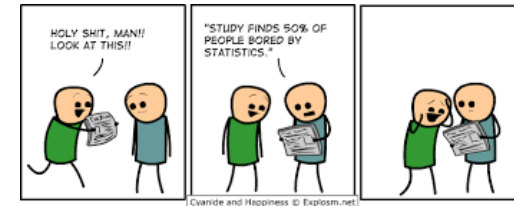


# Applied Linear Modeling

October 22, 2019

## All the things for today

- Type I & II error warm-up
- Mini-lecture on the theory behind the logistic model
- Logistic regression basketball workshop
  - *R package to download: odds.n.ends*
- Slides and workshop packet on GitHub



# What is logistic regression

- A statistical model used to predict or explain a **binary** outcome variable
- For example:
  - What predicts whether or not someone uses the library?
  - What predicts whether or not someone is a smoker?
  - What predicts whether or not someone owns a gun?

# The statistical form of the logistic model

- Because the outcome variable is binary, the linear regression model would not work (it requires a continuous outcome!)
  - Linear model:  $y = b_0 + b_1x_1 + b_2x_2 \dots$
- The linear regression model can be transformed using a *logit transformation* based on the logistic function in order to model binary outcomes
- The logistic function is officially defined:

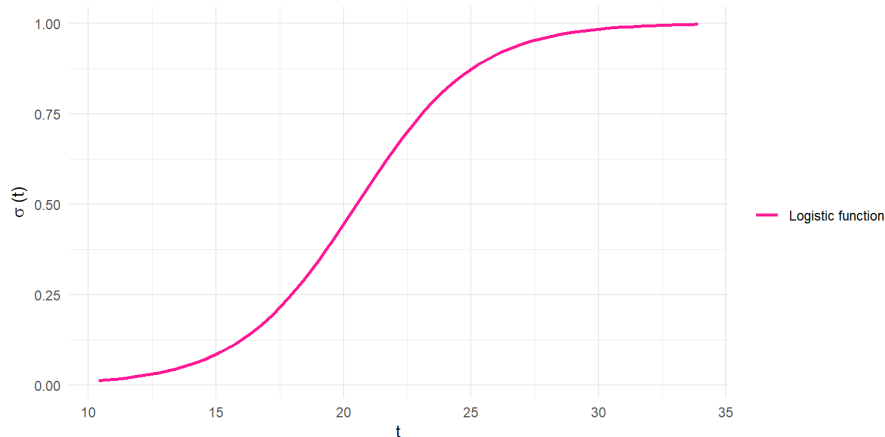
$$\sigma(t) = \frac{e^t}{e^t + 1}$$

- Simplified to:

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

# The logistic function

- When you graph the logistic function, it has a sigmoid shape that stretches from  $-\infty$  to  $\infty$  on the x-axis and from 0 to 1 on the y-axis
- The function  $\sigma(t) = \frac{1}{1+e^{-t}}$  can take any value of  $t$  along the x-axis and give the corresponding value of  $\sigma(t)$  that is between 0 and 1 on the y-axis



# From the logistic function to the logistic model

- Logistic function to logistic model
  - In the case of logistic regression, the value of  $t$  will be the right-hand side of the regression model, which looks something like  $b_0 + b_1x_1 + \dots$*
  - The left hand side of the logistic regression model is the probability of  $y$  or  $p(y)$*
- Start with the logistic function

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

- Substitute in the regression notation of  $x$  and  $y$  and slope and intercept to make the logistic model

$$p(y) = \frac{1}{1 + e^{-(b_0 + b_1x_1 + b_2x_2 \dots)}}$$

# Reading the logistic model

## Logistic model

$$p(y) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2) \dots}}$$

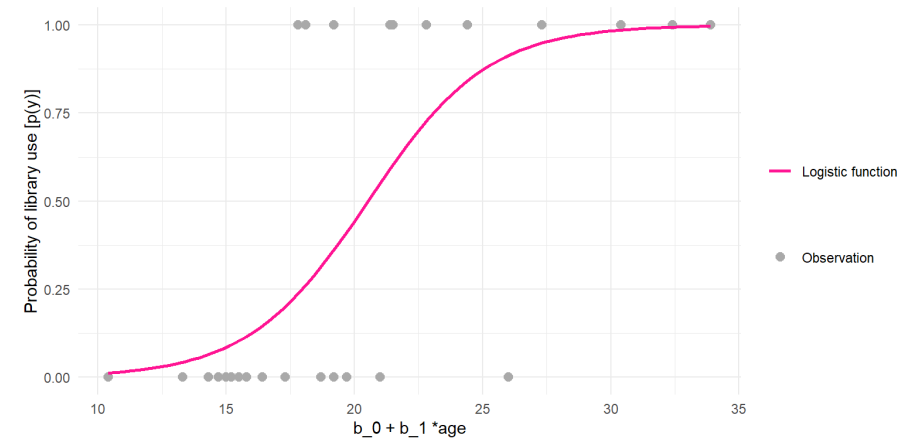
## Where:

- $y$  is the binary outcome variable
- $p(y)$  is the probability of the outcome
- $b_0$  is the  $y$ -intercept
- $x_1, x_2$ , etc are predictors of the outcome
- $b_1, b_2$ , etc are the slopes/coefficients for  $x_1, x_2$ , etc.

# The logistic function with data

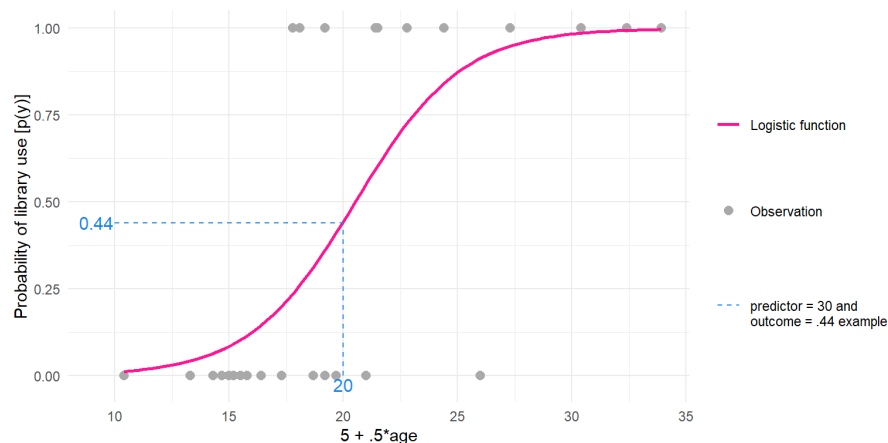
- The data points represent library users ( $y = 1$ ) and non-users ( $y = 0$ )
- The x-axis represents age in years
- The equation would be something like

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



## Example of probability of y for a value of x

- What is the probability of library use for a 30-year old?
  - Substitute 30 into the regression model, starting at this number on the x-axis, trace a straight line up to the logistic function curve and from there look to the y-axis for a value
  - Let's say, for example, that the regression model is  $5 + 0.5 \cdot \text{age}$ , so when age is 30, the value is  $5 + .5 \cdot 30 = 20$
  - For the logistic model represented by this graph, the model would predict a probability of y around .44 or 44%



## Interpreting the probability

- If this were a model predicting library use from age, it would predict a 44% probability of library use for a 30-year-old [ $p(y) = .44$ ]
- Since 44% is lower than a 50% probability of the value of y, the model is predicting that the 30-year-old does not have the outcome
- So, if the outcome is library use, the logistic model would predict this 30-year-old was not a library user

# Let's play logistic regression basketball!