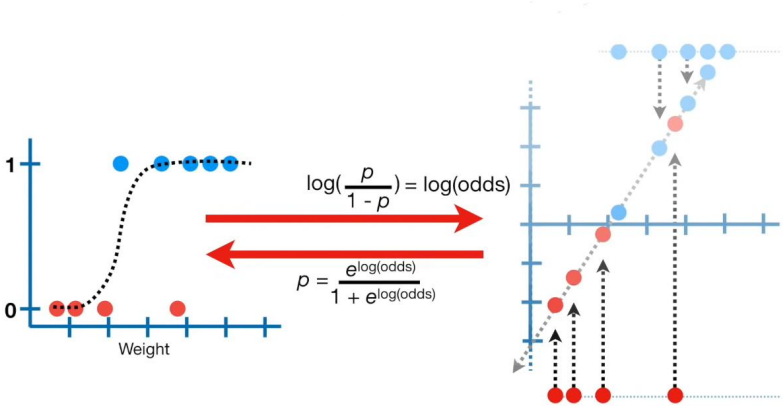
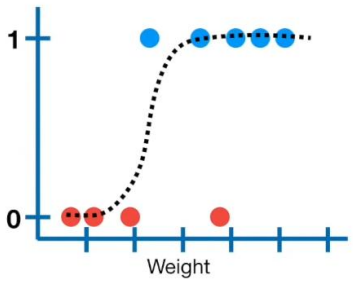
Logistic Regression, Details Part 3: R^2 and p-values

R-squared Calculation



However, we ended with a little bit of a cliff-hanger. We know that the line is the "best fit", but how do we know if it is useful?



In other words, how do we calculate R^2 and a p-value for the relationship between weight and obesity?

Even though pretty much everyone agrees on how to calculate R^2 and the associated p-value for Linear Models…

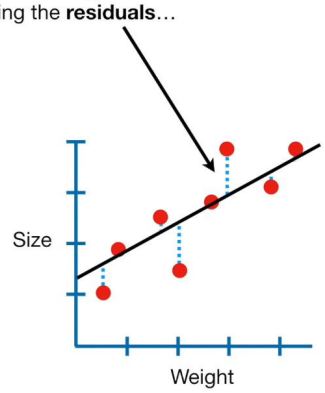
…there is no consensus on how to calculate R^2 for Logistic Regression. There are more than 10 different ways to do it!!!

For this StatQuest, rather than describe every single R^2 for Logistic Regression, I'm focusing on one that is commonly used and is easily calculated from the output that R gives you.

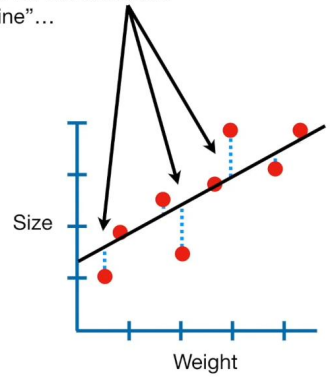
Just so you know, this R^2 is called "McFadden's Pseudo R^2"

Another bonus is that this method is very similar to how R^2 is calculated for regular old Linear Models.

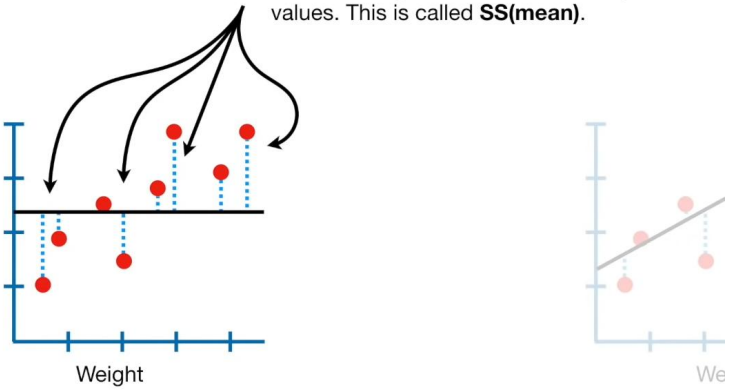
In Linear Regression (and other linear models), R^2 and the related p-value are calculated using the **residuals**…



