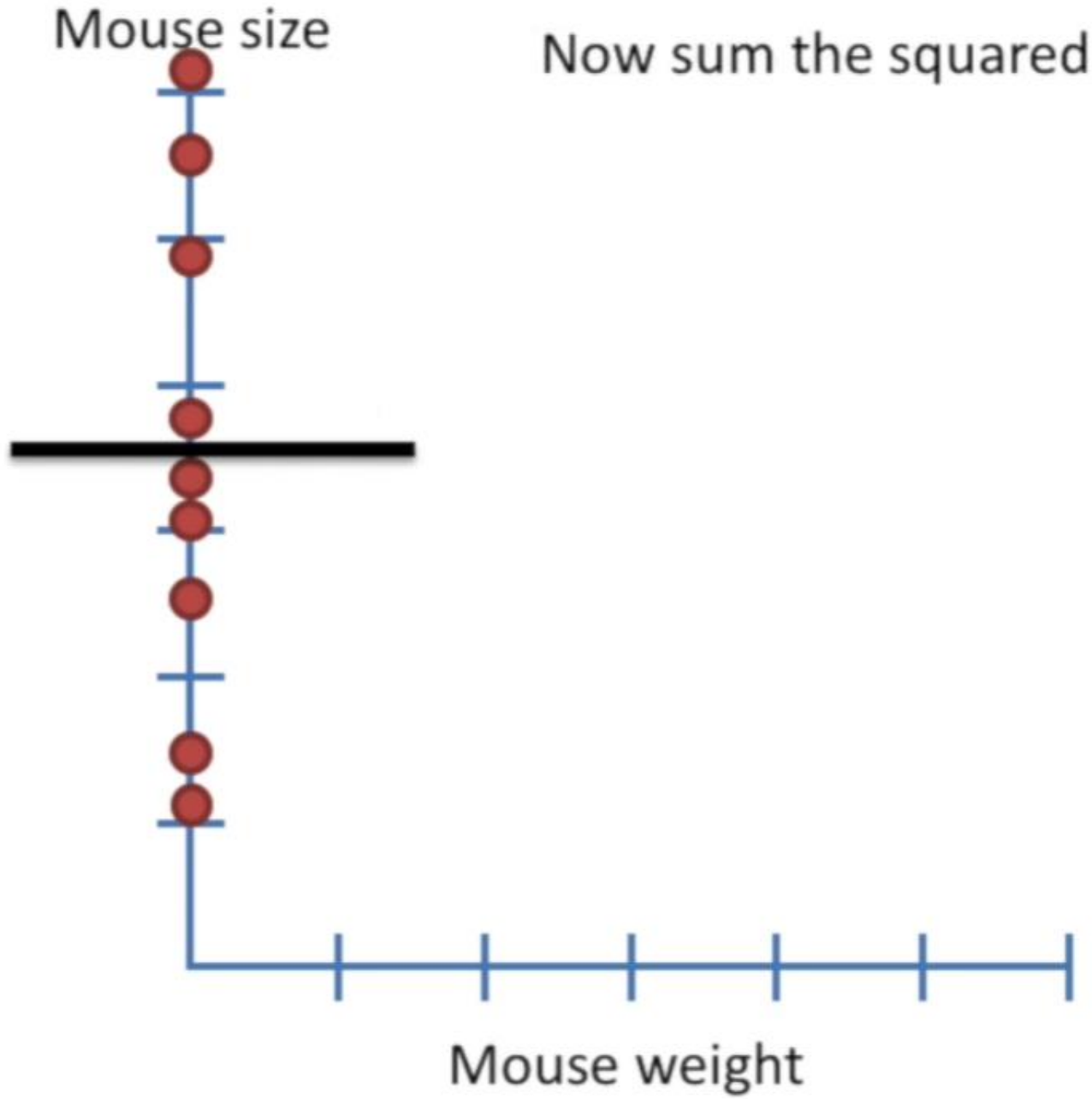
First, calculate the average mouse size.



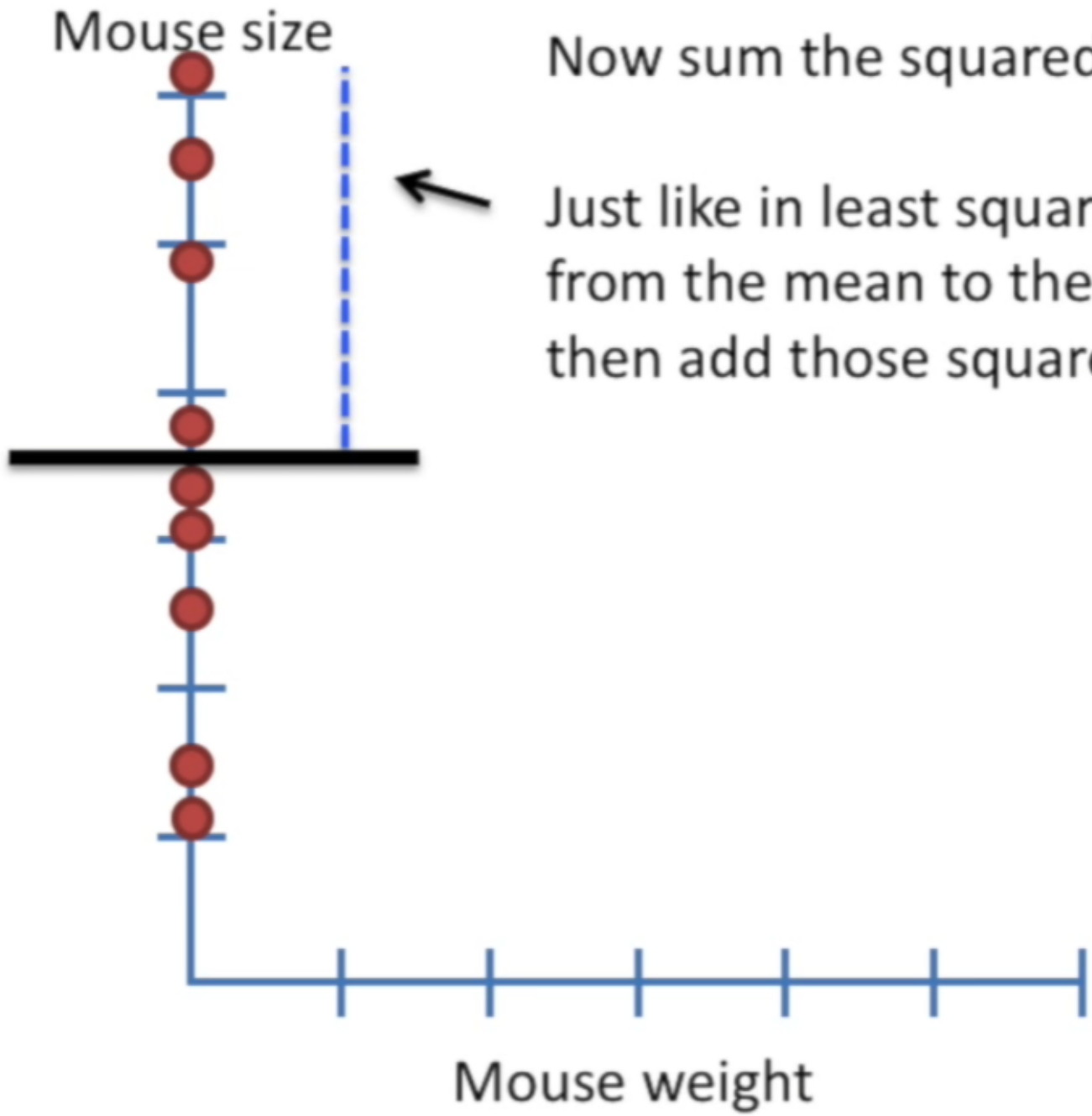
First, calculate the average mouse size.




Mouse size
The average mouse size.
First, calculate the average mouse size.
Mouse weight

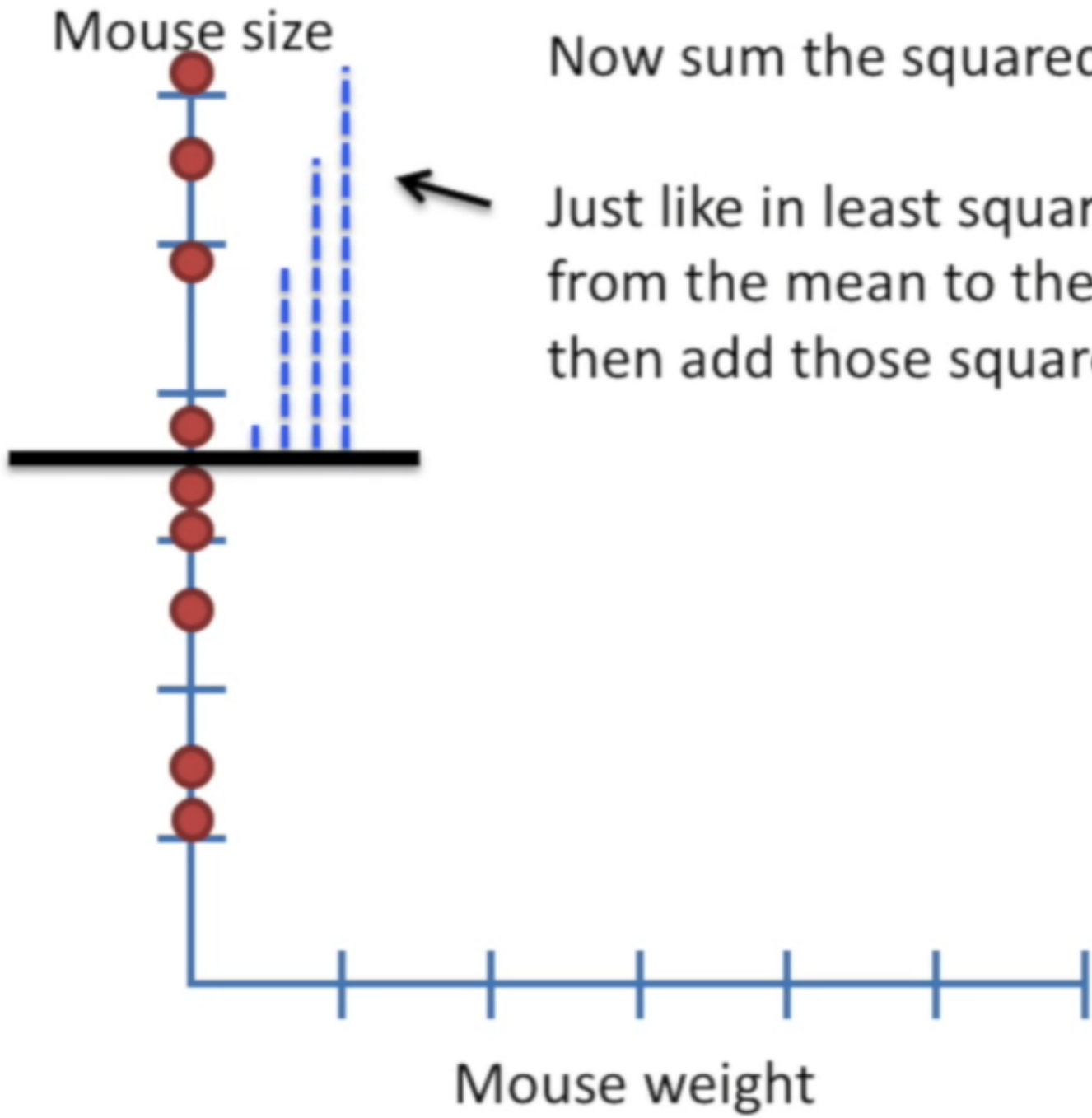


Now sum the squared residuals…



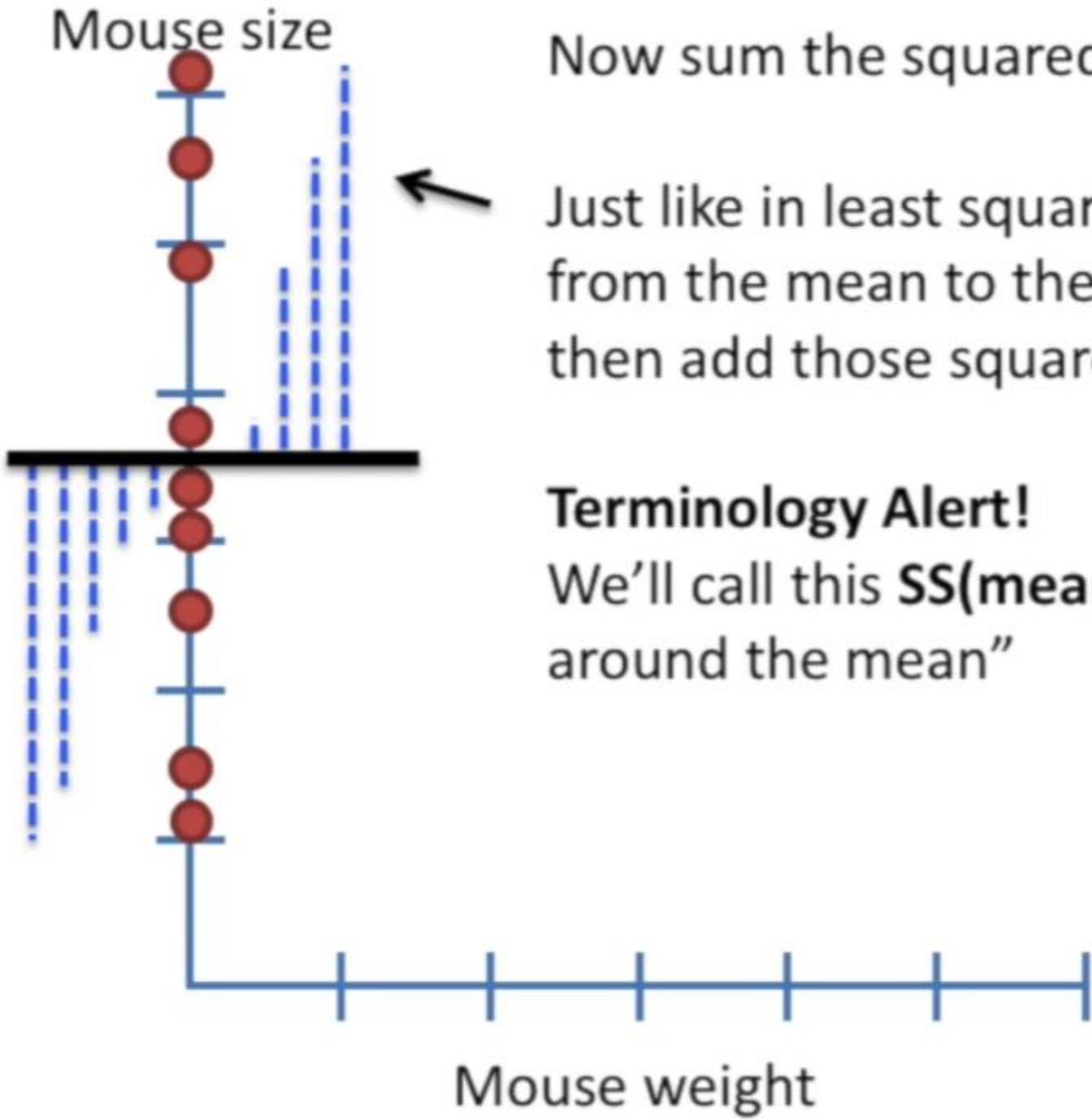
Now sum the squared residuals…

Just like in least squares, we measure the distance from the mean to the data point and square it, then add those squares together.



Now sum the squared residuals…

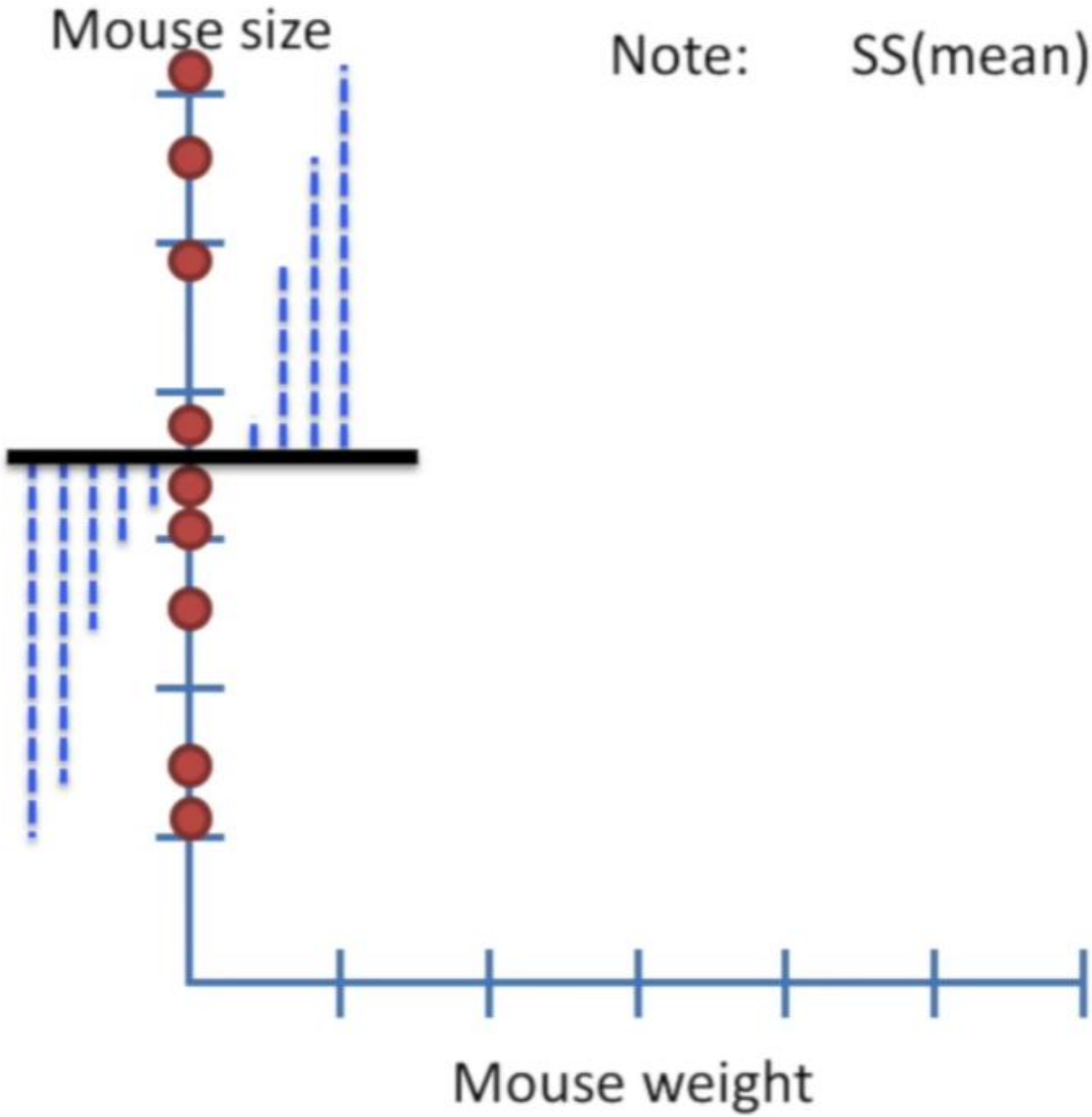
Just like in least squares, we measure the distance from the mean to the data point and square it, then add those squares together.



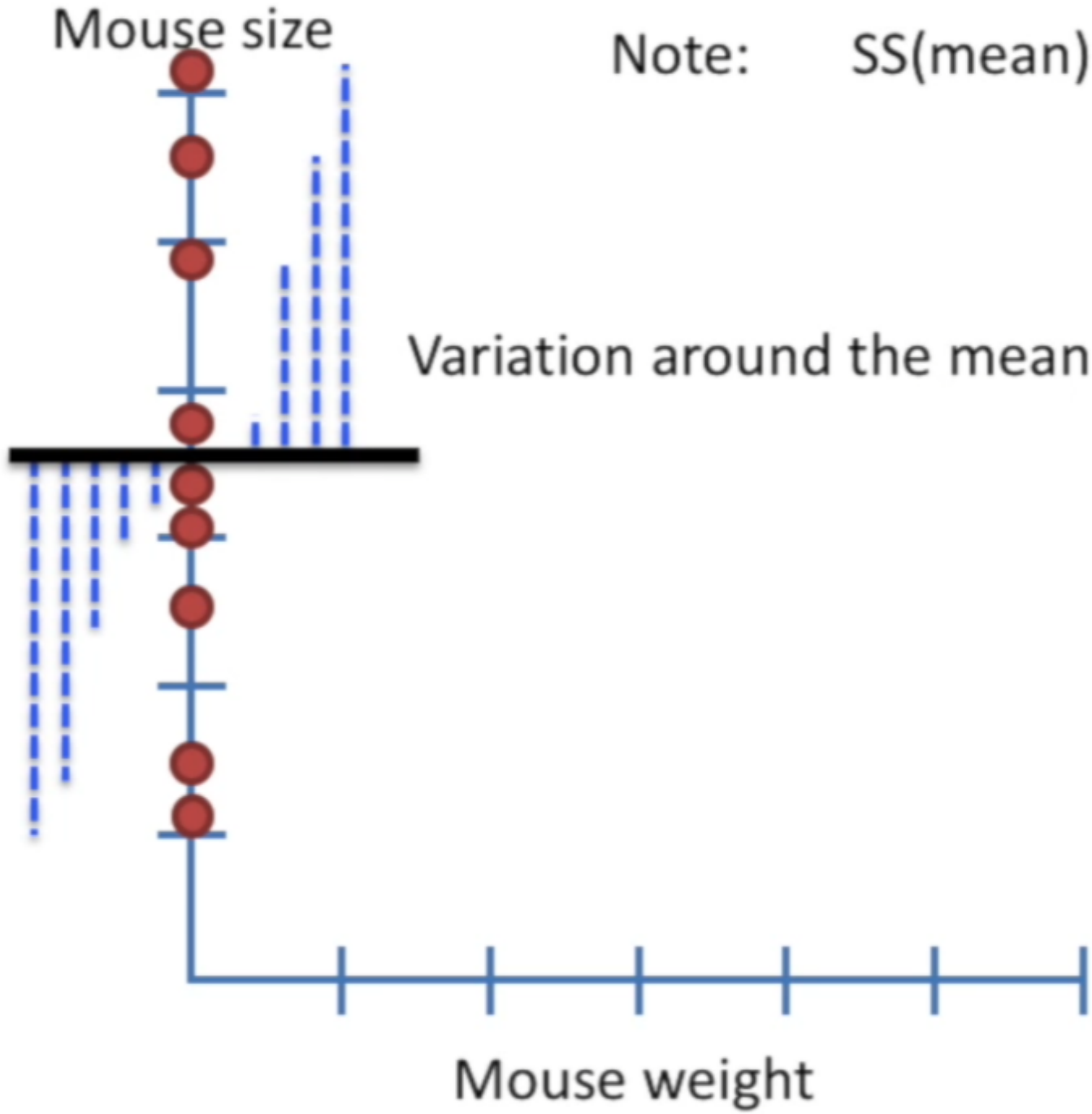
Now sum the squared residuals…

Just like in least squares, we measure the distance from the mean to the data point and square it, then add those squares together.

Terminology Alert!
We'll call this **SS(mean)**, for "sum of squares around the mean"

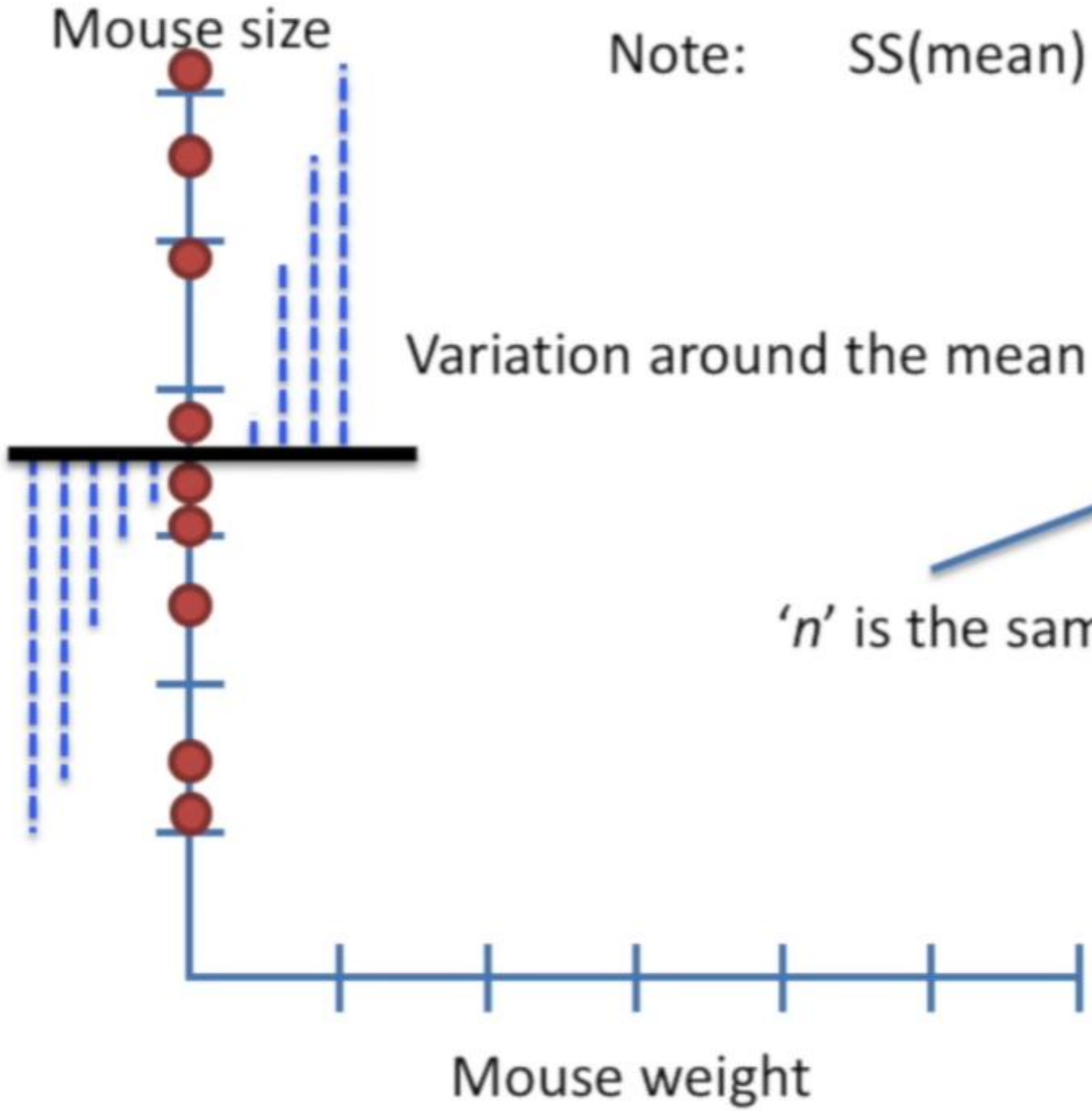


Note: SS(mean) = (data - mean)^2



Note: $SS(\text{mean}) = (\text{data} - \text{mean})^2$

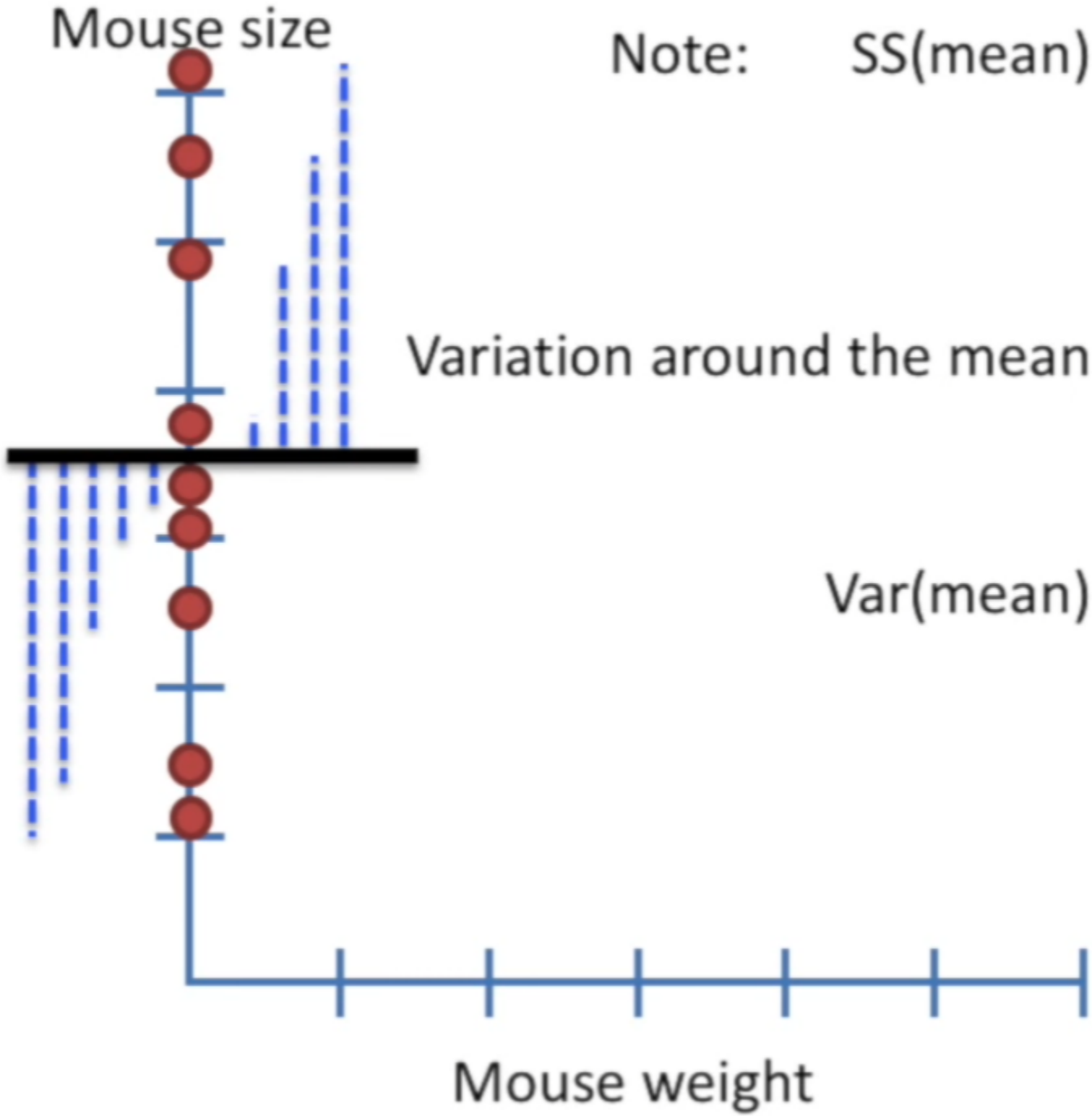
$$\text{Variation around the mean} = \frac{(\text{data} - \text{mean})^2}{n}$$



