Lesson_19_Boardsheet

Kevin Cummiskey

3/3/2020

Does catchers' framing of pitches affect umpires' ball and strike calls?

Read in Statcast data using the baseballr package

```
library(tidyverse)
library(devtools)

#Do this once to install baseballr package
#devtools::install_github("BillPetti/baseballr")

# load baseballr package
library(baseballr)

#read in data
pitches <- scrape_statcast_savant(start_date = "2017-05-01",
                                  end_date = "2017-05-2")
#called strikes and balls only
pitches %>% filter(type %in% c("S","B")) -> taken

#add catcher's name to the taken data.frame
#get master ID list
master_id <- read_csv(file = "https://raw.githubusercontent.com/beanumber/baseball_R/master/data/master
#merge with taken
taken %>%
  left_join(select(master_id,mlb_id,mlb_name),
            by = c("fielder_2_1" = "mlb_id")) %>%
  rename(catcher = mlb_name) -> taken
```

How much variability is there in called strike probability by catcher?

First, let's see how much variability there is in called strike probabilities by catcher. Note the catcher's ID is in the `fielder_2_1` variable.

```
library(ggrepel)

# count balls and strikes by catcher
taken %>%
  group_by(catcher) %>%
  count(type) %>%
  pivot_wider(id_cols = catcher, names_from = type, values_from = n) -> catchers

min.pitches = 100
# filter catchers with less than min.pitches
```

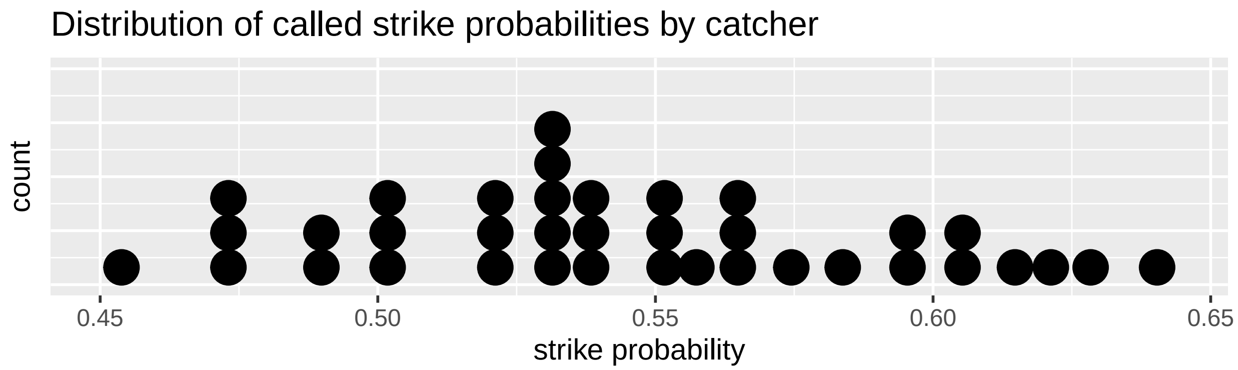
```r
catchers %>% filter(S + B >= min.pitches) -> catchers

# calculate called strike probability and odds
catchers %>%
  mutate(strike.prob = S/(S+B)) %>%
  arrange(desc(strike.prob))-> catchers
catchers %>% head(5)
```

```
## # A tibble: 5 x 4
## # Groups:   catcher [5]
##   catcher             B     S strike.prob
##   <chr>           <int> <int>       <dbl>
## 1 Yasmani Grandal    41    73       0.640
## 2 Austin Barnes      55    93       0.628
## 3 Luke Maile         50    82       0.621
## 4 Sandy Leon         94   150       0.615
## 5 Austin Hedges      50    77       0.606
```

```r
catchers %>% tail(5)
```

```
## # A tibble: 5 x 4
## # Groups:   catcher [5]
##   catcher             B     S strike.prob
##   <chr>           <int> <int>       <dbl>
## 1 J.T. Realmuto     161   153       0.487
## 2 Jonathan Lucroy   109    99       0.476
## 3 Devin Mesoraco     53    48       0.475
## 4 Chris Herrmann     89    79       0.470
## 5 Nick Hundley       71    59       0.454
```

```r
catchers %>%
  ggplot(aes(x = strike.prob)) +
  geom_dotplot() +
  labs(x = "strike probability",
       title = "Distribution of called strike probabilities by catcher") +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())
```



What do these results suggest?