In brief, we square the residuals and then add them up. I call this **SS(fit)**, for "sum of squares for the Residuals around the best fitting line"…

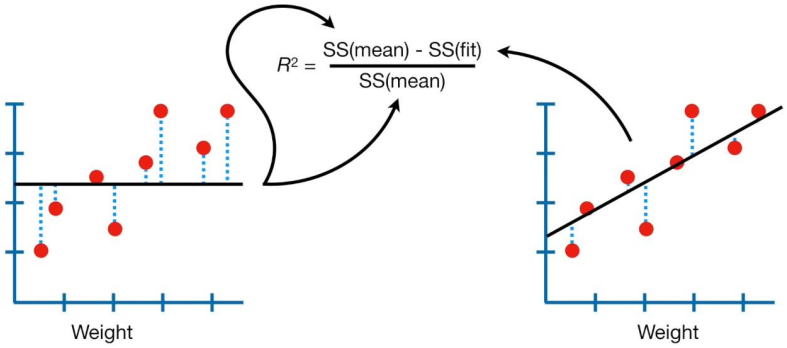


…and we compare that to the sum of squared residuals around the worst fitting line, the mean of the y-axis values. This is called **SS(mean)**.

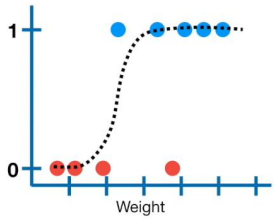


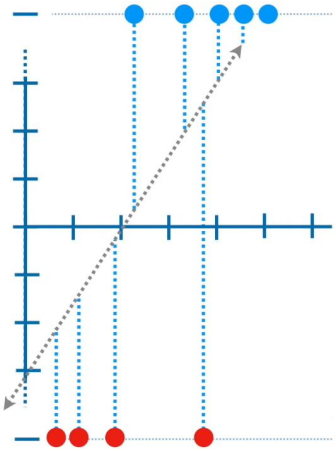
R^2 compares a measure of a good fit, **SS(fit)**…
…to a measure of a bad fit, **SS(mean)**…

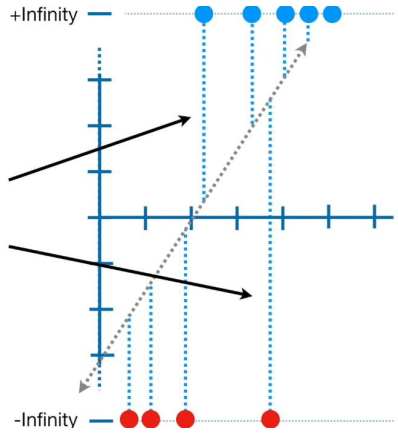
$$R^2 = \frac{\text{SS(mean)} - \text{SS(fit)}}{\text{SS(mean)}}$$



Like linear regression, we need to find a measure of a good fit to compare to a measure of a bad fit…



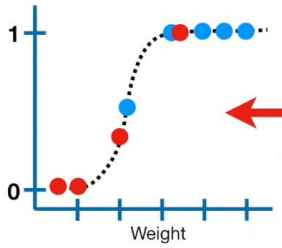




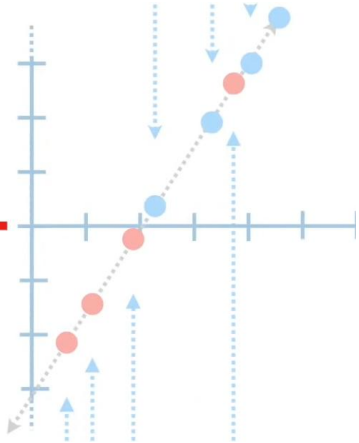
Unfortunately, the residuals for Logistic Regression are all infinite, so we can't use them…

…but we can project the data onto the best fitting line…

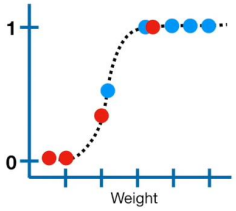
…and then we translate the log(odds) back to probabilities…



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



…and lastly, calculate the log-likelihood of the data given the best fitting squiggle.