Note: $SS(\text{mean}) = (\text{data} - \text{mean})^2$

$$\text{Variation around the mean} = \frac{(\text{data} - \text{mean})^2}{n}$$

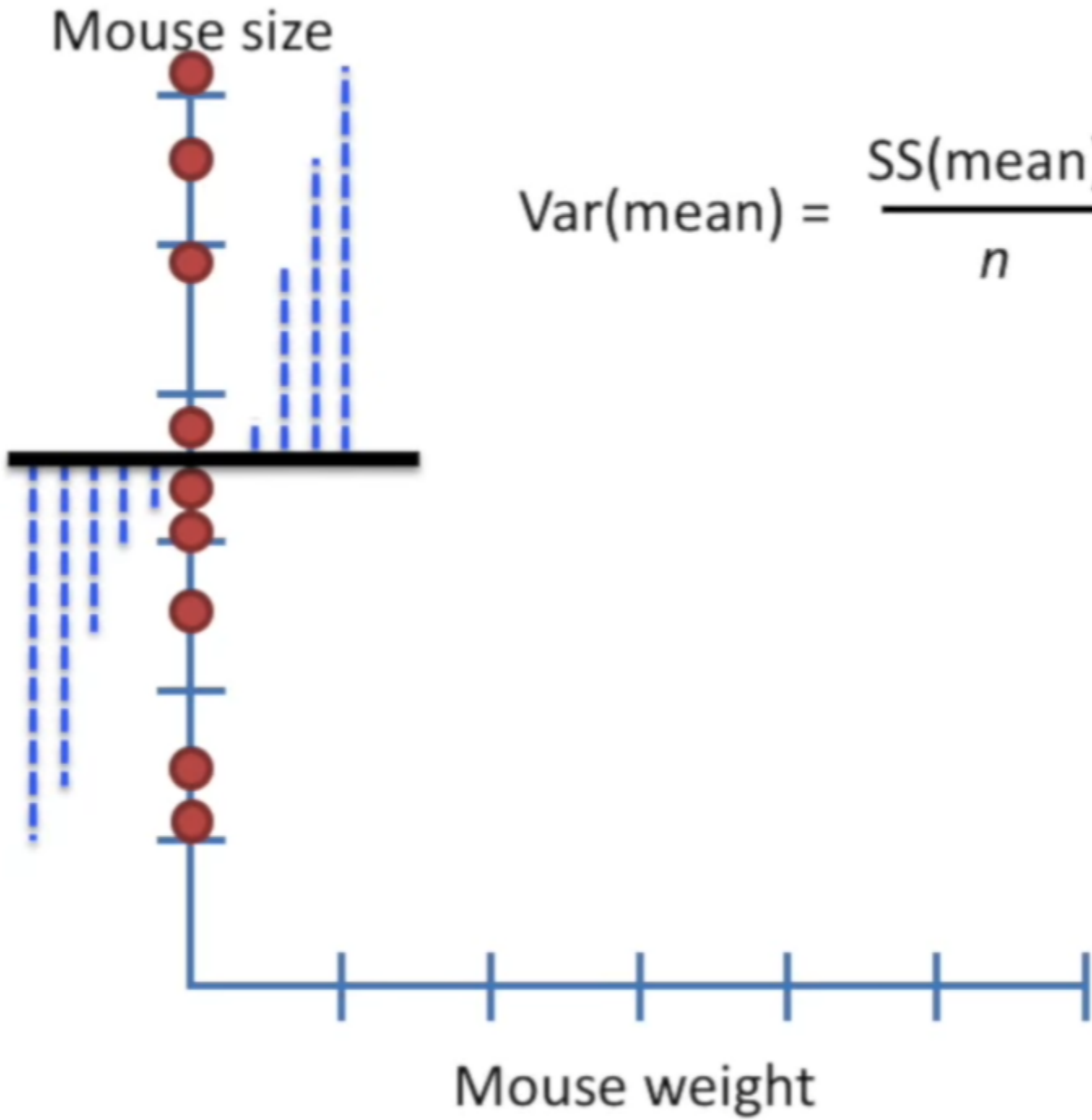
'*n*' is the sample size (in this case, n= 9)



Note: $SS(\text{mean}) = (\text{data} - \text{mean})^2$

$$\text{Variation around the mean} = \frac{(\text{data} - \text{mean})^2}{n}$$

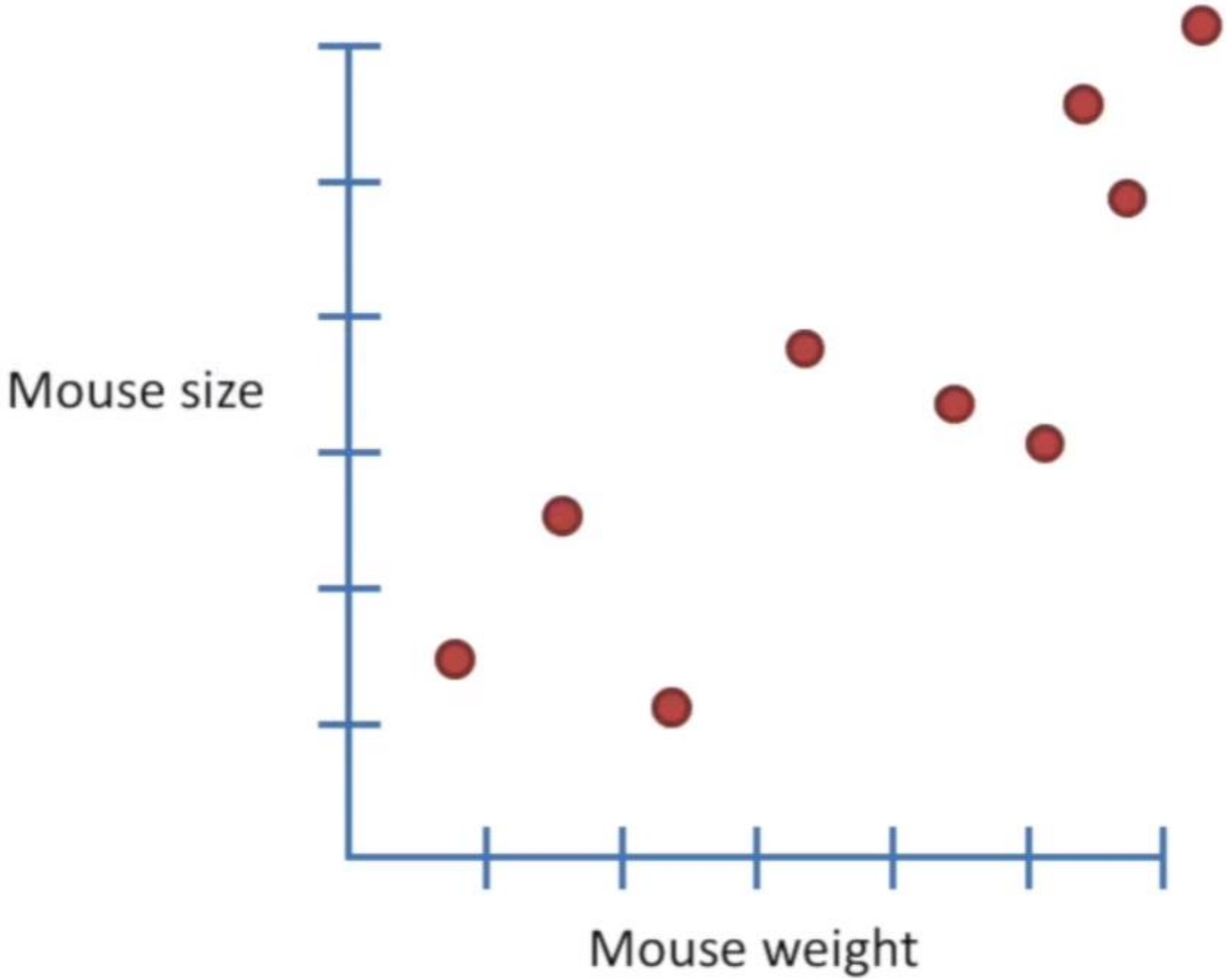
$$\text{Var(mean)} = \frac{SS(\text{mean})}{n}$$



$$\text{Var(mean)} = \frac{\text{SS(mean)}}{n}$$

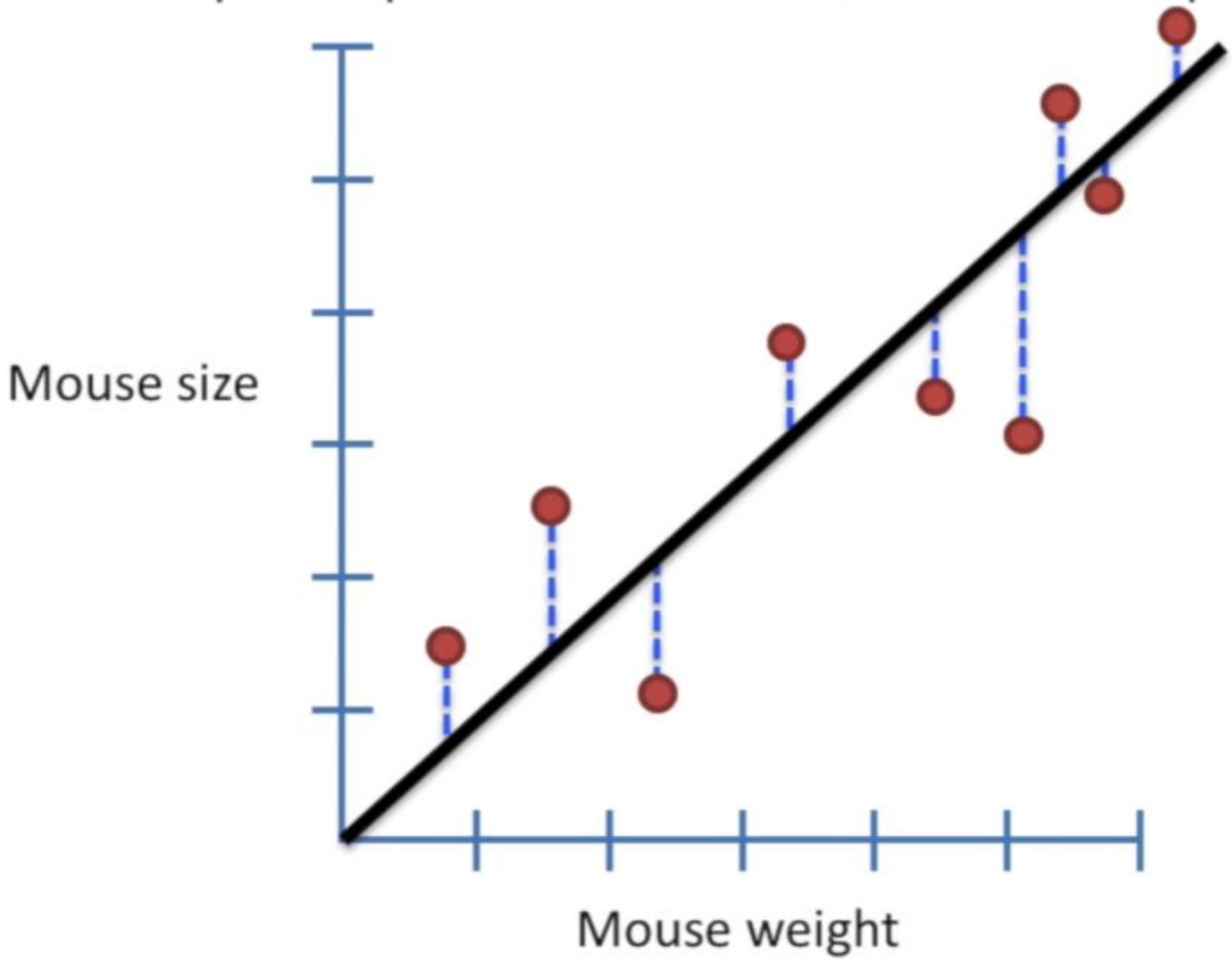
Another way to think about variance is as the average sum of squares per mouse.

Now go back to the original plot.



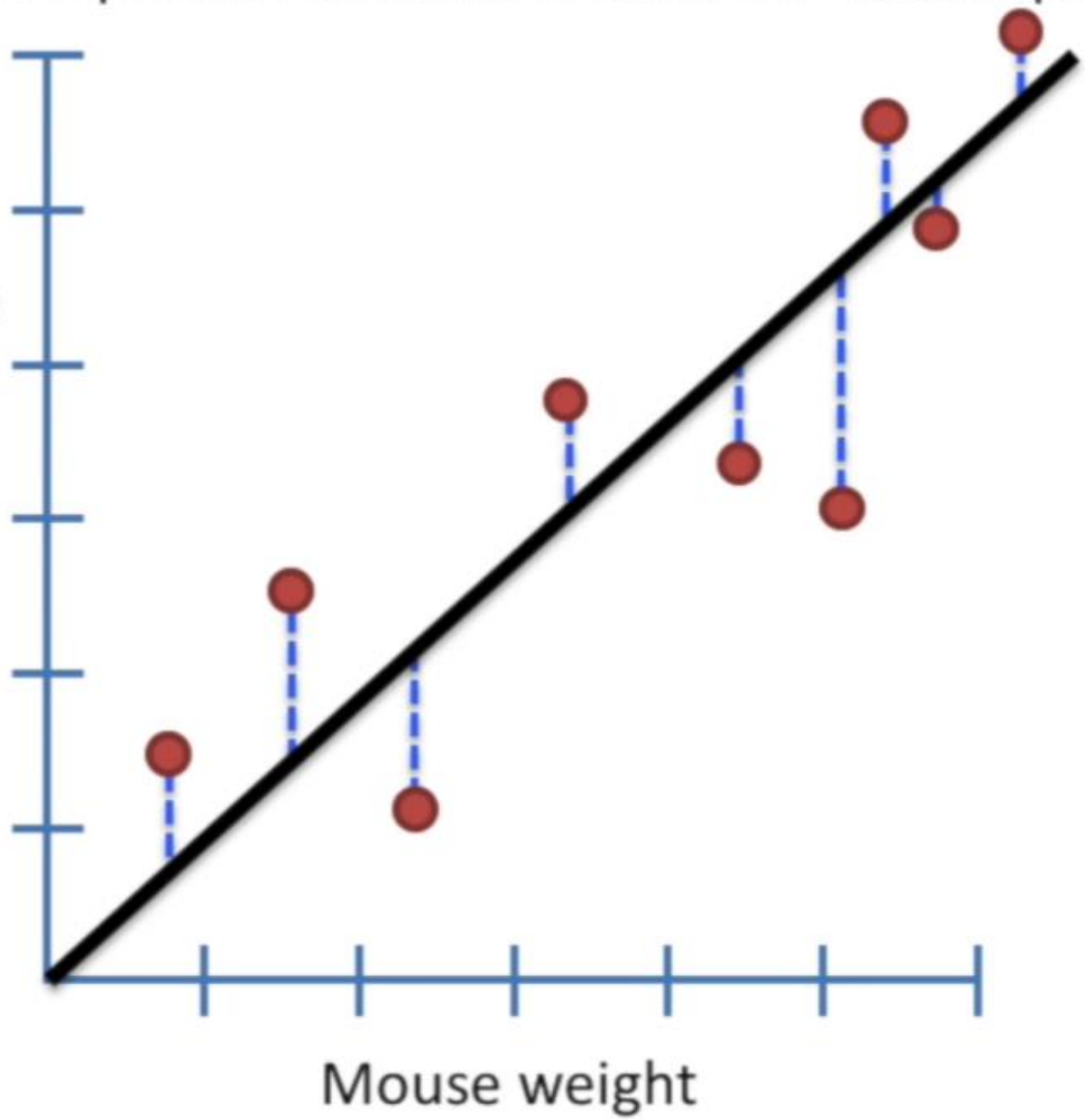
Now go back to the original plot.

Sum up the squared residuals around our least-squares fit.



Now go back to the original plot.
Sum up the squared residuals around our least-squares fit.

We'll call this **SS(fit)**, for the sum of squares around the least-squares fit.

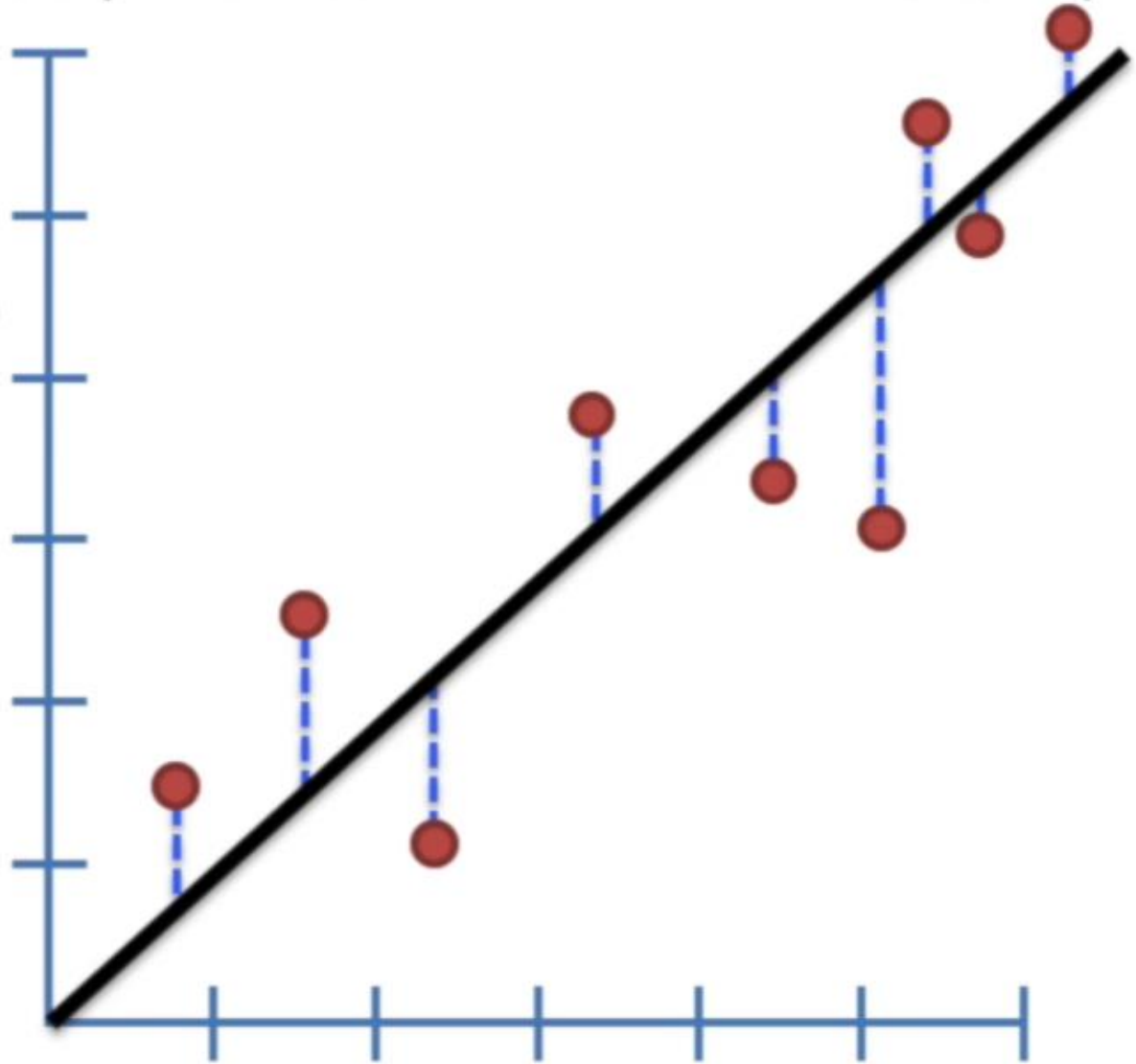


Now go back to the original plot.

Sum up the squared residuals around our least-squares fit.

We'll call this **SS(fit)**, for the sum of squares around the least-squares fit.

$SS(fit) = (data - line)^2$

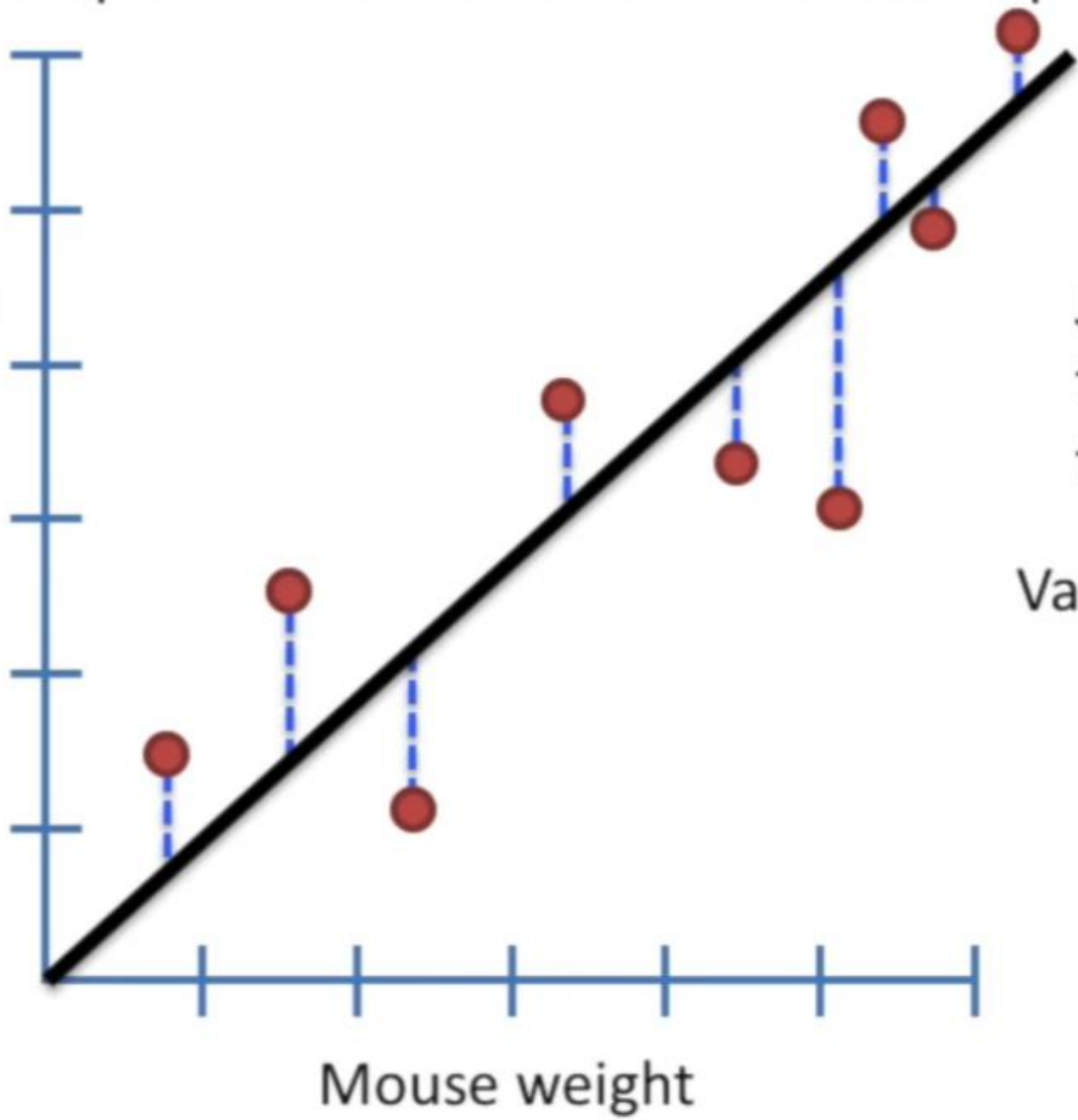


Mouse weight

Now go back to the original plot.
Sum up the squared residuals around our least-squares fit.

We'll call this **SS(fit)**, for the sum of squares around the least-squares fit.

