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 = "2017-05-10")
#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
path = "https://raw.githubusercontent.com/beanumber/baseball_R/master/data/"
file = "masterid.csv"
master_id <- read_csv(paste(path,file, sep = ""))

#merge with taken, note catcher's ID is the fielder_2_1 variable
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.

```
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 = 300
# filter catchers with less than min.pitches
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)
```

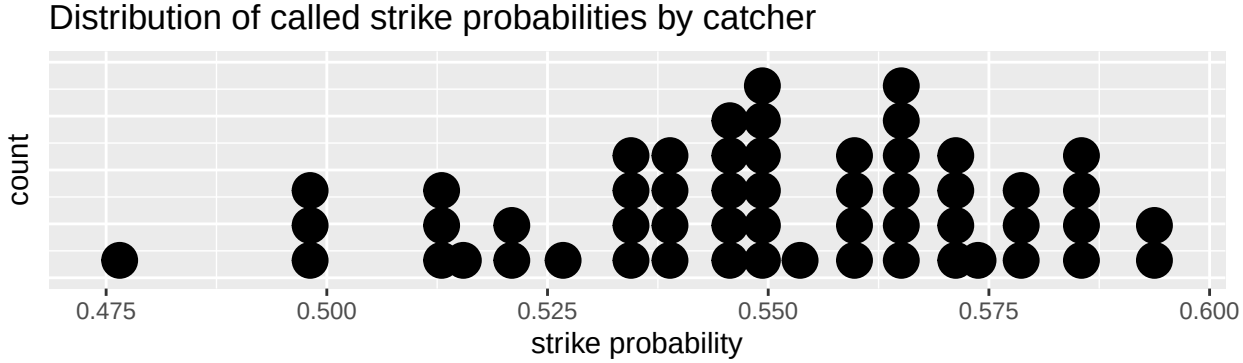
```
## # A tibble: 5 x 4
## # Groups:   catcher [5]
##   catcher               B     S strike.prob
##   <chr>             <int> <int>       <dbl>
## 1 Francisco Cervelli   268   394       0.595
## 2 Yasmani Grandal      298   433       0.592
## 3 Jeff Mathis          130   185       0.587
## 4 Sandy Leon           247   350       0.586
## 5 Salvador Perez       370   524       0.586
```

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

```
## # A tibble: 5 x 4
## # Groups:   catcher [5]
##   catcher            B     S strike.prob
##   <chr>          <int> <int>       <dbl>
## 1 Chris Gimenez    152   159       0.511
## 2 Nick Hundley     180   179       0.499
## 3 Juan Graterol    174   173       0.499
## 4 Andrew Knapp     210   208       0.498
## 5 Devin Mesoraco   290   264       0.477
```

Plot the distribution.

```
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 fixed effects model.

```
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
## -1.345  -1.258   1.034   1.096   1.249
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         0.112478   0.212335   0.530    0.596
## as.factor(catcher)Alex Avila        0.023654   0.241021   0.098    0.922
## as.factor(catcher)Andrew Knapp     -0.122047   0.233786  -0.522    0.602
## as.factor(catcher)Austin Barnes     0.263831   0.244180   1.080    0.280
## as.factor(catcher)Austin Hedges     0.132644   0.224918   0.590    0.555
## as.factor(catcher)Austin Romine     0.085844   0.238388   0.360    0.719
## as.factor(catcher)Brian McCann      0.048452   0.225751   0.215    0.830
## as.factor(catcher)Bruce Maxwell     0.155163   0.233293   0.665    0.506
## as.factor(catcher)Buster Posey      0.225371   0.224934   1.002    0.316
## as.factor(catcher)Caleb Joseph      0.125504   0.220678   0.569    0.570
## as.factor(catcher)Cameron Rupp      0.026205   0.224141   0.117    0.907
## as.factor(catcher)Carlos Ruiz       0.087188   0.234810   0.371    0.710
## as.factor(catcher)Chris Gimenez    -0.067454   0.240737  -0.280    0.779
## as.factor(catcher)Chris Herrmann   -0.053444   0.241272  -0.222    0.825
## as.factor(catcher)Chris Iannetta    0.123244   0.237519   0.519    0.604
## as.factor(catcher)Chris Stewart     0.112533   0.248786   0.452    0.651
## as.factor(catcher)Christian Vazquez 0.152001   0.230639   0.659    0.510
## as.factor(catcher)Derek Norris      0.143455   0.223639   0.641    0.521
## as.factor(catcher)Devin Mesoraco   -0.206410   0.228741  -0.902    0.367
```

