MA388 Sabermetrics: Lesson 14

Intro to Logistic Regression

LTC Jim Pleuss

- Where is a pitch with location *plate_x* = -1 and *plate_z* = 2.5?

Measures of Association for Binary Responses

The goal of our analysis is to assess the effect of the catcher on a called strike. For today's example, let's compare Yadier Molina and Buster Posey.

1. Name the response variable. Classify it as categorical or quantitative.

2. Name the explanatory variable. Classify it as categorical or quantitative.

Here is summary data from May 2019 for Molina and Posey.

```r
library(tidyverse)
library(knitr)
library(broom)

# Retrieve pitch level data from May 2019. (See previous lesson.)
pitches <- readRDS("statcast_may_2019.rds")

# Add catcher's name to the pitch data from the MLB master list.
mlbIDs <- baseballr::chadwick_player_lu() |>
  mutate(
    mlb_name = paste(name_first, name_last),
    mlb_id = key_mlbam
  )


pitches <- pitches |>
  left_join(select(mlbIDs, mlb_name, mlb_id),
            by = c("fielder_2" = "mlb_id")) |>
  rename(catcher_name = mlb_name)

# Look only at pitches taken when Molina or Posey are catching.
pitches_taken_subset <- pitches |>
  filter(catcher_name %in% c("Buster Posey", "Yadier Molina"),
         description %in% c("ball", "called_strike"))

# Form a 2x2 table.
pitches_taken_subset |>
  count(catcher_name, description) |>
  pivot_wider(id_cols = description,
              names_from = catcher_name,
              values_from = n) |>
  kable(caption = "Results of taken pitches (May 2019)")
```

Table 1: Results of taken pitches (May 2019)

description	Buster Posey	Yadier Molina
ball	775	1007
called_strike	364	533

3. Calculate the proportion of called strikes for each catcher.

4. Calculate the odds of a called strike for each catcher.

5. Calculate the log odds of a called strike for each catcher.

6. Calculate the odds ratio for a called strike comparing Yadier Molina to Buster Posey.

7. What are the limitations of this analysis in assessing the effect of catcher on called strikes? What are we failing to consider, and how might a model-based approach improve our analysis?

Logistic Regression

Consider the results in Table 1 again. We can apply the following logistic regression model to our data. Let Y_i be a random variable for whether pitch i was a called strike such that $Y_i \sim \text{Bernoulli}(\pi_i)$ and

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{Molina}_i$$

where $\text{Molina}_i = 1$ if Yadier Molina was the catcher and $\text{Molina}_i = 0$ if Buster Posey was the catcher.

Discuss each component of the model above.

- $Y_i \sim \text{Bernoulli}(\pi_i)$

- $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{Molina}_i$

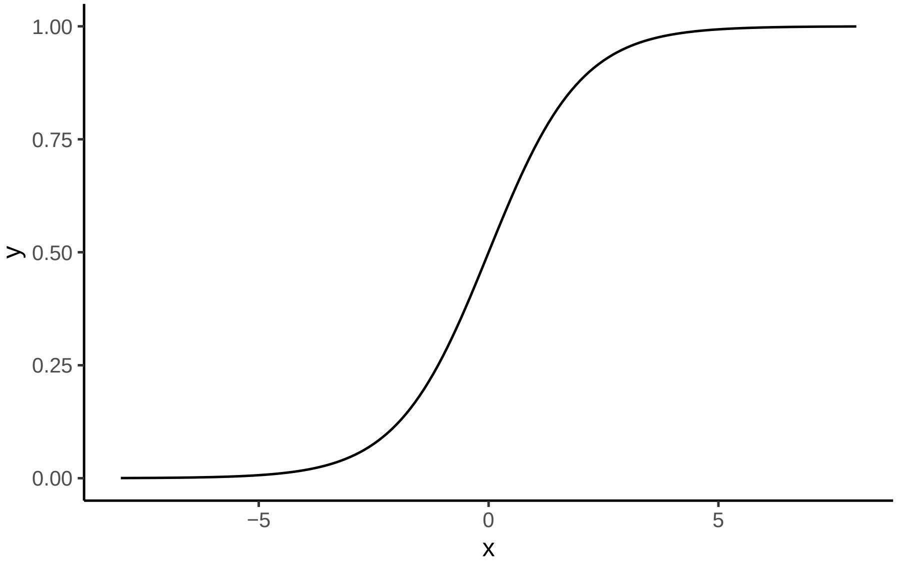
- Interpret β_0.

- Interpret β_1.

Why Log Odds?

$z = \beta_0 + \beta_1 \text{Molina}_i$ Maps the output to a continuous value from $[-\infty, \infty]$. What would a coefficient even mean here?

