STATISTICS AND BIG DATA '25-'26

Very gentle intro to **Logistic Regression**

Niccolò Salvini, PhD

Principles of logistic regression

1 fit a **logistic regression** to data

2 **MLE** estimation

LR Interpretation

3 **LR coefficients**

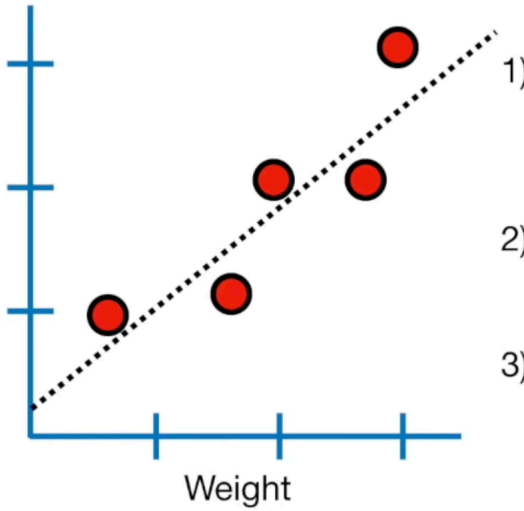
live coding session!

Section 1

recap

UNIVERSITÀ CATTOLICA del Sacro Cuore

recap linear regression



…and with that line, we could do a lot of things:

1) Calculate R^2 and determine if ***weight*** a***size*** are correlated. Large values imply large effect.

2) Calculate a p-value to determine if the *R* value is statistically significant.

3) Use the line to predict ***size*** given ***weigh***

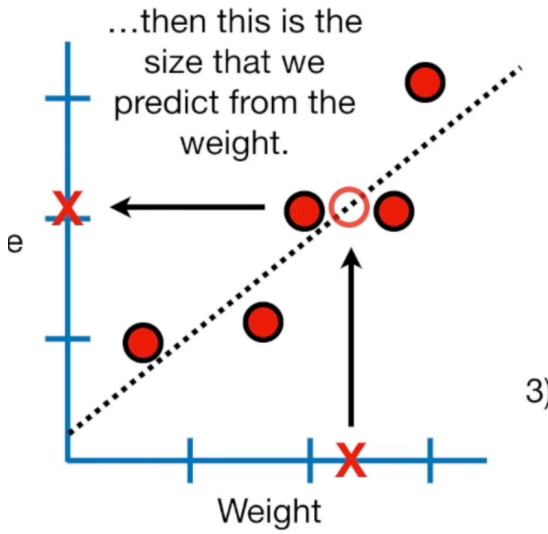
we had some data about **mices**

We fitted a line with Least Square method computing distances

we calculate R2 and R2 pvalues (F statistics)

We saw the summary of the model and we predicted a newly observed mice size.

recap linear regression

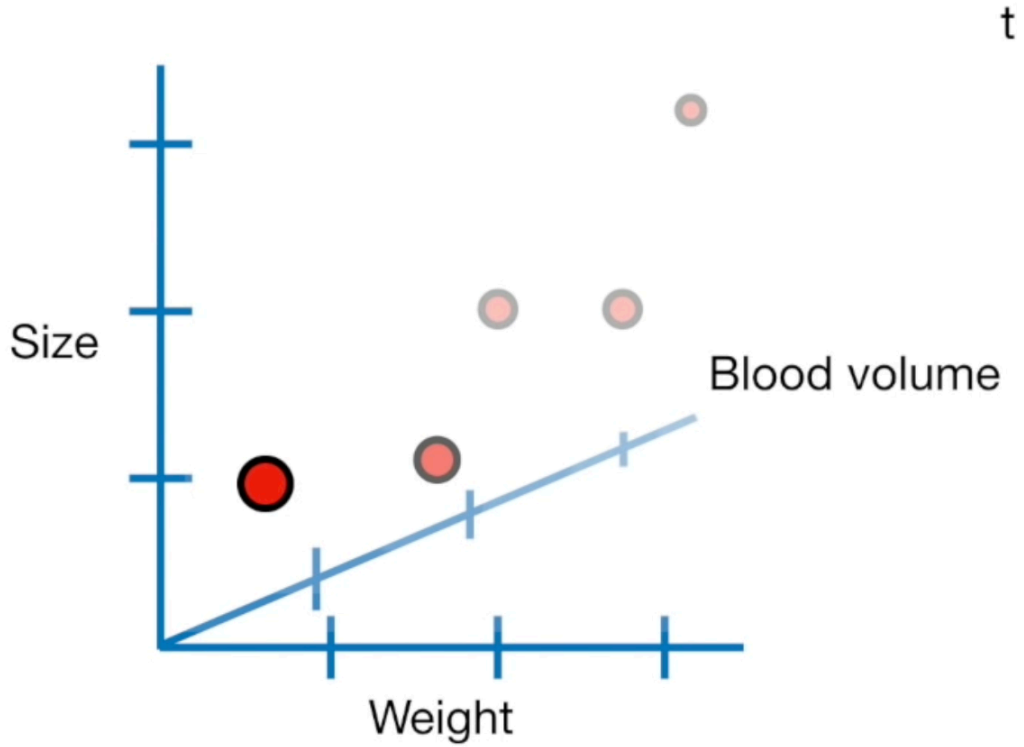


3) Use the line to predict ***size*** given ***weight***

Predicting

Given a weight we can project from the x-axis to the line, them left to the y-axis and know the predicted Size for an unobserved mice weight.

recap linear regression



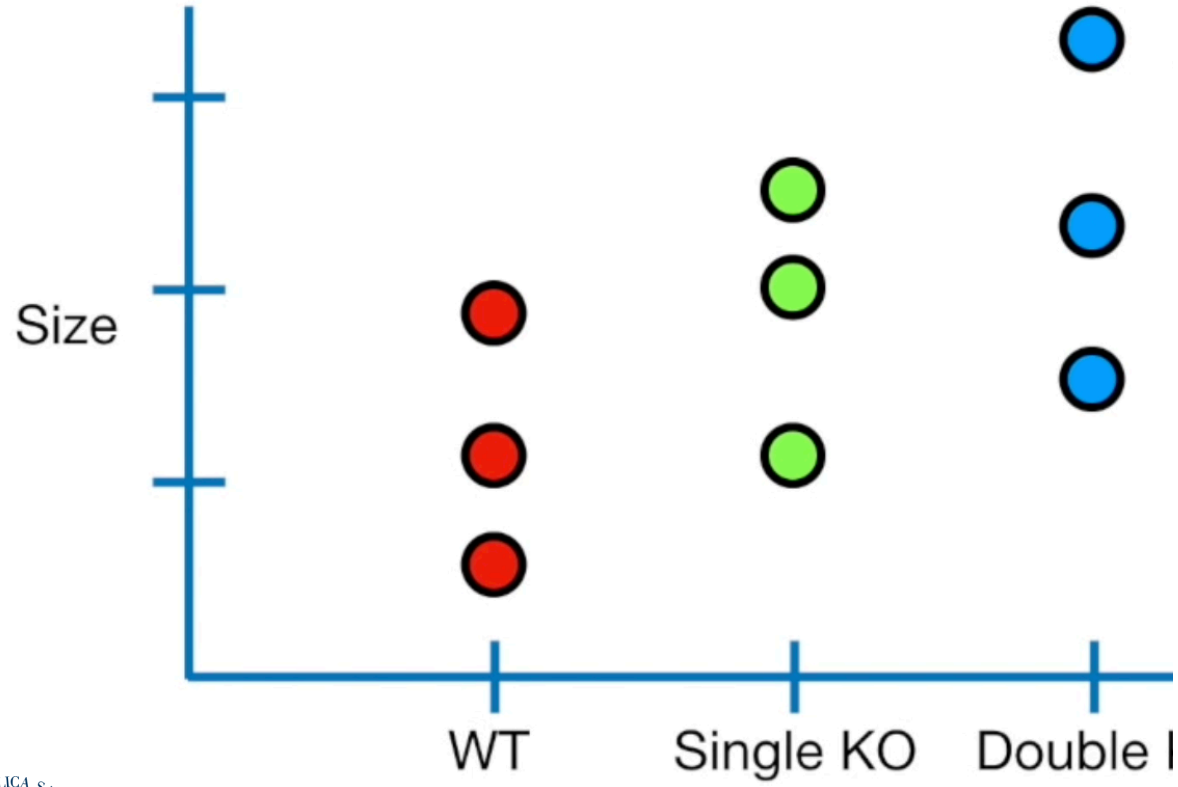
Multiple LR

We extended the concepts to more than 1 X predictors (right end side of ~).

We computed the adj R2 keeping track of the degrees of freedom (params). We know criterias on how to select models and diagnose pathological behaviors, such as:

- **multicollinearity** (vif)
- **nonnormal residuals**
- **heteroscedasticity**

recap linear regression



… also with cateogrical predictors (factors)

we saw how R handles category data and converts them into a 0-1 columns. This is called **econding**, final dataframe is called **design matrix**.

These are genome type. Do not really care what they are but they say something about mouse Size.

btw how would you test if group means are stat sign different wrt Size?

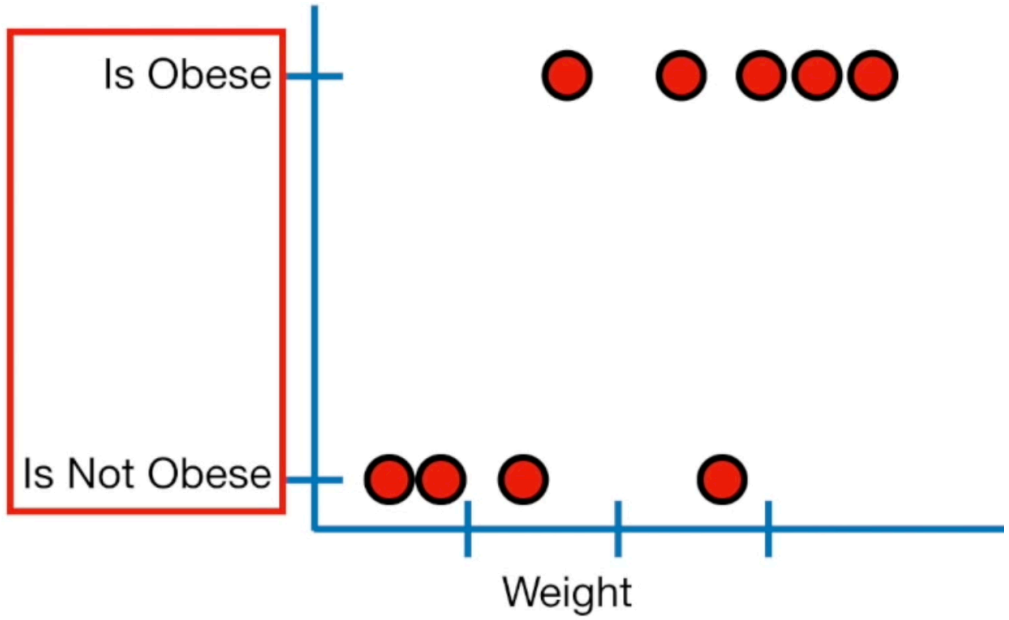
Section 2

Principles of **Logistic Regression**

UNIVERSITÀ CATTOLICA del Sacro Cuore

Fit a **logistic regression** to data

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



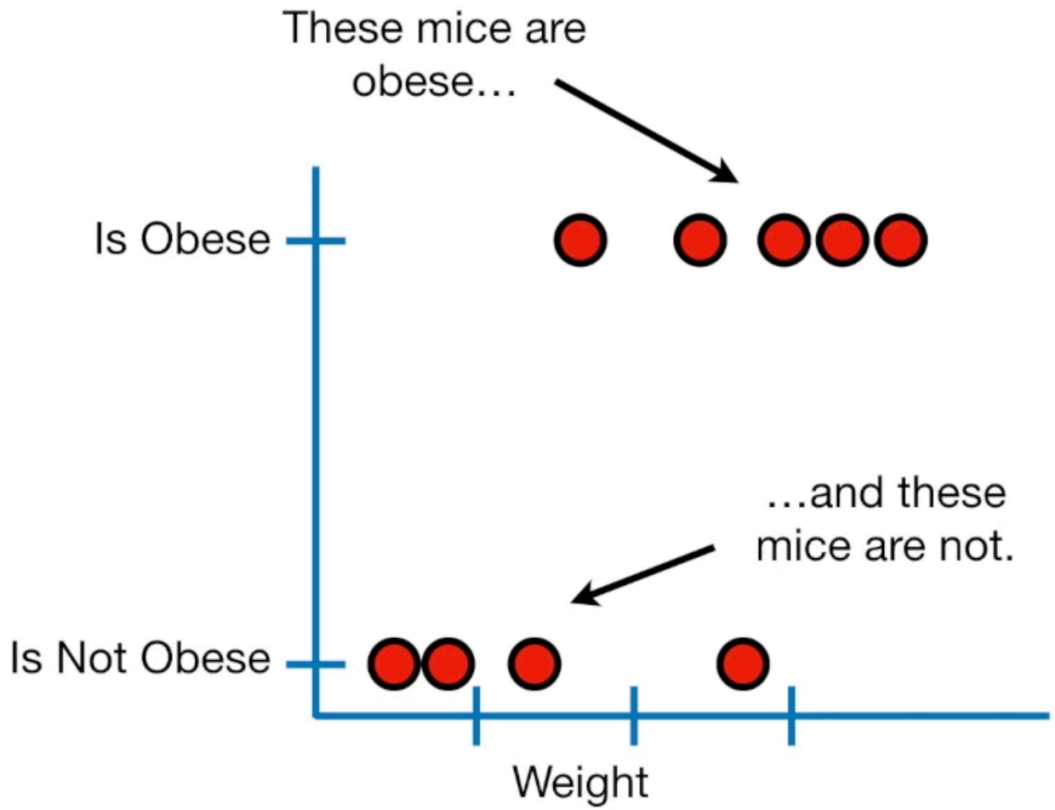
logistic regression

It is conceptually much like linear regression but instead of having a numeric values like **Size at the Y,** you have something like **True or False, Male of Female.**

Here comes the difference betw variables:

- categorical/discrete
- numeric/continuous

Fit a **logistic regression** to data

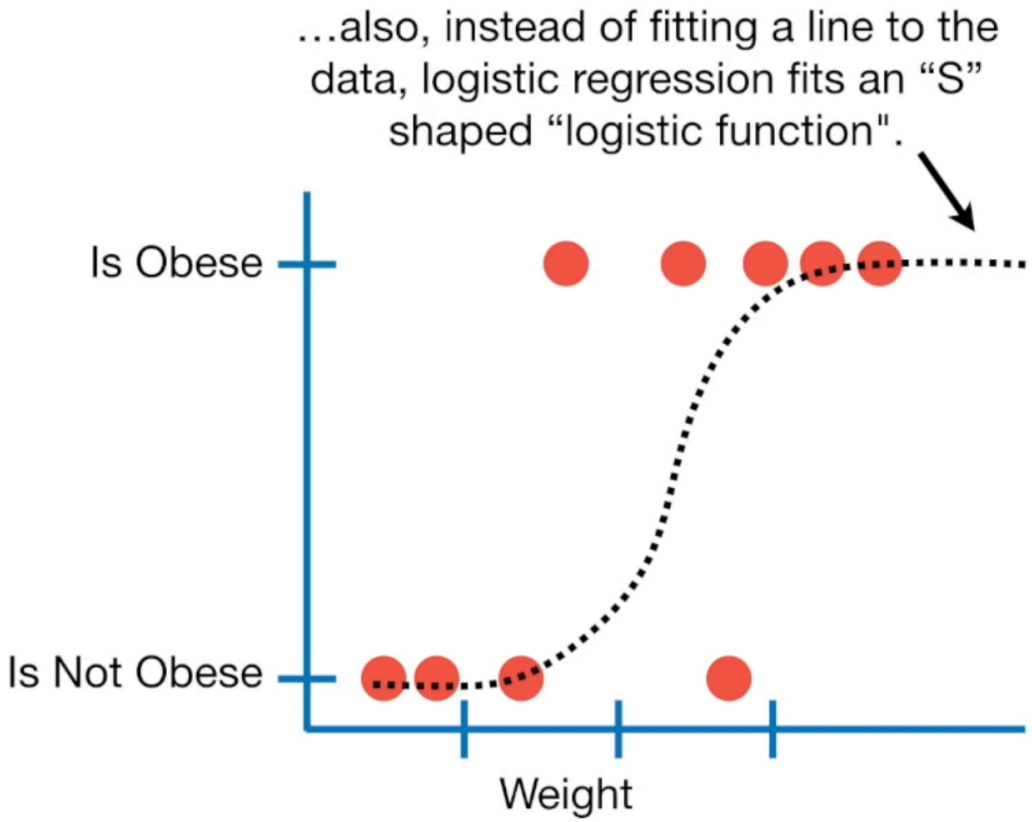


Which are obese and which are not?

