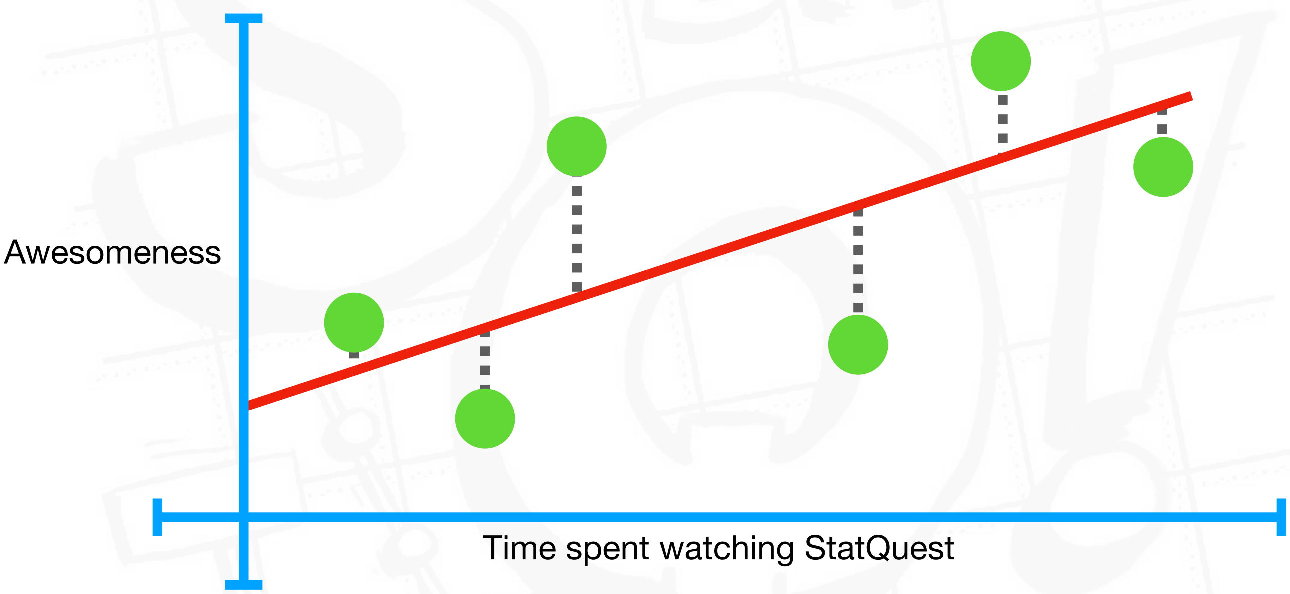
StatQuest!!!

Linear Regression



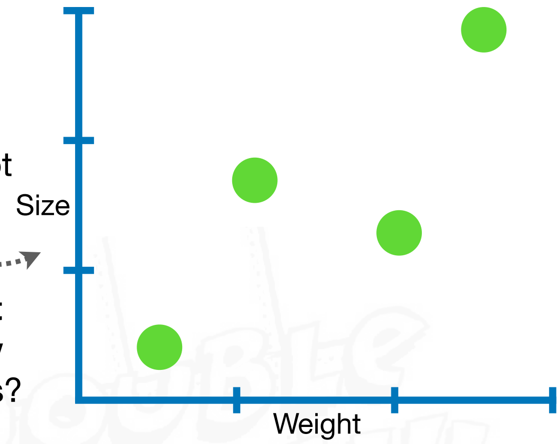
Study Guide!!!

The Problem

	Weight	Size
1	6	8
2	2.5	2
3	13	14
4	10	7

We measured the **Weight** and **Size** of 4 mice…

…and we can plot the data…



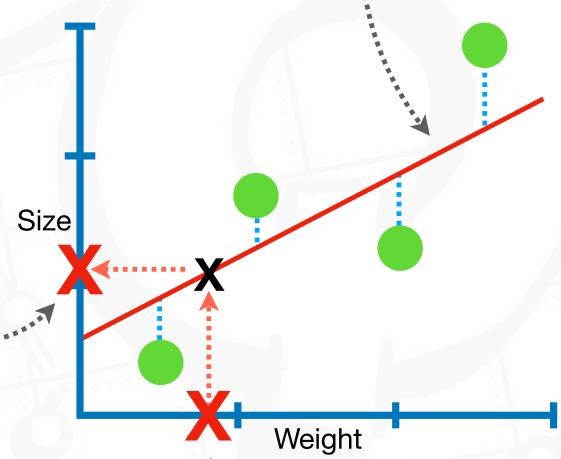
…and it looks like **Weight** and **Size** are related. How can we test this hypothesis? Is the relationship useful?

