Classification: Logistic Regression

Prof. (Dr.) Honey Sharma

Reference Book: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, A
Introduction to Statistical Learning with Applications in R

Classification

Variables can be characterized as either quantitative or qualitative (also known as categorical). We tend to refer to problems with a quantitative response as regression problems, while those involving a qualitative response are often referred to as classification problems.

Classification problems occur often, perhaps even more so than regression problems. Some examples include:

- A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?
- An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.
- On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

Why Not Linear Regression?

Suppose that we are trying to predict the medical condition of a patient in the emergency room on the basis of her symptoms. In this simplified example, there are three possible diagnoses: stroke, drug overdose, and epileptic seizure. We could consider encoding these values as a quantitative response variable, Y , as follows:

$$Y = \begin{cases} 1 & \text{if } \texttt{stroke}; \\ 2 & \text{if } \texttt{drug overdose}; \\ 3 & \text{if } \texttt{epileptic seizure}. \end{cases}$$

Using this coding, least squares could be used to fit a linear regression model to predict Y on the basis of a set of predictors X1,...,Xp. Unfortunately, this coding implies an ordering on the outcomes, putting drug overdose in between stroke and epileptic seizure, and insisting that the difference between stroke and drug overdose is the same as the difference between drug overdose and epileptic seizure. In practice there is no particular reason that this needs to be the case. For instance, one could choose an equally reasonable coding,

$$Y = \begin{cases} 1 & \text{if } \texttt{epileptic seizure}; \\ 2 & \text{if } \texttt{stroke}; \\ 3 & \text{if } \texttt{drug overdose}. \end{cases}$$

which would imply a totally different relationship among the three conditions. Each of these codings would produce fundamentally different linear models that would ultimately lead to different sets of predictions on test observations.

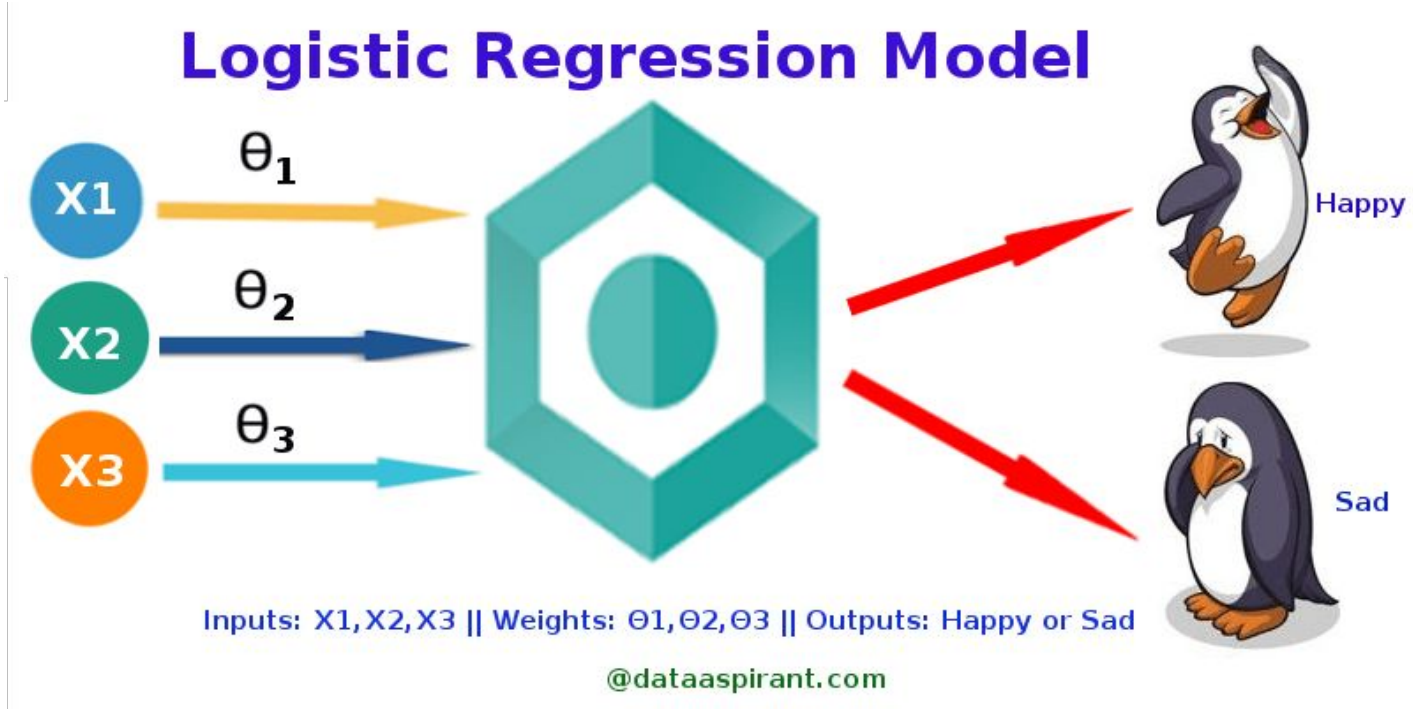
For a binary (two level) qualitative response, the situation is better. For binary instance, perhaps there are only two possibilities for the patient's medical condition: stroke and drug overdose. We could then potentially use the dummy variable approach to code the response as follows:

$$Y = \begin{cases} 0 & \text{if } \texttt{stroke}; \\ 1 & \text{if } \texttt{drug overdose}. \end{cases}$$

We could then fit a linear regression to this binary response, and predict drug overdose if $Y > 0.5$ and stroke otherwise. In the binary case it is not hard to show that even if we flip the above coding, linear regression will produce the same final predictions.

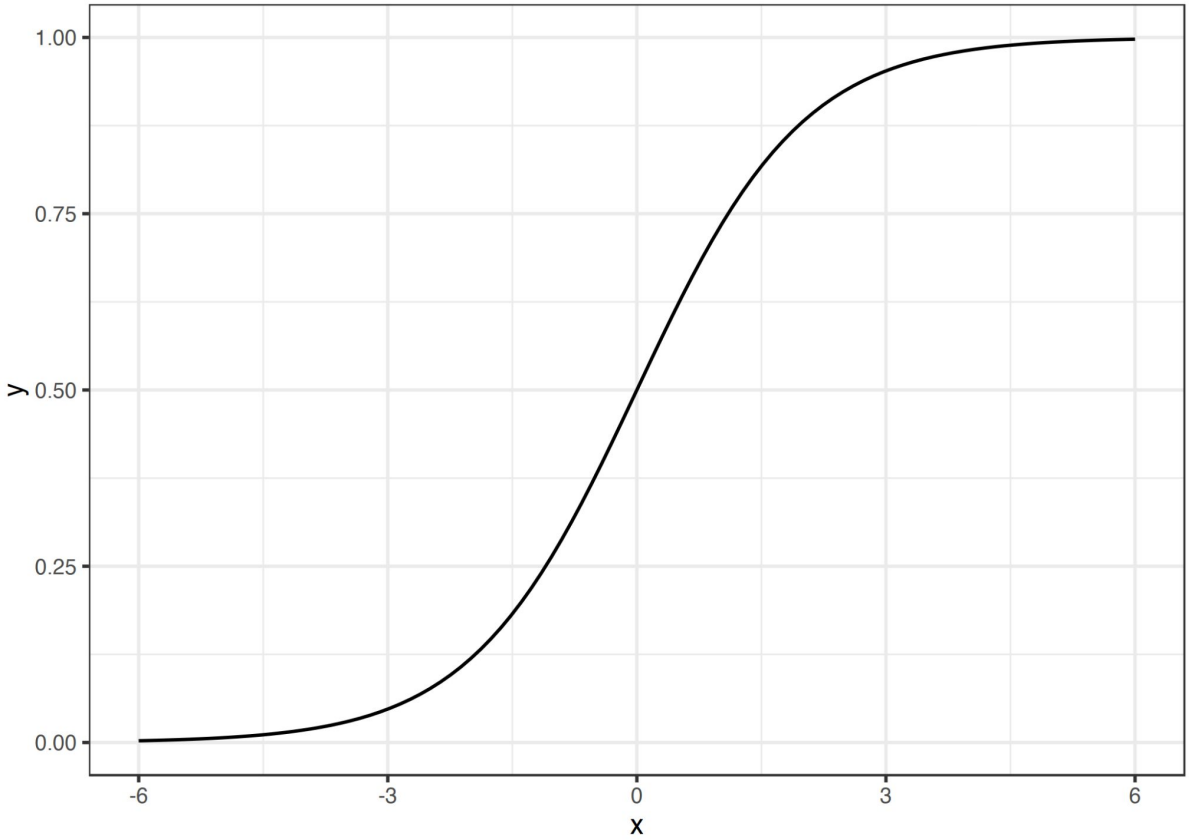
For a binary response with a 0/1 coding as above, regression by least squares does make sense. However, if we use linear regression, some of our estimates might be outside the [0, 1] interval, making them hard to interpret as probabilities!

Logistic Regression



A solution for classification is logistic regression. Instead of fitting a straight line or hyperplane, the logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1. The logistic function is defined as:

$$logistic(\eta) = \frac{1}{1+exp(-\eta)}$$


1.00
0.75
0.50
0.25
0.00
y
-6
-3
0
3
6
x

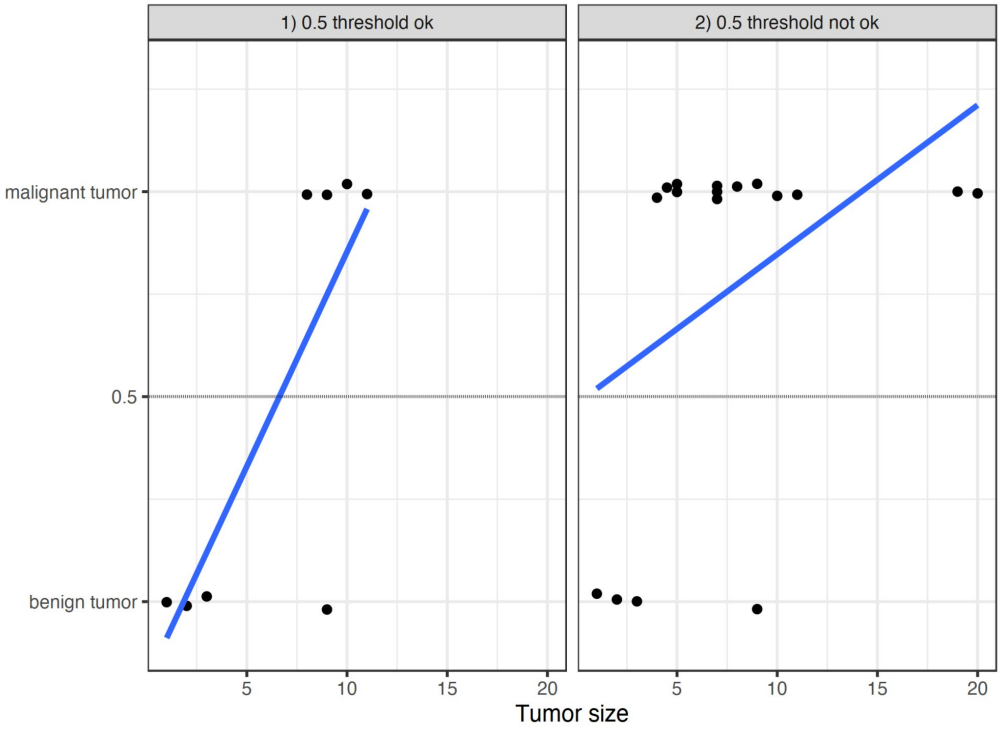
The step from linear regression to logistic regression is kind of straightforward. In the linear regression model, we have modelled the relationship between outcome and features with a linear equation:

$$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}$$

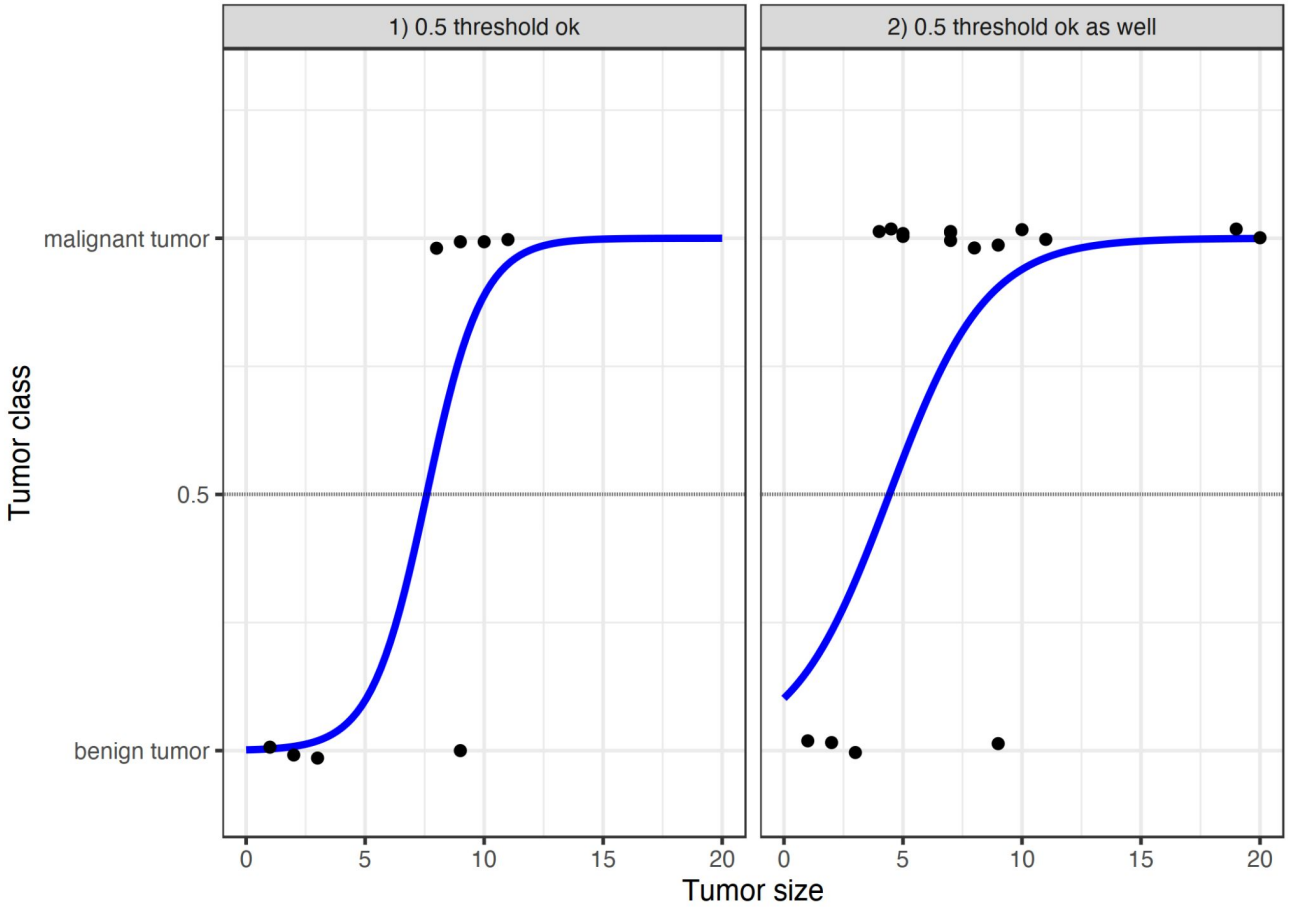
For classification, we prefer probabilities between 0 and 1, so we wrap the right side of the equation into the logistic function. This forces the output to assume only values between 0 and 1.

$$P(y^{(i)} = 1) = \frac{1}{1+exp(-(\beta_0+\beta_1 x_1^{(i)}+\ldots+\beta_p x_p^{(i)}))}$$

A linear model classifies tumors as malignant (1) or benign (0) given their size. The lines show the prediction of the linear model. For the data on the left, we can use 0.5 as classification threshold. After introducing a few more malignant tumor cases, the regression line shifts and a threshold of 0.5 no longer separates the classes. Points are slightly jittered to reduce over-plotting.



The logistic regression model finds the correct decision boundary between malignant and benign depending on tumor size. The line is the logistic function shifted and squeezed to fit the data



Interpretation

The interpretation of the weights in logistic regression differs from the interpretation of the weights in linear regression, since the outcome in logistic regression is a probability between 0 and 1. The weights do not influence the probability linearly any longer. The weighted sum is transformed by the logistic function to a probability.

Therefore we need to reformulate the equation for the interpretation so that only the linear term is on the right side of the formula.

$$\left(\frac{P(y=1)}{1-P(y=1)}\right) = log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

We call the term in the ln() function “odds” (probability of event divided by probability of no event) and wrapped in the logarithm it is called log odds.

This formula shows that the logistic regression model is a linear model for the log odds.

With a little shuffling of the terms, you can figure out how the prediction changes when one of the features xj is changed by 1 unit. To do this, we can first apply the exp() function to both sides of the equation:

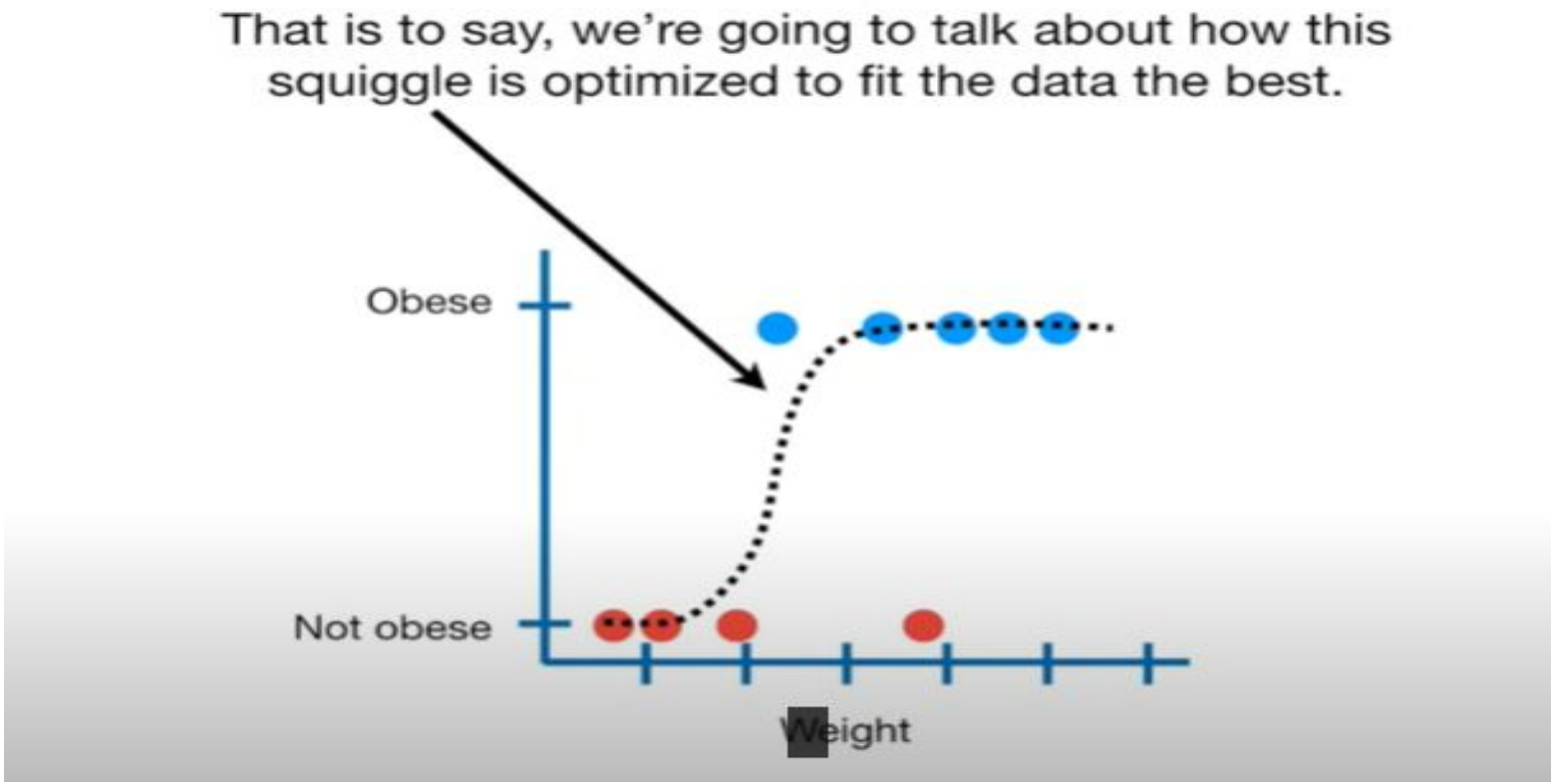
$$\frac{P(y=1)}{1-P(y=1)} = exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\right)$$

$$\frac{odds_{x_j+1}}{odds} = \frac{exp\left(\beta_0+\ldots+\beta_j(x_j+1)+\ldots+\beta_p x_p\right)}{exp\left(\beta_0+\ldots+\beta_j x_j+\ldots+\beta_p x_p\right)}$$

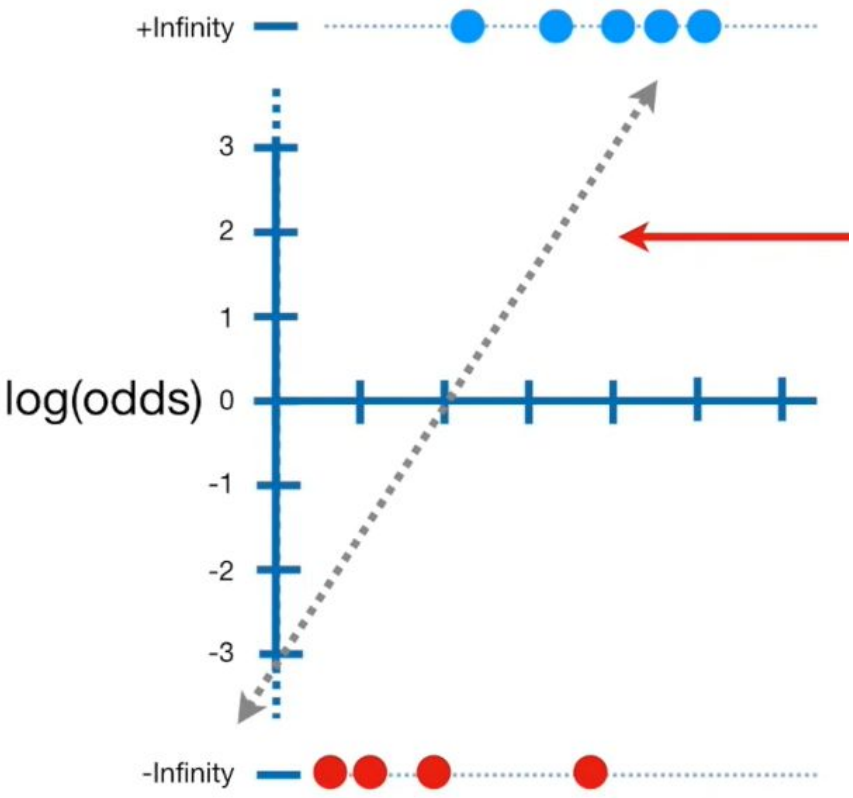
$$= exp\left(\beta_j(x_j+1) - \beta_j x_j\right) = exp\left(\beta_j\right)$$

We could also interpret it this way: A change in xj by one unit increases the log odds ratio by the value of the corresponding weight.

