Lesson 15 Boardsheet

Kevin Cummiskey

2/19/2020

Question: Do umpires call more strikes on left-handed or right-handed hitter?

Compare unadjusted strike probabilities

Below, we will look at 120,000 pitches throw in May 2016.

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

#Run this code once
#library(mlbgameday)
#gamedat <- get_payload(start = "2016-05-01", end = "2016-05-31")
#pitches <- inner_join(gamedat$pitch, gamedat$atbat,
#                      by = c("num","url"))
#recommend writing to someplace on your harddrive
#pitches %>% write_csv("C:/Users/kevin.cummiskey/Data/pitches.csv")

#Read in the pitches you saved above
pitches <- read_csv(file = "C:/Users/kevin.cummiskey/Data/pitches.csv")

#we only want balls and called strikes
#note the types are different that in our text!!
taken <- pitches %>%
  filter(type %in% c("C","B"))

results = taken %>%
  group_by(type, stand) %>%
  count() %>%
  spread(stand, n)

results$type = c("Ball", "Called Strike")

kable(results)
```

type	L	R
Ball	18141	24050
Called Strike	8864	12703

1. Calculate the probability of a called strike for left and right handed batters.

2. Calculate the odds of a called strike for left and right handed batters.

3. Calculate the odds ratio and log odds ratio for a called strike comparing left and right handed hitters.

git **Note, we could also get the same results using the following model:**

$$\text{Strike}_i \sim \text{Bernoulli}(p_i)$$
$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Stand_i$$

where $Stand_i$ is whether batter i is left or right handed.

```
library(mgcv)

#fit the model above
model_stand <- gam(type == "C" ~ stand, family = binomial, data = taken)
summary(model_stand)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## type == "C" ~ stand
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.71618    0.01296 -55.264  < 2e-16 ***
## standR       0.07788    0.01698   4.587 4.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.000314   Deviance explained = 0.0258%
## UBRE = 0.27949  Scale est. = 1         n = 63758
```

Using the model results, calculate the probability of a strike for a right-handed hitter.

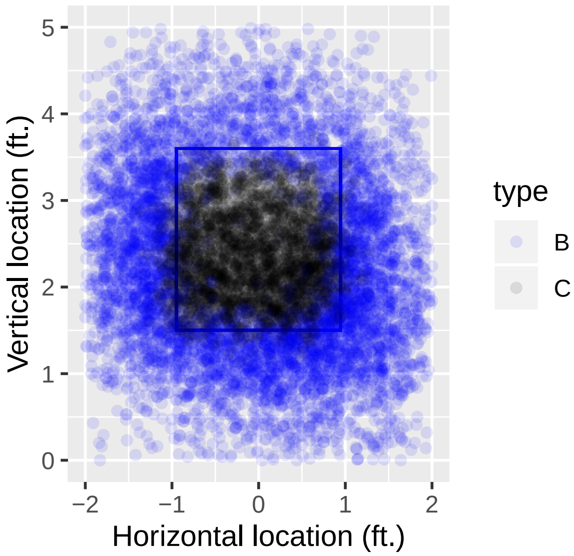
If the goal of your analysis is to see whether umpires are more likely to call *similar* pitches strikes on righties than lefties, how could we improve upon our analysis above?

Adjusting for pitch location

Let's look at the pitch locations.

```
#if there is code you use a lot, you can put it in a separate file and
#source it.  The file kzoneplot contains the code on page 165 for creating
# the strike zone
source("kzoneplot.R")

k_zone_plot %+% sample_n(taken, 10000) +
  aes(color = type) +
  geom_point(alpha = 0.1) +
  scale_color_manual(values = c("blue", "black"))
```



Generalized Additive Models

Last lesson, we discussed models for the probability of a called strike based on the location of the pitch when it crosses the plate.

1. Discuss two limitations of using a linear regression model in this situation.

2. Linear regression is fundamentally flawed as a model for location and strike probability because pitches near the middle of the strike zone are nearly always called strikes and pitches far away from the strike zone are nearly always called balls. A one inch difference in the middle of the plate is not as important as a one inch difference on the corner of the plate. Therefore, we want to consider more general models of the form:

$$\text{Strike}_i \sim \text{Bernoulli}(p_i)$$
$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + f(px_i, pz_i)$$

In R, the `gam` function of the `mgcv` package will fit thin plate regression splines for $f(px, py)$. Deriving how to fit these splines by hand is outside the scope of this course. However, you should understand they represent a flexible way of modeling nonlinear relationships. In future analyses, we will use these models to adjust for pitch location when assessing the relationship between other variables (count, catcher, etc) and called strikes. For more information on thin plate regression splines, see https://www.mailman.columbia.edu/research/population-health-methods/thin-plate-spline-regression.

Fit the model

Let's fit model above to this data.

