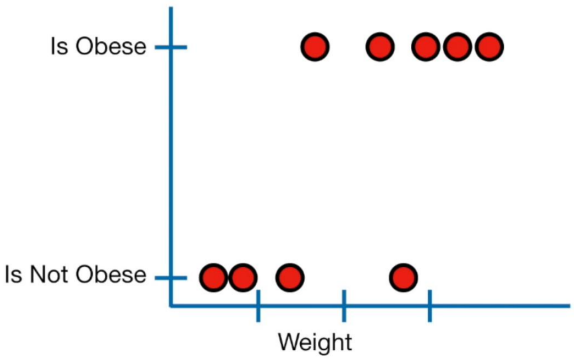
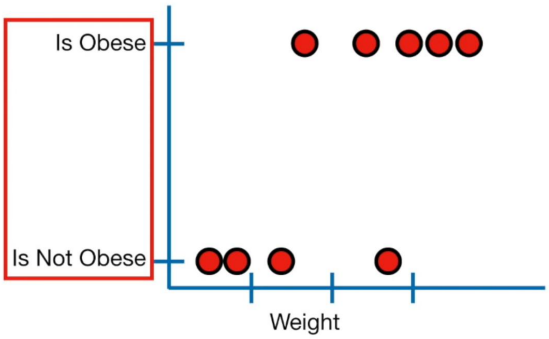
StatQuest:
Logistic Regression

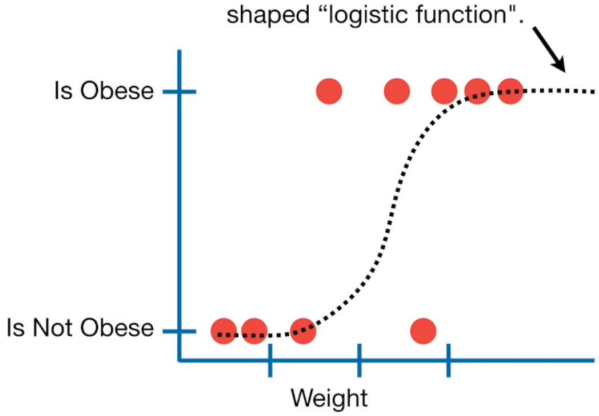
Logistic regression is similar to linear regression, except…

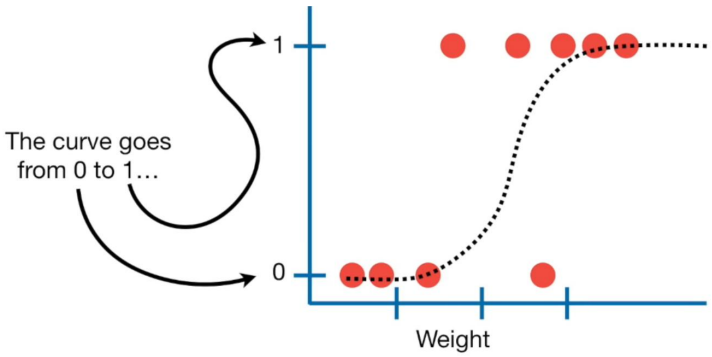


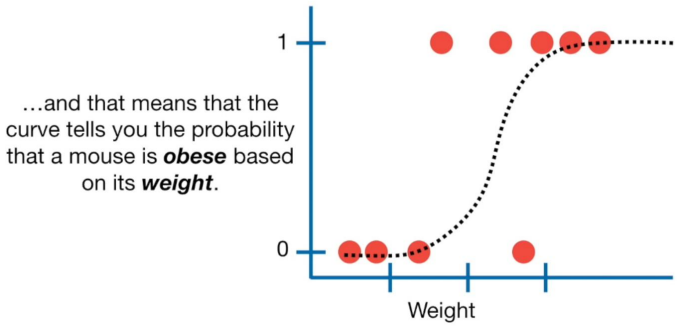
Logistic regression predicts whether something is ***True*** or ***False***, instead of predicting something continuous like ***size***.



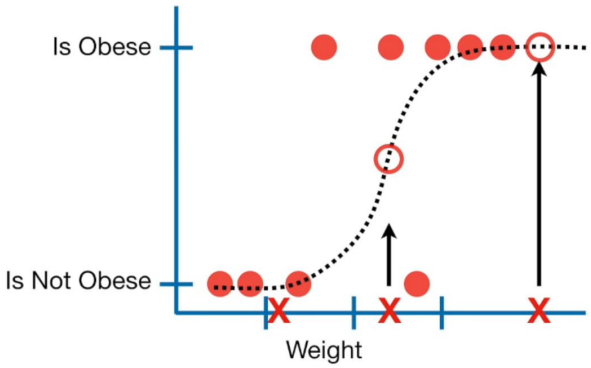
…also, instead of fitting a line to the data, logistic regression fits an "S" shaped "logistic function".



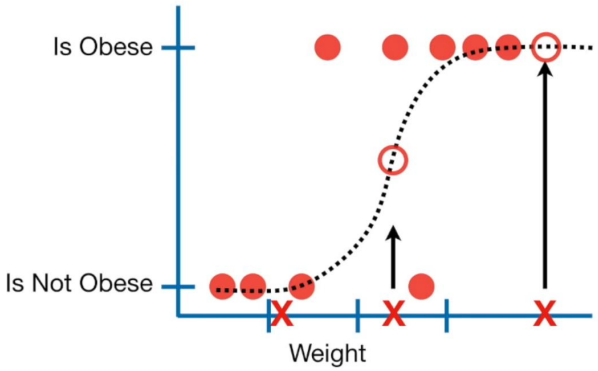




Although logistic regression tells the probability that a mouse is obese or not, it's usually used for classification.