The Solution - Linear Regression

The slope of this regression line tells us that as **Weight** increases, so does **Size**. This suggests there is a relationship between **Weight** and **Size**.

R^2 tells us the strength of the relationship between **Weight** and **Size**.

The **p-value** for ***R^2*** tells us if the relationship could be the result of random chance or not.



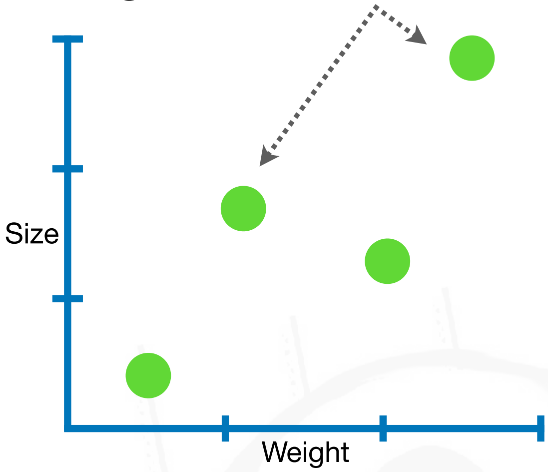
We can use the the regression line to make predictions. For example, this **Weight**…

…corresponds to this predicted **Size**.

BAM!

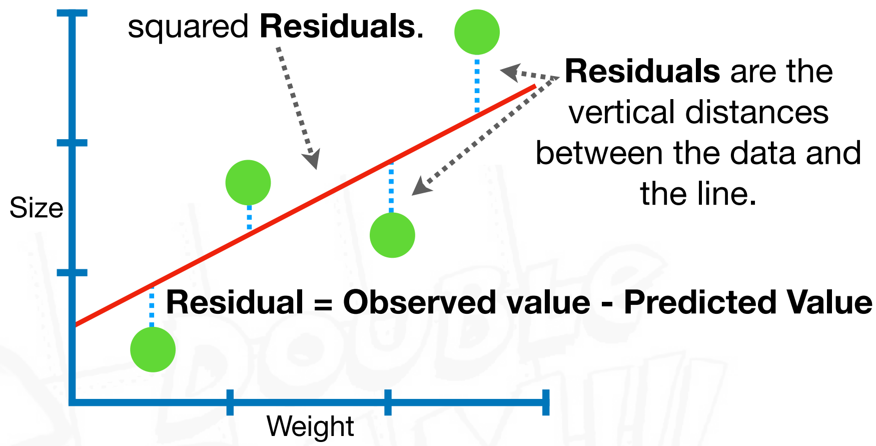
NOTES:

Fitting a line to the data



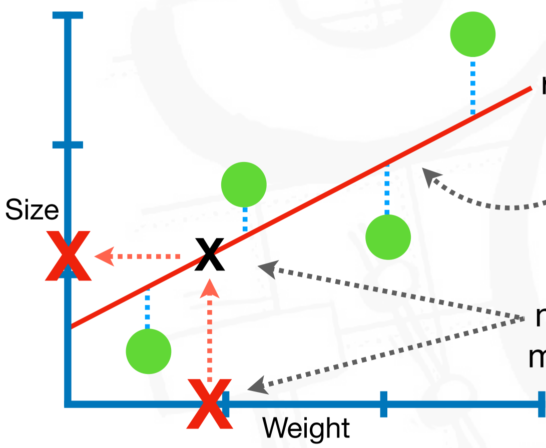
The **Regression Line** minimizes the sum of the squared **Residuals**.