Effect of catcher - Fixed effects

Thus far in your education, when dealing with regression, you've probably only encountered "fixed effects". Here is a fixed effects model we might fit to this data.

$$Strike_i \sim \text{Bernoulli}(p_i)$$

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 catcher_1 + \beta_2 catcher_2 + \cdots + \beta_{m-1} catcher_{m-1}$$

where $catcher_j$ is an indicator (1/0) of whether catcher j is the catcher on pitch i.

How many parameters are in this model? \

What kinds of questions can we answer with this fixed effects model?

Let's fit the model.

```r
model.fixed <- glm(type == "S" ~ as.factor(catcher),
                   family = "binomial",
                   data = taken)
summary(model.fixed)
```

```
##
## Call:
## glm(formula = type == "S" ~ as.factor(catcher), family = "binomial",
##     data = taken)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4301  -1.2278   0.9864   1.1196   1.3072
##
## Coefficients:
##                                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)                            0.0141846  0.1684346   0.084   0.9329
## as.factor(catcher)Austin Barnes        0.5110817  0.2393841   2.135   0.0328 *
## as.factor(catcher)Austin Hedges        0.4175978  0.2477039   1.686   0.0918 .
## as.factor(catcher)Austin Romine       -0.0141846  0.2218187  -0.064   0.9490
## as.factor(catcher)Brian McCann         0.0754275  0.2414144   0.312   0.7547
## as.factor(catcher)Buster Posey         0.1399660  0.2589608   0.540   0.5889
## as.factor(catcher)Caleb Joseph         0.0811255  0.2178558   0.372   0.7096
## as.factor(catcher)Cameron Rupp         0.1234367  0.2090689   0.590   0.5549
## as.factor(catcher)Carlos Ruiz          0.1864861  0.3596948   0.518   0.6041
## as.factor(catcher)Chris Herrmann      -0.1333732  0.2286141  -0.583   0.5596
## as.factor(catcher)Chris Stewart        0.4092990  0.2417784   1.693   0.0905 .
## as.factor(catcher)Derek Norris         0.2490064  0.2176864   1.144   0.2527
## as.factor(catcher)Devin Mesoraco      -0.1132755  0.2609051  -0.434   0.6642
## as.factor(catcher)Drew Butera          0.1955359  0.2419719   0.808   0.4190
## as.factor(catcher)Dustin Garneau       0.3912805  0.4089732   0.957   0.3387
## as.factor(catcher)Elias Diaz           0.1811241  0.2531938   0.715   0.4744
## as.factor(catcher)Evan Gattis          0.3593289  0.2454081   1.464   0.1431
## as.factor(catcher)Geovany Soto       0.2425352  0.2676348   0.906   0.3648
## as.factor(catcher)J.T. Realmuto     -0.0651511  0.2027741  -0.321   0.7480
## as.factor(catcher)James McCann       0.1090480  0.2434264   0.448   0.6542
## as.factor(catcher)Jason Castro       0.2163390  0.2481826   0.872   0.3834
## as.factor(catcher)Jett Bandy         0.2859200  0.2395466   1.194   0.2326
## as.factor(catcher)Jonathan Lucroy   -0.1104127  0.2182786  -0.506   0.6130
## as.factor(catcher)Jose Lobaton       0.1360976  0.2313258   0.588   0.5563
## as.factor(catcher)Juan Graterol     -0.2120104  0.2919953  -0.726   0.4678
## as.factor(catcher)Kevin Plawecki     0.2371298  0.3941502   0.602   0.5474
## as.factor(catcher)Kurt Suzuki        0.2417487  0.2556024   0.946   0.3443
## as.factor(catcher)Kyle Higashioka    0.1035984  0.3272215   0.317   0.7515
## as.factor(catcher)Kyle Schwarber     0.2089589  0.5033589   0.415   0.6780
## as.factor(catcher)Luke Maile         0.4805116  0.2461003   1.953   0.0509 .
## as.factor(catcher)Manny Pina        -0.0444900  0.2422400  -0.184   0.8543
## as.factor(catcher)Martin Maldonado  -0.1152808  0.2624454  -0.439   0.6605
## as.factor(catcher)Mike Zunino        0.1002257  0.2388205   0.420   0.6747
## as.factor(catcher)Nick Hundley      -0.1993271  0.2437291  -0.818   0.4135
## as.factor(catcher)Omar Narvaez      -0.2654991  0.2658915  -0.999   0.3180
## as.factor(catcher)Roberto Perez     -0.0003913  0.2365539  -0.002   0.9987
## as.factor(catcher)Russell Martin     0.3849713  0.2447374   1.573   0.1157
## as.factor(catcher)Salvador Perez     0.1077052  0.2515595   0.428   0.6685
## as.factor(catcher)Sandy Leon         0.4531559  0.2137175   2.120   0.0340 *
## as.factor(catcher)Stephen Vogt       0.1121091  0.2540672   0.441   0.6590
## as.factor(catcher)Stuart Turner     -0.3142892  0.3397164  -0.925   0.3549
## as.factor(catcher)Tony Wolters       0.0293005  0.2680840   0.109   0.9130
## as.factor(catcher)Travis d'Arnaud    0.1434443  0.2257920   0.635   0.5252
## as.factor(catcher)Tucker Barnhart    0.2051782  0.2497792   0.821   0.4114
## as.factor(catcher)Tyler Flowers      0.2515185  0.2584106   0.973   0.3304
## as.factor(catcher)Willson Contreras  0.0599233  0.2117870   0.283   0.7772
## as.factor(catcher)Yadier Molina      0.3239283  0.2192286   1.478   0.1395
## as.factor(catcher)Yan Gomes          0.3744734  0.2651780   1.412   0.1579
## as.factor(catcher)Yasmani Grandal    0.5627027  0.2577966   2.183   0.0291 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8999.9  on 6520  degrees of freedom
## Residual deviance: 8936.2  on 6472  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 9034.2
##
## Number of Fisher Scoring iterations: 4
```