```
## as.factor(catcher)Drew Butera         0.184774   0.241853   0.764    0.445
## as.factor(catcher)Dustin Garneau     -0.005309   0.231681  -0.023    0.982
## as.factor(catcher)Elias Diaz          0.106736   0.250528   0.426    0.670
## as.factor(catcher)Eric Fryer          0.138836   0.277186   0.501    0.616
## as.factor(catcher)Evan Gattis         0.154023   0.235926   0.653    0.514
## as.factor(catcher)Francisco Cervelli  0.272886   0.226618   1.204    0.229
## as.factor(catcher)Francisco Pena     -0.255579   0.284576  -0.898    0.369
## as.factor(catcher)Gary Sanchez        0.205561   0.230162   0.893    0.372
## as.factor(catcher)Geovany Soto        0.146384   0.241888   0.605    0.545
## as.factor(catcher)Hector Sanchez     -0.015628   0.278244  -0.056    0.955
## as.factor(catcher)J.T. Realmuto      -0.050603   0.220286  -0.230    0.818
## as.factor(catcher)James McCann        0.122898   0.227973   0.539    0.590
## as.factor(catcher)Jason Castro        0.062726   0.227175   0.276    0.782
## as.factor(catcher)Jeff Mathis         0.240343   0.241213   0.996    0.319
## as.factor(catcher)Jesus Sucre         0.167691   0.235794   0.711    0.477
## as.factor(catcher)Jett Bandy          0.041673   0.229032   0.182    0.856
## as.factor(catcher)Jonathan Lucroy    -0.024371   0.223891  -0.109    0.913
## as.factor(catcher)Jose Lobaton        0.069844   0.233956   0.299    0.765
## as.factor(catcher)Josh Phegley       -0.112478   0.277186  -0.406    0.685
## as.factor(catcher)Juan Graterol      -0.118242   0.237936  -0.497    0.619
## as.factor(catcher)Kevin Plawecki      0.108894   0.247038   0.441    0.659
## as.factor(catcher)Kurt Suzuki         0.032004   0.233991   0.137    0.891
## as.factor(catcher)Kyle Higashioka    -0.101428   0.259203  -0.391    0.696
## as.factor(catcher)Kyle Schwarber      0.110666   0.519698   0.213    0.831
## as.factor(catcher)Luis Torrens       -0.014360   0.254443  -0.056    0.955
## as.factor(catcher)Luke Maile          0.207292   0.231140   0.897    0.370
## as.factor(catcher)Manny Pina          0.080588   0.231371   0.348    0.728
## as.factor(catcher)Martin Maldonado   -0.032847   0.224118  -0.147    0.883
## as.factor(catcher)Matt Wieters        0.081194   0.223793   0.363    0.717
## as.factor(catcher)Miguel Montero      0.202291   0.234704   0.862    0.389
## as.factor(catcher)Mike Ohlman         0.170385   0.280731   0.607    0.544
## as.factor(catcher)Mike Zunino         0.087582   0.239358   0.366    0.714
## as.factor(catcher)Nick Hundley       -0.118049   0.237125  -0.498    0.619
## as.factor(catcher)Omar Narvaez       -0.055966   0.229402  -0.244    0.807
## as.factor(catcher)Rene Rivera         0.179611   0.226403   0.793    0.428
## as.factor(catcher)Roberto Perez       0.069844   0.230341   0.303    0.762
## as.factor(catcher)Robinson Chirinos   0.040827   0.236811   0.172    0.863
## as.factor(catcher)Russell Martin      0.090463   0.229465   0.394    0.693
## as.factor(catcher)Ryan Hanigan        0.074506   0.231592   0.322    0.748
## as.factor(catcher)Salvador Perez      0.235511   0.222929   1.056    0.291
## as.factor(catcher)Sandy Leon          0.236067   0.228017   1.035    0.301
## as.factor(catcher)Stephen Vogt        0.038206   0.226986   0.168    0.866
## as.factor(catcher)Stuart Turner      -0.279532   0.317852  -0.879    0.379
## as.factor(catcher)Tony Wolters       -0.068993   0.297632  -0.232    0.817
## as.factor(catcher)Travis d'Arnaud     0.031561   0.267144   0.118    0.906
## as.factor(catcher)Tucker Barnhart     0.102792   0.232251   0.443    0.658
## as.factor(catcher)Tuffy Gosewisch     0.153225   0.240599   0.637    0.524
## as.factor(catcher)Tyler Flowers       0.078577   0.227928   0.345    0.730
## as.factor(catcher)Willson Contreras   0.019128   0.221835   0.086    0.931
## as.factor(catcher)Yadier Molina       0.181239   0.222064   0.816    0.414
## as.factor(catcher)Yan Gomes           0.172151   0.228493   0.753    0.451
## as.factor(catcher)Yasmani Grandal     0.261166   0.225280   1.159    0.246
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 45363  on 32955  degrees of freedom
## Residual deviance: 45266  on 32885  degrees of freedom
##   (42 observations deleted due to missingness)
## AIC: 45408
##
## 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.

```
library(lme4)

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
##  45359.8  45376.6 -22677.9  45355.8    32954
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.1435 -1.1017  0.8806  0.9069  0.9527
```

```
##
## Random effects:
##  Groups  Name        Variance Std.Dev.
##  catcher (Intercept) 0.003782 0.0615
## Number of obs: 32956, groups:  catcher, 71
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.19667    0.01373   14.33   <2e-16 ***
## ---
## 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 = levels(grp),
            effect = condval) %>%
  arrange(desc(effect)) -> catcher_effects
catcher_effects %>% head(5)
```

