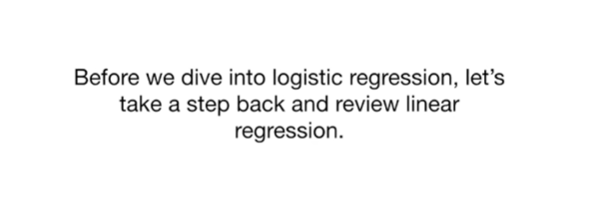
<https://www.youtube.com/watch?v=yIYKR4sgzI8&list=PLblh5JKOoLUKxzEP5HA2d-Li7IJkHfXSe&index=1>



Today we're going to talk about logistic regression.

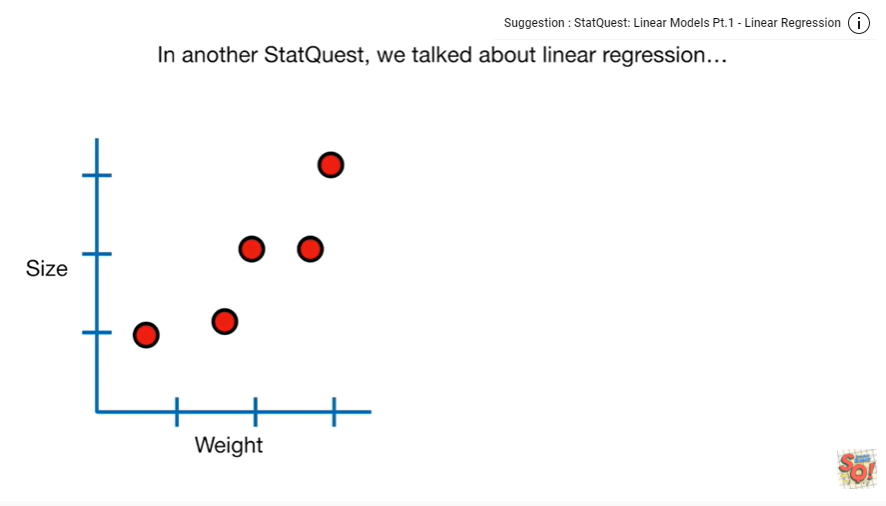
This is a technique that can be used for traditional statistics as well as machine learning.

So let's get right to it.

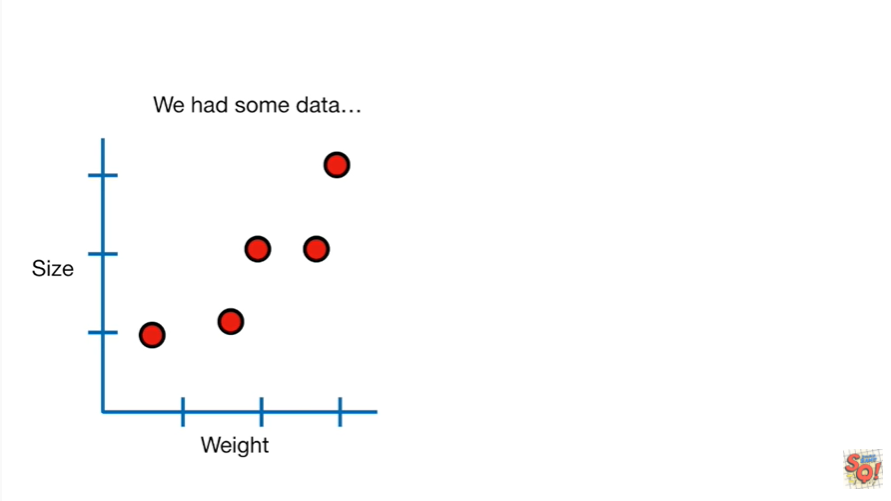


Before we dive into logistic regression.

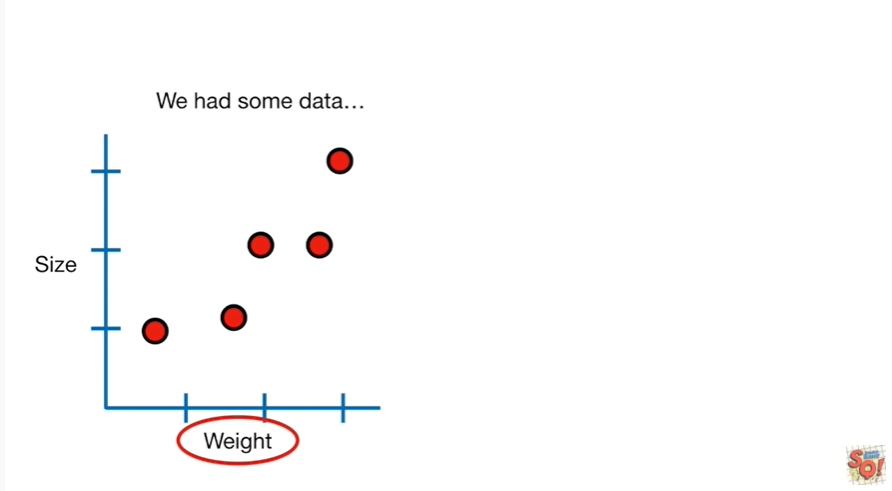
Let's take a step back and review linear regression.



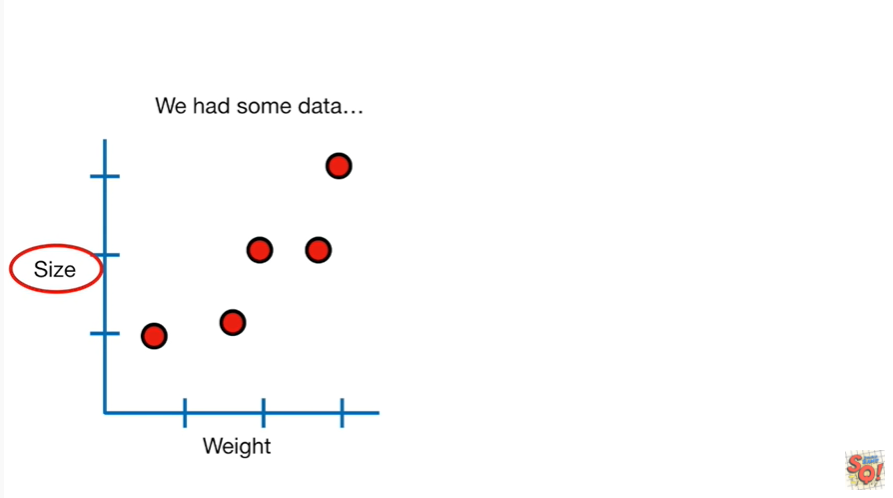
In another stat quest we talked about linear regression.



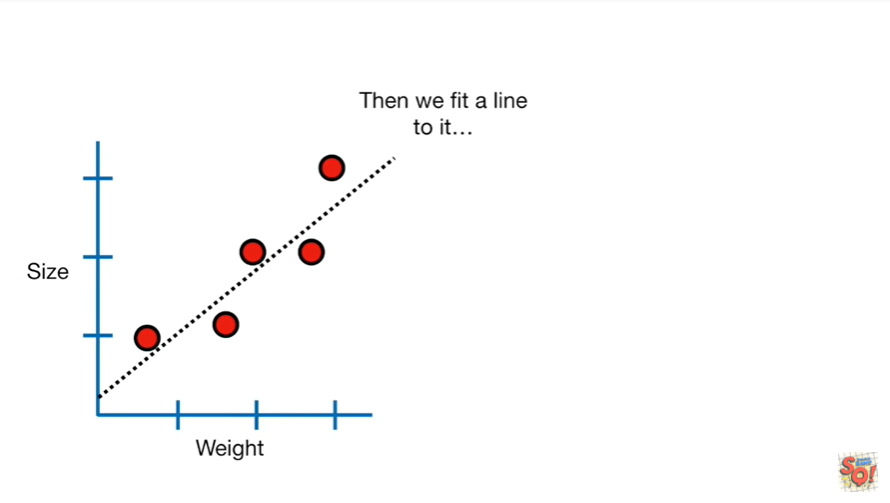
We had some data



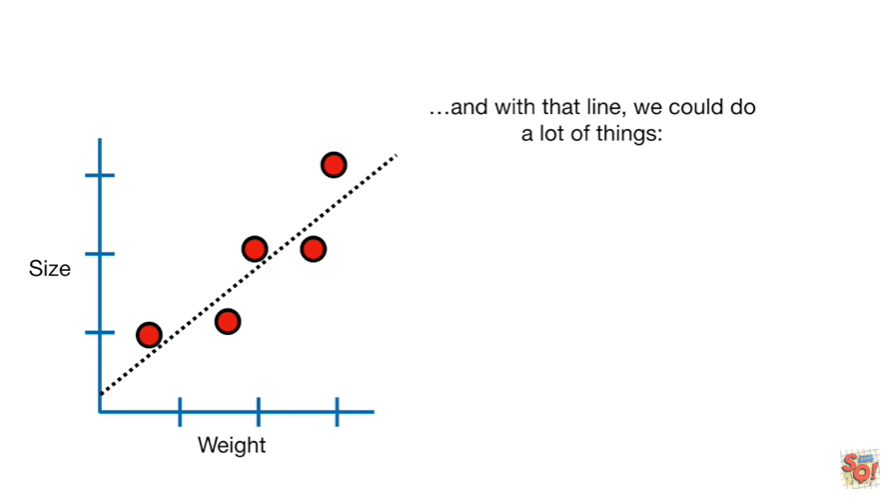
weight



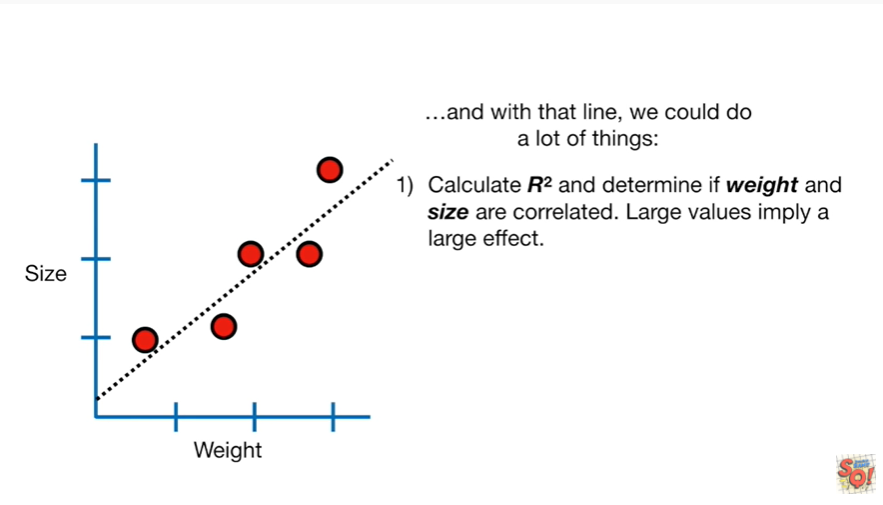
and size



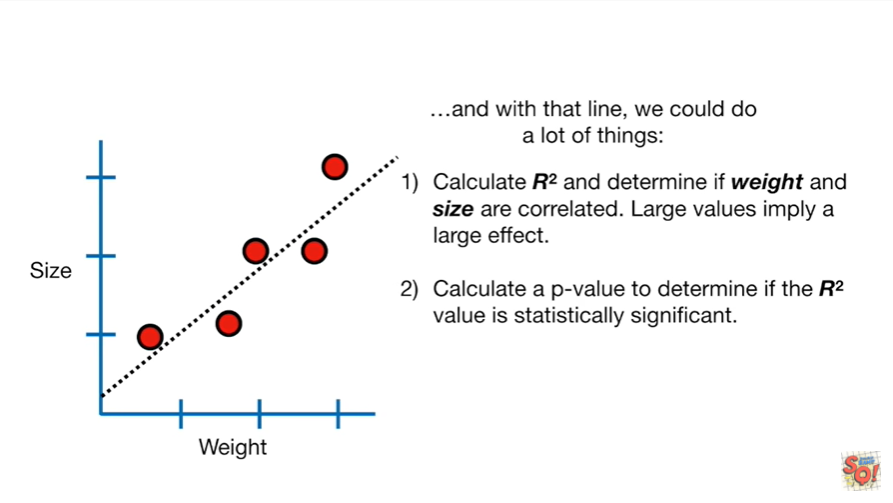
then we fit a line to it



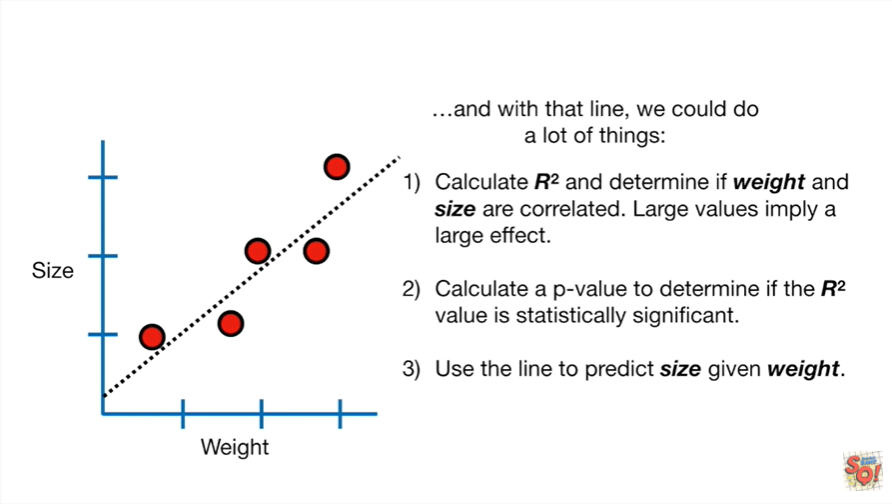
and with that line we could do a lot of things.



