LOGISTICAL REGRESSION

ISL Chapter 4

In an experiment, the variables that you measure typically fall into a given category:

- Nominal variables: This is a variable that falls into a category.

  eye-color, race, blood type

- Ordinal variables: This is a variable that has a rank or an order.

  education level,

  low-middle-high income level

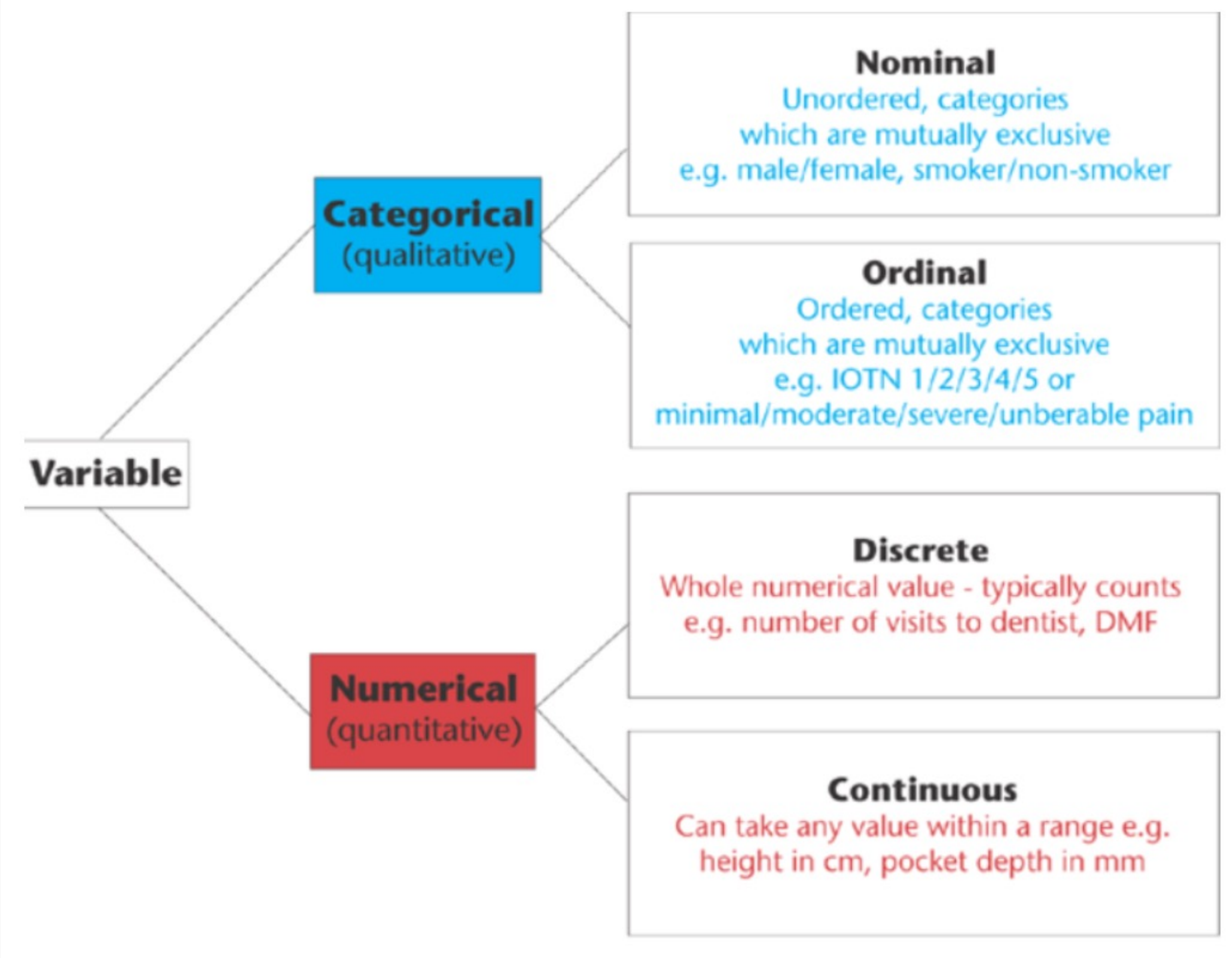
- Interval variables: These variables have a measurable difference but no real zero.

  Temperature, SAT score

- ratio-level intervals: These variables have a difference and a real zero.

  Chemical concentration,

  pulse, weight



Variable

Categorical (qualitative)

Nominal
Unordered, categories which are mutually exclusive e.g. male/female, smoker/non-smoker

Ordinal
Ordered, categories which are mutually exclusive e.g. IOTN 1/2/3/4/5 or minimal/moderate/severe/unberable pain

Numerical (quantitative)

Discrete
Whole numerical value - typically counts e.g. number of visits to dentist, DMF

Continuous
Can take any value within a range e.g. height in cm, pocket depth in mm

Logistic regression is the appropriate analysis to conduct when the dependent variable is binary.

Simple Linear Regression

highest degree → salary

Multiple Linear Regression

highest degree, years of experience, age → salary

Logistic Regression

highest degree, years of experience, age → hired/not hired

Like all regression analyses, the logistic regression is a predictive analysis.

The independent variable(s) can be nominal, ordinal, interval, or ratio-level.

Logistic Regression is used for predicting dichotomous variables, 0 or 1.

The probability of occurrence of characteristic 1 (present) is estimated from a logistic curve.

For instance, predictors such as age, gender, smoking status predict presence/absence of disease.

Logistical Regression answers questions like:

> How does the probability of getting COVID (yes or no) change for every additional pound a person is overweight and for every pack of cigarettes smoked per day?
>
> Do body weight, calorie intake, fat intake, and age have an influence on the probability of having a heart attack (yes or no)?

