Week Six

Last Week

- Contingency Tables
- Simpson's Paradox
- Fisher's Exact Test

This Week: Generalized Linear Models

Today:

- Activity:
  - Generative models for binary data
  - MLE for logistic regression
  - Bayesian estimation for logistic regression
- Thursday: Lab

Next Week: Generalized Linear Models: Binary Data

Logistic Regression

Recall the logistic regression framework, which satisfies the three elements of a GLM (random component, systematic component, link function)

$$
\begin{aligned}
y &\sim Bernoulli(\pi) \\
\pi &= \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \\
\pi &= logit^{-1}(\beta_0 + \beta_1 x)
\end{aligned}
$$

Logistic Regression Activity: Continuous Predictor

We are going to focus on the generative process we assume underlies logistic regression (with a single continuous covariate).

1. Simulate 100 covariate values. This isn't necessary, but assume they are equally spaced between -3 and 3.

2. The β values will change the shape of our expected relationship. Using the following values below, create figures of π vs x.

   i. $\beta_0 = 0$, $\beta_1 = 1$
   ii. $\beta_0 = 0$, $\beta_1 = -1$
   iii. $\beta_0 = 1$, $\beta_1 = 1$
   iv. $\beta_0 =$ -1, $\beta_1 = 1$
   v. $\beta_0 = 0$, $\beta_1 = 3$
   vi. $\beta_0 = 0$, $\beta_1 = -3$

3. Based on the figure provide an intuitive summary of how β_0 and β_1 impact the curve.

- β_0:
- β_1:
- $\beta_0 + \beta_1 x$:

4. Simulate a binary outcome at each x value. Update the figure from part to to include these data points.

5. Use MLE to estimate the coefficients in each of these six settings. Report point estimates and uncertainty. You'll want to use the following formulation `glm(y~x, family = binomial, data =)`.

6. Use Bayesian estimation for the coefficients in each of these six settings. Report point estimates and uncertainty. You'll want to use the following formulation `stan_glm(y~x, family = binomial, refresh = 0, data =)` which is the `rstanarm` package. *Note this has a weakly informative prior distribution embedded in the function.*

7. How do values from parts 6 and 7 compare with each other? Do the values match your expectation?

Logistic Regression Activity: Binary Predictor

Now let's consider a data structure that we've already seen, one binary predictor and one binary covariate.

There are two formulations of this model, the first is known as the reference case model.

$$
\begin{aligned}
y &\sim Bernoulli(\pi) \\
\pi &= \frac{\exp(\beta_0 + \beta_1 I_{x=1})}{1 + \exp(\beta_0 + \beta_1 I_{x=1})} \\
\pi &= logit^{-1}(\beta_0 + \beta_1 I_{x=1})
\end{aligned}
$$

The second is the cell means model

$$
\begin{aligned}
y &\sim Bernoulli(\pi) \\
\pi &= \frac{\exp(\beta_0 I_{x=0} + \beta_1 I_{x=1})}{1 + \exp(\beta_0 I_{x=0} + \beta_1 I_{x=1})} \\
\pi &= logit^{-1}(\beta_0 I_{x=0} + \beta_1 I_{x=1})
\end{aligned}
$$

what is the difference?

Let's repeat similar steps to the continuous setting. Use the cell means formulation for this question.

1. Let there be a total of 100 observations, 50 from $x = 1$ and 50 from $x = 2$

2. The β values will change our expected relationship. Using the following values below, create figures of π vs x.

i. $\beta_0 = 0$, $\beta_1 = 1$
ii. $\beta_0 = 0$, $\beta_1 = -1$
iii. $\beta_0 = 1$, $\beta_1 = 1$
iv. $\beta_0 =$ -1, $\beta_1 = 1$
v. $\beta_0 = 0$, $\beta_1 = 3$
vi. $\beta_0 = 0$, $\beta_1 = -3$

3. Simulate a binary outcome at each x value. Update the figure from part to to include these data points.