$SS(fit) = (data - line)^2$



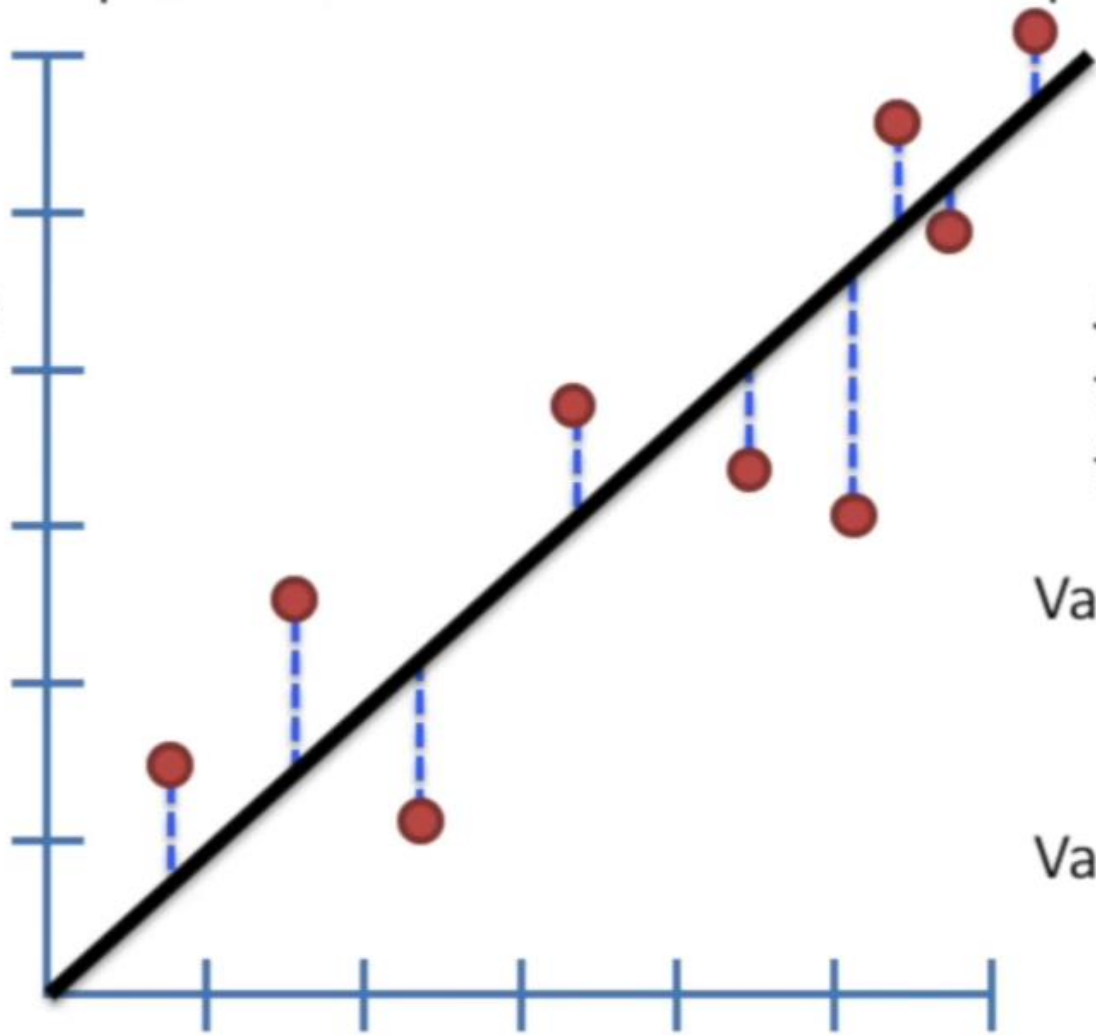
Just like with the mean, the variance around the fit…

$$Var(fit) = \frac{(data - line)^2}{n}$$

Now go back to the original plot.
Sum up the squared residuals around our least-squares fit.

We'll call this **SS(fit)**, for the sum of squares around the least-squares fit.

$SS(fit) = (data - line)^2$



Just like with the mean, the variance around the fit…

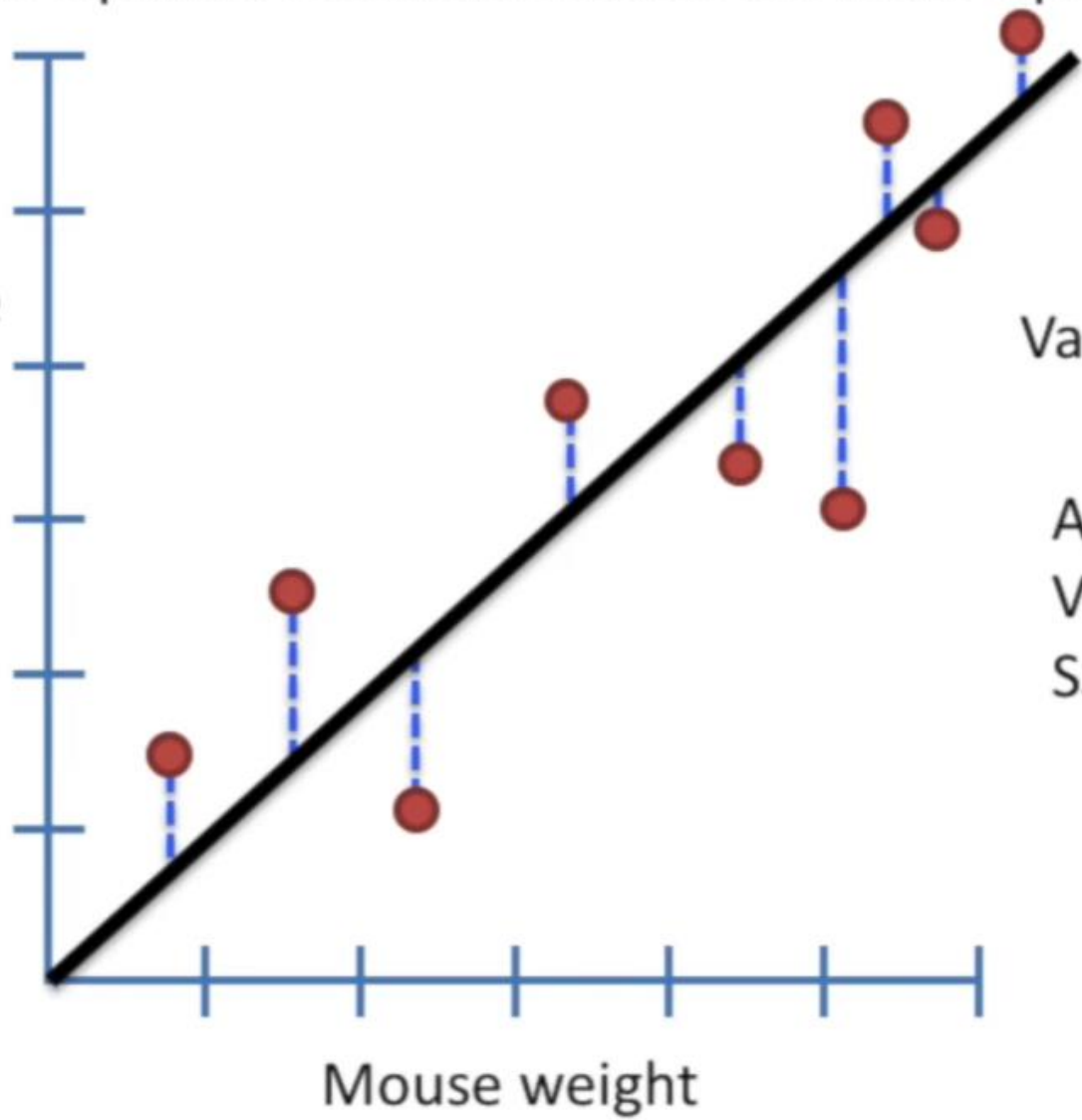
$$Var(fit) = \frac{(data - line)^2}{n}$$

$$Var(fit) = \frac{SS(fit)}{n}$$

Now go back to the original plot.
Sum up the squared residuals around our least-squares fit.

We'll call this **SS(fit)**, for the sum of squares around the least-squares fit.

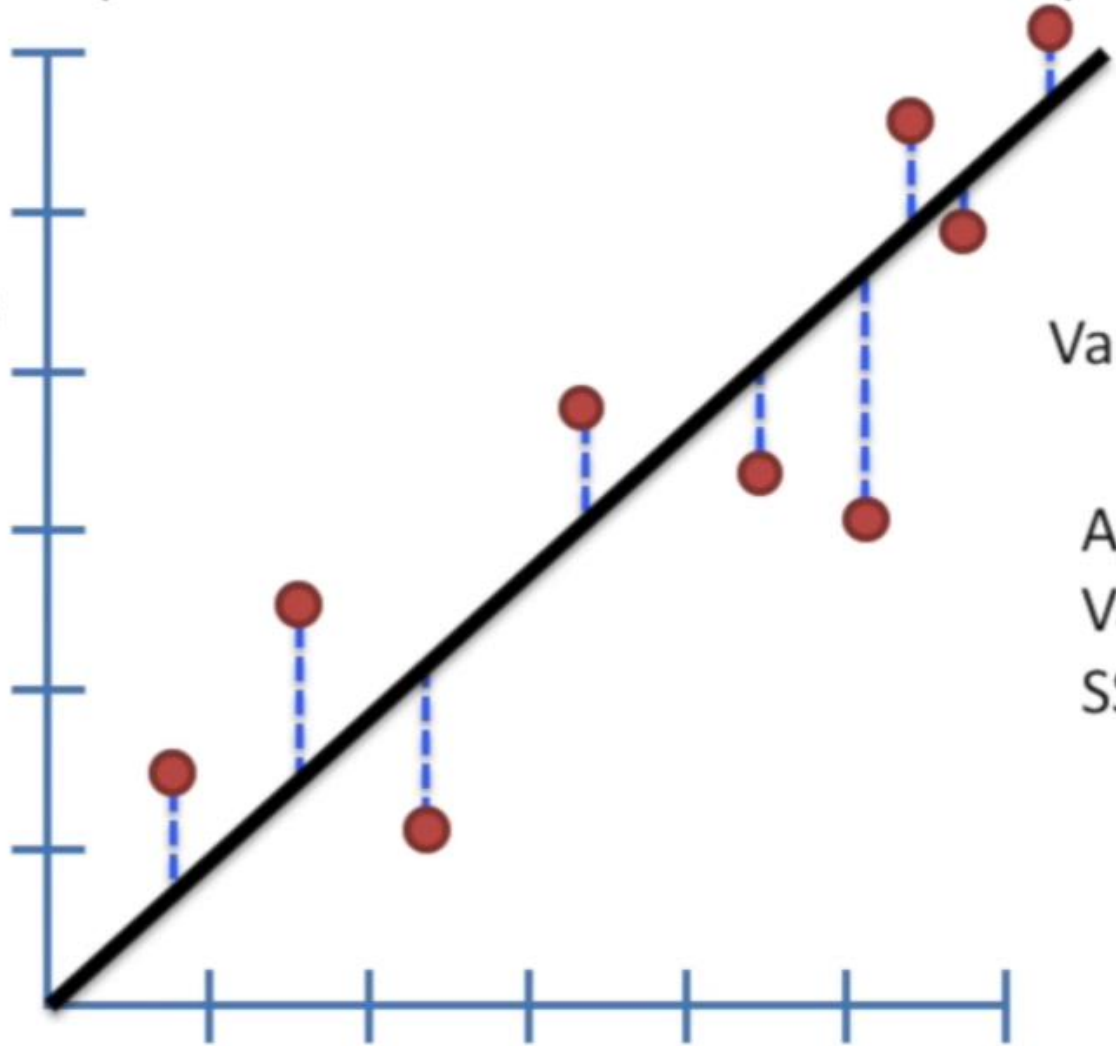
$SS(fit) = (data - line)^2$



$$Var(fit) = \frac{SS(fit)}{n}$$

Again, we can think of Var(fit) as the average SS(fit) for each mouse.

Now go back to the original plot.
Sum up the squared residuals around our least-squares fit.



We'll call this **SS(fit)**, for the sum of squares around the least-squares fit.

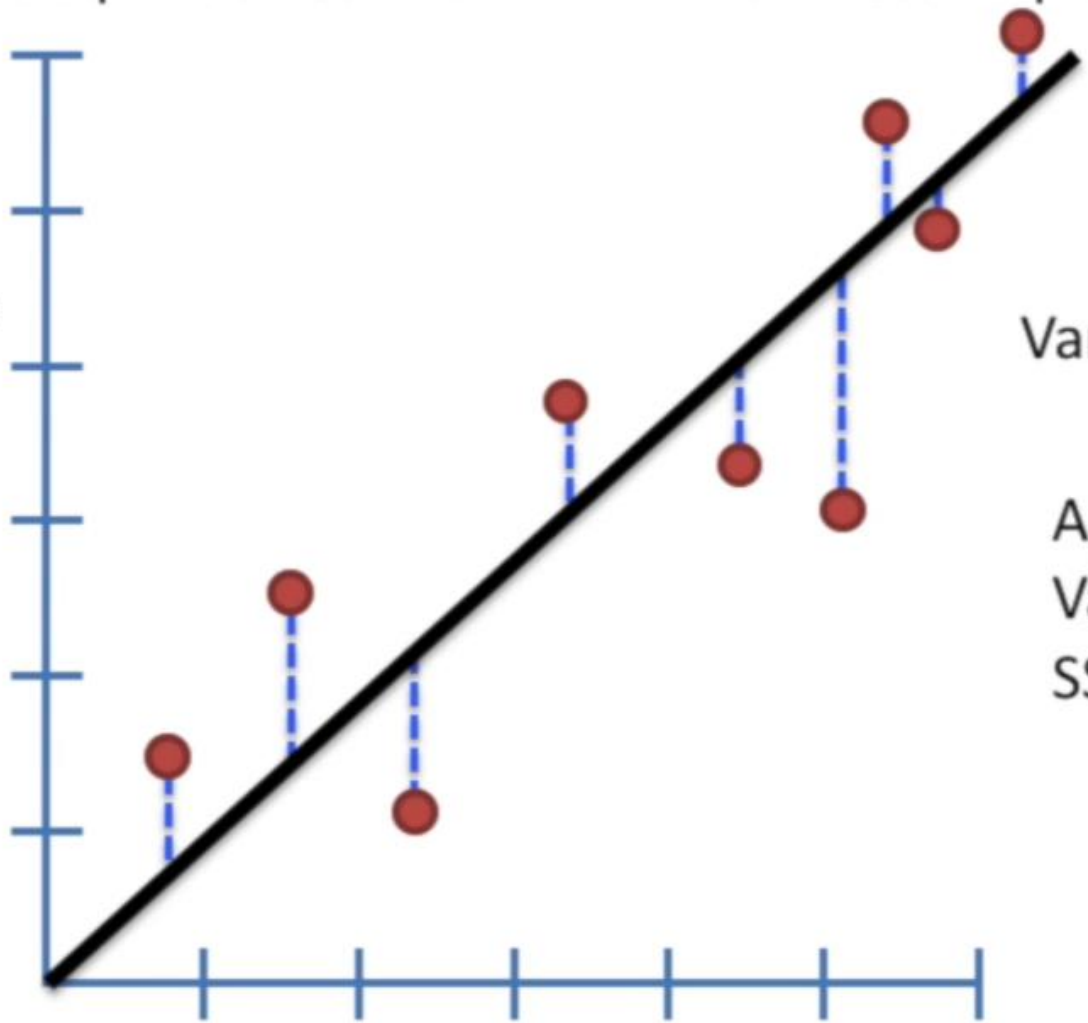
$SS(fit) = (data - line)^2$

$$Var(fit) = \frac{SS(fit)}{n}$$

Again, we can think of Var(fit) as the average SS(fit) for each mouse.

$$\text{In general: Variance(something)} = \frac{\text{Sums of squares}}{\text{The number of those things}}$$

Now go back to the original plot.
Sum up the squared residuals around our least-squares fit.



We'll call this **SS(fit)**, for the sum of squares around the least-squares fit.

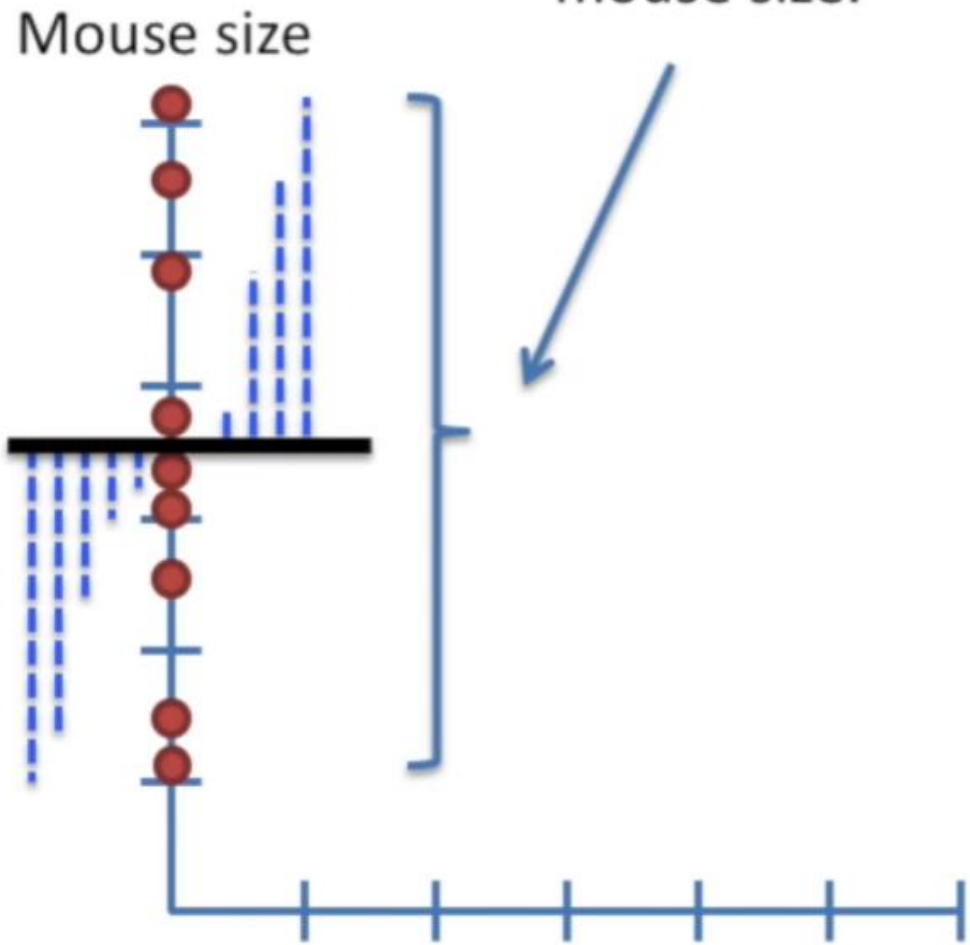
$SS(fit) = (data - line)^2$

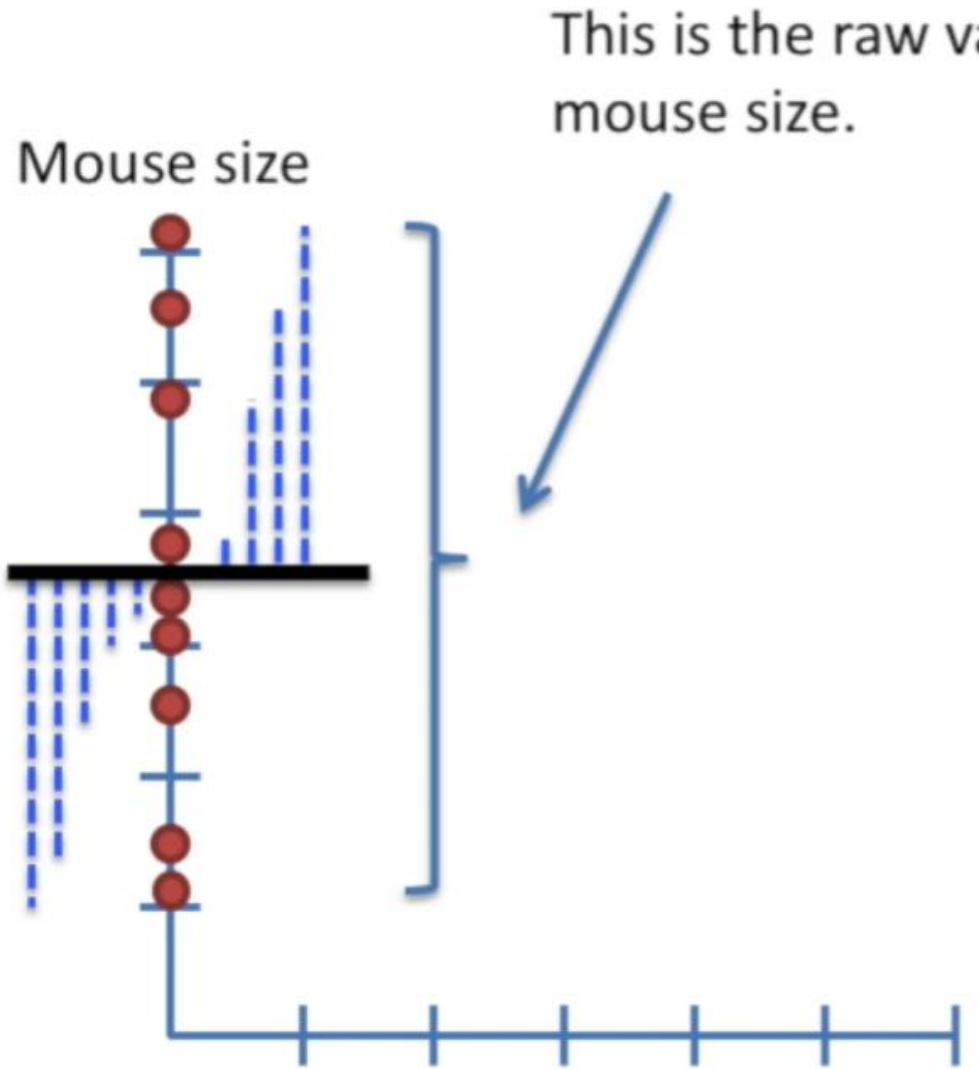
$$Var(fit) = \frac{SS(fit)}{n}$$

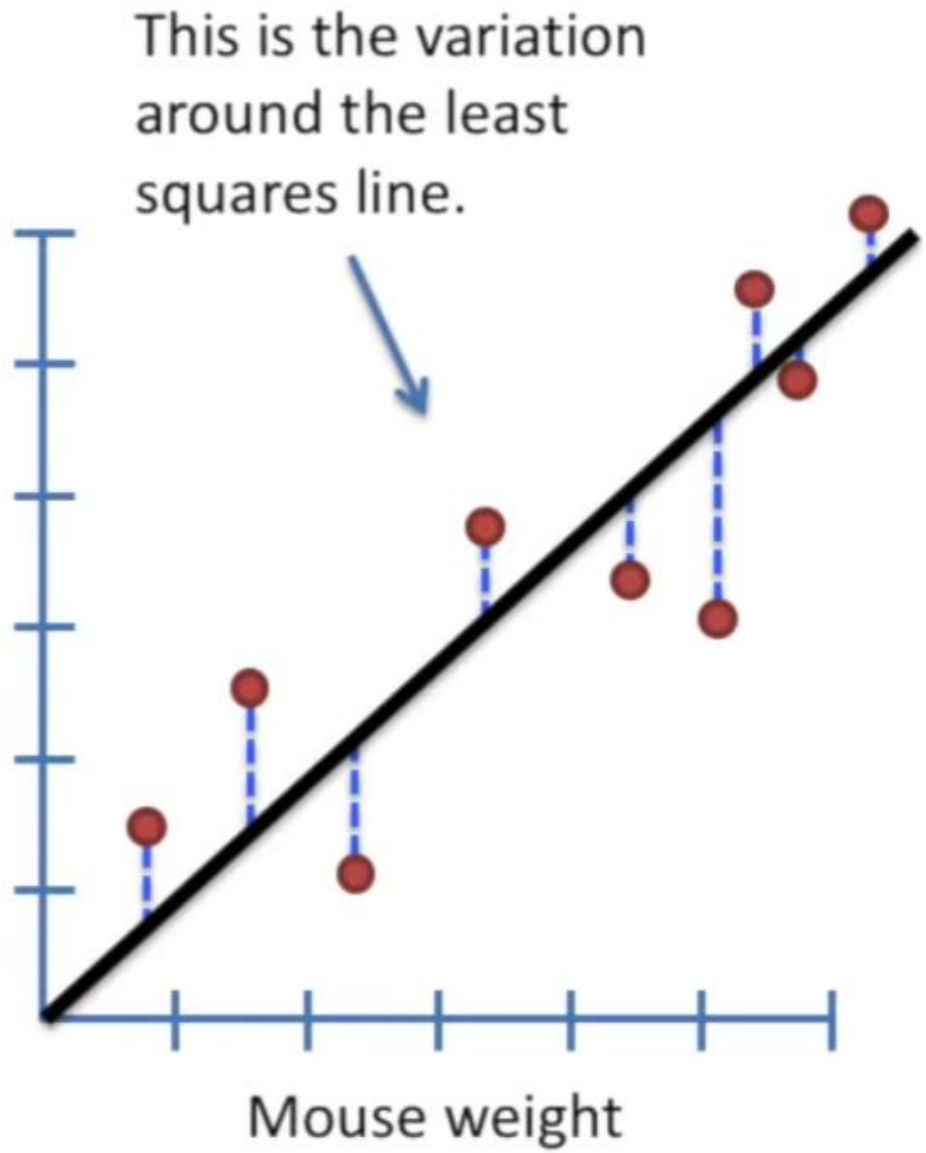
Again, we can think of Var(fit) as the average SS(fit) for each mouse.

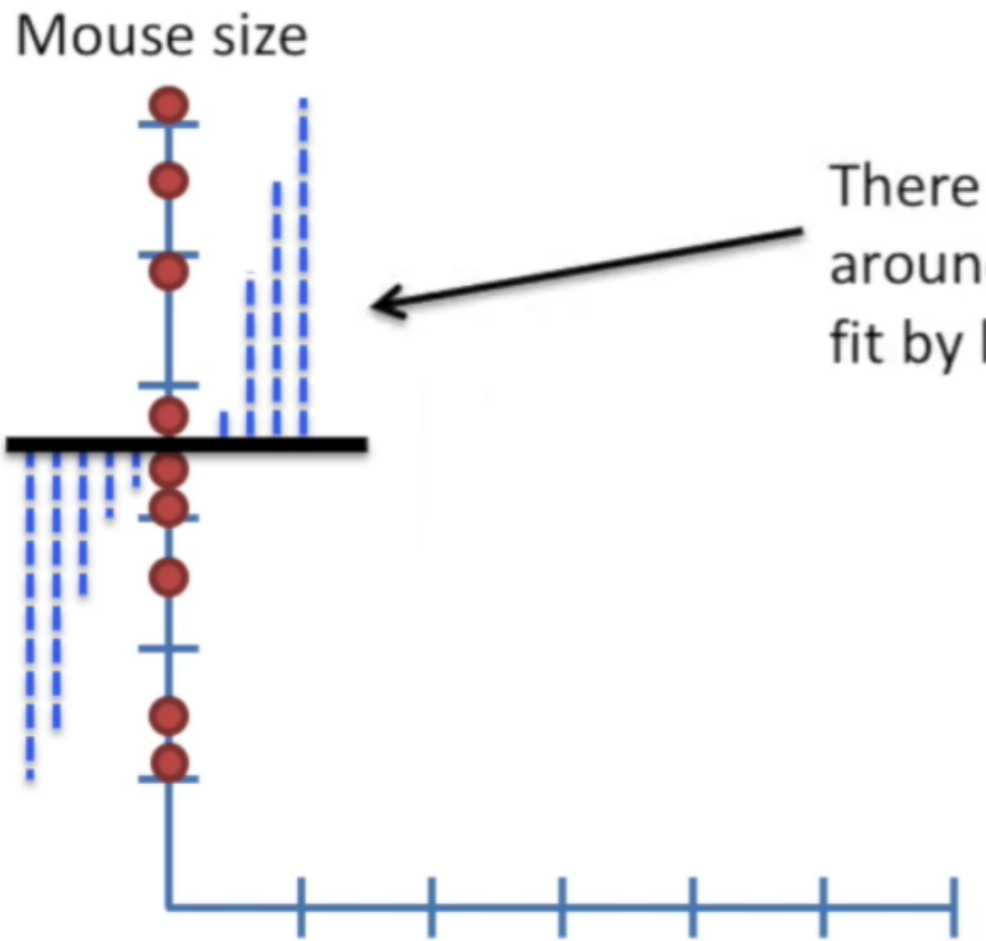
$$\text{In general: Variance(something)} = \frac{\text{Sums of squares}}{\text{The number of those things}} = \text{Average sum of squares.}$$

This is the raw variation in mouse size.