For example, if the probability a mouse is obese is > 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…

Obesity is predicted by **Weight**

…or more complicated models…

Obesity is predicted by **Weight + Genotype**

…or more complicated models…

Obesity is predicted by **Weight + Genotype + Age**

…or more complicated models…

Obesity is predicted by **Weight + Genotype + Age + Astrological Sign**

Obesity is predicted by **Weight + Genotype + Age + 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**.

Obesity is predicted by **Weight + Genotype + Age + Astrological Sign**

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).

Obesity is predicted by **Weight + Genotype + Age + Astrological Sign**

~~vs.~~

Obesity is predicted by **Weight + Genotype + Age**

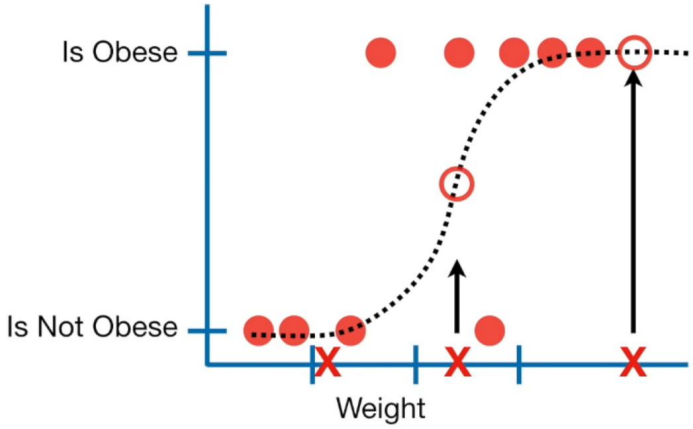
Instead, we just test to see if a variable's effect on the prediction is significantly different from 0.

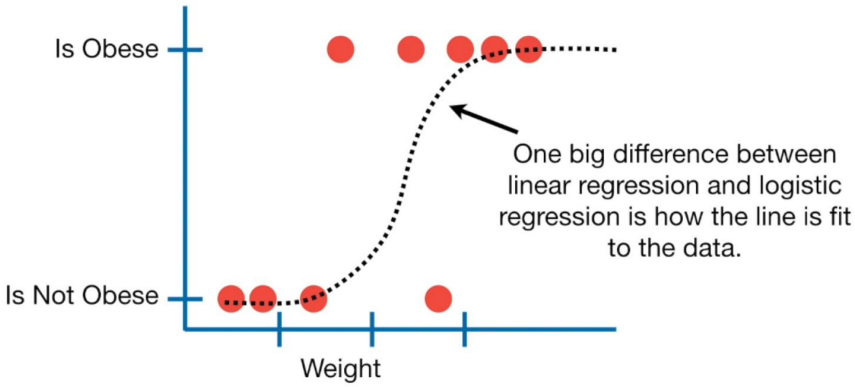
Obesity is predicted by **Weight + Genotype + Age + Astrological Sign**

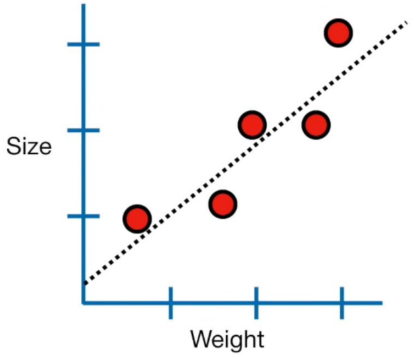
If not, it means the variable is not helping the prediction.

(psst… we use "Wald's Test" to figure this out. We'll talk about that in another StatQuest)

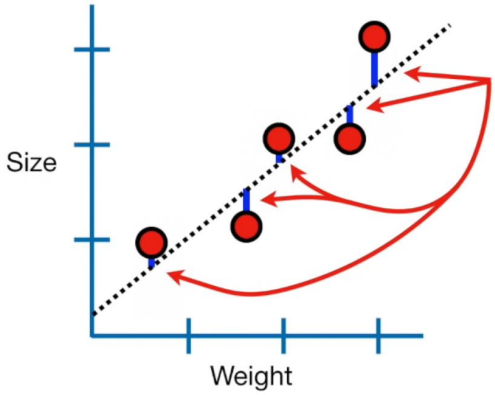
Logistic regression's ability to provide probabilities and classify new samples using continuous and discrete measurements makes it a popular machine learning method.







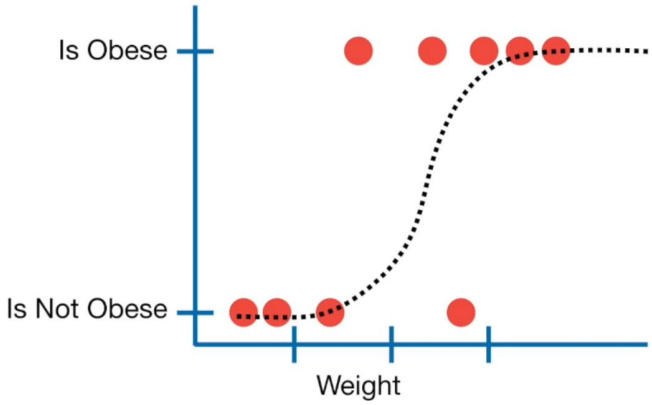
With linear regression, we fit the line using "least squares".



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^2 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^2.



Instead it uses something called "maximum likelihood".