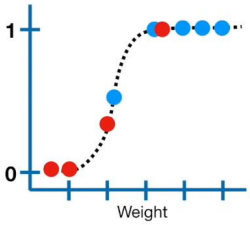


log(likelihood of data given the squiggle) =

log(0.49) + log(0.9) + log(0.91) + log(0.91) +
log(0.92) + log(1 - 0.9) + log(1 - 0.3) +
log(1 - 0.01) + log(1 - 0.01)

= -3.77

We can call this **LL(fit)**, for the log-likelihood of the fitted line, and use it as a substitute for **SS(fit)**.



log(likelihood of data given the squiggle) =

log(**0.49**) + log(**0.9**) + log(**0.91**) + log(**0.91**) +
log(**0.92**) + log(**1 - 0.9**) + log(**1 - 0.3**) +
log(**1 - 0.01**) + log(**1 - 0.01**)

LL(fit) = -3.77

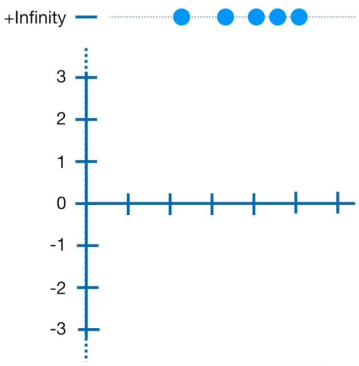
Now we need a measure of a poorly fitted line that is analogous to SS(mean)…

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

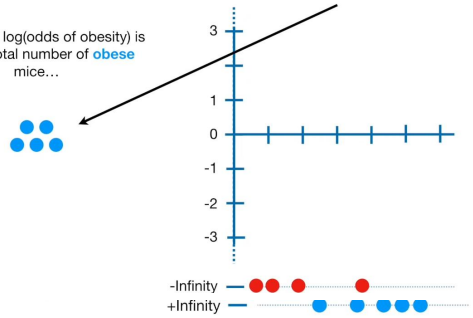
Vs

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

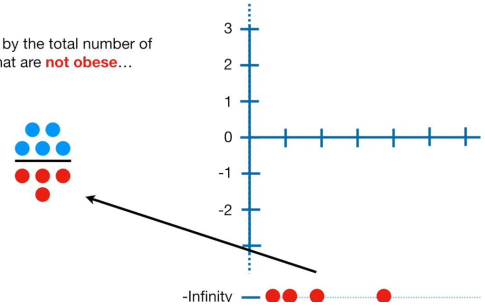
We do this by calculating the log(odds of obesity) without taking weight into account.