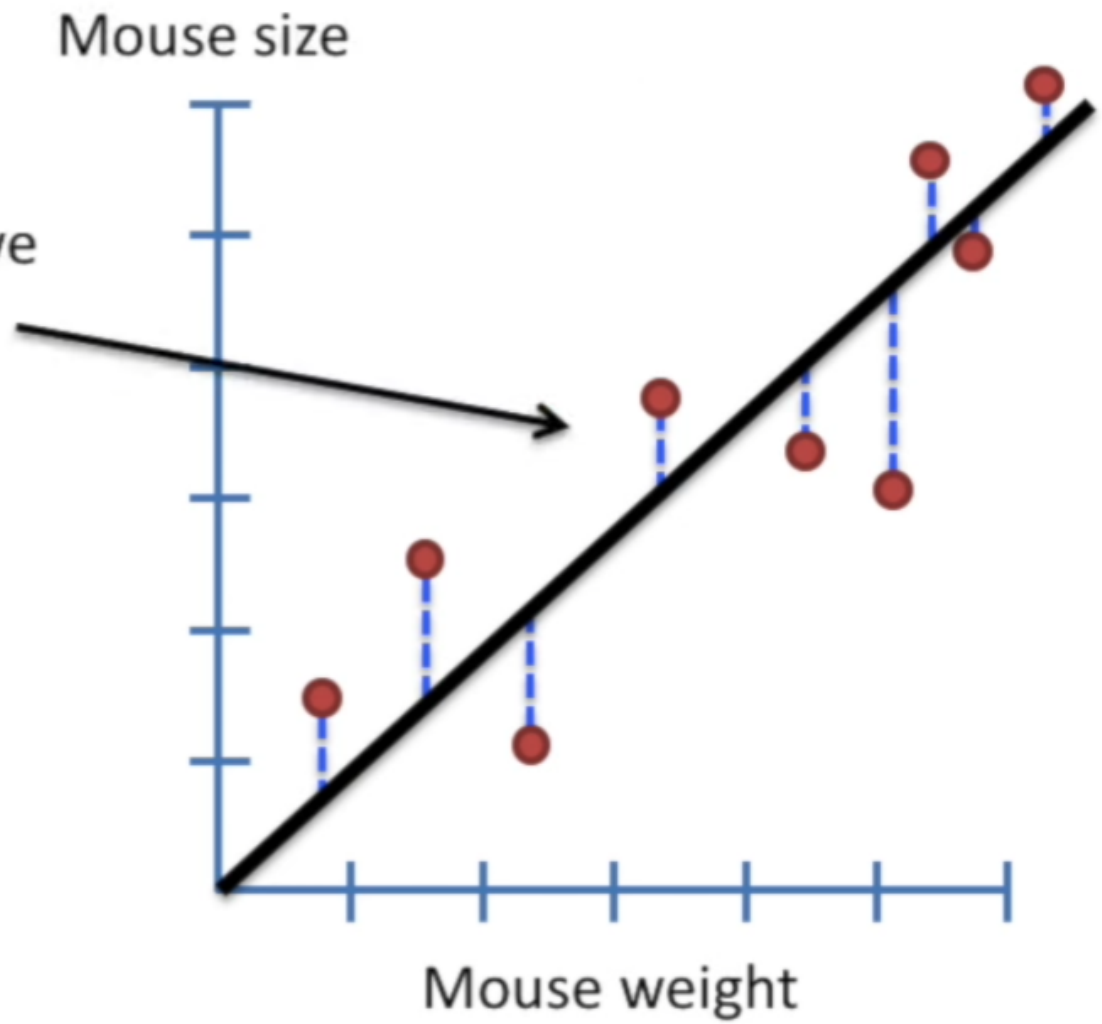


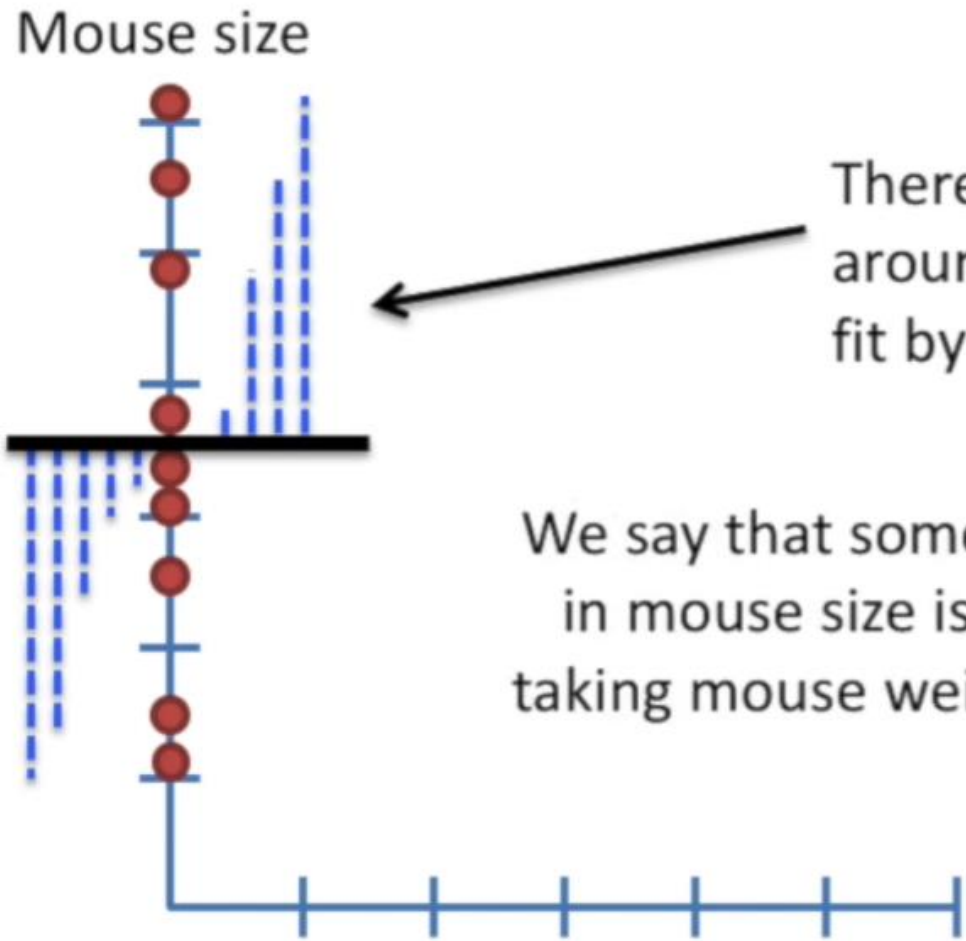

This is the raw variation in mouse size.
Mouse size


This is the variation around the least squares line.
Mouse weight



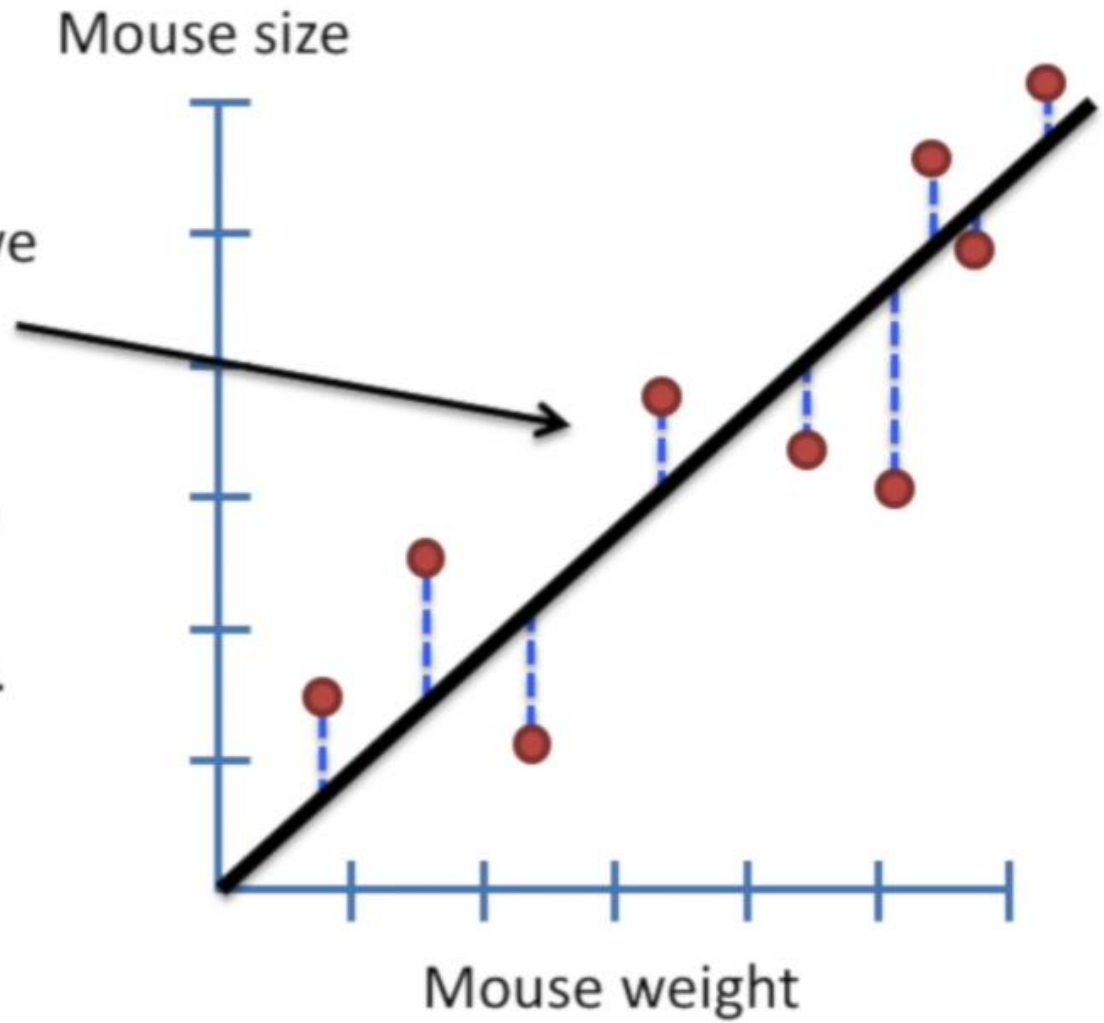
There is less variation around the line that we fit by least-squares.

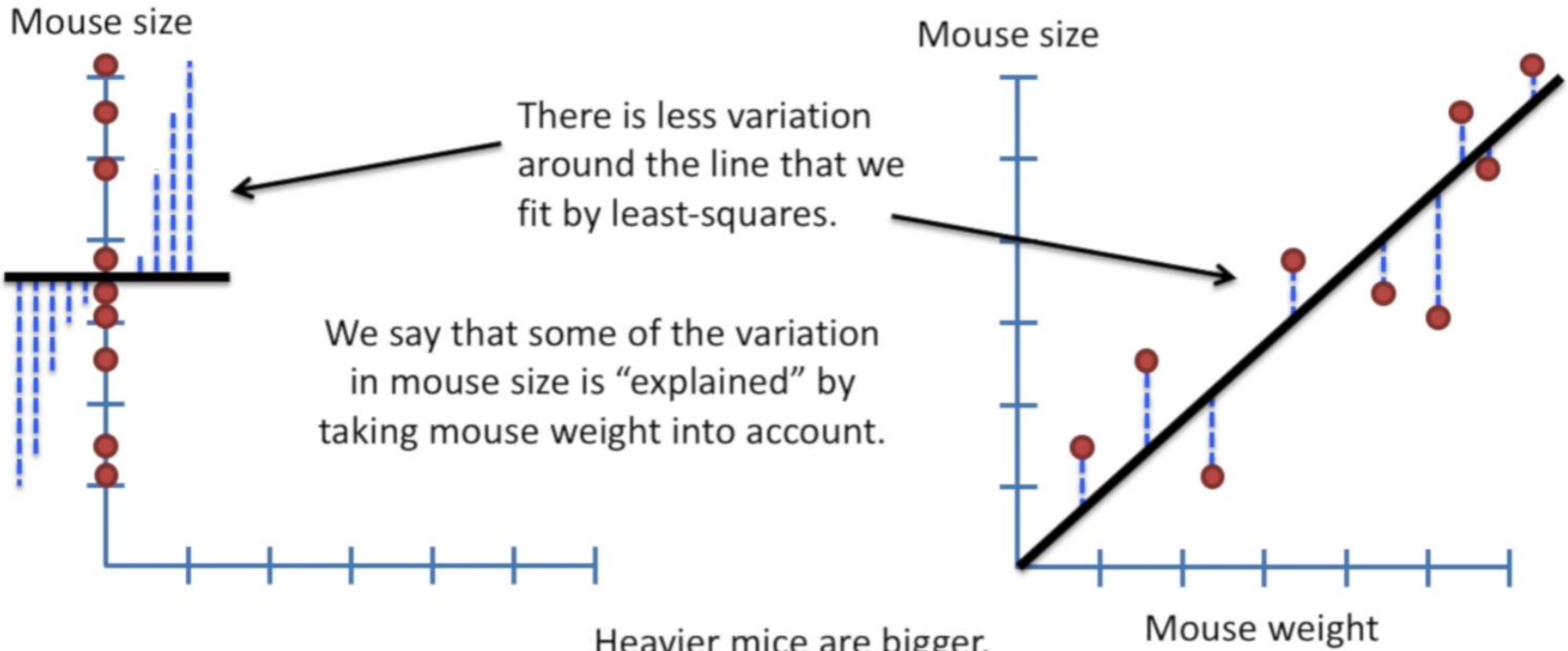


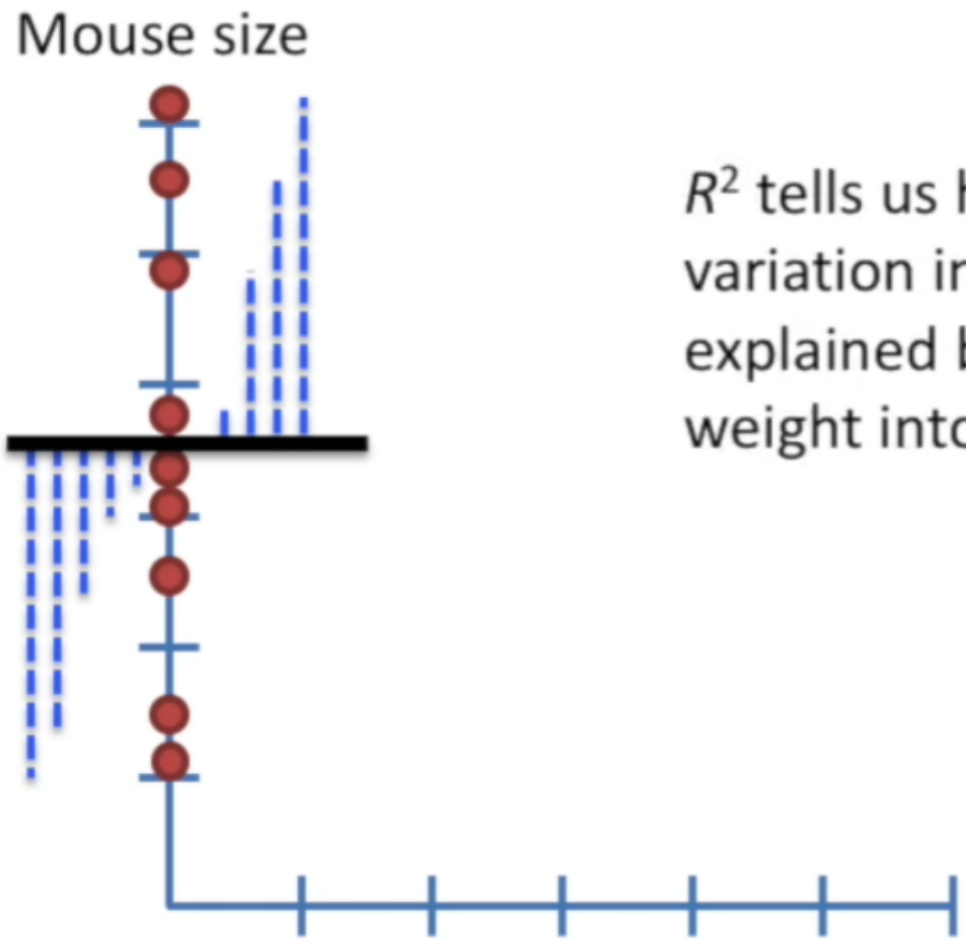


There is less variation around the line that we fit by least-squares.

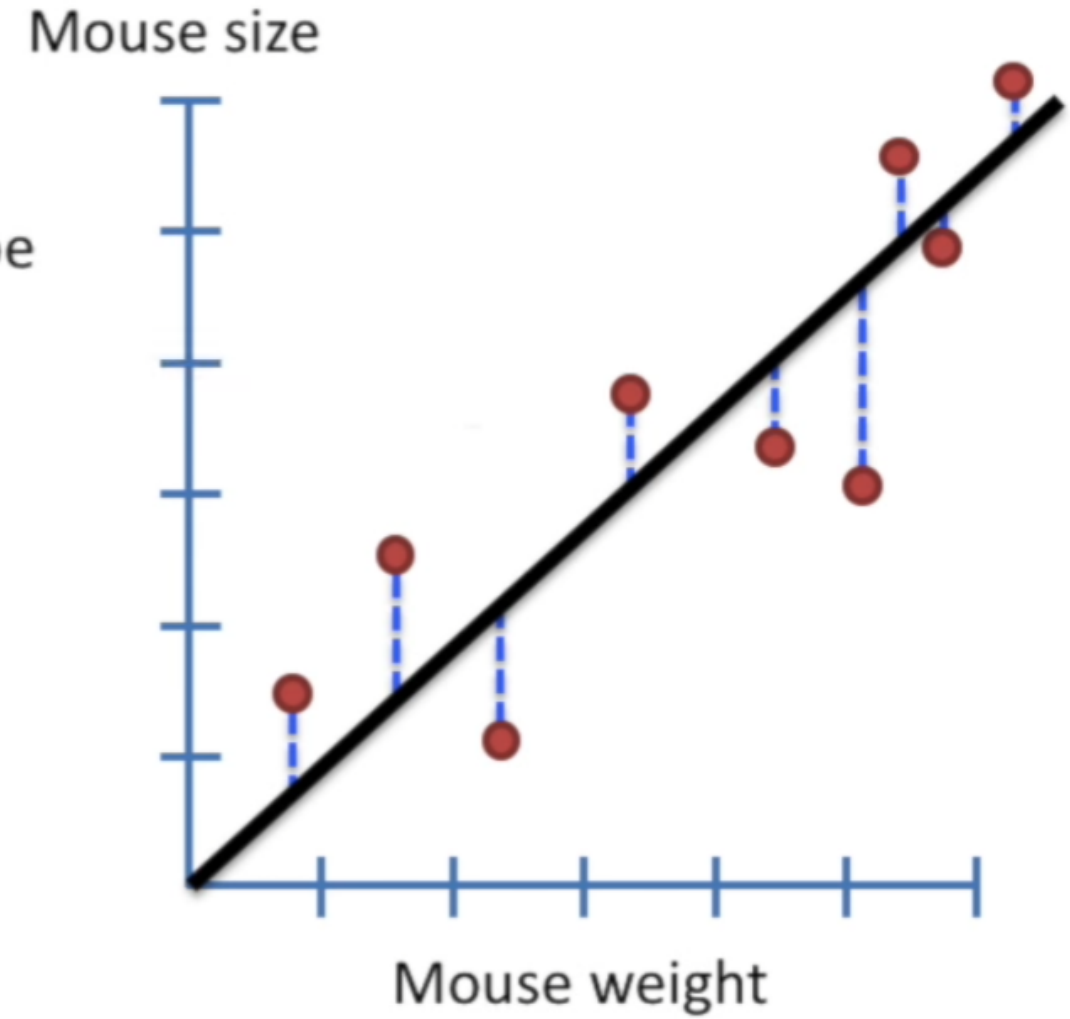
We say that some of the variation in mouse size is “explained” by taking mouse weight into account.

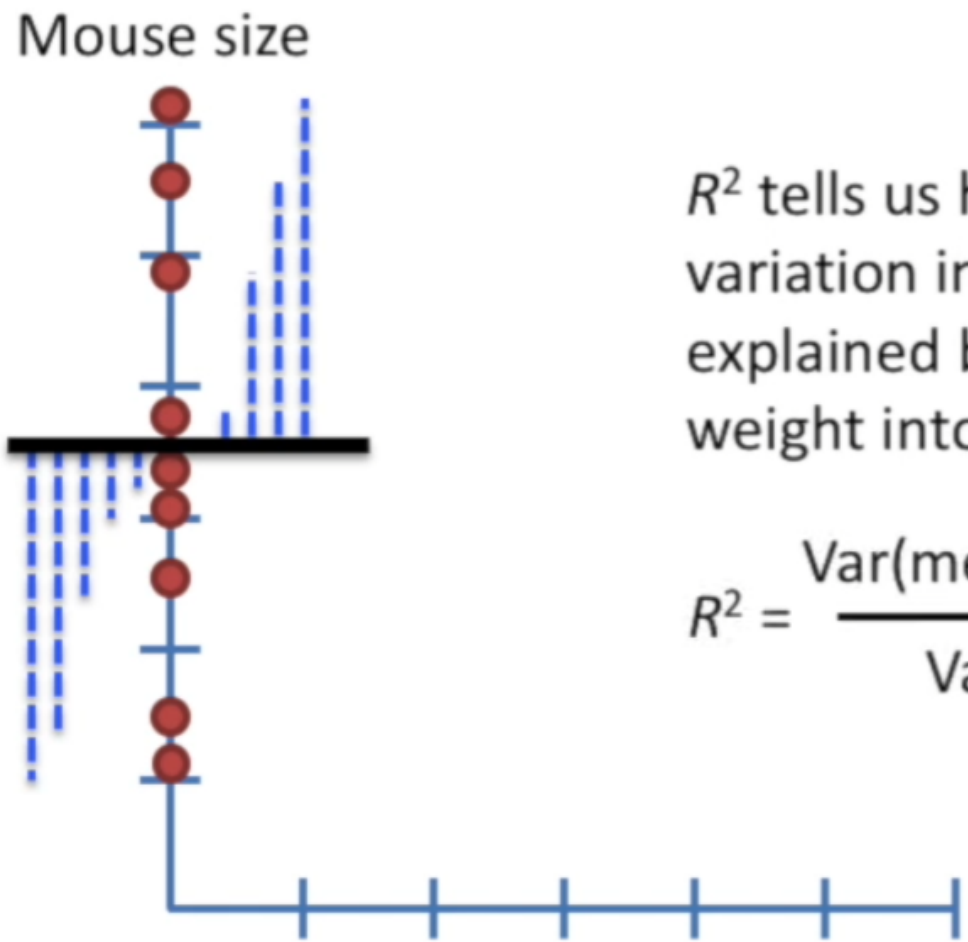



Mouse size
There is less variation around the line that we fit by least-squares.
Mouse size
We say that some of the variation in mouse size is “explained” by taking mouse weight into account.
Mouse weight
Heavier mice are bigger.
Lighter mice are smaller.



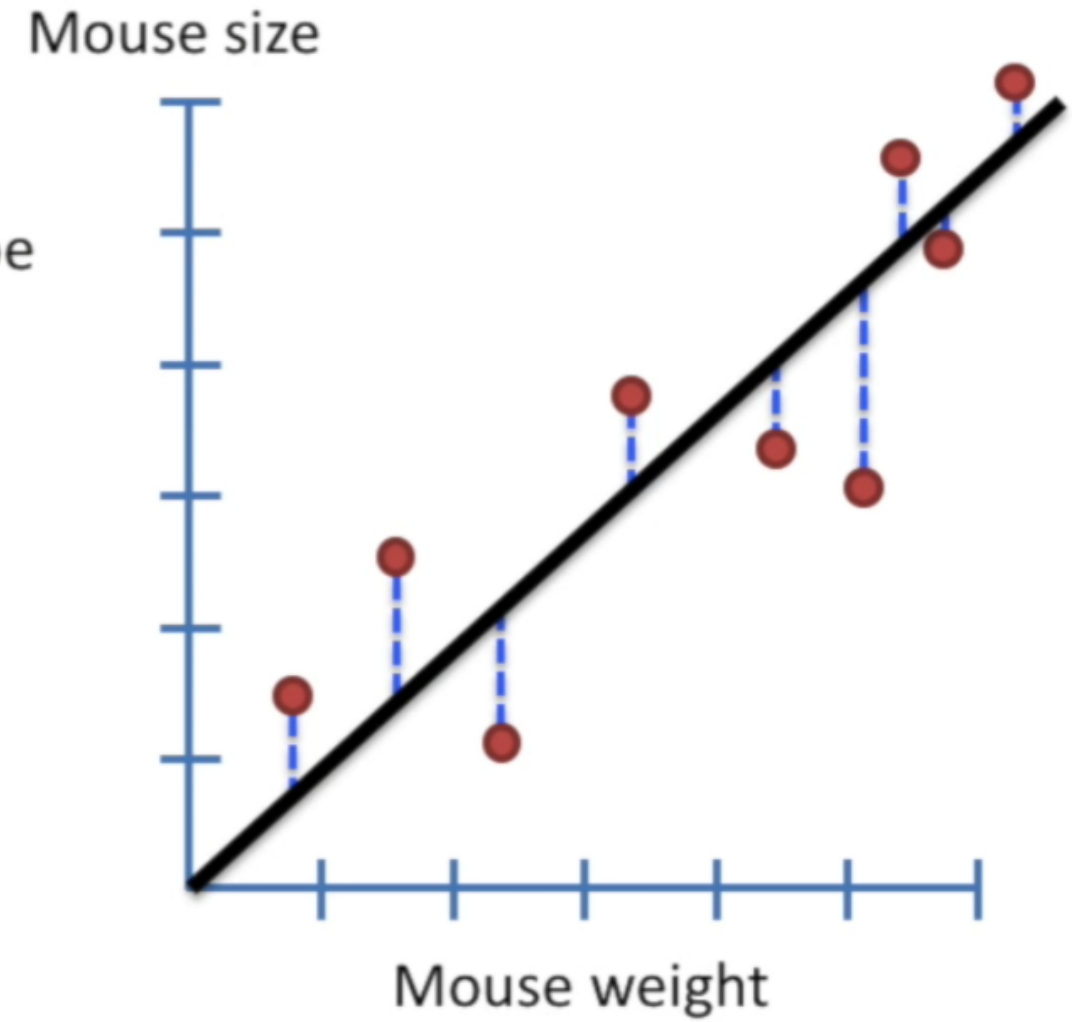
R^2 tells us how much of the variation in mouse size can be explained by taking mouse weight into account.





