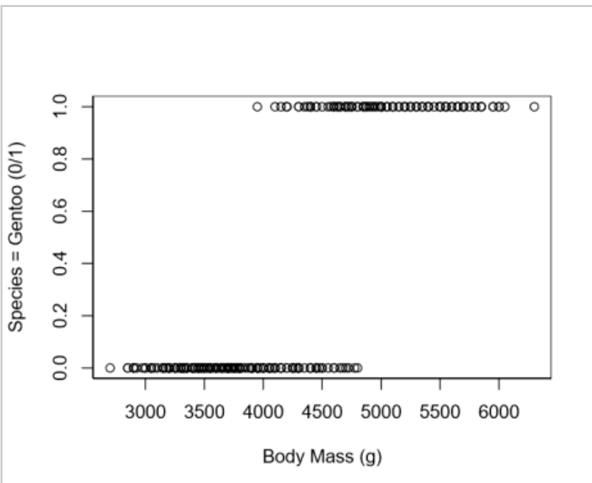
Logistic Regression

Previously

Regression: Describe average value of a continuous response variable Y from one or more variables $X_1, X_2, X_3, \ldots$

numeric (discrete or continuous)

What if the variable Y were categorical rather than numeric? (Note: we focus on binary variables)



Bodymass as a predictor of Species being Gentoo.

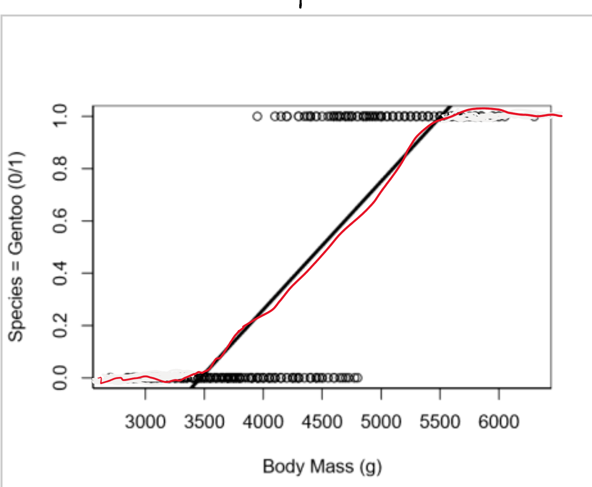
Chinstrap or Adelie

$Y = \{0, 0, 1, 1, 0, \ldots\}$

Gentoo

What if we just change nothing?

$\hat{y} = b_0 + b_1 \cdot BM$ ← Lets fit this model.



This is our fit.

How can we fix it?

$\hat{y} > 1 \to$ replace with 1

$\hat{y} < 0 \to$ replace with 0

Not a statistically motivated solution

This does not work.

The problem is $\hat{y} = \underbrace{b_0 + b_1 BM}_{\text{Takes values on real number line}}$ $\mathbb{R} = (-\infty, +\infty)$

The problem is $y = b_0 + b_1 BM \quad \mathbb{R} = (-\infty, +\infty)$

y ← mean for a given BM

y ← Takes values of 0 or 1.

We want to replace the mean value $\hat{y}$ with a prob (Specifically $P(\text{Species} = \text{Gentoo})$)

$$p = \underbrace{b_0 + b_1 BM}_{(-\infty, \infty)}$$

p: $[0,1]$

<u>2 options</u>

Both sides of equation are in range $(-\infty, \infty)$

or

Both sides in range $(0,1)$

(We will do both)

What we will fit is

$$\text{function}(p) = b_0 + b_1 \cdot BM$$

We need this function to map values of p from $(0,1) \to (-\infty, \infty)$

$p = P(\text{Gentoo})$

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 \cdot BM$$

What is $\frac{p}{1-p}$?