Mices in the lower part are **obese**
Mices in the upper part **are not**.

In the x axis you still have continuous variable Weight.

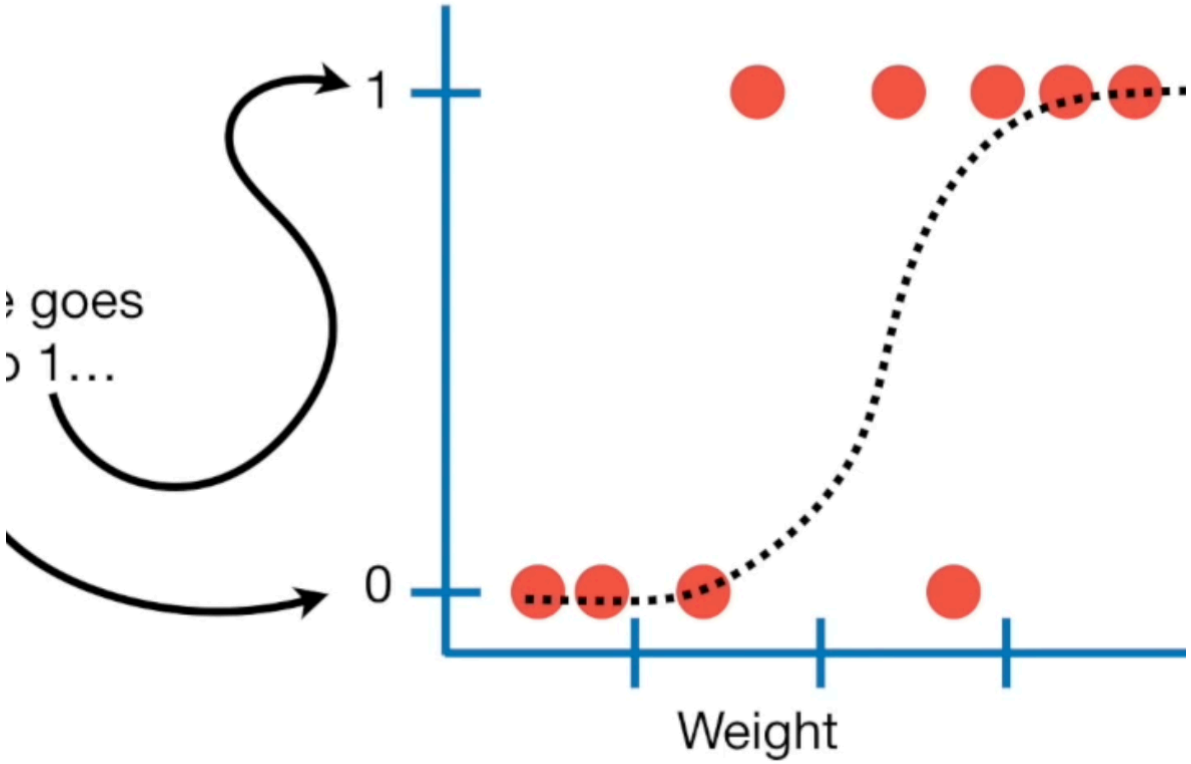
Fit a **logistic regression** to data



S-shaped

Instead of fitting a straight line to data which seems counterintuitive, logistic regression fits a S shaped curve.
This is called Logistic Function (you will see that in R later).

Fit a **logistic regression** to data

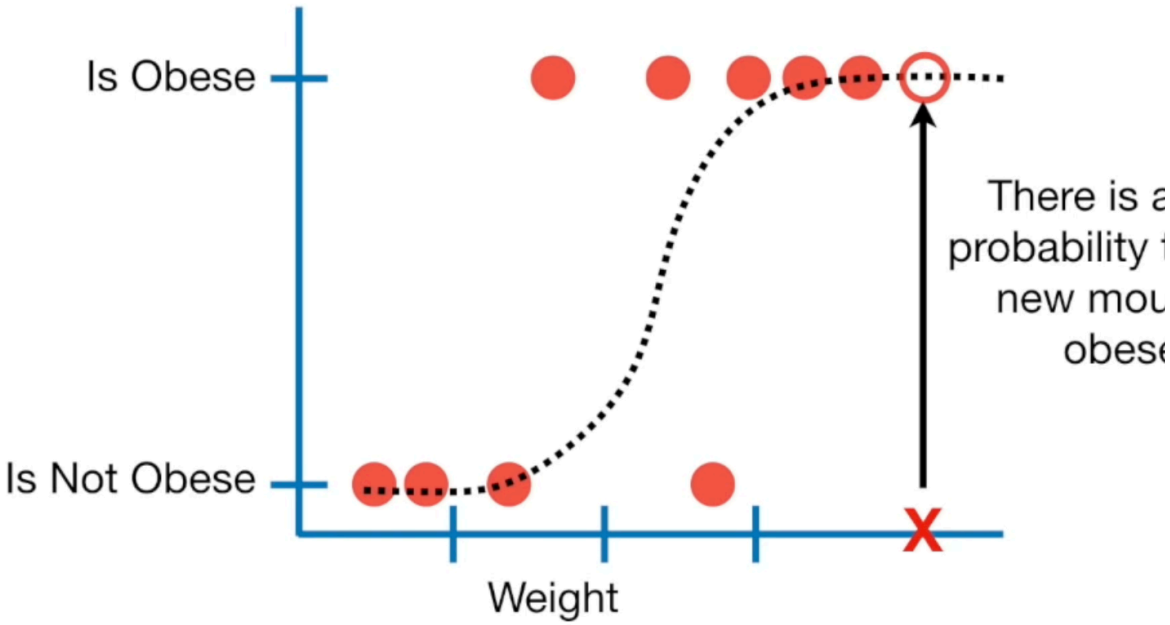


What does the curve say?

The curve tells you the probability of a mouse being obese given its weight.

Note that the y axis ranges from from 0 to 1, like any probability.

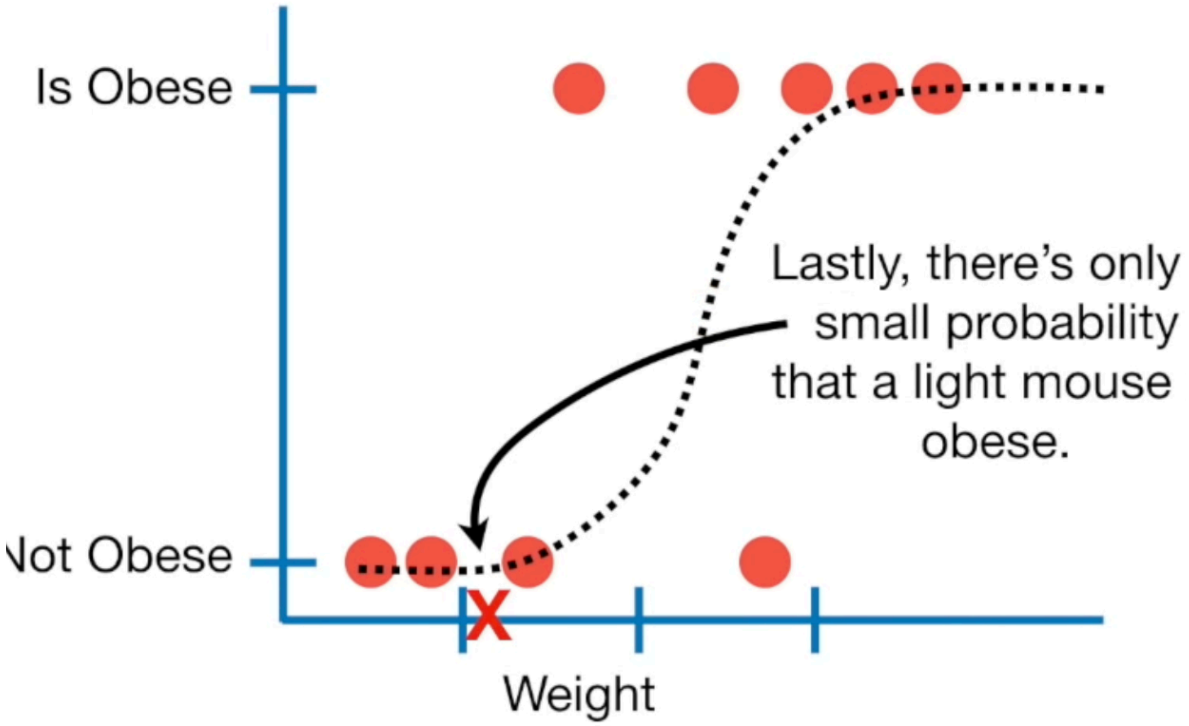
Fit a **logistic regression** to data



Mouse #1

- take a x_1 **weight** (red cross)
- go **up** to the S-line (red circle)
- then project it to the **left** (category)
- Is it obese or not? **yes it its**
- "If we observe a weight of ~4Kg mouse then there is a high probability that this mouse is obese"

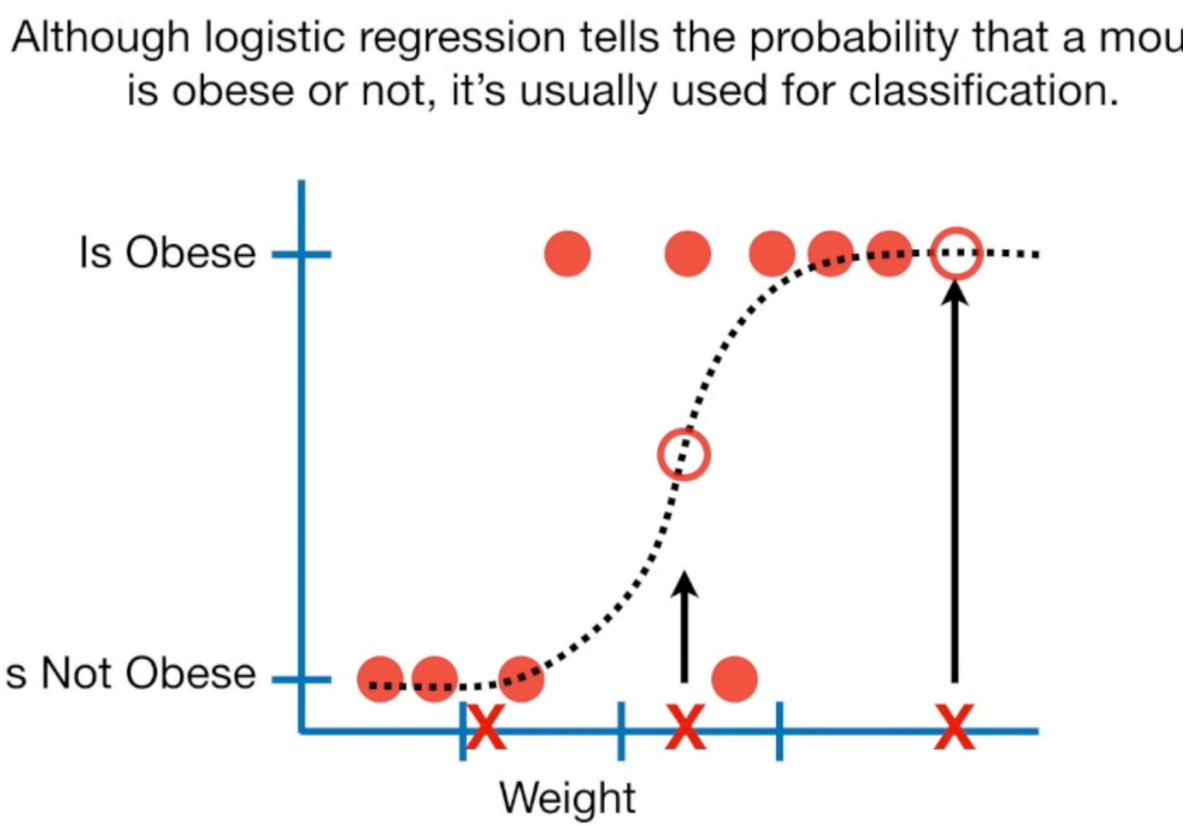
Fit a **logistic regression** to data



mouse #2

- take a x_2 **weight** (red cross)
- go **up** to the S-line (red circle)
- then project it to the **left**
- Is it obese or not? Hell no!
- "If we observe a mouse whose weight is 1.1 Kg then there is a low probability that it is obese"

Fit a **logistic regression** to data



mouse #3

- take a x_3 **weight** (red cross)
- go **up** to the S-line (red circle)
- then project it to the left
- Is it obese or not? Well, not sure.
- "If we observe a mouse whose weight is 2.1 Kg then there is a ~50% probability of it being obese."

MLE estimation

ust like with linear regression, we can make simp
models…

Obesity is predicted by **Weight**

now we introduce “tail length”

Just as in traditional linreg if we want to fit models with other predictors we can. In fact let assume that **Obesity** is now predicted by **Weight**.

MLE estimation

…or more complicated models…

sity is predicted by **Weight + Genotype +** *A*

… Or more complicated models

Obesity is now predicted by Weight and Age.

MLE estimation

…or more complicated models…

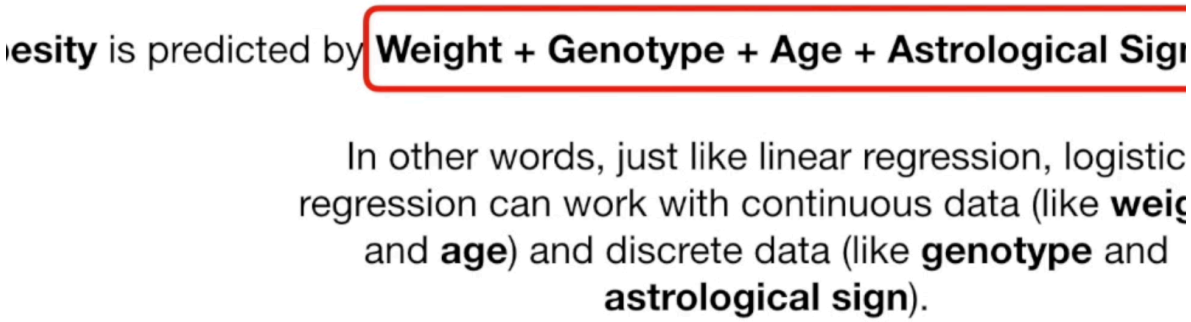
besity is predicted by **Weight + Genotype + Age + Astrological Sig**

… let's add a further predictor

Just like linreg (either single and multiple) can work with discrete and continuous predictors. In this case is **Astrological Sign (discrete)**

Did you remember in which case we discussed Astrological Sign?

MLE estimation



comparing models

However unlike linreg we can easily compare the complicated (the one with many predictors) model with the simple one.

Indeed we can see if a variable's effect on the prediction is significantly different from **0.**

MLE estimation

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).