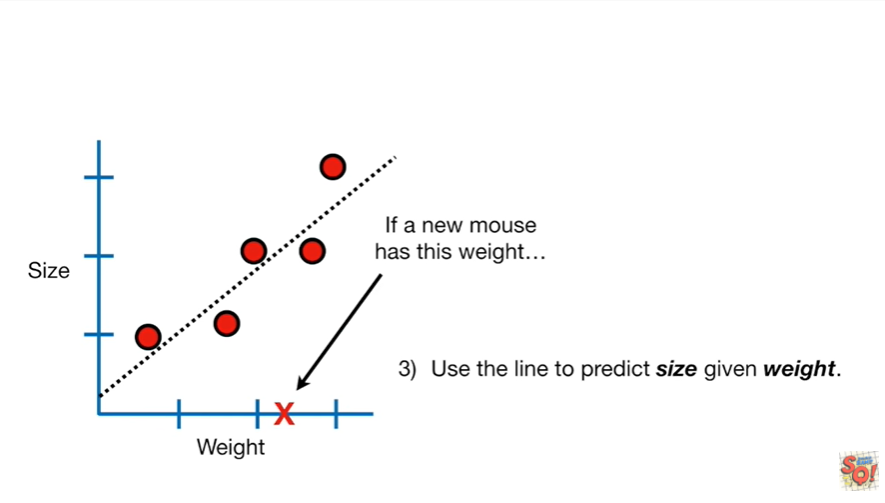
First we could calculate r-squared and determine if weight and size are correlated large values imply a large effect.



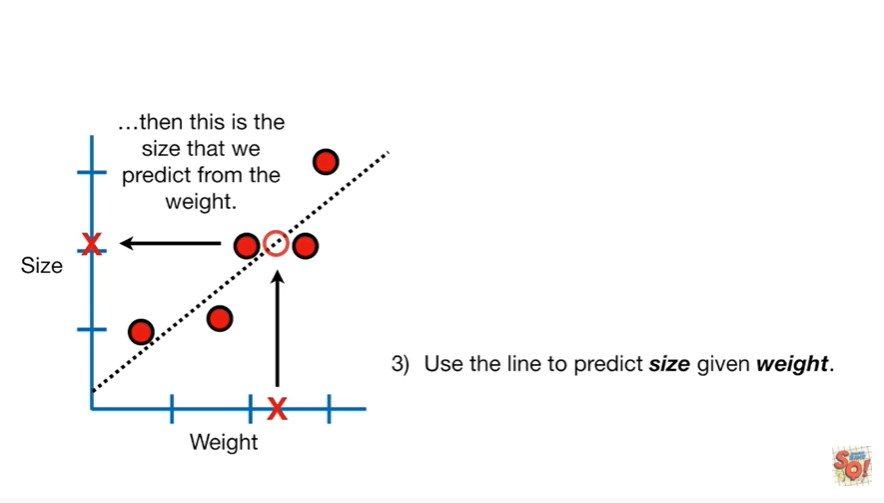
And second calculate a p-value to determine if the r-squared value is statistically significant.



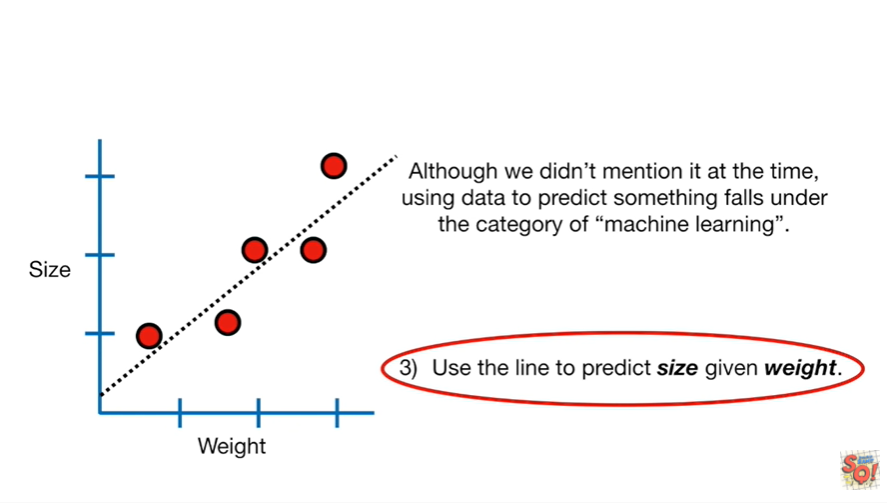
And third we could use the line to predict size given weight.



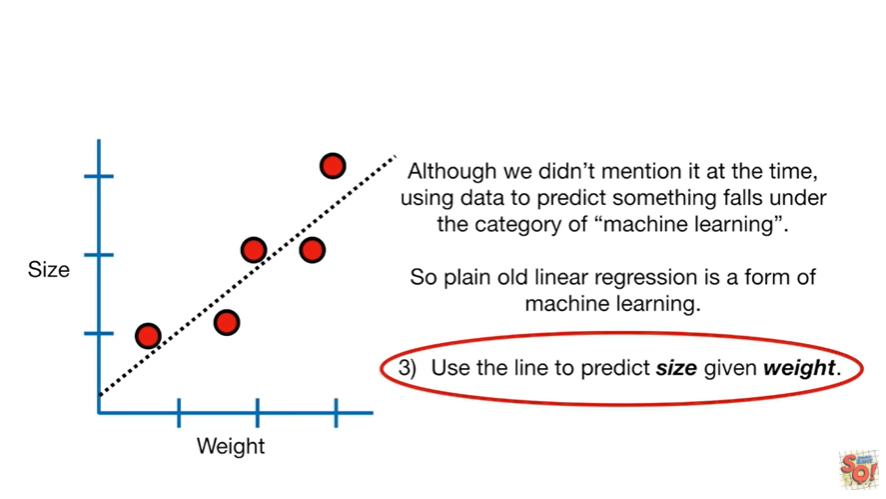
If a new mouse has this weight



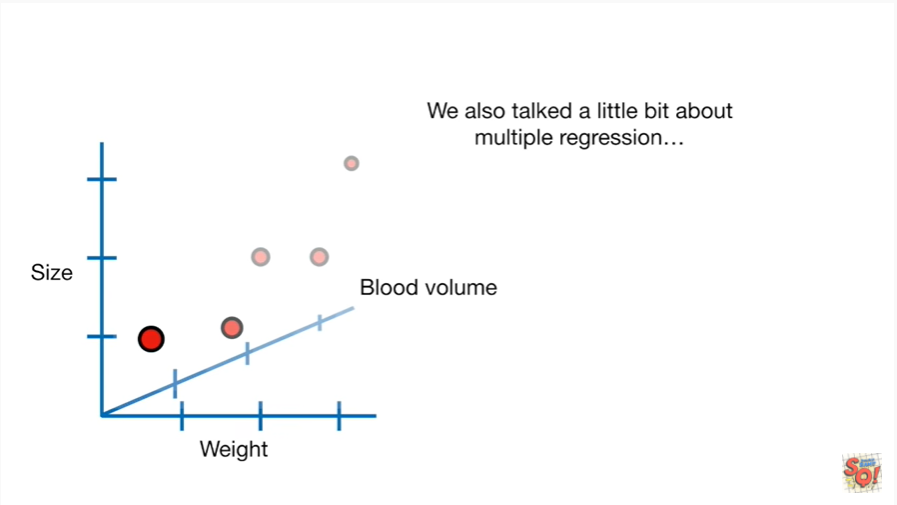
then this is the size that we predict from the weight.



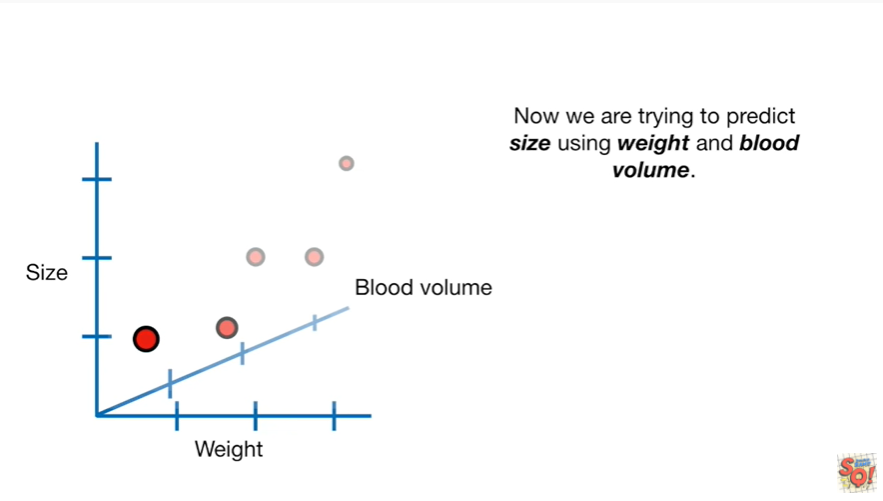
Although we didn't mention it at the time using data to predict something falls under the category of machine learning.



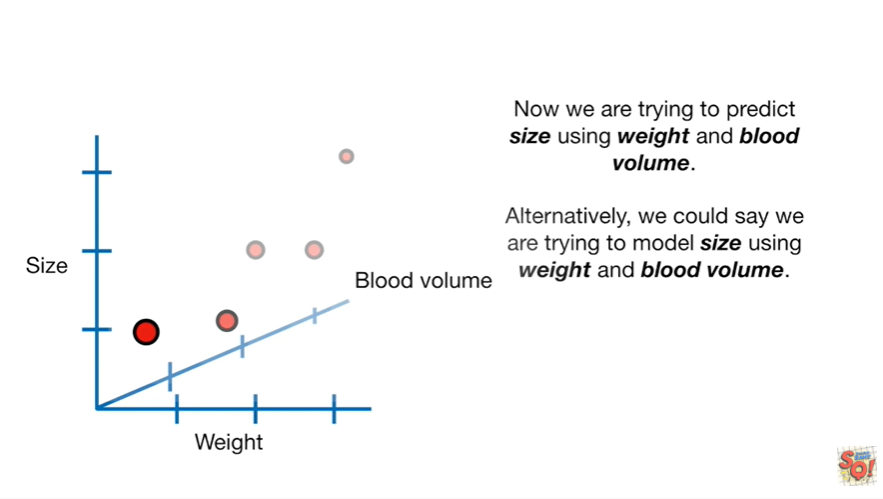
So plain old linear regression is a form of machine learning.



