SASL Project

Samuel Brown

2022-09-19

Logistic regression(or knn classification) using sz_top/sz_bottom, plate_x, plate_z, to make expected called strikes. Maybe try to make separate regressions for different pitch types and pitcher/batter handedness, also counts.

```
# Make the logit model
logreg <- glm(called_strike ~ plate_x + plate_z, family = binomial, data = all_pitches)
summary(logreg)
```

```
##
## Call:
## glm(formula = called_strike ~ plate_x + plate_z, family = binomial,
##     data = all_pitches)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8116  -0.7880  -0.7676   1.6199   1.6620
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.068451   0.012680 -84.265  < 2e-16 ***
## plate_x     -0.038226   0.005534  -6.907 4.94e-12 ***
## plate_z      0.013304   0.004961   2.682  0.00732 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 236248  on 205599  degrees of freedom
## Residual deviance: 236186  on 205597  degrees of freedom
## AIC: 236192
##
## Number of Fisher Scoring iterations: 4
```

```
names(logreg)
```

```
##  [1] "coefficients"      "residuals"         "fitted.values"
```

```
##  [4] "effects"           "R"                 "rank"
##  [7] "qr"                "family"            "linear.predictors"
## [10] "deviance"          "aic"               "null.deviance"
## [13] "iter"              "weights"           "prior.weights"
## [16] "df.residual"       "df.null"           "y"
## [19] "converged"         "boundary"          "model"
## [22] "call"              "formula"           "terms"
## [25] "data"              "offset"            "control"
## [28] "method"            "contrasts"         "xlevels"
```

```
logreg$contrasts
```

```
## NULL
```

```
table(all_pitches$called_strike, logreg$fitted.values > .5)
```

```
##
##          FALSE
##   FALSE 151853
##   TRUE   53747
```

```
all_pitches$log_prob <- logreg$fitted.values
```

```
# Make a GAM
gam <- gam(called_strike ~ s(plate_x) + s(plate_z), family = binomial, data = all_pitches)
summary(gam)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## called_strike ~ s(plate_x) + s(plate_z)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.12034    0.08165  -50.47   <2e-16 ***
## ---
## 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(plate_x) 8.561  8.917  26606  <2e-16 ***
## s(plate_z) 8.801  8.980  24706  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.638   Deviance explained = 61.3%
## UBRE = -0.55458  Scale est. = 1         n = 205600
```

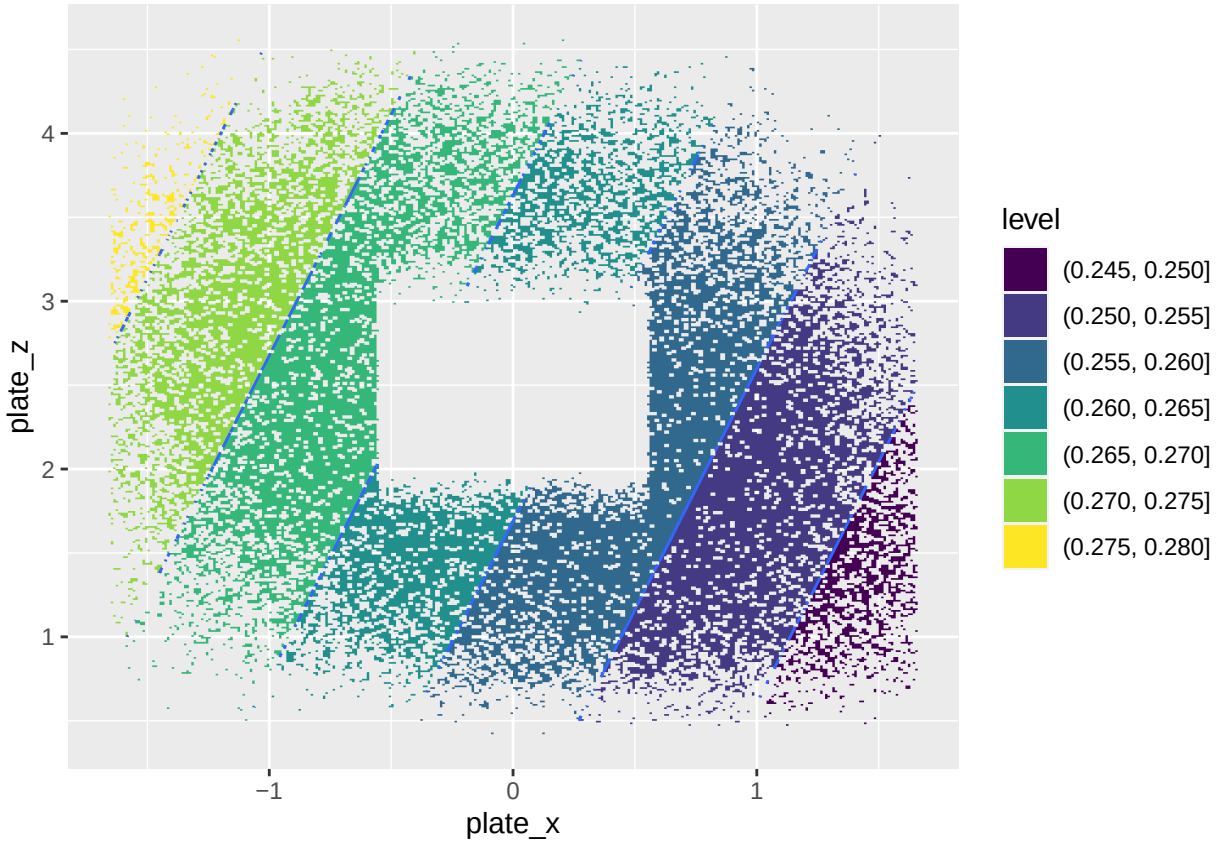
```
table(all_pitches$called_strike, gam$fitted.values > .5)
```

```
##
##           FALSE   TRUE
##   FALSE 141098  10755
##   TRUE    9670  44077
```