)besity is predicted by **Weight + Genotype + Age + Astrological Sigr**

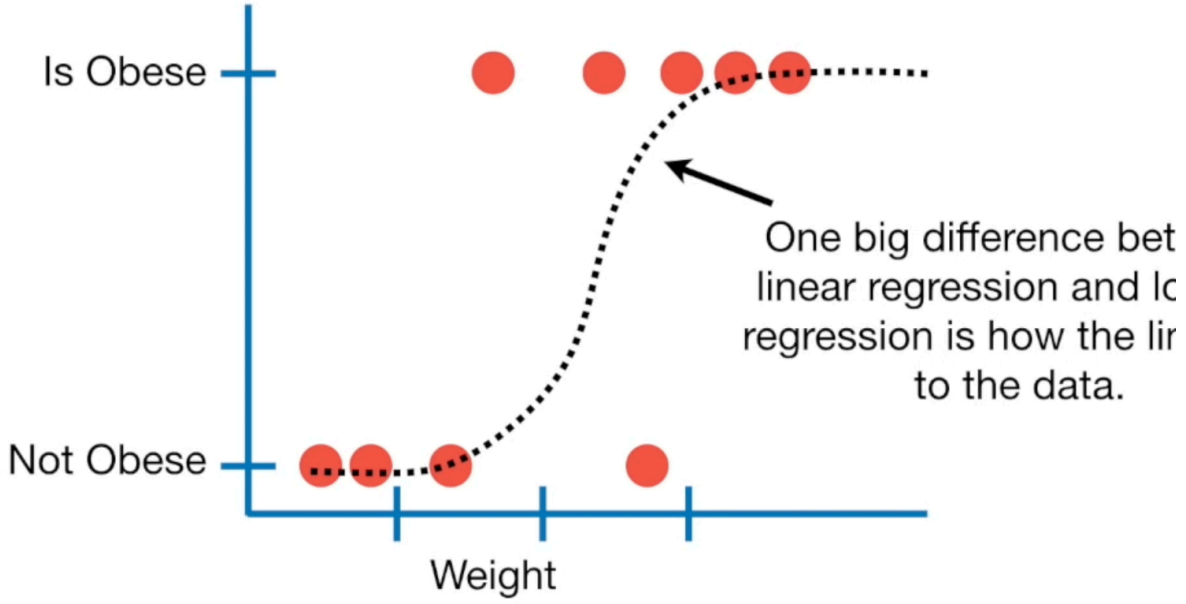
~~Vs.~~

)besity is predicted by **Weight + Genotype + Age**

WE DO EXACTLY AS BEFORE

In this case Astrological sign as you may understand does not help predicting mouse Obesity. We are not *Paolo Fox*.

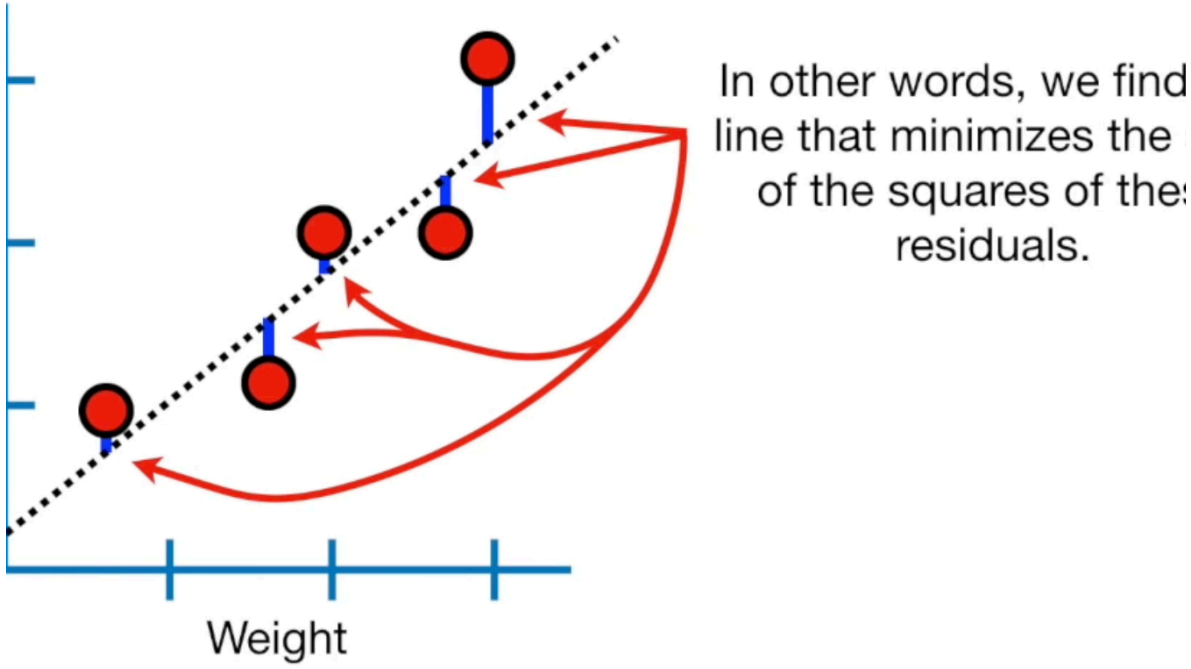
MLE estimation



the way to fit the line to data

The other major difference between **logreg** and **linreg** is the way the S line is fitted to data.

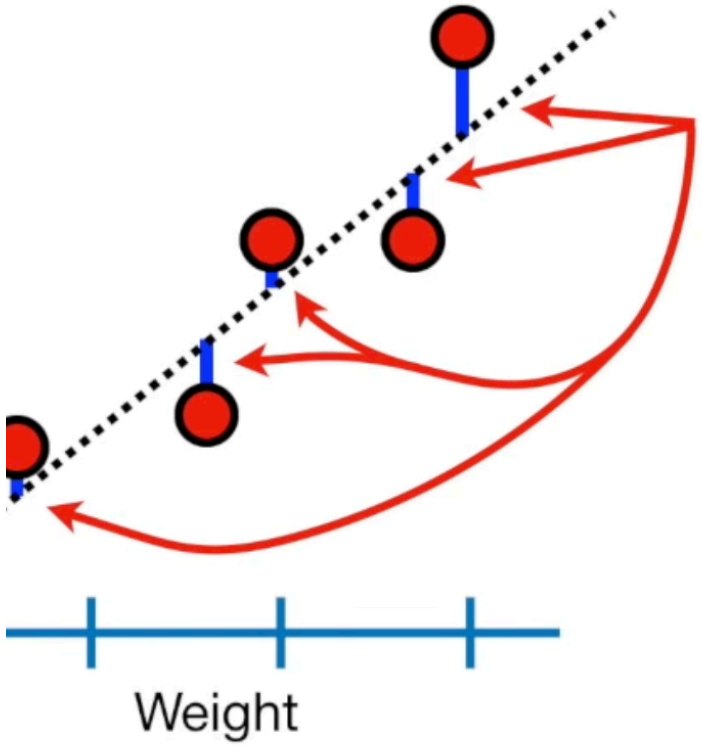
MLE estimation



OLS for linreg

We adopted OLS, minimizing the SS distance between the line and data.

MLE estimation

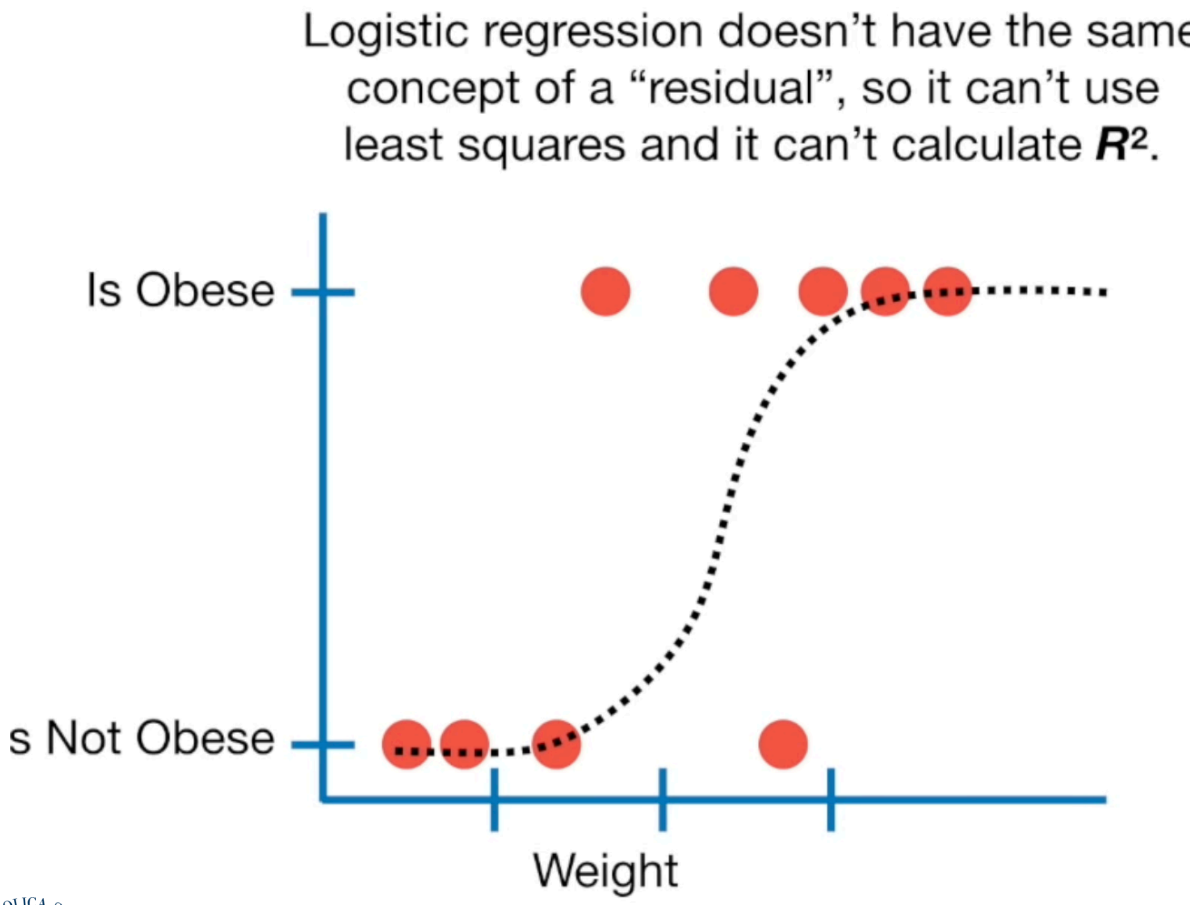


We also use the re
to calculate R^2 a
compare simple m
complicated mo

R2 for linreg

We calculated the residuals also with the aim to compute R2 which was useful as a metric to separate which model is good and which is not.

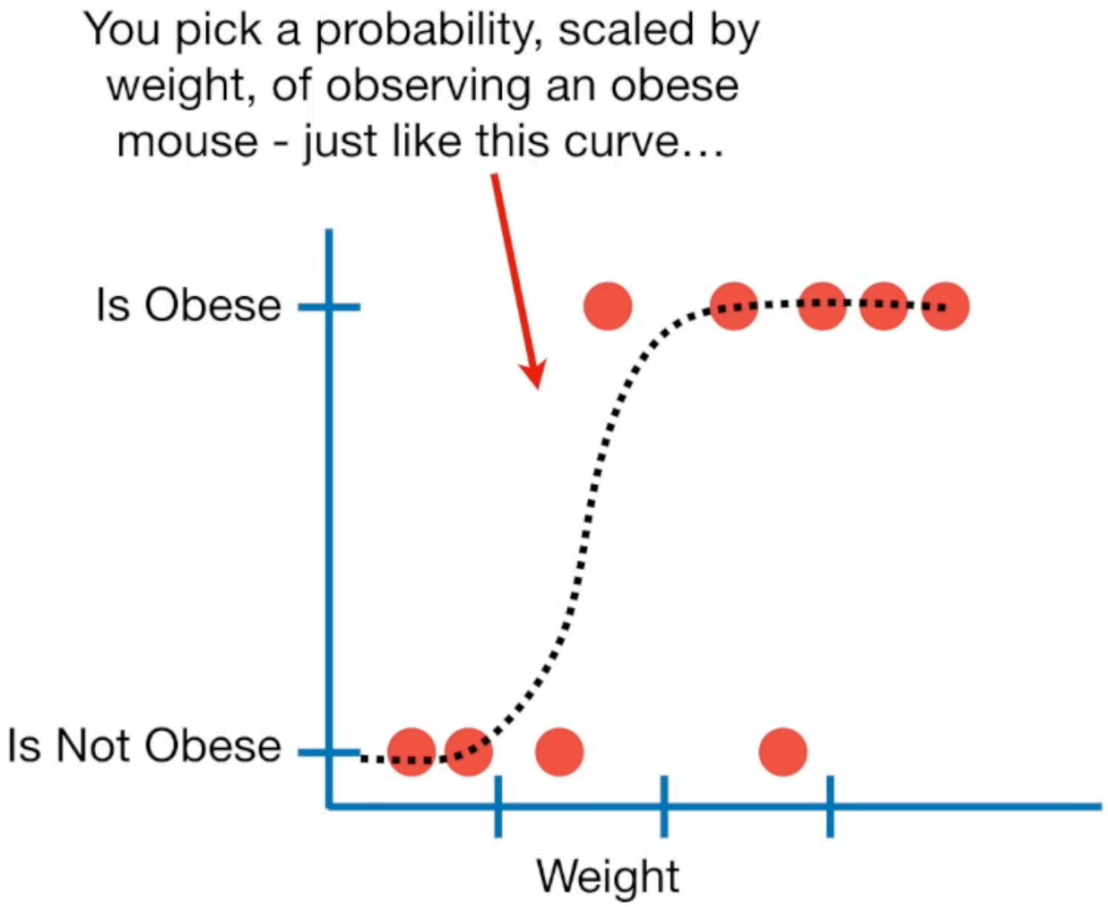
MLE estimation



R2 does not exist for Log reg

Logistic regression does not have the same concept of residuals, *remember we are talking about probabilities*.

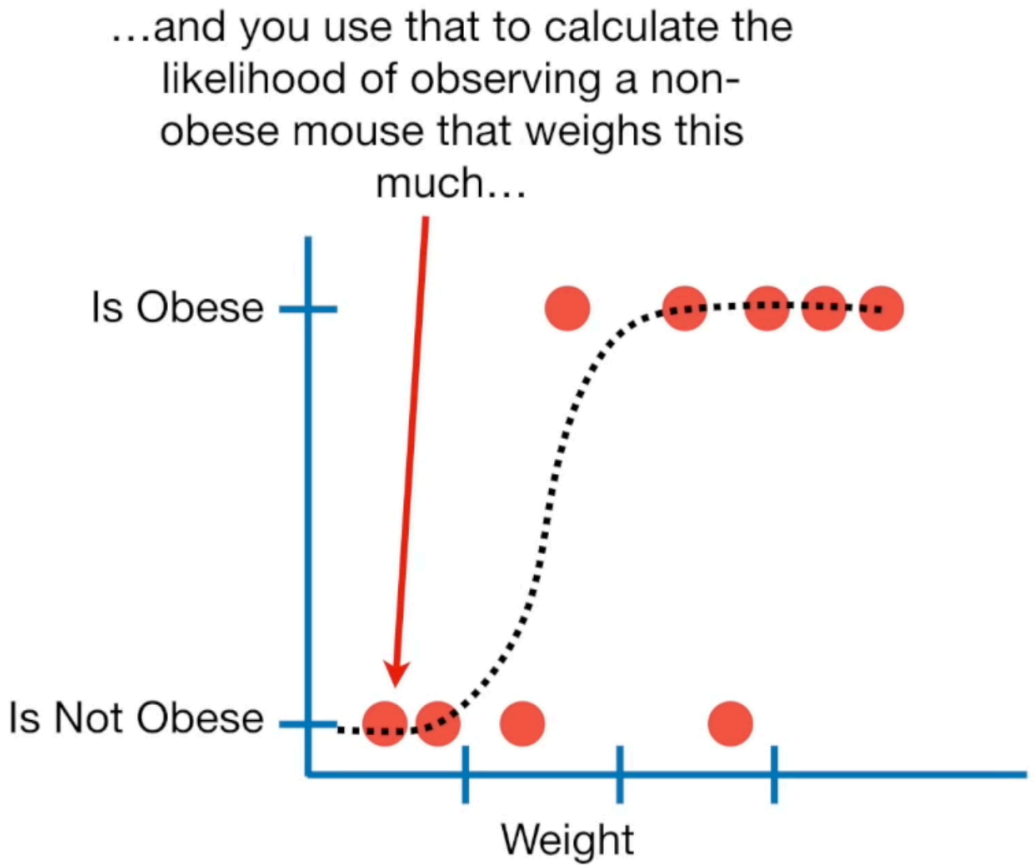
MLE estimation



ML i.e. Maximum Likelihood

Instead of OLS, Log Reg uses this method called Maximum Likelihood estimation MLE which in a nutshell