Residuals are the vertical distances between the data and the line.



Residuals are squared for two reasons: 1) So that negative values do not cancel out positive values when we add them together, 2) Unlike the absolute value, the derivative of a squared number exists for all values, making the math easier.

Interpreting the Regression Line

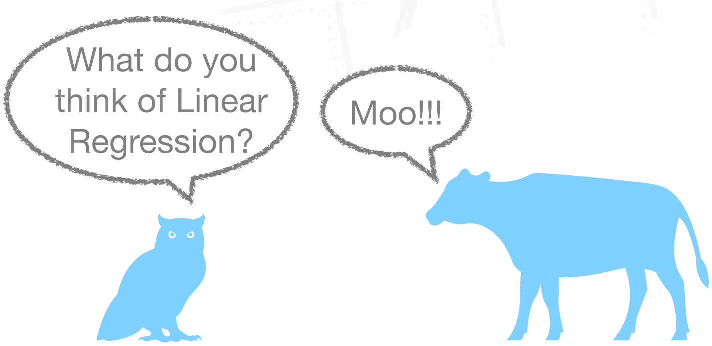


The line suggests a relationship between **Weight** and **Size**. The larger the **Weight**, the larger the **Size**.

In other words, knowing a mouse's **Weight** will help us make a guess about its **Size**.

We quantify how good that guess will be with R^2, which tells us the accuracy of the guess, and its **p-value**, which tells us if we should believe the guess in the first place.

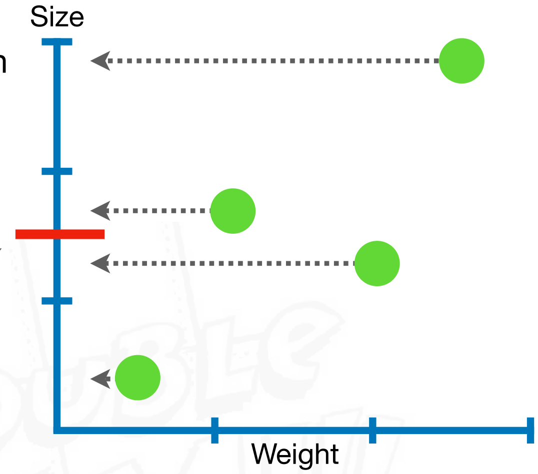
NOTES:



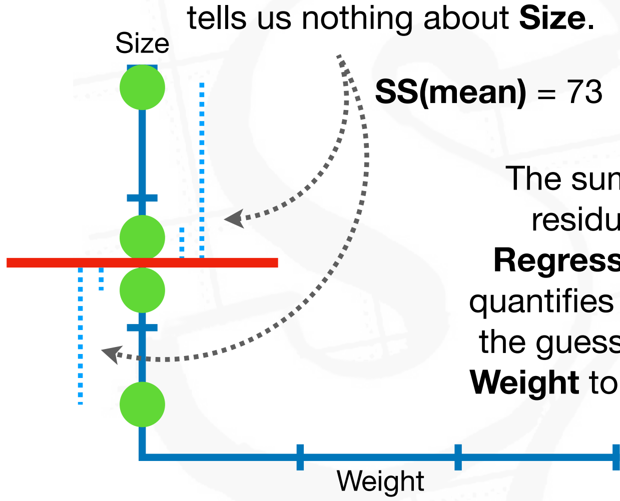
Calculating R^2

If there was no relationship between **Weight** and **Size**, and knowing **Weight** would not help you predict **Size**…

…then the best prediction of **Size** would be the average **Size**.