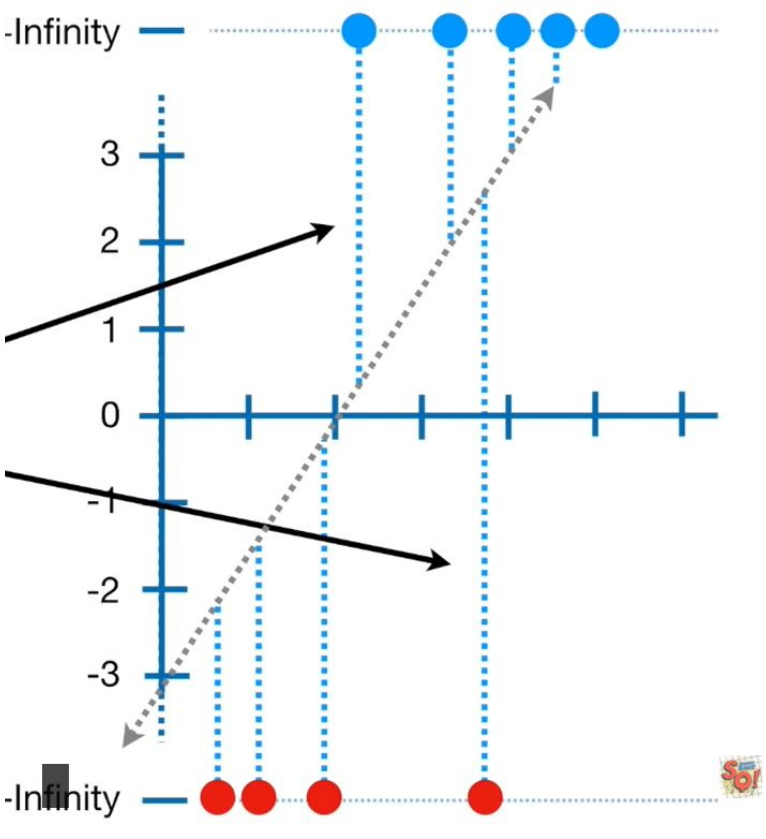
Maximum Likelihood Estimation



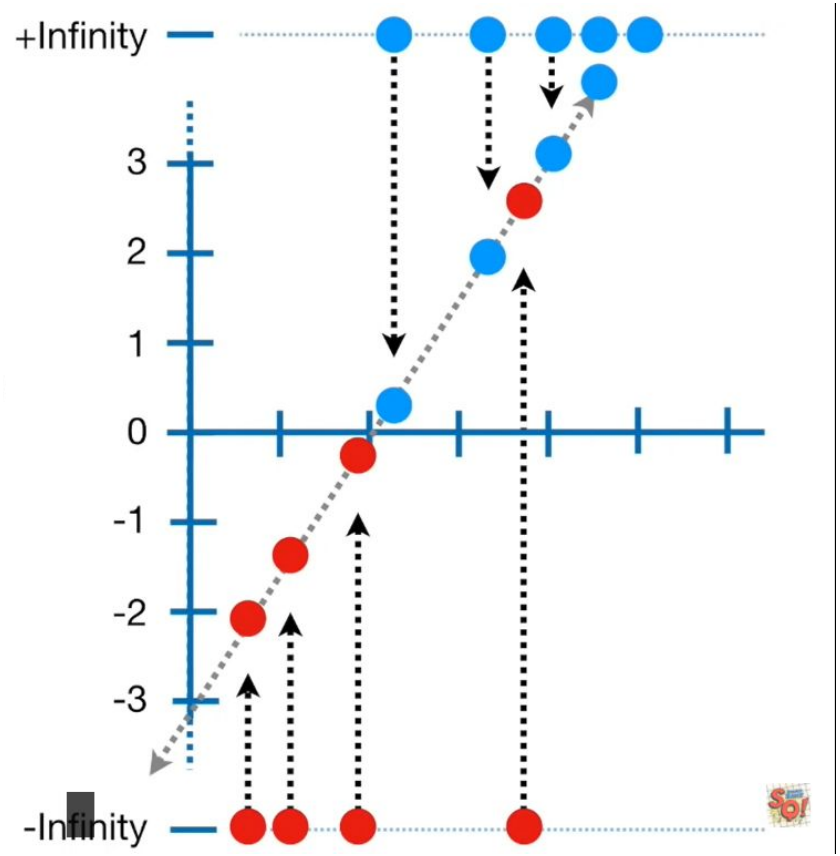
- Coefficient of Logistic Regression is very similar to Linear Regression Model.
- The only difference is that in Logistic Regression we use log(odds) on y axises

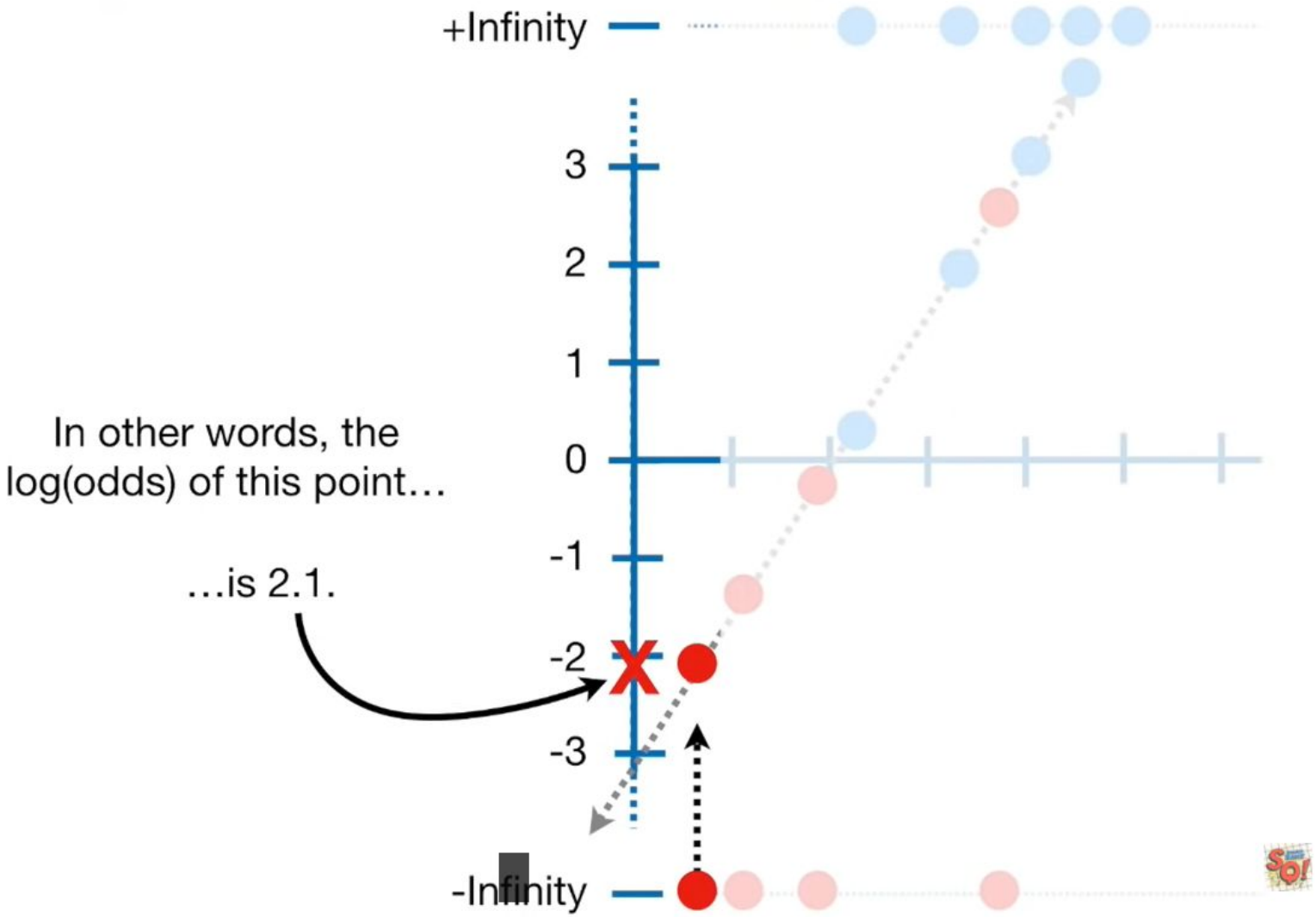


- In Linear Regression best line is obtained using the least of the sum of Squares of Residues.
- In case of Logistic Regression, the only problem is that the transformation pushes the data to positive and negative infinity.
- This means residuals (distance of the points from line) are positive and negative infinity.
- This means we cannot use the least squares to find the best fit.



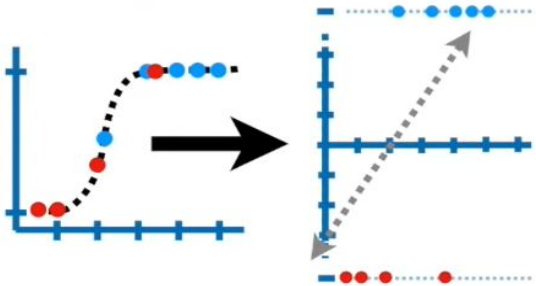
- We use the Maximum Likelihood.
- Firstly we project the original data points on the line.
- This gives each sample a log(odd) values



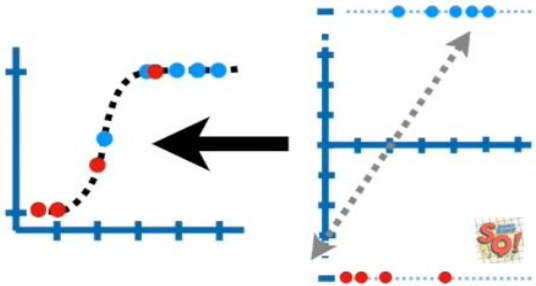

+Infinity
3
2
1
0
-1
-2
-3
-Infinity
In other words, the
log(odds) of this point…
…is 2.1.

