PRIOR – KNOWLEDGE

STATQUEST:LOGISTIC REGRESSION:

We have some data (Weight ~ Size), and we fit a line to it. With that line, we could do a lot of things:

- Calculate R^2 and determine if **weight** and **size** are correlated. **Large values imply a large effect.**
- Calculate a p-value to determine if the R2R2 value is statistically significant.
- Use the line to predict **size** given **wight**.

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

Instead of fitting a line to the data, logistic regression **fits an "S" shaped "logistic function**".

Logistic regression is usually used for **classification**.

For logistic regression, unlike normal regression, we can't easily compare the complicated model to the simple model. Instead, we just to see if a variable's effect on the prediction is significantly different 0. If not, it means the variable is not helping the prediction. (We use "Wald's Test" to figure this out.)

Logistic regression's 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". BUt in logistic regression, we use something called "maximum likelihood".

We find the line that minimizes the sum of the squares of these residuals. We also use the residuals to calculate R2R2 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 R2R2.

Coefficients

> Logistic Regression is a specific type of **Generalized Linear Model** (often abbreviated **GLM**). And GLM are a generalization of the concepts and abilities of regular Linear Models.

Just like with linear regression, the best fitting line has a y-axis intercept and a slope. The coefficients for the line are what you get when you do logistic regression.

$$y=-3.48+1.83\times \text{weight}$$

y=−3.48+1.83×weight

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.475	2.364	-1.471	0.1414
weight	1.825	1.088	1.678	0.0934

- The first coefficient is the y-axis intercept when $\text{weight}=0$weight=0, it means that when $\text{weight}=0$weight=0, the $\log(\text{odds of obesity})$log(odds of obesity) are -3.48−3.48.
- The z-value is the estimated intercept divided by the standard error. (it's the number of standard deviations the estimated intercept is away from 0 on a standard normal curve.) That means that this is the Wald's Test that we talked about in the odds ratio StatQuest. Since the estimate is less than 2 standard deviations away from 0, we know it is not statistically significant.
- The second coefficient is the slope. It means that for every one unit of weight gained, the $\log(\text{odds of obesity})$log(odds of obesity) increased by 1.825.

Maximum Likelihood

The algorithm tha finds the line with the maximum likelihood is pretty smart - each time it rotates the line, it does so in a way that increases the log-likelihood. Thus, the algorithm can find the optimal fit after a few rotation.