MLE estimation

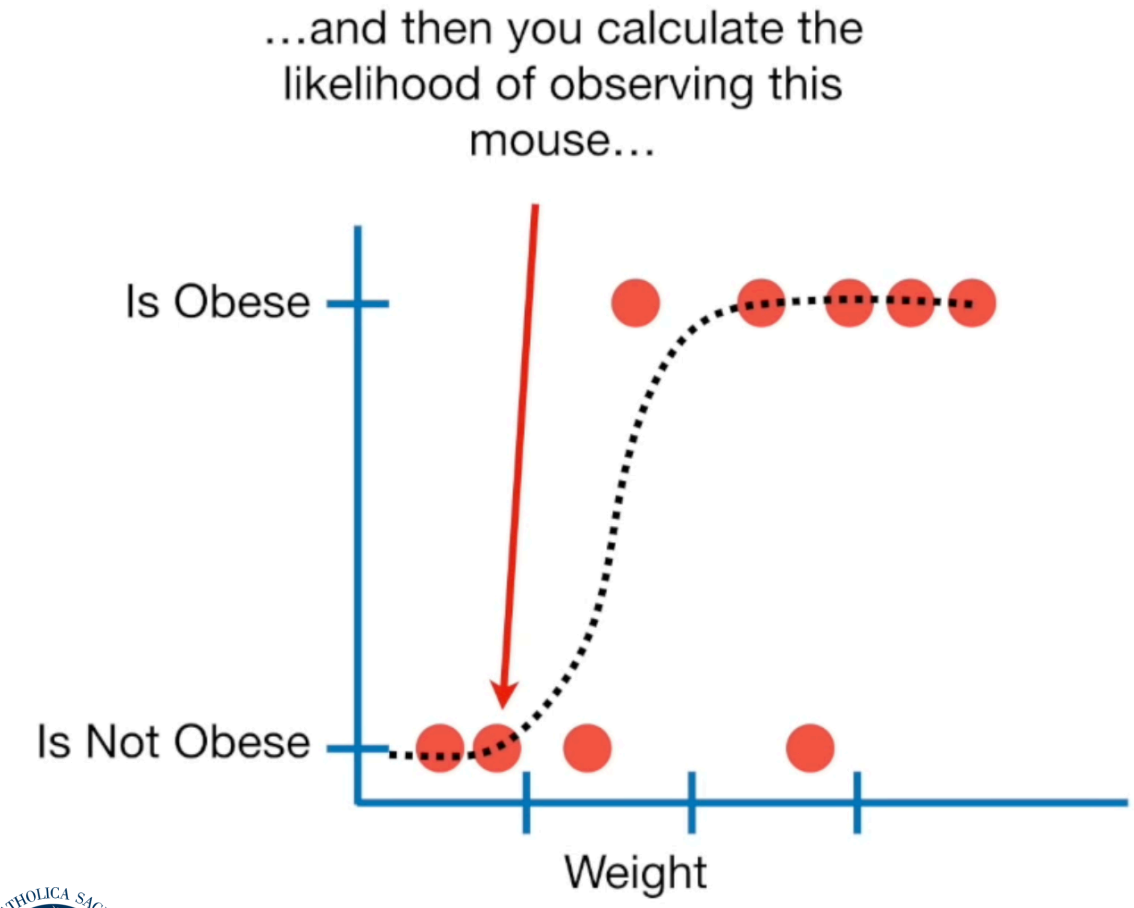


ML in a nutshell

We start by drawing a S line and calculating the likelihood of observing a non obese mouse that weights exactly as the red arrow.

remember the process: *up to the line, then left to the right.*

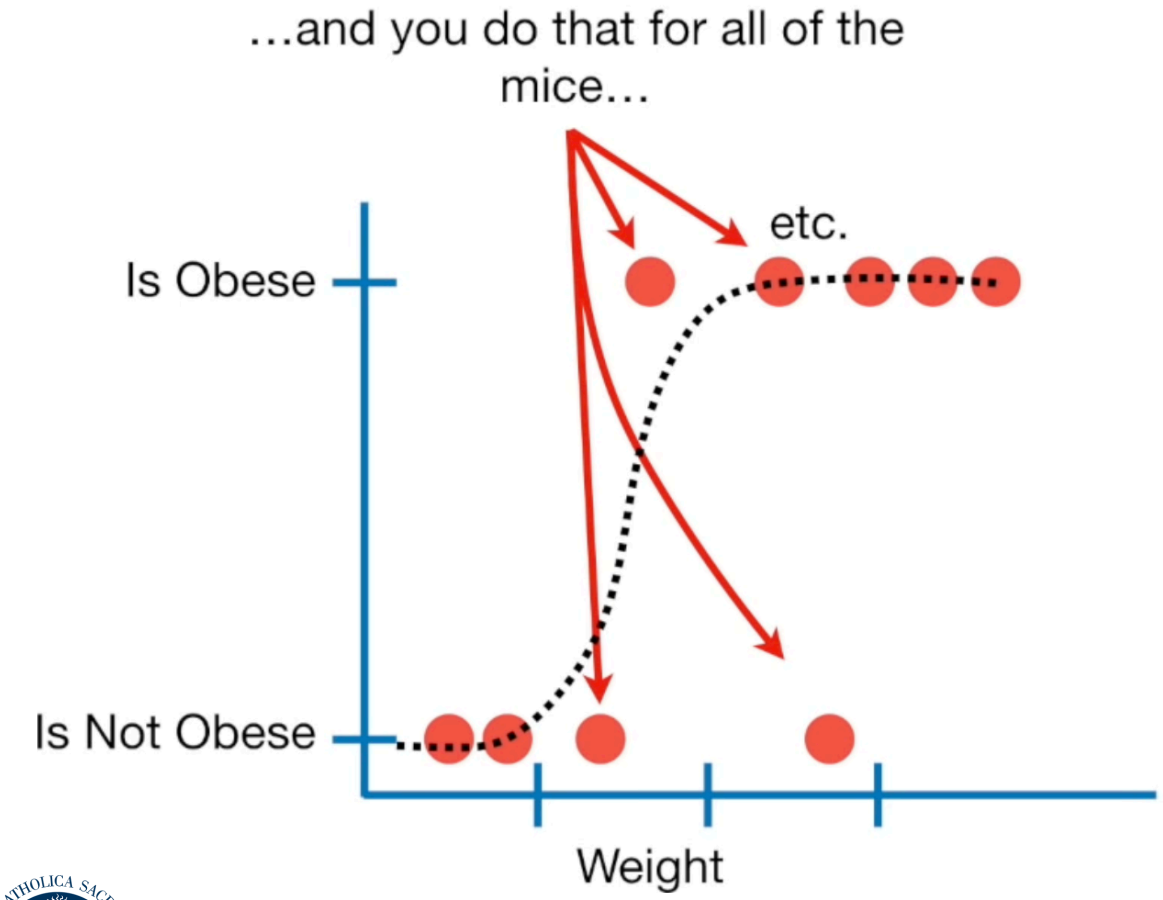
MLE estimation



2nd mouse

we keep doing that for each of the mouses…

MLE estimation



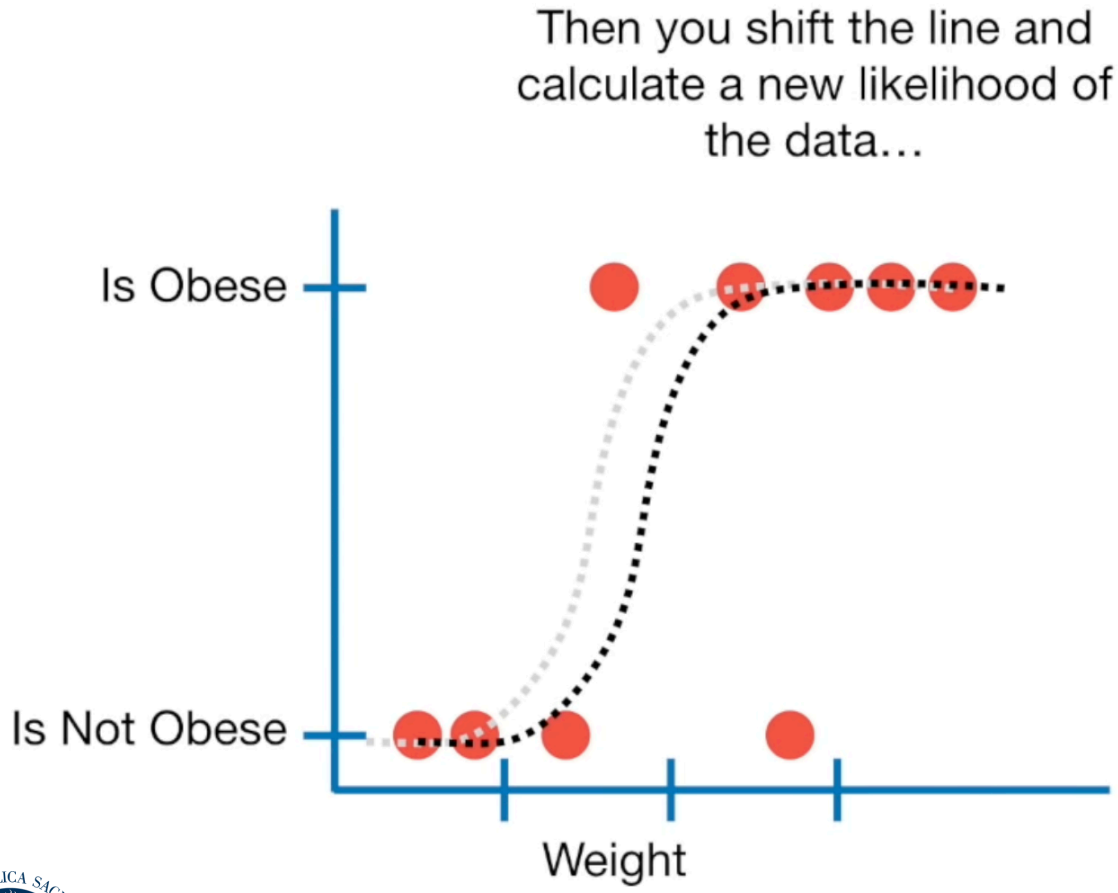
… for all the mouses

then you multiply all of this likelihood together, Remember likelihood = probability, When you multiply a probability you are looking at the combined event.

Inter is winning the match against Atalanta this Sunday 90%
Fiorentina is winning the match against Milan this Sunday 20%

combined event result is 90%*20% (joint probability)

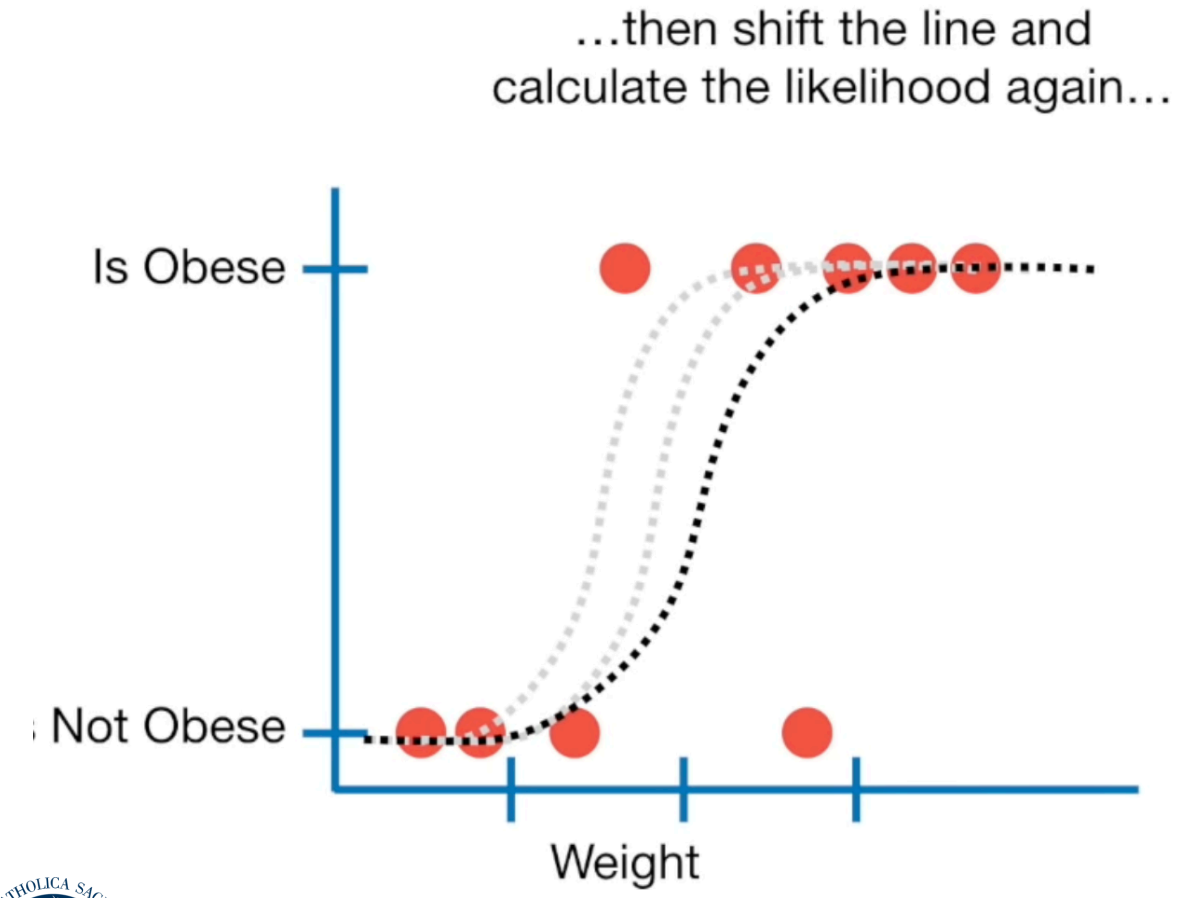
MLE estimation



What you obtain is the likelihood of this data given this line

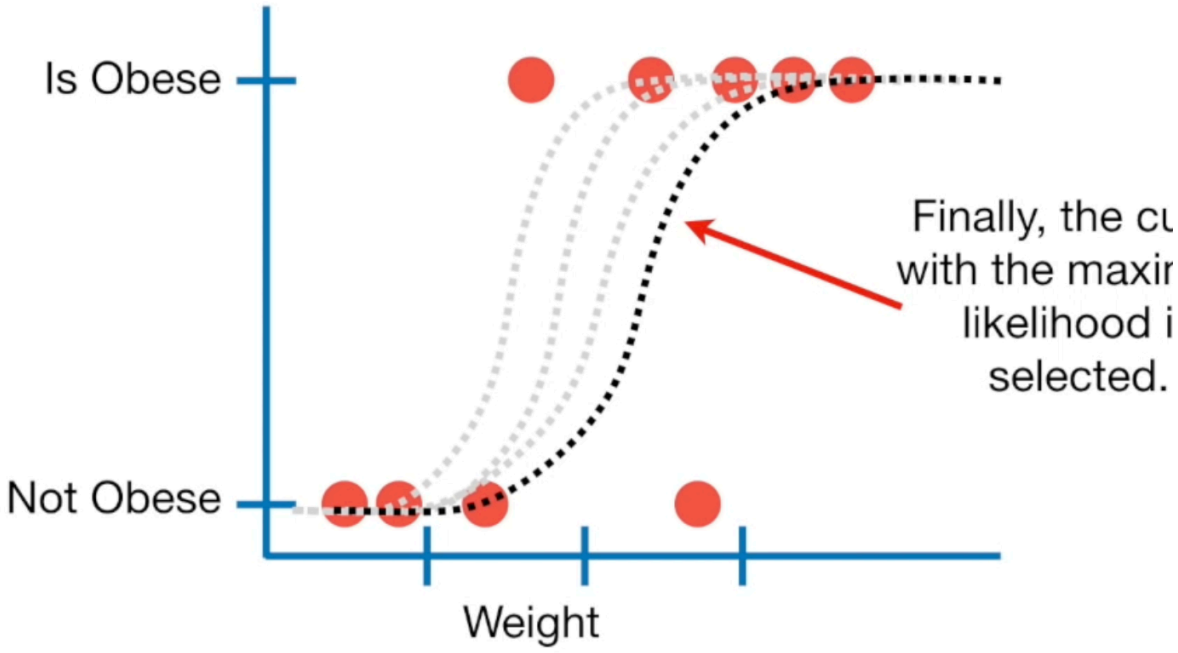
What about all of the other lines that can be drawn?