R^2 tells us how much of the variation in mouse size can be explained by taking mouse weight into account.

$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

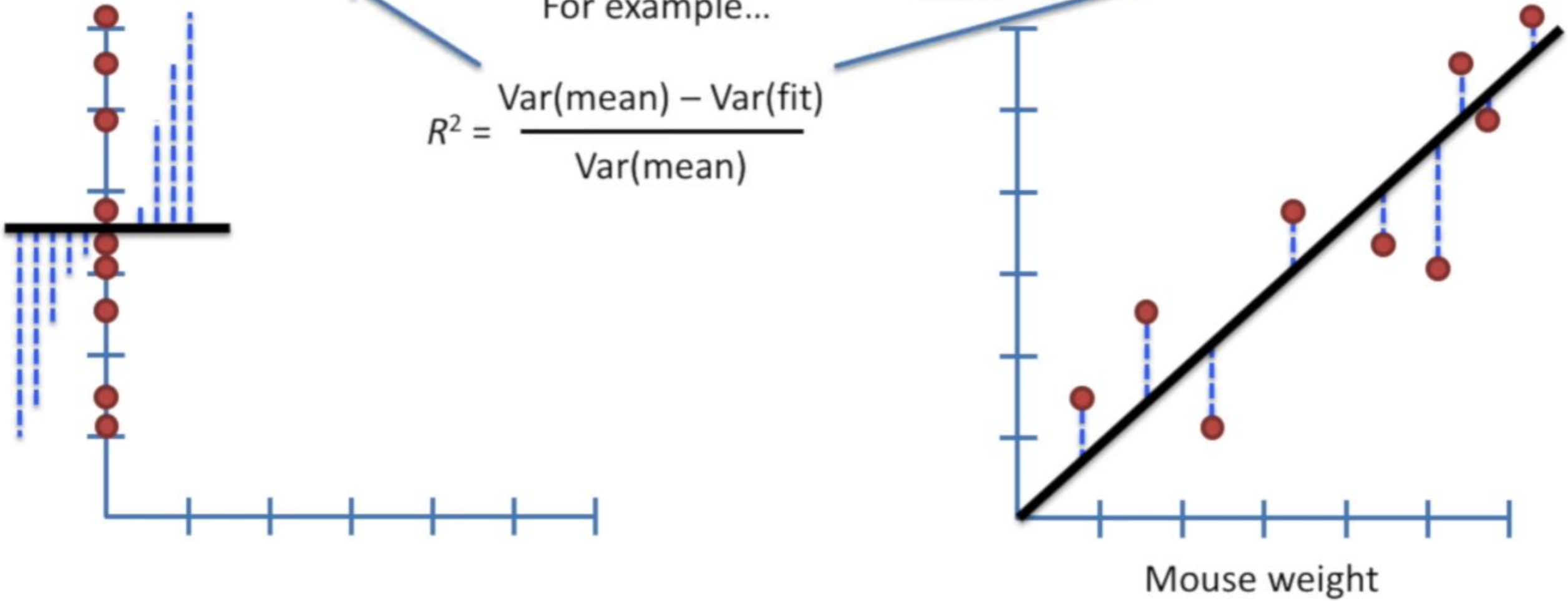


Var(mean) = 11.1

For example…

$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

Var(fit) = 4.4



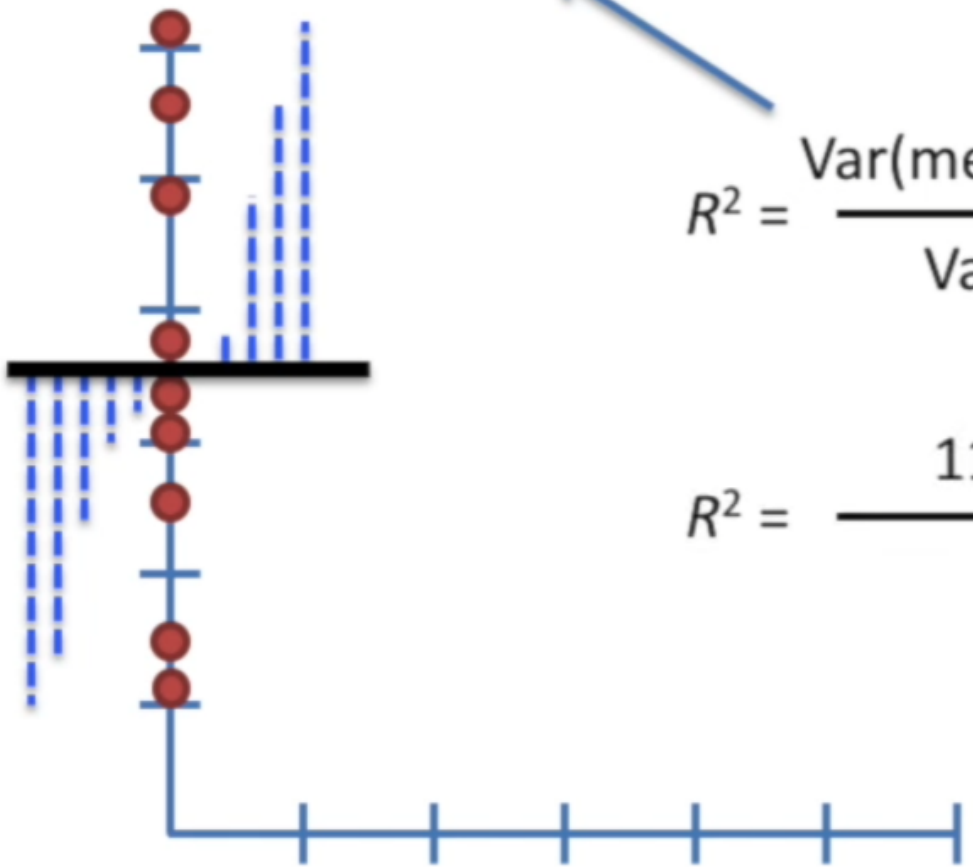
Var(mean) = 11.1

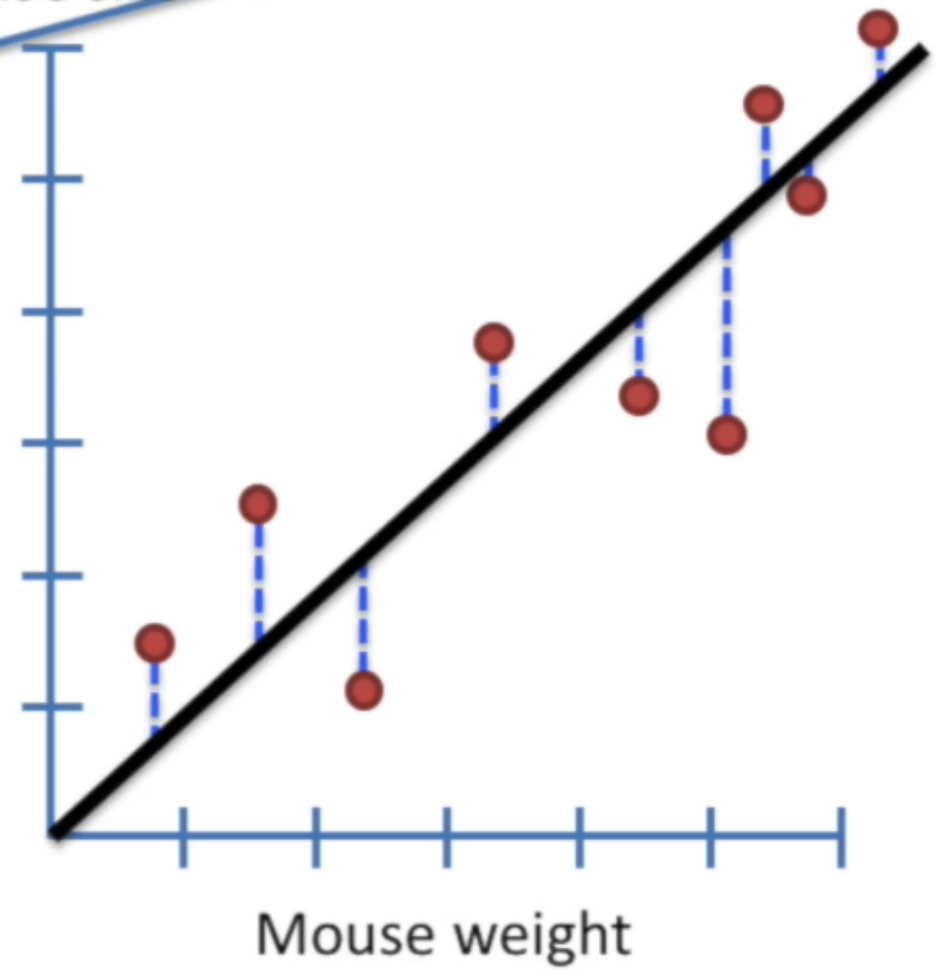
Var(fit) = 4.4

Mouse size

$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

$$R^2 = \frac{11.1 - 4.4}{11.1}$$





Mouse weight

Var(mean) = 11.1

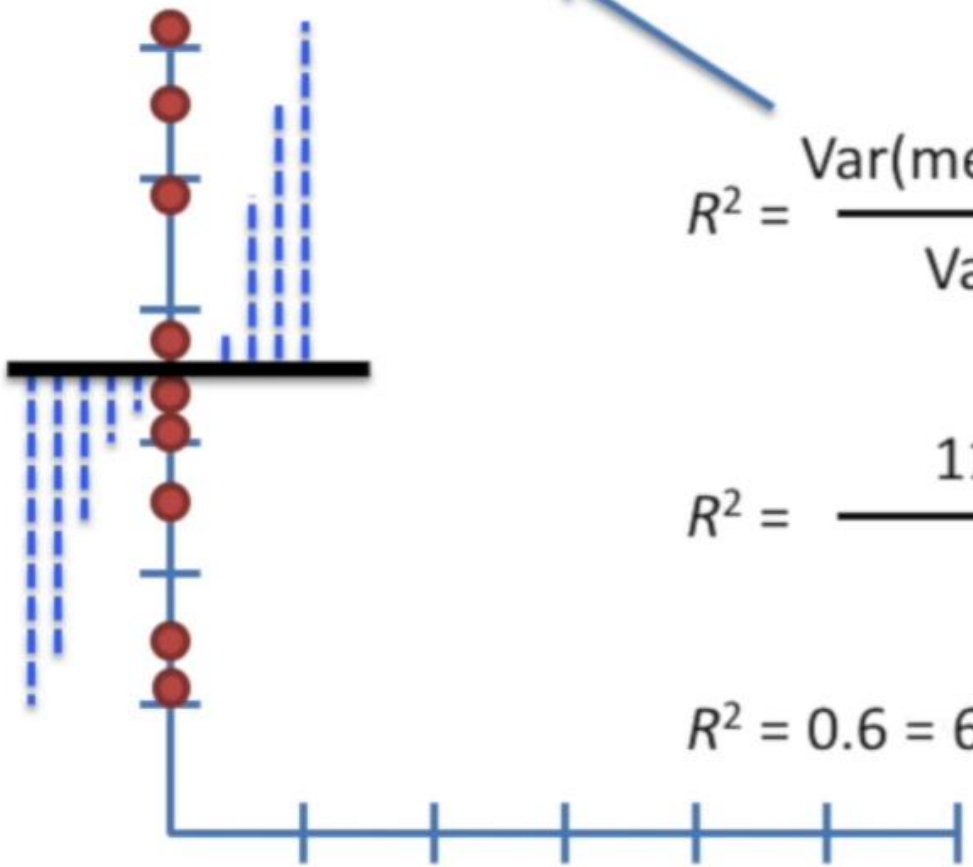
Var(fit) = 4.4

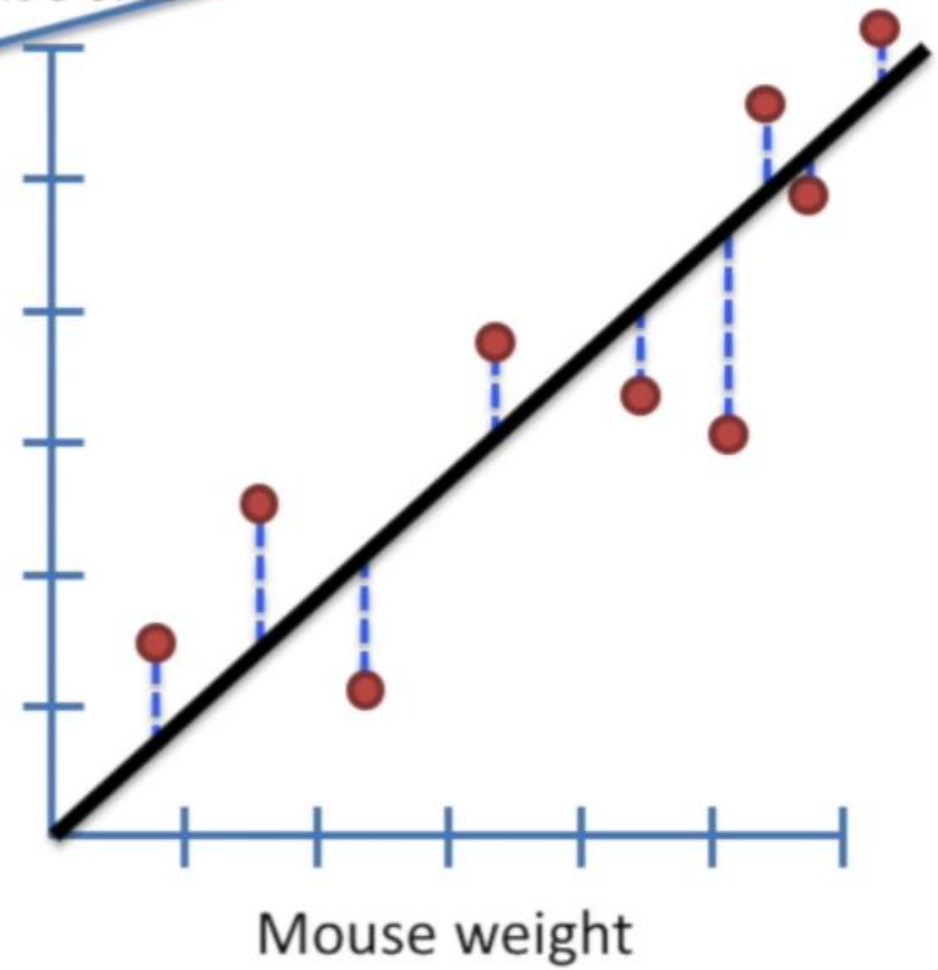
Mouse size

$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

$$R^2 = \frac{11.1 - 4.4}{11.1}$$

$R^2 = 0.6 = 60\%$





Mouse weight

Var(mean) = 11.1

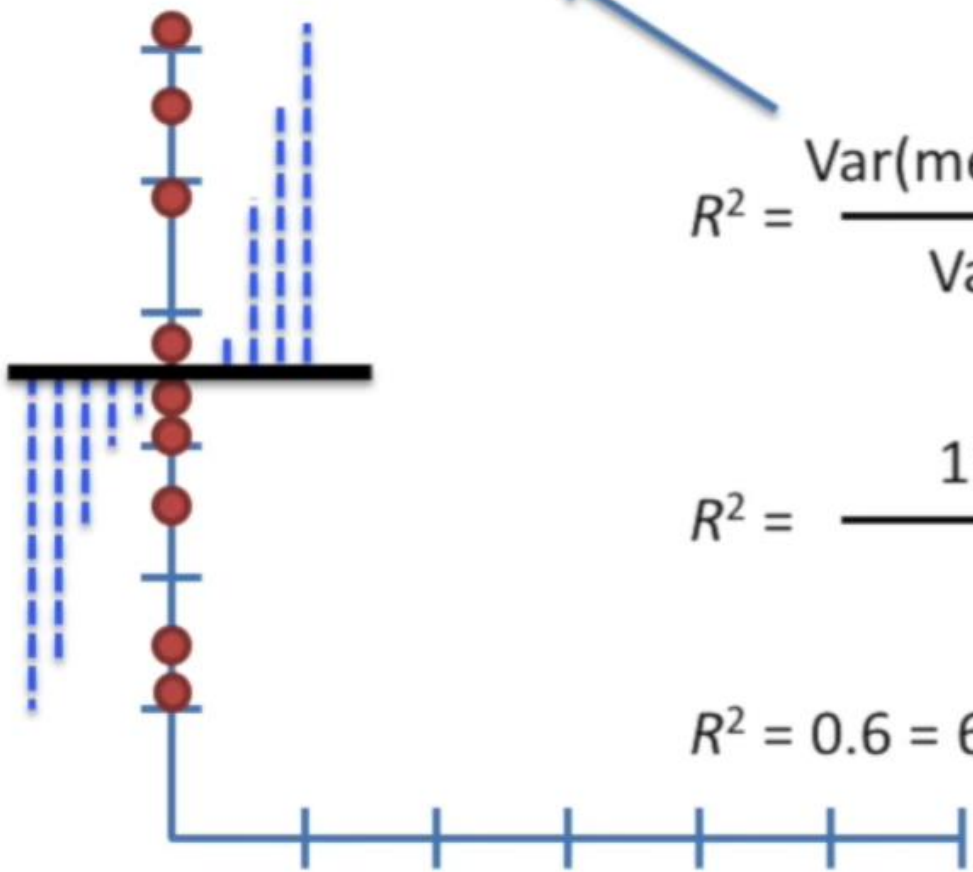
Var(fit) = 4.4

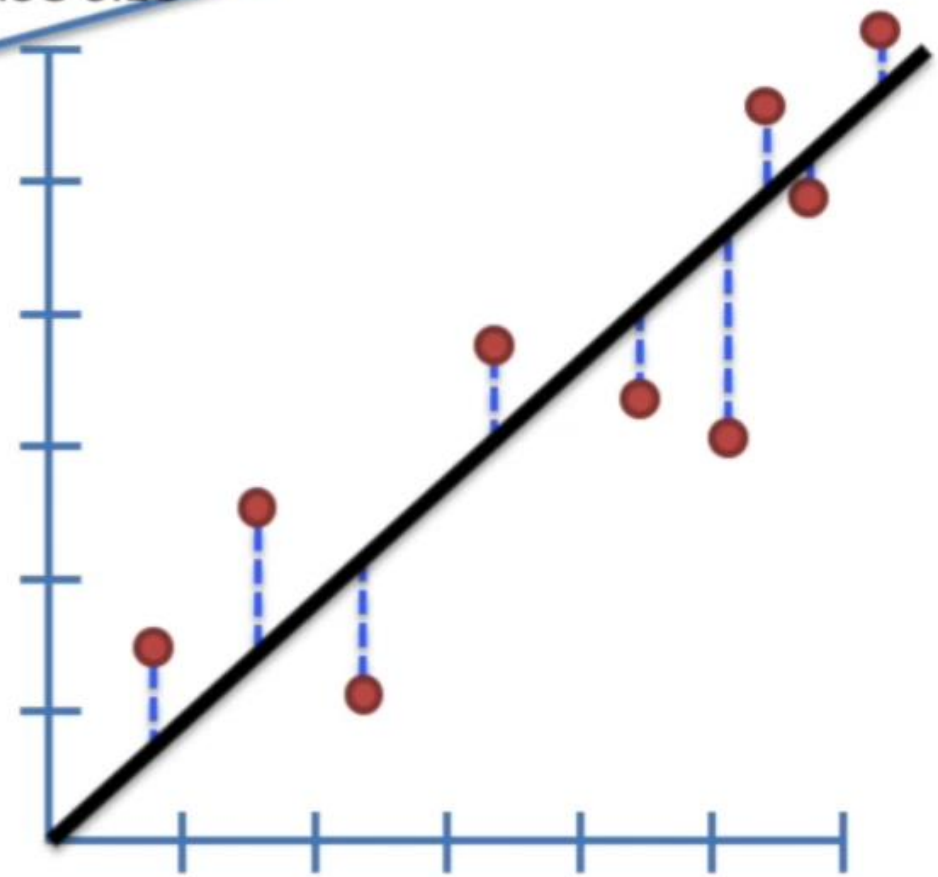
Mouse size

$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

$$R^2 = \frac{11.1 - 4.4}{11.1}$$

$R^2 = 0.6 = 60\%$

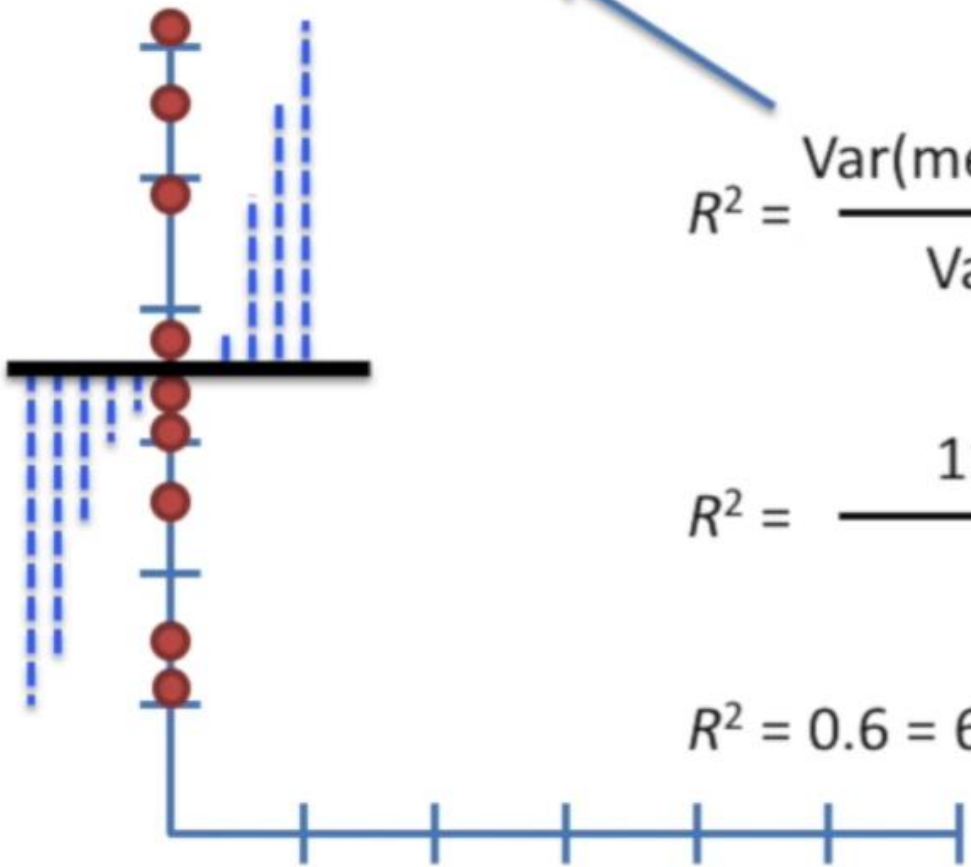




Mouse weight

There is a 60% reduction in variance when we take the mouse weight into account.

Var(mean) = 11.1



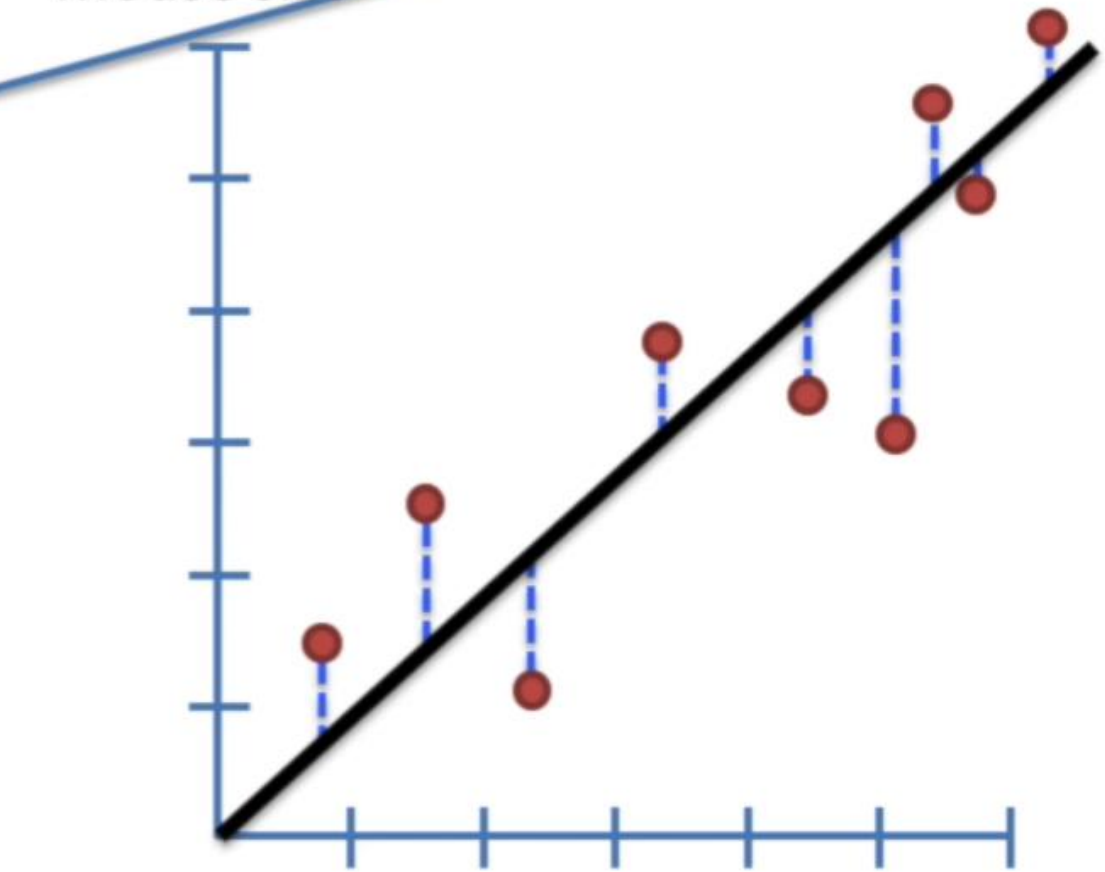
$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

$$R^2 = \frac{11.1 - 4.4}{11.1}$$

$R^2 = 0.6 = 60\%$

There is a 60% reduction in variance when we take the mouse weight into account.