What do we conclude from these results?

Effect of catcher - random effects

The fixed effect model doesn't directly answer the question we are interested in. Instead, we can say the catcher effect is itself a random variable.

$$Strike_i \sim \text{Bernoulli}(p_i)$$
$$\log\left(\frac{p}{1-p}\right) = \beta_0 + catcher_j$$
$$catcher_j \sim \text{Normal}(0, \sigma^2)$$

where $catcher_j$ is the random effect of catcher j.

How many parameters are in this model? \

What kinds of questions can we answer with this model?

Which parameter answers these questions? \

Let's fit the random effects model.

```r
library(lme4)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```r
model.random <- glmer(type == "S" ~ (1|catcher),
                      family = "binomial",
                      data = taken)
summary(model.random)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: type == "S" ~ (1 | catcher)
##    Data: taken
##
##      AIC      BIC   logLik deviance df.resid
##   9000.9   9014.4  -4498.4   8996.9     6519
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.1493 -1.0776  0.8810  0.9272  0.9691
##
## Random effects:
##  Groups  Name        Variance Std.Dev.
##  catcher (Intercept) 0.01074  0.1036
## Number of obs: 6521, groups:  catcher, 49
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.15866    0.02953   5.373 7.75e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What are the estimates of the model parameters?

What do we conclude from this analysis?

Let's look at the catcher random effects.