- Then we transform log(odds) to probability

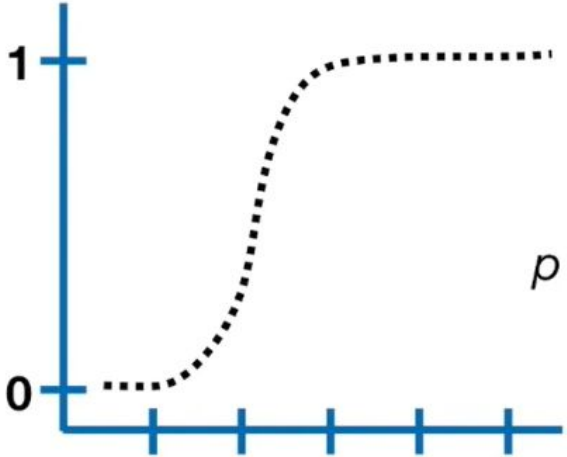
$$\log\left(\frac{p}{1-p}\right) = \log(\text{odds})$$



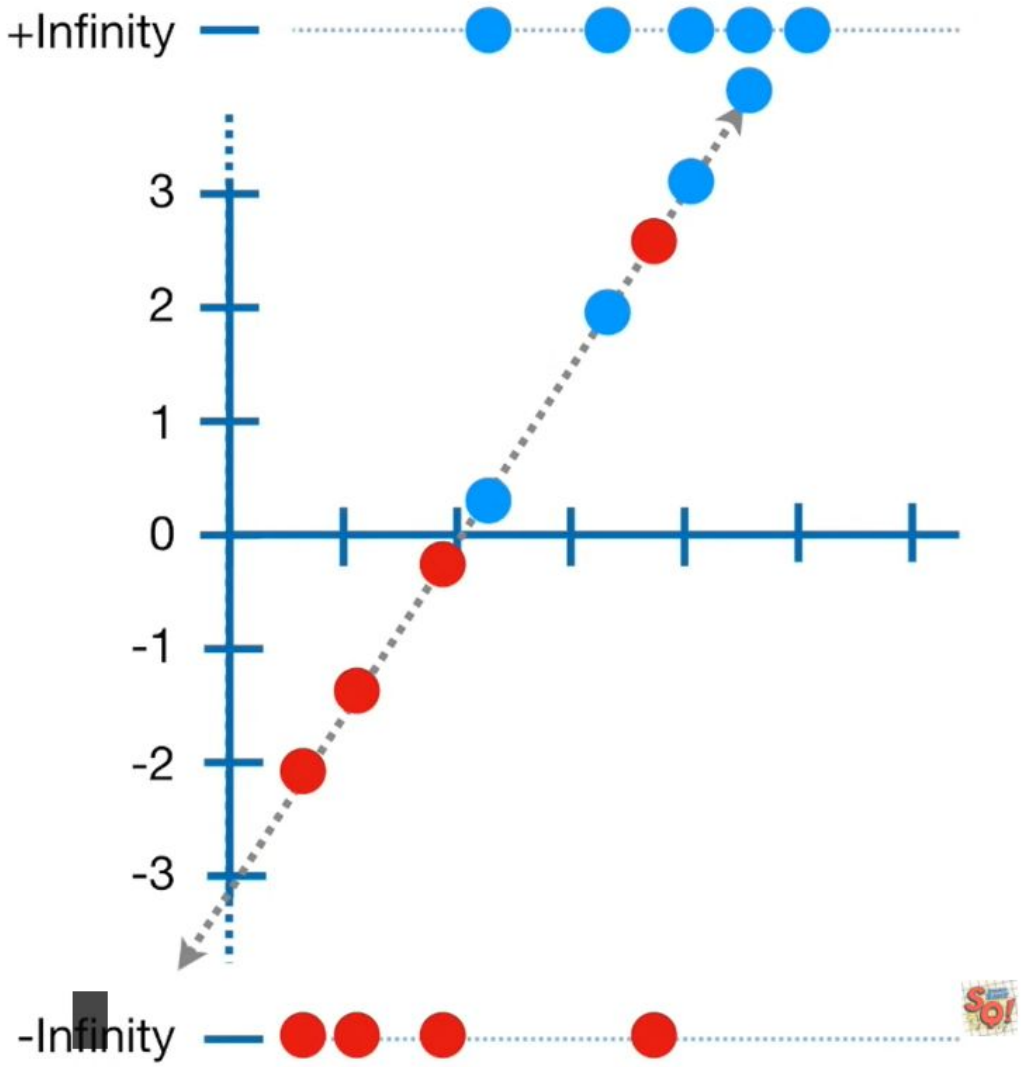
$$p = \frac{e^{\log(\text{odds})}}{1 + e^{\log(\text{odds})}}$$

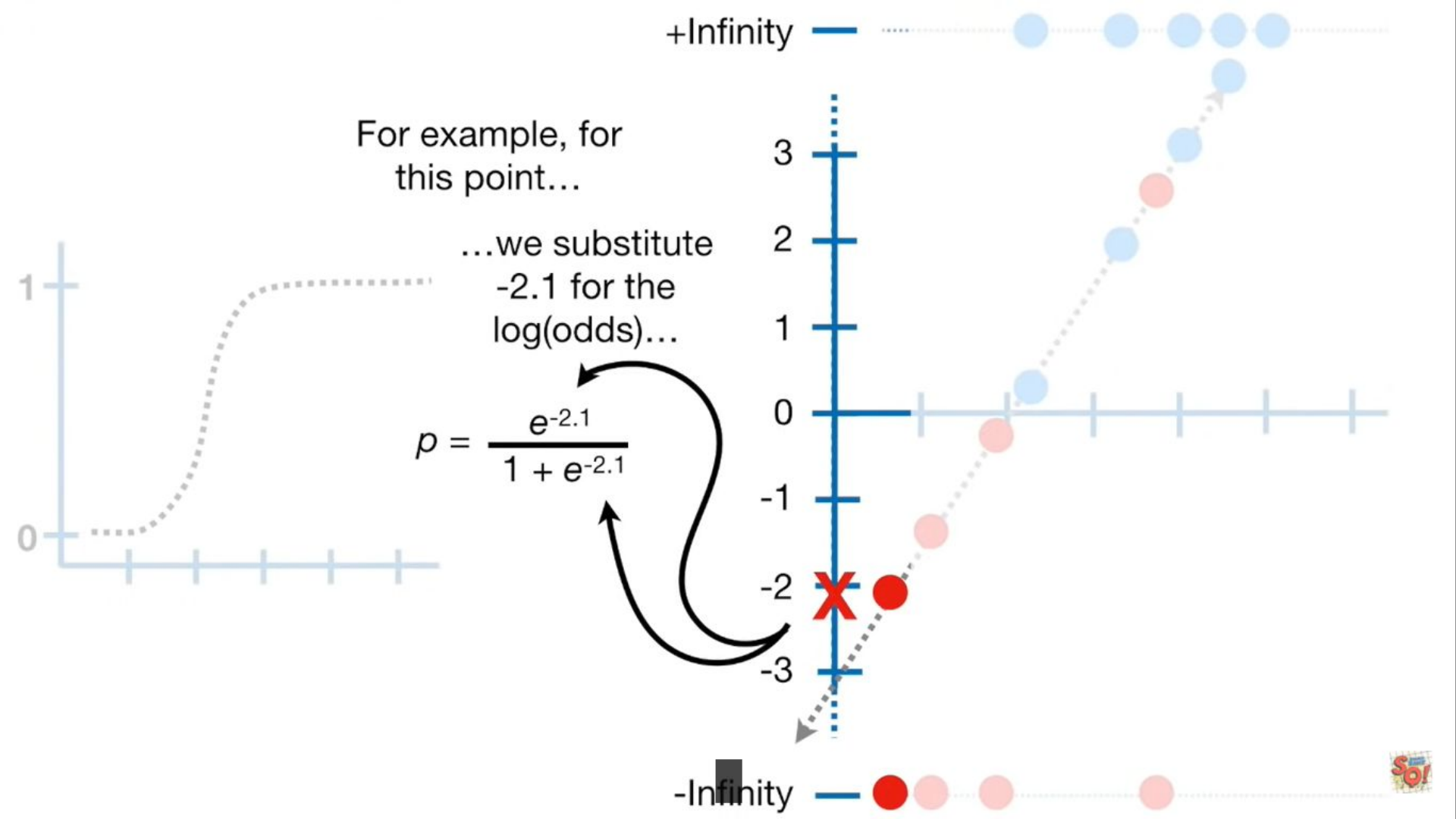


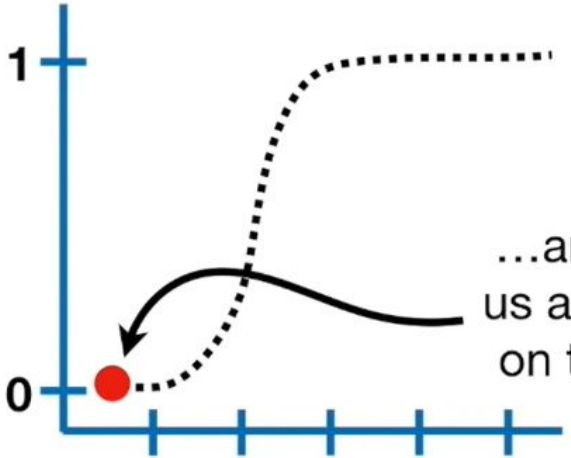
Now let's see this fancy equation in action!!!!



$$p = \frac{e^{\log(odds)}}{1 + e^{\log(odds)}}$$



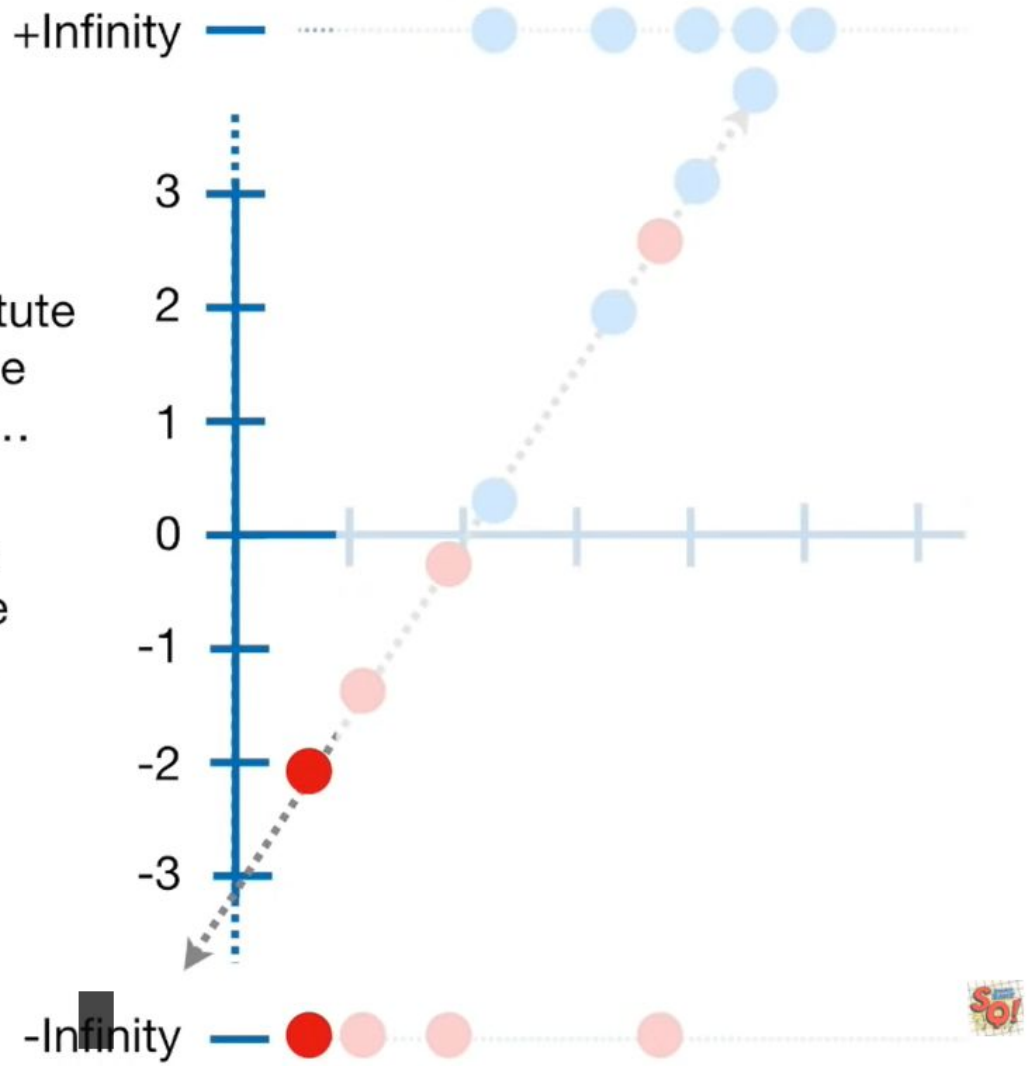

+Infinity
For example, for this point…
…we substitute -2.1 for the log(odds)…
$p = \frac{e^{-2.1}}{1 + e^{-2.1}}$
3
2
1
0
-1
-2
-3
1
0
-Infinity



