Statcast Analysis - MA678 Midterm Project Report

Andrew Sisitzky

10 December 2021

ABSTRACT

In 2015, Major League Baseball (MLB) implemented the Statcast system with the purpose of measuring previously unquantifiable aspects of the game. A significant and unprecedented aspect of Statcast is its ability to measure statistics related to the quality of contact a batter makes with the baseball. In recent seasons, however, reports have surfaced about the MLB making changes to the composition of the baseballs used in their games. With the shifting weight, size, and density of the baseball over the previous six years of competetion, it would be helpful to account for these year to year changes when using quality of contact Statcast stats to predict traditional baseball statistics. In this report, I will use a multilevel model to predict BABIP (batting average on balls in play) using Statcast data from 2015 to 2021. To address the issue of the changes in the ball, I will be grouping by year and using varying slopes and varying intercepts for all statistics that could be affected by these changes. The model found that Statcast metrics can adequately predict BABIP and that the year to year changes in the balls can be accounted for using a mixed effects model.

INTRODUCTION

BABIP, or batting average on balls in play, is a statistic in baseball which measures a player's batting average exclusively on balls hit into the field of play. Historically, many have considered luck to play a large role in this stat; once the ball leaves the player's bat, numerous factors which the batter has no control over come into play, such as defensive positioning and the range/ability of the opposing fielders. Both of these factors, among others, can inflate or deflate a batter's BABIP. However, it has long been asserted that the quality of contact by the batter and their running ability could be effective in predicting this stat, although prior to the 2015 season, there was no large scale effort to quantify and/or track these statistics. With the implementation of the Statcast system, such prediction became possible.

METHOD

Data Processing

The data that I used for this project was gathered from Baseball Savant's custom leaderboards. I made the decision to not include the data from the shortened 2020 season as the distribution of some statistics during this season appeared to differ greatly from the other years I observed. Therefore, the final dataset contained 835 observations of 38 variables, 4 of which would remain in my final model. The variables included in the BABIP model are listed in the table below.

Statistic	Description
exit_velocity_avg	How fast, in miles per hour, a ball was hit by a batter.
launch_angle_avg	How high/low, in degrees, a ball was hit by a batter.
sweet_spot_percent	How often a player produces a batted-ball event in the launch angle sweet-spot zone of 8-32 degrees.
sprint_speed	A measurement of a player's top running speed, expressed in "feet per second in a player's fastest one-second window."

Before getting further into the model itself, I will take a step back to show my exploratory data analysis and highlight how each variable and the model itself was chosen.

Exploratory Data Analysis

In an effort to identify correlation between BABIP and the statcast metrics, I created numerous scatterplots of this variable and the quality of contact statistics. To condense the data, I took the logs of both variables for each plot.

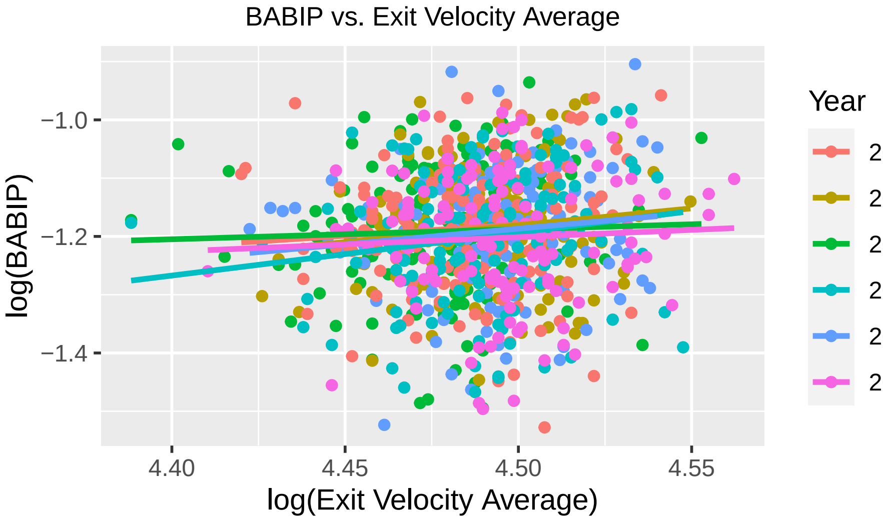


Figure 1

As seen above in figure 1, there appears to be a slightly positive correlation between average exit velocity and BABIP. While this positive correlation holds true for all groups, the slopes and intercepts for each group appear to vary slightly.