Var(fit) = 4.4



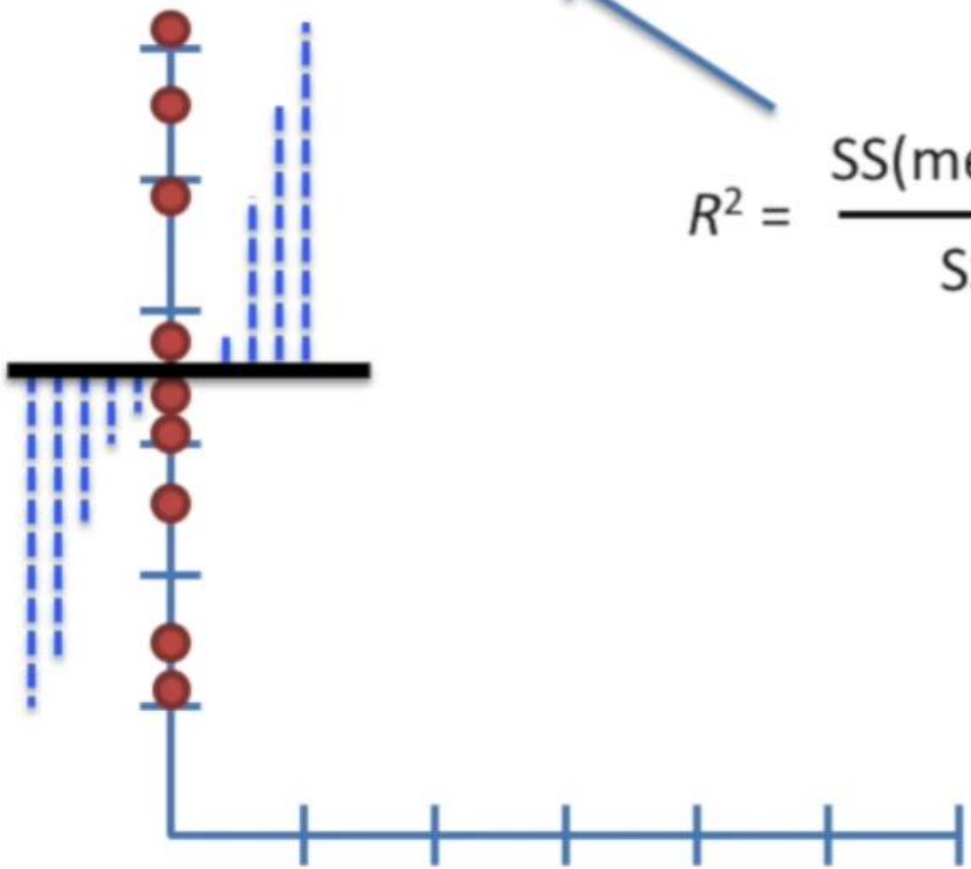
Alternatively, we can say that mouse weight "explains" 60% of the variation in mouse size.

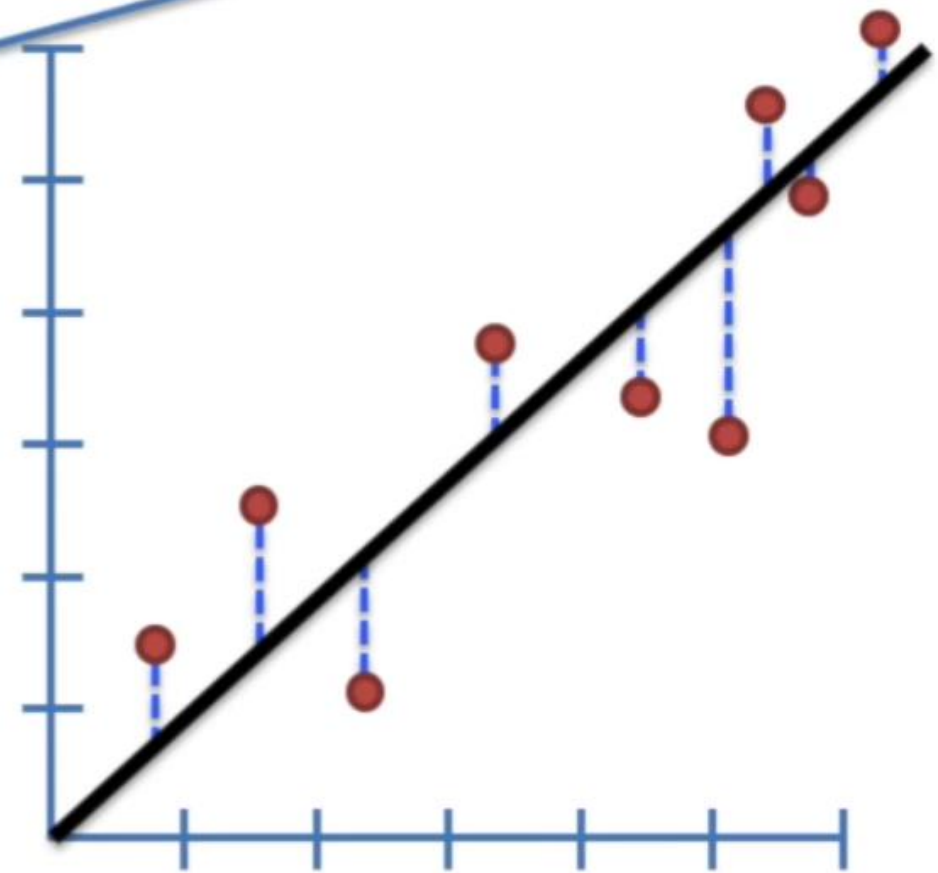
We can also use the sums of squares to make the same calculation.

SS(mean) = 100

SS(fit) = 40

$$R^2 = \frac{SS(mean) - SS(fit)}{SS(mean)}$$





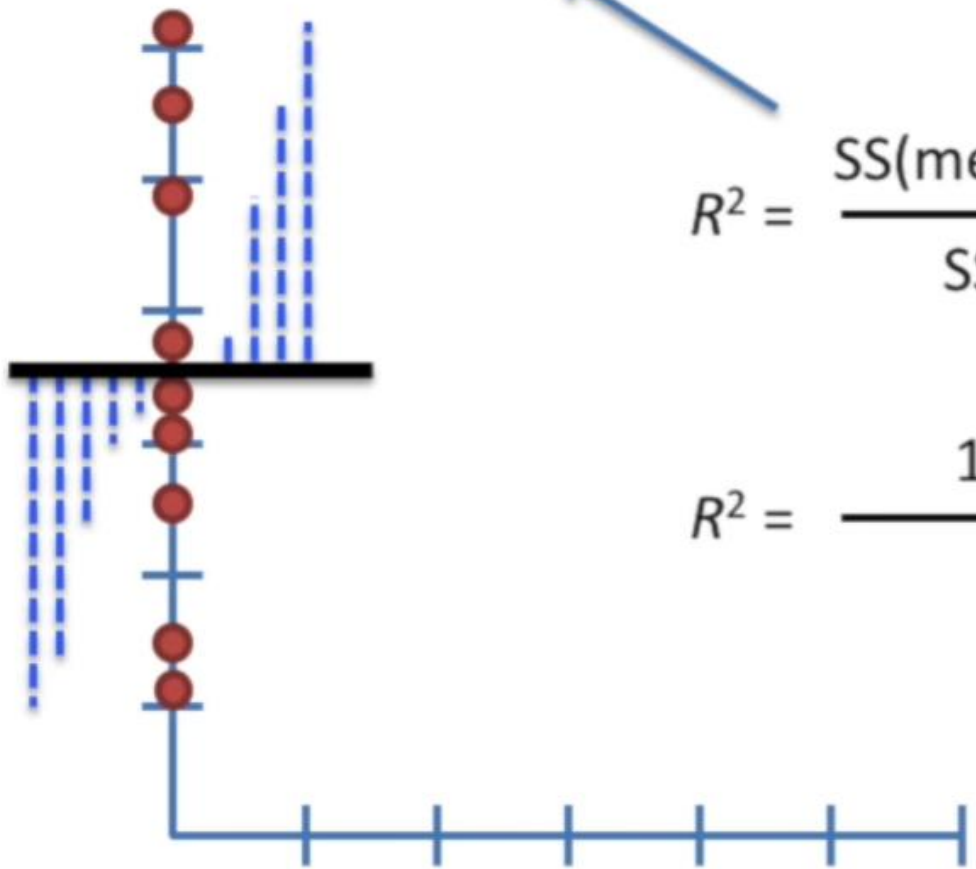
We can also use the sums of squares to make the same calculation.

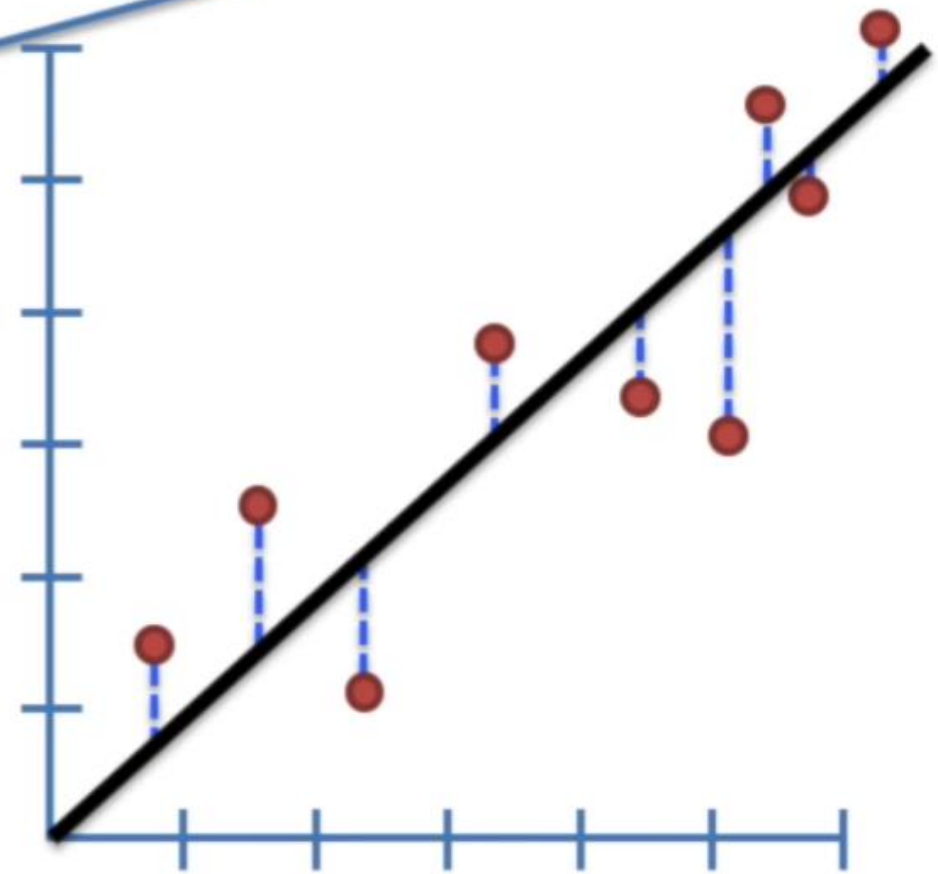
SS(mean) = 100

SS(fit) = 40

$$R^2 = \frac{SS(mean) - SS(fit)}{SS(mean)}$$

$$R^2 = \frac{100 - 40}{100}$$

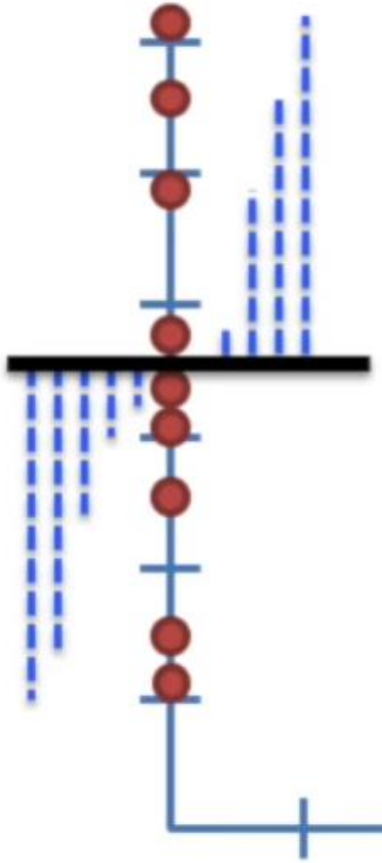




We can also use the sums of squares to make the same calculation.

SS(mean) = 100

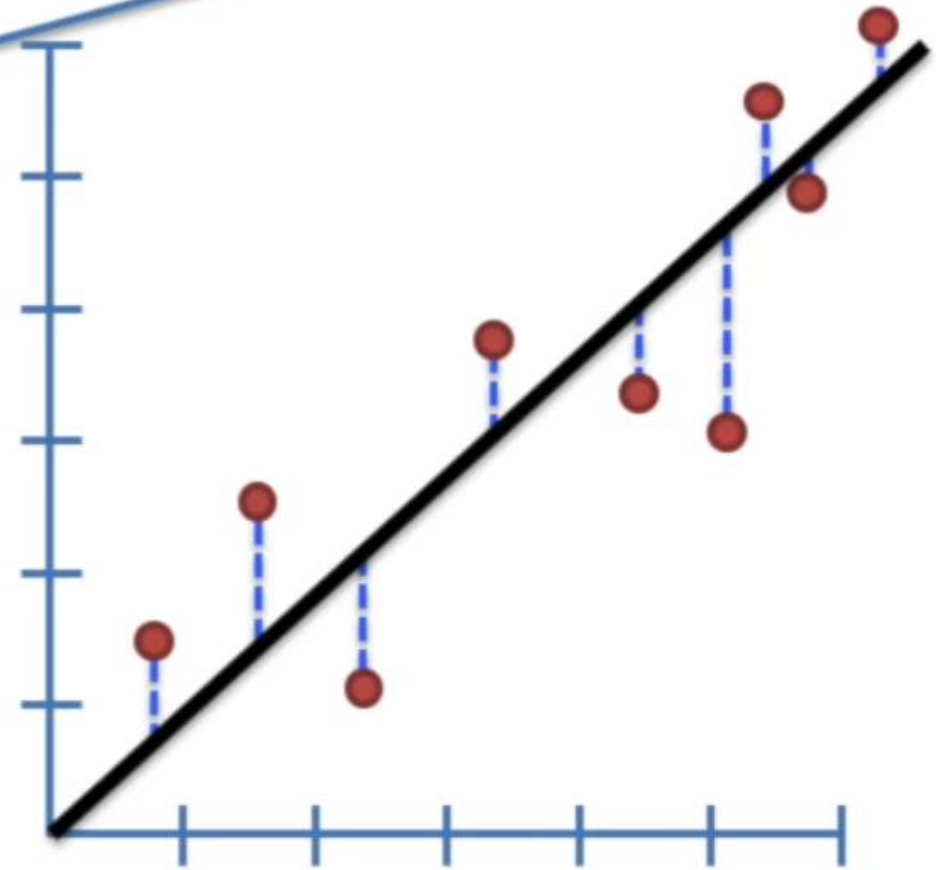
SS(fit) = 40



$$R^2 = \frac{SS(mean) - SS(fit)}{SS(mean)}$$

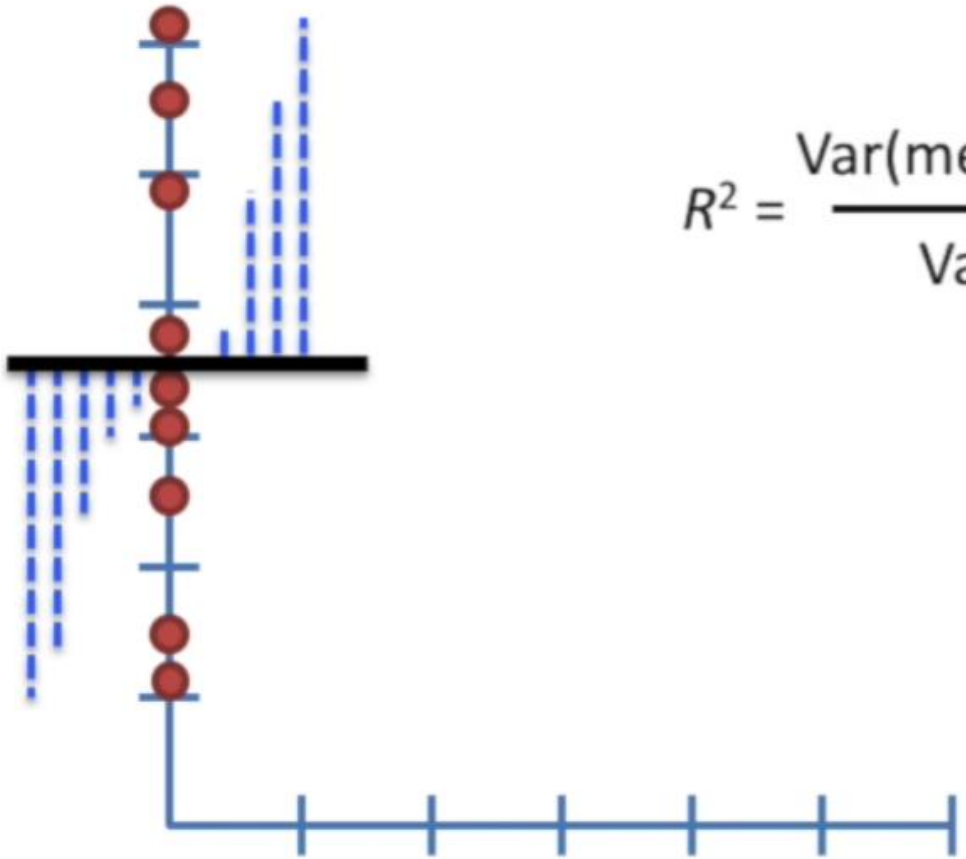
$$R^2 = \frac{100 - 40}{100}$$

$R^2 = 0.6 = 60\%$

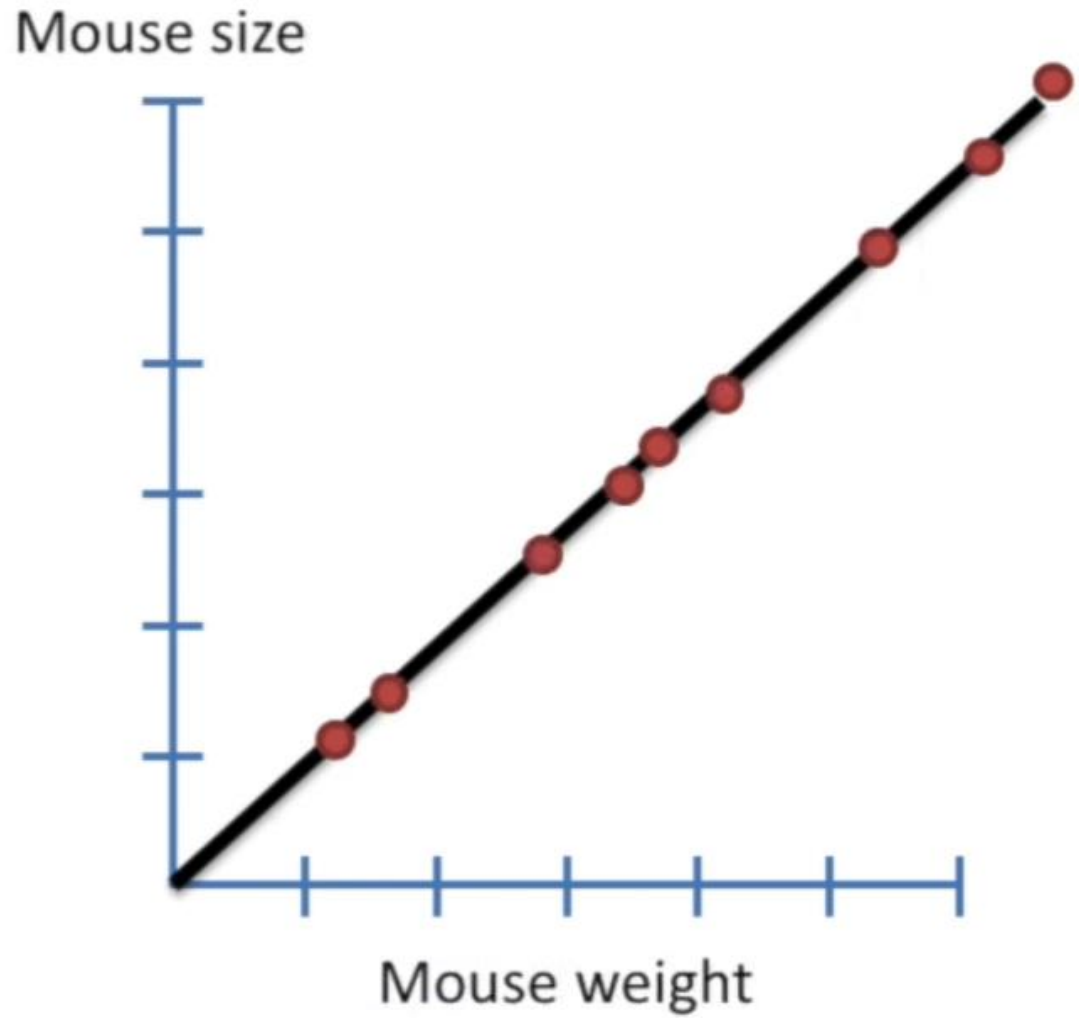


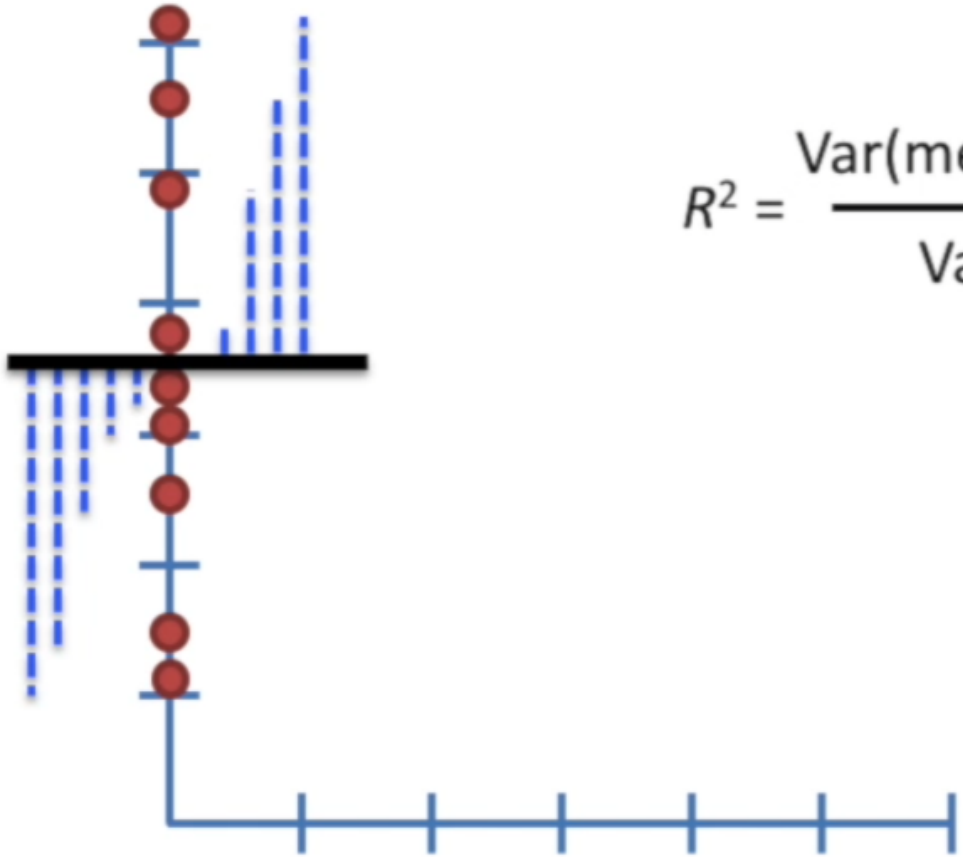
60% of the sums of squares of the mouse size can be explained by mouse weight..

Another Example…



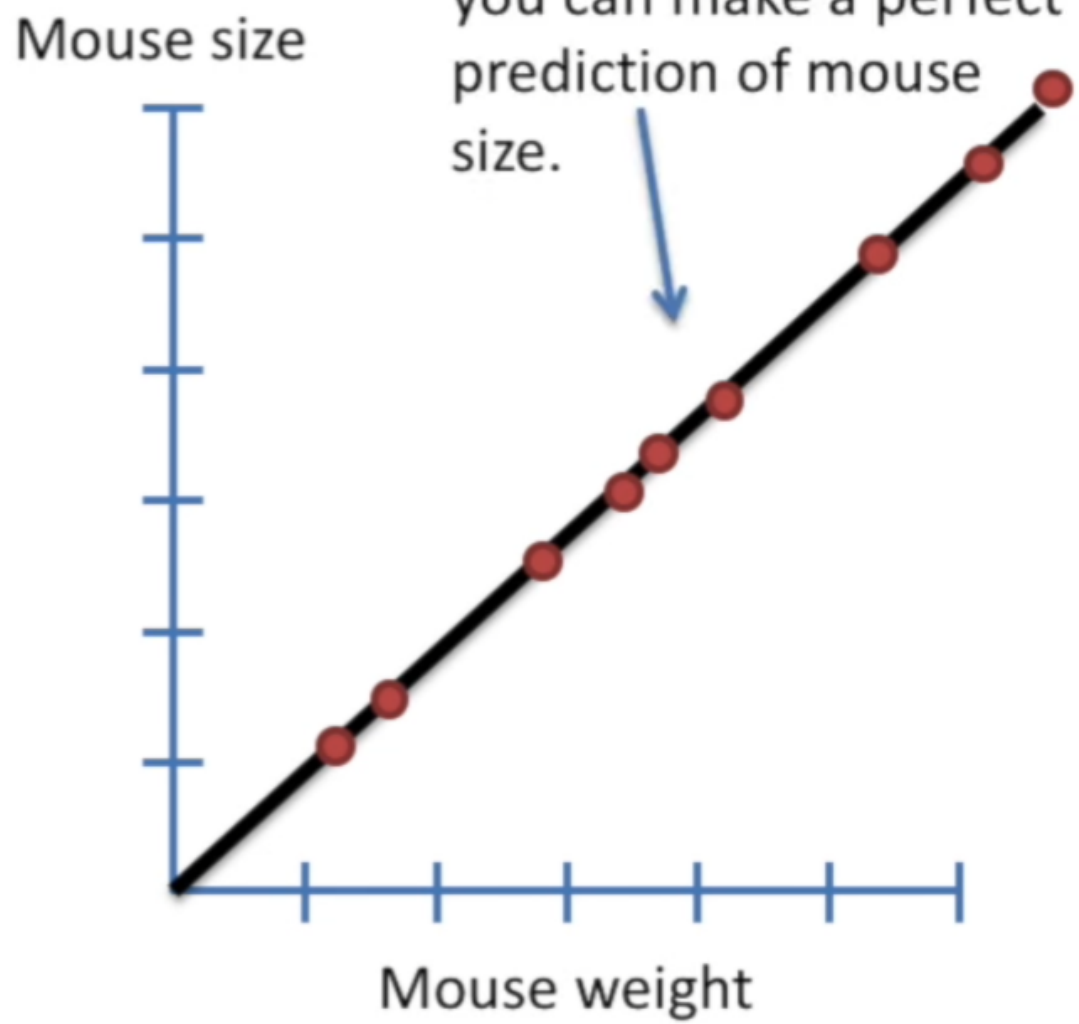
$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$





$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

In this case, knowing mouse weight means you can make a perfect prediction of mouse size.