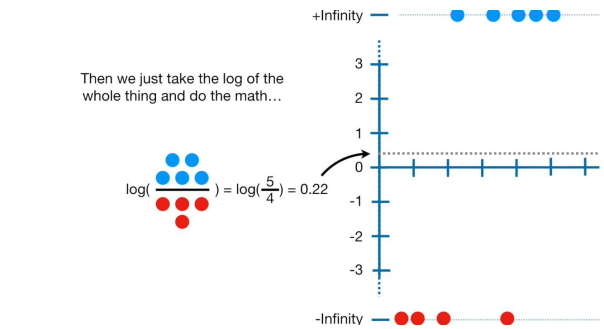


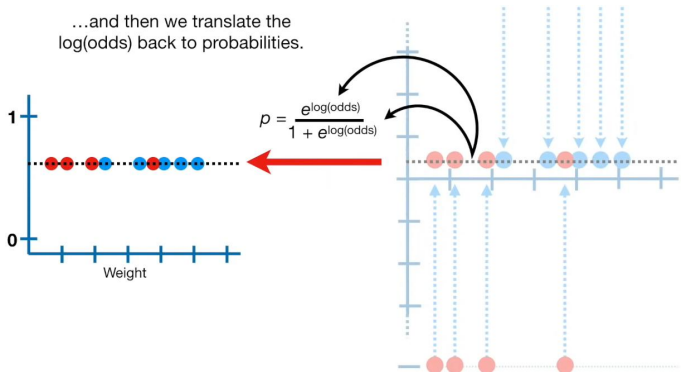
The overall log(odds of obesity) is just the total number of **obese** mice…



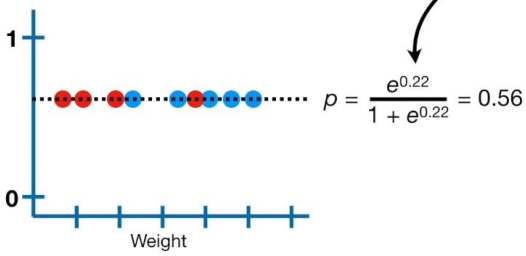
…divided by the total number of mice that are **not obese**…



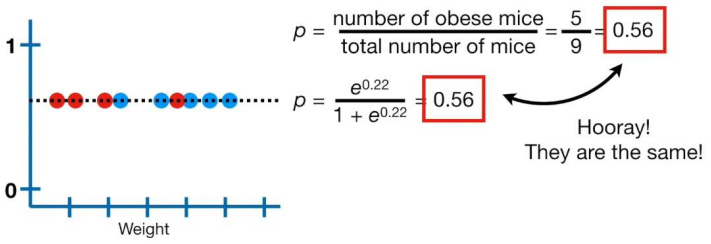




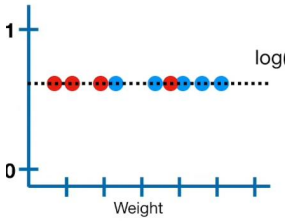
NOTE: The overall log(odds), 0.22, *translates* to the overall probability of being obese, 0.56.



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

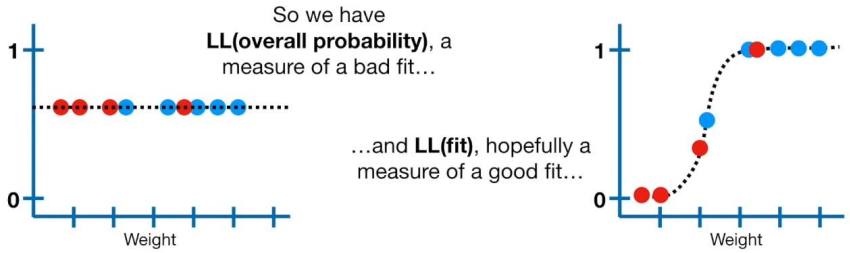


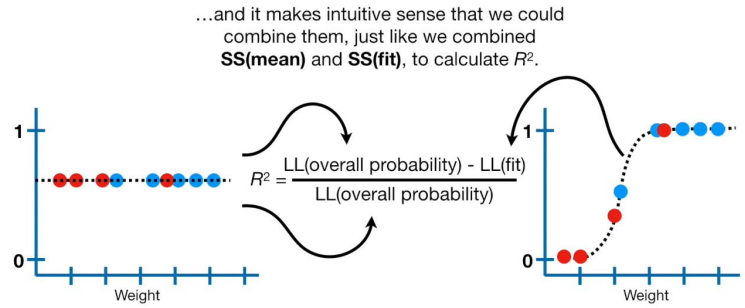
We'll call this **LL(overall probability)** and use it as a substitute for **SS(mean)**.