Logistic Regression assumptions

The dependent variable should be binary in nature (true/false, present/absent, 1/0)

There should be no outliers

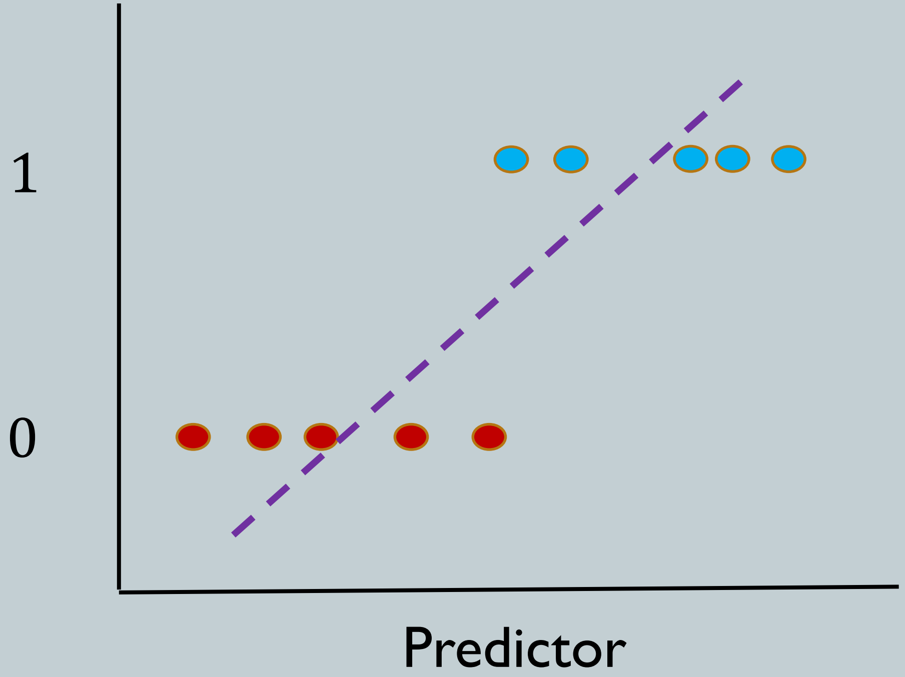
There should be no multicollinearity

Why not use linear regression?

Linear regression cannot be used since binary data does not have a normal distribution. A line indicates that values can go less than 0 and greater than 1. Problem.

The goal of logistic regression is to estimate the probability of occurrence, not the value of the variable itself.

$$\hat{y} = \beta_0 + \beta_1 x_1 + +\beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \epsilon$$

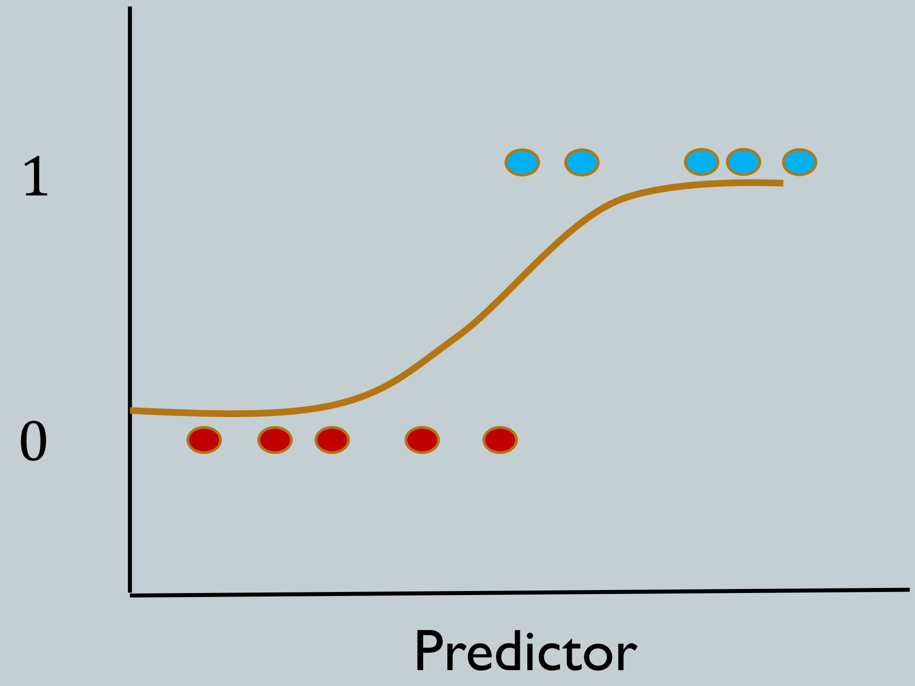


Why not use linear regression?

The goal of logistic regression is to estimate the probability of occurrence, not the value of the variable itself.

The range of values for the prediction is restricted to the range between 0 and 1. Since only values between 0 and 1 are possible, the logistic function f is used.

$$f(z) = \frac{1}{1 + e^{-\boldsymbol{z}}}$$



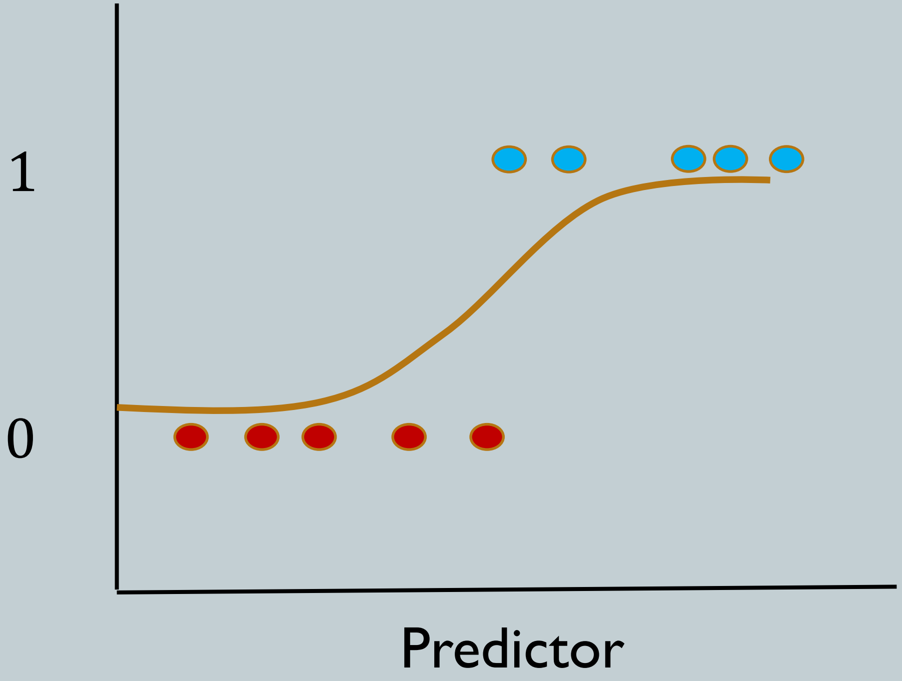
$$\boldsymbol{z} = \beta_0 + \beta_1 x_1 + +\beta_2 x_2 \ +\beta_3 x_3 + \cdots + \beta_k x_k + \epsilon$$