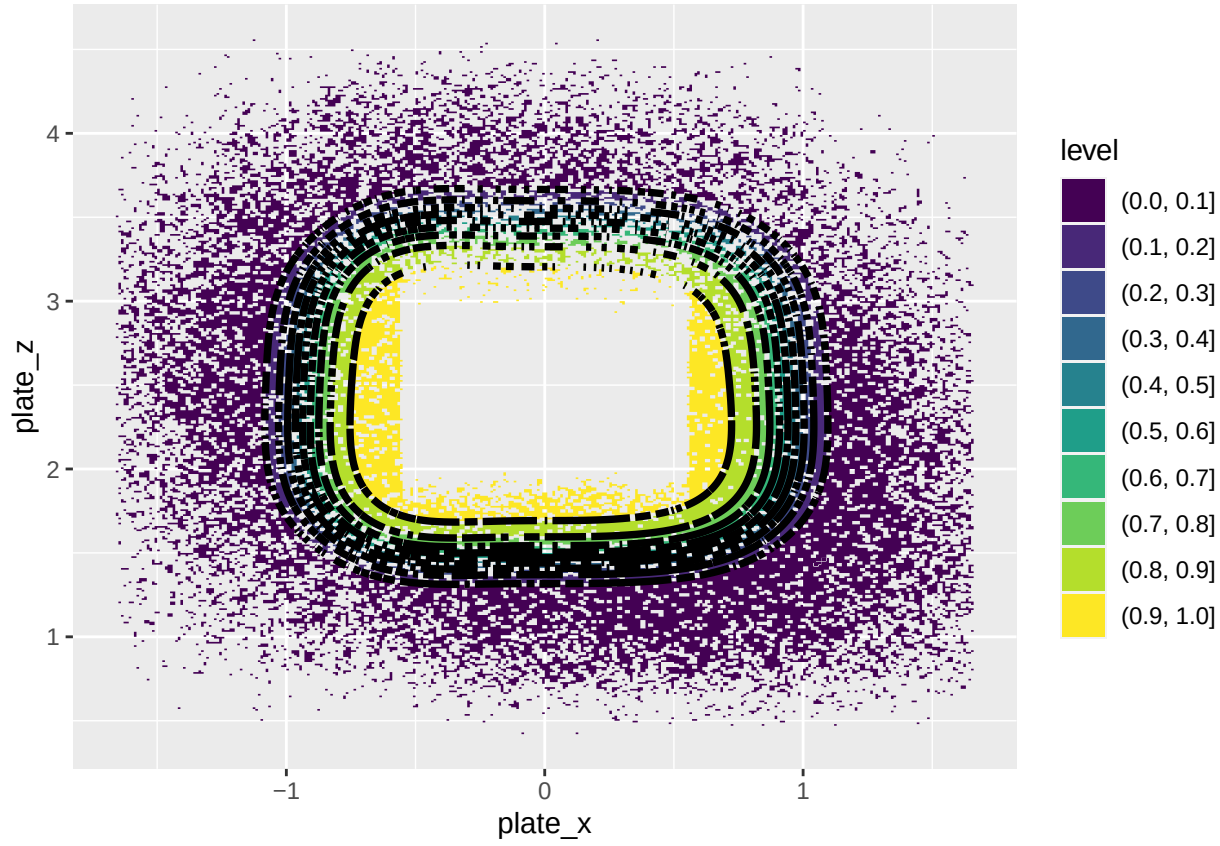
```
all_pitches$gam_prob <- gam$fitted.values
```

```
# KNN Classification
```

```
### Make a strike zone plot of each
# Logit
ggplot(all_pitches, aes(x = plate_x, y = plate_z, z = log_prob)) +
  geom_contour_filled() +
  geom_contour()
```



