Intro to supervised learning and Linear Regression – Topics

Machine Learning:

- Intro to machine learning, learning from data.
- Supervised and Unsupervised learning, , train - test data.
- Overfitting and Under fitting

Linear Regression:

- Linear relation between two variables, measures of association – correlation and covariance.
- A simple fit, best fit line – measure of a regression fit.
- Multiple regression
- R squared.

Machine Learning

- The ability of a computer to do some task without being explicitly programmed.
- The ability to do the tasks come from the underlying model which is the result of the learning process.
- The model is generated by learning from huge volume of data, huge both in breadth and depth reflecting the real world in which the processes are performed.

What machine learning algorithms do?

- Search through the data to look for patterns in form of trends, cycles, associations, etc.
- Express these patterns as mathematical structures.

Supervised Machine Learning

- Class of machine learning that work on externally supplied instances in form of predictor attributes and **associated target values**.

- The target values are the 'correct answers' for the predictor model which can either be a regression model or a classification model (classifying data into classes.)

- The model learns from the training data using these 'correct answers/target variables' as reference variables.

- The model thus generated is used to make predictions about data not seen by the model before.
  - Ex1 : *model to predict the resale value of a car based on its mileage, age, color etc.*
  - *Ex2 : model to determine the type of a tumor.*

- If the model does very well with the training data but fails with test data(unseen data), overfitting is said to have taken place. However, if the data does not capture the features of train data itself, we term it as under fitting.

Linear Regression

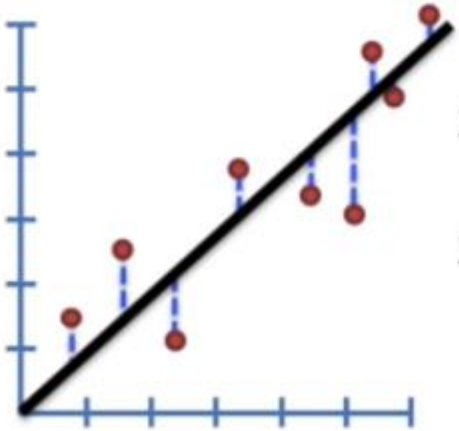
- The term “Regression” generally refers to predicting a target value, which is generally a real number, for a data point based on its attributes.

- The term “linear” in linear regression refers to the fact that the method models data with linear combination of the explanatory variables (attributes).

- In case of linear regression with a single explanatory variable, the linear combination can be expressed as :
  - response = intercept + constant*explanatory variable

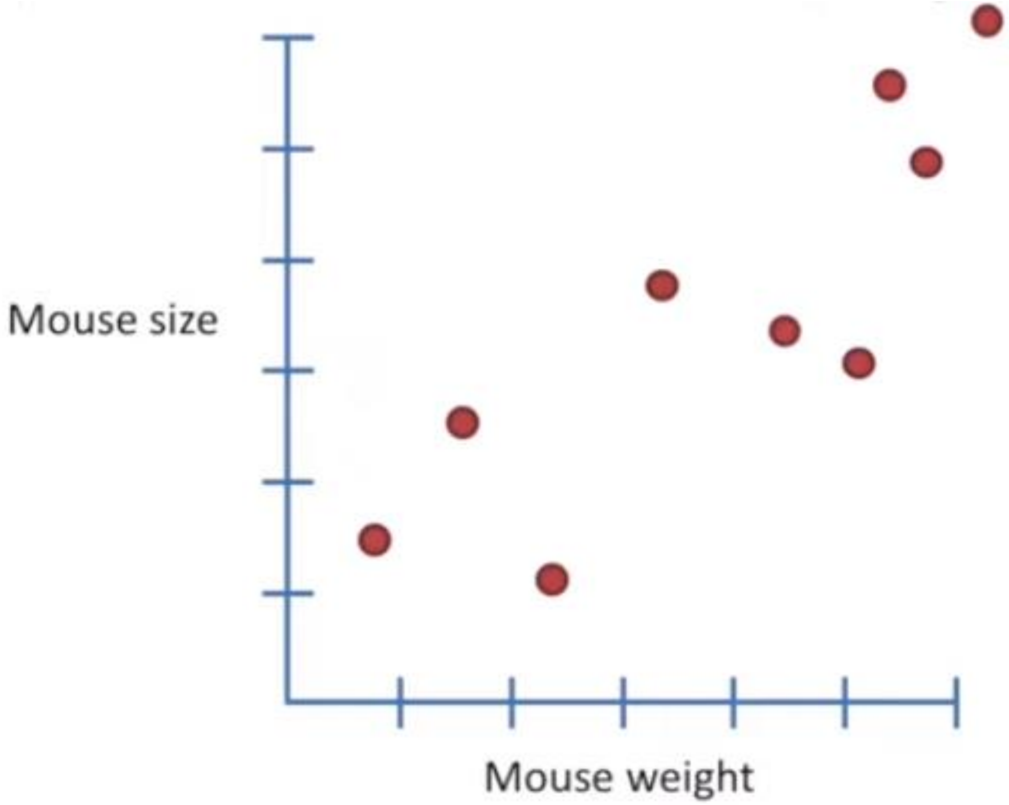
The Main Ideas!

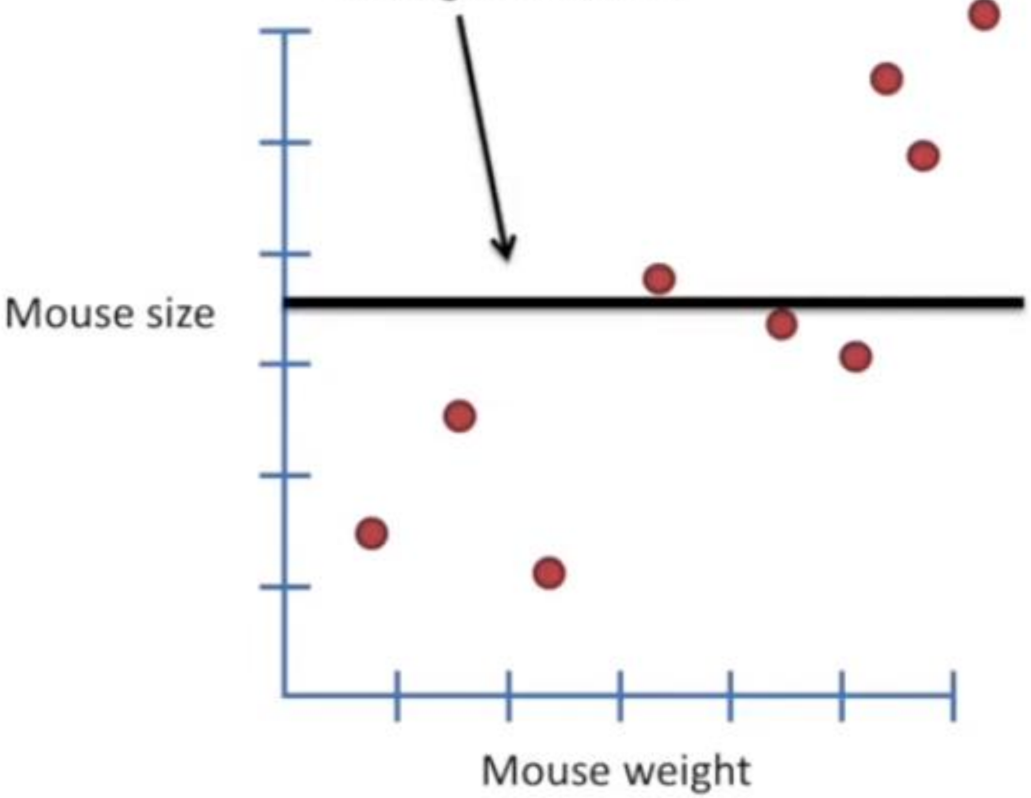
1) Use least-squares to fit a a line to the data.



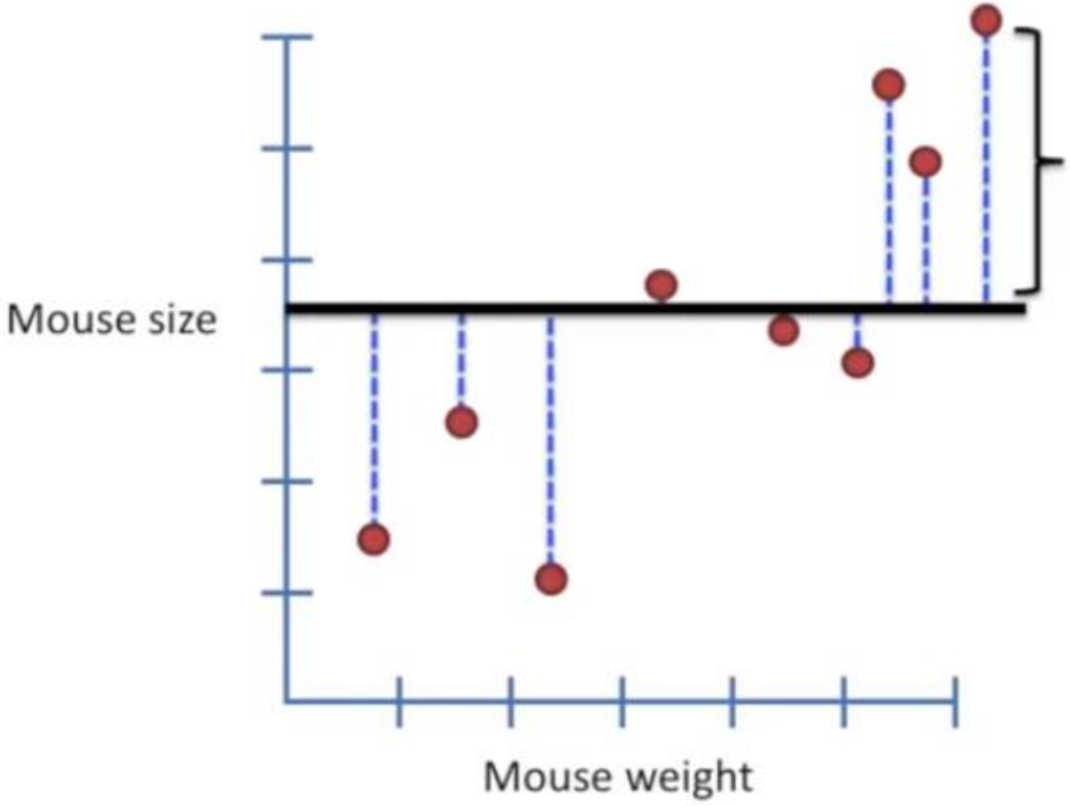
2) Calculate R^2

3) Calculate a *p*-value for R^2

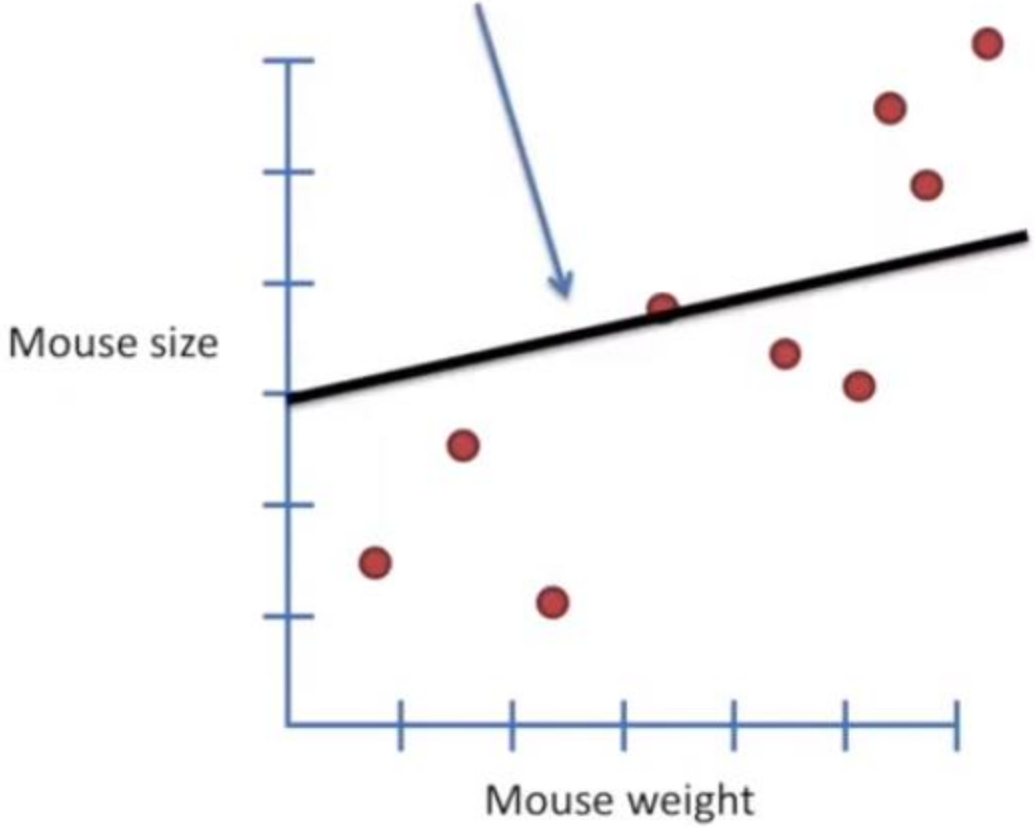

Mouse size
Mouse weight



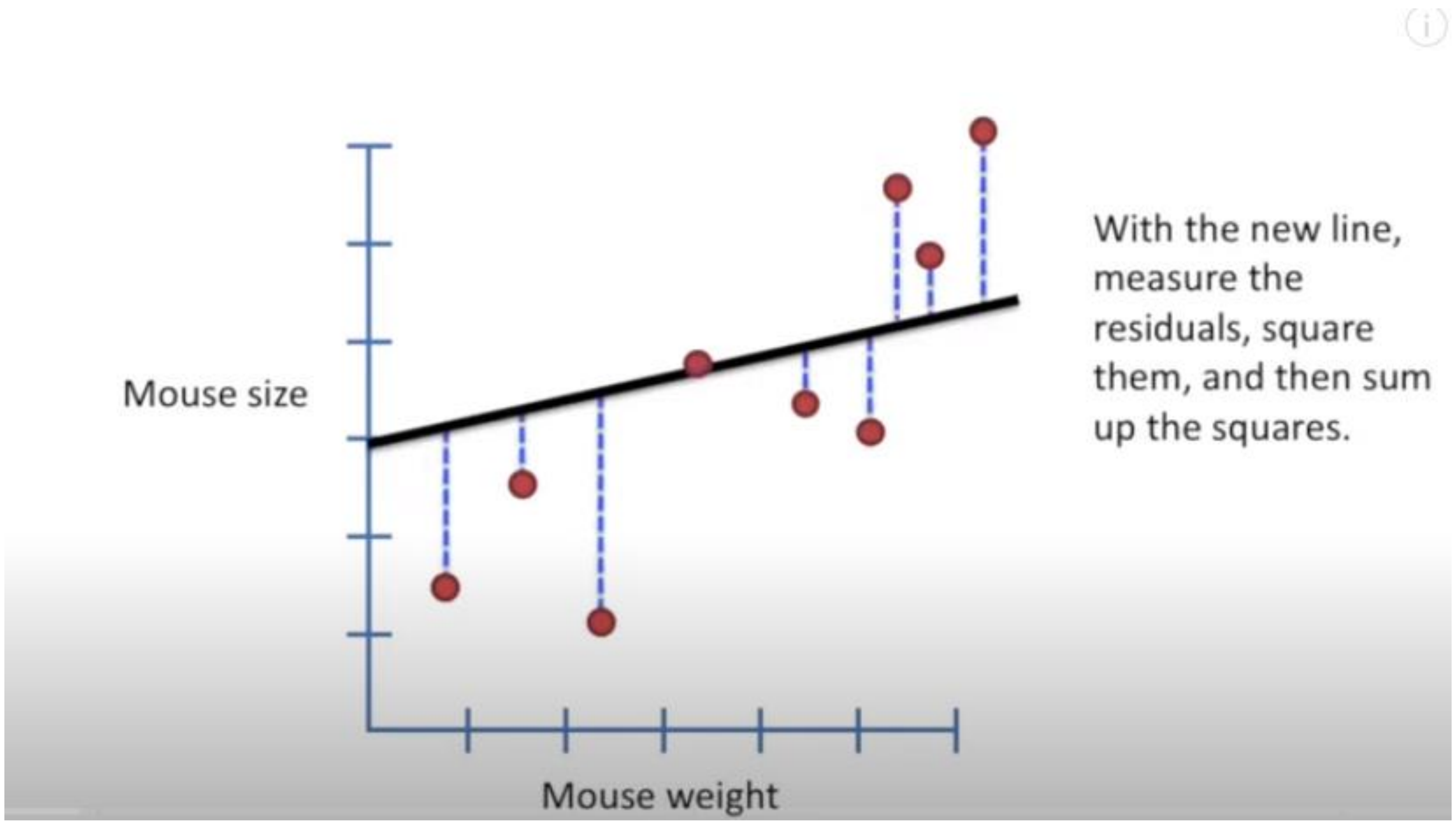
First, draw a line through the data...