In the logistic function only values between 0 and 1 are possible.

$$\lim_{z \to -\infty} f(z) = 0$$

$$\lim_{z \to \infty} f(z) = 1$$

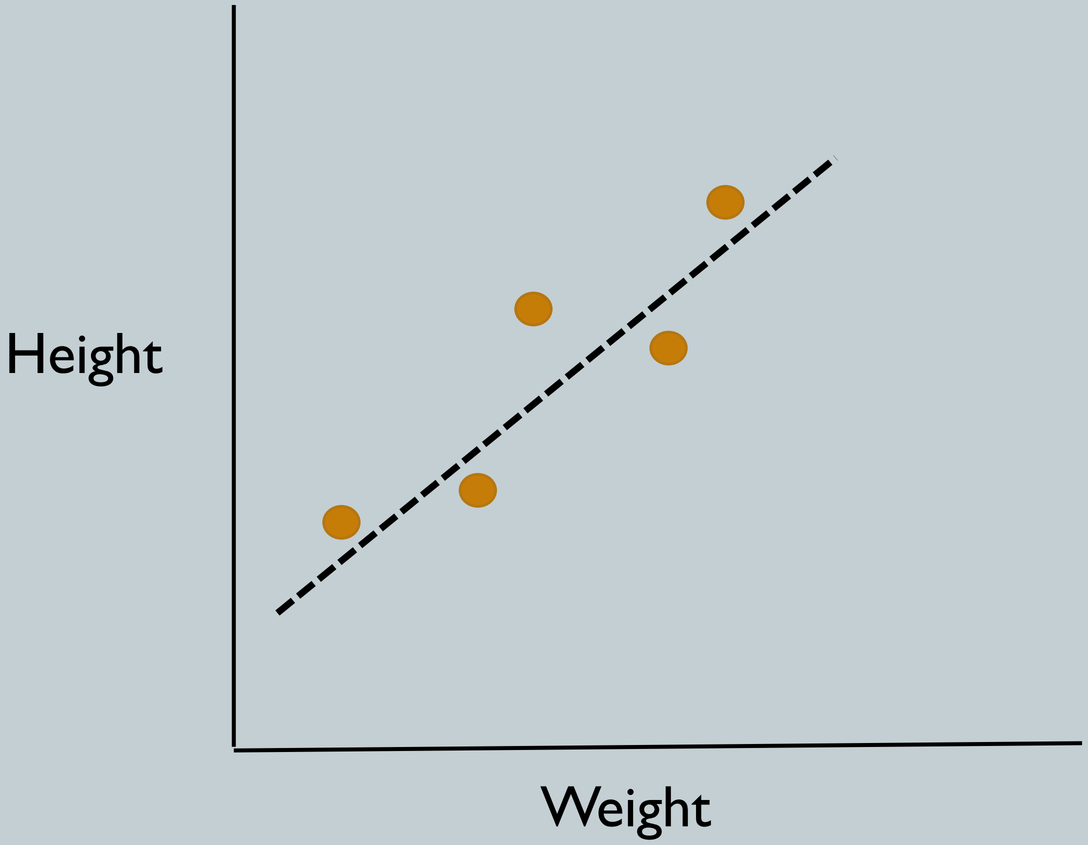
$$f(z) = \frac{1}{1 + e^{-\boldsymbol{z}}}$$



$$\boldsymbol{z} = \beta_0 + \beta_1 x_1 + +\beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \epsilon$$

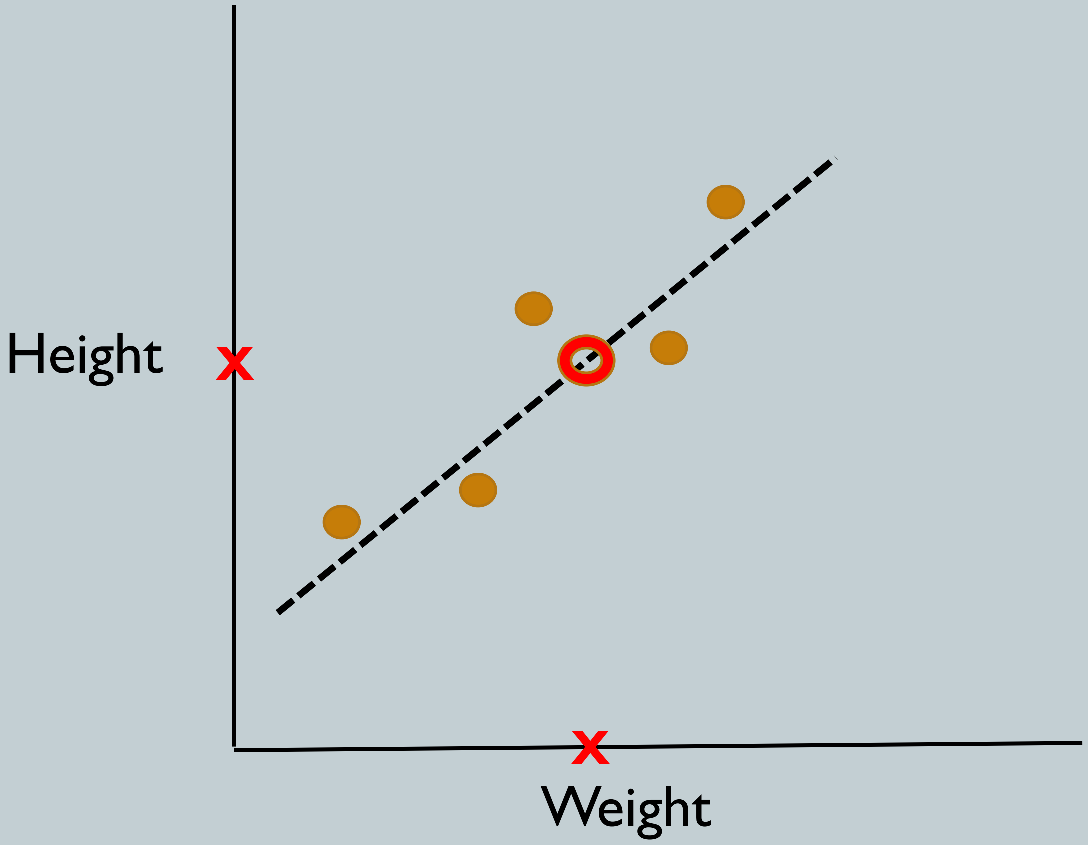
Recall SLR

We created a model, using least squares, that generally fit the data.