We can use the sigmoid function, $\sigma(z) = p = \frac{1}{1+e^{-z}}$ to map any continuous value to a number between $[0, 1]$ (i.e. a probability).



Ok, but now we have an even more complicated function with z buried in the denominator. Let's get z by itself.

$$\begin{aligned}
\frac{1}{p} &= 1 + e^{-z} \\
\frac{1}{p} - 1 &= e^{-z} \\
\frac{1-p}{p} &= e^{-z}
\end{aligned}$$

Take the recipricals of both sides and we get:

$$\text{odds}(p) = \frac{p}{1-p} = e^{z}$$

How do we get z by itself?

$$\log(\frac{p}{1-p}) = \log(e^{z}) = z$$

So, in logistic regression, the linear predictor is

$$z = \beta_0 + \beta_1 x.$$

On the probability scale, the model is

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

On the log-odds (logit) scale, the model is

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

Let's fit the model above.

```
catcher_model <-
  pitches_taken_subset |>
  glm(description == "called_strike" ~ catcher_name,
      data = _,
      family = "binomial")

catcher_model |>
  tidy() |>
  kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.7557092	0.0635419	-11.89308	0.0000000
catcher_nameYadier Molina	0.1194997	0.0831071	1.43790	0.1504626

Based on this model, is there evidence Molina has more called strikes in the long term?

Based on this model, is there evidence Molina caused the increase in called strikes (for example, by framing pitches)?

Name other variables to control for to provide stronger evidence Molina caused the increase.

Explain why count might be a confounder of the relationship between catcher and called strikes.

Let's investigate how strong these relationships are:

```
pitches_taken_subset <- pitches_taken_subset |>
  mutate(count = paste(balls,strikes, sep = "-"))

pitches_taken_subset |>
  count(catcher_name, count) |>
  group_by(catcher_name) |>
  mutate(prop = n/sum(n)) |>
  pivot_wider(id_cols = count, names_from = catcher_name, values_from = prop) |>
  kable(digits = 3)
```

count	Buster Posey	Yadier Molina
0-0	0.330	0.347
0-1	0.140	0.123
0-2	0.066	0.060
1-0	0.103	0.111
1-1	0.090	0.084
1-2	0.075	0.066
2-0	0.031	0.039
2-1	0.043	0.051
2-2	0.066	0.059
3-0	0.012	0.016
3-1	0.010	0.023
3-2	0.035	0.022

```
pitches_taken_subset |>
  count(description, count) |>
  group_by(count) |>
  mutate(prop = n/sum(n)) |>
  pivot_wider(id_cols = description, names_from = count, values_from = prop) |>
  kable(digits = 3)
```

description	0-0	0-1	0-2	1-0	1-1	1-2	2-0	2-1	2-2	3-0	3-1	3-2
ball	0.528	0.748	0.922	0.604	0.766	0.882	0.4	0.701	0.855	0.205	0.652	0.865
called_strike	0.472	0.252	0.078	0.396	0.234	0.118	0.6	0.299	0.145	0.795	0.348	0.135

Next, let's adjust for count in our model for called strikes. Here is the updated model.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{Molina}_i + \beta_2 \text{count0-1}_i + ... + \beta_{12} \text{count3-2}_i$$

where countX-Y is an indicator of whether the pitch had X balls and Y strikes.

```
pitches_taken_subset |>
  glm(description == "called_strike" ~ catcher_name + count,
      data = _,
      family = "binomial") |>
  tidy() |>
  kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.1429731	0.0843064	-1.6958749	0.0899096
catcher_nameYadier Molina	0.0525685	0.0884053	0.5946308	0.5520903
count0-1	-0.9728490	0.1400530	-6.9462937	0.0000000
count0-2	-2.3582783	0.2963634	-7.9573883	0.0000000
count1-0	-0.3111663	0.1375763	-2.2617723	0.0237115
count1-1	-1.0737305	0.1690577	-6.3512664	0.0000000
count1-2	-1.8946577	0.2365813	-8.0084856	0.0000000
count2-0	0.5153006	0.2197385	2.3450634	0.0190238
count2-1	-0.7404951	0.2048653	-3.6145463	0.0003009
count2-2	-1.6638610	0.2304900	-7.2187982	0.0000000
count3-0	1.4640085	0.4021148	3.6407728	0.0002718
count3-1	-0.5257107	0.3170033	-1.6583763	0.0972415
count3-2	-1.7377286	0.3466271	-5.0132513	0.0000005

Based on this count-adjusted model, is there evidence Molina has more called strikes in the long term?