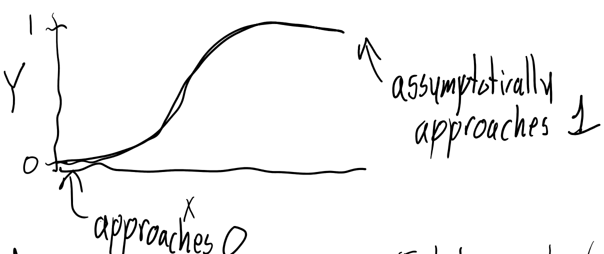
This is the odds.

p is a probability which measures the number of times an event will occur given all possible events

$P(\text{Coin} = \text{heads}) = 1 \text{ in } 2 = 1/2$

$P(\text{Die Roll} = 3) = 1 \text{ in } 6 = 1/6$

<u>Sigmoid</u>



<u>Sigmoid</u>

$$Y = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x}$$

Euler's constant ≈ 2.71

<u>Note</u>

$\log(a) = b$

$a = e^b$

$e^{a+b} = e^a \cdot e^b$

$e^{a-b} = e^a / e^b$

If we solve for X....

$$\log\left(\frac{y}{1-y}\right) = X$$

natural $\log_e$

$p/1-p$ is an odds, which measures the number of times an event occurs compared to the number of times it doesn't occur.

$\text{Odds}(\text{Coin} = \text{heads}) = 1 \text{ to } 1 = 1$

$\text{Odds}(\text{Die Roll} = 3) = 1 \text{ to } 5 = 1/5$

$P(\text{Die Roll}=3) = 1 \text{ in } 6 = 1/6$ ← $[0,1]$

Odds(Coin-heads) ...

$\text{Odds}(\text{Die Roll}=3) = 1 \text{ to } 5 = 1/5$ ← $(-\infty,\infty)$

We define the log-odds using a linear equation

$$\log\left(\frac{p}{1-p}\right) = \ln\left(\frac{p}{1-p}\right) = b_0 + b_1 BM$$

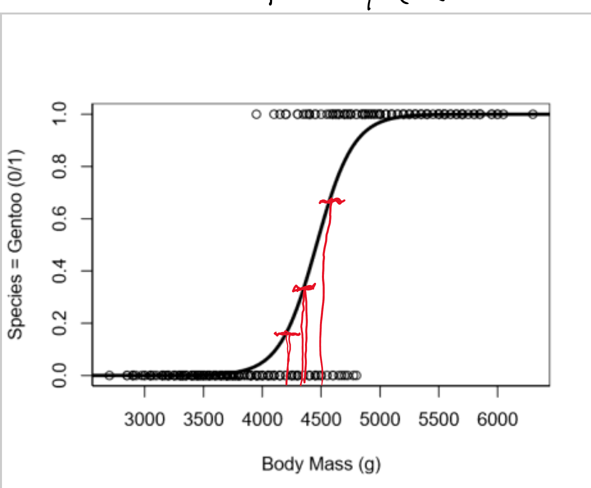
← units of grams

Using R: $\hat{b}_0 = -28.41806$

$\hat{b}_1 = 0.00637 \frac{1}{g}$

$$p = \frac{1}{1+e^{-(b_0+b_1 x)}} = \frac{e^{b_0+b_1 x}}{1+e^{b_0+b_1 x}}$$

$$p = \frac{1}{1+\exp\left(-\left(-28.418 + 0.00637\frac{1}{g}\cdot BM\right)\right)}$$



Interpretation of terms is complex, but easiest if we look at odds.

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 \times BM$$

$$\frac{p}{1-p} = e^{b_0 + b_1 \times BM} = e^{b_0} \times e^{b_1 \cdot BM}$$

Before: 1 unit increase in BM would result in a $\hat{b}_1$ increase in Y on average.

Here: 1 unit increase in BM increases the odds multiplicatively by $e^{\hat{b}_1}$

A one unit increase in body mass is associated with a 0.6% increase in odds.

Penguin: BM = 4000 and BM = 4001

$e^a / e^b = e^{a-b}$

$$\left(\frac{p}{1-p}\right)_{4001} \Big/ \left(\frac{p}{1-p}\right)_{4000} = e^{\hat{b}_0 + \hat{b}_1 \cdot 4001 - (\hat{b}_0 + \hat{b}_1 4000)} = e^{\hat{b}_1}$$

$$e^{0.00637} = 1.00639$$

... BM is associated

in odds.

e = 1.00639

A ten unit increase in BM is associated
with a 1.0658 times increase in odds
(6.58%)

A 100 unit increase in BM ....
...... 89.1% increase (1.891 times)

1000 unit increase .....
... 58300% increase

1kg