We then used the line to predict height given weight.

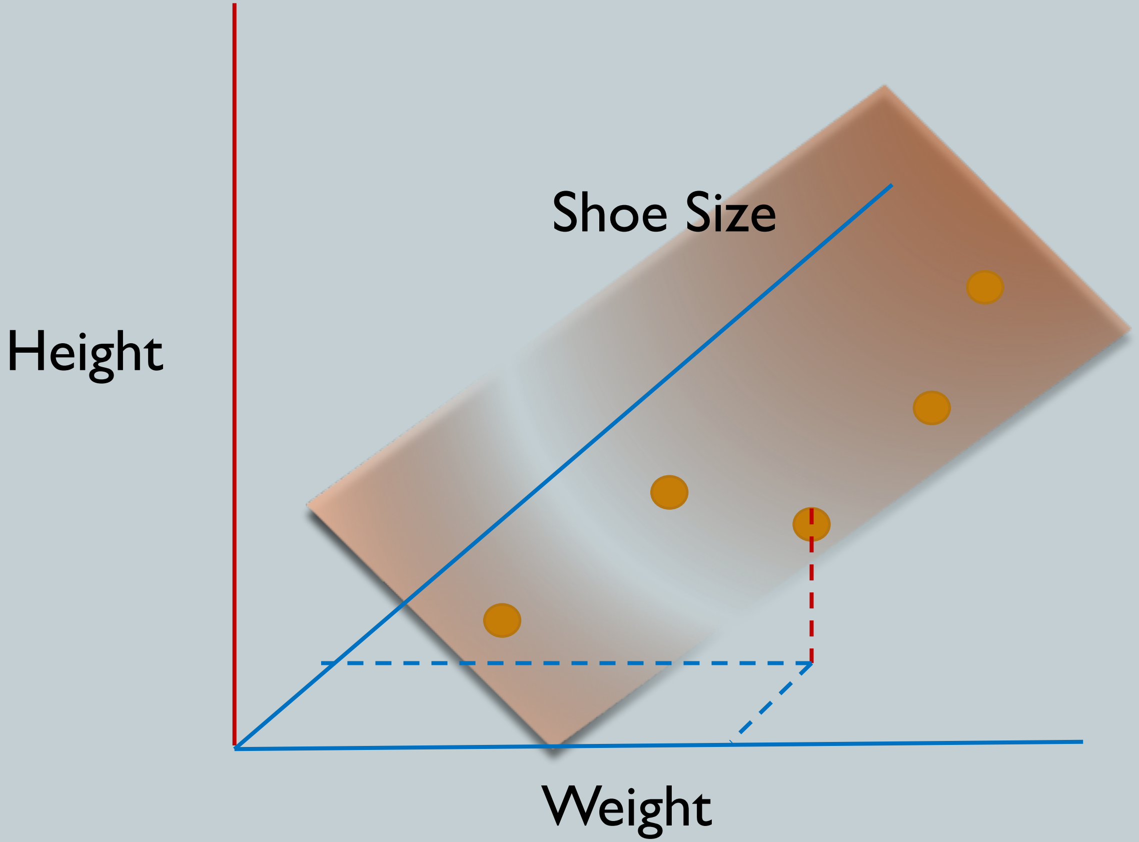
This type of linear regression is also called machine learning.



Recall MLR

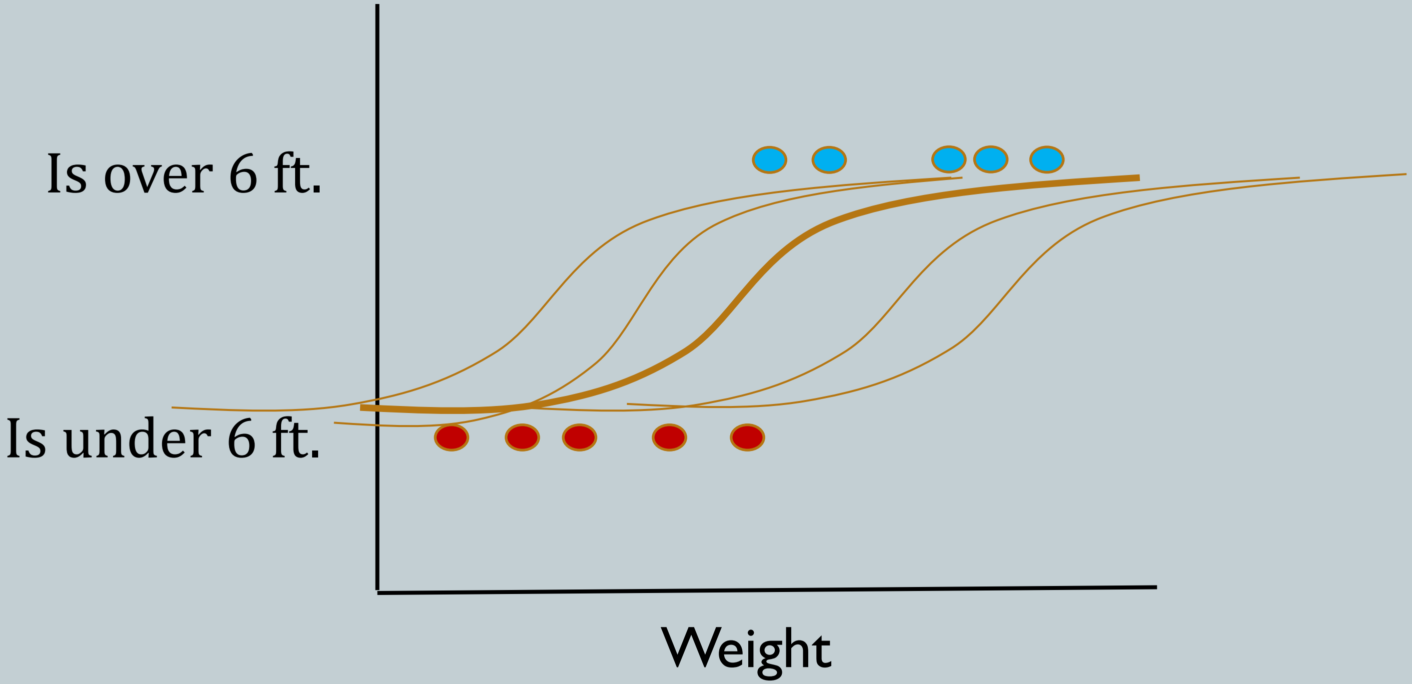
Let's predict height from weight **and** shoe size