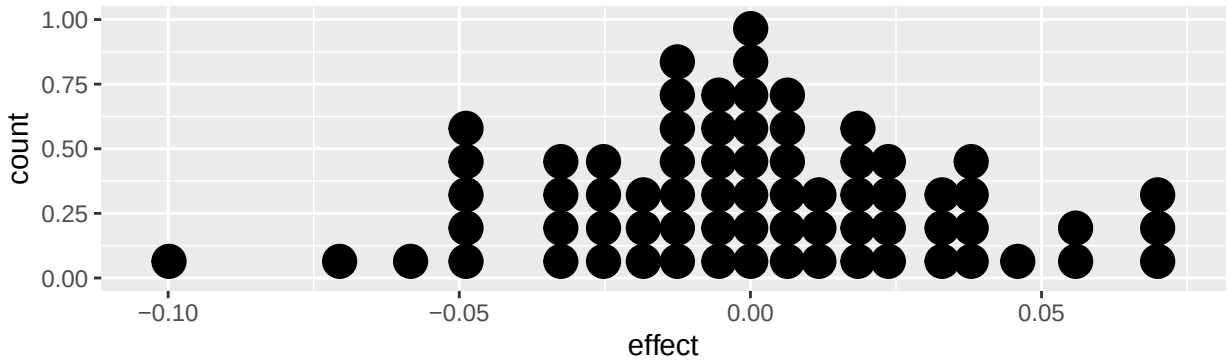
```
## # A tibble: 5 x 2
##   id                 effect
##   <chr>               <dbl>
## 1 Robinson Chirinos  0.0714
## 2 Francisco Cervelli 0.0712
## 3 Yan Gomes          0.0684
## 4 Chris Gimenez      0.0577
## 5 Miguel Montero     0.0540
```

```
catcher_effects %>% tail(5)
```

```
## # A tibble: 5 x 2
##   id             effect
##   <chr>           <dbl>
## 1 Carlos Ruiz   -0.0499
## 2 Austin Hedges -0.0512
## 3 Andrew Knapp  -0.0584
## 4 Roberto Perez -0.0706
## 5 Luis Torrens  -0.0999
```

```
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 adjusting for strike probability with a catcher random effect.

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)
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
##  24378.5  24403.7 -12186.2  24372.5    32910
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.3864 -0.4045  0.1895  0.2683  4.3786
##
## Random effects:
##  Groups  Name        Variance Std.Dev.
##  catcher (Intercept) 0.04138  0.2034
## Number of obs: 32913, groups:  catcher, 71
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.91745    0.04115   -70.9   <2e-16 ***
## strike_prob  6.11920    0.05939   103.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## strike_prob -0.673
```

```
#get random effects
model.random.adj %>%
  ranef() %>%
  as_tibble() %>%
  transmute(id = levels(grp),
            effect = condval) %>%
  arrange(desc(effect)) -> catcher_effects.adj
catcher_effects.adj %>% head(5)
```

```
## # A tibble: 5 x 2
##   id            effect
##   <chr>          <dbl>
## 1 Chris Stewart  0.339
## 2 Mike Zunino    0.285
## 3 Jason Castro   0.257
## 4 Jeff Mathis    0.249
## 5 Sandy Leon     0.247
```

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

```
## # A tibble: 5 x 2
##   id              effect
```

```
##   <chr>            <dbl>
## 1 Manny Pina      -0.222
## 2 Drew Butera     -0.285
## 3 Andrew Knapp    -0.309
## 4 Tyler Flowers   -0.400
## 5 Tuffy Gosewisch -0.445
```

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