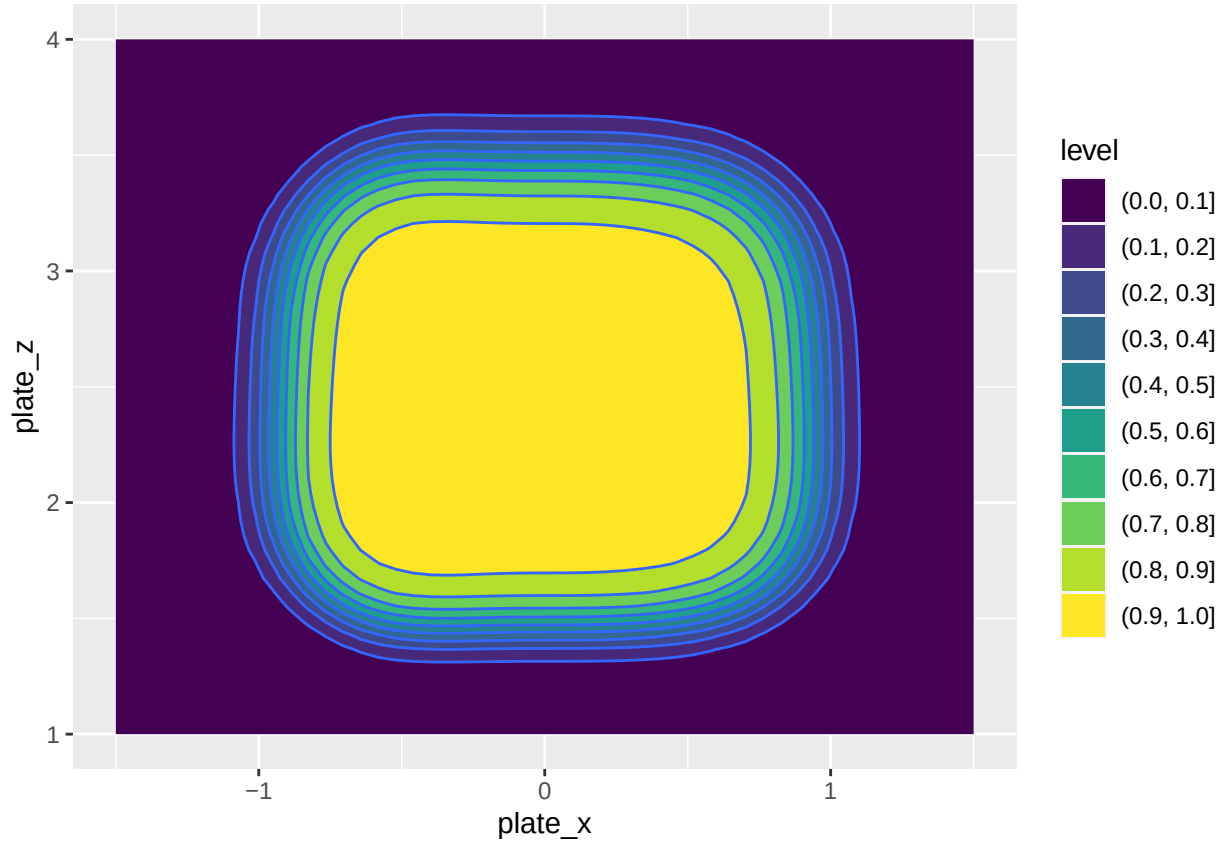

```
# GAM
ggplot(all_pitches, aes(x = plate_x, y = plate_z, z = gam_prob)) +
  geom_contour_filled() +
  geom_contour(size = 1.3, color = 'black')
```




```
# Fancy grid
grid_predict <- function(fit){
  grid <- expand.grid(plate_x = seq(-1.5, 1.5, length=50),
                      plate_z = seq(1, 4, length=50))
  grid$lp <- predict(fit, grid, type = "response")
  grid
}

test <- grid_predict(gam)

ggplot(test, aes(x = plate_x, y = plate_z, z = lp)) +
  geom_contour_filled() +
  geom_contour()
```