MLE estimation



… and this line?

MLE estimation



… and this line?

Finally the curve with the **maximum value** is selected.

Meaning we are selecting the S line which better fits the probabilities given data that mouses are obese.

Section 3

Logistic Regression
coefs Interpretation

glm **Context**

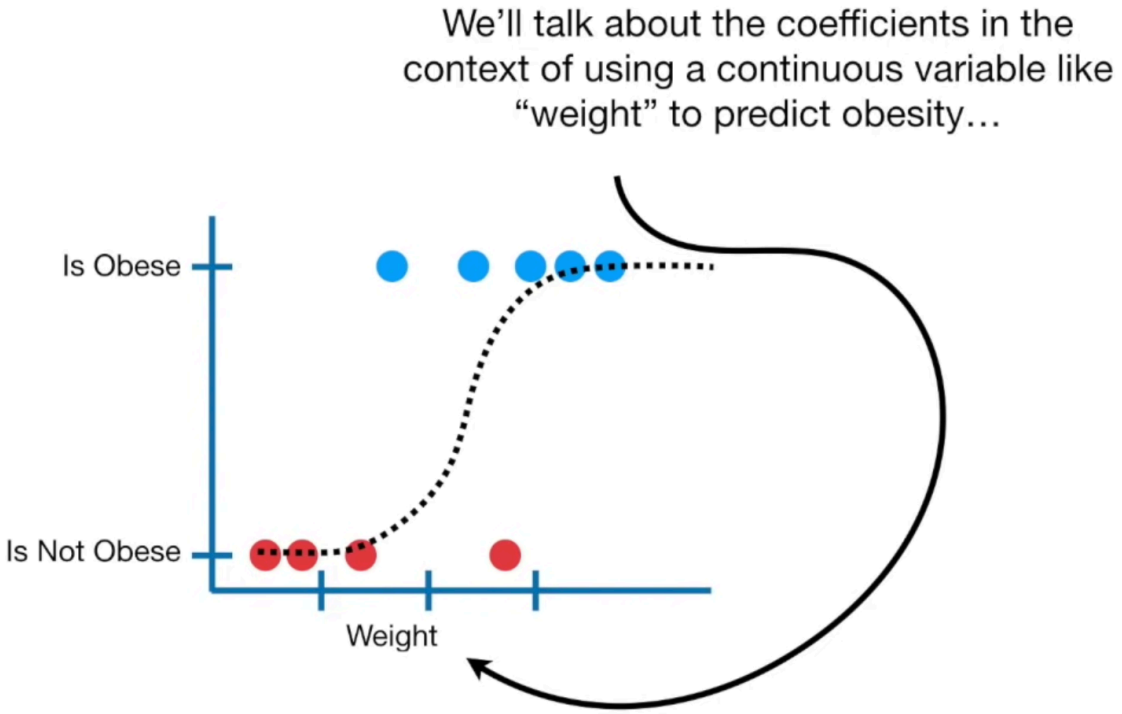
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.476      2.364  -1.471   0.1414
weight          1.825      1.088   1.678   0.0934
```

The coefficients have a critical interpretation

.. and it is hard one too!

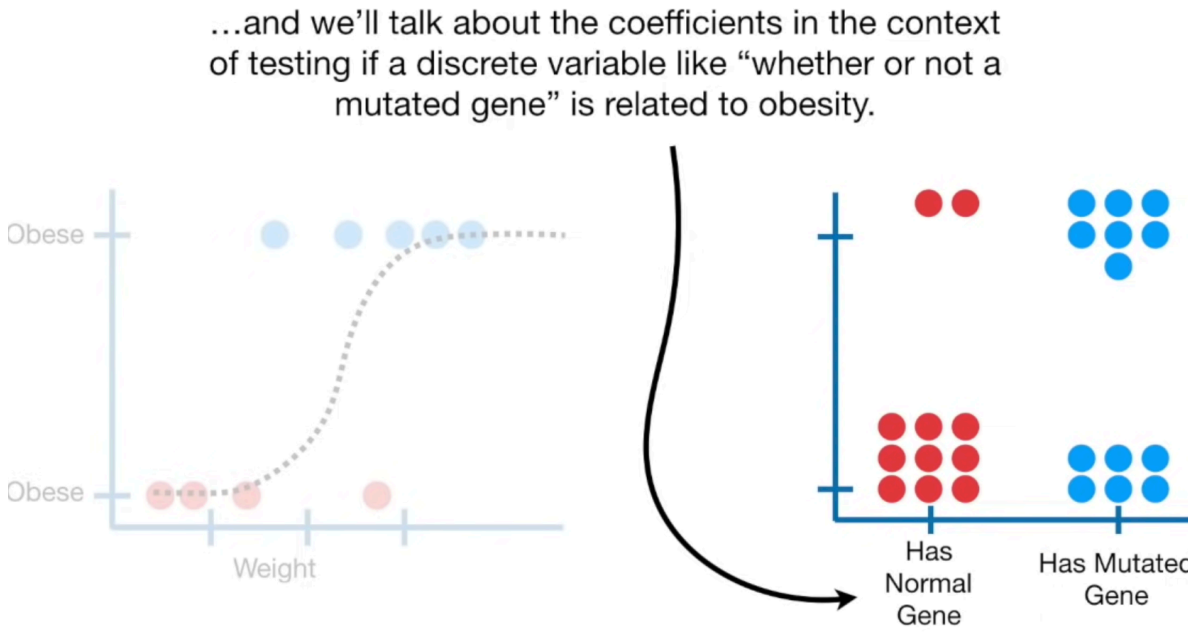
buckle your seat belts.

glm **Context**



Coefficients in the context of continuous var.

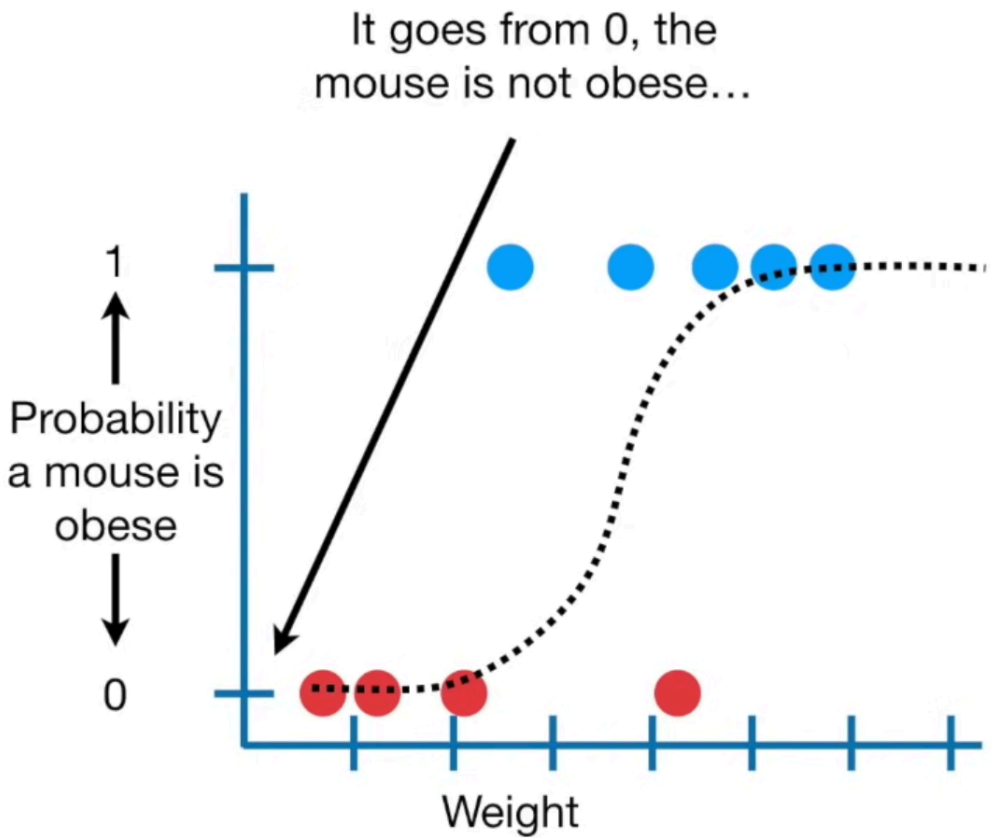
glm **Context**



coefficients in the context of discrete var

if there's time for that. This is the case in which Y is False - True or Male - Female and X variables (i.e. predictors) are lke that too, in this case "has normal genes" vs "has mutated genes"

glm **Context**

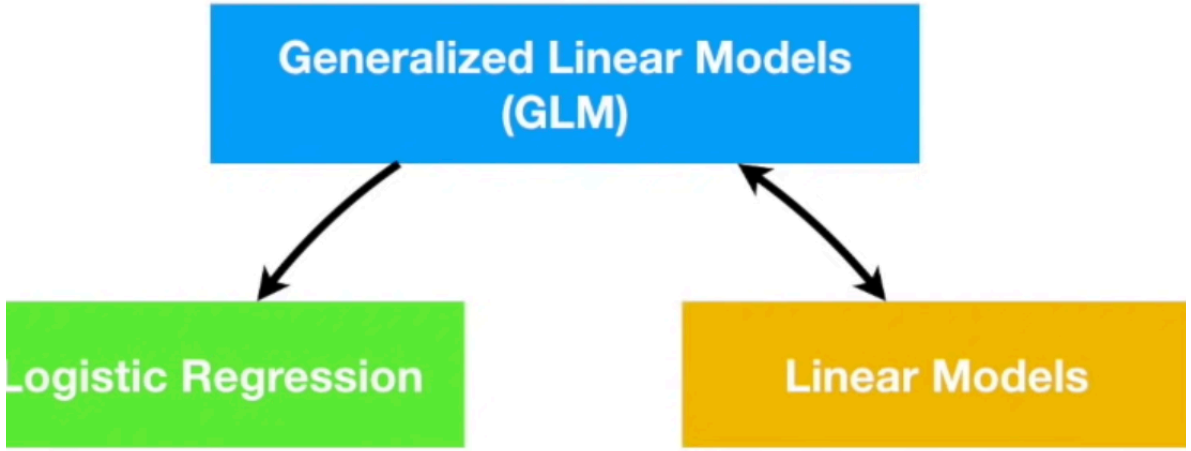


Quick review main ideas

Y axis is a probability a mouse is obese and goes from 0 to 1.

Let's take one of the mouse and calculate the probability is obese or not.

glm **Context**



class of glm

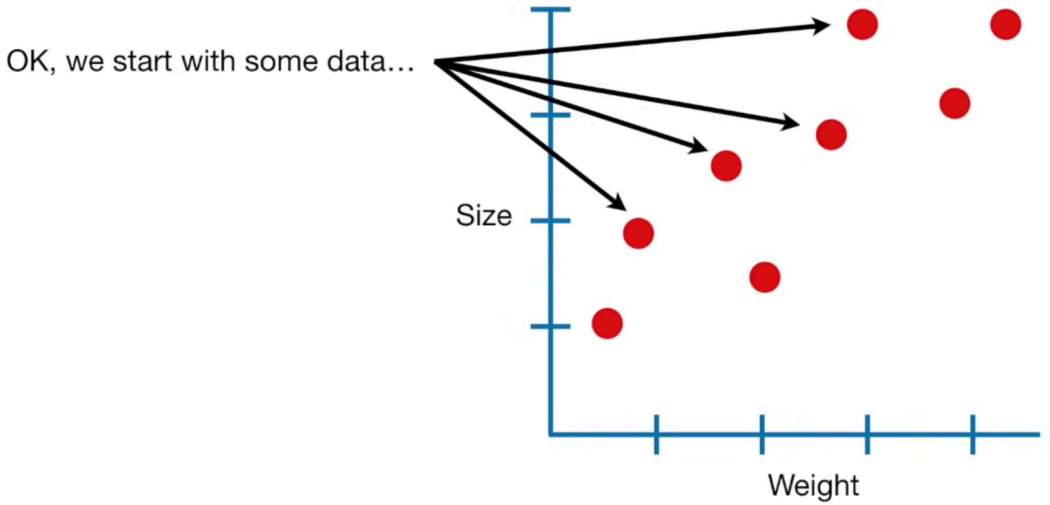
in R glm() you will see deep down in some minutes.

However Linear Regression and Linear model are special case of GLM.

Generalized linear models are a generalization of the concepts and abilitie of regular linear models. that mean if you are faimilar with lienar models, then you can understand Log Reg.

Logistic Regression interpret

with continuous variable predictor



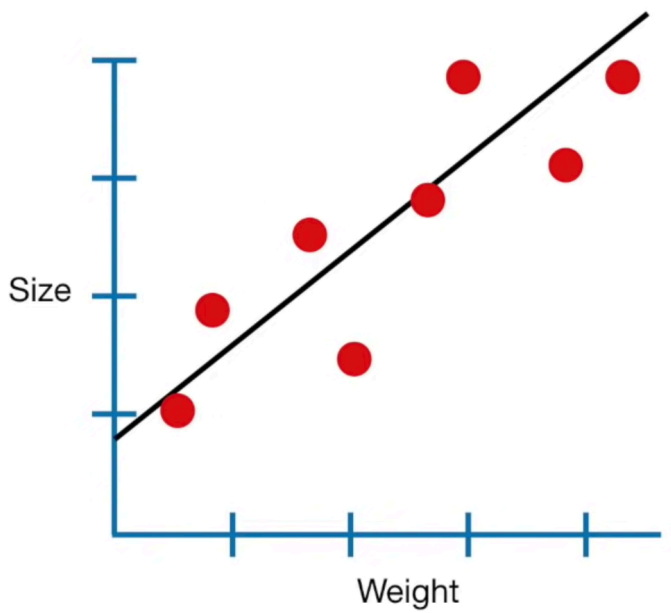
Some data here

Logistic Regression interpret

with continuous variable predictor

size = 0.86 + 0.7 × weight

… to get predicted values for size.



fit a line to data as usual

We have an intercept and a slope.