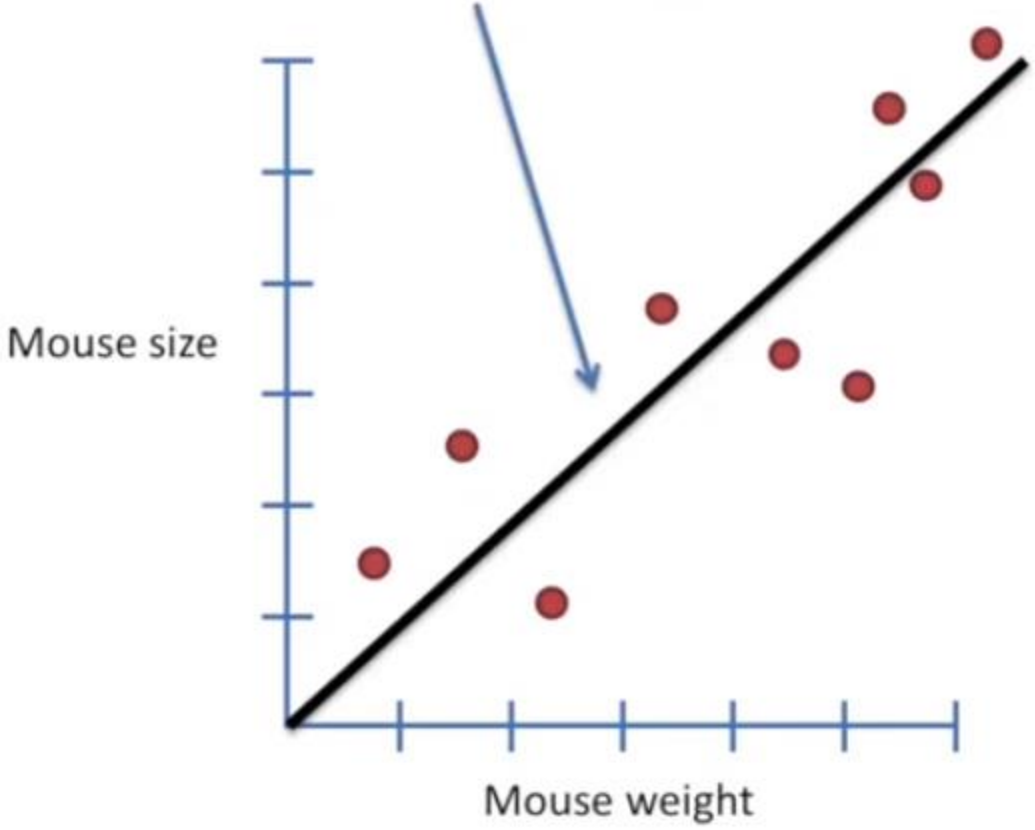
Second, measure the distance from the line to the data, square each distance, and then add them up.



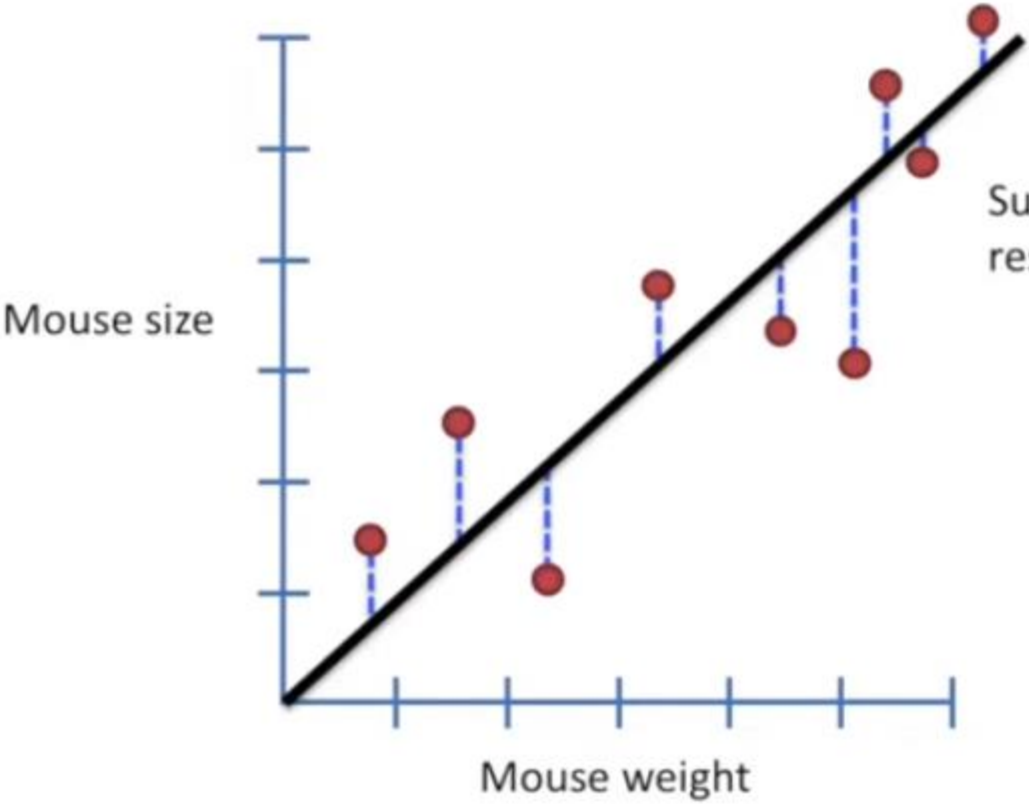
Third, rotate the line a little bit...



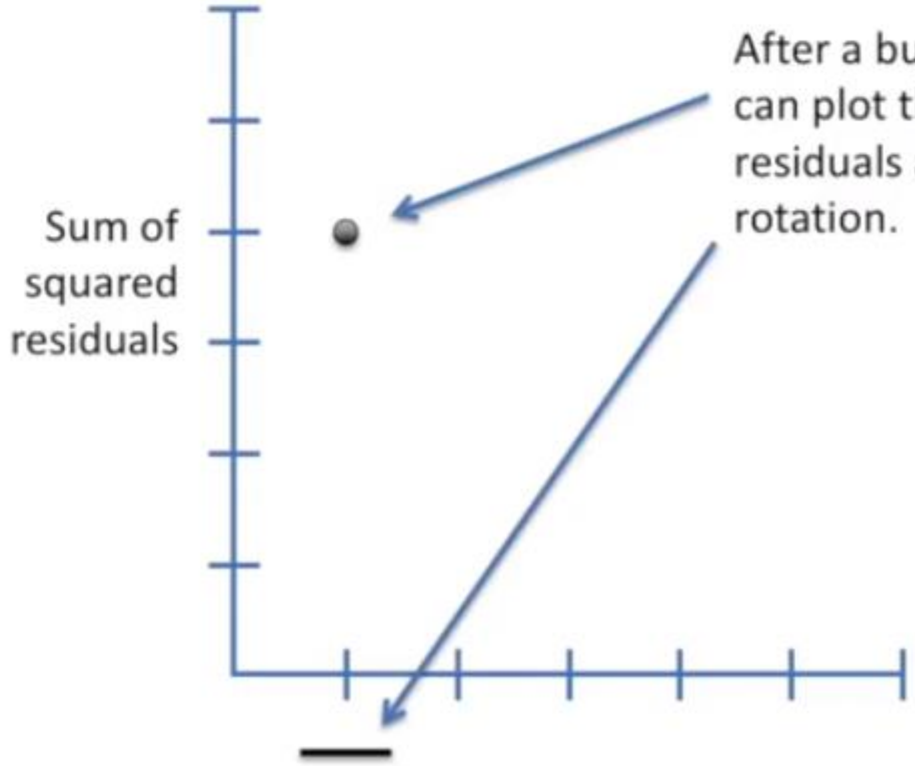
With the new line, measure the residuals, square them, and then sum up the squares.



Rotate the line a little bit more...

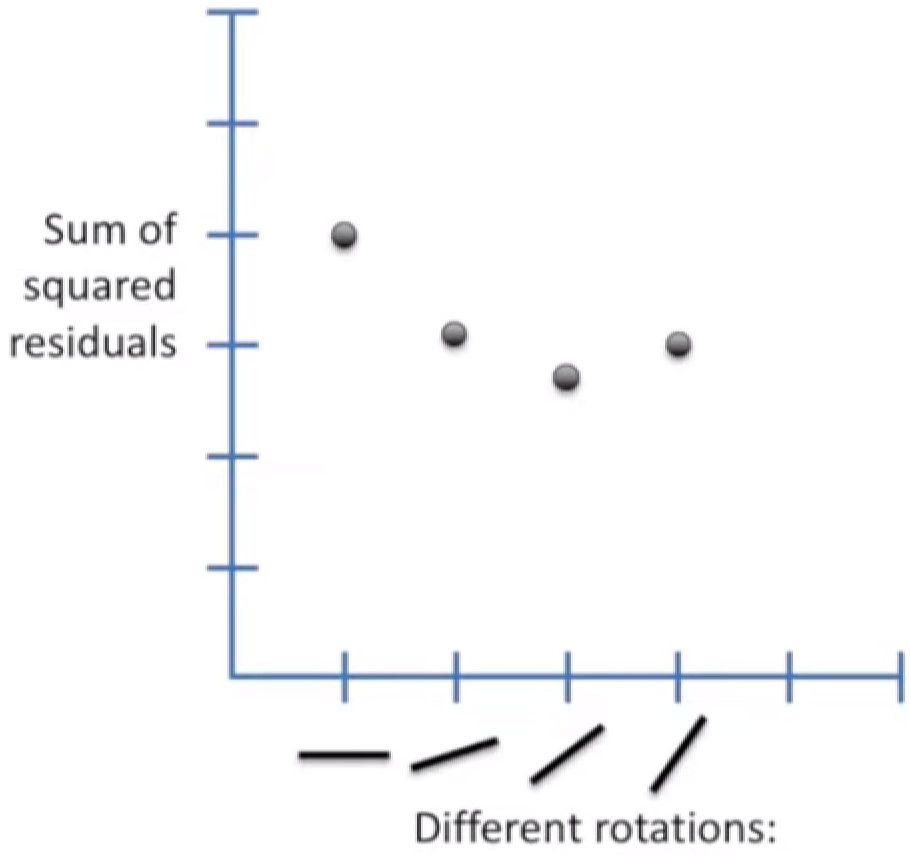


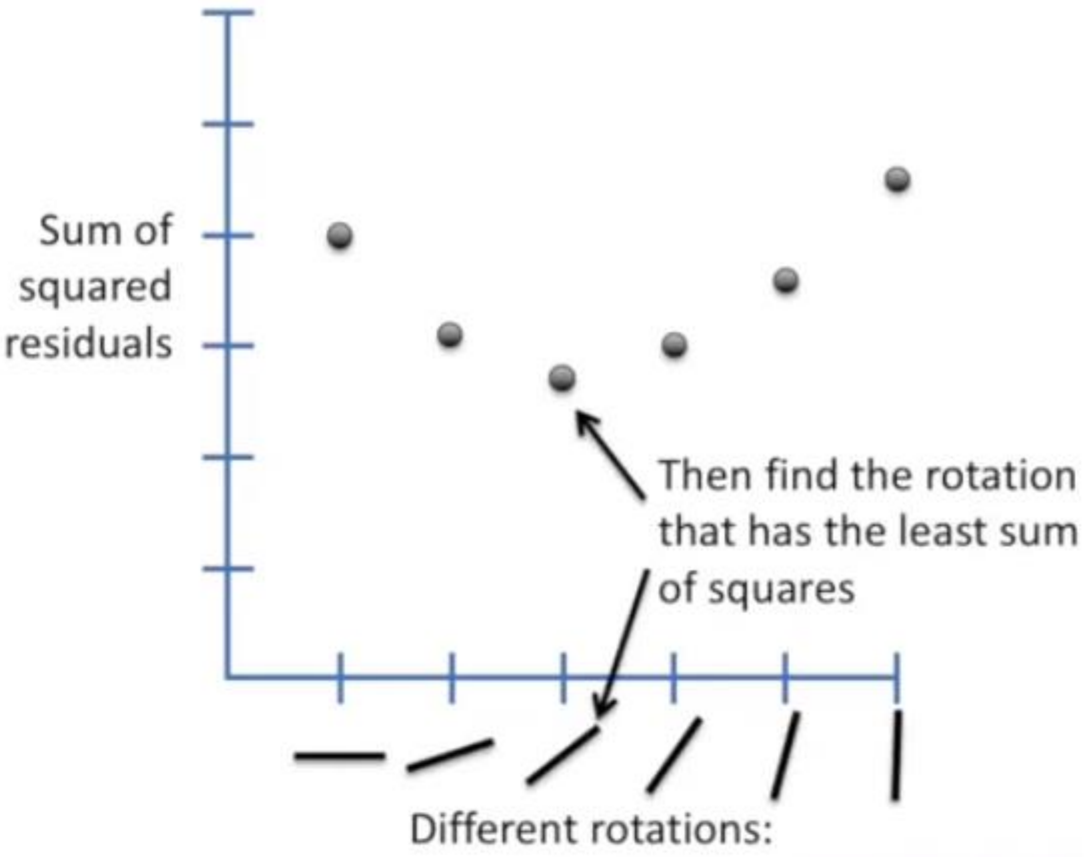
Sum up the squared residuals...