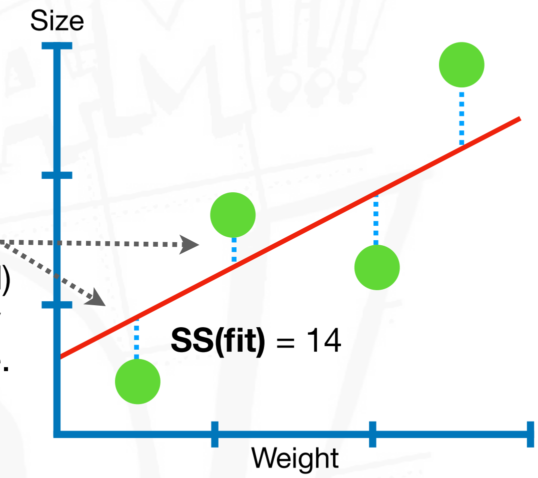


The sum of the squared residuals around the mean for **Size**, **SS(mean)**, quantifies how bad (or good) the guess is when we assume that **Weight** tells us nothing about **Size**.



SS(mean) = 73

The sum of the squared residuals around the **Regression Line**, **SS(fit)**, quantifies how bad (or good) the guess is when we allow **Weight** to tell us about **Size**.



SS(fit) = 14

R^2 compares the guesses using only the mean, **SS(mean)**, to the guesses made with the regression line, **SS(fit)**

By dividing **SS(mean)** and **SS(fit)** by the number of observations, n, we get the variances.

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

Plugging in **SS(mean) = 73** and **SS(fit) = 14** from the data…

$$R^2 = \frac{73 - 14}{73} = \frac{\frac{73}{4} - \frac{14}{4}}{\frac{73}{4}} = \frac{18 - 3.5}{18} \boxed{= 0.81}$$

When R^2 = **0**, the average **Size** (or average y-axis variable, whatever that happens to be) is the best guess we can make. When R^2 **= 1**, the **Regression Line** makes perfect guesses.

This tells us that there is **81%** less variation around the **Regression Line** than around the mean value for **Size**.

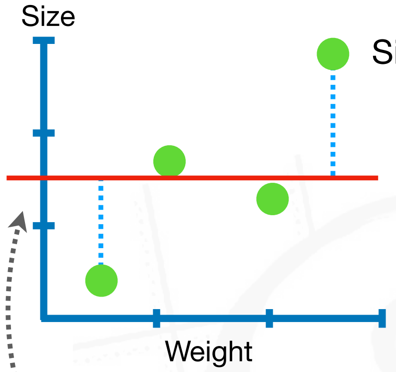
Alternatively, we can say that **Weight** “explains” **81%** of the variation in **Size**.

Calculating the p-value for R^2

When we ignore **Weight** when predicting **Size**, the equation for making predictions is simply…

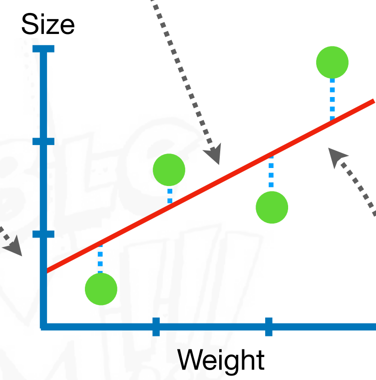
Size = y-intercept

…where the y-intercept is the average **Size.**



When we use **Weight** predict **Size**, the equation for making predictions is…

Size = y-intercept + (slope × Weight)



We use an **F-statistic** to compare the equation that *excludes* **Weight** as a predictor to the equation that *includes* **Weight** as a predictor.

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

NOTE: SS(mean) and **SS(fit)** are defined in **Calculating R^2.**

p_{mean} **= 1** because the y-intercept is the only parameter for this line.

p_{fit} **= 2** because the y-intercept and slope are parameters for this line.

The Numerator is the average variance explained per extra parameter that the **Regression Line** uses.

$$\frac{\boxed{SS(\text{mean}) - SS(\text{fit})}}{\boxed{p_{\text{fit}} - p_{\text{mean}}}}$$

The top is proportional to the amount of variance explained by the **Regression Line.** (see **Calculating R^2** for details)

This is the extra number of parameters used for the **Regression Line.**

The Denominator is the variation in size not explained by the **Regression Line.**

$$SS(\text{fit}) / \boxed{(n - p_{\text{fit}})}$$

Dividing **SS(fit)** by n-p_{fit} compensates for the extra variables in the **Regression Line.**