```
library(mgcv)
strike_mod <- gam(type == "C" ~ s(px, pz),
                  family = binomial,
                  data = taken)
```

How can we use this model?

First, let's look at predicted probabilities of some pitches.

```
# a pitch right down the middle
predict(strike_mod,
        newdata = data.frame(px = 0, pz = 2.5),
        type = "response")
```

```
##         1
## 0.9992354
```

```
# a pitch on the inside corner for a right-handed batter
predict(strike_mod,
        newdata = data.frame(px = -1, pz = 2.5),
        type = "response")
```

```
##         1
## 0.6193452
```

```
# a pitch on the outside corner for a right-handed batter
predict(strike_mod,
        newdata = data.frame(px = 1, pz = 2.5),
        type = "response")
```

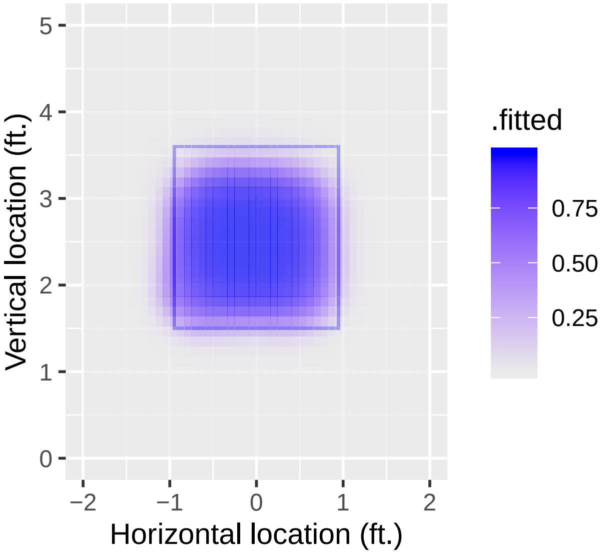
```
##         1
## 0.2661631
```

Visualizing the estimated surface

```
library(modelr)
library(broom)
#create a grid of points in the strike zone
grid <- taken %>%
  data_grid(px = seq_range(px, n = 100),
            pz = seq_range(pz, n = 100))

#get predicted values from the model on the grid
grid_hats <- strike_mod %>%
  augment(type.predict = "response", newdata = grid)

#plot the results
tile_plot <- k_zone_plot %+% grid_hats +
  geom_tile(aes(fill = .fitted),alpha = 0.7) +
  scale_fill_gradient(low = "gray92", high = "blue")
tile_plot
```



Adjusting for pitch location

Now, we can reconsider our question of called strikes for righties and lefties after adjusting for pitch location. Let's fit a model that includes `stand`.

$$\text{Strike}_i \sim \text{Bernoulli}(p_i)$$
$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + f(px_i, pz_i) + \beta_1 stand_i$$

How do we intepret β_1 in this model?

```r
hand_mod <- gam(type == "C" ~ s(px,pz) + stand,
                family = binomial,
                data = taken)
summary(hand_mod)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## type == "C" ~ s(px, pz) + stand
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.25567    0.27849 -15.281  < 2e-16 ***
## standR       0.17822    0.03456   5.158  2.5e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##            edf Ref.df Chi.sq p-value
## s(px,pz) 28.02  28.82  10738  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.734   Deviance explained = 70.1%
## UBRE = -0.61654  Scale est. = 1         n = 63697
```

How do we interpret these results?

Let's look at some pitches.