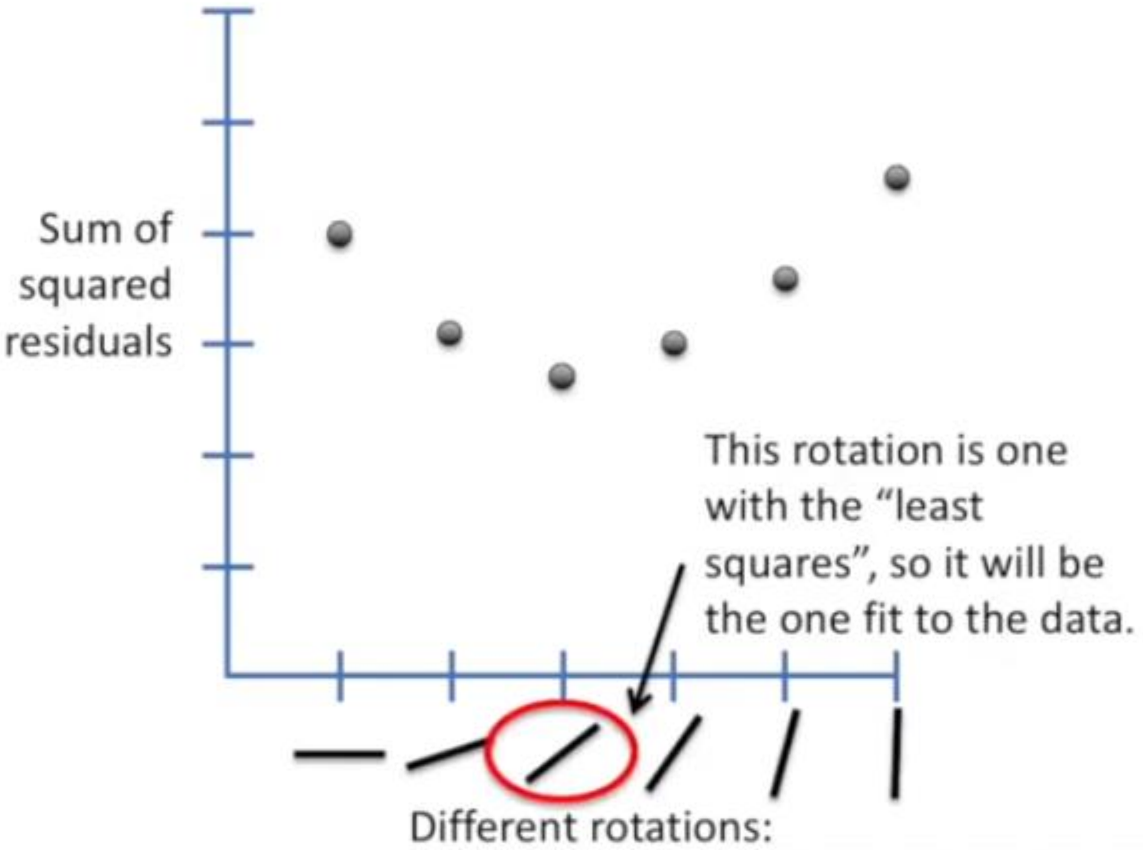
After a bunch of rotations, you can plot the sum of squared residuals and corresponding rotation.

Different rotations:

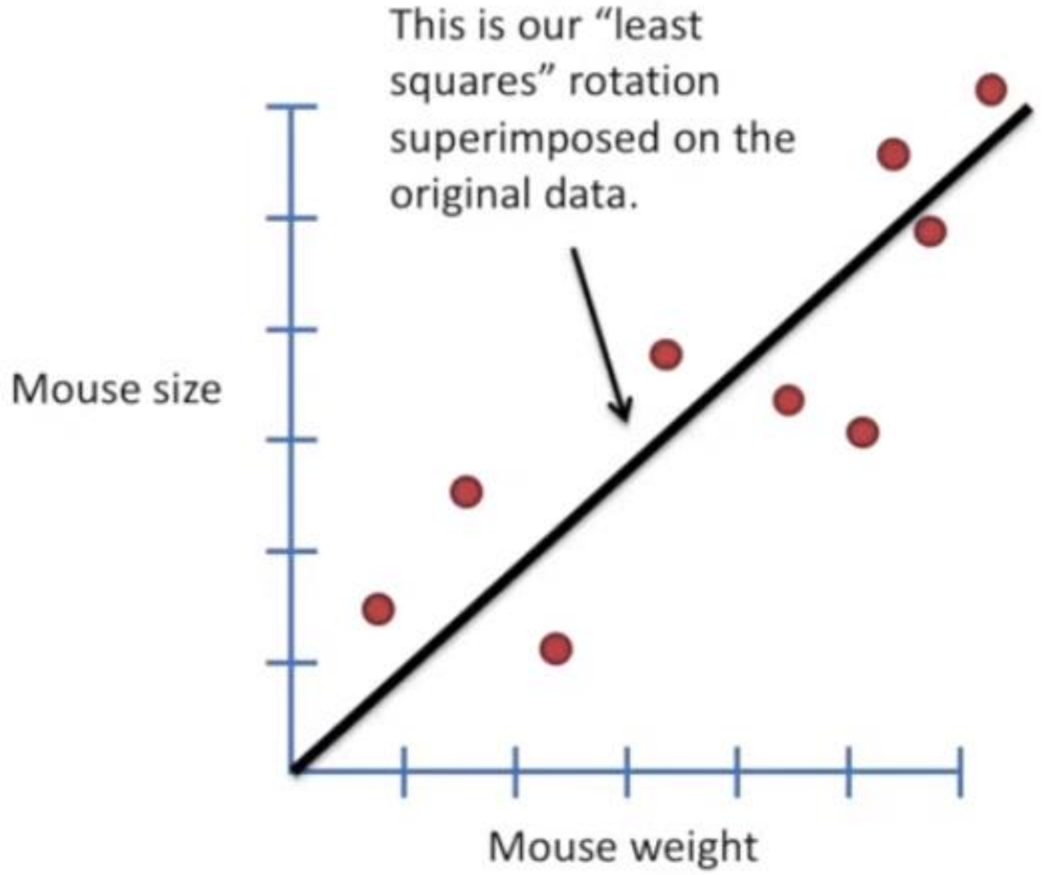

Sum of squared residuals
Different rotations:



Then find the rotation that has the least sum of squares

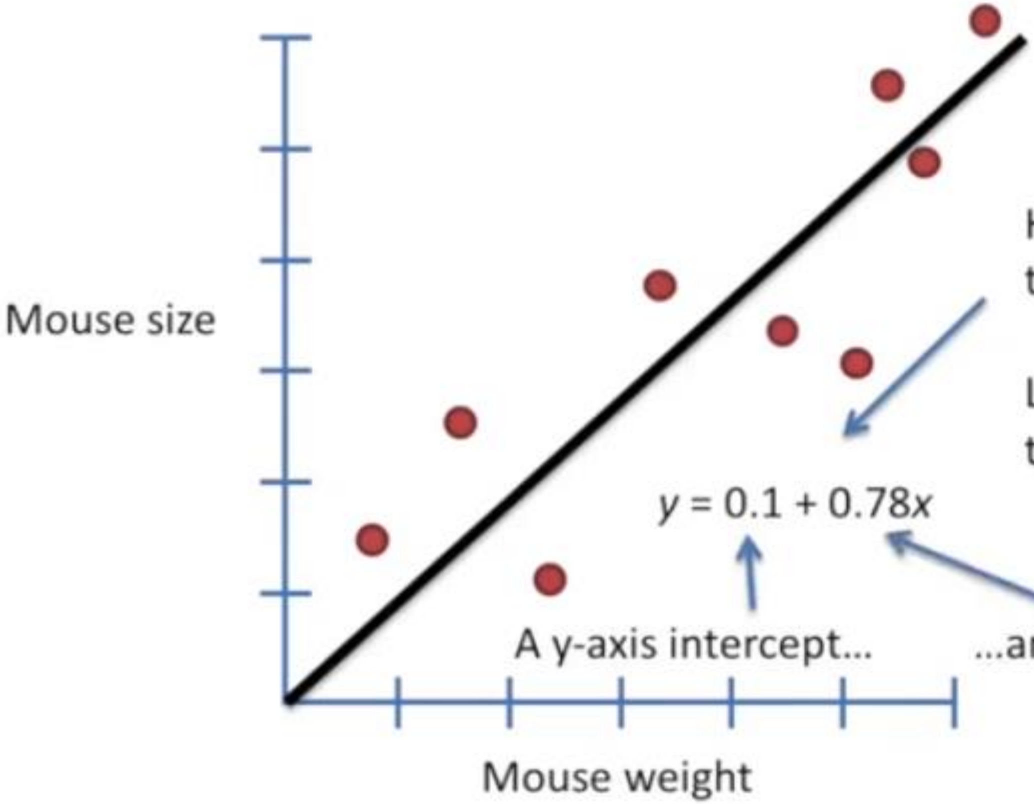


This rotation is one with the “least squares”, so it will be the one fit to the data.



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

Now we have fit a line to the data!



Here's the equation for the line.

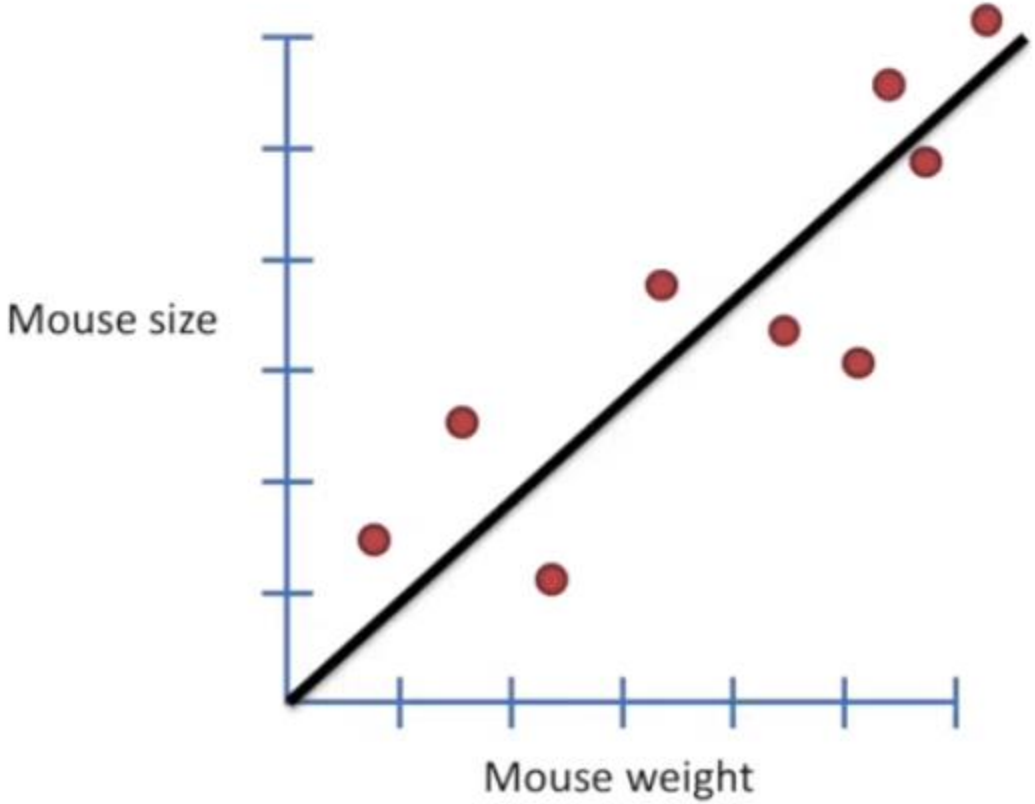
Least-squares estimated two parameters:

$y = 0.1 + 0.78x$

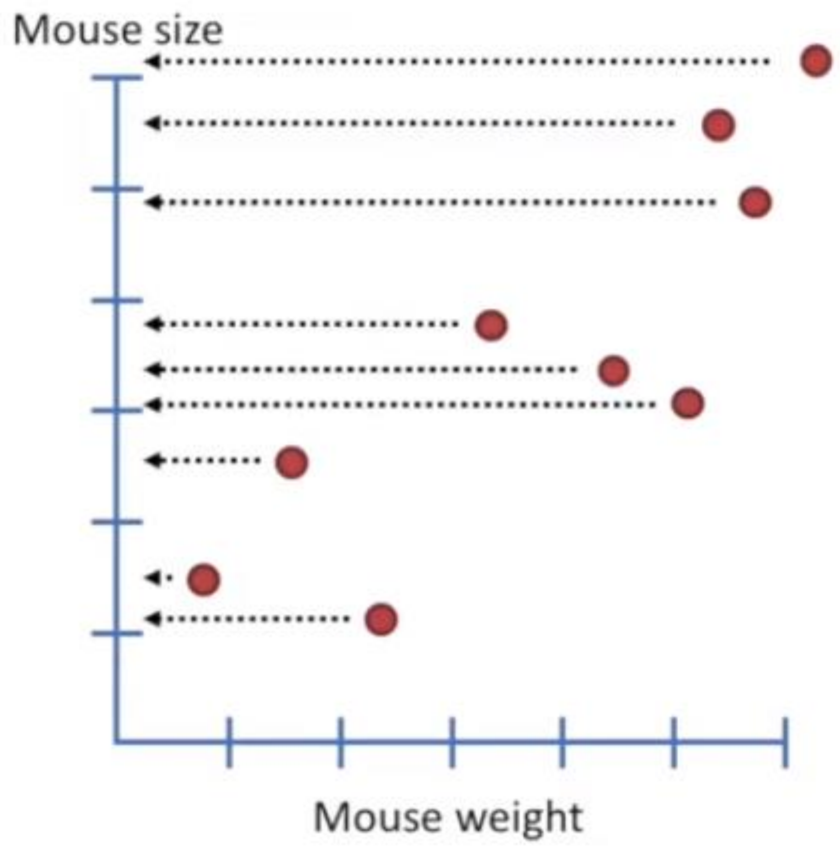
A y-axis intercept...

...and a slope

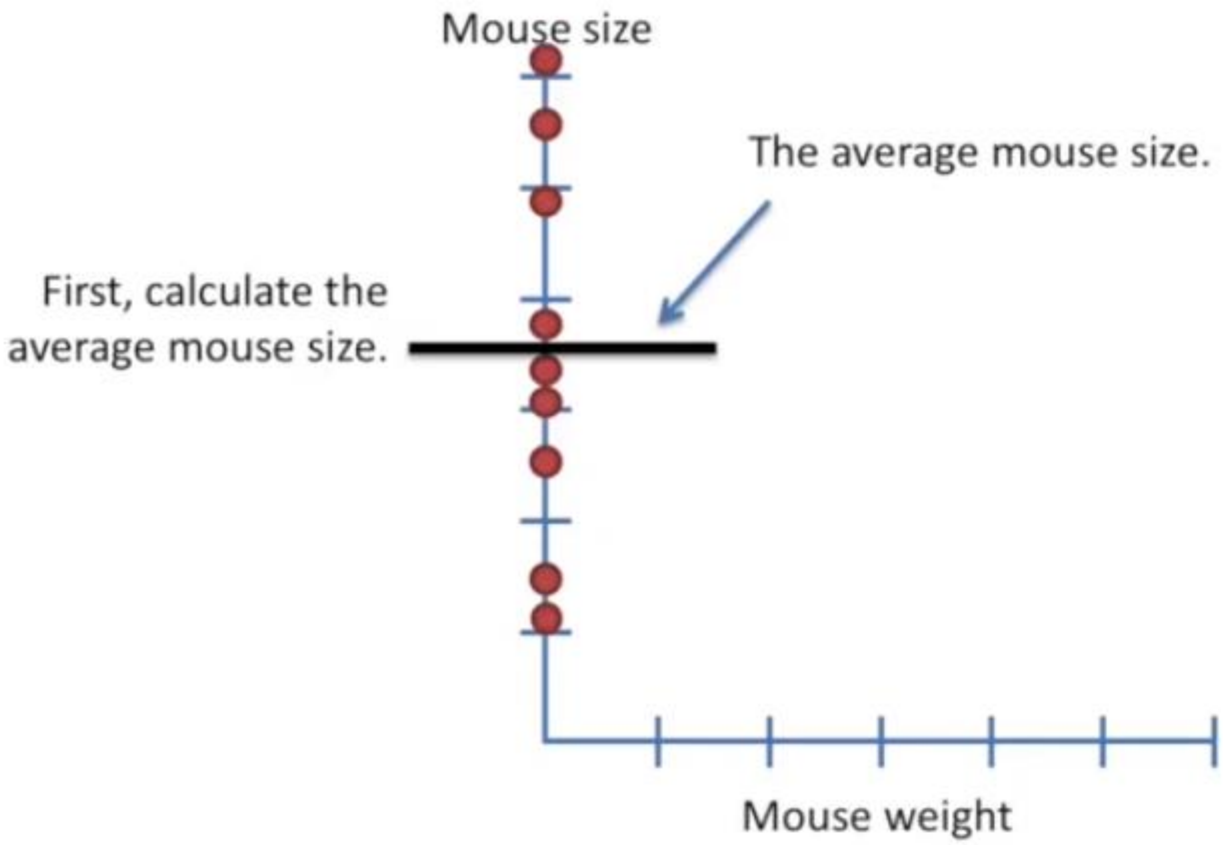
Calculating R^2 is the first step in determining how good that guess will be.