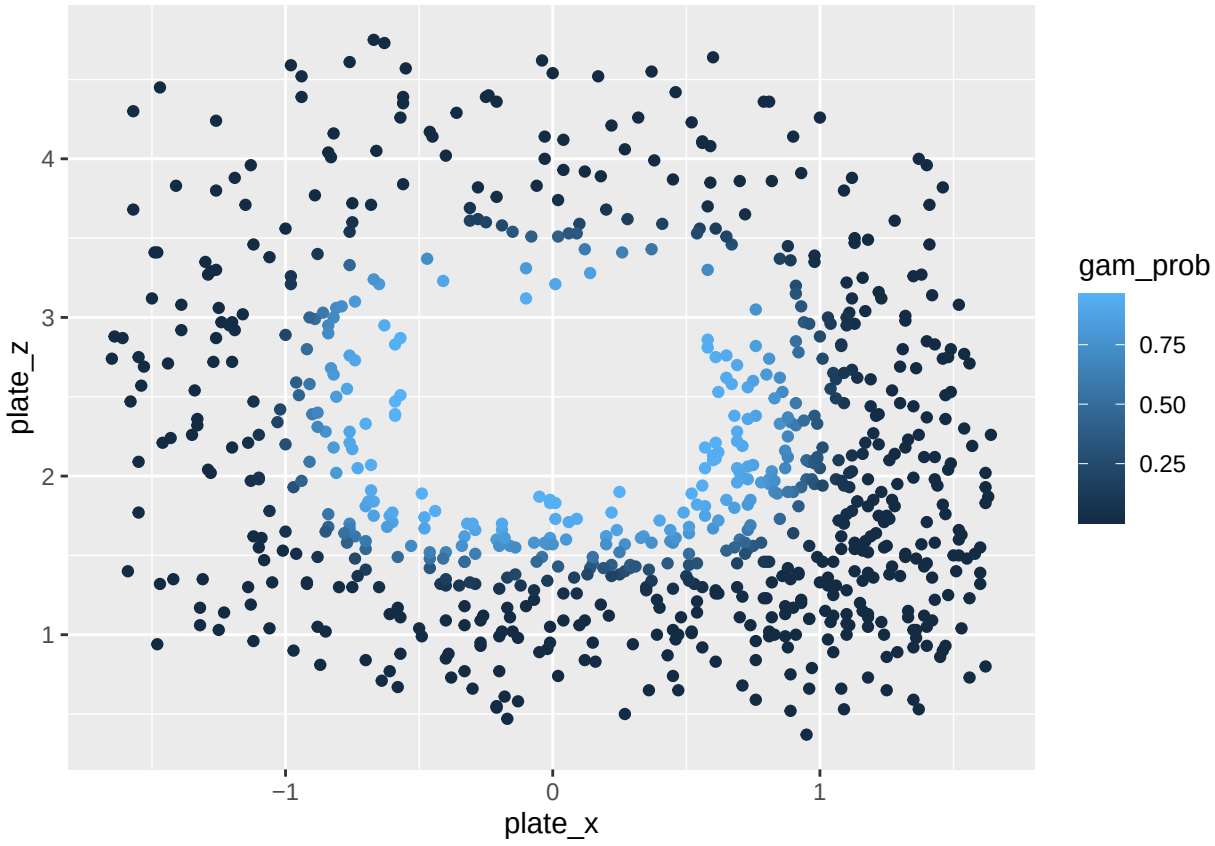



```
# Hitter table
all_pitches %>%
  group_by(player_name) %>%
  summarize(exp_strikes_aa = sum(gam_prob) - sum(called_strike)) %>%
  mutate(exp_strikes_aa = round(exp_strikes_aa, 2)) %>%
  arrange(exp_strikes_aa)
```