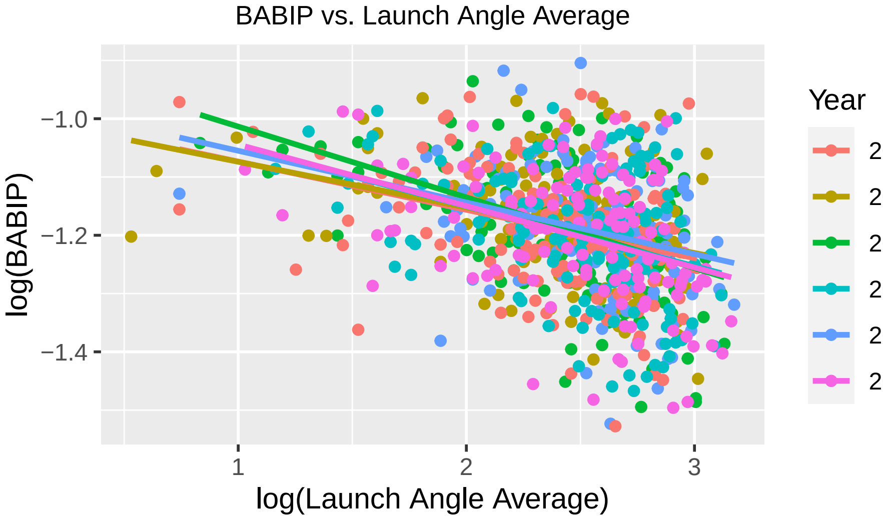


Figure 2

In figure 2, we see a plot of launch angle vs BABIP on the log scale. Similar to the exit velocity plot, slight variation in the slopes and intercepts of each group can be seen. In this case we see a rather clear negative correlation between the two variables. This is to be expected as players with higher average launch angles tend to hit more fly balls, which have a much lower probability of leading to hits as ground balls or line drives.

Not displayed among these graphs are the final two variables, sweet spot percentage and sprint speed. Sweet spot percentage, which is calculated using launch angle, shows a similar trend of varying slopes and intercepts and this will be accounted for in the model. Sprint speed, which has been determined to be a viable predictor of BABIP, will not have a varying slope or intercept in the model as it is not a quality of contact statistic. The year to year changes in the construction of the baseball would not have an effect on the measure of sprint speed, therefore I will not consider random effects for this variable.

Model Fitting

As previously mentioned, the model that I chose to fit on the data was a multilevel model. From my exploratory data analysis, I have concluded that varying slopes and varying intercepts for the quality of contact statistics is logical as there is slight variation in the data among the years considered. The final model is printed below.

```r
model <- lmer(b_babip ~ sprint_speed + exit_velocity_avg + launch_angle_avg +
             sweet_spot_percent + (1+exit_velocity_avg|year) + (1+launch_angle_avg|year) +
             (1+sweet_spot_percent|year), data = df_no_2020, REML = FALSE)
```

After fitting this model, the following fixed effects were returned. ($\alpha = 0.05$)

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-0.2306	0.045200	27.16	-5.106	2.25e-05 ***
sprint_speed	0.0074	0.000599	831.8	12.398	< 2e-16 ***
exit_velocity_avg	0.0025	0.000420	18.93	6.058	8.06e-06 ***
launch_angle_avg	-0.0045	0.000219	27.36	-20.688	< 2e-16 ***
sweet_spot_percent	0.0049	0.000330	19.71	14.726	4.26e-12 ***

RESULTS

Model Coefficients

The model returned the following coefficients for each year.

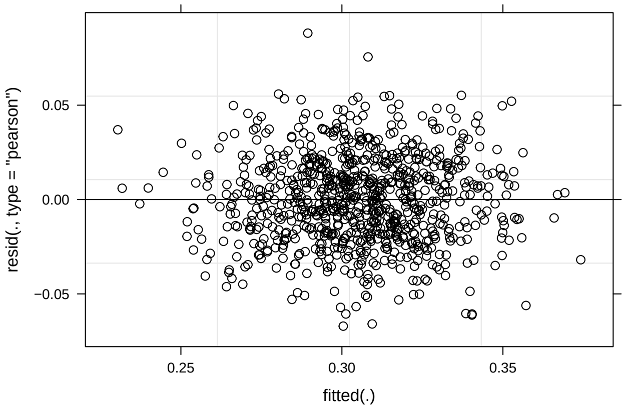
	(Intercept)	sprint_speed	exit_velocity	launch_angle	sweet_spot
2015	-0.1789903	0.00742076	0.002375287	-0.004519693	0.004593839
2016	-0.1658082	0.00742076	0.002331760	-0.004523871	0.004699817
2017	-0.1911803	0.00742076	0.002415534	-0.004560878	0.004693095
2018	-0.2844739	0.00742076	0.002723599	-0.004493228	0.005151112
2019	-0.2309935	0.00742076	0.002546998	-0.004509899	0.004945222
2021	-0.3324017	0.00742076	0.002881859	-0.004519432	0.005058736