Plugging in: $n = 4$

SS(mean) = 73 $\quad p_{mean} = 1$

SS(fit) = 14 $\quad p_{fit} = 2$

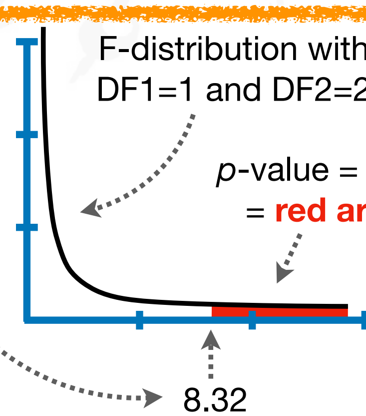
Degrees of Freedom:

DF1 = $p_{mean} - p_{fit} = 1$

DF2 = $n - p_{fit} = 2$

$$F = \frac{\dfrac{73 - 14}{2 - 1}}{14 / (4 - 2)} = 8.32$$

The **F-statistic**, **8.32**, tells us that the ***p*-value**, **0.1**, is the area under the curve from **8.32** to infinity.



Using more than one variable to make predictions:

When we use more than one variable to make predictions, we compare a **Simple Model** to a **Fancy Model**.

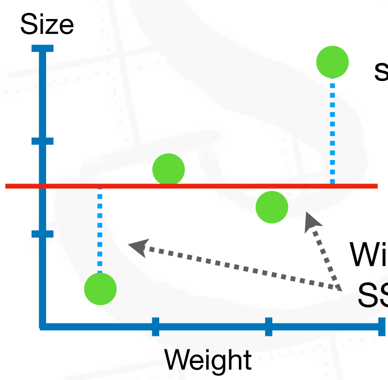
$$R^2 = \frac{SS(\text{mean}) - SS(\text{fancy})}{SS(\text{mean})}$$

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

Example 1

In this example, the **Simple Model** has **1** parameter, the y-intercept, so $\boldsymbol{p}_{\textbf{simple}} = \mathbf{1}$...

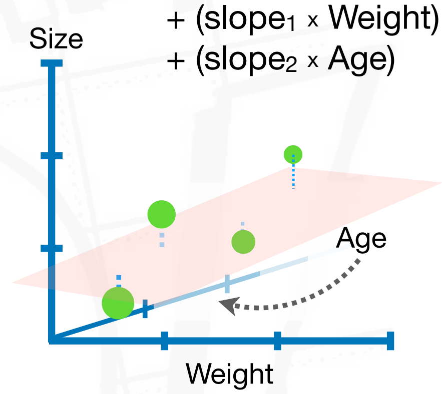
...while this **Fancy Model** has **3** parameters, the y-intercept and **2** slopes, so $\boldsymbol{p}_{\textbf{fancy}} = \mathbf{3}$.



size = y-intercept

With just the y-intercept, SS(mean) = SS(simple).

size = y-intercept
+ (slope$_1$ × Weight)
+ (slope$_2$ × Age)



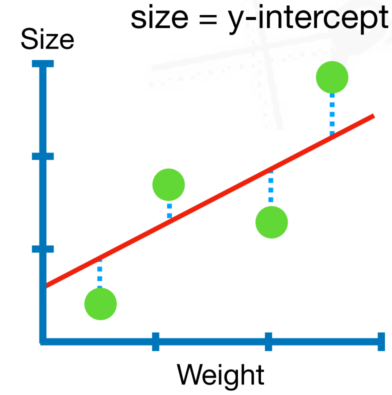
NOTE: R^2 is typically only calculated between a fancy model and the simplest model that only contains the y-intercept.

Example 2

In this example, the **Simple Model** has **2** parameters, the y-intercept and a slope, so $\boldsymbol{p}_{\textbf{simple}} = \mathbf{2}$...

...while this **Fancy Model** has **3** parameters, the y-intercept and **2** slopes, so $\boldsymbol{p}_{\textbf{fancy}} = \mathbf{3}$.

size = y-intercept + (slope × Weight)



NOTE: The **Simple Model** should always be a subset of the **Fancy Model**.

BAM!!!

size = y-intercept
+ (slope$_1$ × Weight)
+ (slope$_2$ × Age)