```
model.random %>%
  ranef() %>%
  as.tibble() %>%
  transmute(id = as.numeric(levels(grp)),
            effect = condval) -> catcher_effects
```

```
## Warning: `as.tibble()` is deprecated, use `as_tibble()` (but mind the new semantics).
## This warning is displayed once per session.
```

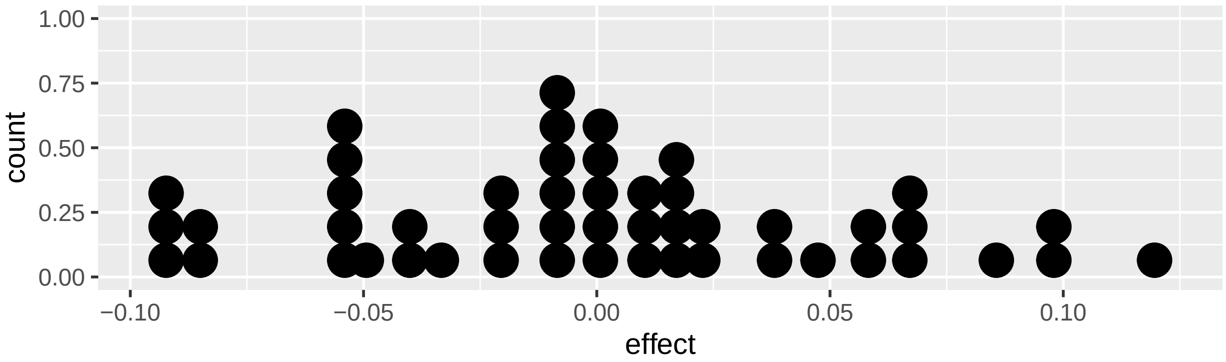
```
## Warning: NAs introduced by coercion
```

```
catcher_effects %>% head(5)
```

```
## # A tibble: 5 x 2
##      id  effect
##   <dbl>   <dbl>
## 1    NA -0.0396
## 2    NA  0.101
## 3    NA  0.0680
## 4    NA -0.0539
## 5    NA -0.0182
```

```
catcher_effects %>%
  ggplot(aes(x = effect)) +
  geom_dotplot()
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



Confounding

OK, so far we've concluded the catcher makes a difference in called strike probability. Let's say we compare a catcher with a called strike probability of 0.5 to another catcher with 0.6. Would you conclude the difference is evidence one catcher is better at framing than the other? Explain.

List variables would you adjust for to make better conclusions about catcher framing.

Adjusting for pitch location.

Write a model for called strike probability with a catcher random effect that adjusts for pitch location.

Instead of fitting the model above directly, we will first fit the pitch location model and then use predictions from it in the random effects model.

```
library(mgcv)
```

```
## Loading required package: nlme

##
## Attaching package: 'nlme'

## The following object is masked from 'package:lme4':
##
##     lmList

## The following object is masked from 'package:dplyr':
##
##     collapse
## This is mgcv 1.8-31. For overview type 'help("mgcv-package")'.
```

```r
library(broom)
#fit pitch location model (plate_x = px, plate_z = pz)
model.location <- gam(type == "S" ~ s(plate_x, plate_z),
                      family = "binomial",
                      data = taken)
#get predictions from location model
taken %>%
  mutate(strike_prob = predict(model.location, newdata = .,
                               type = "response")) -> taken

#fit random effects model adjusting for pitch location
model.random.adj <- glmer(type == "S" ~ strike_prob + (1|catcher),
                          family = "binomial",
                          data = taken)
summary(model.random.adj)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: type == "S" ~ strike_prob + (1 | catcher)
##    Data: taken
##
##      AIC      BIC   logLik deviance df.resid
##   5051.8   5072.1  -2522.9   5045.8     6513
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.2178 -0.4380  0.1848  0.2781  4.3265
##
## Random effects:
##  Groups  Name        Variance Std.Dev.
##  catcher (Intercept) 0.08331  0.2886
## Number of obs: 6516, groups:  catcher, 49
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.89196    0.08394  -34.45   <2e-16 ***
## strike_prob  6.09463    0.13600   44.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## strike_prob -0.741
```

```r
#get random effects
model.random.adj %>%
  ranef() %>%
  as.tibble() %>%
  transmute(id = as.numeric(levels(grp)),
            effect = condval) -> catcher_effects.adj
```