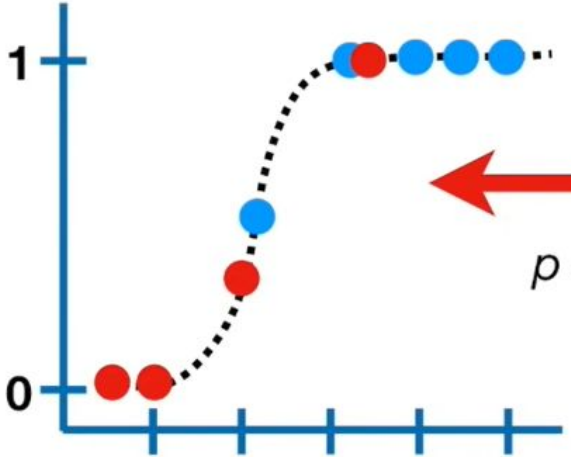
For example, for this point…

…we substitute -2.1 for the log(odds)…

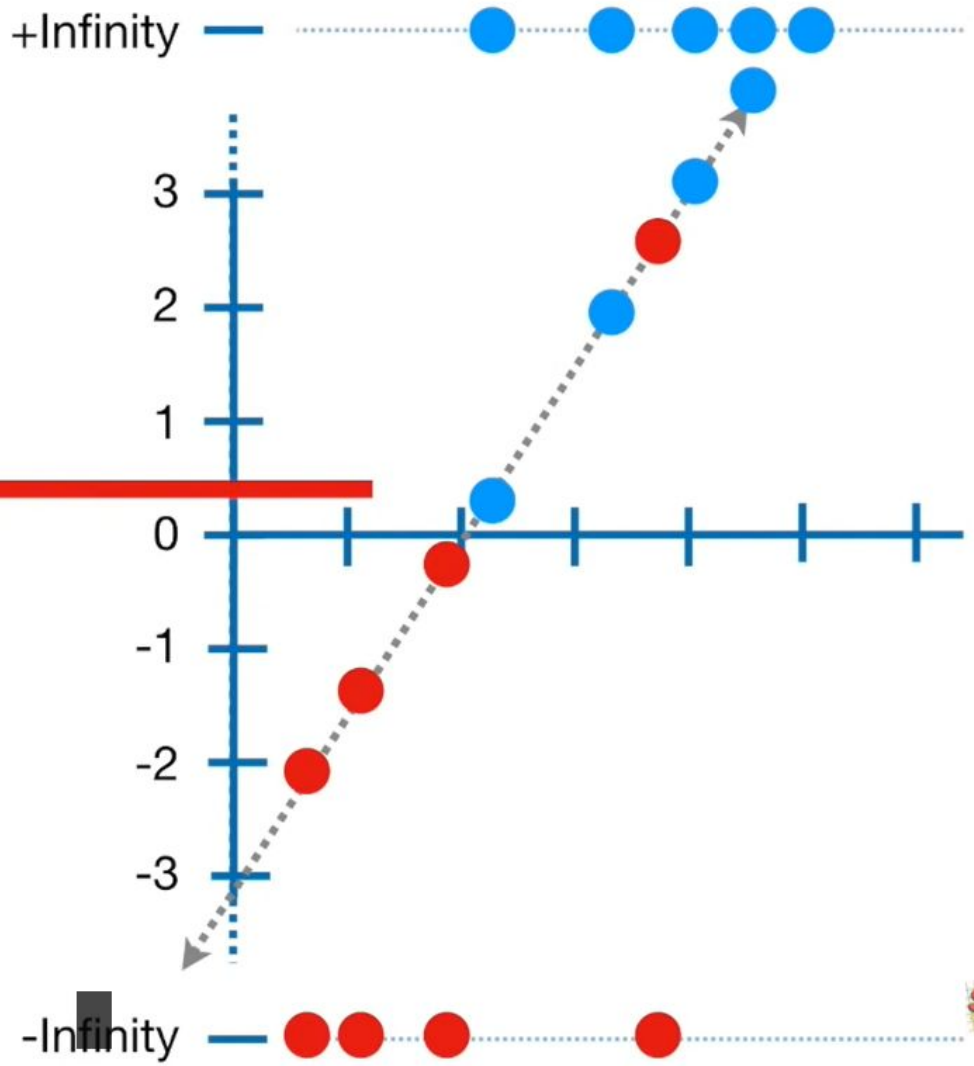
…and that gives us a y-coordinate on the squiggle.



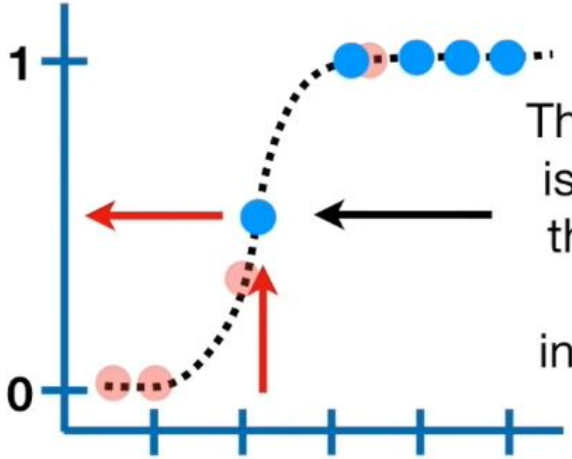
…and we do the same thing for all of the points.



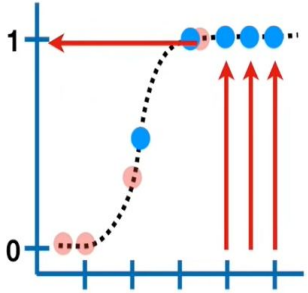
$$p = \frac{e^{\log(\text{odds})}}{1 + e^{\log(\text{odds})}}$$



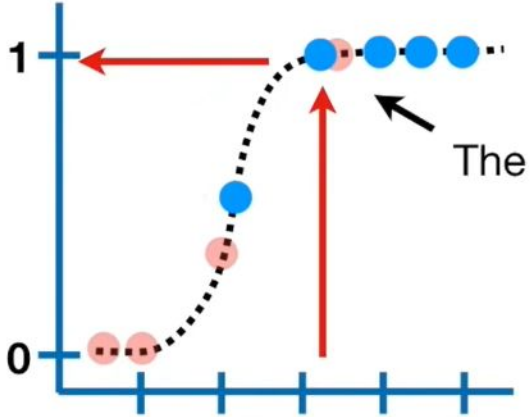
- Calculate likelihood of obese mice
- In the curve probability is not calculated as area under the curve, but instead it is y-axis value.



The likelihood that this mouse is **obese**, given the shape of the squiggle, is the value on the y-axis where point intersects the squiggle, **0.49**.

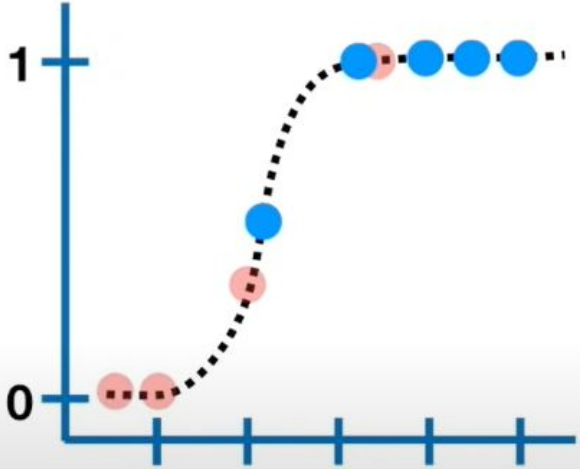


The likelihoods that these mice are **obese** are **0.91**, **0.91** and **0.92**

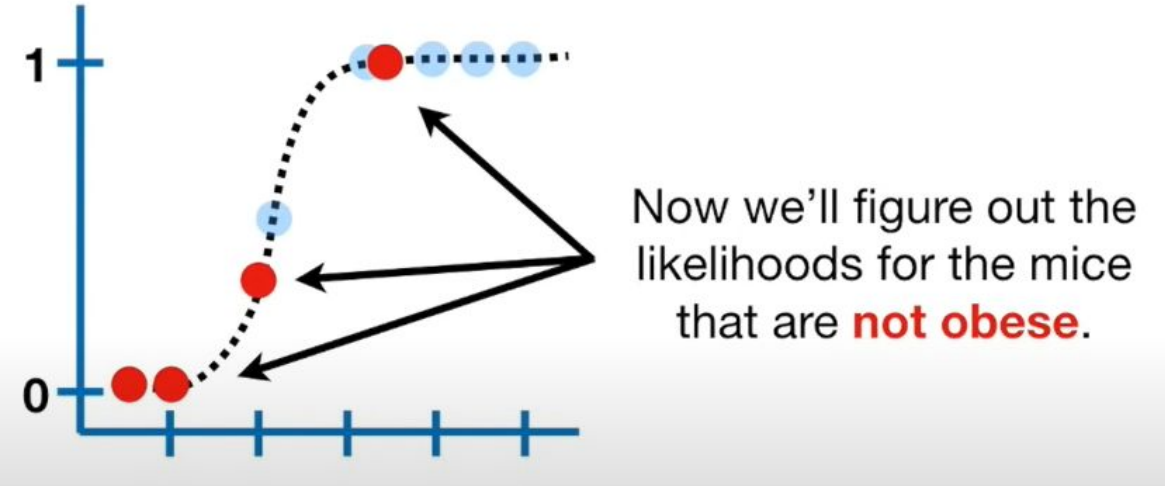


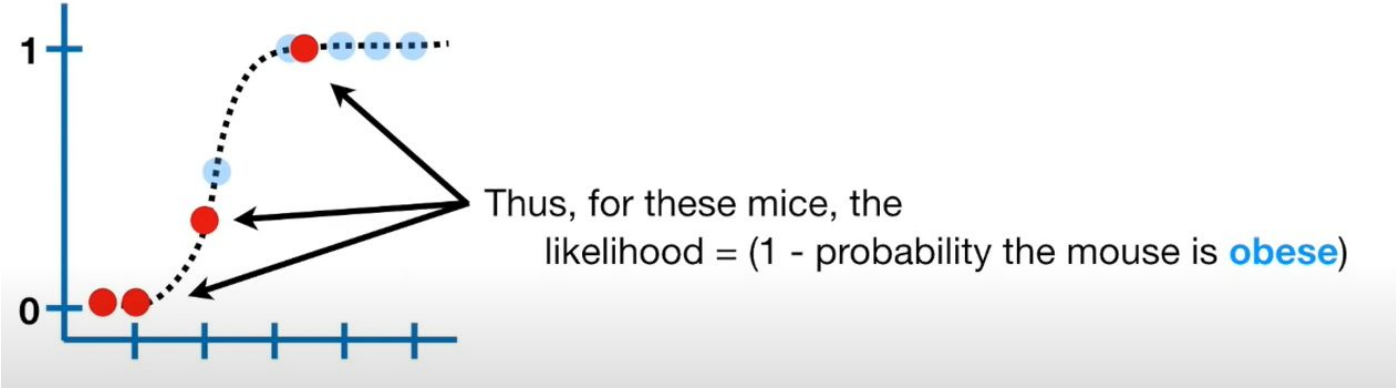
The likelihood that this mouse is **obese** is **0.9**

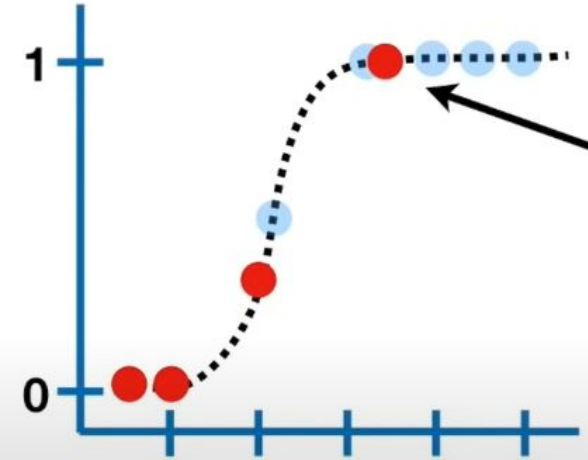
likelihood of data given the squiggle = **0.49** × **0.9** × **0.91** × **0.91** × **0.92** × …



The likelihood for all of the **obese** mice is just the product of the individual likelihoods.