R^2

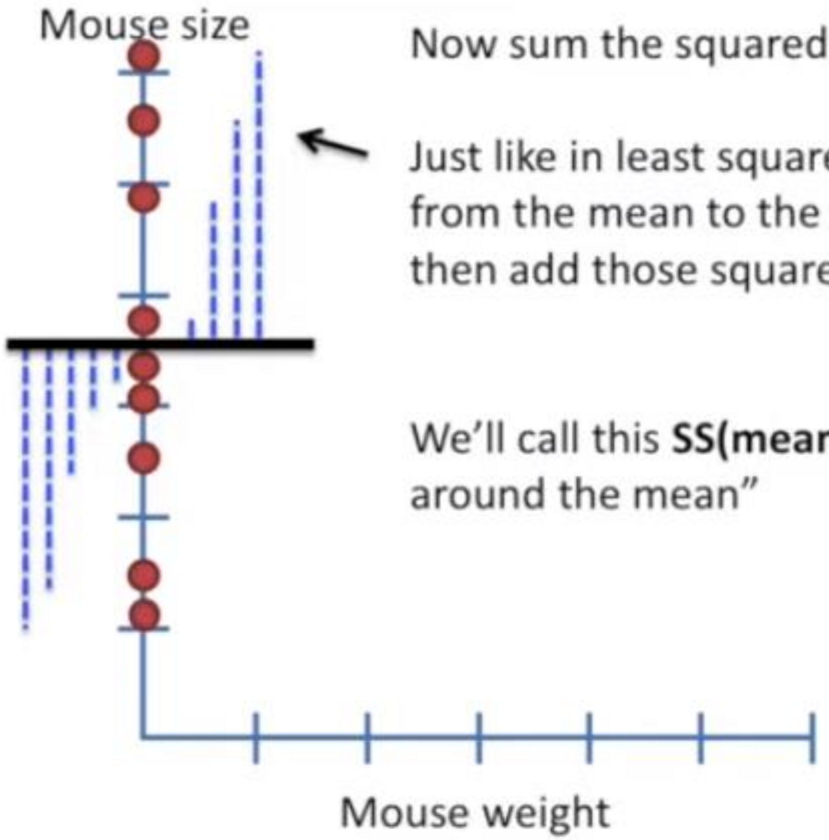


First, calculate the average mouse size.



The average mouse size.

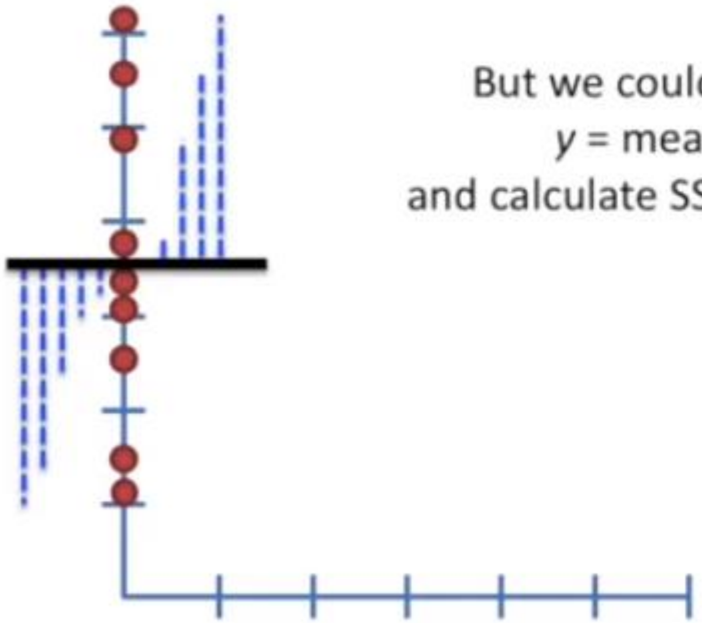
First, calculate the average mouse size.



Now sum the squared residuals…

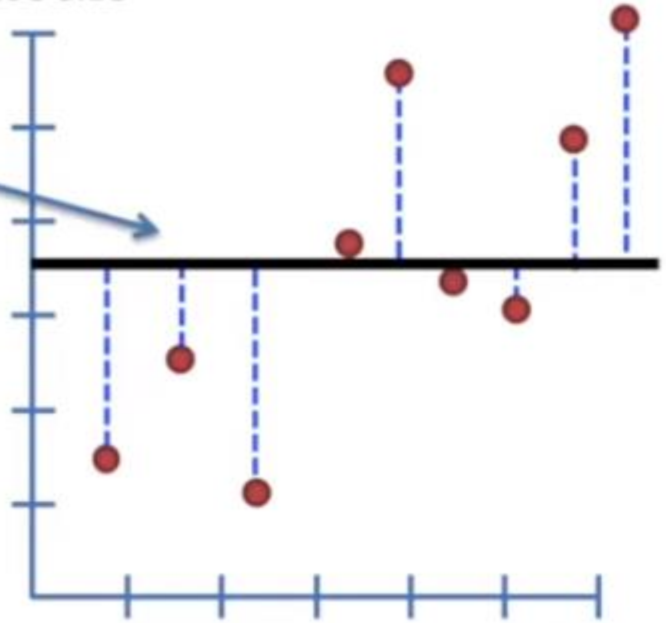
Just like in least squares, we measure the distance from the mean to the data point and square it, then add those squares together.

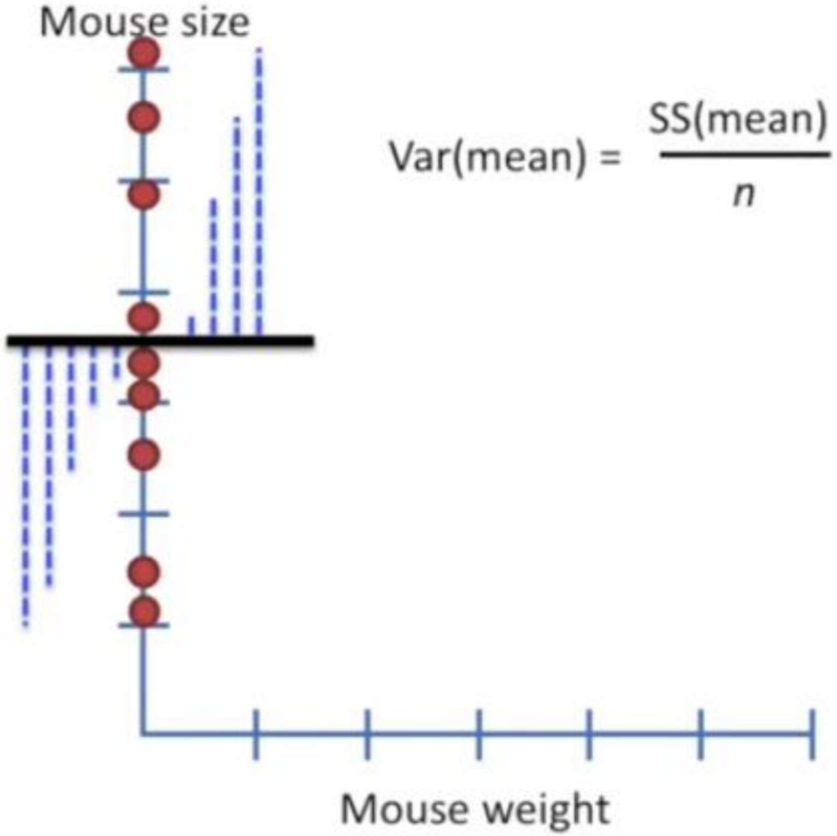
We'll call this **SS(mean)**, for "sum of squares around the mean"



SS(mean)

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



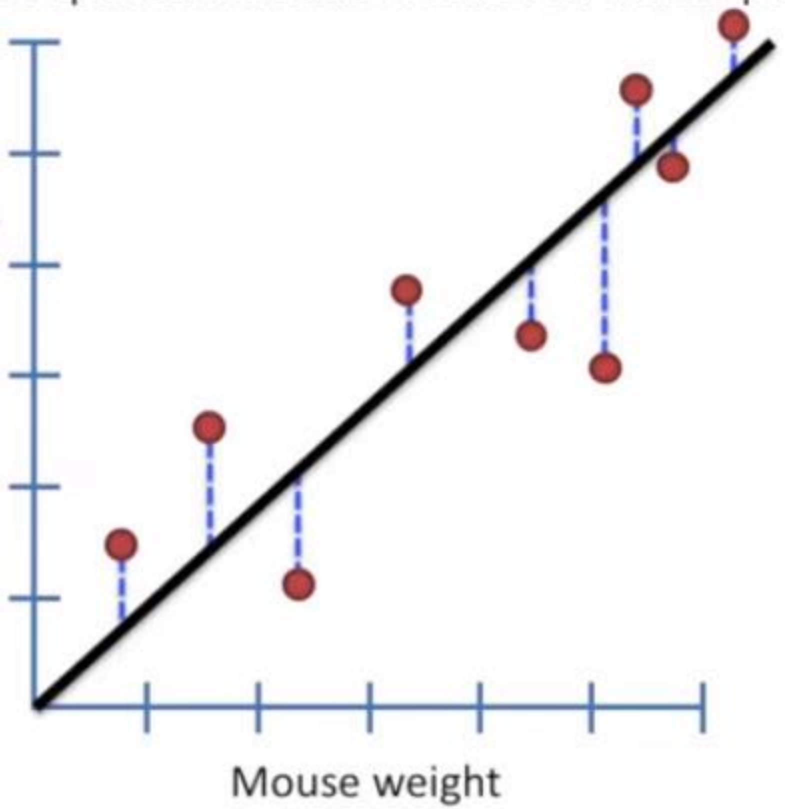


$$\mathrm{Var(mean)} = \frac{\mathrm{SS(mean)}}{n}$$

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

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

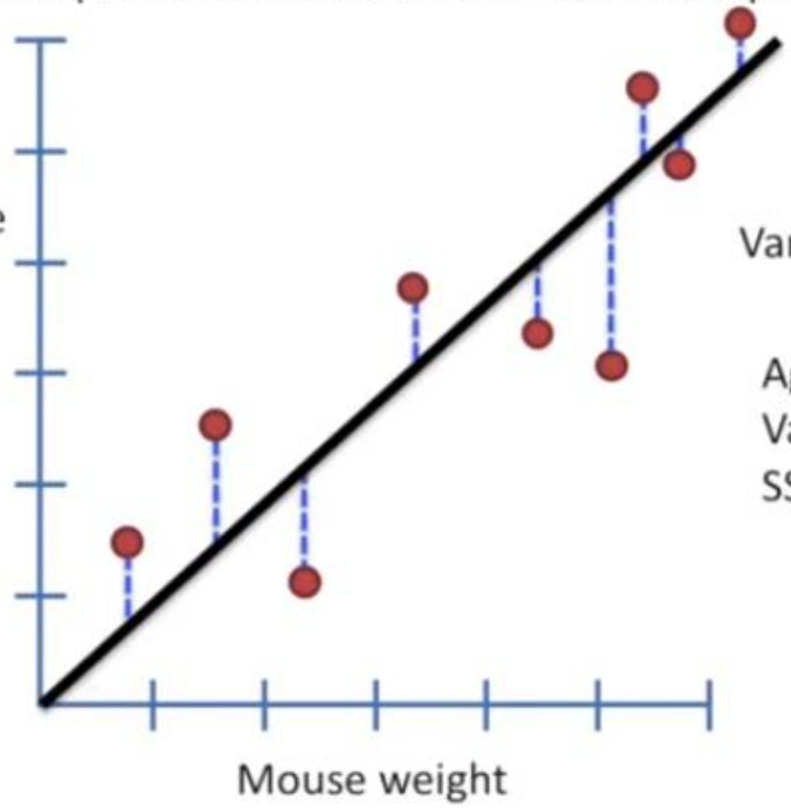
We'll call this **SS(fit)**, for the sum of squares around the least-squares fit.



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

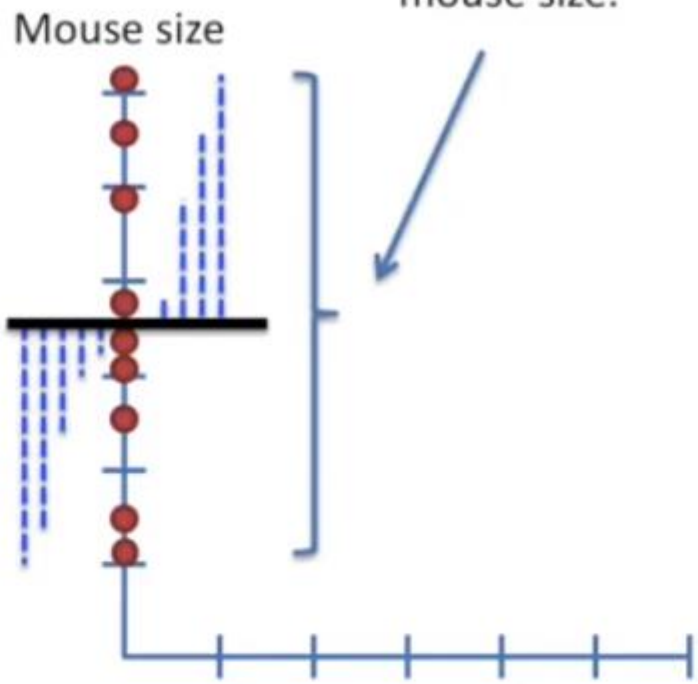
We’ll call this **SS(fit)**, for the sum of squares around the least-squares fit.