The intercept is where the line crosses the y axis and the slope is steepness of the line.

upper left the equation.

Logistic Regression interpret

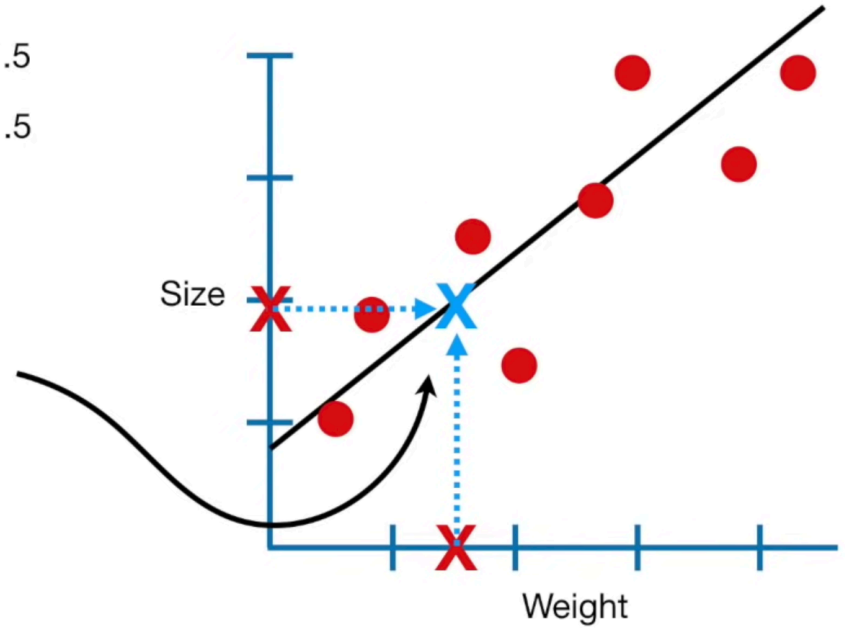
with continuous variable predictor

size = 0.86 + 0.7 × weight

size = 0.86 + 0.7 × 1.5

1.91 = 0.86 + 0.7 × 1.5

…and a mouse with Weight 1.5 and Size 1.91 would end up on the line at this point.

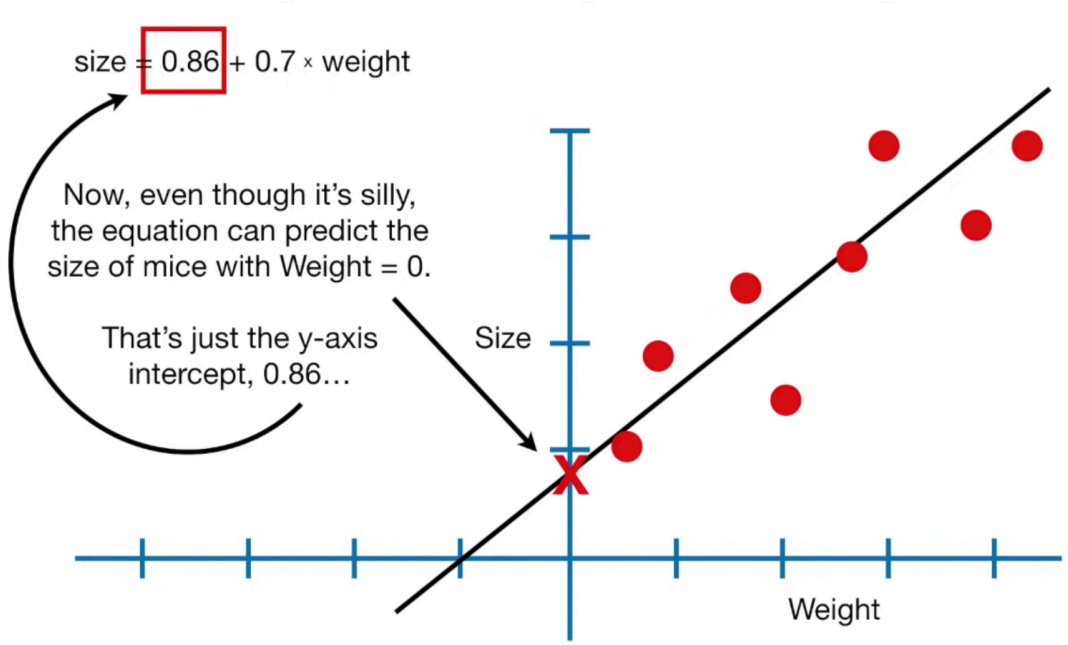


if we would like to have a new mouse weight = 1.5 Kg

then we can predict its **Size** by plugging in the **Weight** in the estimate equatio, resulting in **1.91 Size**

Logistic Regression interpret

with continuous variable predictor

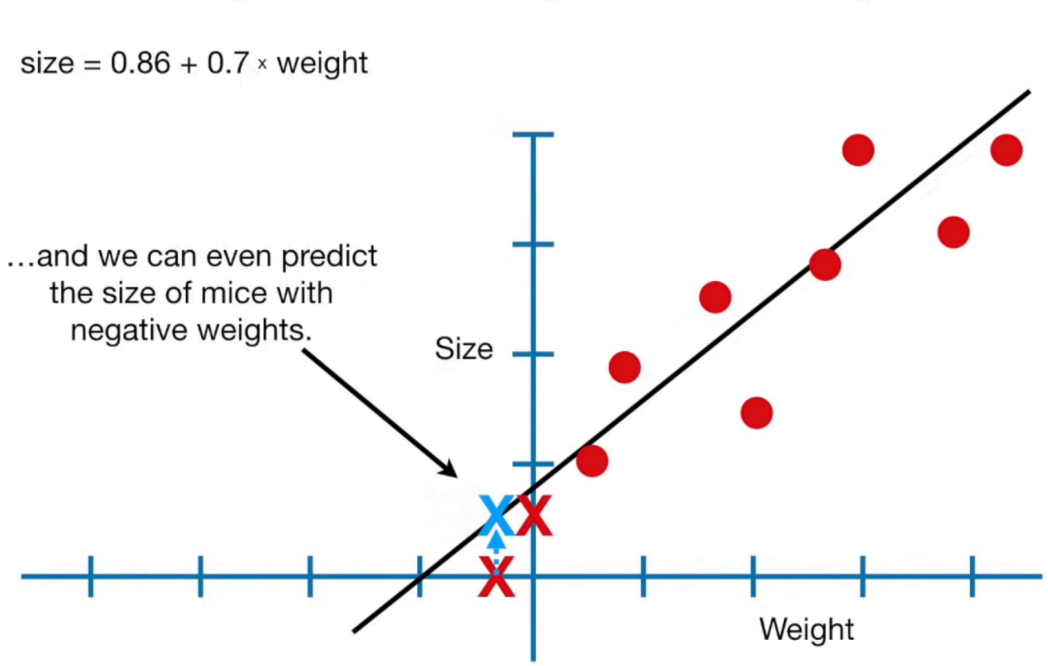


if weight is 0?

Is it non sensical? How could a mouse weight 0. This is just the Y intercept **0.86.**

Logistic Regression interpret

with continuous variable predictor



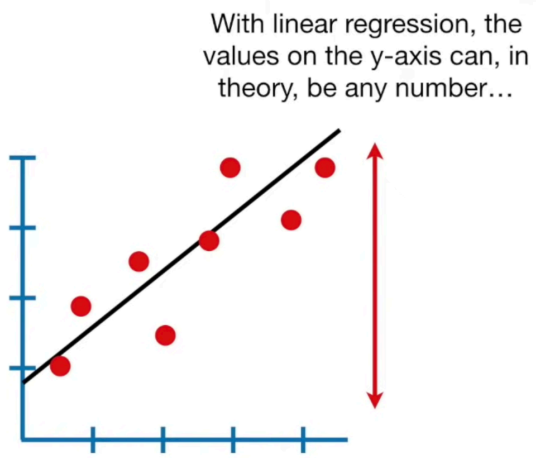
If weight < 0?

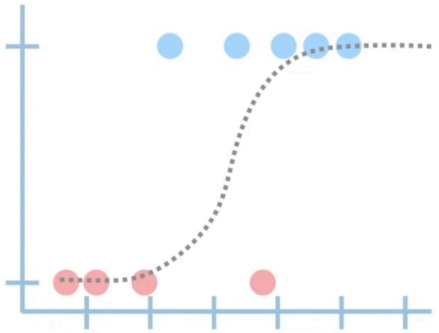
we can keep on with non sense and predict a mouse Size given a negative weight.

We are seeing this because the fact that we are not limiting the equation to a specific domain (weight being > 0) make it easier to solve..

Logistic Regression interpret

with continuous variable predictor



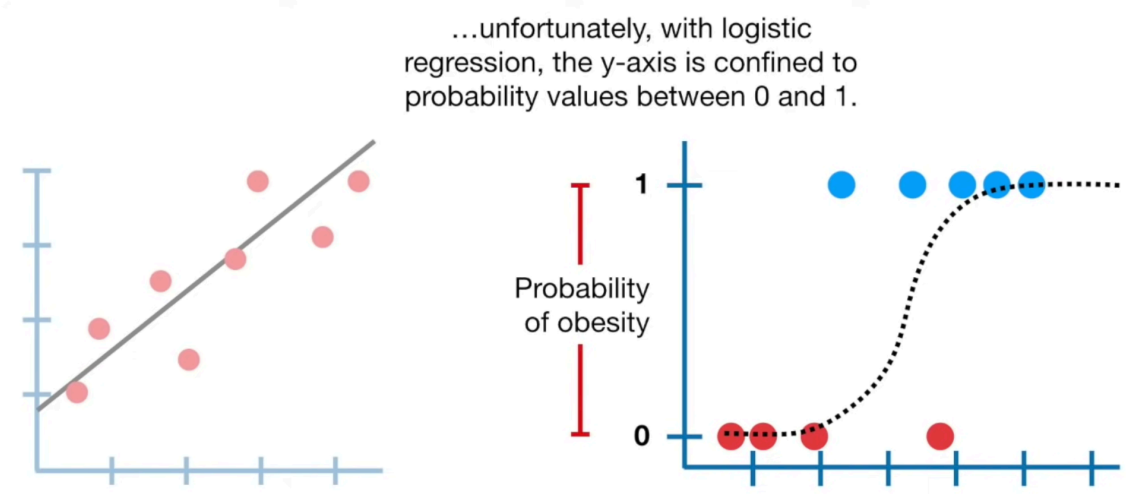


Left hand side

With lienar regression the values on the y axis can be in theory any number.

Logistic Regression interpret

with continuous variable predictor

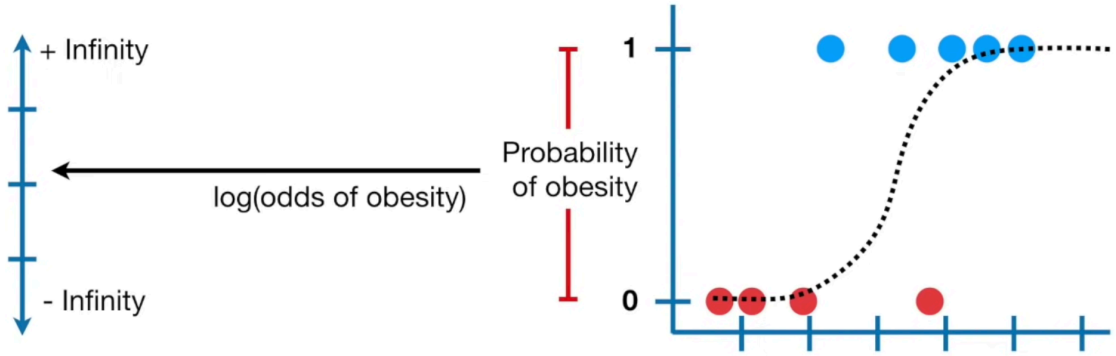


right hand side

Values in logistic regression can be only limited from 0 to 1 since they are probabilites, in our case the probabolity of being obese.

Logistic Regression interpret

with continuous variable predictor

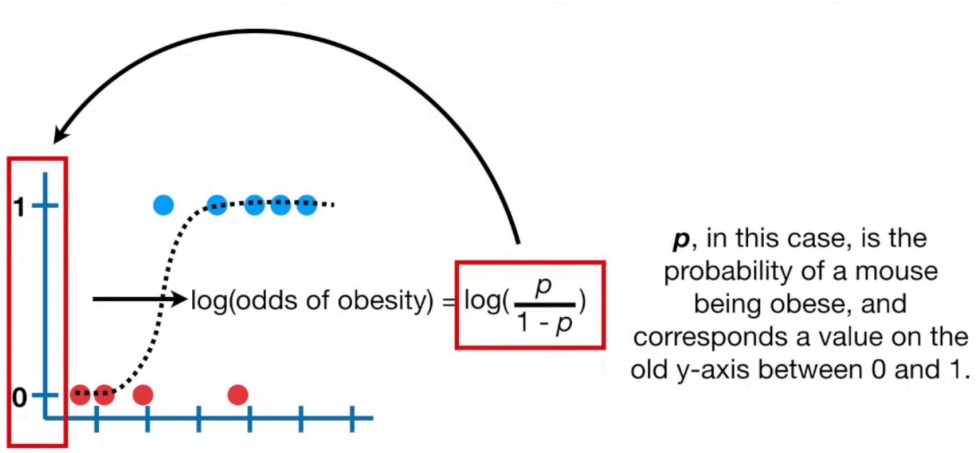


log odds transf

To solve for this problem the Y axis in the logistic regression is transformed from the "probability of obesity" to the "log(odds) of obesity" so just like as the linear regression it can go from -infinity to +infinity.