```
#LEFT CORNER STRIKE
#right-handed batter
predict(hand_mod,
        newdata = data.frame(px = -1, pz = 2.5, stand = "R"),
        type = "response")
```

```
##         1
## 0.6446487
```

```
#left-handed batter
predict(hand_mod,
        newdata = data.frame(px = -1, pz = 2.5, stand = "L"),
        type = "response")
```

```
##         1
## 0.6028551
```

```
#RIGHT CORNER STRIKE
#right-handed batter
predict(hand_mod,
        newdata = data.frame(px = 1, pz = 2.5, stand = "R"),
        type = "response")
```

```
##         1
## 0.2742944
```

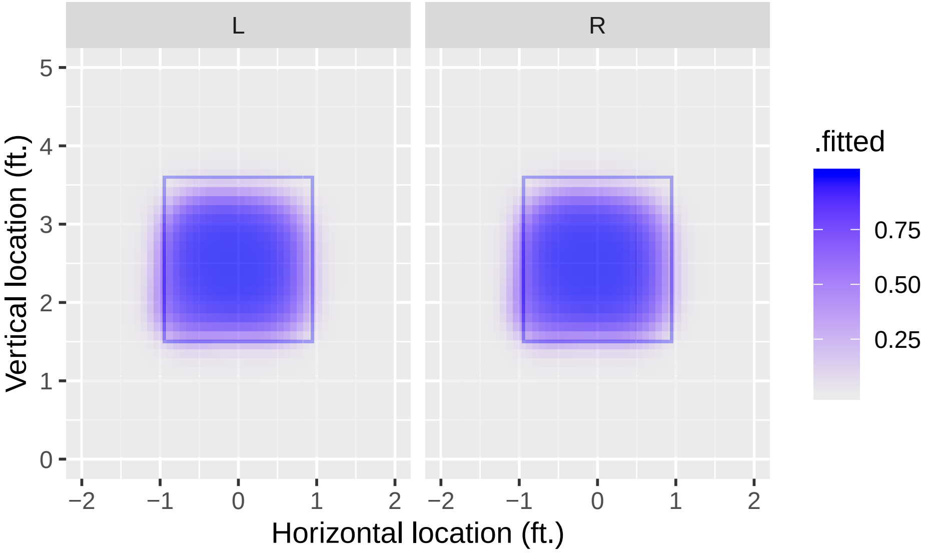
```
#left-handed batter
predict(hand_mod,
        newdata = data.frame(px = 1, pz = 2.5, stand = "L"),
        type = "response")
```

```
##         1
## 0.2402763
```

Let's visualize the results.

```
#create a grid of points
hand_grid <- taken %>%
  data_grid(px = seq_range(px, n = 100),
            pz = seq_range(pz, n = 100),
            stand)
#get model predictions on the grid
hand_grid_hats <- hand_mod %>%
  augment(type.predict = "response",
          newdata = hand_grid)
#plot predictions by handedness
tile_plot %+% hand_grid_hats +
  facet_grid(. ~ stand)
```

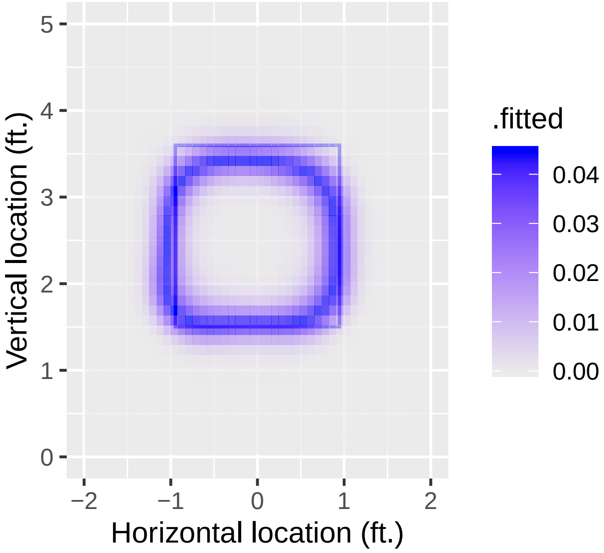
```
## Warning: Removed 15872 rows containing missing values (geom_tile).
```



Let's see where the predictions differ the most.

```
#this code calculates the difference in predictions for
#left and right handed hitters at the same pitch location
diffs <- hand_grid_hats %>%
  group_by(px,pz) %>%
  summarize(N = n(), .fitted = diff(.fitted))
tile_plot %+% diffs
```

```
## Warning: Removed 7936 rows containing missing values (geom_tile).
```


Vertical location (ft.)
Horizontal location (ft.)
.fitted