Logistic Regression

The logistic regression model

Jenine Harris
Brown School

Importing and cleaning the data

```
# import the libraries cleaned file
libraries <- read.csv("/Users/harrisj/Box/teaching/Teaching/Fall2020/dat

# change data types
library(package = "tidyverse")
libraries.cleaned <- libraries %>%
  mutate(age = as.numeric(age))
```

Understanding the binary logistic regression statistical model

- Binary logistic regression follows a similar format and process as linear regression but the outcome or dependent variable is *binary* (e.g., library use, smoking status, voting).
- Because the outcome is binary, or categorical consisting of two categories, the model predicts the probability that a person is in one of the categories.
- For example, a logistic regression might predict the probability someone is a smoker or not, is incarcerated or not, votes or not, or any other outcome with two categories.
- In this case we are predicting what is associated with whether or not someone uses the library.
- 49.47% of people in the sample use the library.

The statistical form of the model

- Because the outcome variable is binary, the linear regression model would not work since it requires a continuous outcome.
- However, the linear regression statistical model can be transformed using the *logit transformation* in order to be useful for modeling binary outcomes.
- The statistical model for the logistic:
  - $p(y) = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2)}}$
- Each part:
  - y is the binary outcome variable (e.g., library use)
  - p(y) is the probability of the outcome (e.g., probability of library use)
  - b_0 is the y-intercept
  - x_1 , x_2 , etc are predictors of the outcome (e.g., age, rurality)
  - b_1 , b_2 , etc are the slopes for x_1 x_2

The logistic function

- Examining a graph of the logistic function might help.
- The logistic function has a sigmoid shape that stretches from - ∞ to ∞ on the x-axis and from 0 to 1 on the y-axis.
- The function can take any value along the x-axis and give the corresponding value between 0 and 1 on the y-axis.

Defining the logistic function

- The logistic function is defined:
  - $\sigma(t) = \frac{e^t}{e^{t+1}}$
- Which can be simplified to:
  - $\sigma(t) = \frac{1}{1+e^{-t}}$
- Where t is the value along the x-axis of the function and $\sigma(t)$ is the value of y for a specific value of t, or the probability of y given t.
- In the case of logistic regression, the value of t will be the *right-hand side of the regression model*, which looks something like $\beta_0 + \beta_1 x$ where x is an independent variable, β_1 is the coefficient for that variable, and β_0 is the slope.
- Substituting this regression model for t in the logistic function gives:
  - $p(y) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$
- This is useful because it returns a probability of the outcome happening for any value of an independent predictor or set of independent predictors.

Visualizing the logistic function

- To visualize how it works, add some example data points (i.e., observations) representing the values of a binary outcome variable to the logistic graph.

Interpreting a value on the logistic function

- By starting at 20 on the x-axis, trace a straight line up to the logistic function curve and look to the y-axis for a value.
- For x = 20 in these data, the model would predict a probability of y around .44 or 44%.

Translating the predicted value

- If this were a model predicting library use from age, it would predict a 44% probability of library use for a 20-year-old.
- Since 44% is lower than a 50% probability of the value of y, the model is predicting that the 20-year-old does not have the outcome.
- So, if the outcome is library use, the logistic model would predict this 20-year-old was not a library user.
- While finding the predicted probability of having the outcome (e.g., using the library) is interesting, it is not as useful when there are multiple variables in the model.
- At that point, knowing the influence of each variable on the probability of having the outcome would be better.
- This would be similar to having the coefficients in linear regression that allow interpretation of how much the value of the outcome changes with each increase or decrease in the value of a predictor.
- With transformation of the outcome variable, there is no direct interpretation for how the value of the coefficient of each predictor is related to the value of the outcome.
- Luckily, someone figured this out already and there is a way to transform the model results to get a more interpretable value to describe the relationship between each independent variable and the outcome.

Getting the odds from a model

- First, the logistic model can be transformed to show odds on the left-hand side of the equal sign.
- Odds are related to probability as:
  - $odds = \frac{probability}{1+probability}$
- Substituting the logistic model in to the odds Equation results in:
  - $odds = \frac{\frac{1}{1+e^{-(\beta_0+\beta_1 x)}}}{1+\frac{1}{1+e^{-(\beta_0+\beta_1 x)}}}$
- Which simplifies to:
  - $odds = e^{\beta_0+\beta_1 x}$

Finding the odds ratio from the odds

- Once β_0 and β_1 are estimated, which would make them b_0 and b_1 since they are sample values rather than population values, this equation can be used to determine the odds of the outcome for a given value of the independent variable x.
- To be equivalent to the interpretation of the coefficients in linear regression, however, there is one more step.
- That is, what is the *increase or decrease* in the odds of the outcome with a one-unit increase in x?
- There is a little more math to determine the **odds ratio**:
  - $OR = \frac{e^{b_0+b_1(x+1)}}{e^{b_0+b_1x}} = e^{b_1}$
- This shows that, for every one unit increase in the independent variable x, the odds of the outcome increase or decrease by e^{b_1} .
- Taking e to the power of b_1 is referred to as **exponentiating** b_1 .
- After a model is estimated, the analyst will usually exponentiate the b value(s) in order to report odds ratios describing the relationships between each predictor and the outcome.