comparing the simple model to the complicated model tells us if we need to measure weight AND shoe size to accurately predict height or if we can just get away with weight as a predictor.



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

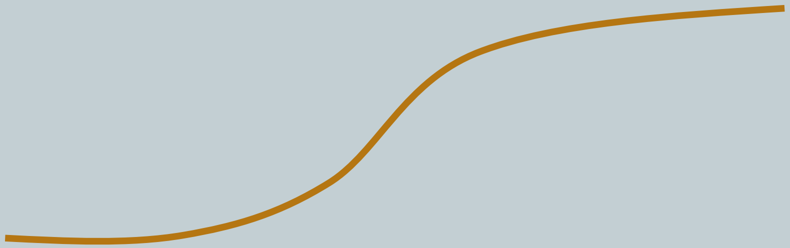
Instead of fitting a line, we will be fitting an S-shaped logistic regression function.



This S-shaped regression function is called the sigmoid function.

The sigmoid function and its properties

$sigmoid(z) = f(z) = \frac{1}{1+e^{-z}}$



$\lim_{z\to\infty} f(z) = 1$ $\lim\limits_{z\to-\infty} f(z) = 0$ $f(0) = \frac{1}{2}$

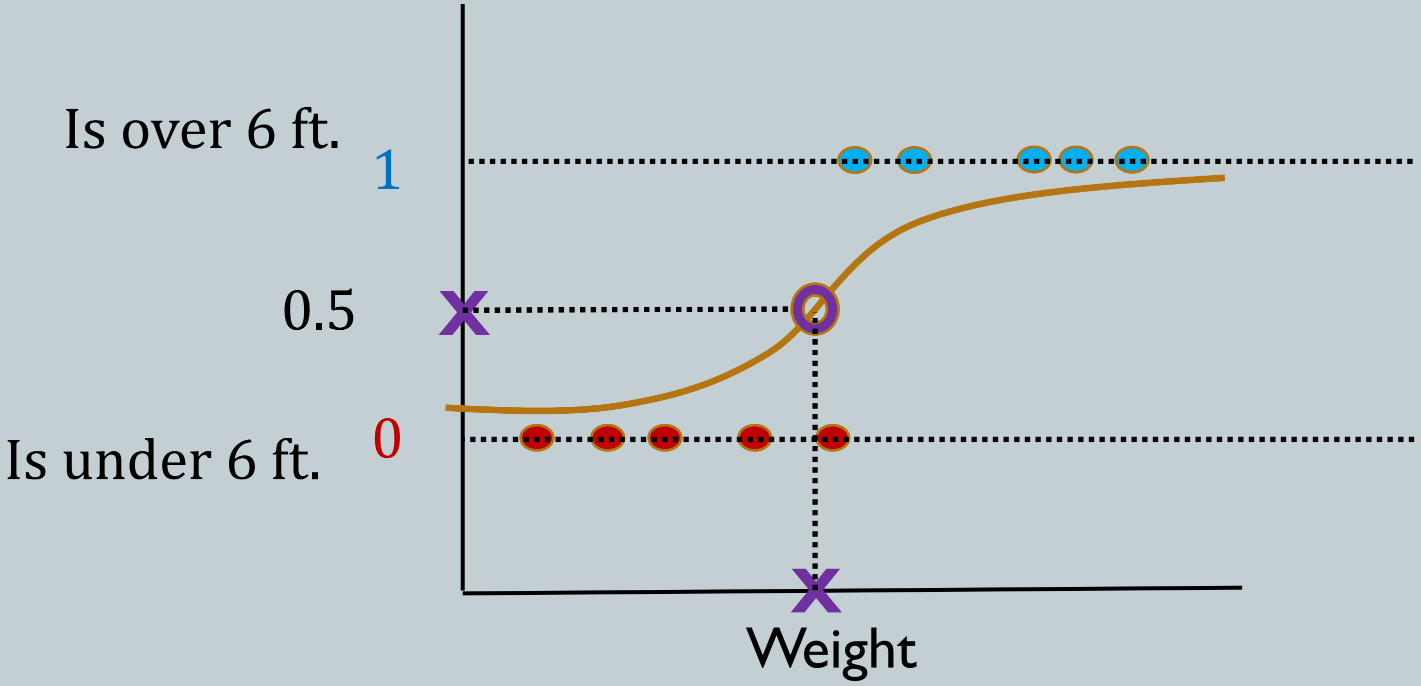
Behaves like probability

Distinguishes between points

Really smooth, important for calculus reasons, derivatives,…

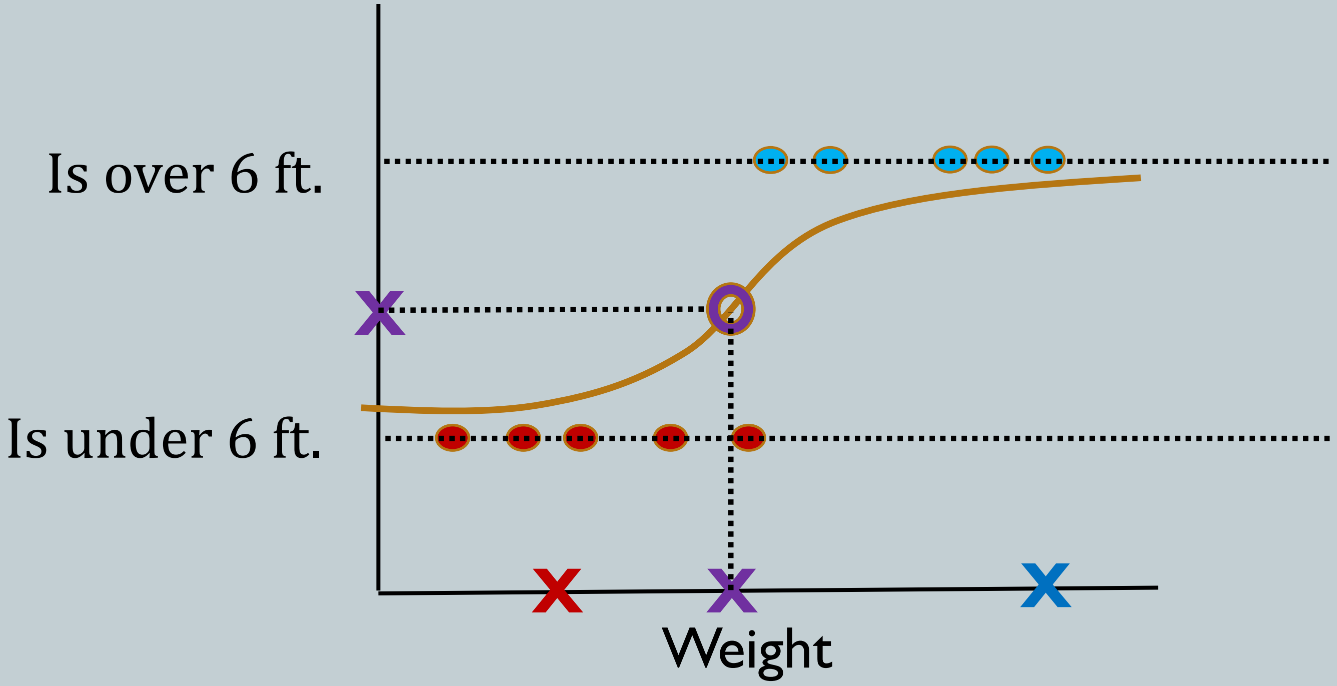
Using the sigmoid