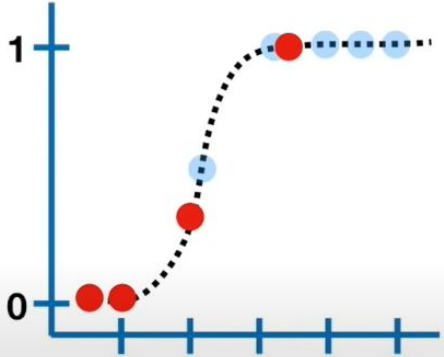

1
0
Now we'll figure out the likelihoods for the mice that are **not obese**.


1
0
Thus, for these mice, the likelihood = (1 - probability the mouse is **obese**)

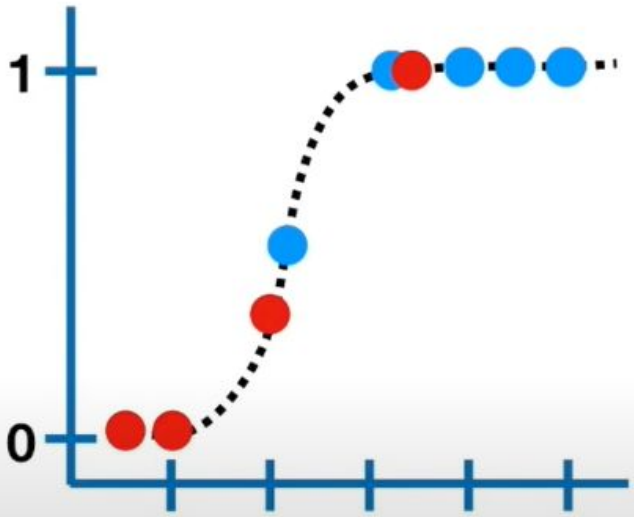


The probability that this mouse is **obese** is 0.9, so the probability and likelihood that it is **not obese** is (**1 - 0.9**)

likelihood of data given the squiggle = **0.49** × **0.9** × **0.91** × **0.91** × **0.92** ×
(**1 - 0.9**) × (**1 - 0.3**) × (**1 - 0.01**) × (**1 - 0.01**)



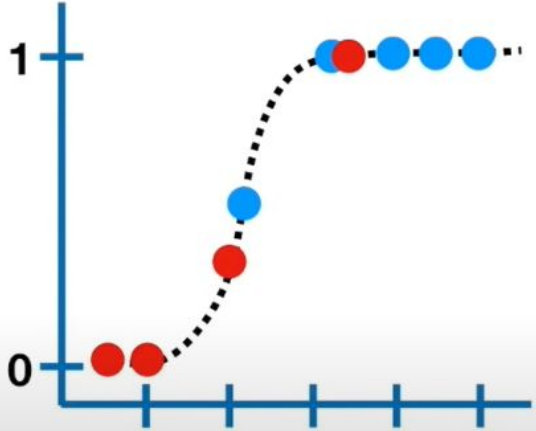
Now we can include the individual likelihoods for the mice that are **not obese** to the equation for the overall likelihood.



NOTE: Although it is possible to calculate the likelihood as the product of the individual likelihoods, statisticians prefer to calculate the **log of the likelihood** instead.

Either way works because the squiggle that maximizes the likelihood is the same one that maximizes the log of the likelihood.

likelihood of data given the squiggle = 0.49 × 0.9 × 0.91 × 0.91 × 0.92 ×
(1 - 0.9) × (1 - 0.3) × (1 - 0.01) × (1 - 0.01)

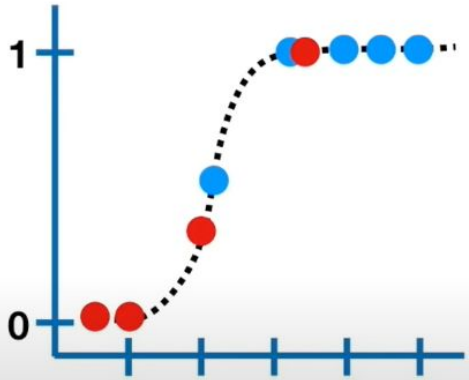


NOTE: Although it is possible to calculate the likelihood as the product of the individual likelihoods, statisticians prefer to calculate the **log of the likelihood** instead.

Either way works because the squiggle that maximizes the likelihood is the same one that maximizes the log of the likelihood.

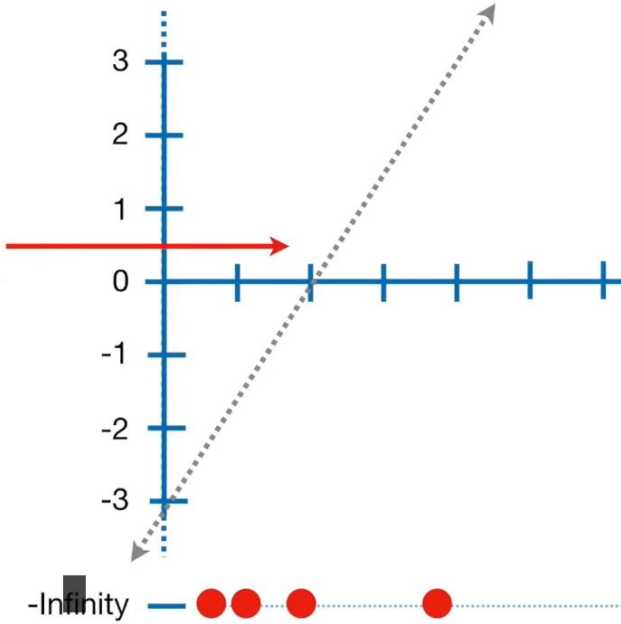
Logistic Regression Details Pt 2: Maximum Likelihood

log(likelihood of data given the squiggle) = -3.77

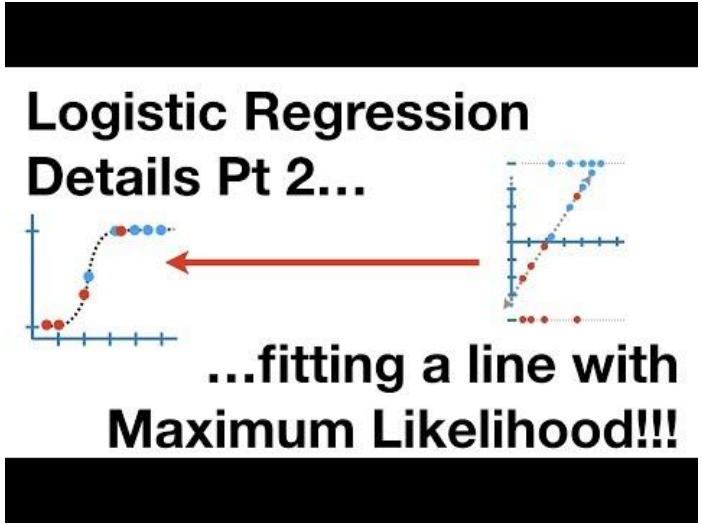


Thus, the log-likelihood of the data given the squiggle is -3.77…

…and this means that the og-likelihood of the original line is -3.77.

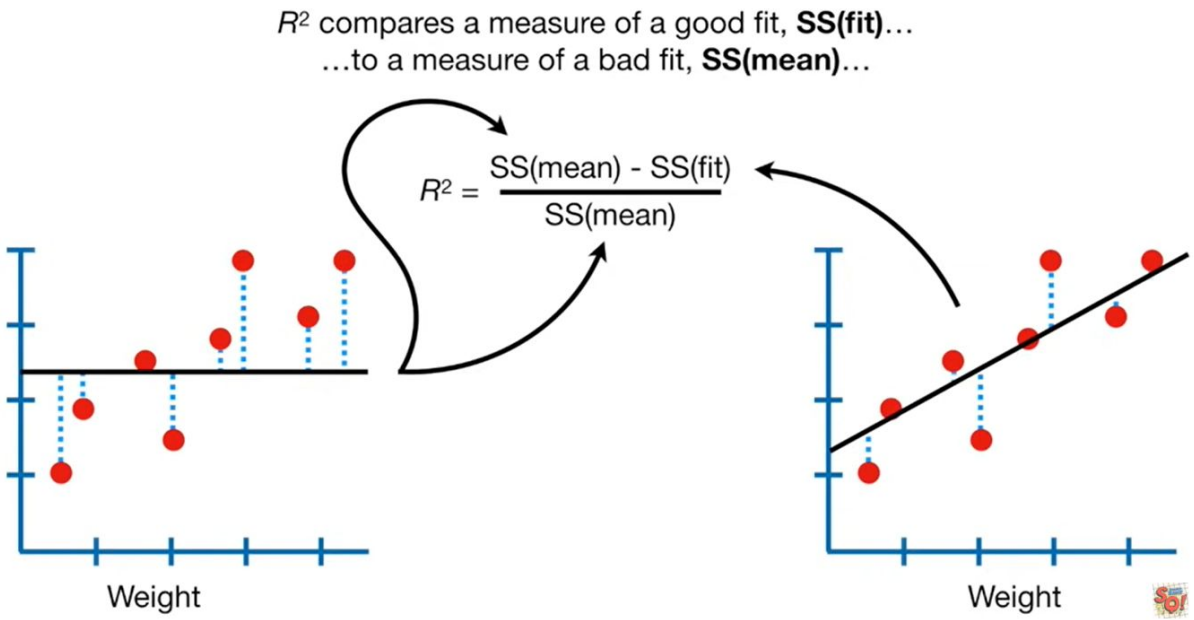


- Rotate the line, calculate log likelihood by projecting the data onto it and transforming it into probabilities.
- The algorithm that finds the line with maximum likelihood is pretty smart- each times it rotate the line, it does so in a way that increase the log likelihood. This algorithm can find a optimal fit after a few rotations.