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

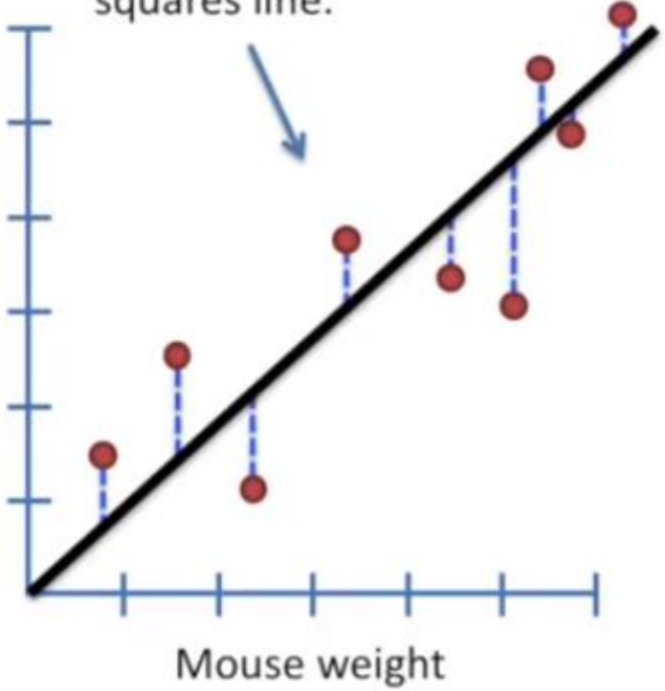


$$Var(fit) = \frac{SS(fit)}{n}$$

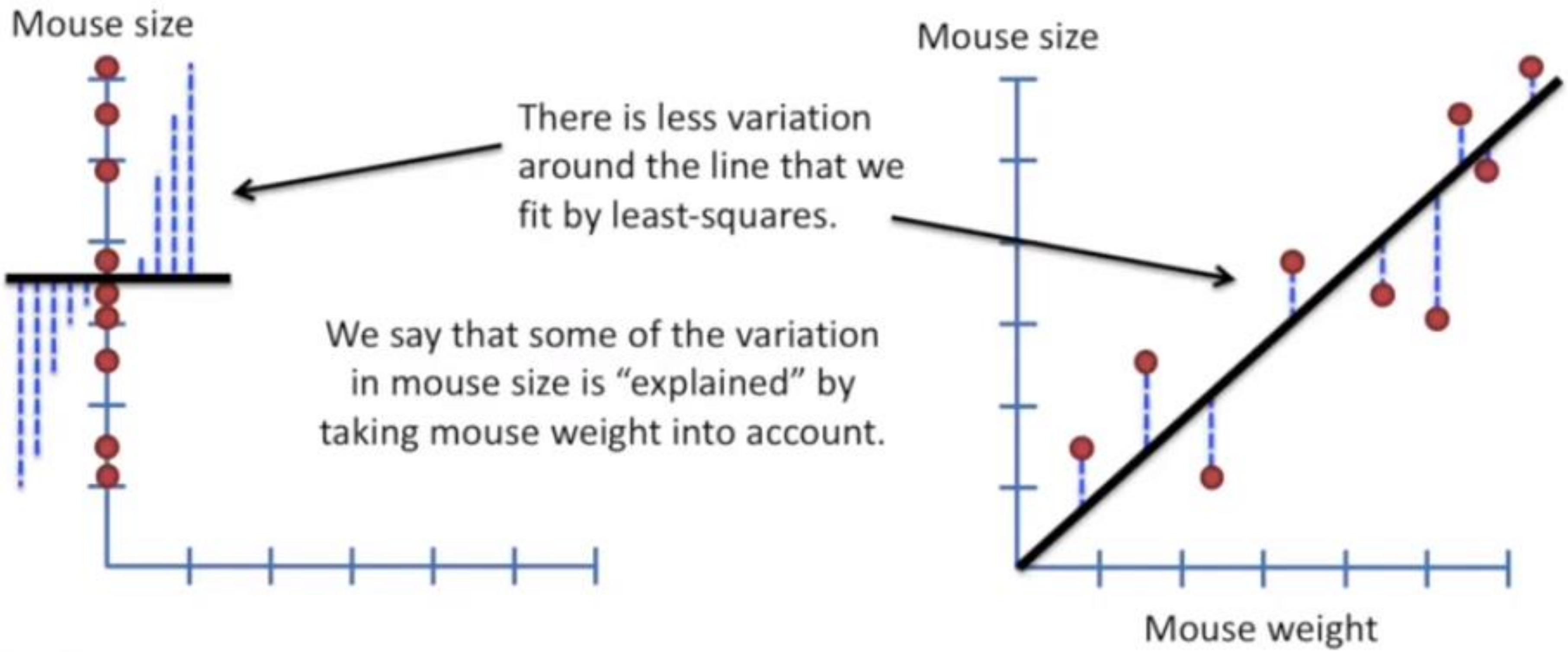
Again, we can think of Var(fit) as the average SS(fit) for each mouse.



This is the raw variation in mouse size.

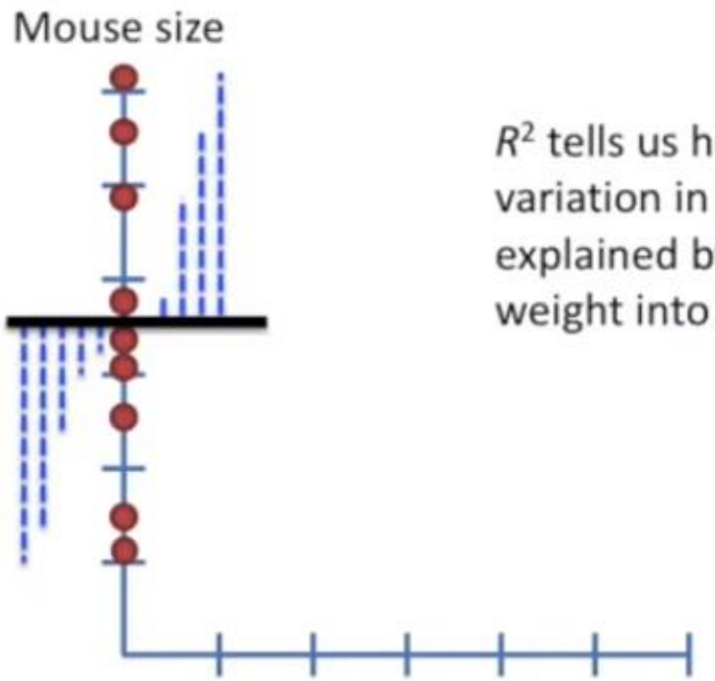


This is the variation around the least squares line.

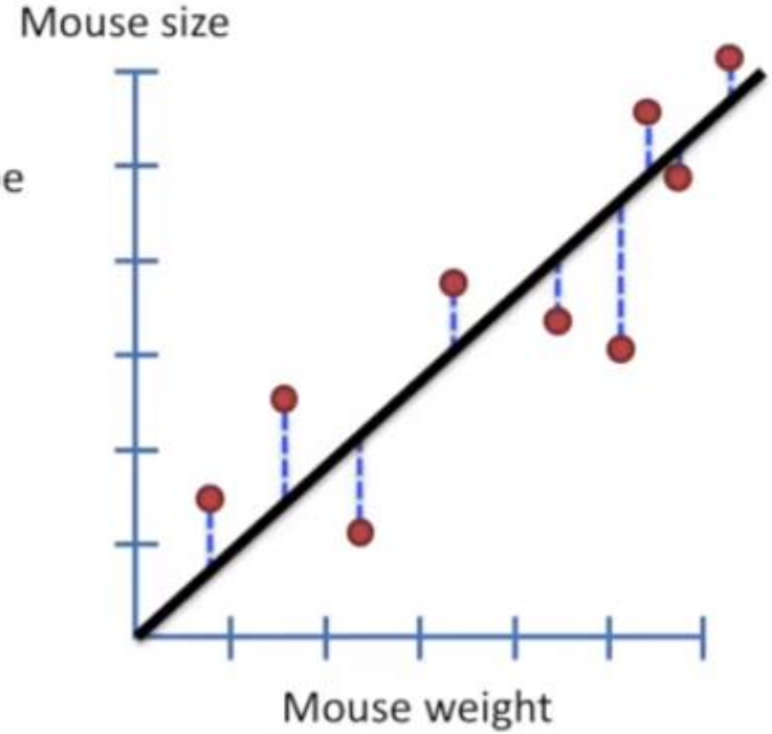


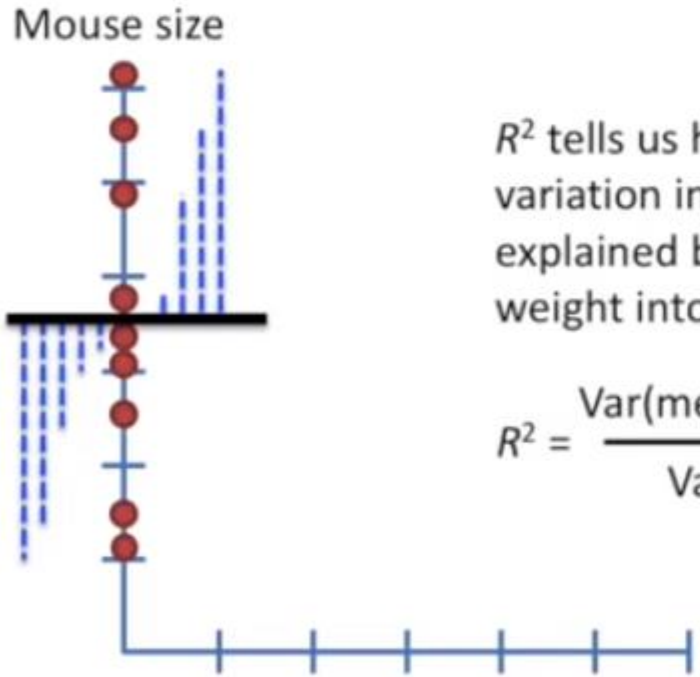
There is less variation around the line that we fit by least-squares.

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



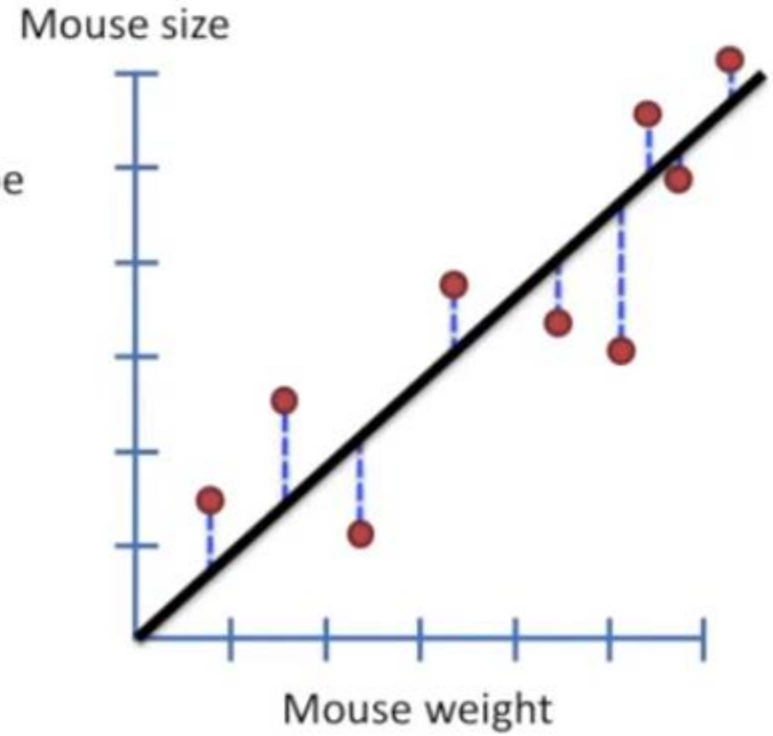
R^2 tells us how much of the variation in mouse size can be explained by taking mouse weight into account.



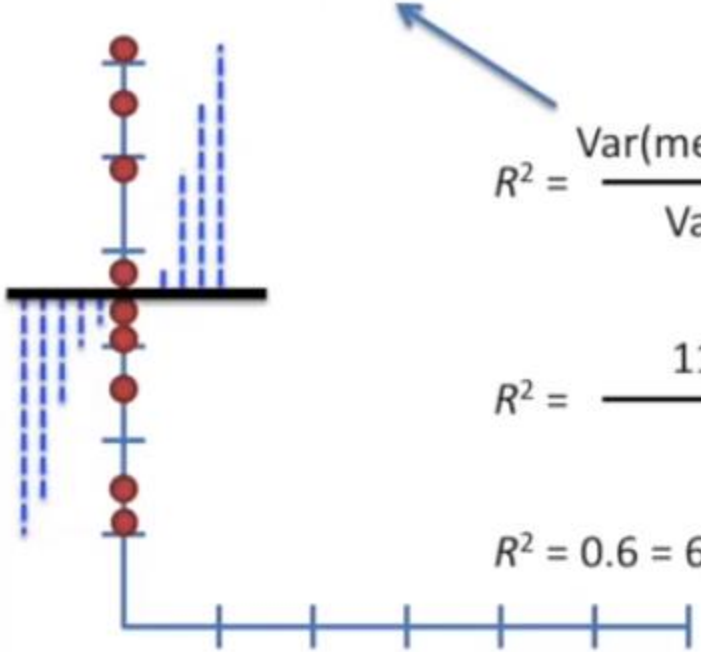


R^2 tells us how much of the variation in mouse size can be explained by taking mouse weight into account.

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



Var(mean) = 11.1

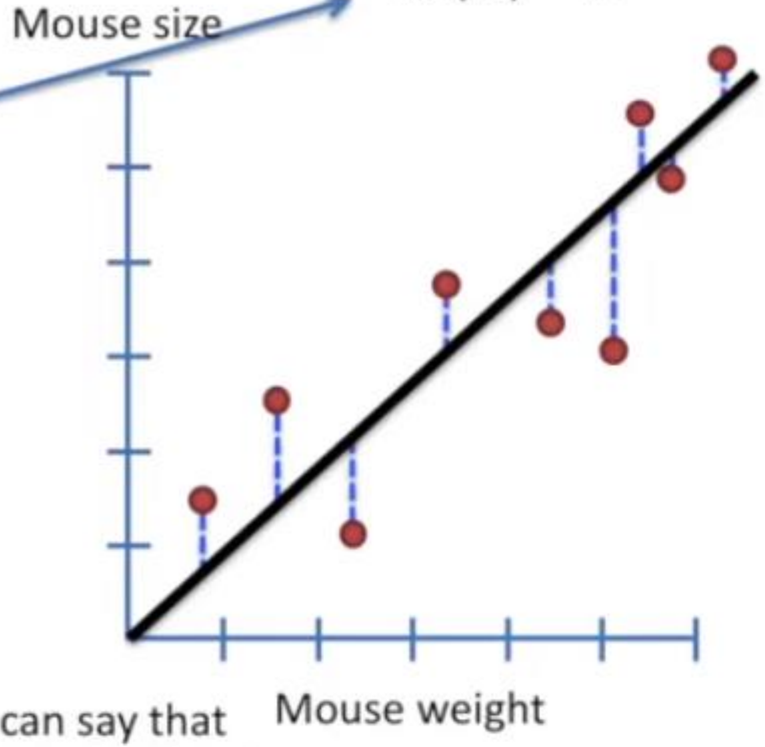


$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

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

$R^2 = 0.6 = 60\%$

Var(fit) = 4.4

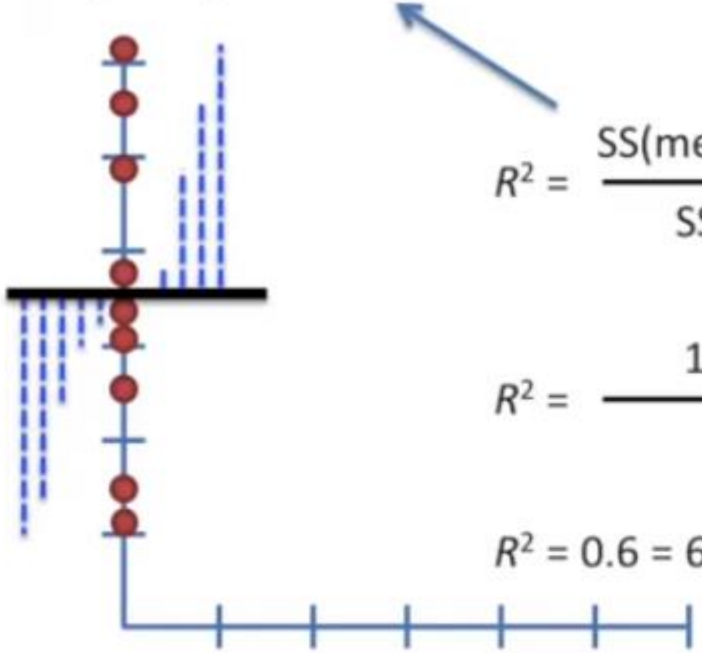


There is a 60% reduction in variance when we take the mouse weight into account.