There is only a 50% chance this person is over 6 ft. tall.



Although logistic regression will tell us the probability that a person is over 6 ft. tall, it is usually just used for classification.

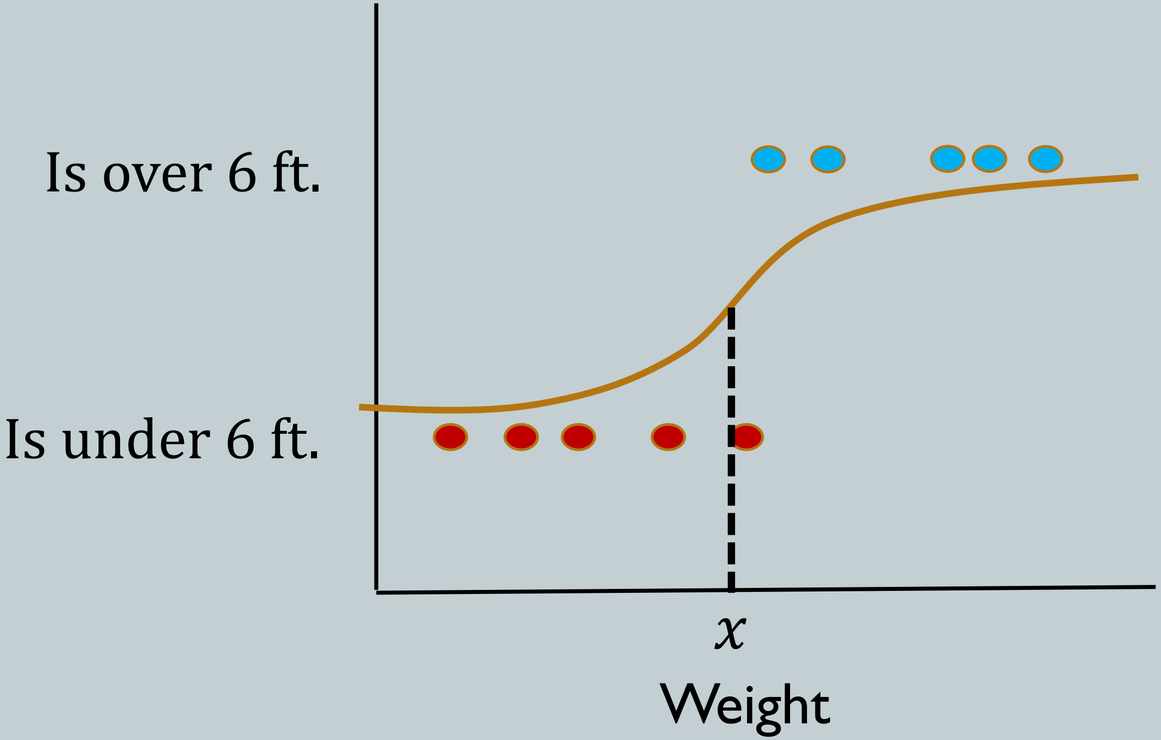
We can set a threshold and then declare anyone past the threshold will be categorized as a 1.



Now, suppose $z = \beta_0 + \beta_1 x$ in $f(z) = \frac{1}{1+e^{-z}}$

$$p(y = 1 \mid x) = f(\beta_0 + \beta_1 x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x\ )}}$$

What if $\frac{1}{1+e^{-(\beta_0+\beta_1 x\ )}} = \frac{1}{2}$ i.e., solve for x.



$x = \frac{\beta_0}{\beta_1}$ is the ‘threshold’

Over 6 ft. could also be predicted by weight, genetics, age, and month born

In other words, just like linear regression, logistic regression can work with continuous data (weight and age) and discrete data (genetics, month born).