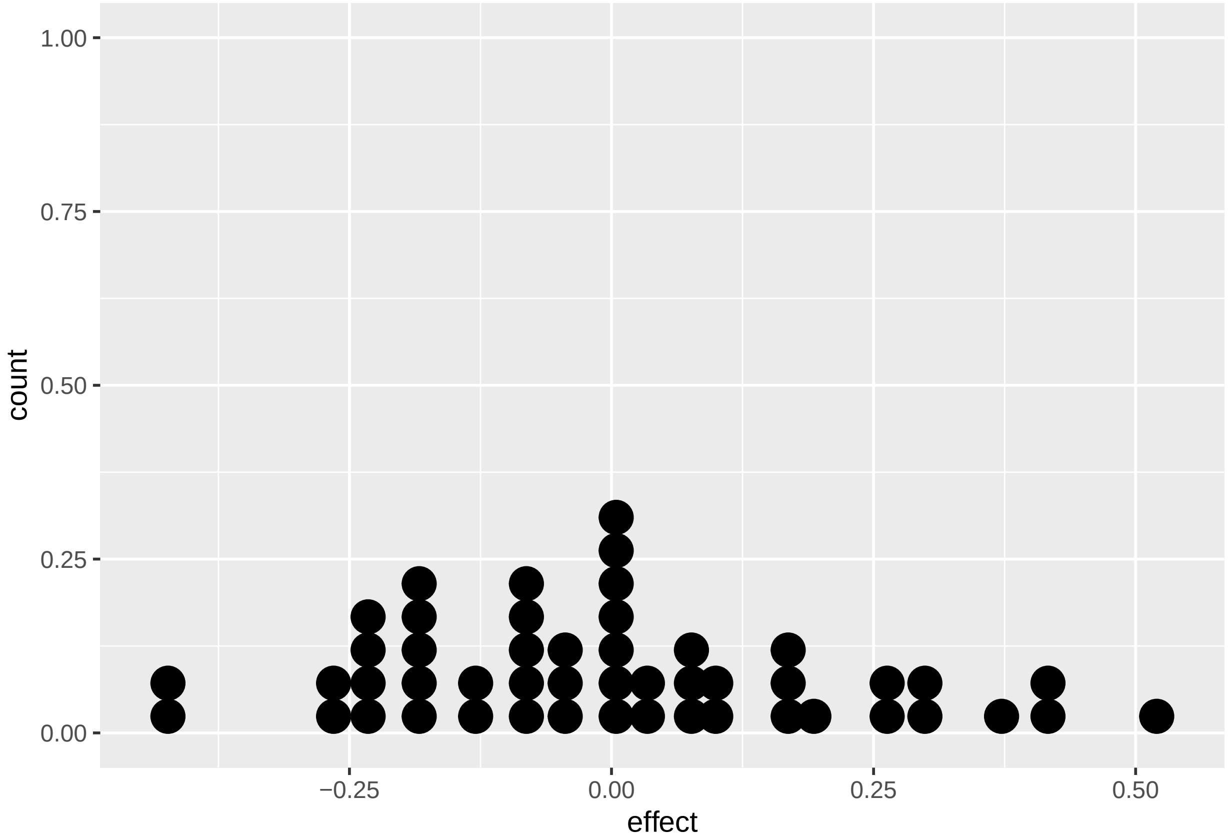
```
## Warning: NAs introduced by coercion
```

```
catcher_effects.adj %>% head(5)
```

```
## # A tibble: 5 x 2
##      id   effect
##   <dbl>    <dbl>
## 1    NA -0.258
## 2    NA  0.426
## 3    NA  0.372
## 4    NA -0.00472
## 5    NA  0.0127
```

```
catcher_effects.adj %>%
  ggplot(aes(x = effect)) +
  geom_dotplot()
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```


count
effect