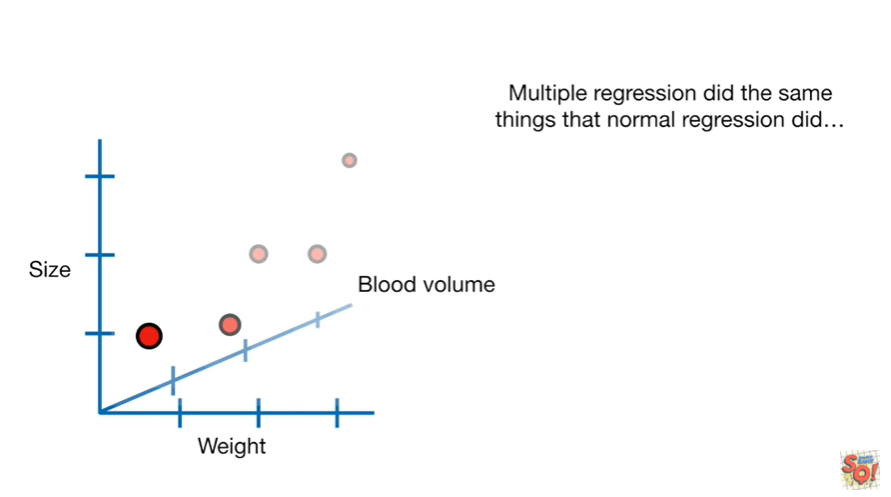
We also talked a little bit about multiple regression.



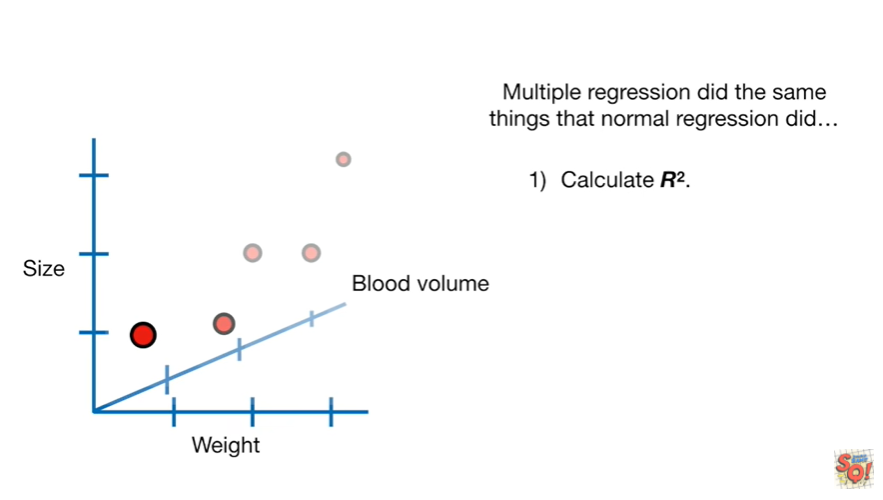
Now we are trying to predict size using weight and blood volume.



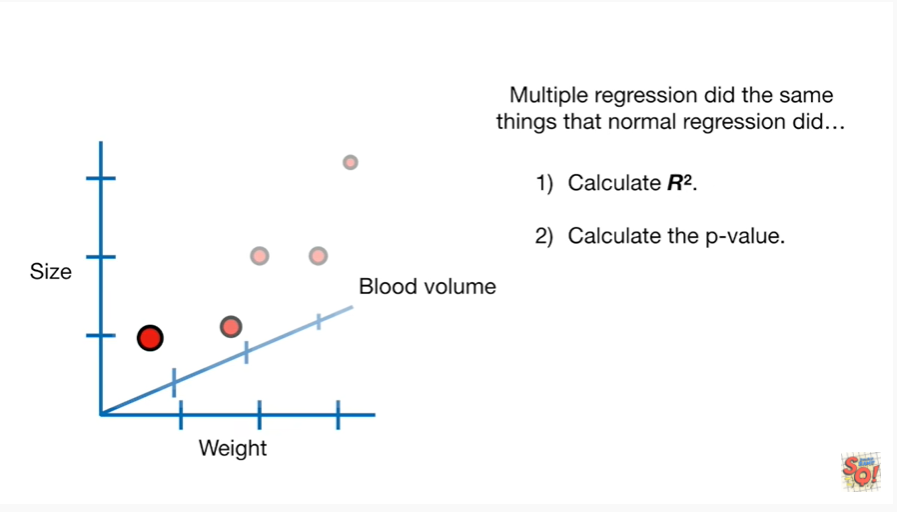
Alternatively we could say that we are trying to model size using weight and blood volume.



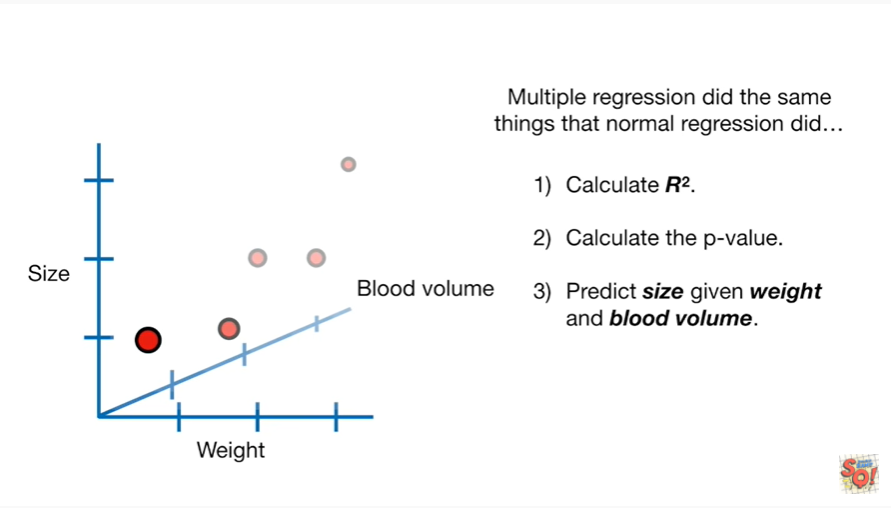
Multiple regression did the same things that normal regression did.



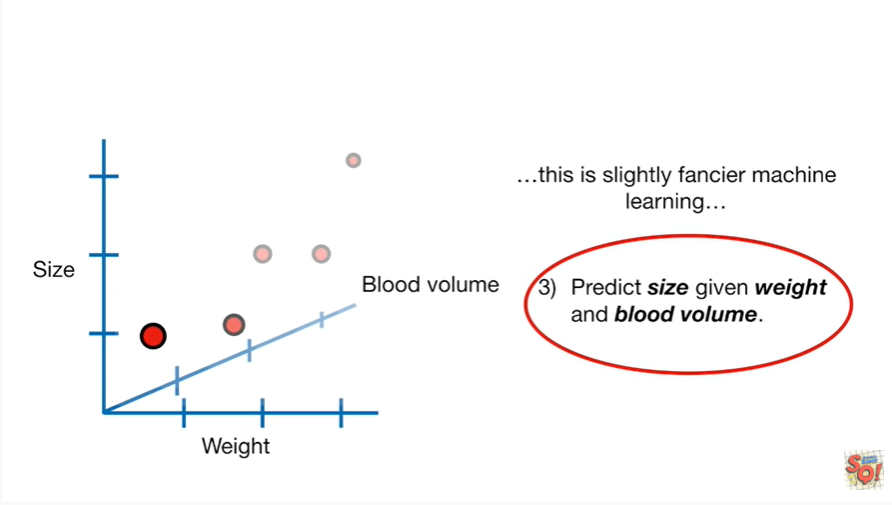
We calculated r-squared



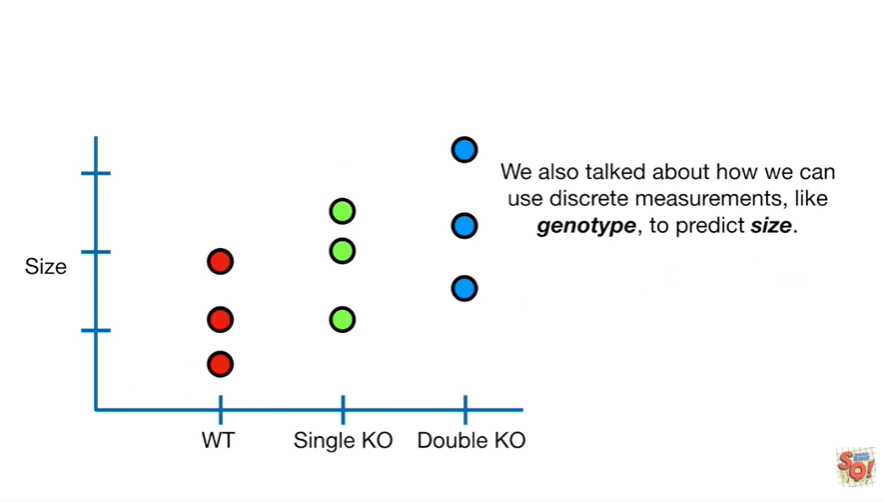
and we calculated the p-value



and we could predict size using weight and blood volume.



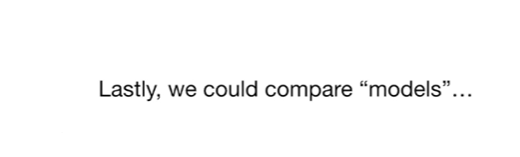
And this makes multiple regression a slightly fancier machine learning method.



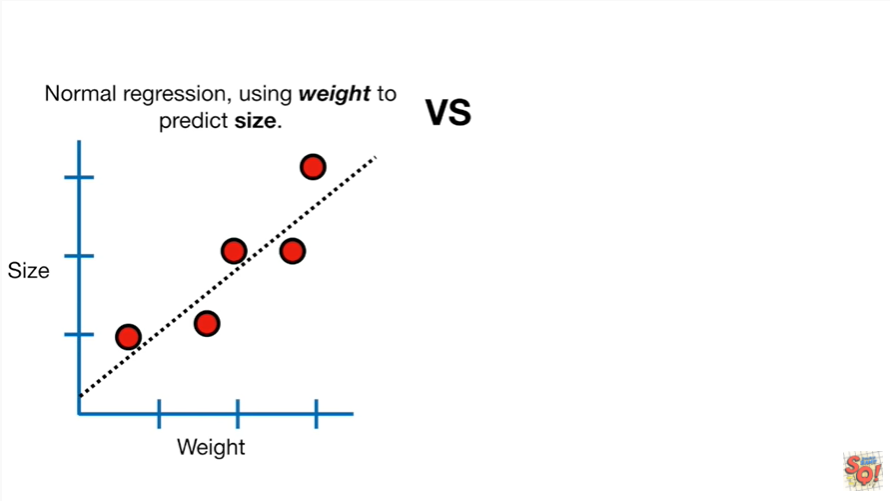
We also talked about how we can use discrete measurements like genotype to predict size.

If you're not familiar with the term genotype don't freak out, it's no big deal.

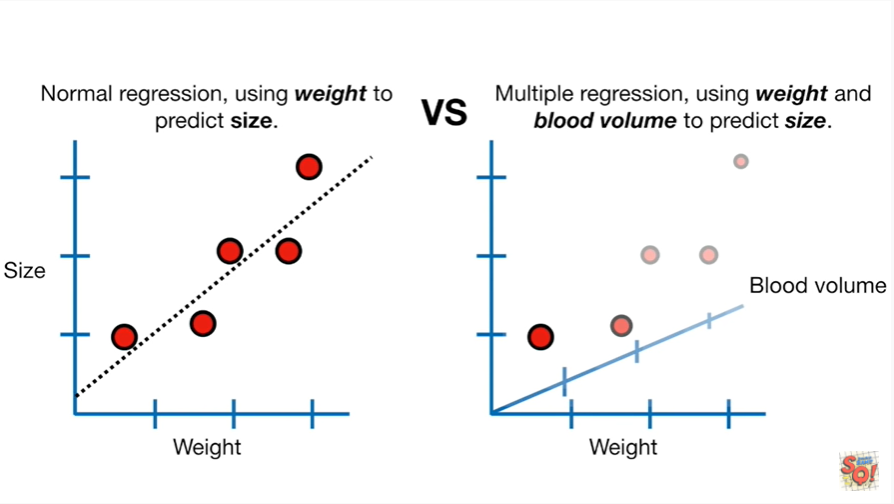
Just know that it refers to different types of mice.