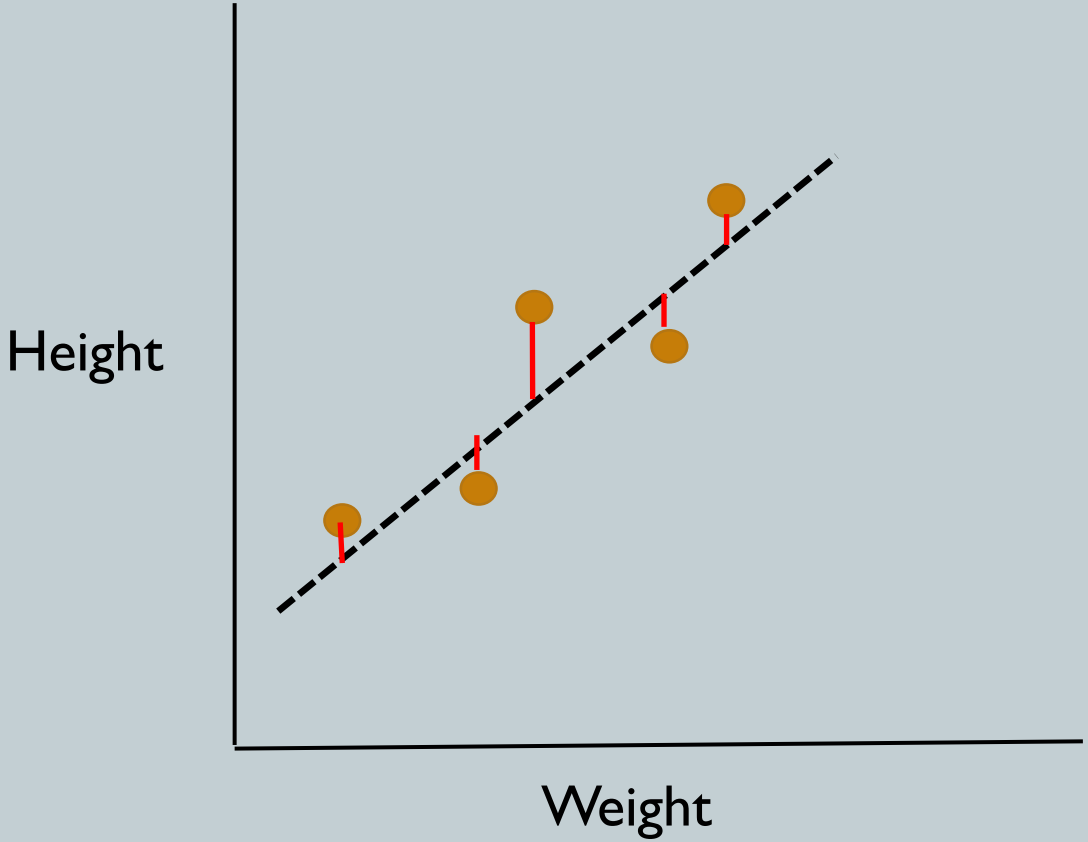
Unlike normal regression, we cannot easily compare the complicated model to the simple model.

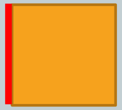
over 6ft. tall predicted by weight, genetics, age, month born
versus
over 6ft. tall predicted by weight, genetics, age.

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

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

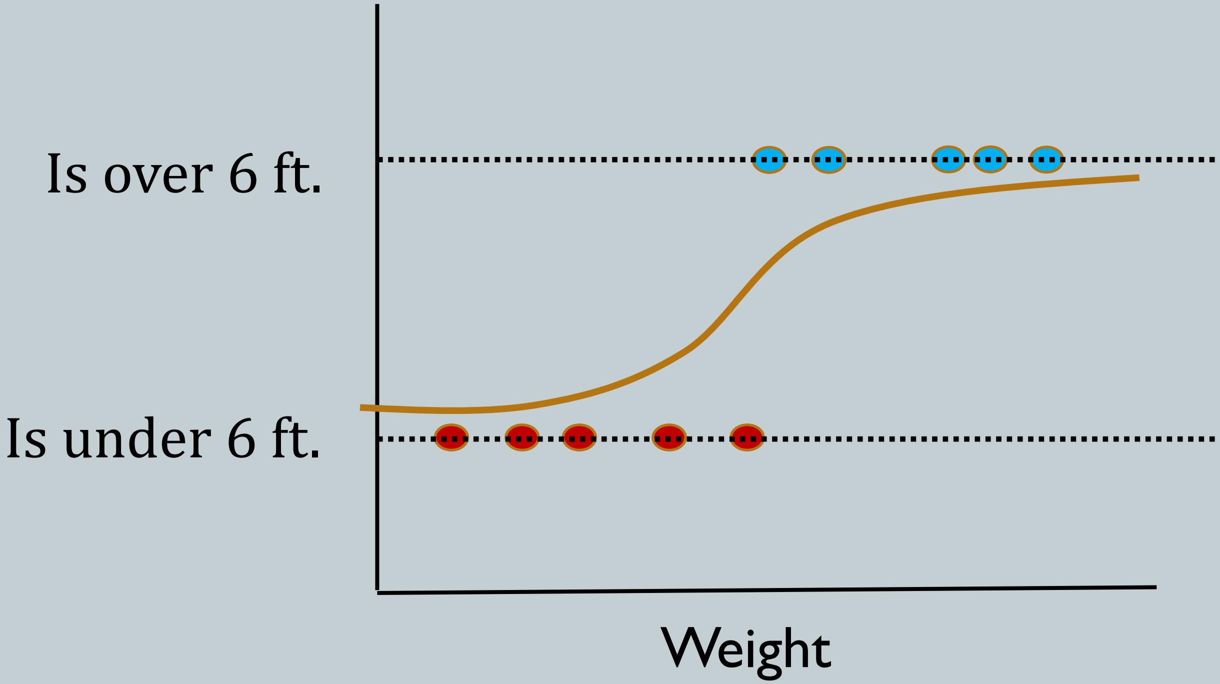
With SLR we found the line that minimizes the sum of the squares of the residuals.





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 cannot calculate R^2.