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

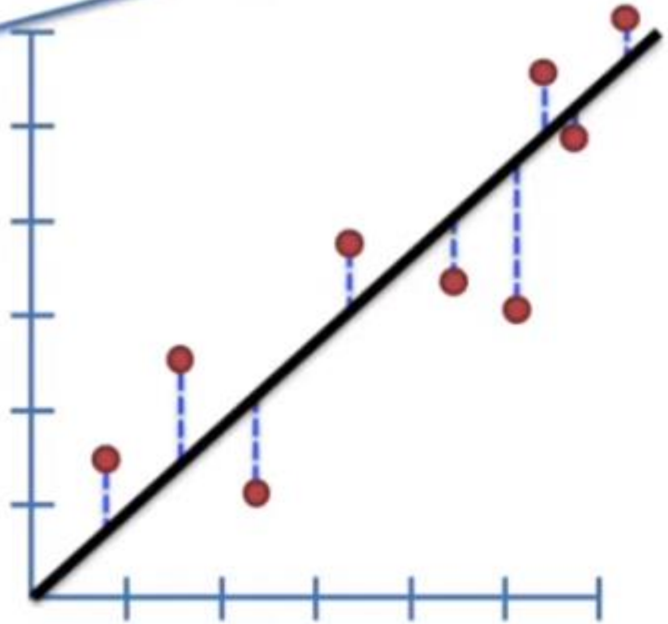


$$R^2 = \frac{SS(\text{mean}) - SS(\text{fit})}{SS(\text{mean})}$$

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

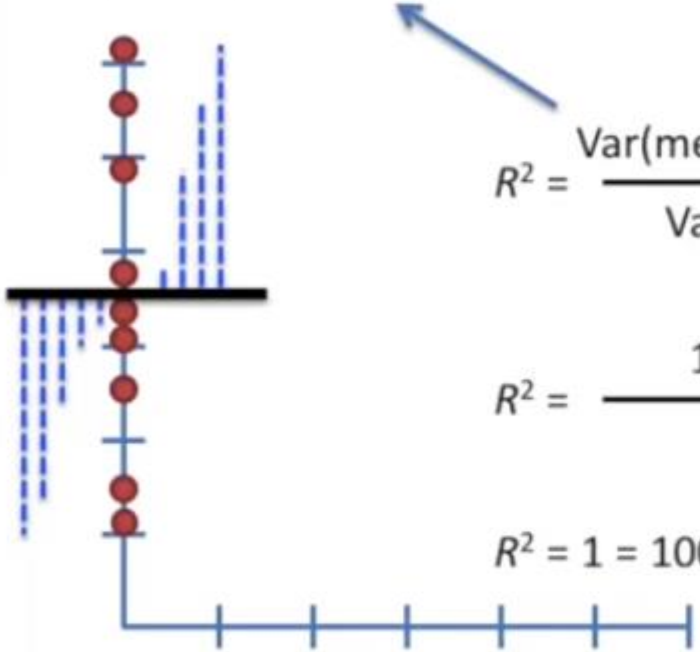
$$R^2 = 0.6 = 60\%$$

SS(fit) = 40



Var(mean) = 11.1

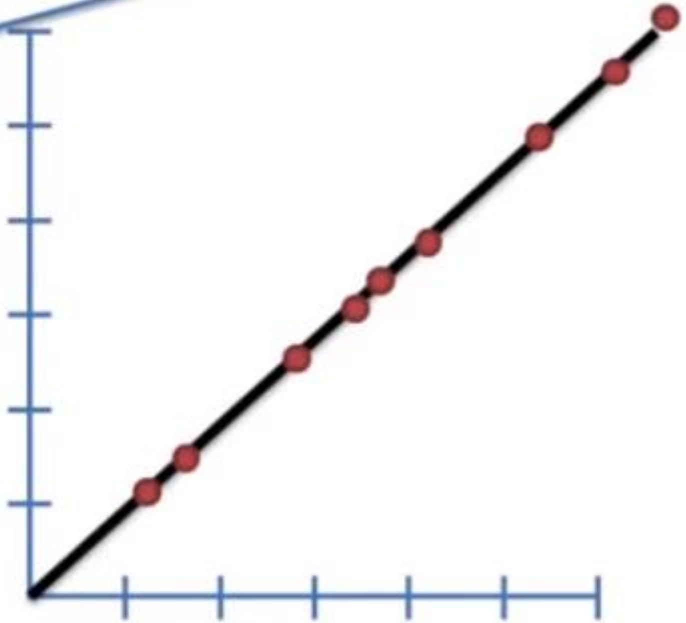
Var(fit) = 0



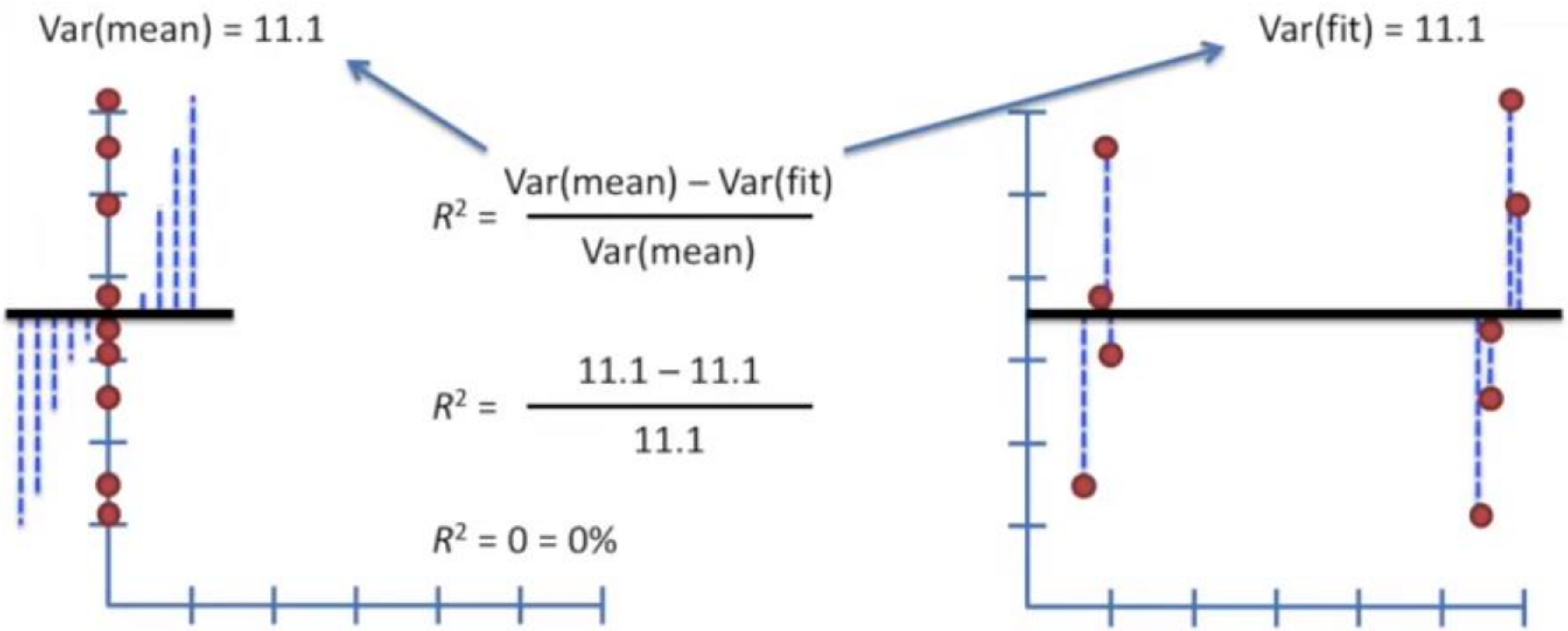
$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

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

$R^2 = 1 = 100\%$



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



Var(mean) = 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(fit) = 11.1

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

Multiple regression

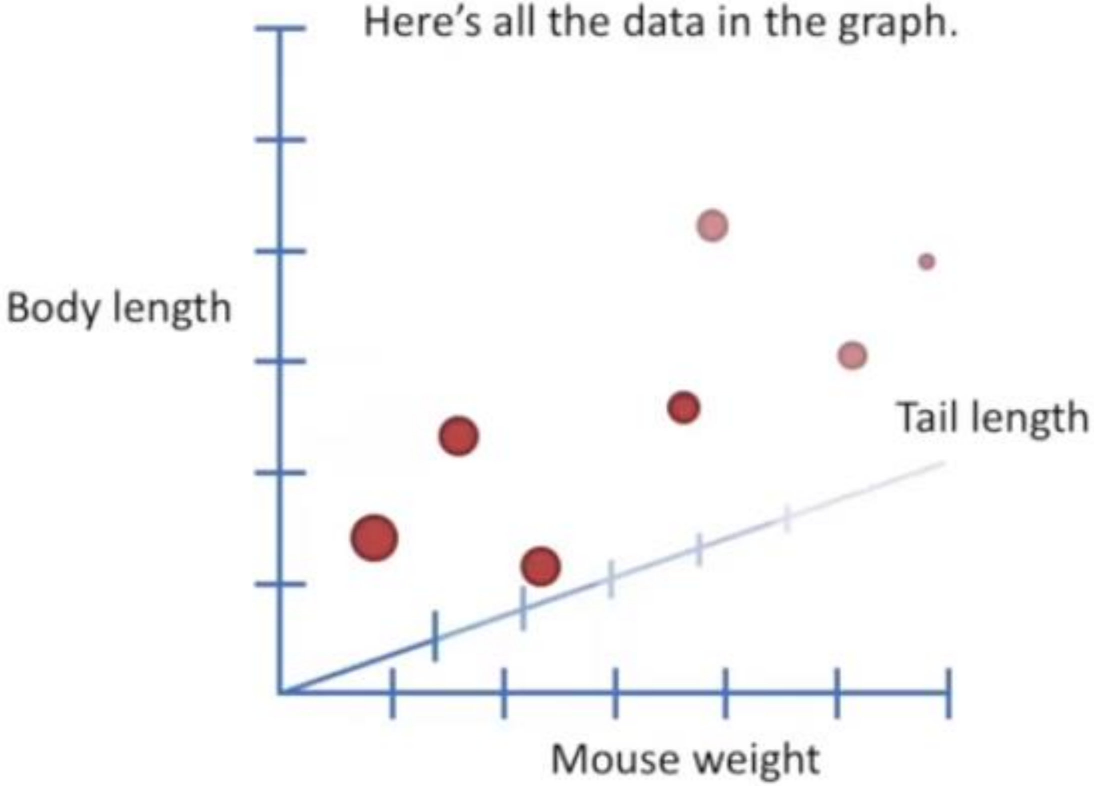
- Till now we have seen a simple regression where we have one attribute or independent variable.
- However, in the real world, a data point has various important attributes and they need to be catered to while developing a regression model.

  - Ex: predicting price of a house, we need to consider various attributes related with this house. Such a regression problem is an example of a multiple regression.
  - This can be represented by :

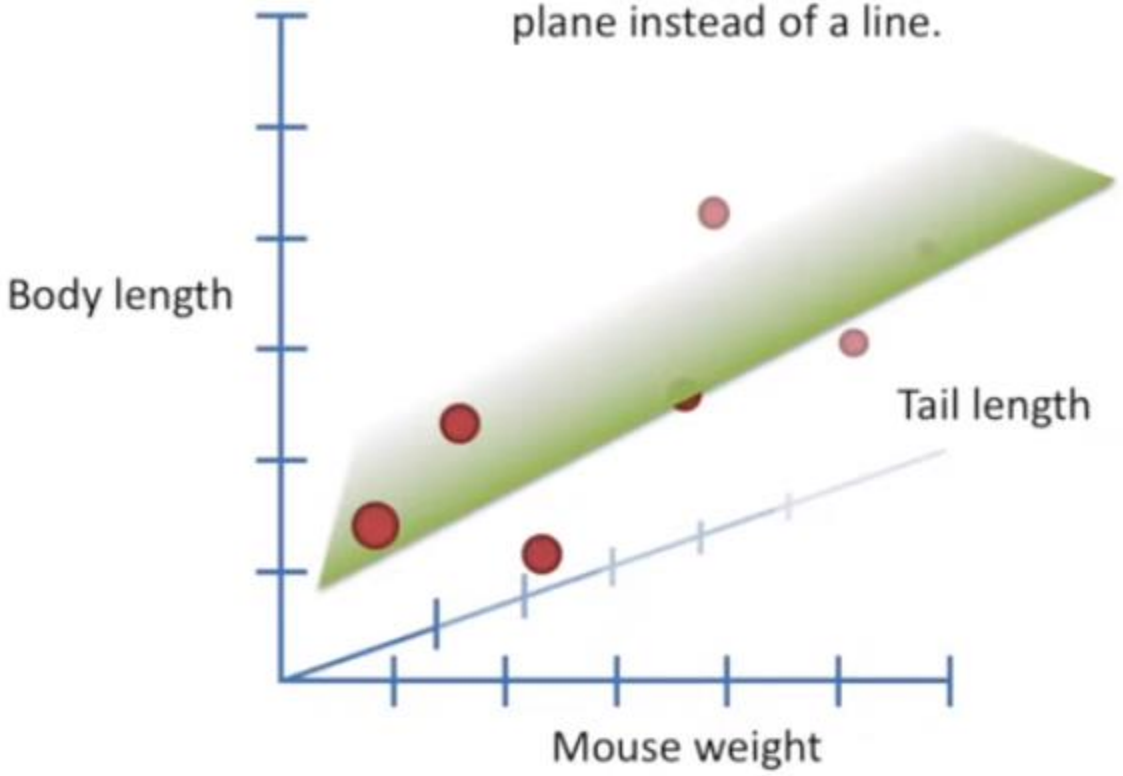
target = constant1*feature1 + constant2*feature2 + constant3*feature3 + …..+ intercept

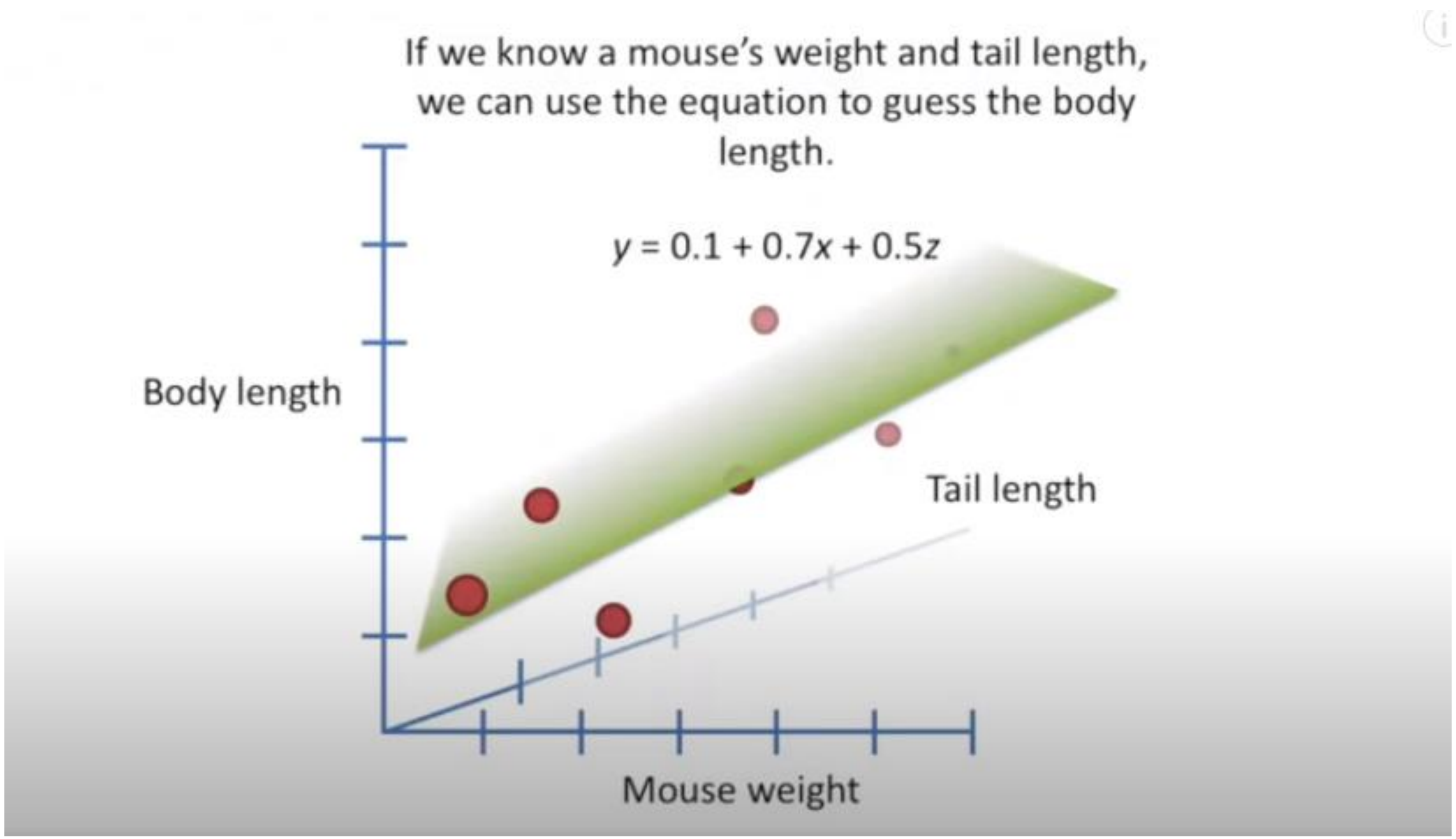
The model aims to find the constants and intercept such that this line is the best fit.