R-Squared & p-values

In Linear Regression R-Squared is calculated as



- In case of Logistic Regression, R-squared is called McFadden's Pseudo R-Square.
- Unfortunately the residues in log(odds) graph are at infinity

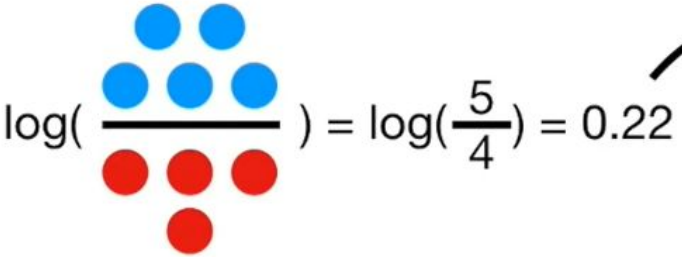
Now we need a measure of a poorly fitted line that is analogous to SS(mean)…

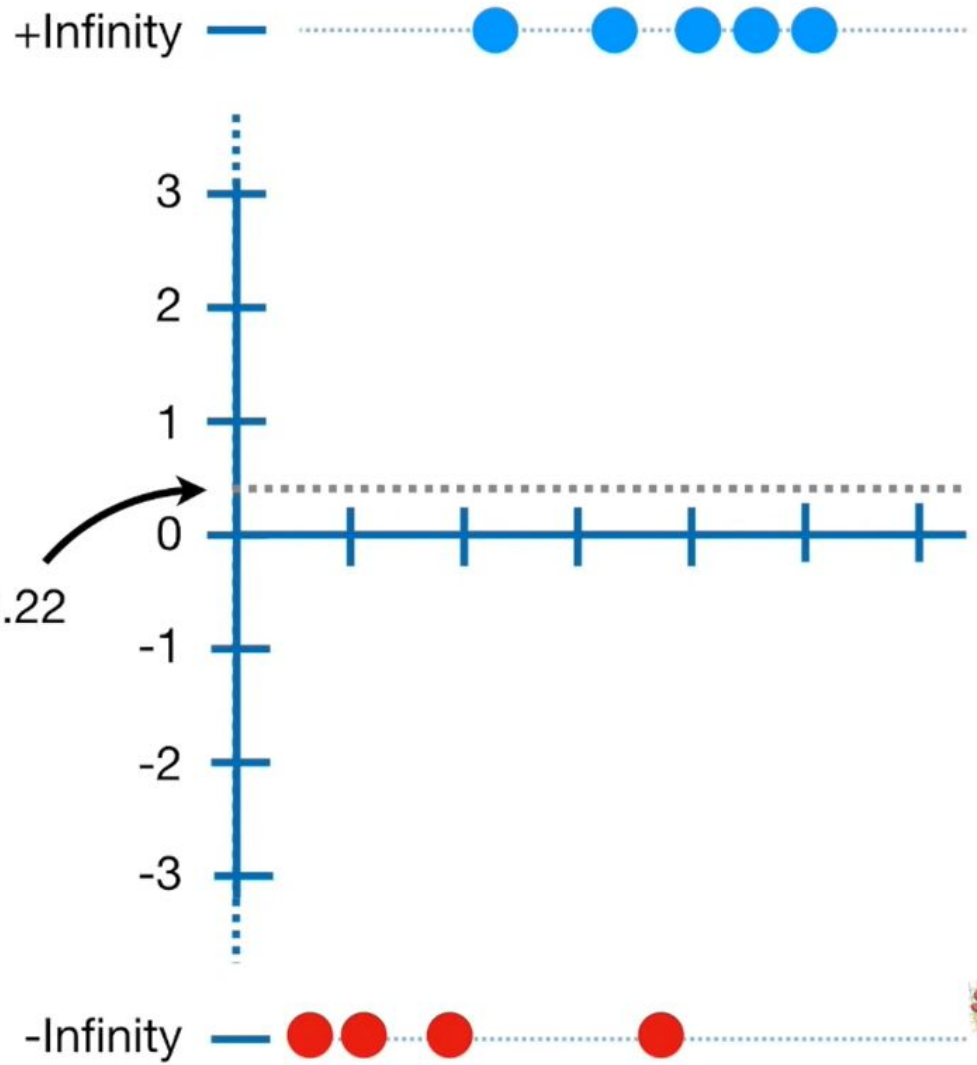
$$R^2 = \frac{SS(mean) - SS(fit)}{SS(mean)}$$

Vs

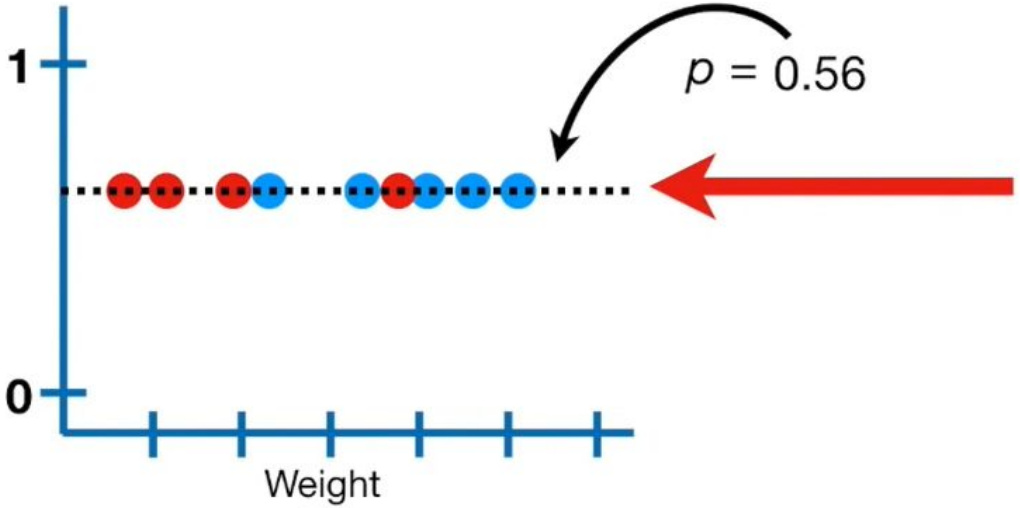
$$R^2 = \frac{??? - LL(fit)}{???}$$

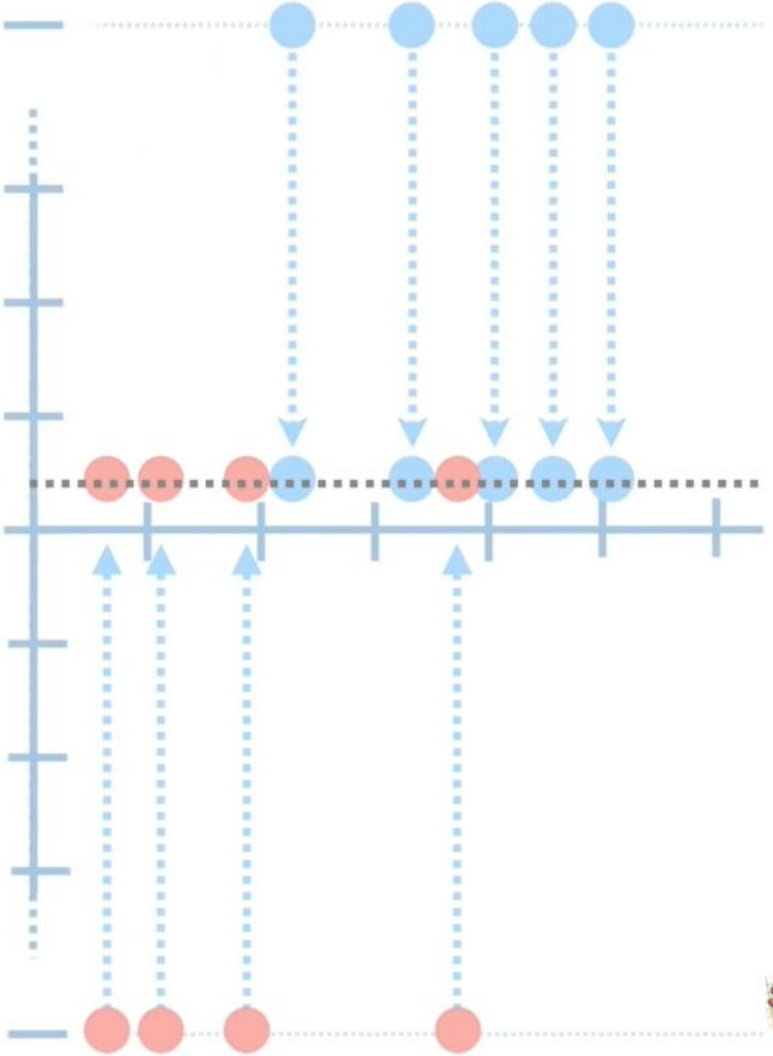
We do this by calculating the log(odds of obesity) without taking weight into account.



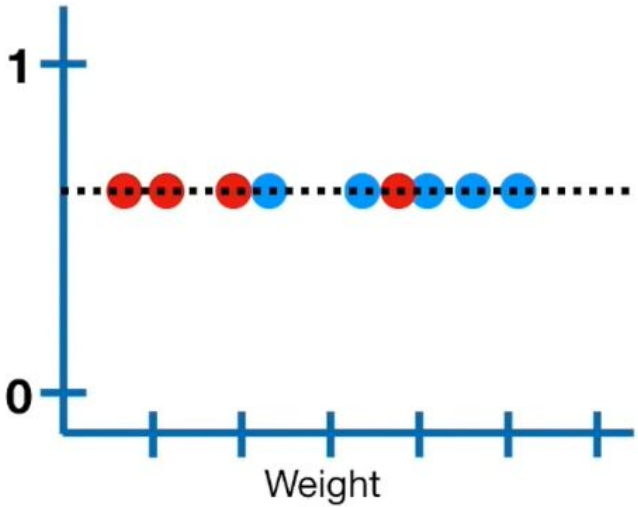


…and then we translate the log(odds) back to probabilities.





In other words, we can arrive at the same solution by calculating the overall probability of obesity.

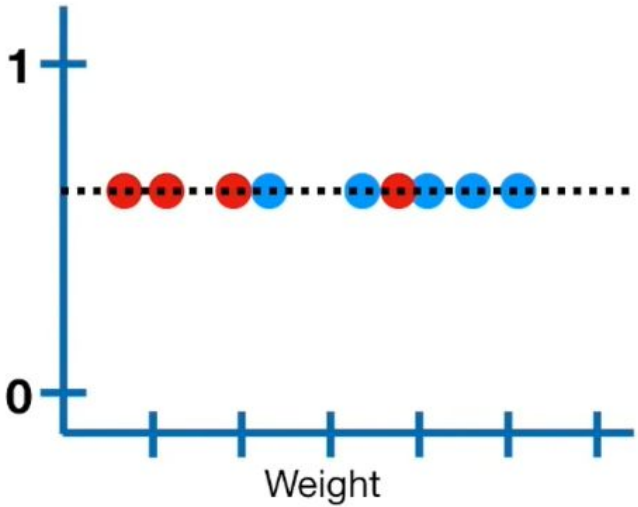


$$p = \frac{\text{number of obese mice}}{\text{total number of mice}} = \frac{5}{9} = \boxed{0.56}$$

$$p = \frac{e^{0.22}}{1 + e^{0.22}} = \boxed{0.56}$$

Hooray!
They are the same!

In other words, we can arrive at the same solution by calculating the overall probability of obesity.