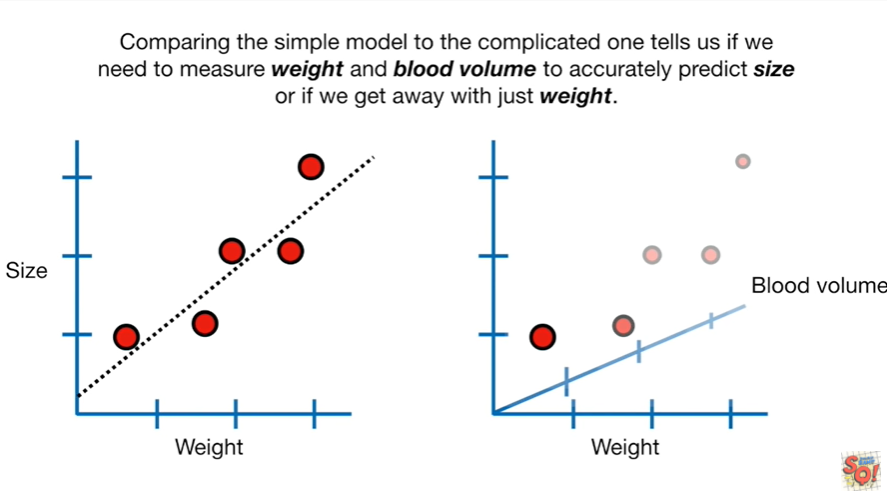
Lastly we could compare models.



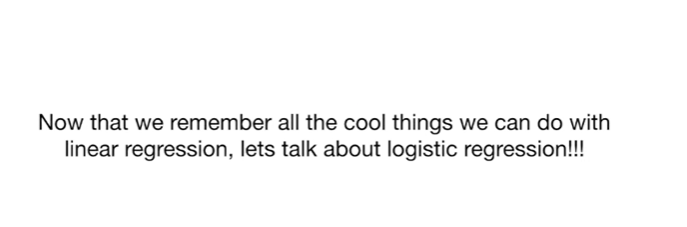
So on the left side we've got normal regression using weight to predict size.



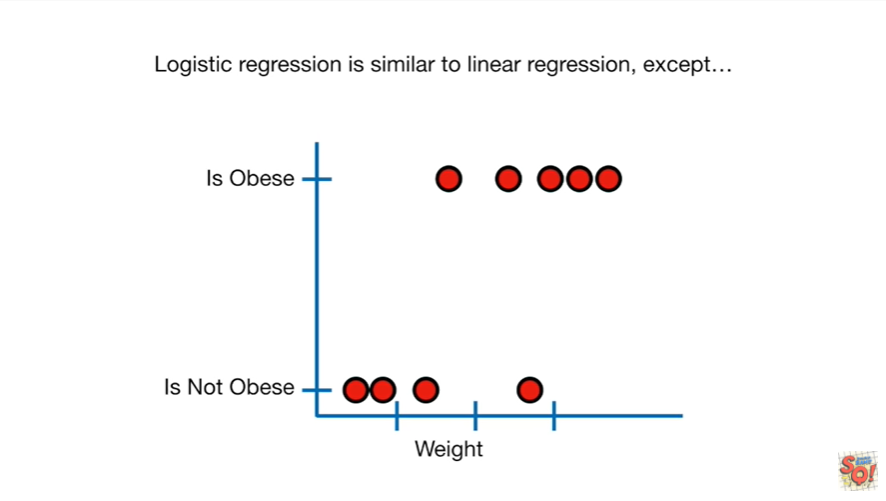
And we can compare those predictions to the ones we get from multiple regression, where we're using weight and blood volume to predict size.



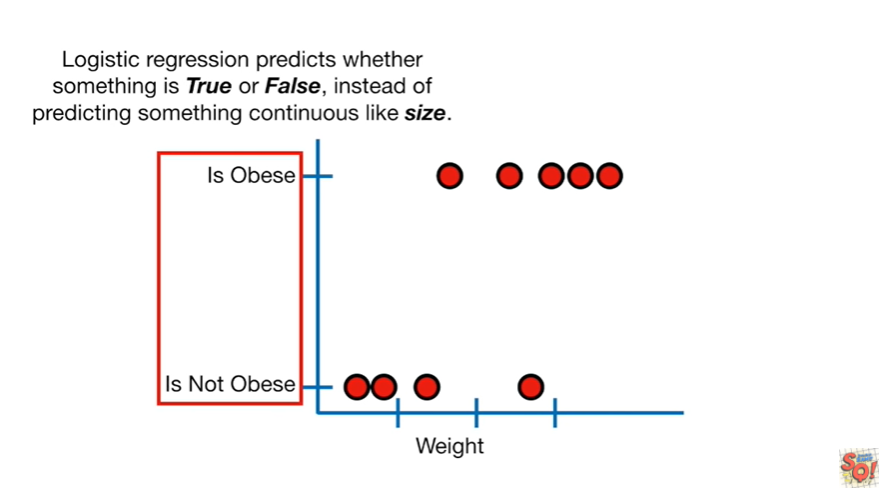
Comparing the simple model to the complicated one tells us if we need to measure weight and blood volume to accurately predict size or if we can get away with just weight.



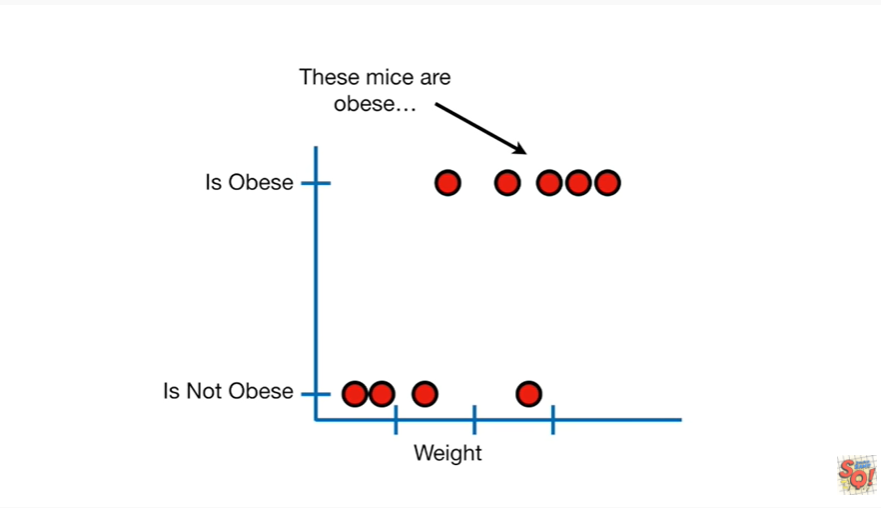
Now that we remember all the cool things we can do with linear regression, let's talk about logistic regression.



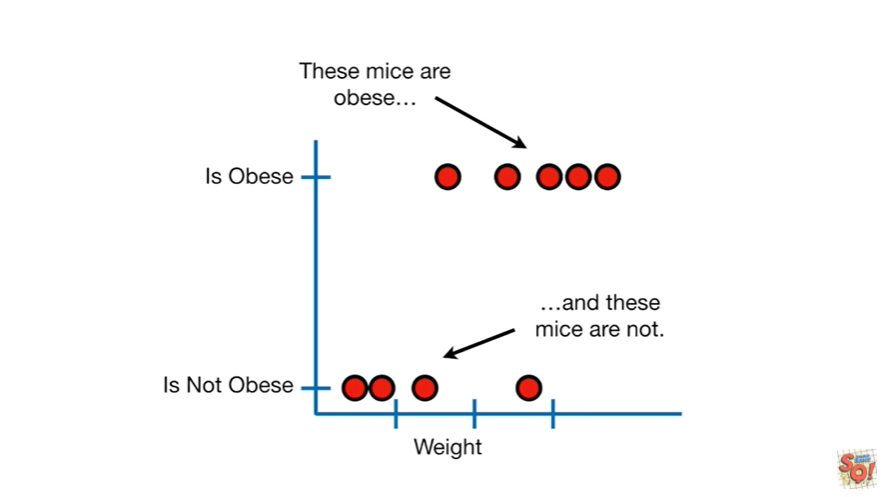
Logistic regression is similar to linear regression except



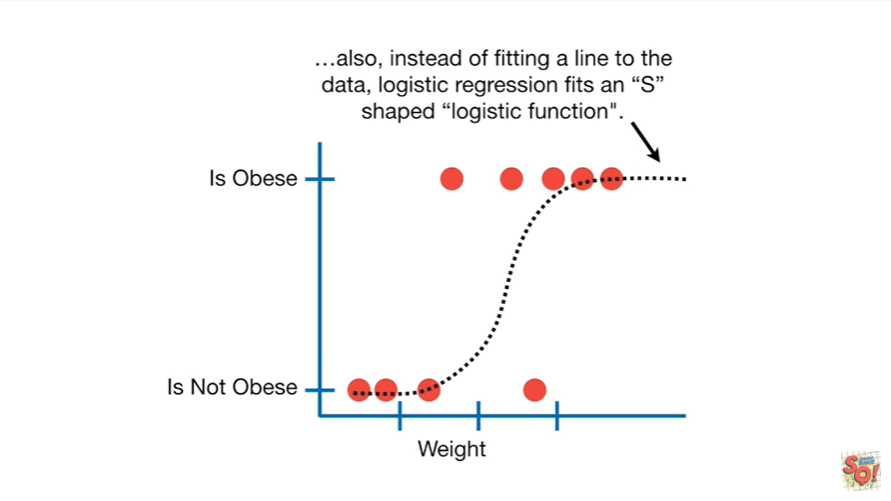
Logistic regression predicts whether something is true or false, instead of predicting something continuous like size.



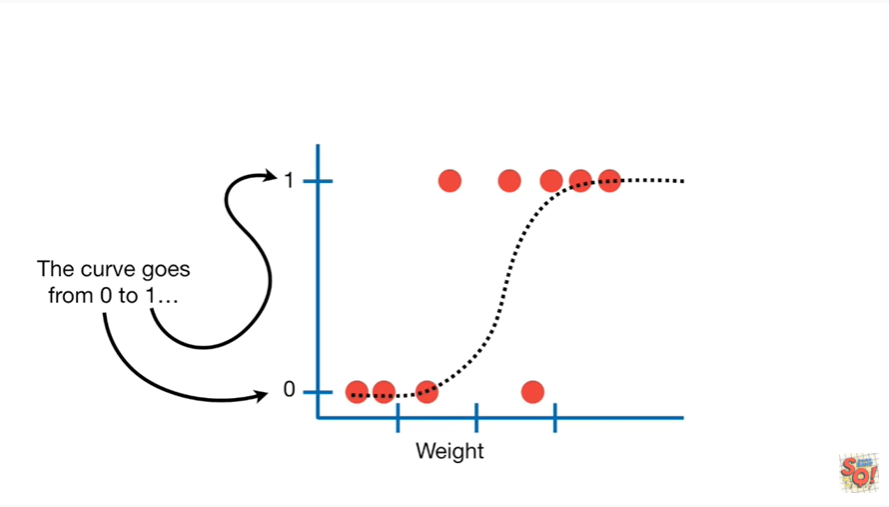
These mice are obese



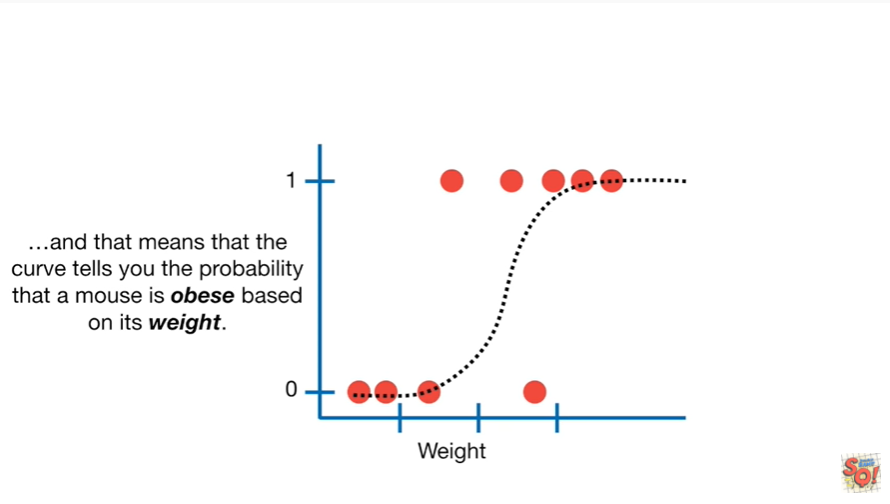
and these mice are not.