Here’s all the data in the graph.



Now we do a least-squares fit. Since we have the extra term in the equation, we fit a plane instead of a line.





If we know a mouse’s weight and tail length, we can use the equation to guess the body length.

$y = 0.1 + 0.7x + 0.5z$

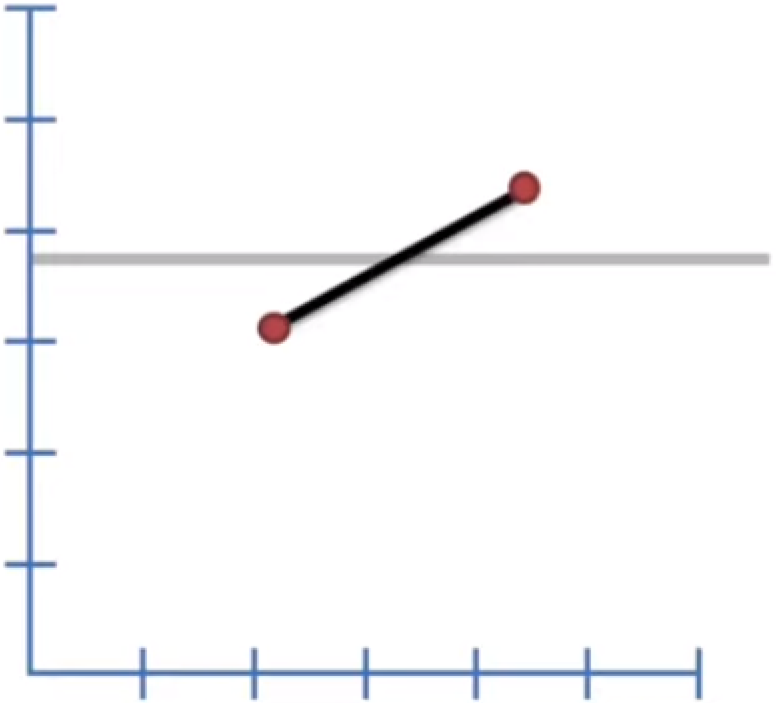
R^2 is awesome, but it's missing something…

SS(mean) = 10

SS(fit) = 0

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

$$= \frac{100 - 0}{100} = 100\%$$



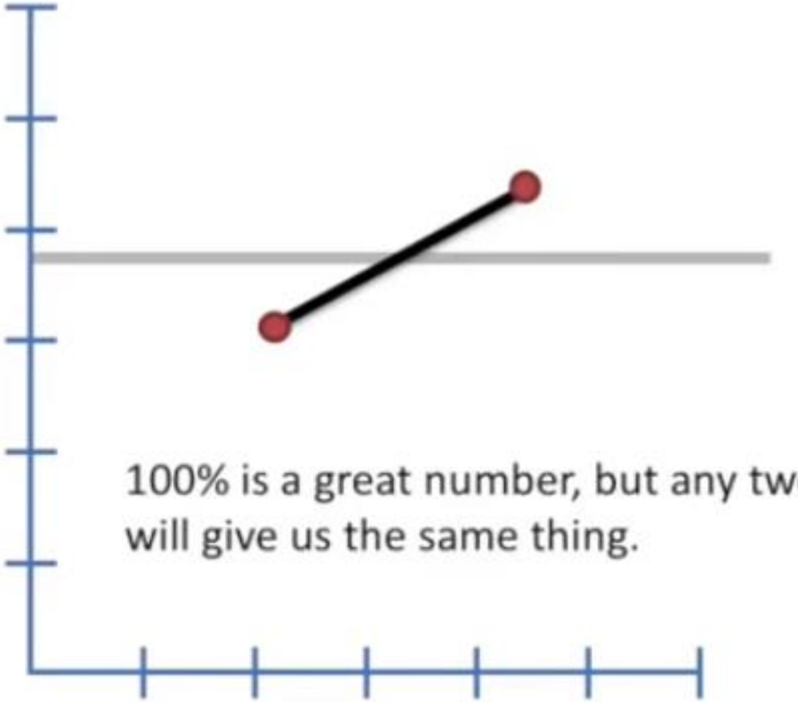
R^2 is awesome, but it's missing something…

SS(mean) = 10

SS(fit) = 0

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

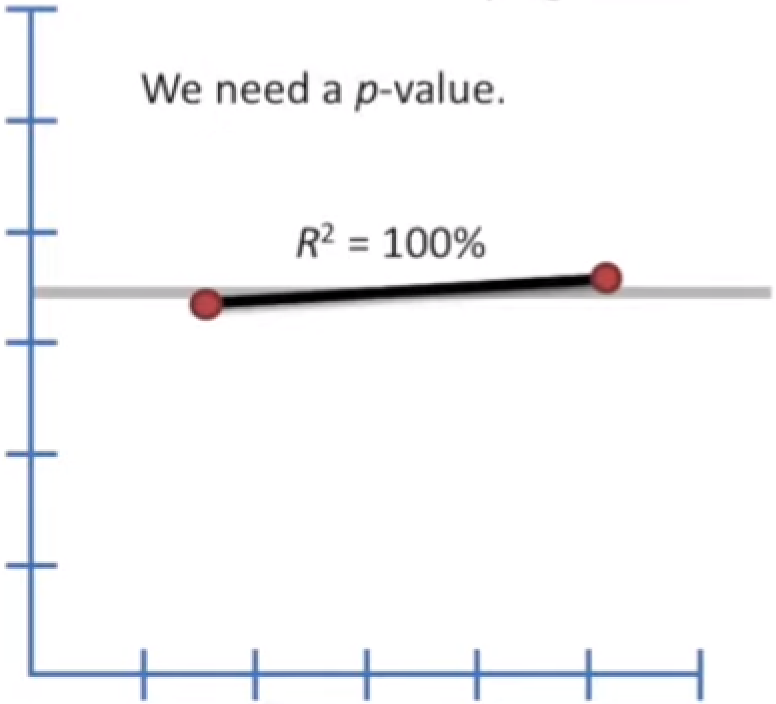
$$= \frac{100 - 0}{100} \boxed{= 100\%}$$

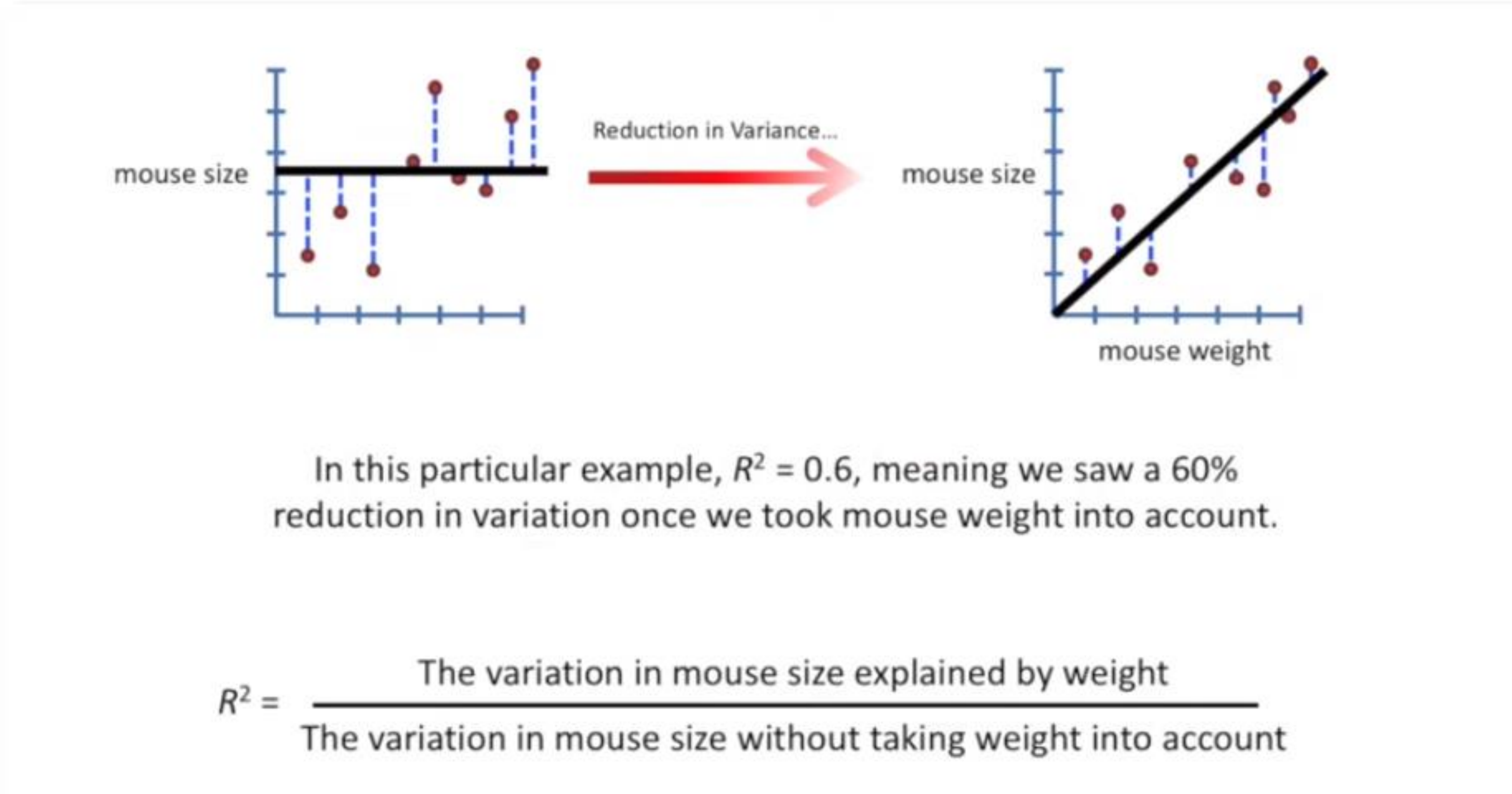


100% is a great number, but any two random points will give us the same thing.

We need a way to determine if the R^2 value is statistically significant.

We need a *p*-value.





In this particular example, $R^2 = 0.6$, meaning we saw a 60% reduction in variation once we took mouse weight into account.

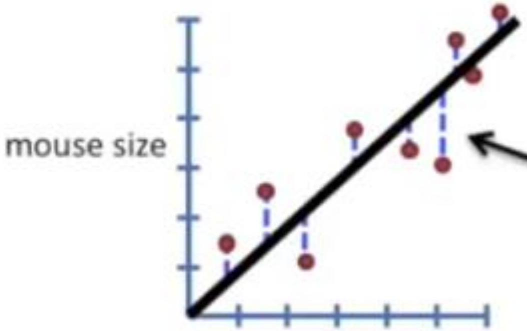
$$R^2 = \frac{\text{The variation in mouse size explained by weight}}{\text{The variation in mouse size without taking weight into account}}$$

$$R^2 = \frac{\text{The variation in mouse size explained by weight}}{\text{Variation in mouse size without taking weight into account}}$$

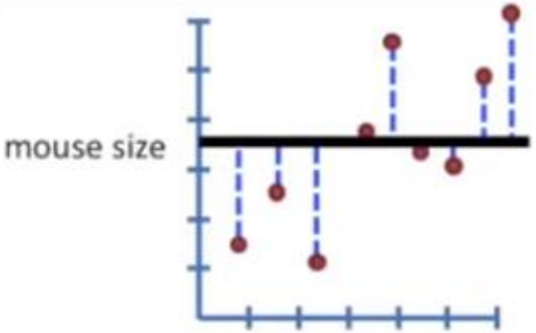
$$F = \frac{\text{The variation in mouse size explained by weight}}{\text{The variation in mouse size not explained by weight}}$$

The *p*-value for R^2 comes from something called "*F*"

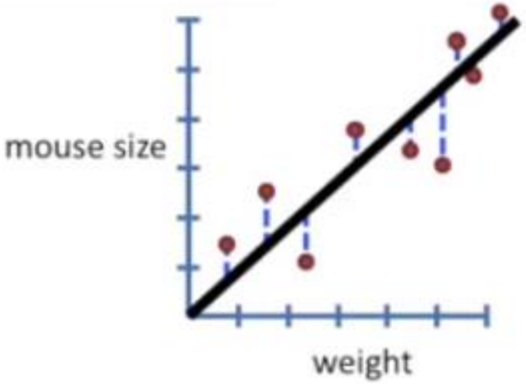
$$F = \frac{\text{The variation in mouse size explained by weight}}{\text{The variation in mouse size not explained by weight}}$$



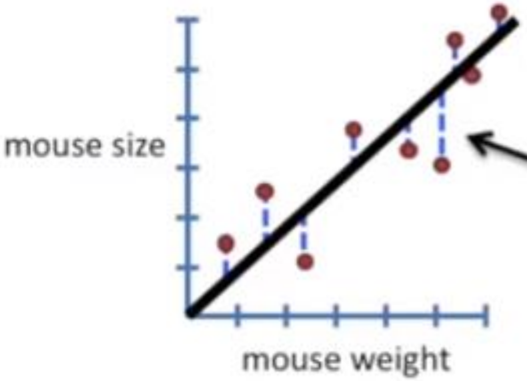
These dotted lines (residuals) represent the variation that remains after fitting the line. **This is the variation that is not explained by weight.**



Reduction in variance when we take weight into account.



$$F = \frac{\text{The variation in mouse size explained by weight}}{\text{The variation in mouse size not explained by weight}}$$



These dotted lines (residuals) represent the variation that remains after fitting the line. **This is the variation that is not explained by weight.**

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

$$\boxed{F = \frac{\text{SS(mean)} - \text{SS(fit)} \,/\, (p_{\text{fit}} - p_{\text{mean}})}{\text{SS(fit)} \,/\, (n - p_{\text{fit}})}}$$

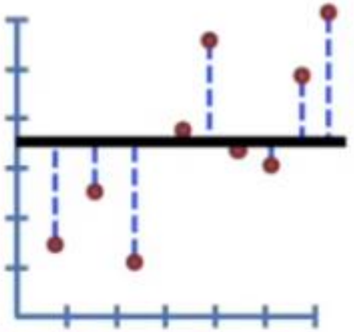
This equation will tell us if R^2 is significant.

$$F = \frac{\mathrm{SS(mean)} - \mathrm{SS(fit)} \;/\; (p_{\mathrm{fit}} - p_{\mathrm{mean}})}{\mathrm{SS(fit)} \;/\; (n - p_{\mathrm{fit}})}$$

↑

These numbers over here are the “degrees of freedom”.