Logistic Regression interpret

with continuous variable predictor



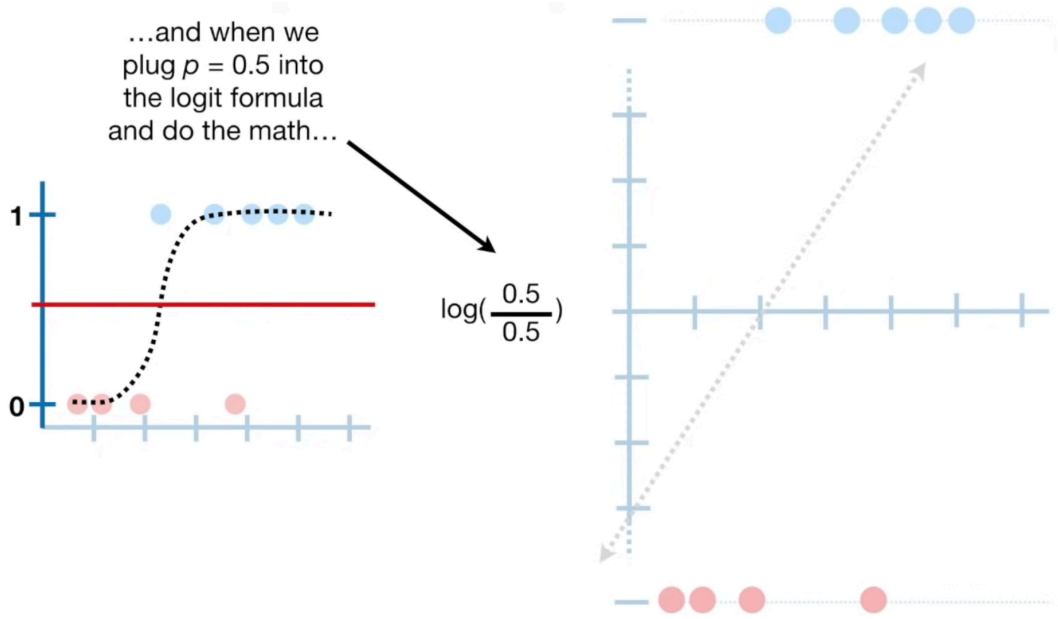
p, in this case, is the probability of a mouse being obese, and corresponds a value on the old y-axis between 0 and 1.

let's try to apply it

lets try to transform the Y axis to the log(odds) and w do that with the logit function (the one in the square red).
p in this case is the probability of a mouse being obese (from 0 to 1)

Logistic Regression interpret

with continuous variable predictor



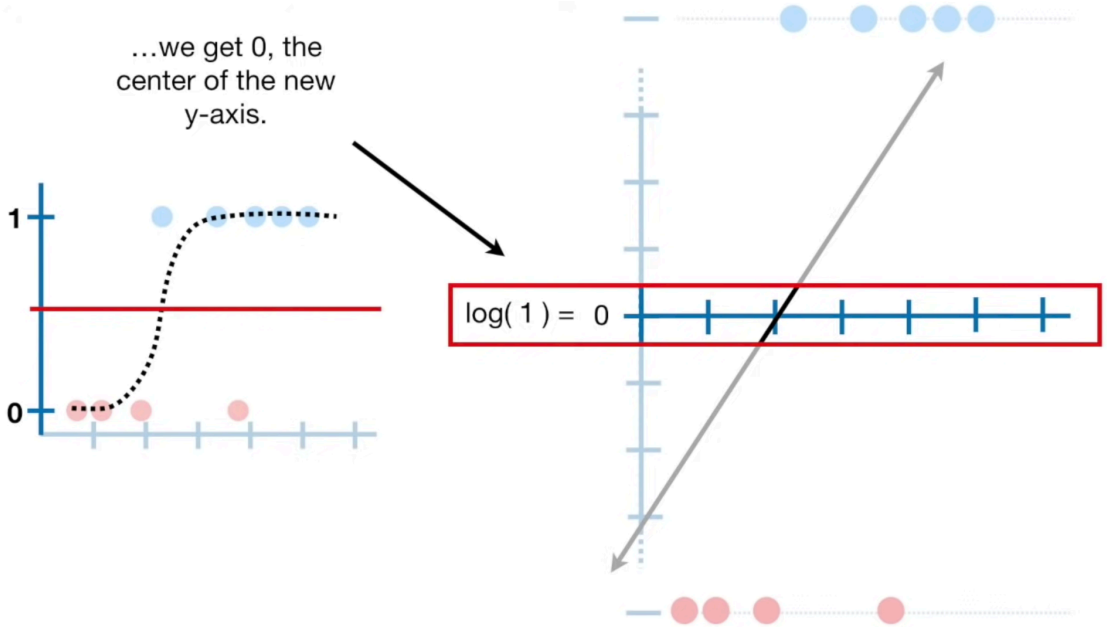
let's logit(p = 0.5)

remember that p is a probability.

we take 0.5 (i.e. 50%) probability and we transform it with the logistic function and we obtain…

Logistic Regression interpret

with continuous variable predictor

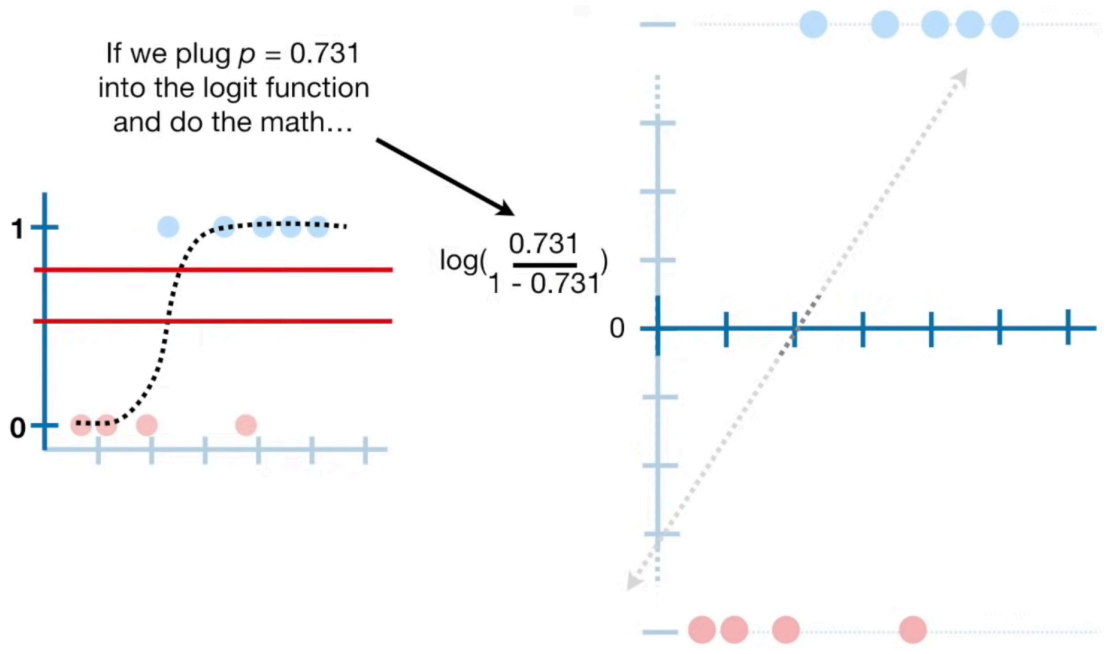


… on the new Y-axis

log(1) = **0**

Logistic Regression interpret

with continuous variable predictor

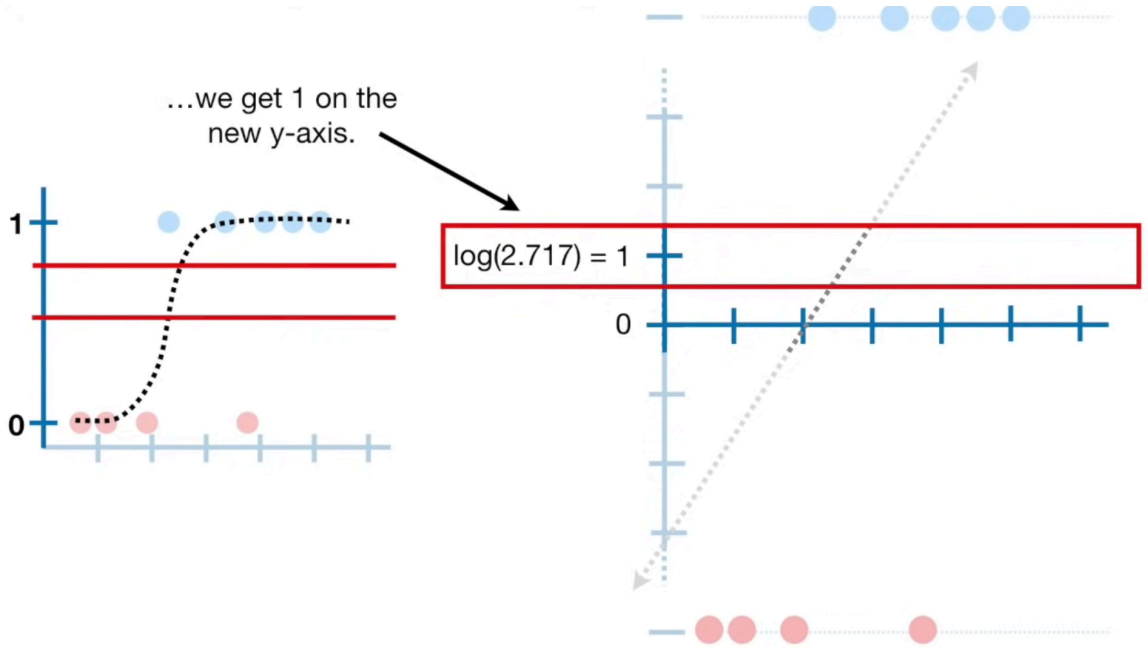


now let's logit(p = 0.731)

we plug the probability 0.731 into the logistic function and…

Logistic Regression interpret

with continuous variable predictor

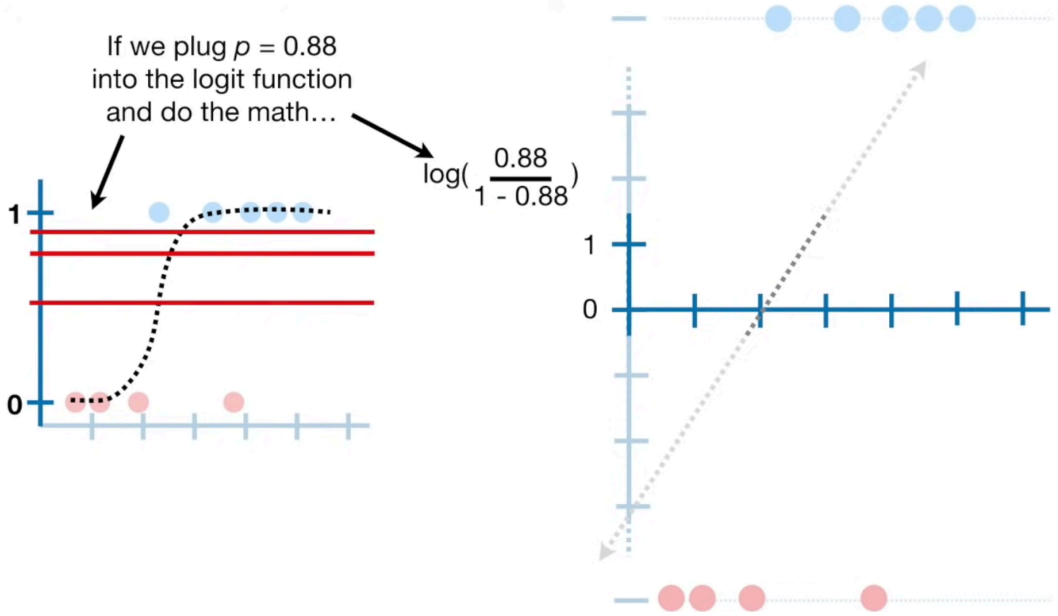


... on the new Y axis

and we obtain **1.**

Logistic Regression interpret

with continuous variable predictor

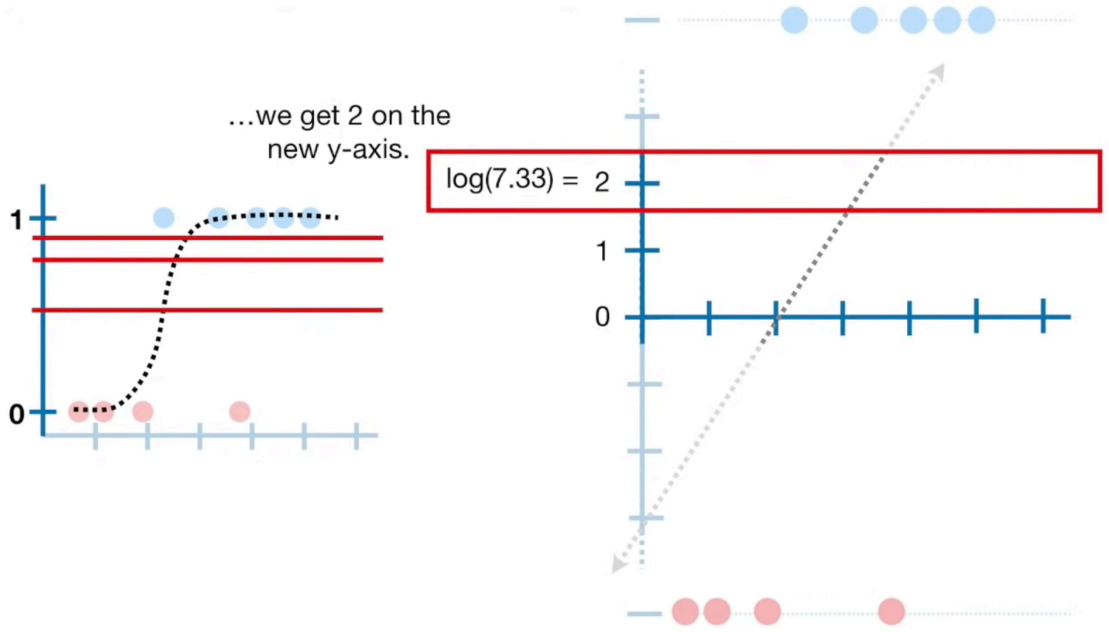


lets logit(p = 0.88)

we do the math and we obtain…

Logistic Regression interpret

with continuous variable predictor

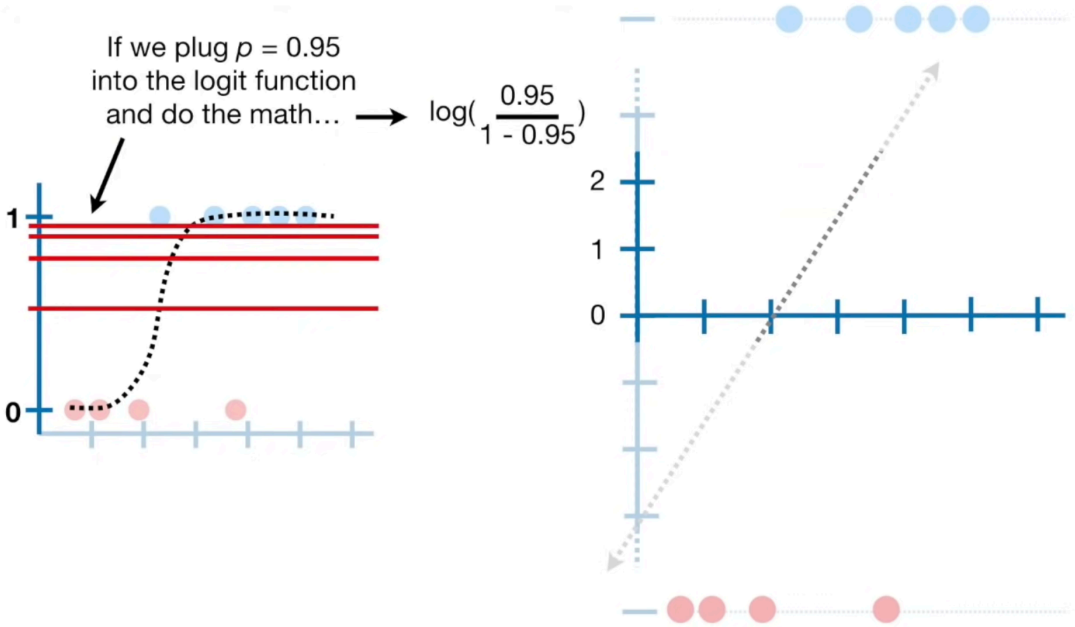


… on the Y axis

log(7.33) = **2**

Logistic Regression interpret

with continuous variable predictor

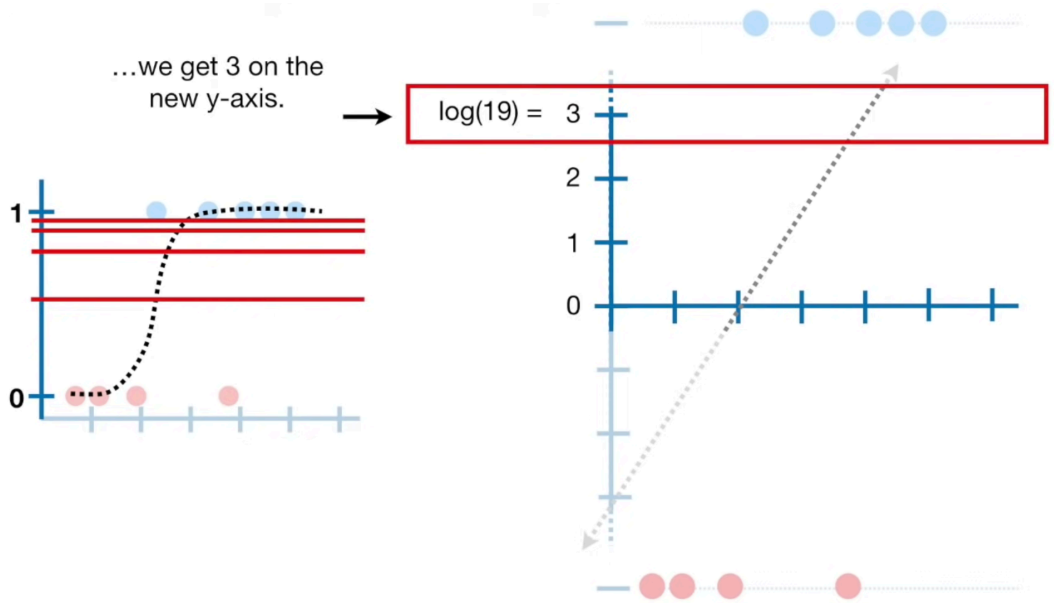


lets logit(p = 0.95)

we do the math and we obtain„

Logistic Regression interpret

with continuous variable predictor

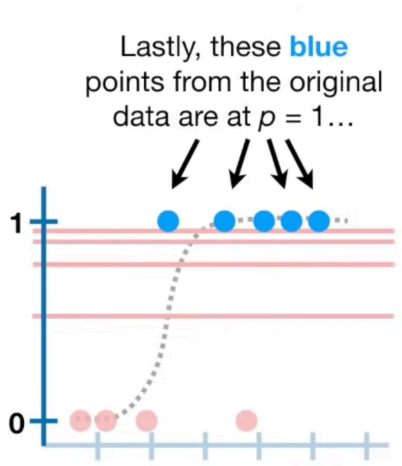


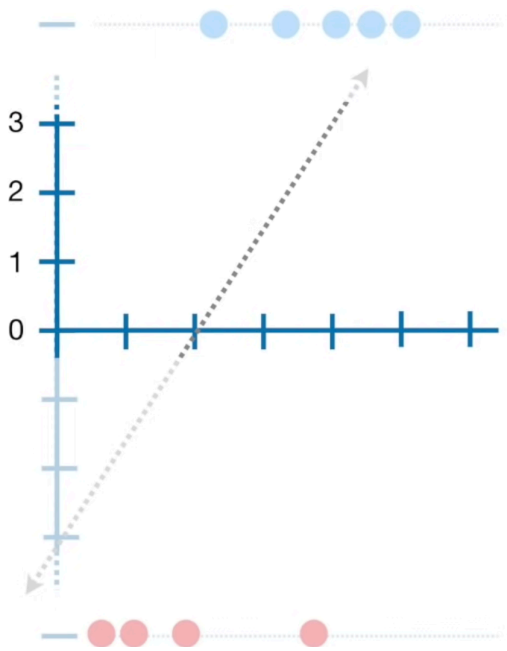
... on the new Y axis