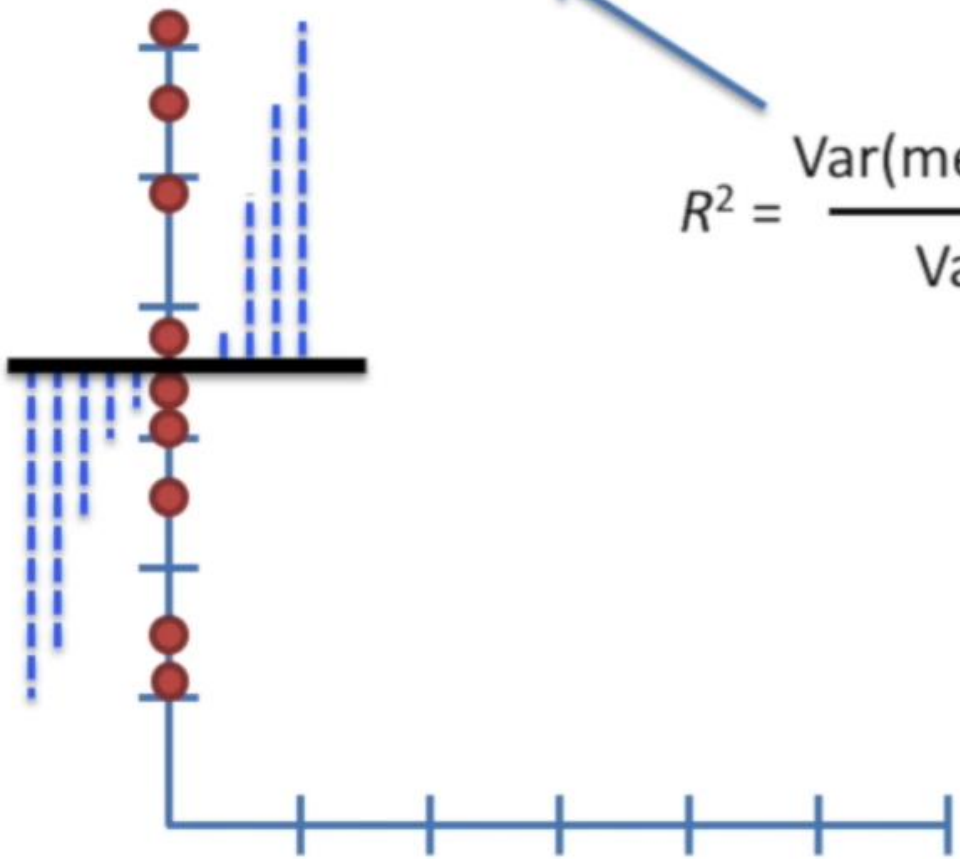


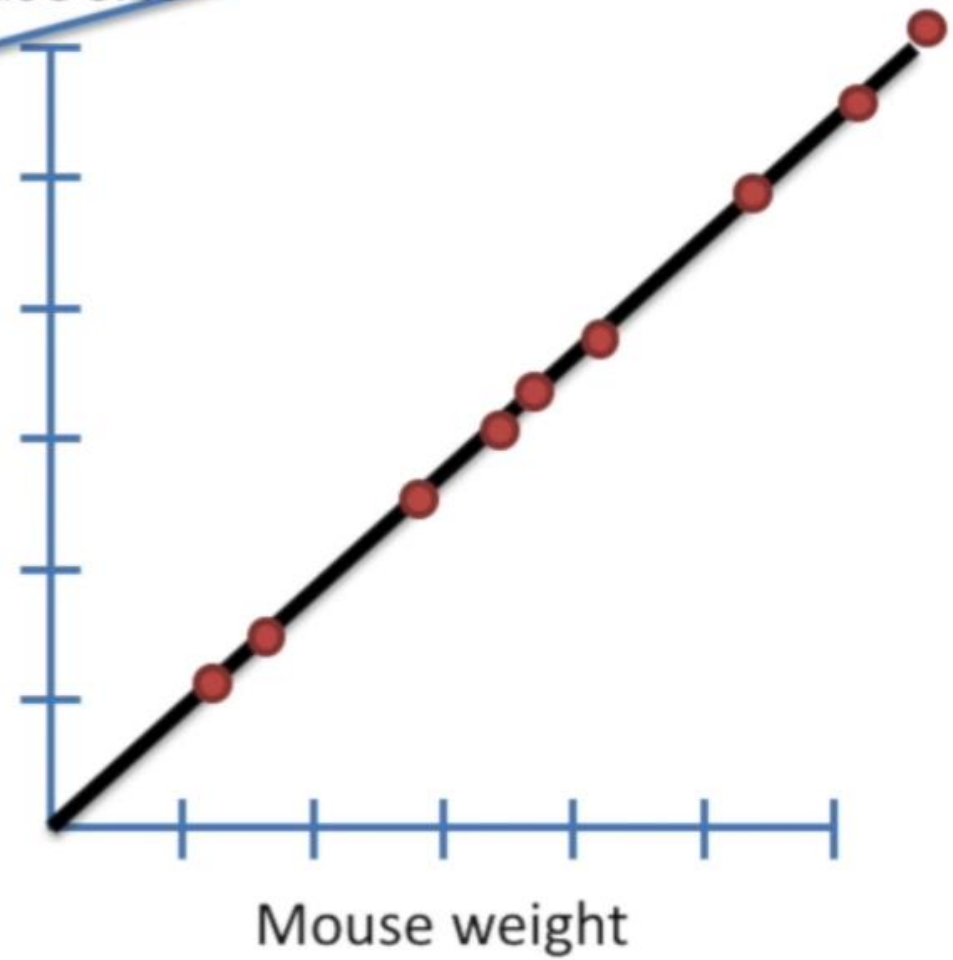
Var(mean) = 11.1

Var(fit) = 0

Mouse size

$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$





Mouse weight

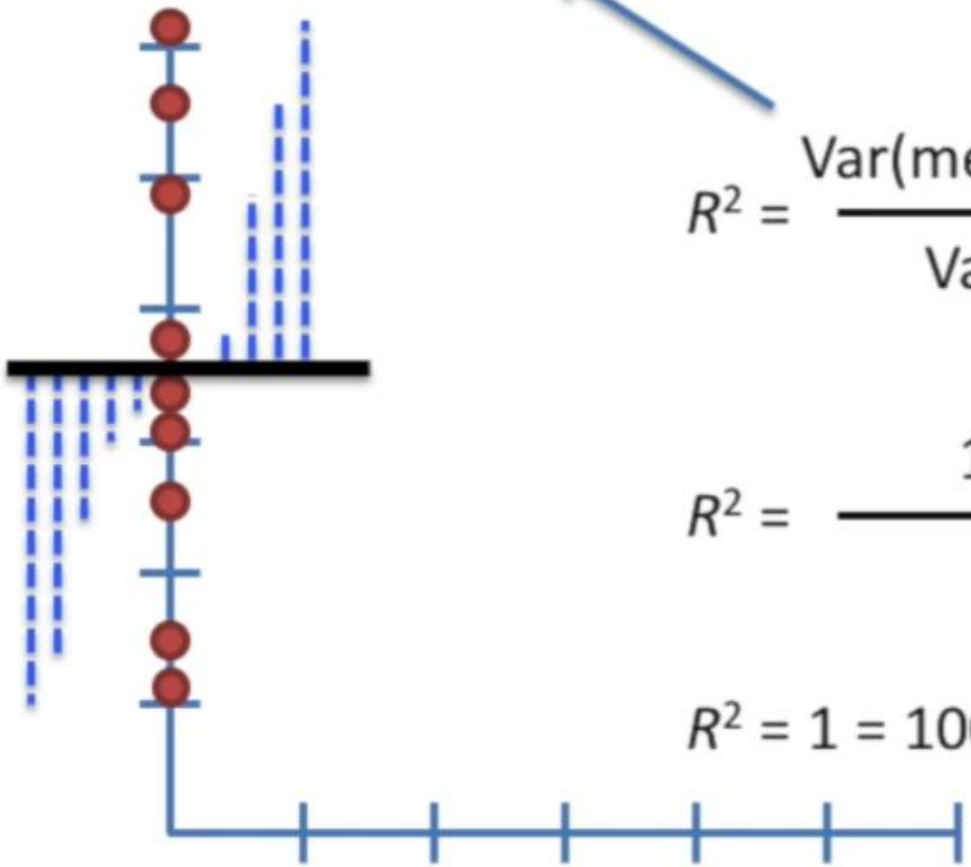
Var(mean) = 11.1

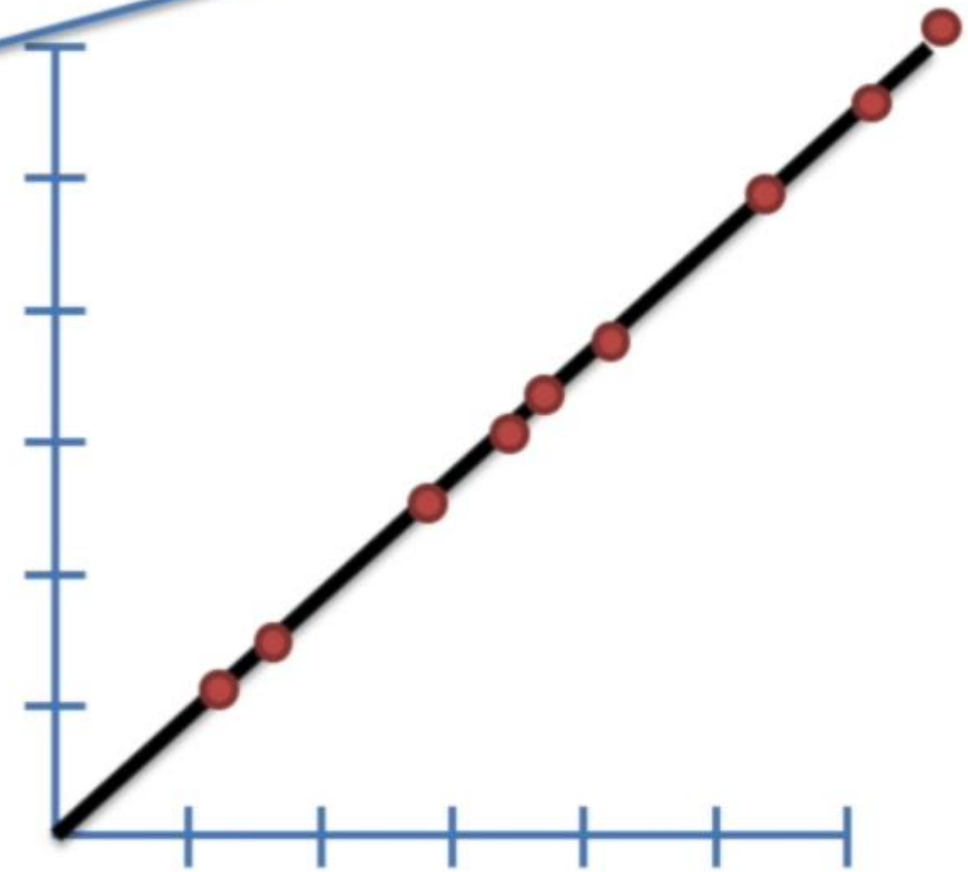
Var(fit) = 0

$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

$$R^2 = \frac{11.1 - 0}{11.1}$$

$R^2 = 1 = 100\%$

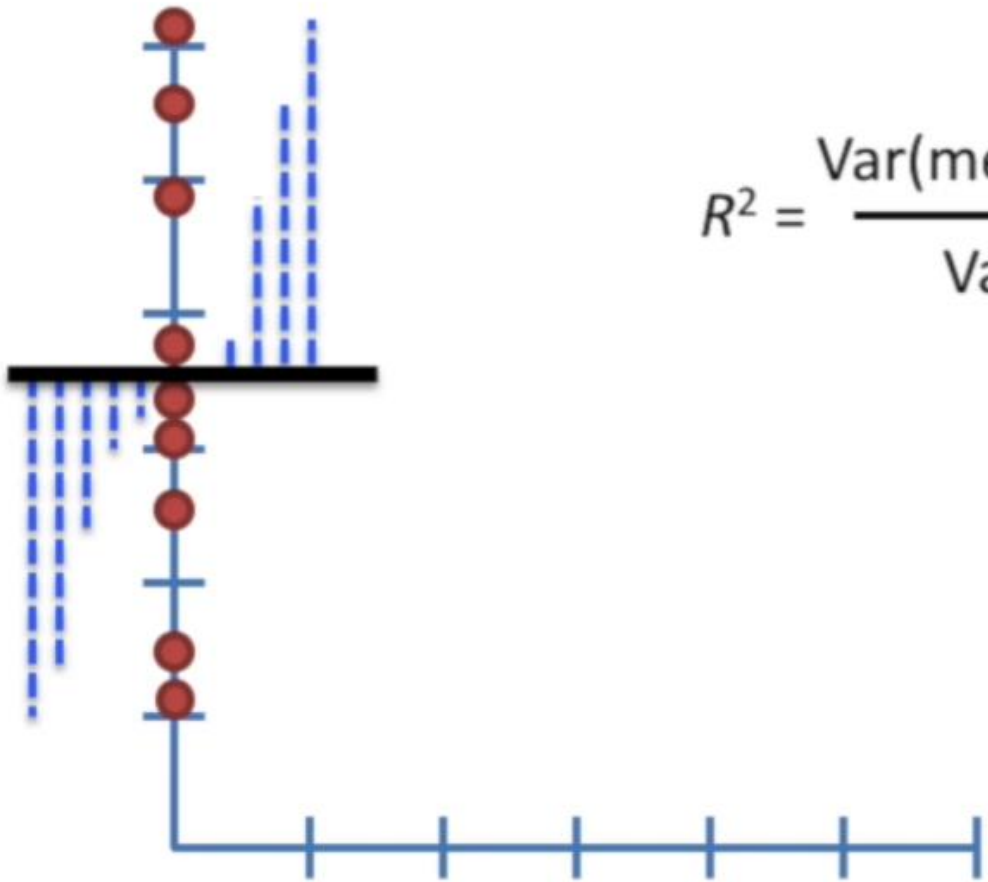




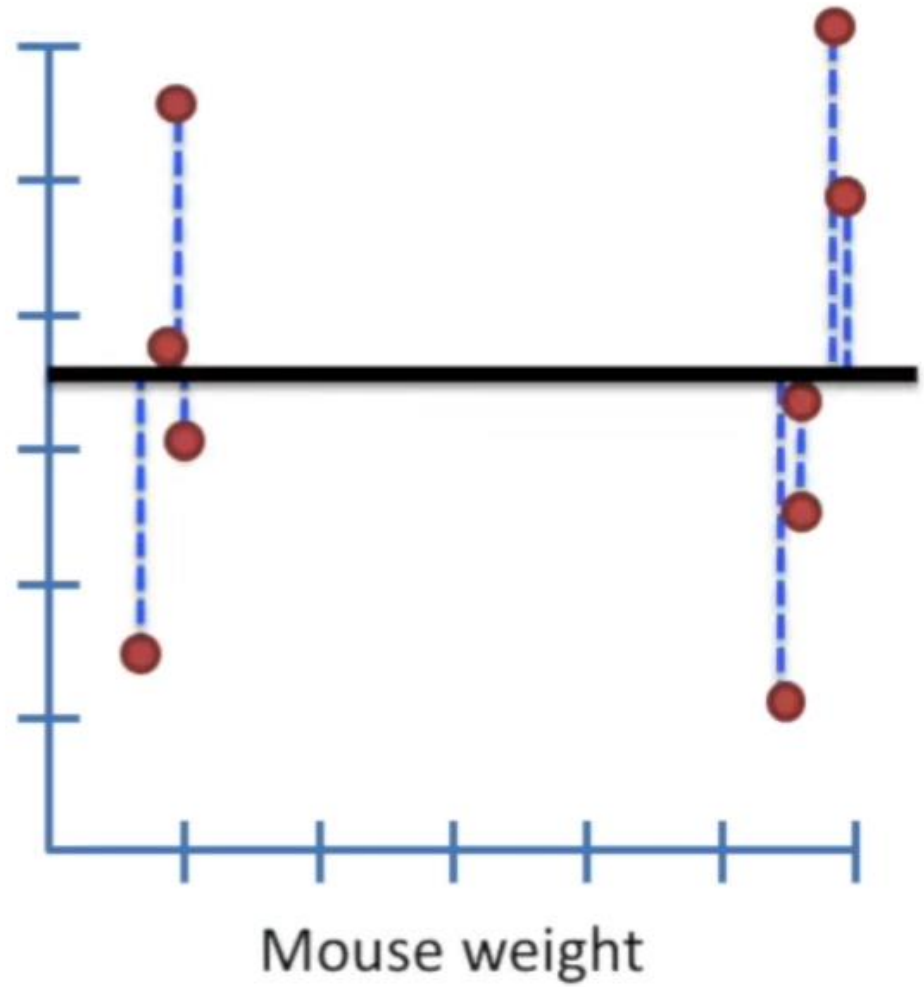
In this case, mouse weight “explains” 100% of the variation in mouse size.

One last example…

Var(mean) = 11.1

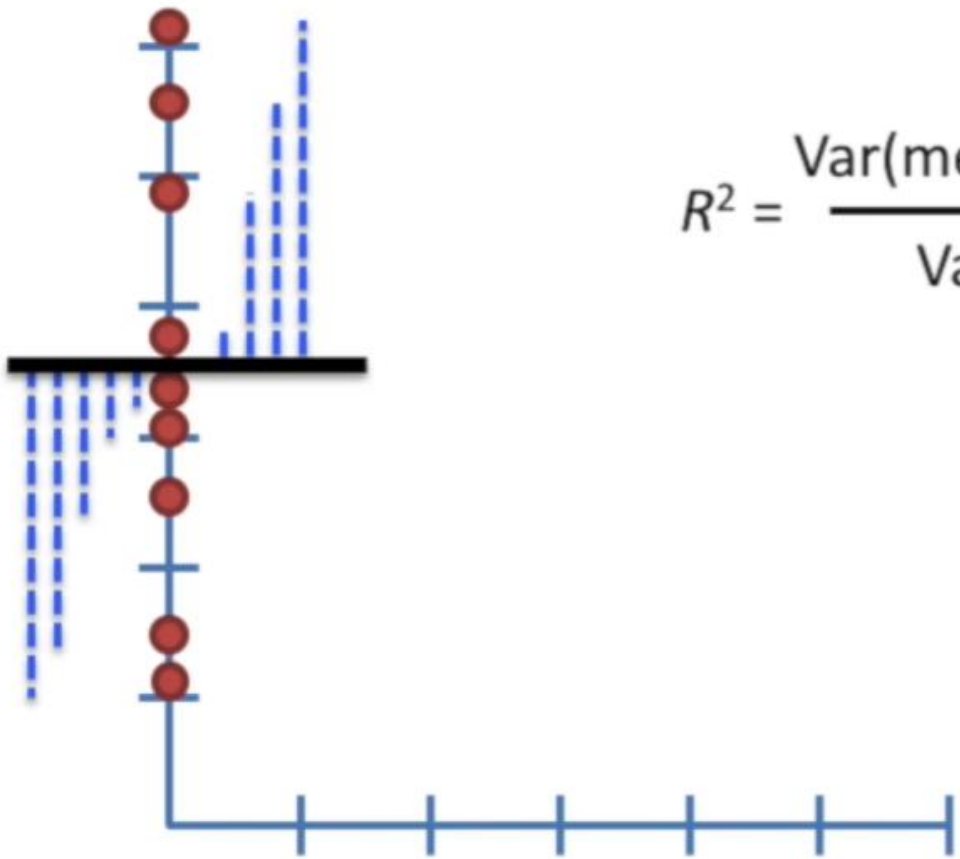


$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$



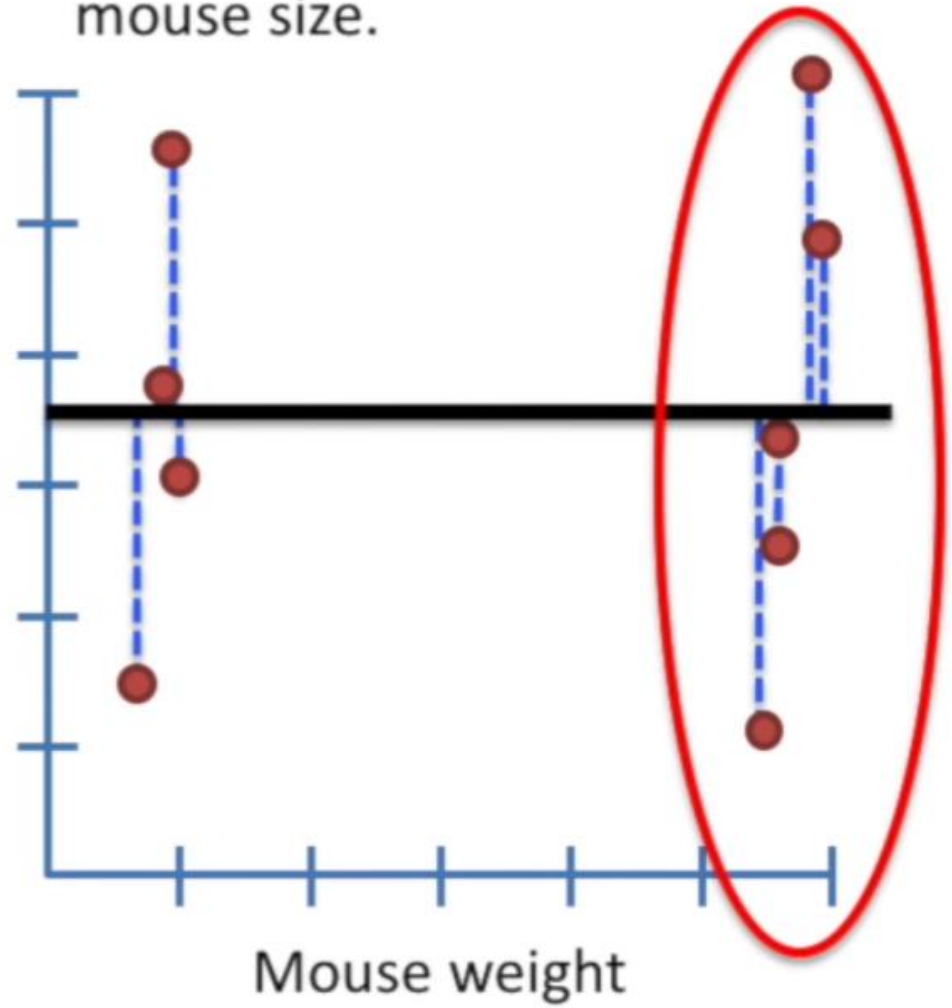
Mouse weight

Var(mean) = 11.1



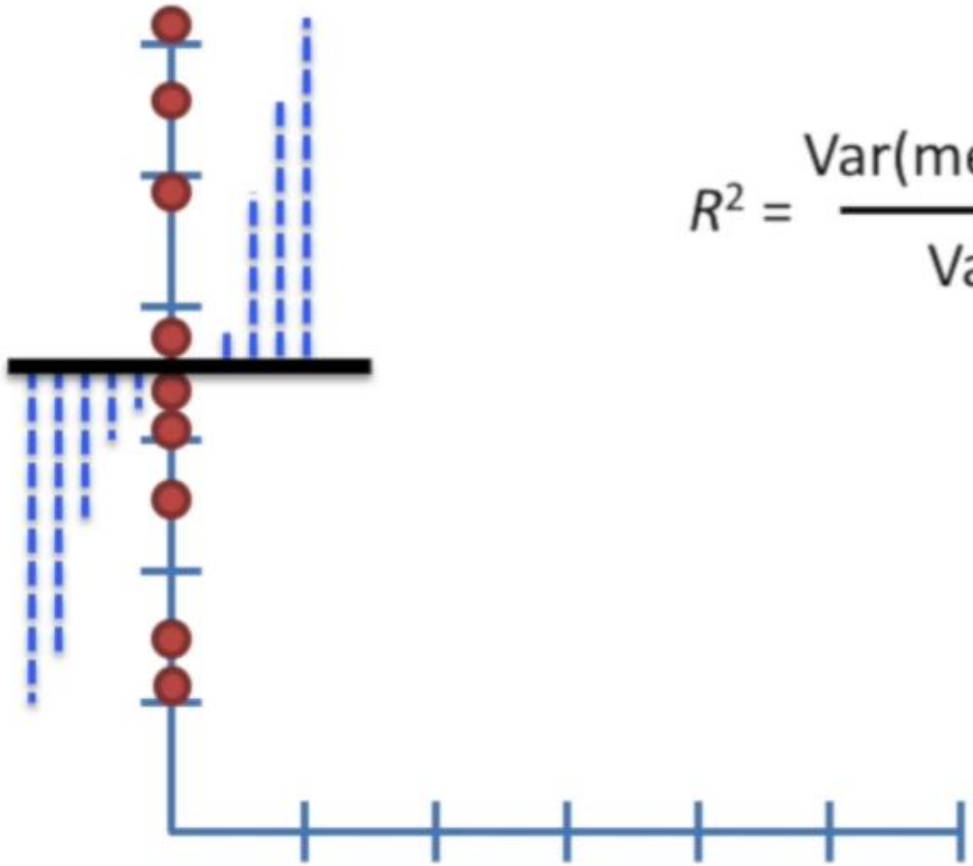
$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

In this case, knowing mouse weight doesn't help us predict mouse size.



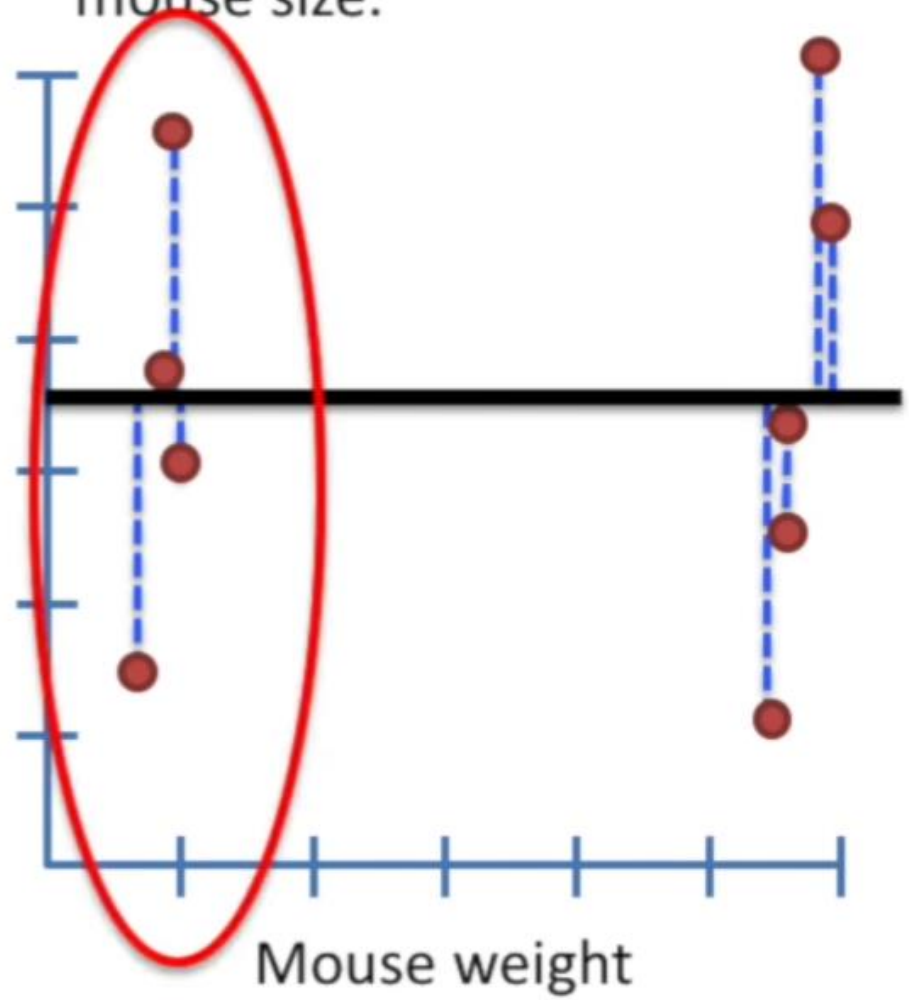
Mouse weight

Var(mean) = 11.1



$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

In this case, knowing mouse weight doesn't help us predict mouse size.