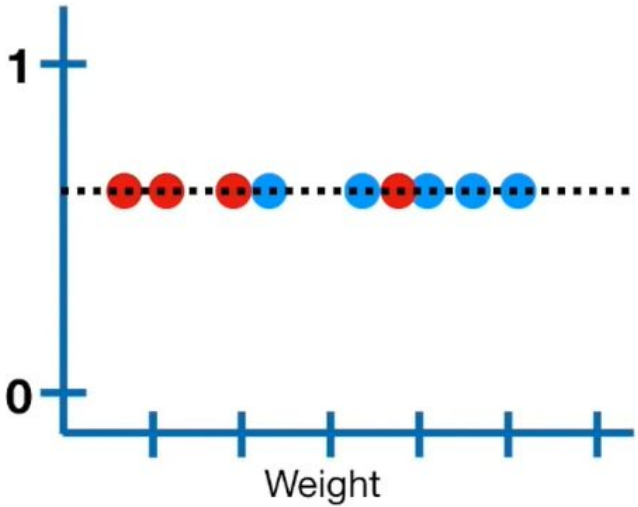


$$p = \frac{\text{number of obese mice}}{\text{total number of mice}} = \frac{5}{9} = \boxed{0.56}$$

$$p = \frac{e^{0.22}}{1 + e^{0.22}} = \boxed{0.56}$$

Hooray!
They are the same!

In other words, we can arrive at the same solution by calculating the overall probability of obesity.

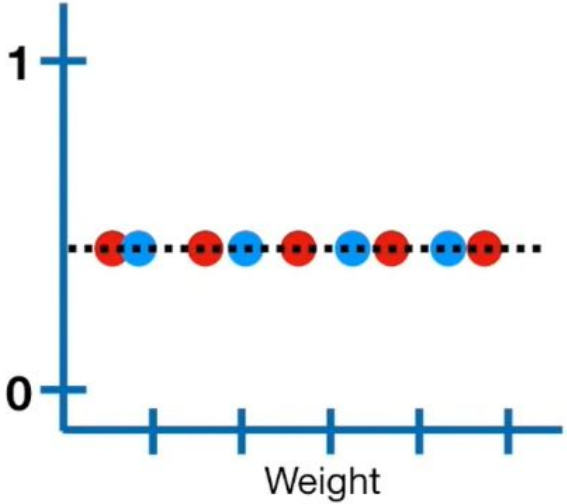


$$p = \frac{\text{number of obese mice}}{\text{total number of mice}} = \frac{5}{9} = \boxed{0.56}$$

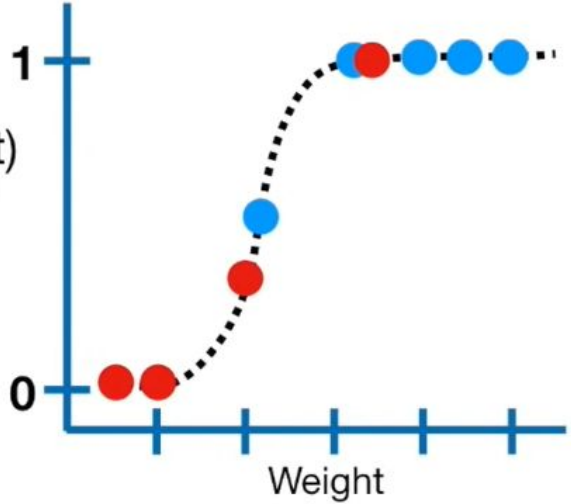
$$p = \frac{e^{0.22}}{1 + e^{0.22}} = \boxed{0.56}$$

Hooray!
They are the same!

The log-likelihood R^2 values go from 0, for poor models, to 1, for good models.



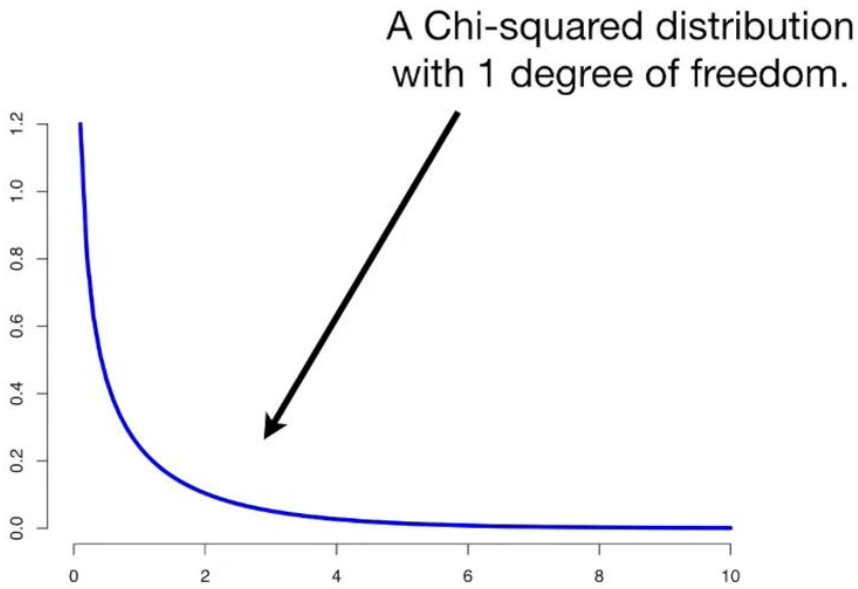
$$R^2 = \frac{\text{LL(overall probability)} - \text{LL(fit)}}{\text{LL(overall probability)}}$$



p-values

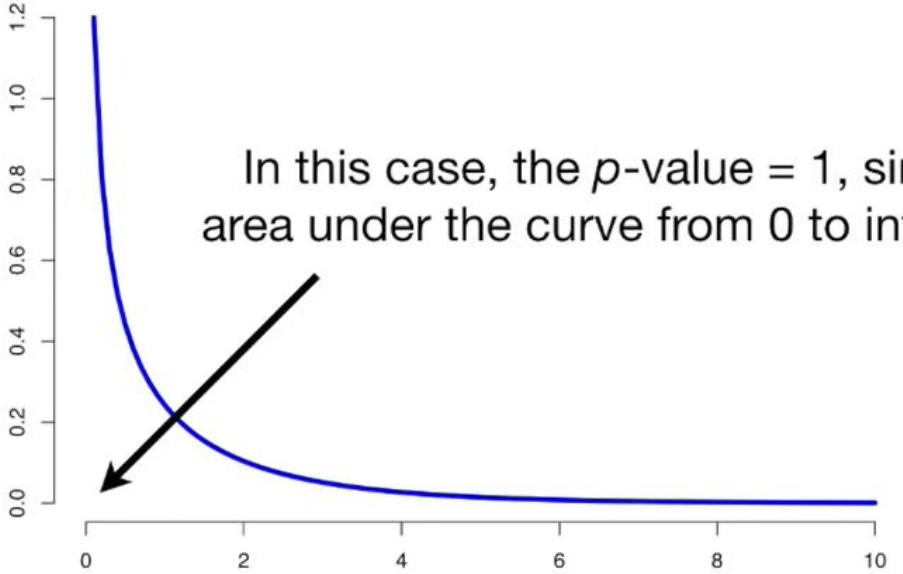
It is simple to calculate the p-value

2(LL(fit)-LL(Overall Probability))=A Chi squared value with degrees of freedom 1



In the worst case scenario,
LL(fit) = LL(overall probability) and the
whole thing = 0.

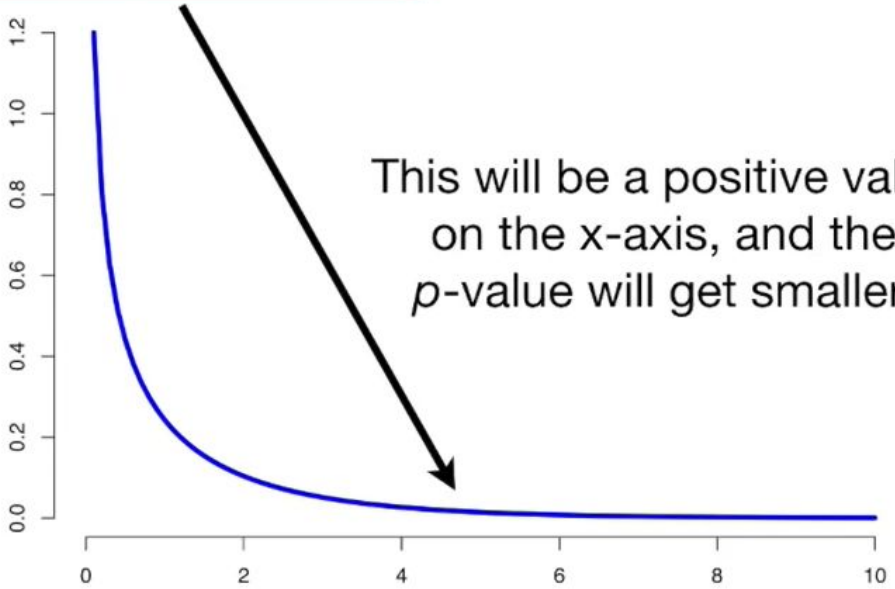
2(LL(fit) - LL(overall probability)) = 0



In this case, the *p*-value = 1, since the
area under the curve from 0 to infinity = 1.

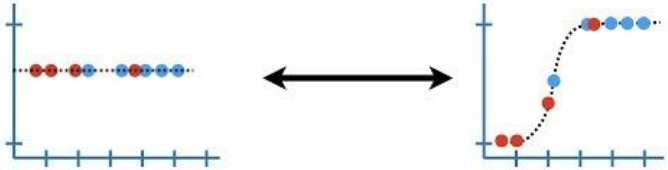
However, most of the time LL(fit) will be closer to 0 than LL(overall probability), and since the log-likelihoods are negative….

$\boxed{2(\text{LL(fit)} - \text{LL(overall probability)})}$ = A Chi-squared value.



This will be a positive value on the x-axis, and the *p*-value will get smaller.

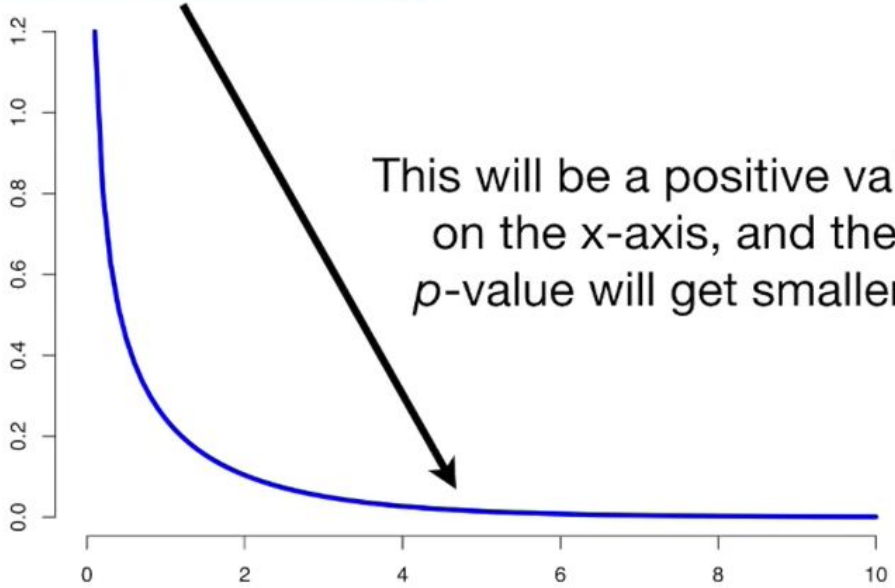
Logistic Regression Details Pt 3...



...R^2 and its *p*-value!!!

However, most of the time LL(fit) will be closer to 0 than LL(overall probability), and since the log-likelihoods are negative....

2(LL(fit) - LL(overall probability)) = A Chi-squared value.



This will be a positive value on the x-axis, and the *p*-value will get smaller.

Decision Tree