Side note. What is the relationship between 'Probability' and 'odds'?

$$odds = \frac{p}{1 - p}$$

Example:

If you have a 75% chance of winning, then your odds are $\frac{.75}{1-.75} = \frac{3/4}{1/4} = \frac{3}{1}$, or the odds are 3 to 1 in your favor.

If you have a 10% chance of winning, then your odds are $\frac{.1}{1-.1} = \frac{1/10}{9/10} = \frac{1}{9}$, or the odds are 9 to 1 against you.

$$\text{odds} = \frac{\text{Number of ways to win}}{\text{Number of ways to lose}}$$

If the probability that $y = 1$, given x, is $sigmoid(x)$, (recall sigmoid is a probability)
then what are the odds that $y = 1$ given x ?

$$p(y = 1 \mid x) = f(\beta_0 + \beta_1 x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x)}}$$

$odds = \frac{p}{1-p}$ implies

$$\frac{\frac{1}{1+e^{-(\beta_0+\beta_1 x)}}}{1 - \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \cdot \frac{1+e^{-(\beta_0+\beta_1 x)}}{1+e^{-(\beta_0+\beta_1 x)} - 1} = \frac{1}{e^{-(\beta_0+\beta_1 x)}} = e^{(\beta_0+\beta_1 x)}$$

The odds that $y = 1$ given x are $e^{(\beta_0+\beta_1 x)}$

So, the odds that $y = 1$ given x are $e^{(\beta_0+\beta_1 x)}$.

$odds = \ e^{(\beta_0+\beta_1 x)}$

Taking the natural log of both sides: $\log(odds) = \beta_0 + \beta_1 x$

Meaning, we are actually doing SLR on $\log(odds)$ instead of probability

With multiple predictors the boundary is no longer a point:

$$\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}} = \frac{1}{2}$$

Solve for x_2

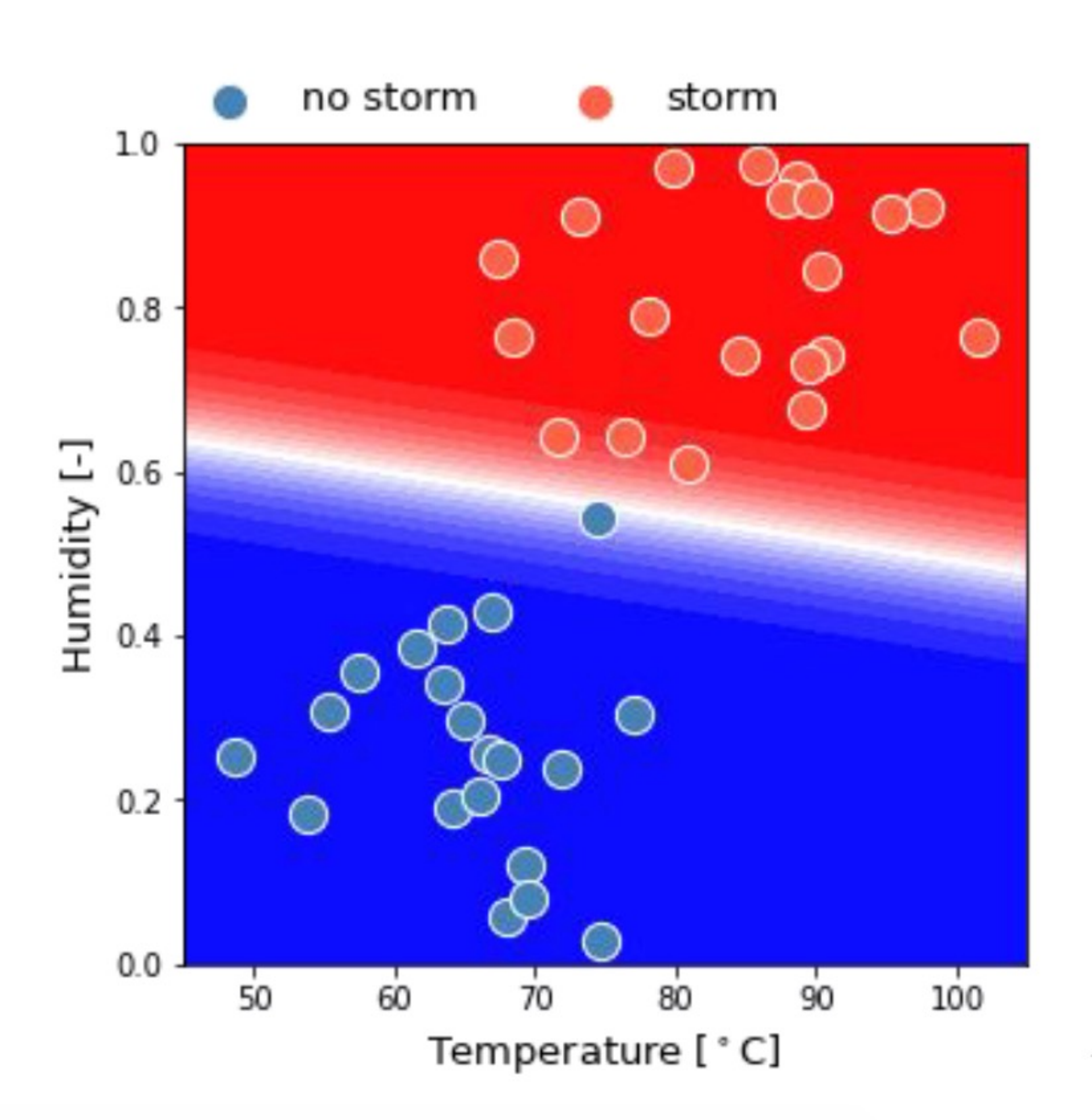
The logreg (logarithmic regression) model with a single feature looks like:

$$p(y = 1 \mid x) = sigmoid(\beta_0 + \beta_1 x_1)$$

What about when we are looking at several features ?

$$p(y = 1 \mid x) \\ = sigmoid(\beta_0 + \beta_1 x_1 + +\beta_2 x_2 \\ + \beta_3 x_3 + \cdots + \beta_k x_k)$$

Graphically, a decision boundary is created.



Next time:

k