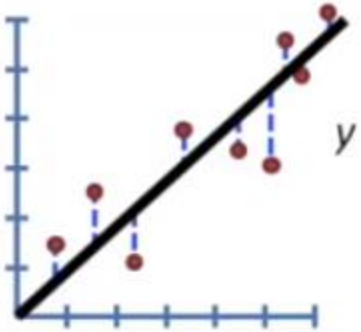
They turn the sums of squares into variances.



y = y-intercept

1 parameter

$p_{\text{mean}} = 1$



y = y-intercept + slope x

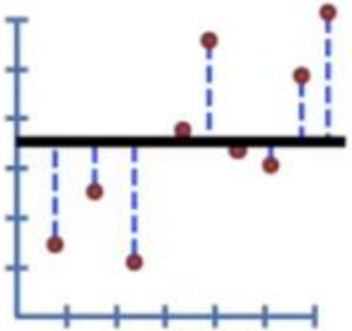
2 parameters

$p_{\text{fit}} = 2$

$$F = \frac{\text{SS(mean)} - \text{SS(fit)} / (p_{\text{fit}} - p_{\text{mean}})}{\text{SS(fit)} / (n - p_{\text{fit}})}$$

p_{fit} is the number of parameters in the fit line…

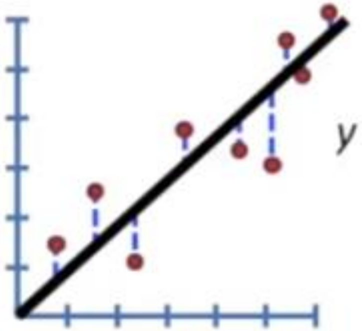
p_{mean} is the number of parameters in the mean line.



y = y-intercept

1 parameter

$p_{\text{mean}} = 1$



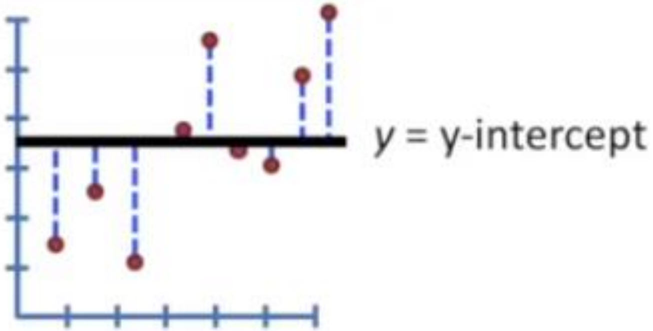
y = y-intercept + slope x

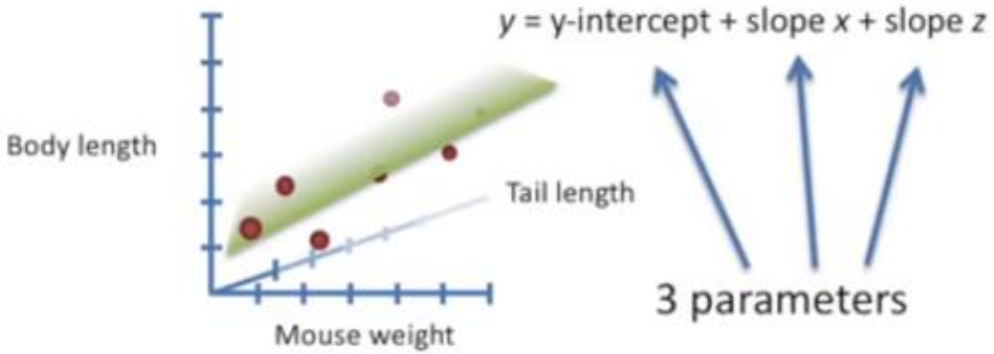
2 parameters

$p_{\text{fit}} = 2$

$$F = \frac{\boxed{SS(\text{mean}) - SS(\text{fit}) / (p_{\text{fit}} - p_{\text{mean}})}}{SS(\text{fit}) / (n - p_{\text{fit}})}$$

Thus, the numerator is the variance explained by the extra parameter. In our example, that's the variance in mouse size explained by mouse weight.





y = y-intercept + slope x + slope z

3 parameters

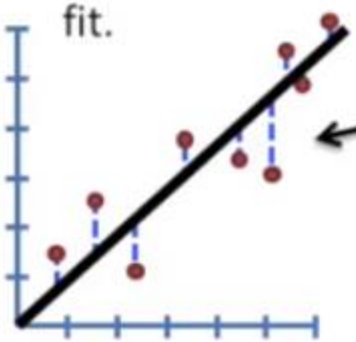
$$F = \frac{\text{SS(mean)} - \text{SS(fit)} / (p_{\text{fit}} - p_{\text{mean}})}{\text{SS(fit)} / (n - p_{\text{fit}})}$$

$p_{\text{fit}} = 3$

$(p_{\text{fit}} - p_{\text{mean}}) = (3 - 1) = 2 =$ Now the fit has two extra parameters, mouse weight and tail length.

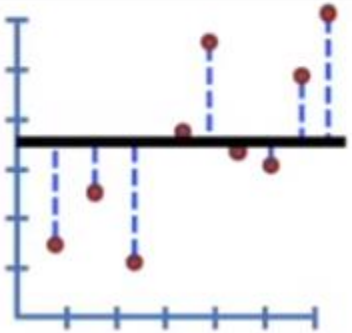
The variation in mouse size not explained by the fit.

$$F = \frac{SS(mean) - SS(fit) / (p_{fit} - p_{mean})}{SS(fit) / (n - p_{fit})}$$

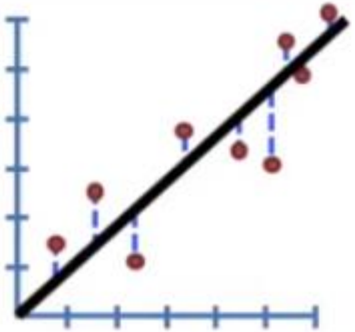


Why divide SS(fit) by $n - p_{fit}$ instead of just n?

Intuitively, the more parameters you have in your equation, the more data you need to estimate them. For example, you only need two points to estimate a line, but you need 3 points to estimate a plane.



If the “fit” is good, then…

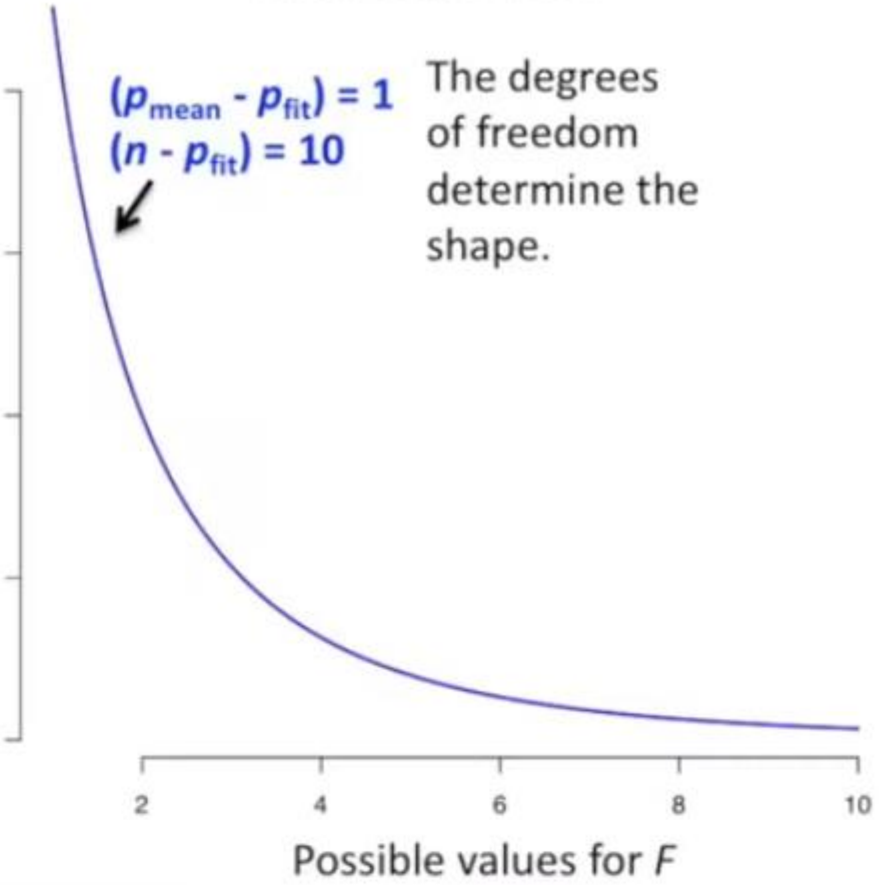


$$F = \frac{\text{The variation explained by the extra parameters in the “fit”}}{\text{The variation not explained by the extra parameters in the “fit”.}} \rightarrow \frac{\text{large number}}{\text{small number}}$$

F = really large number

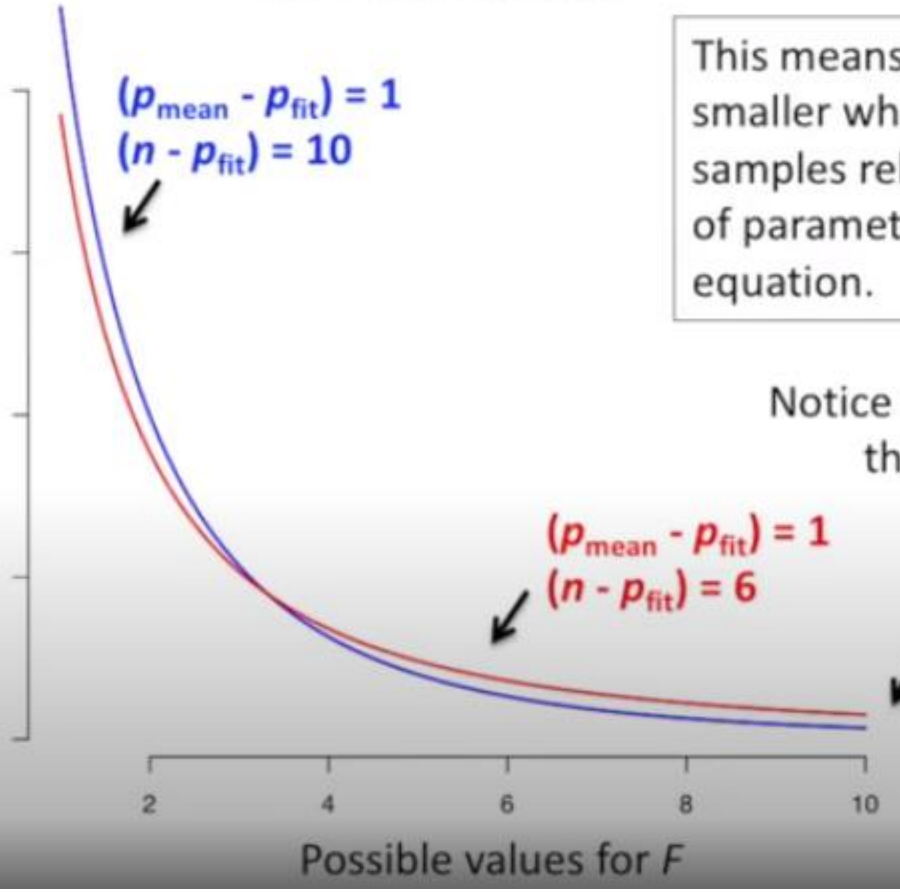
How do we turn this number in to a *p*-value?

One *F*-distribution



The degrees of freedom determine the shape.

Two F-distributions



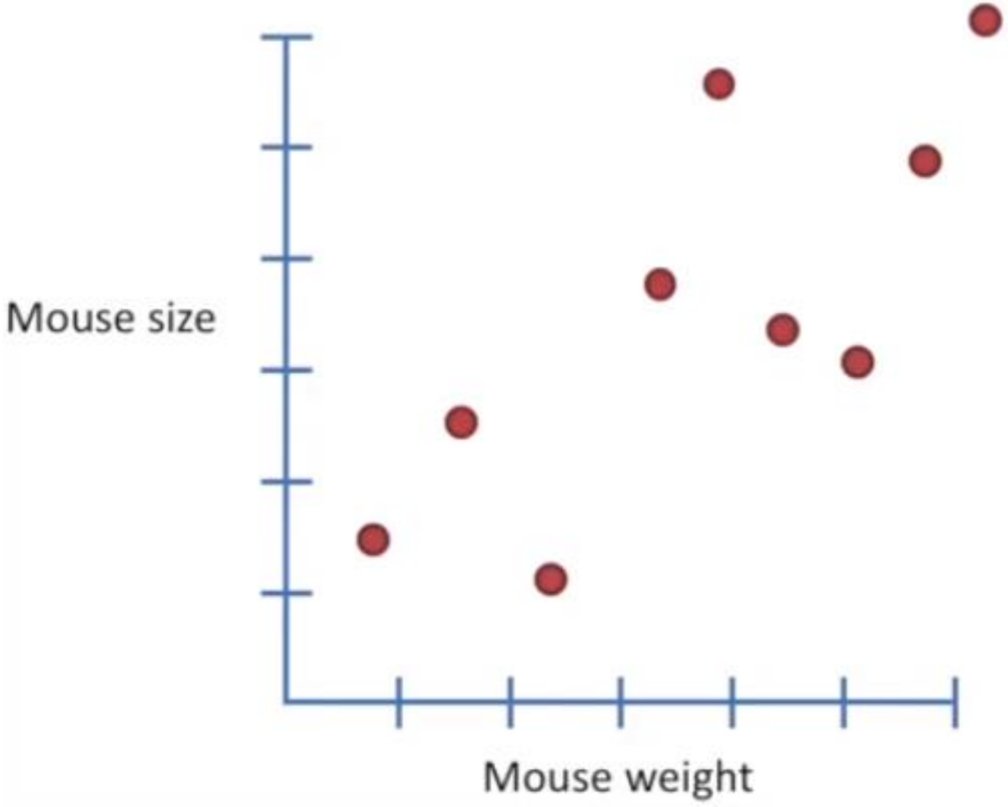
$(p_{\text{mean}} - p_{\text{fit}}) = 1$
$(n - p_{\text{fit}}) = 10$

$(p_{\text{mean}} - p_{\text{fit}}) = 1$
$(n - p_{\text{fit}}) = 6$

This means the p-value will be smaller when there are more samples relative to the number of parameters in the fit equation.

Notice that when $(n - p_{\text{fit}}) = 10$, the distribution tapers off faster.

Given some data that you think are related...



Linear regression:

1) Quantifies the relationship in the data (this is R^2).
    1) This needs to be large.

2) Determines how reliable that relationship is (this is the *p*-value that we calculate with *F*).
    1) This needs to be small.

You need both to have an interesting result!!!

Pros and Cons of Linear Regression

Advantages

□ Simple to implement and easier to interpret the outputs coefficient.

Disadvantages

□Assumes a linear relationships between dependent and independent variables.

□Outliers can have huge effects on regression.

□Linear Regression assume independence between attributes.