Logistic Regression

Max and Marc

Used to classify

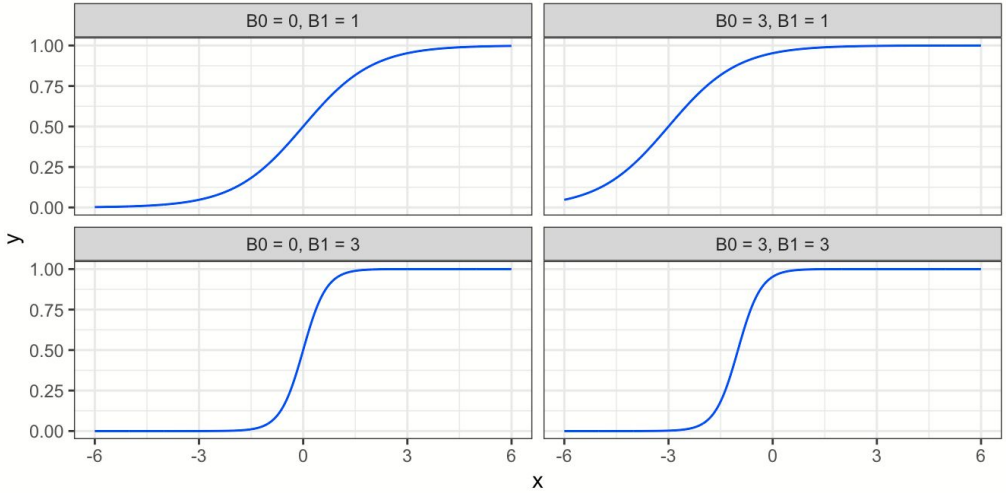
Linear regression only lets us use quantitative Y

Logistic regression is used for qualitative or categorical Y

-> This is what's called "classification", we classify our observations into categorical responses

"A person arrives at the emergency room with a set of symptoms that can be attributed to one of two possible medical conditions. Which of the two conditions does the individual have?"

The logistic model



Linear regressions:

- We model Y directly from X

Logistic regression:

- We model the probability of Y from X
- Binary response 0 and 1

The probability of falling into 0 or 1 is given by a logistic function

β_0 and β_1 are the coefficients we have to estimate

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

The logistic model - Log Formula

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

Through math magic we find:

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}.$$

Odds

Odds range from 0 to ∞, indicating very low and very high probabilities for X

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

Log-odds or Logit
We see that it's linear in X!

$$f(X) = \beta_0 + \beta_1 X_1$$

Recall the linear regression formula!

Maximum Likelihood to find β_0 and β_1

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})).$$

We search for estimates of B0 and B1 that when plugged into our model gives us a number close to 1 for success and 0 for failure, they are chosen to maximize the likelihood function.

Once We know the estimates we can predict stuff

Multiple logistic regression, and more categories?

If we have more X predictors we can modify the formula:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

If we have more Y categories → Linear Discriminant Analysis