Linear Regression

Ahman Smith

Main Ideas

1. Use least-squares in order to fit a line to the data
2. Calculate R^2
3. Calculate p-value for R^2

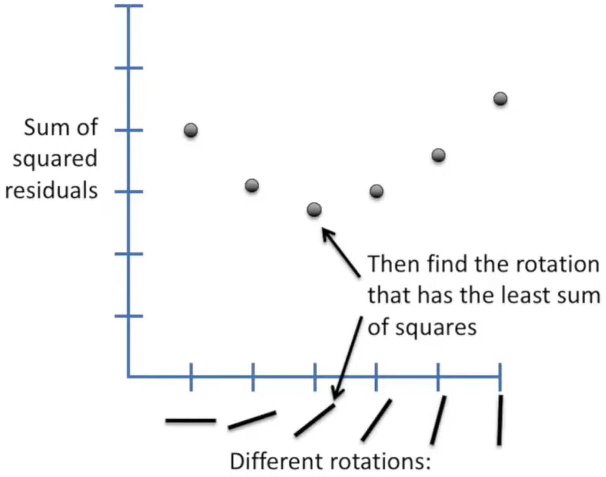
Fitting a Line — Least Squares

Draw a line to the data. Measure the distance from line to data — square each distance and add them up.

Residual: distance from datapoint to line

Rotate the line a bit, with the new line, measure the residuals, square them, then sum up the squares.

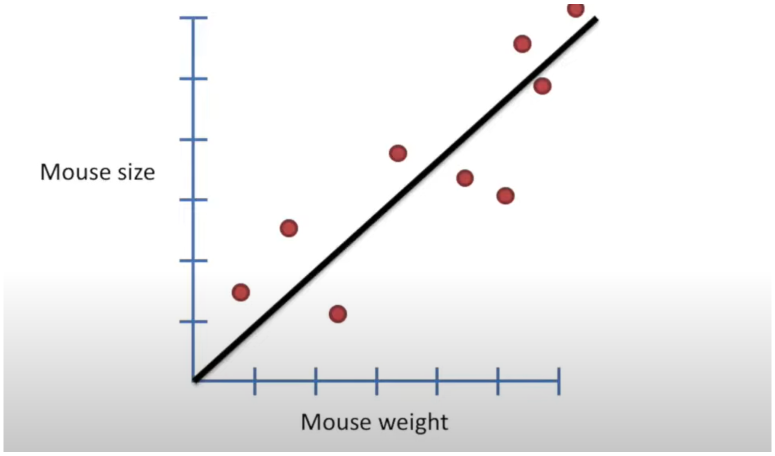
Rotate the line, sum up squares, etc.



Find the rotation that has the lowest value. It’ll be the one to fit to the data

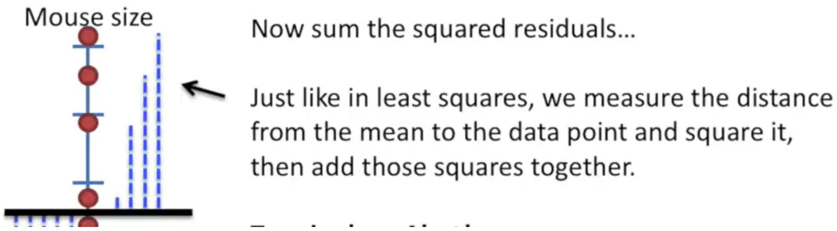
Equation: y intercept + slope

How good is our equation?



In this example, first, calculate average mouse size

- Sum squared residuals



Now sum the squared residuals…

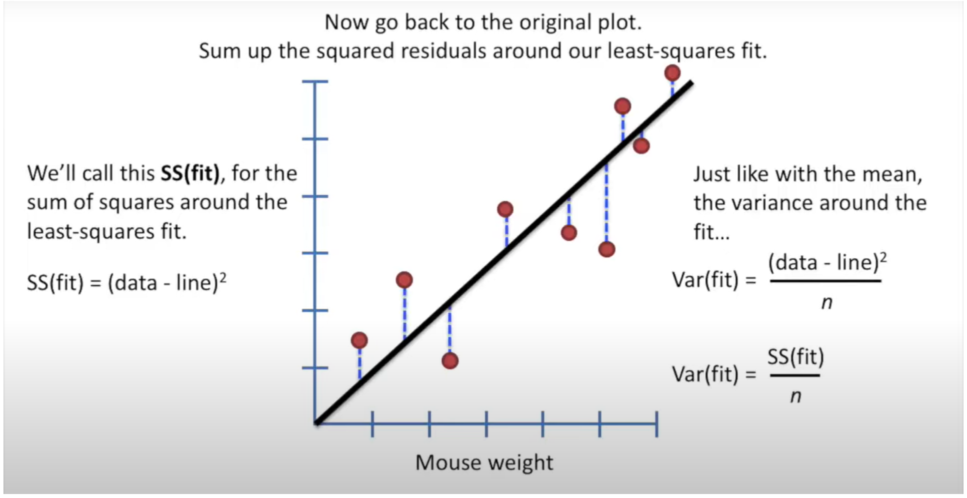
Just like in least squares, we measure the distance from the mean to the data point and square it, then add those squares together.