log(19) = **3**

Logistic Regression interpret

with continuous variable predictor



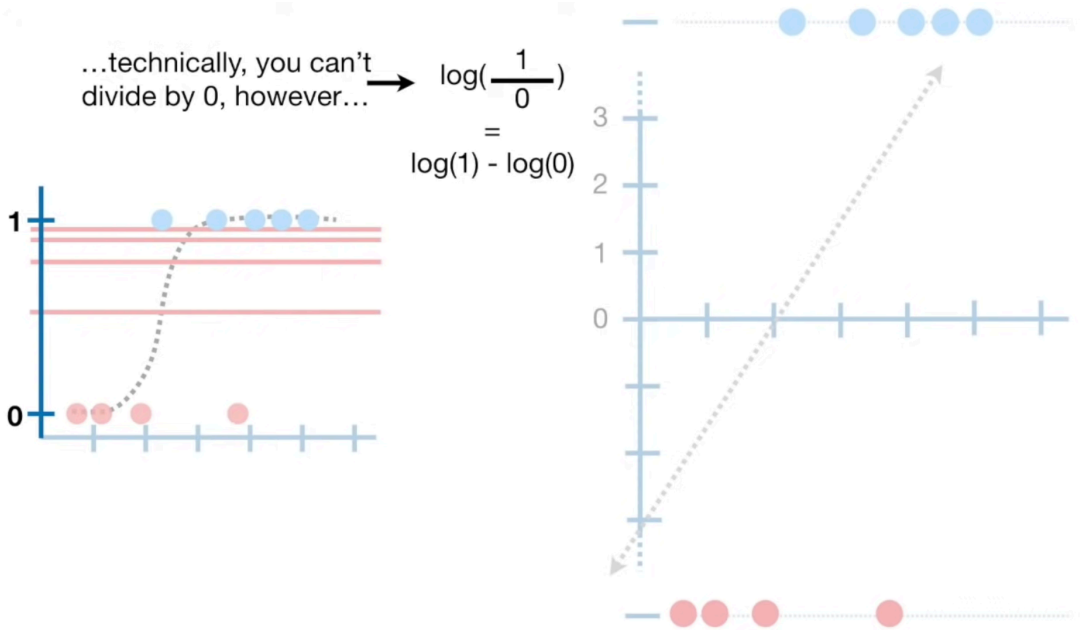


Now the blue dots

these are 1 i.e. 100% probability being obese. We plung them into the logistic function and…

Logistic Regression interpret

with continuous variable predictor

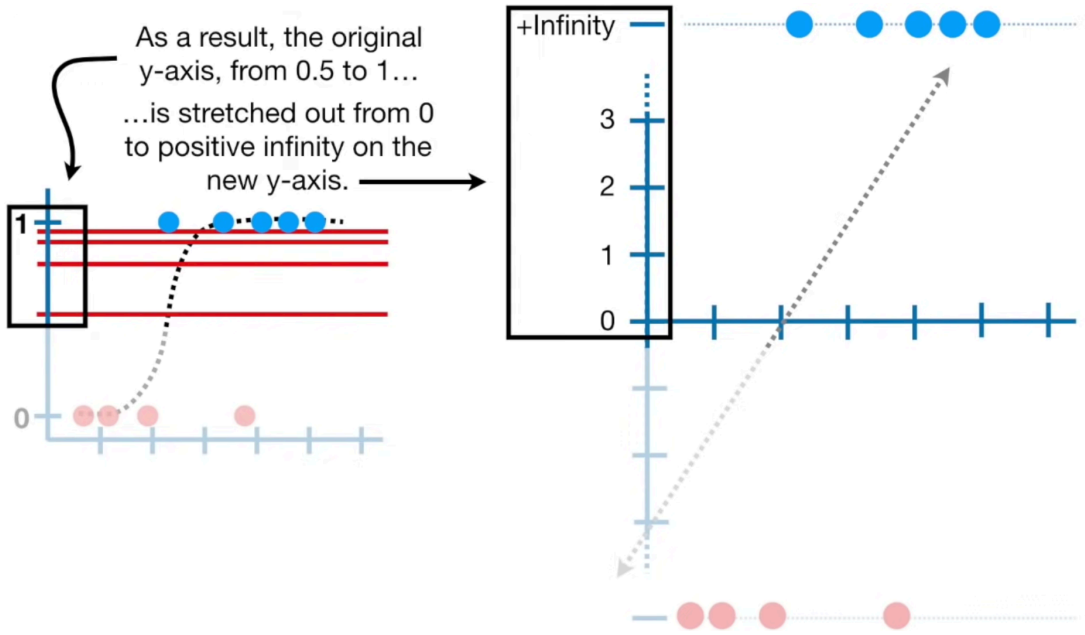


… on the new Y axis

log(1) - log(0) which by rudimental algebra you know is **+ infinity**.

Logistic Regression interpret

with continuous variable predictor

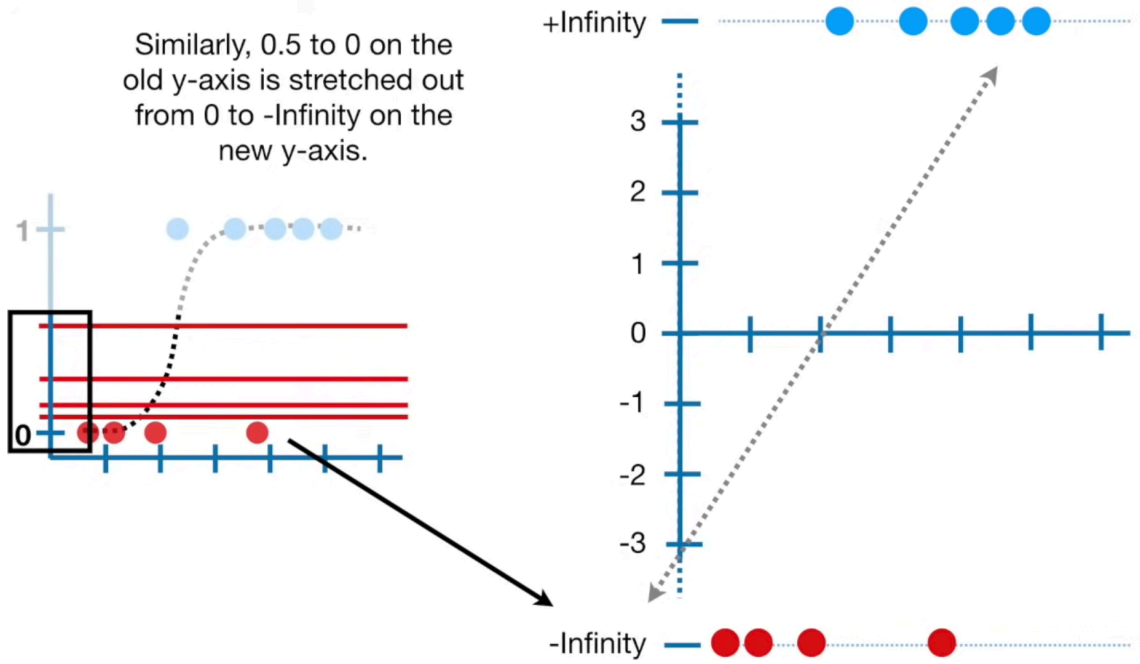


map from **old** to **new**

so now each result is mapped from old Y axis ranging **from 0 to 1**, to a new one which ranges **from -infinity to +infinity**

Logistic Regression interpret

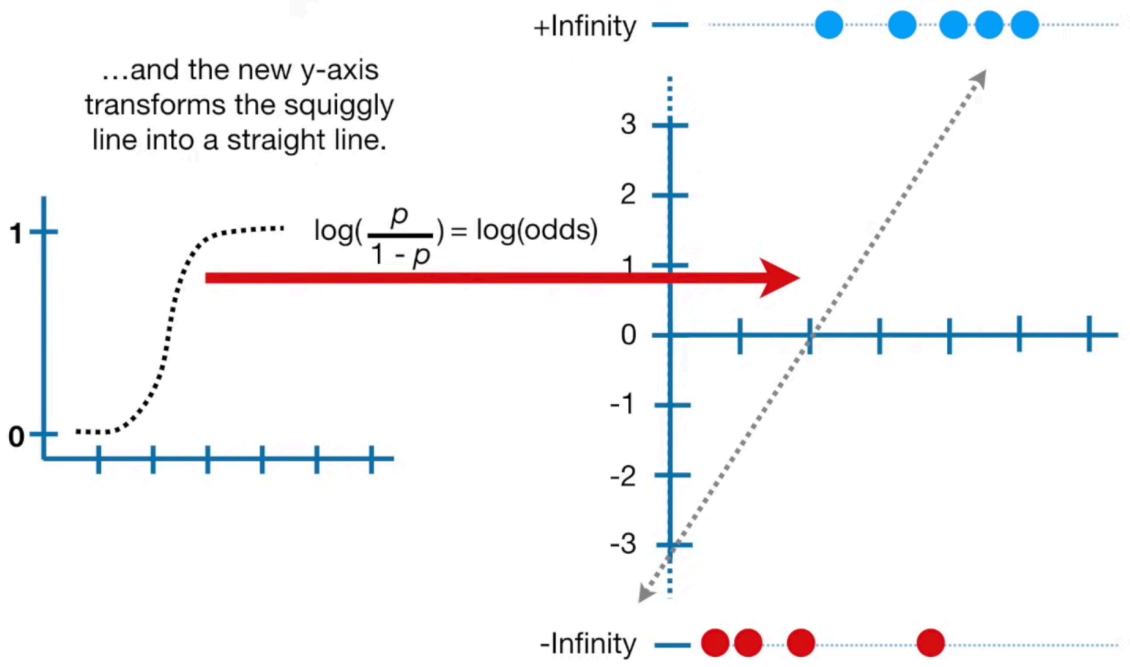
with continuous variable predictor



this is also true for negative values…

Logistic Regression interpret

with continuous variable predictor



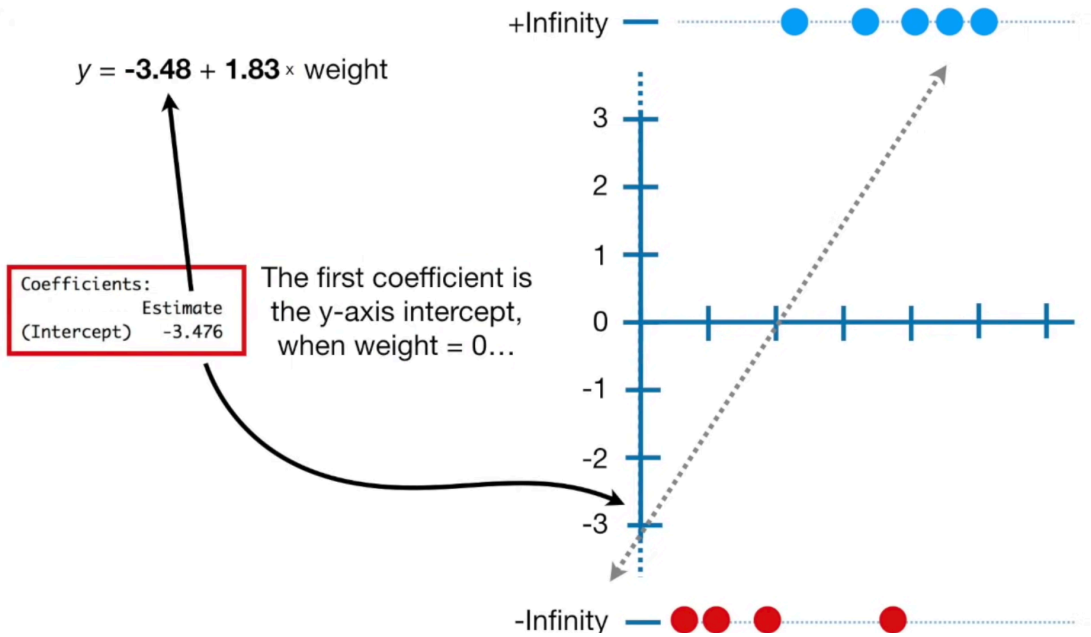
from S line to straight line

when you apply logistic function you pass from an S line in the old Y axis to a straight line to the new Y axis.

The important thing to know is even though the graph with a S line is what we associate with logistic regressio the coefficients are presented in terms of the odds graph.

Logistic Regression interpret

with continuous variable predictor



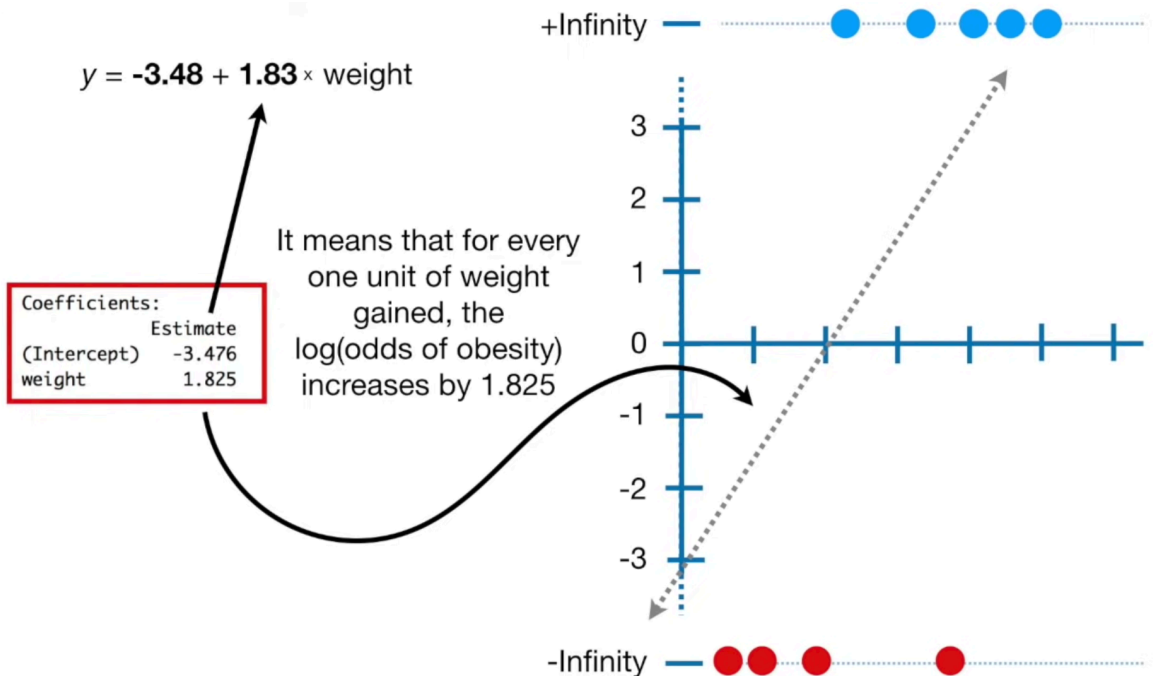
Intercept on the new Y axis

Now that we have a way to transform the S line to the straight line just as in linear regression we have an intercept and a slope.

When weight is equal 0 i.e. the intercept the line founds **−3.48 log odds. in other words if**

Logistic Regression interpret

with continuous variable predictor



slope on the new Y axis.

this means that the log of the odds of obesity increase of 1.825 when the weight increases to 1 Kg

Section 4

Live coding session!

JUMP TO RSTUDIO!

UNIVERSITÀ CATTOLICA del Sacro Cuore