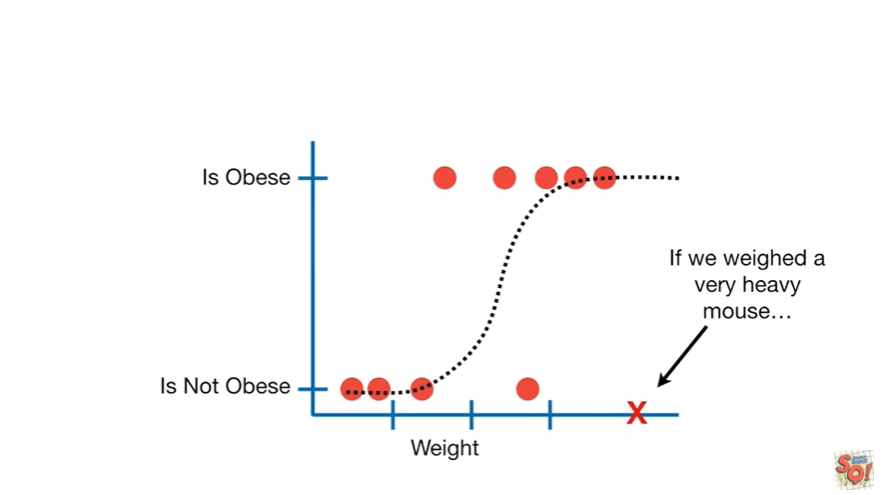
Also instead of fitting a line to the data logistic regression fits an s-shaped logistic function.



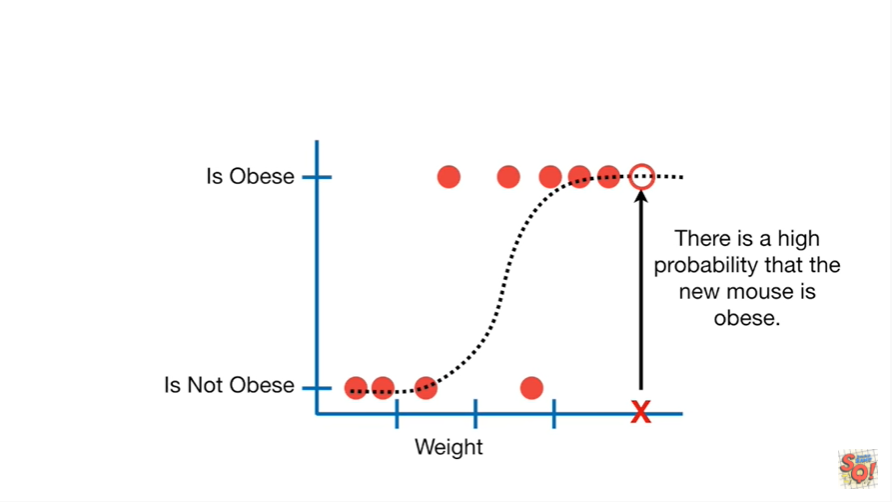
The curve goes from zero to one.



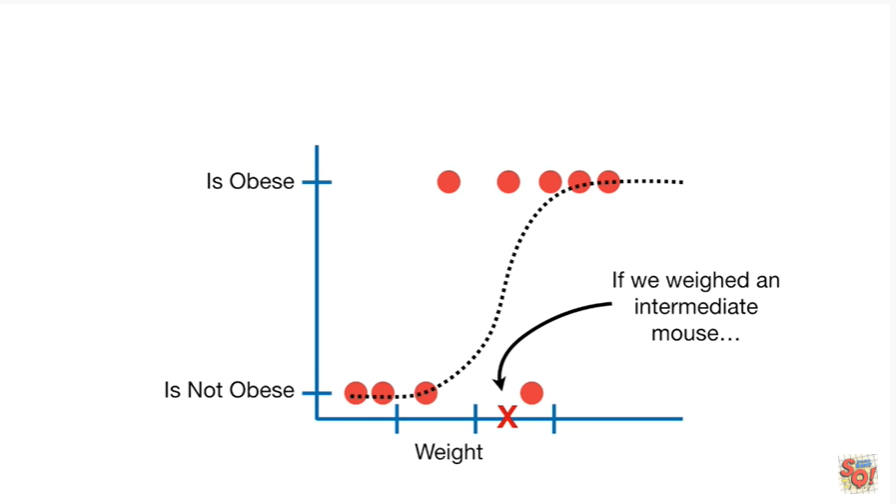
And that means that the curve tells you the probability that a mouse is obese based on its weight.



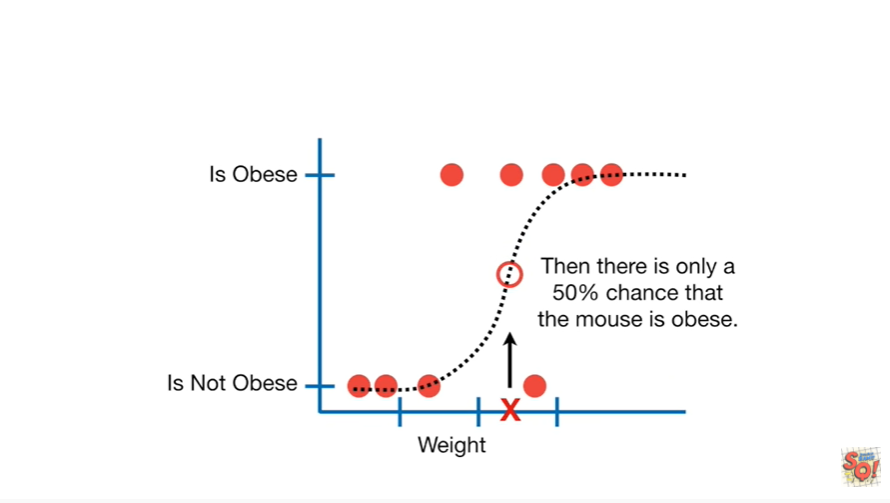
If we weighed a very heavy Mouse



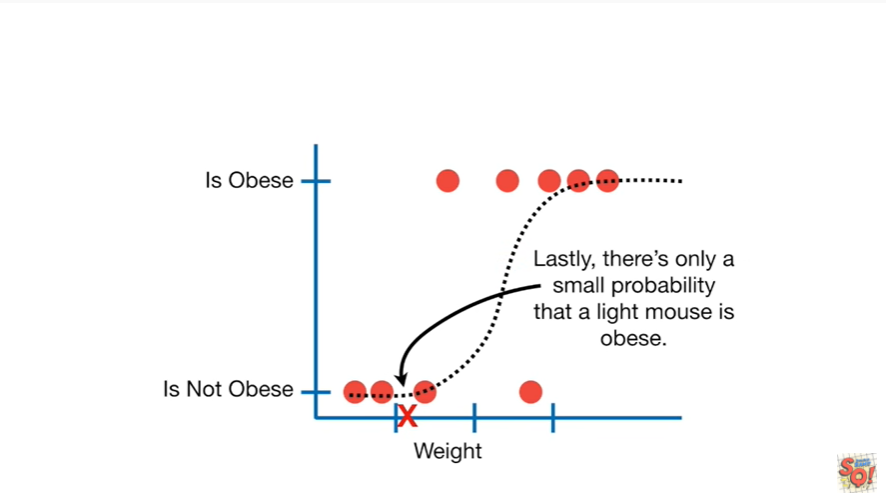
there is a high probability that the new Mouse is obese.



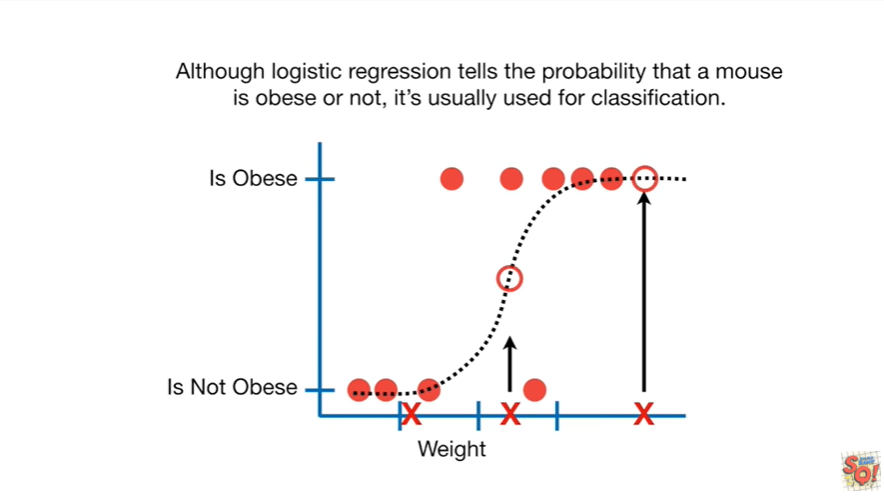
If we weighed an intermediate Mouse



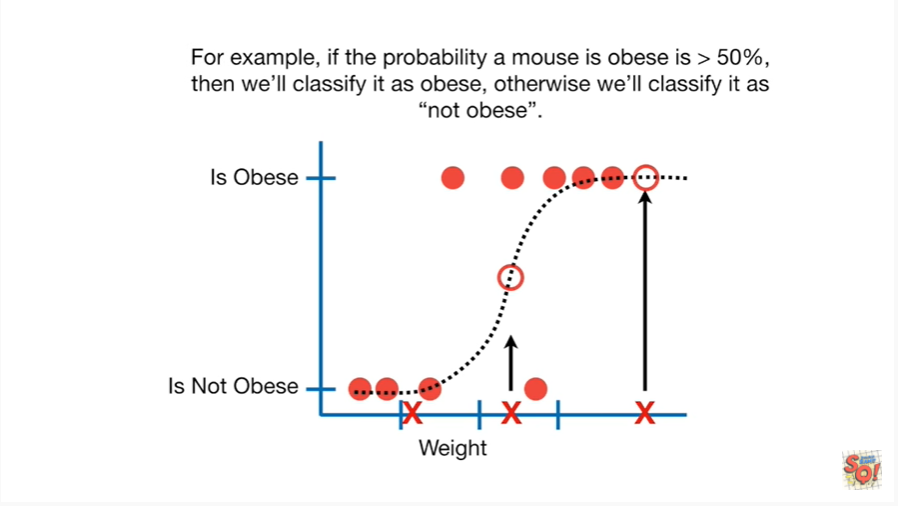
then there is only a 50% chance of the mouse is obese.



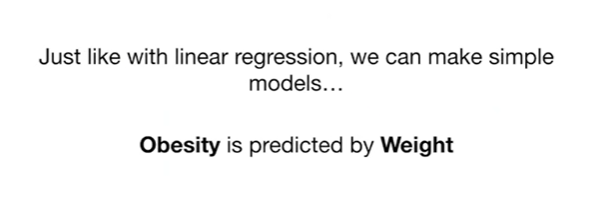
Lastly there's only a small probability that a light Mouse is obese.



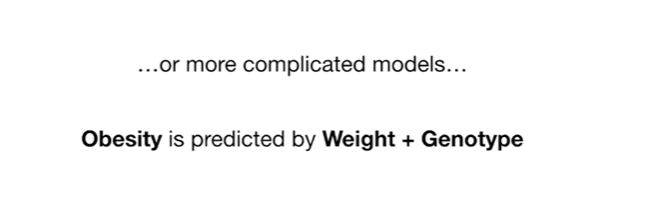
Although logistic regression tells the probability that amounts is obese or not, it's usually used for classification.



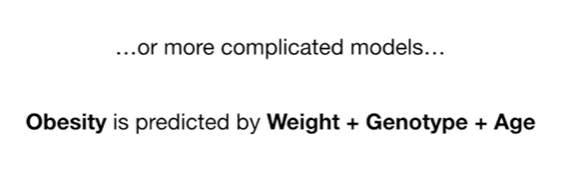
For example if the probability of Mouse is obese is greater than 50% then we'll classify it as obese, otherwise we'll classify it as not obese.