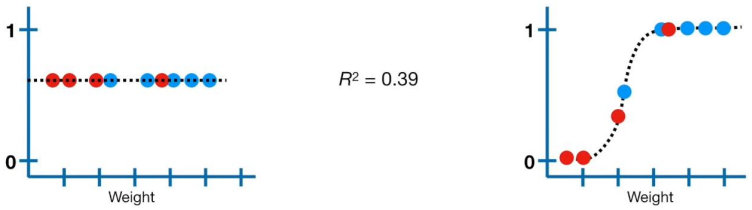
log(likelihood of data given overall probability of obesity) =
log(0.55) + log(0.55) + log(0.55) + log(0.55) +
log(0.55) + log(1 - 0.55) + log(1 - 0.55) +
log(1 - 0.55) + log(1 - 0.55)

LL(overall probability) = -6.18





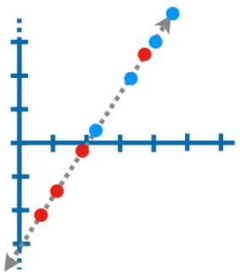
BAM!!!



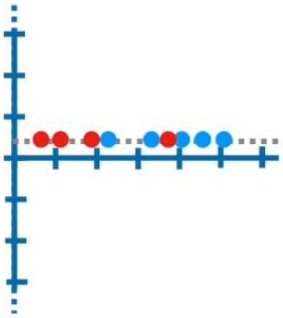
P-value Calculation

The good news is that calculating the *p*-value is pretty straight forward.

2(LL(fit) - LL(overall probability)) = A Chi-squared value with degrees of freedom equal to the difference in the number of parameters in the two models.



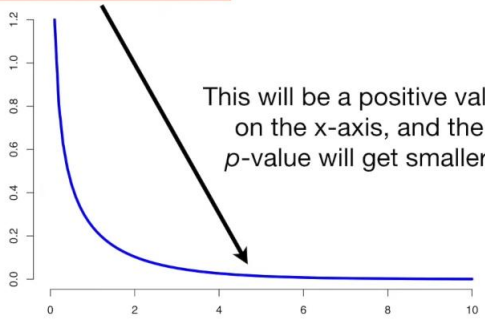
LL(fit) has 2 parameters since it needs estimates for a y-axis intercept and a slope.



LL(overall probability) has 1 parameter since it only needs an estimate for a y-axis intercept.

Degrees of freedom = 2 - 1 = 1

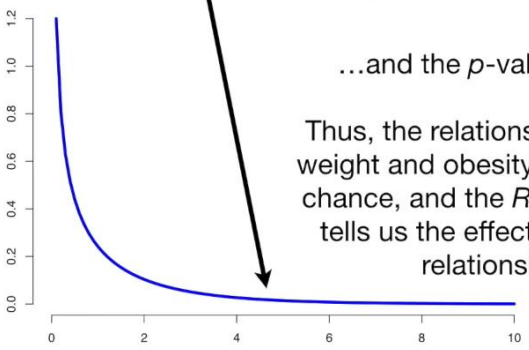
2(LL(fit) - LL(overall probability)) = A Chi-squared value.



P-value is the area under the curve on the right of the Chi-squared value

2(-3.77 + 6.18) = 4.82

A Chi-squared value with 1 degree of freedom.



…and the *p*-value = 0.03.

Thus, the relationship between weight and obesity is not due to chance, and the R^2 value, 0.39, tells us the effect size of this relationship.

R-squared Formula for all situations

When you see these formulas for R^2 and the associated *p*-value out in the wild…

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

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

…they will look more like this…

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

$$\text{2((LL(saturated model) - LL(fit)) - (LL(saturated - model) - LL(overall probability))} = \text{A Chi-square value}$$

This is because these formulas usually include terms for the **saturated model**.

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

$$\text{2((LL(saturated model) - LL(fit)) - (LL(saturated - model) - LL(overall probability))} = \text{A Chi-square value}$$

I'll talk about the Saturated Model in another StatQuest. For now, however, just know that when doing Logistic Regression, the log-likelihood of the **saturated model** = 0, so we can omit it…

However, the log-likelihood of the saturated model isn't always 0 when it is used for other "Generalized Linear Models"…