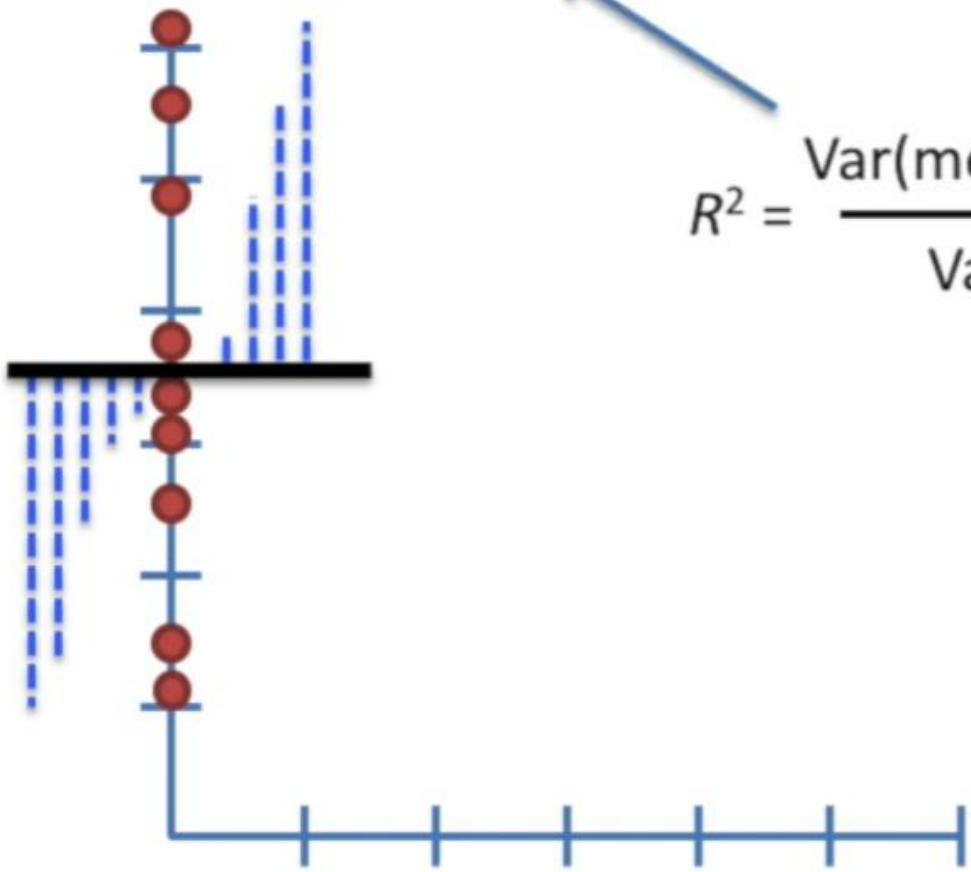


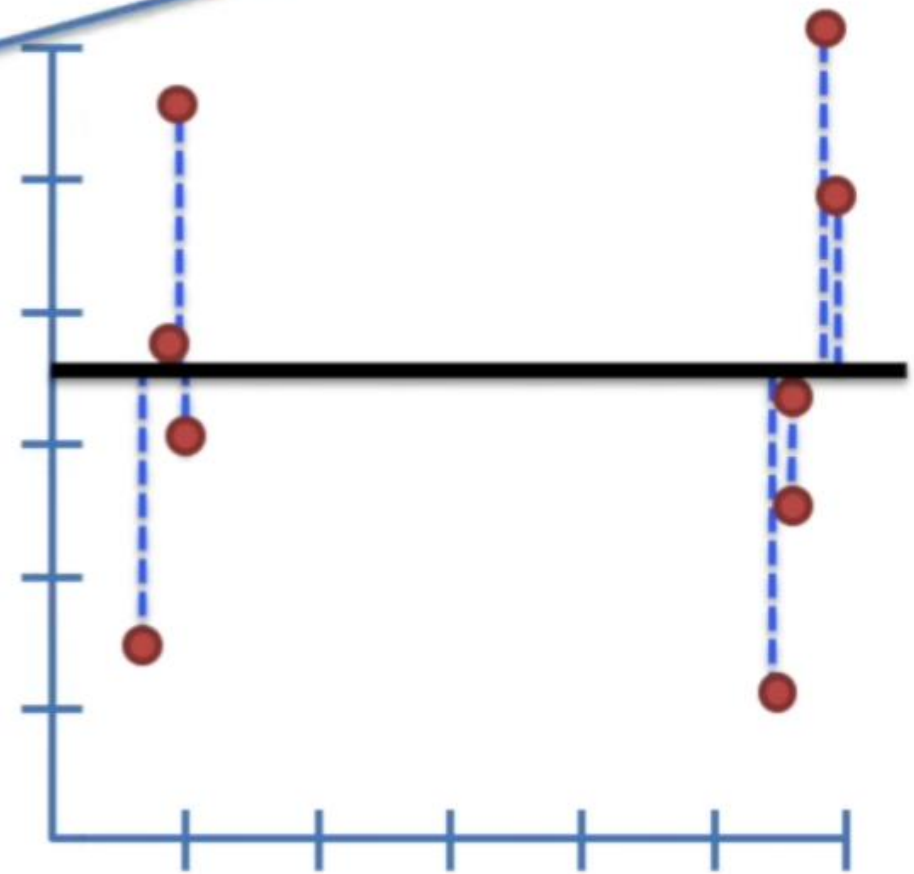
Mouse weight

Var(mean) = 11.1

Var(fit) = 11.1

$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$





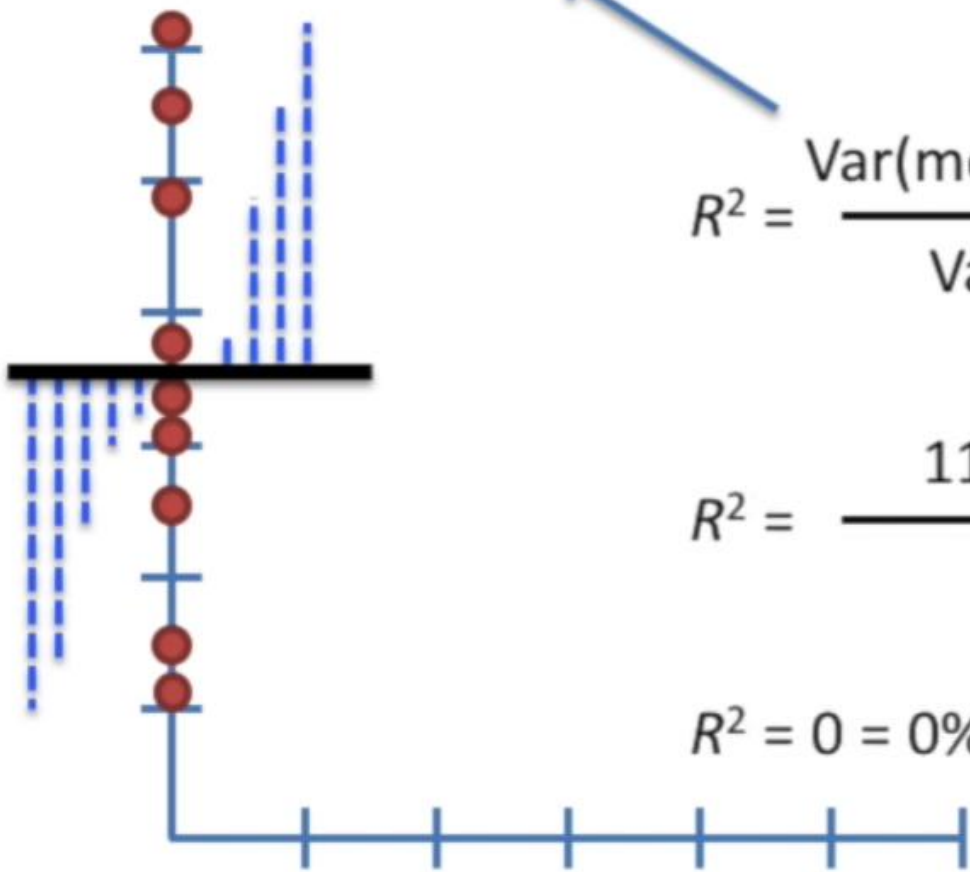
Var(mean) = 11.1

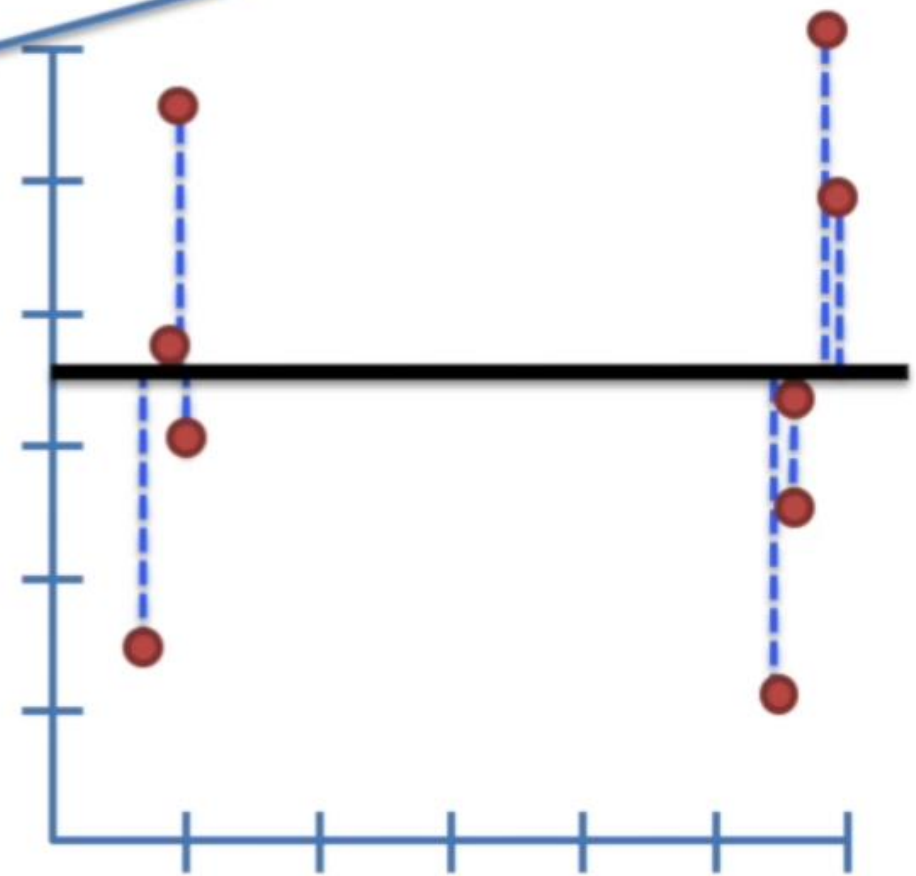
Var(fit) = 11.1

$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

$$R^2 = \frac{11.1 - 11.1}{11.1}$$

$R^2 = 0 = 0\%$





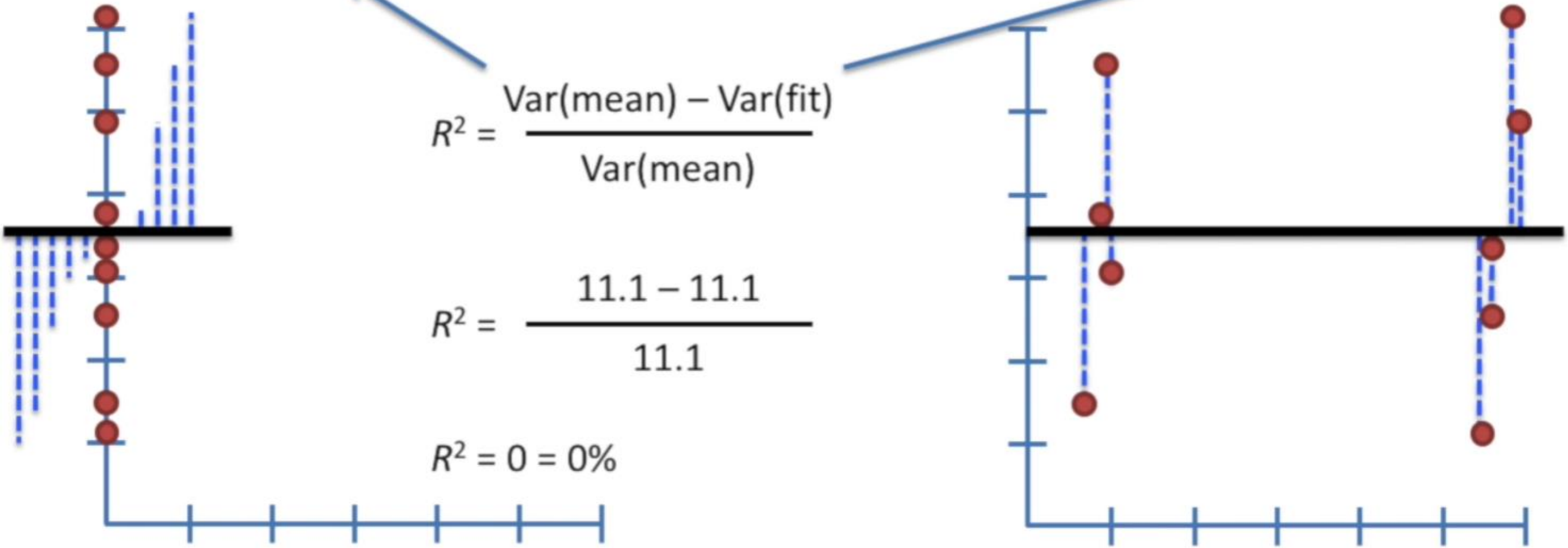
Var(mean) = 11.1

Var(fit) = 11.1

$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

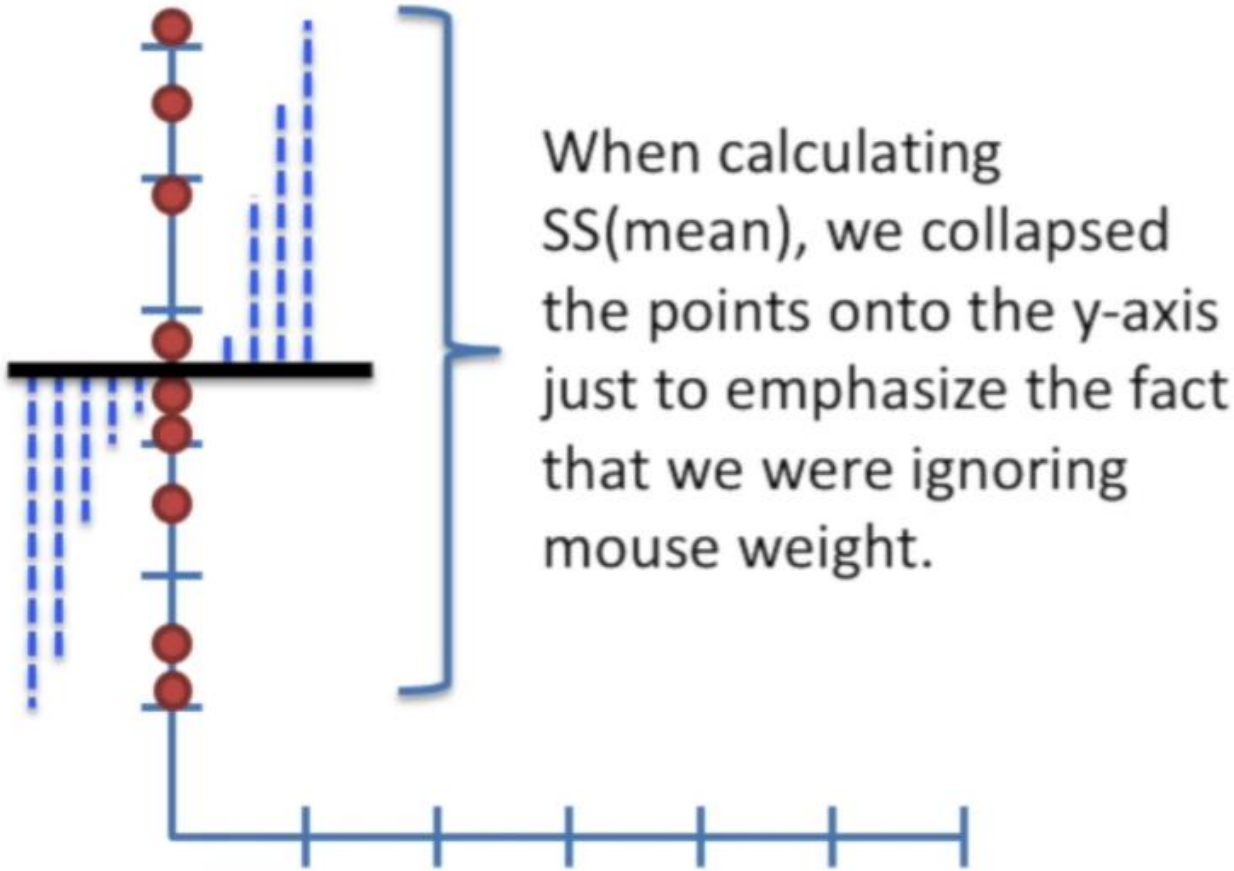
$$R^2 = \frac{11.1 - 11.1}{11.1}$$

$R^2 = 0 = 0\%$



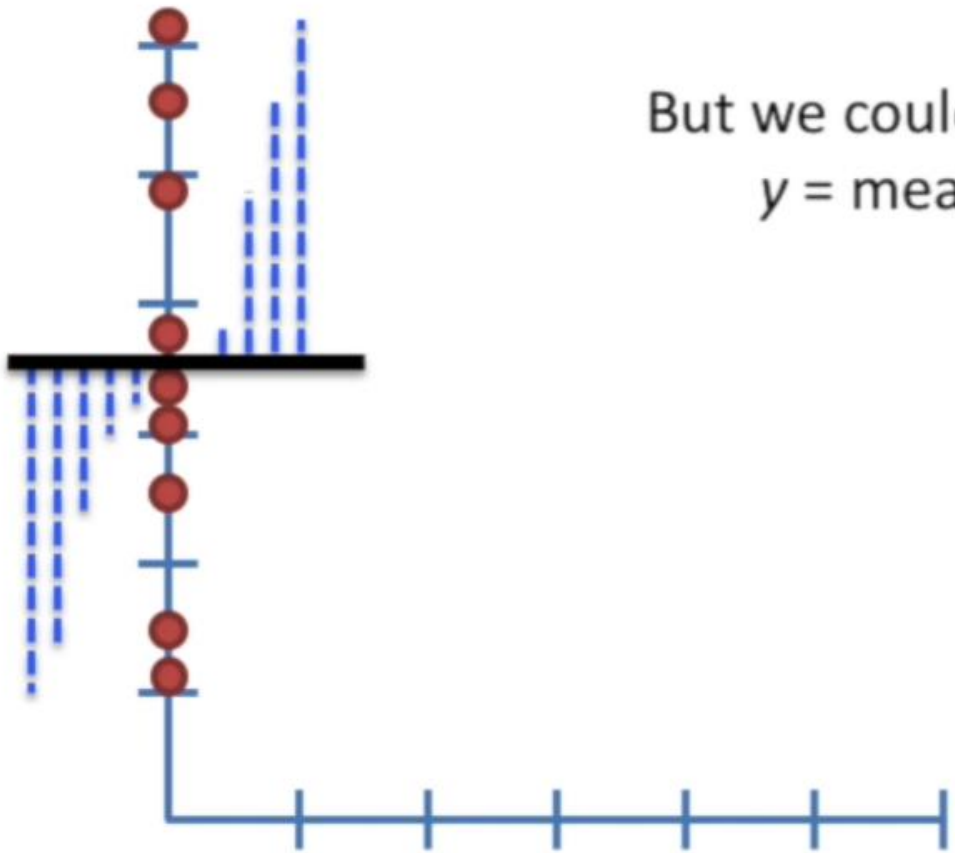
In this case, mouse weight doesn't "explain" any of the variation around the mean.

SS(mean)



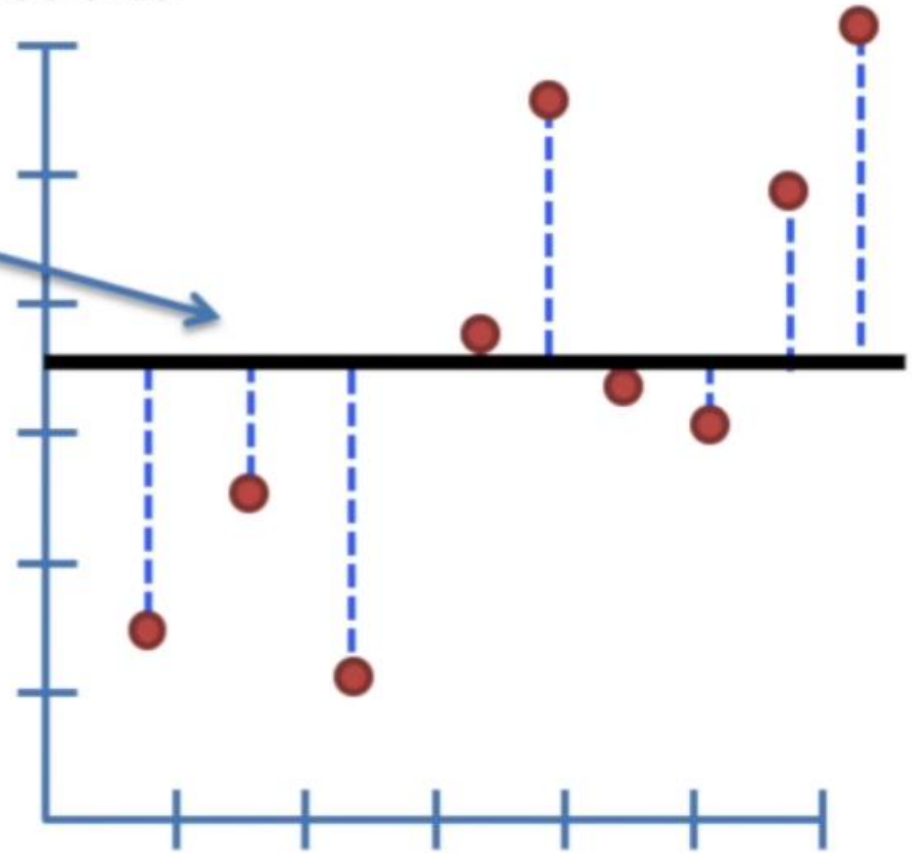
When calculating SS(mean), we collapsed the points onto the y-axis just to emphasize the fact that we were ignoring mouse weight.

SS(mean)



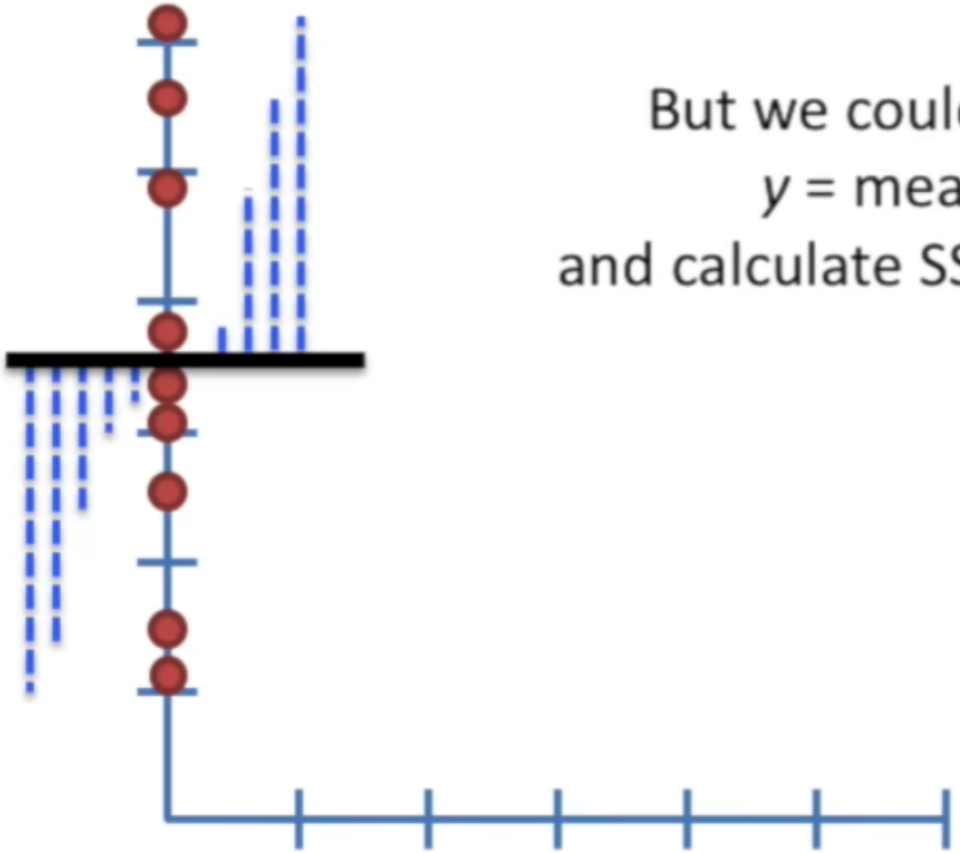
But we could just draw a line,
y = mean mouse size

Mouse size



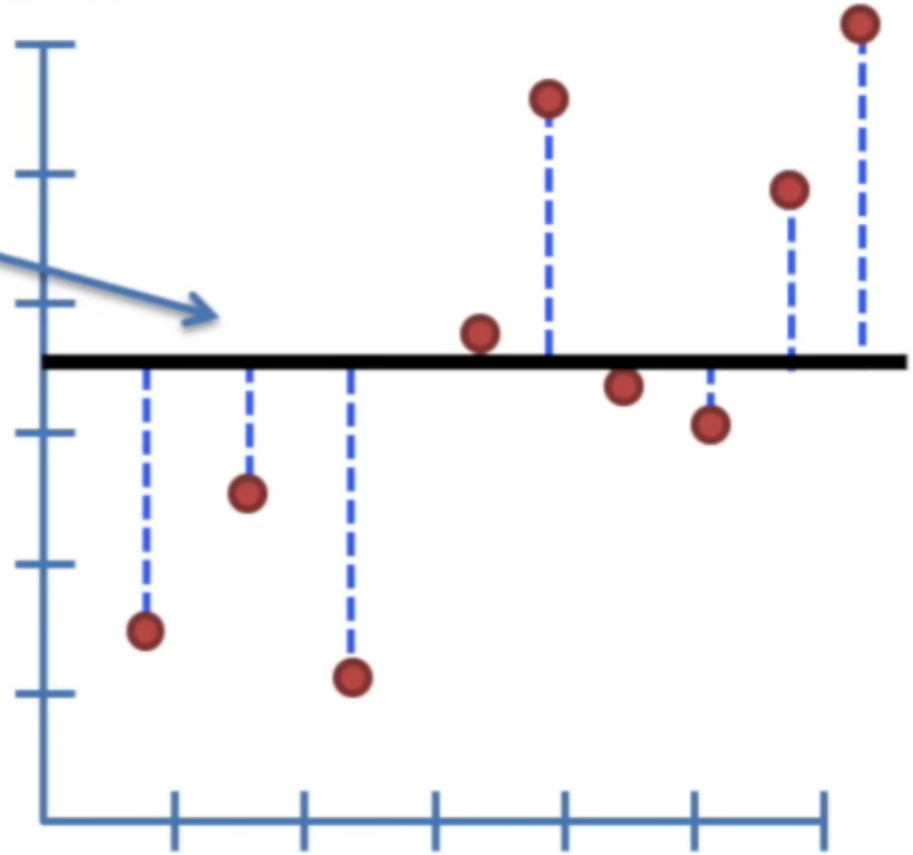
Mouse weight

SS(mean)



But we could just draw a line,
y = mean mouse size
and calculate SS(mean) around that.

Mouse size



Mouse weight