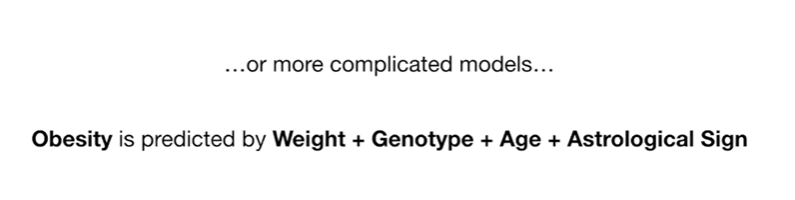
Just like with linear regression we can make simple models in this case we can have obesity predicted by weight



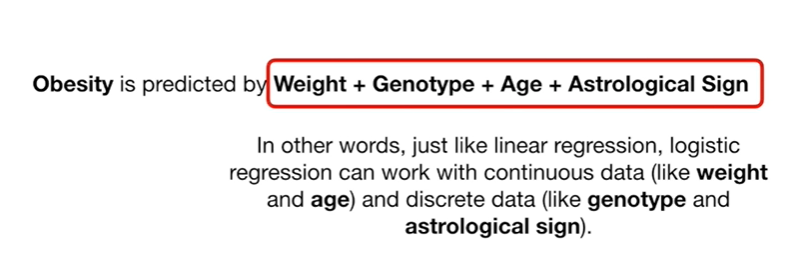
or more complicated models in this case obesity is predicted by weight and genotype.



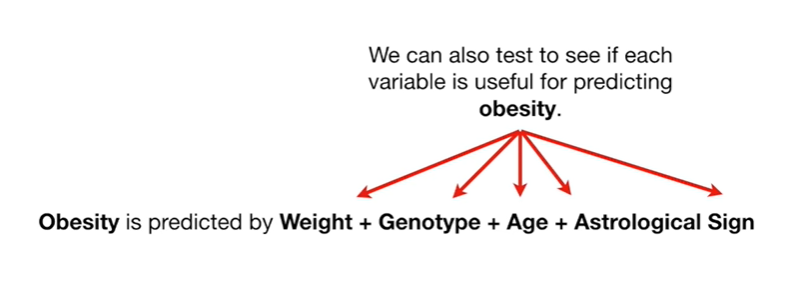
In this case obesity is predicted by weight and genotype and age.



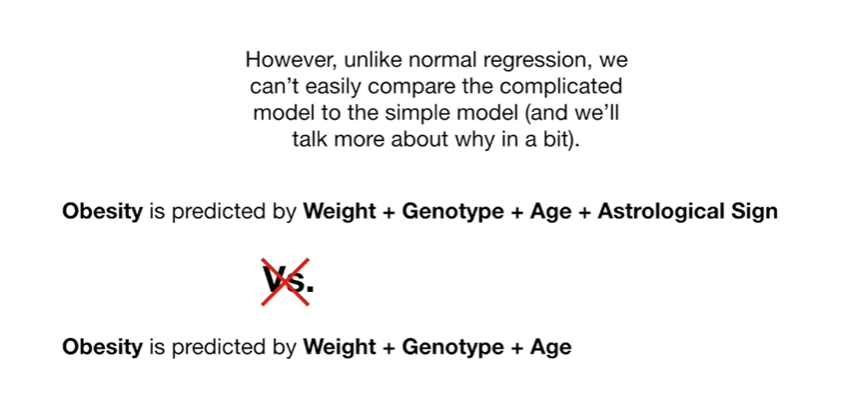
And lastly obesity is predicted by weight genotype age and astrological sign.



In other words just like linear regression logistic regression can work with continuous data like weight and age, and discrete data like genotype and astrological sign.

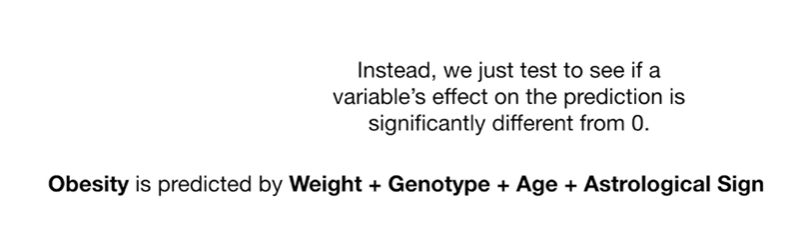


We can also test to see if each variable is useful for predicting obesity.

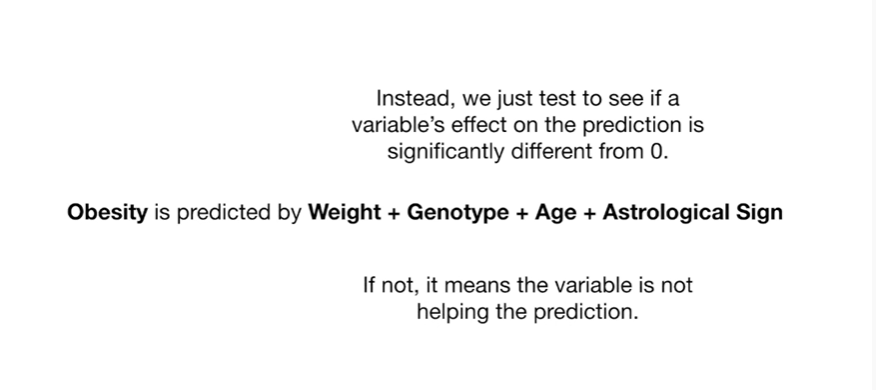


However unlike normal regression we can't easily compare the complicated model to the simple model.

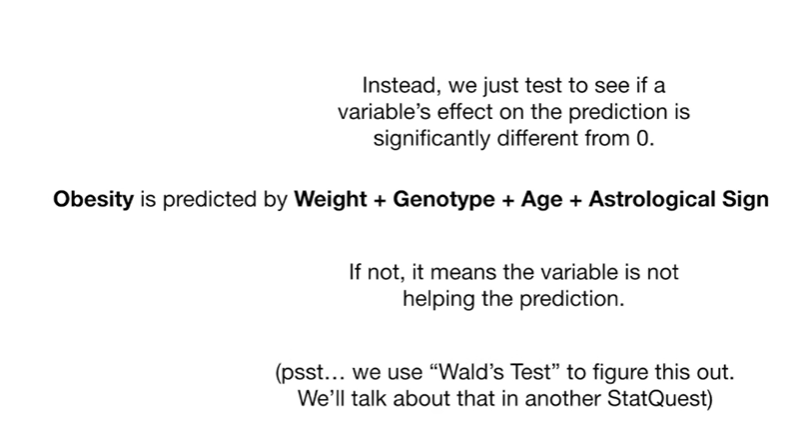
And we'll talk more about why in a bit.



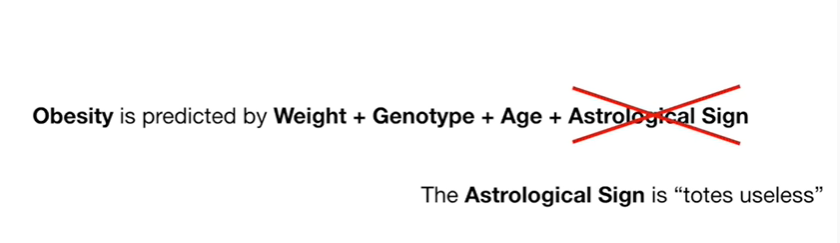
Instead we just test to see if a variables affect on the prediction is significantly different from zero.



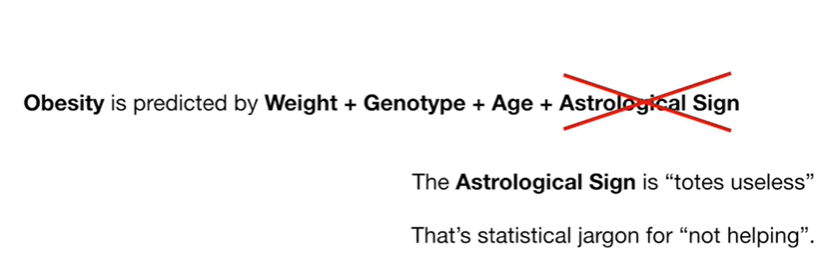
If not it means that the variable is not helping the prediction.



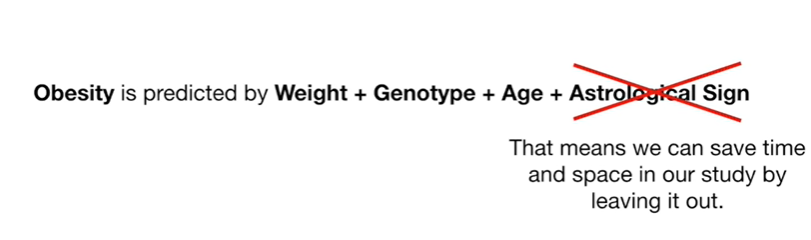
We use Wald's tests to figure this out, we'll talk about that in another stat quest.



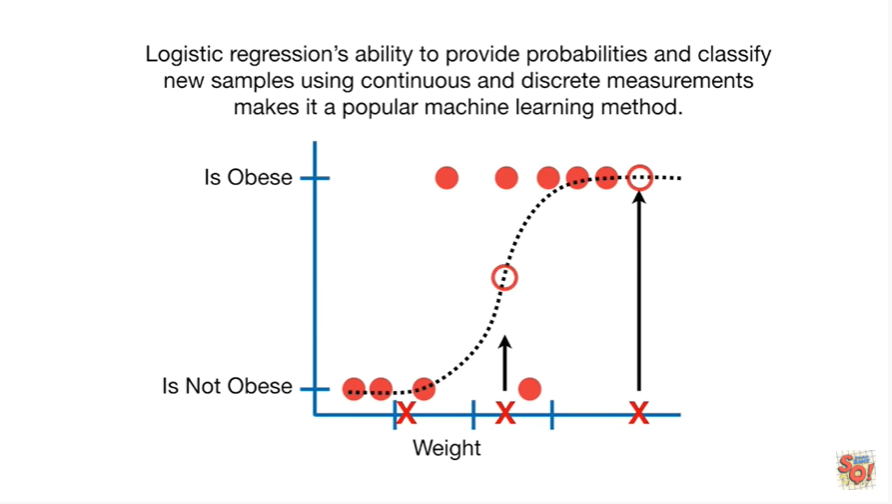
In this case the astrological sign is totes useless.



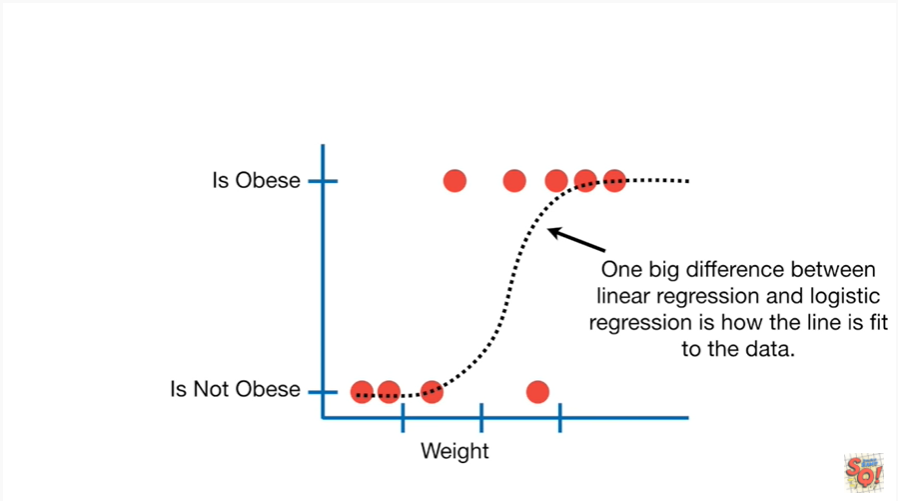
Best statistical jargon for not helping.