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

Var(mean) = SS(mean) / n (sample size)

Then, calculate the squared residuals around least-squares fit

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



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…

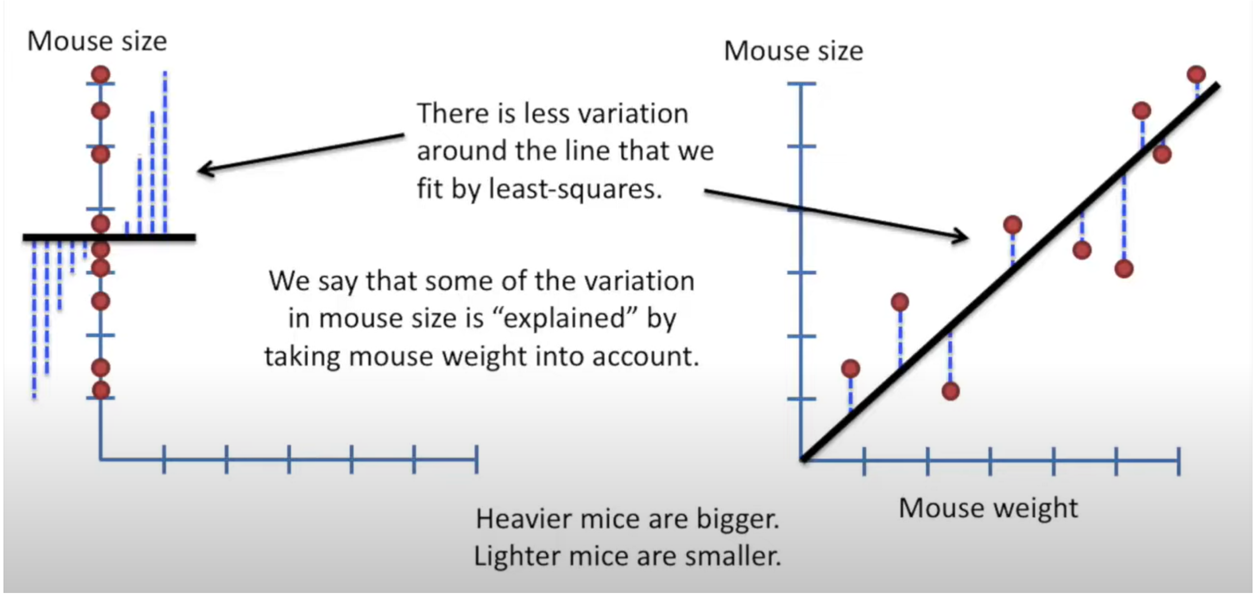
$$\text{Var(fit)} = \frac{(\text{data - line})^2}{n}$$

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

Important

In general, variance(x) = sum of squares / len(x) —> average sum of squares

Continuation



- Determines that there is a correlation between the two variables

R^2

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

R^2 = var(mean) - var(fit) / var(mean)

In this given example, var(mean) = 11.1 and var(fit) = 4.4

11.1 - 4.4 / 11.1 = 0.6, meaning that there is a 60% reduction in variance if we take mouse weight into account

- Mouse weight “explains” 60% of the variation in mouse size

Different Approach

To make the same calculation, we can use the raw sum of squares

SS(mean) = 100

SS(fit) = 40

100 - 40 / 100 = 0.6 or 60%

R^2 shortcoming

What if we only had 2 points?

ss(mean) = 10

ss(fit) = 0

ss(fit) = 0 because you can always draw a straight line to connect any 2 points

r^2 = 100%

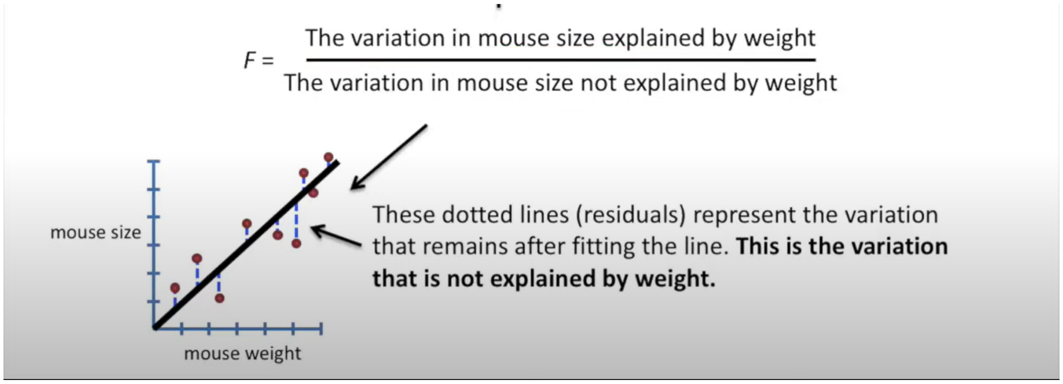
This explains the variation, but any two points will give us the exact same thing.

Goal: Strategy to determine whether R^2 is statistically significant

Solution: Determine a p-value

What is a p-value?

- comes from something called **F**
- F = the variation in mouse size explained by weight / the variation in mouse size not explained by weight
- numerators are the same
- reduction in variance when we take the 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.**

P-values

- look at the likelihood of observing the relationship between predictor variables and the response variable **simply by chance**
- small p-value suggests that there is strong evidence of a correlative relationship
- large p value means weak evidence

Linear Regression is used to predict a continuous numerical outcome. It is not used for classification