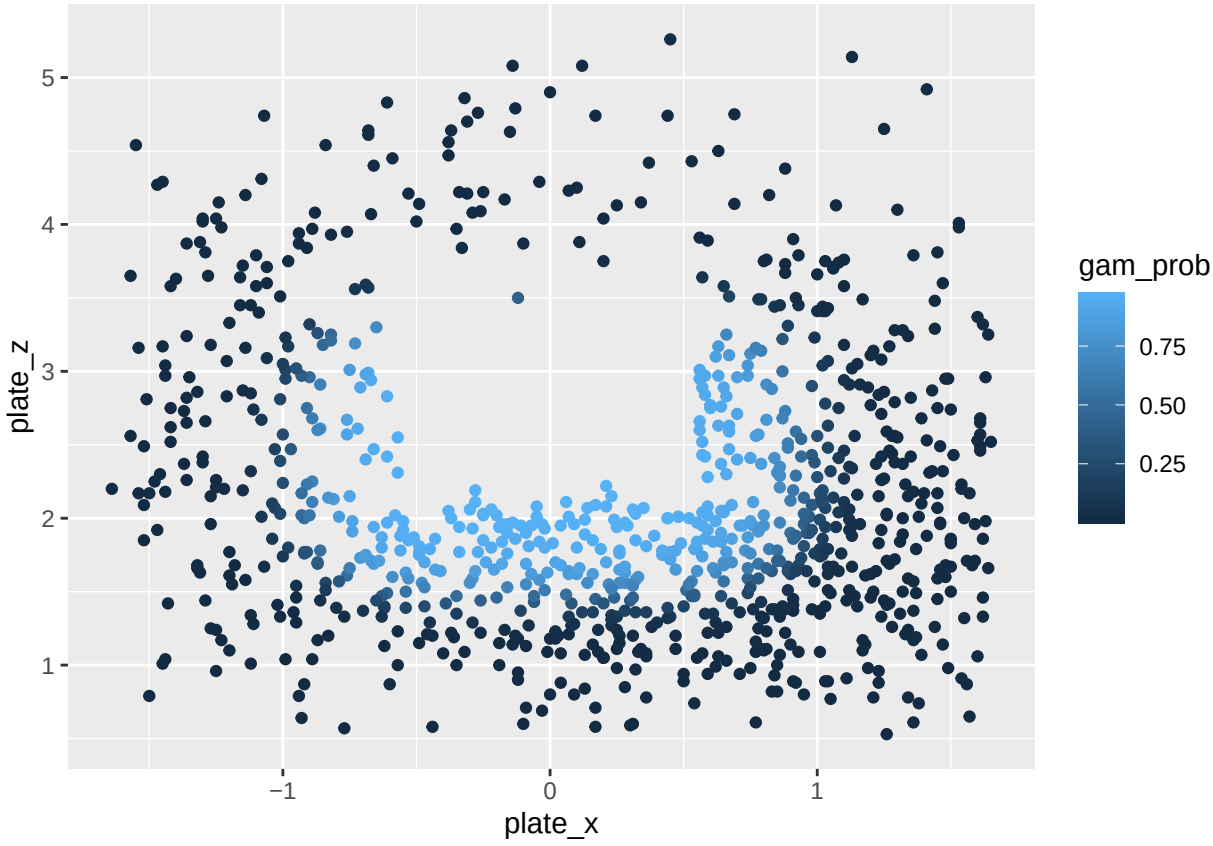
```
## # A tibble: 650 x 2
##    player_name      exp_strikes_aa
##    <chr>                     <dbl>
##  1 Bogaerts, Xander          -31.0
##  2 Trout, Mike               -26.0
##  3 Semien, Marcus            -24.4
##  4 Flores, Wilmer            -22.4
##  5 Iglesias, José            -18.6
##  6 Kwan, Steven              -18.1
##  7 Castillo, Diego           -17.8
##  8 Witt Jr., Bobby           -17.8
##  9 Rodríguez, Julio          -16.8
## 10 Marte, Starling           -15.9
## # ... with 640 more rows
```

Find expected strikes above average in relation to different characteristics?

```
# Plot Bogaerts' heat map with line for 50% called
all_pitches %>%
  filter(player_name == "Bogaerts, Xander") %>%
  ggplot(aes(x = plate_x, y = plate_z, color = gam_prob)) +
  geom_point()
```




```
all_pitches %>%
  filter(player_name == "Judge, Aaron") %>%
  ggplot(aes(x = plate_x, y = plate_z, color = gam_prob)) +
  geom_point()
```



Sources

1. https://baseballsavant.mlb.com/statcast_search?hfPT=&hfAB=&hfGT=R%7C&hfPR=ball%7Ccalled%5C.%5C.strike%7C&hfZ=&hfVenue=&hfBBL=&hfNewZones=11%7C12%7C13%7C14%7C16%7C17%7C18%7C19%7C21%7C22%7C23%7C24%7C26%7C27%7C28%7C29%7C&hfPull=&hfC=&hfSea=2022%7C&hfSit=&player_type=batter&hfOuts=&hfOpponent=&pitcher_throws=&batter_stands=&hfSA=&game_date_gt=&game_date_lt=&hfMo=&hfTeam=&home_road=&hfRO=&position=&hfInfield=&hfOutfield=&hfInn=&hfBBT=&hfFlag=&metric_1=&group_by=name&min_pitches=0&min_results=0&min_pas=0&sort_col=pitches&player_event_sort=api_p_release_speed&sort_order=desc#results