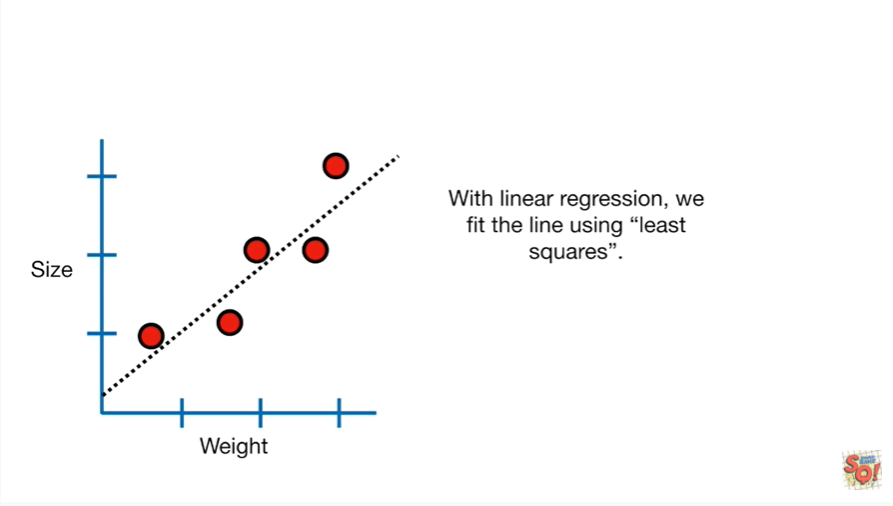
That means we can save time and space in our study by leaving it out.



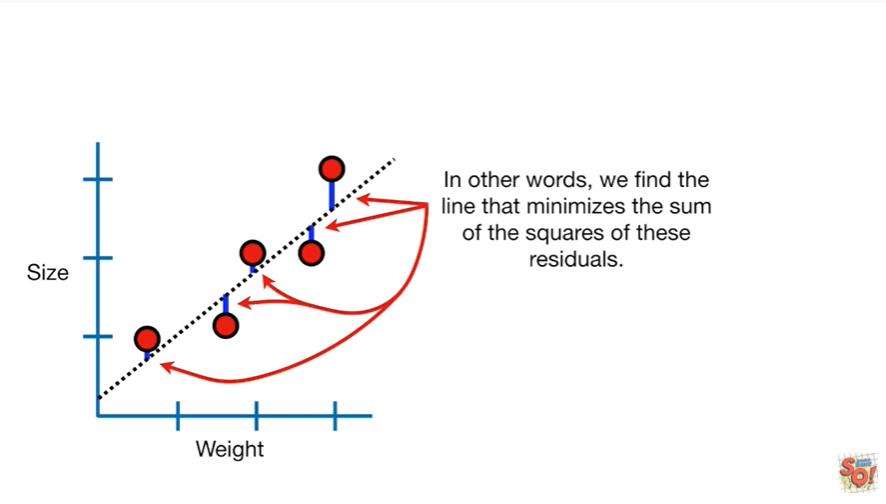
Logistic regressions ability to provide probabilities and classify new samples using continuous and discrete measurements makes it a popular machine learning method.



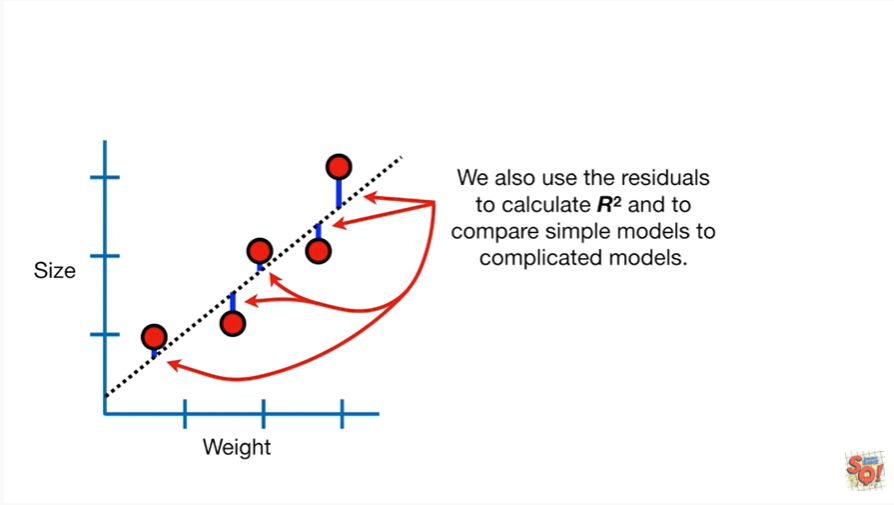
One big difference between linear regression and logistic regression is how the line is fit to the data.



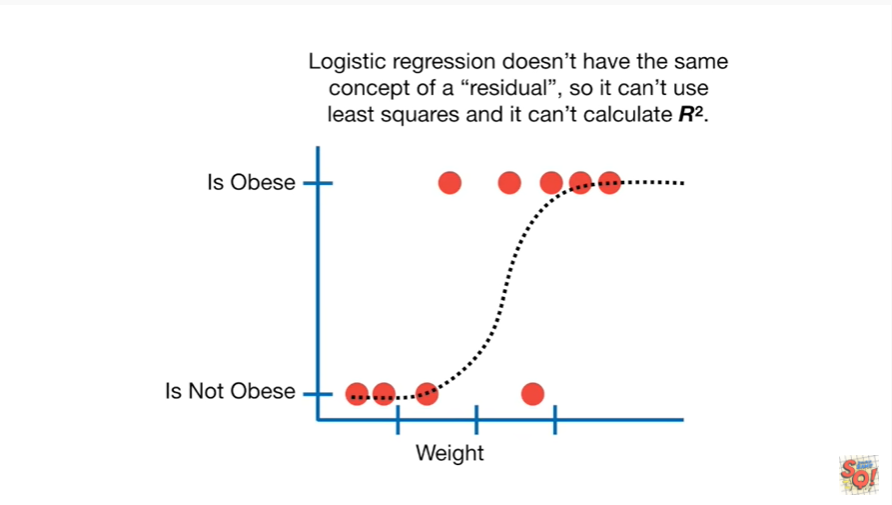
With linear regression we fit the line using least squares.



In other words we find the line that minimizes the sum of the squares of these residuals.



We also use the residuals to calculate r-squared and to compare simple models to complicated models.

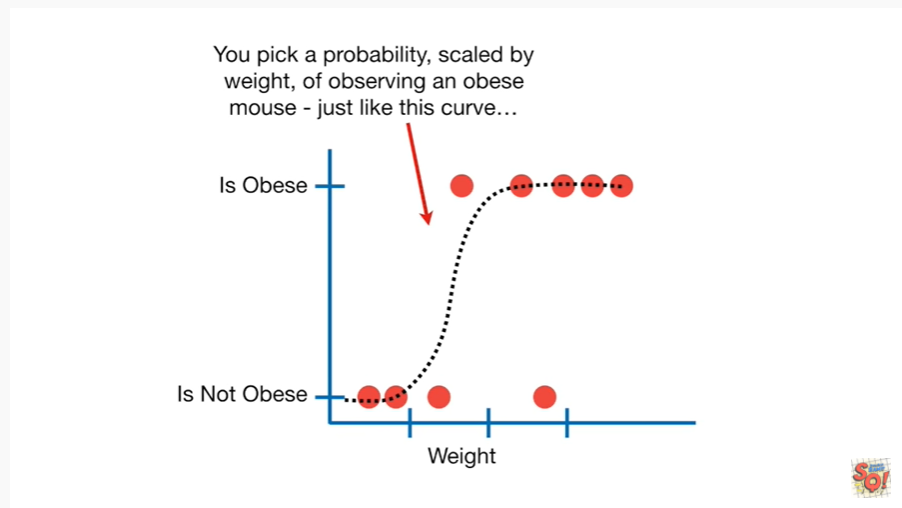


Logistic regression doesn't have the same concept of a residual so it can't use least squares and it can't calculate R squared.

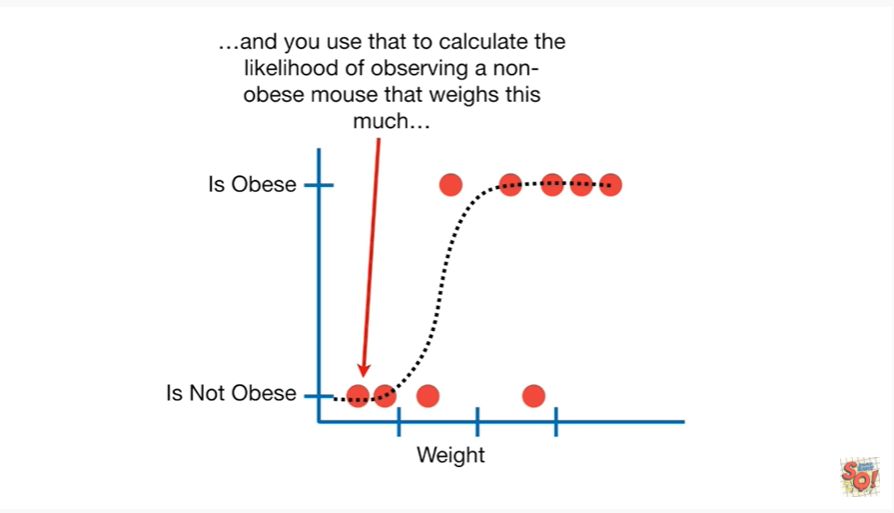


Instead it uses something called maximum likelihood.

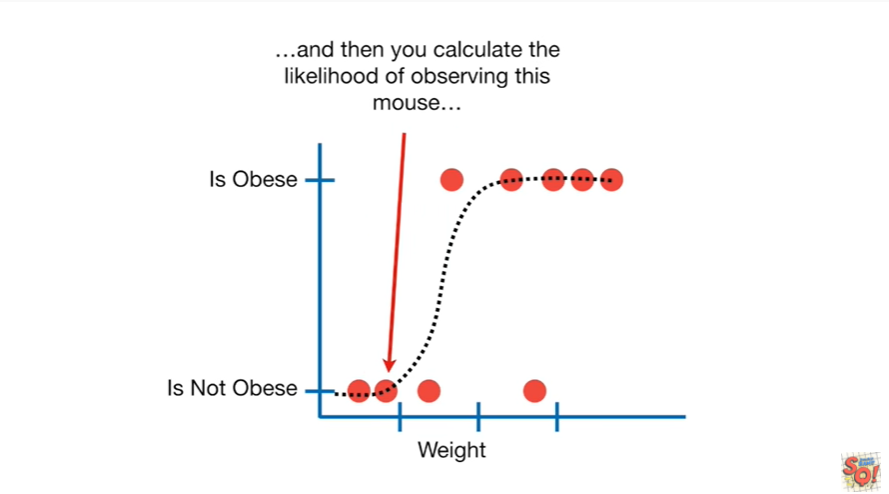
There's a whole stat quest on maximum likelihood so see that for details but in a nutshell.



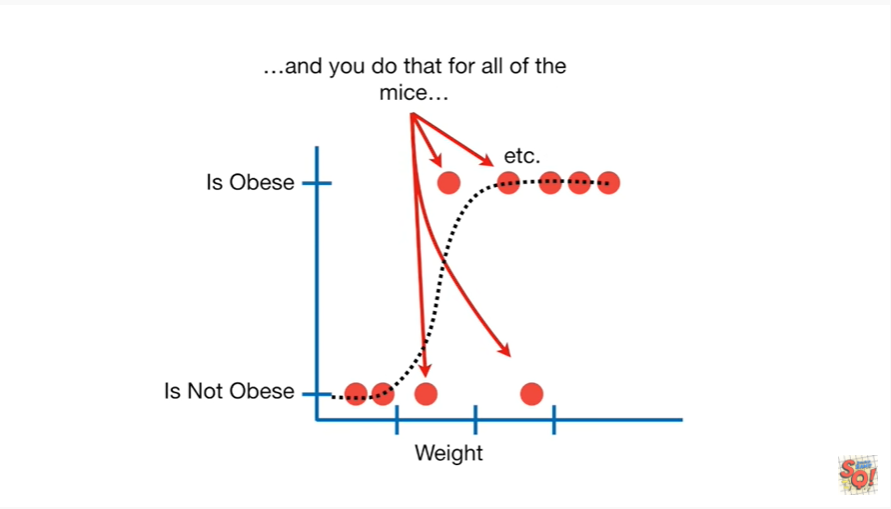
You pick a probability scaled by weight of observing an obese Mouse just like this curve.



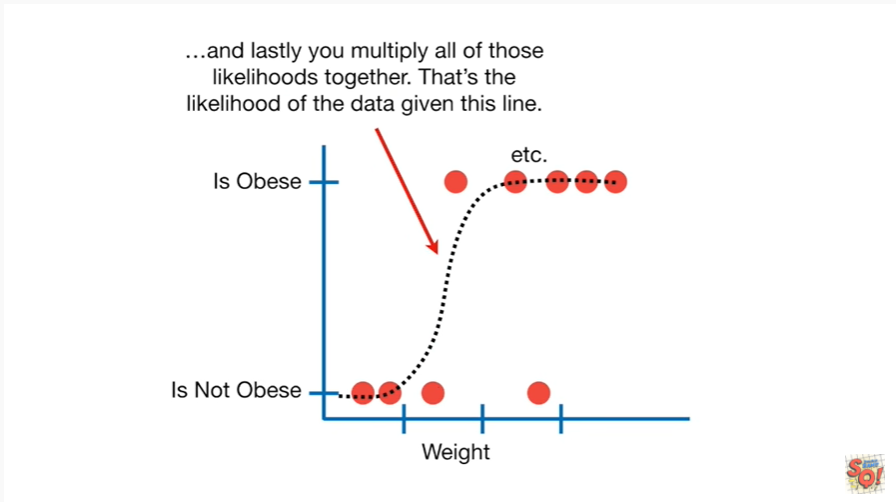
And you use that to calculate the likelihood of observing a non obese Mouse that wastes this much.



And then you calculate the likelihood of observing this Mouse.

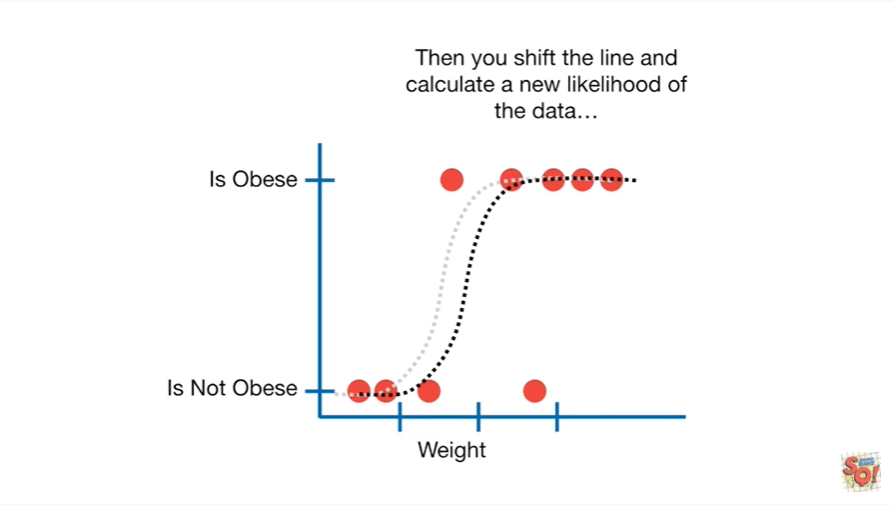


And you do that for all of the mice.

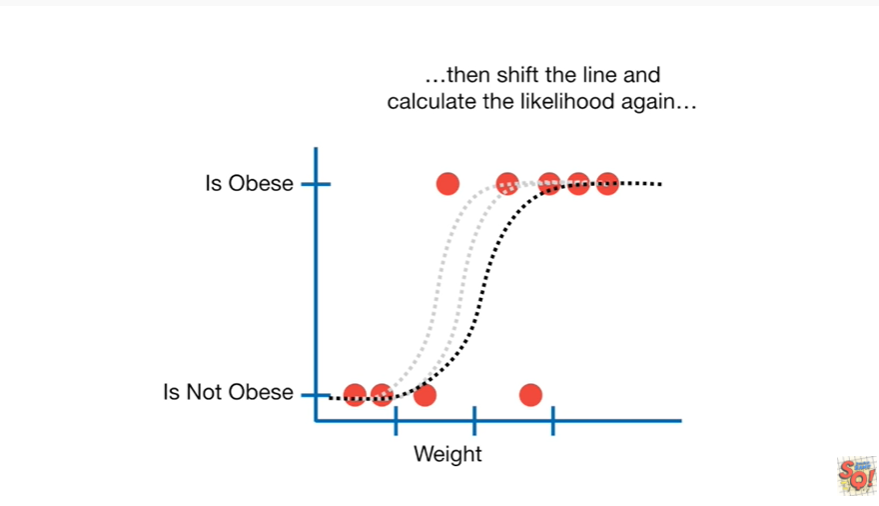


And lastly you multiply all of those likelihoods together.

That's the likelihood of the data given this line.



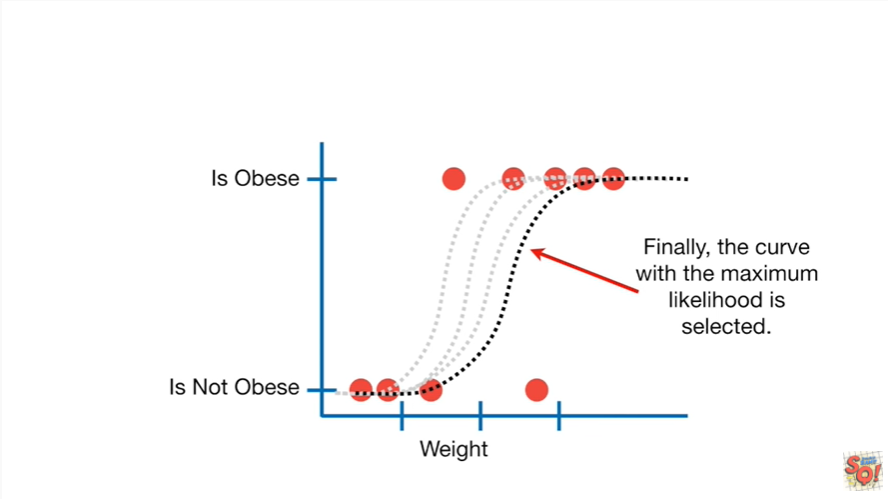
Then you ship the line and calculate a new likelihood of the data.



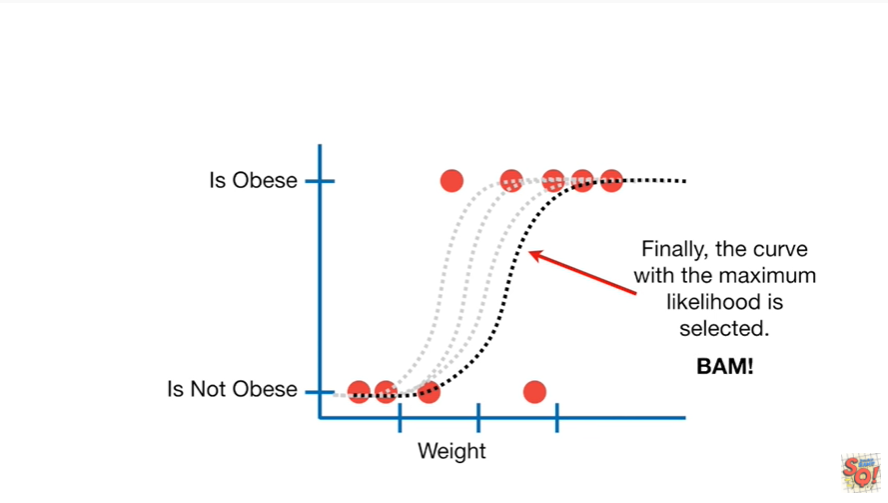
And then ship the line and calculate the likelihood again.



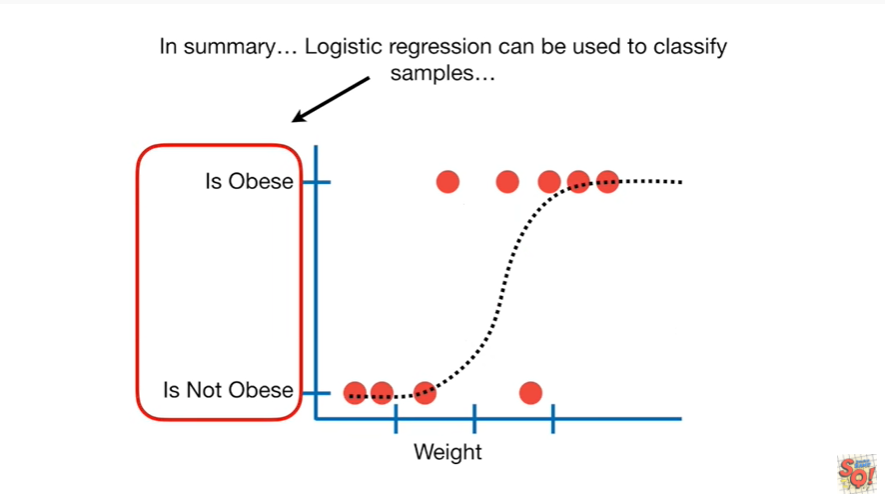
And again.



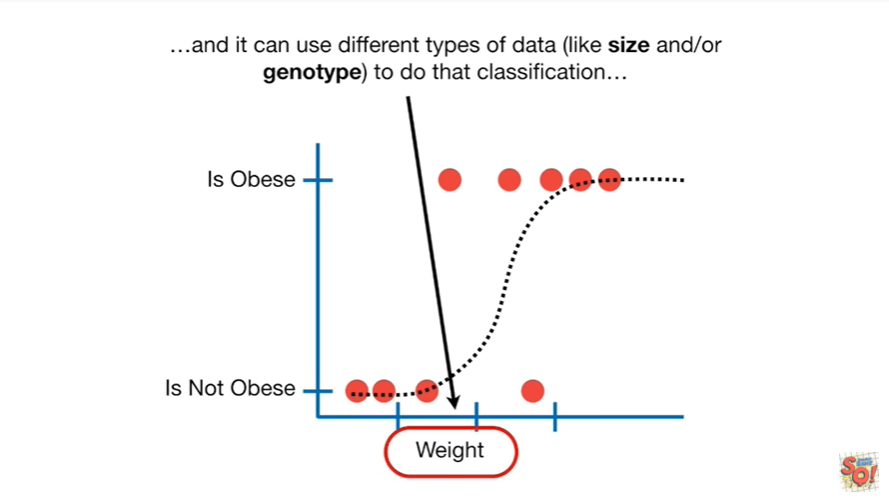
Finally the curve with the maximum value for the likelihood is selected.



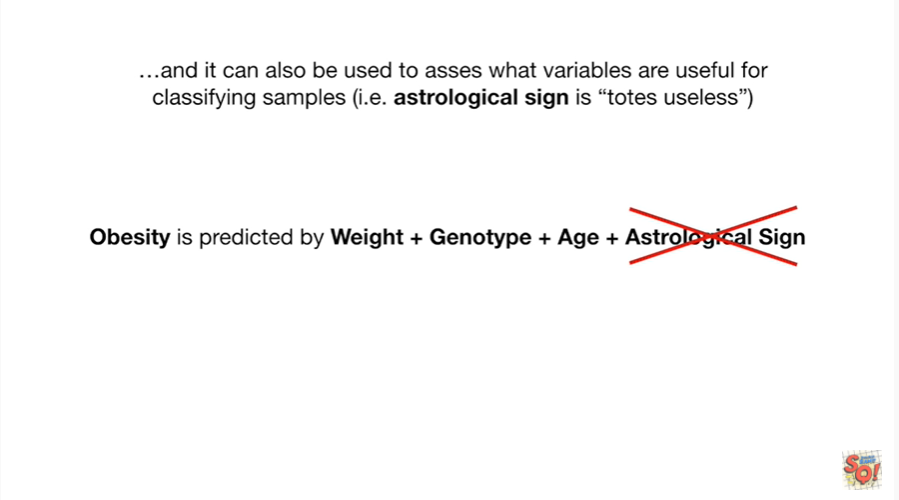
BAM.



In summary logistic regression can be used to classify samples.



And it can use different types of data like size and/or genotype to do that classification.



And it can also be used to assess what variables are useful for classifying samples ie astrological sign is totes useless.