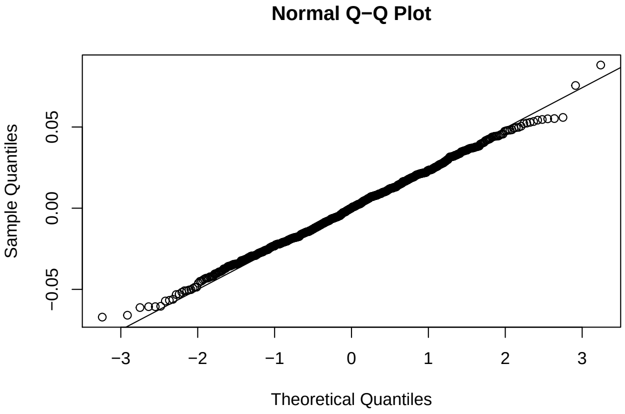
Analyzing the above coefficients, we first notice a large shift in the data beginning in 2017 and high variation in the intercept and exit velocity continues on after this point. In 2017, a record number of home runs were hit with 6,105 being hit across all 30 teams. For the previous 20 years, the number of home runs hit across the league had hovered between 4000 and the low 5000s. This led to speculation that the balls were being altered to allow for more offense in the game, specifically more home runs. The varying numbers for exit velocity suggest that the baseballs may have been being altered to affect the velocity off the bat (matching the speculation of many). The correlation between exit velocity and BABIP is likely why we can see the effects of these changes in this multilevel model.

The formula that this model suggests should be used to predict BABIP using the 2017 data is printed below.

$BABIP$ = -0.1911803 + 0.00742076 * $sprint\_speed$ + 0.002415534 * $exit\_velocity$ + -0.004560878 * $launch\_angle$ + 0.004693095 * $sweet\_spot$

Model Validation




Normal Q−Q Plot

Looking at the residuals vs fitted plot, we see that there are no clear patterns in the residuals and they are rather evenly distributed around 0. This suggests that the assumption that the relationship is linear is reasonable and that the variance of the error terms are equal. In the normal QQ plot, the points follow the line relatively well, suggesting that the assumtion of normality is valid.

DISCUSSION

CITATIONS

APPENDIX

Model Coefficients

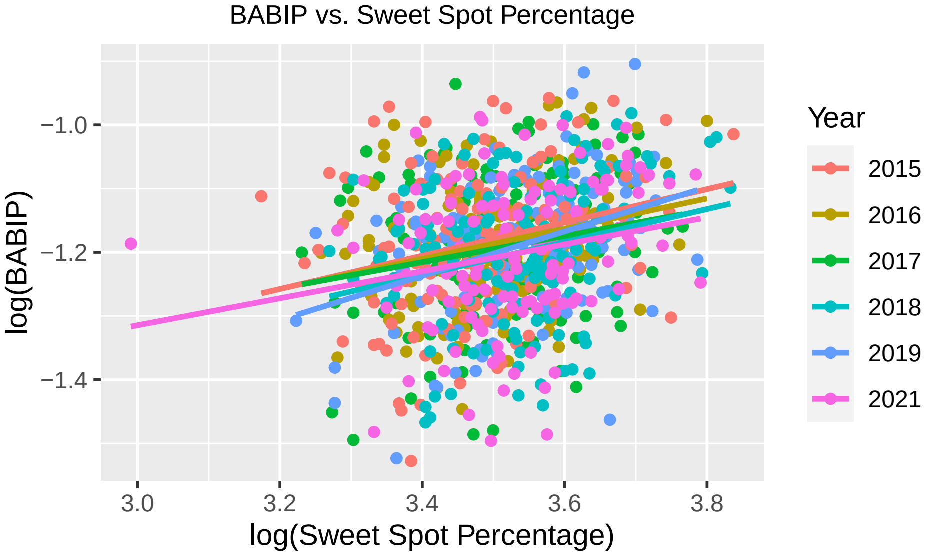
```
## $year
##      (Intercept) sprint_speed exit_velocity_avg launch_angle_avg
## 2015  -0.1789903   0.00742076       0.002375287     -0.004519693
## 2016  -0.1658082   0.00742076       0.002331760     -0.004523871
## 2017  -0.1911803   0.00742076       0.002415534     -0.004560878
## 2018  -0.2844739   0.00742076       0.002723599     -0.004493228
## 2019  -0.2309935   0.00742076       0.002546998     -0.004509899
## 2021  -0.3324017   0.00742076       0.002881859     -0.004519432
##      sweet_spot_percent
## 2015        0.004593839
## 2016        0.004699817
## 2017        0.004693095
## 2018        0.005151112
## 2019        0.004945222
## 2021        0.005058736
```

Random Effects of Model

```
## $year
##        (Intercept) exit_velocity_avg   (Intercept) launch_angle_avg
## 2015  0.0172169963     -1.705527e-04 -2.601473e-05     1.474122e-06
## 2016  0.0216110379     -2.140792e-04  5.498789e-05    -2.704542e-06
## 2017  0.0131536610     -1.303054e-04  7.637189e-04    -3.971118e-05
## 2018 -0.0179441918      1.777593e-04 -5.384774e-04     2.793843e-05
## 2019 -0.0001173775      1.158699e-06 -2.169855e-04     1.126819e-05
## 2021 -0.0339201260      3.360193e-04 -3.722915e-05     1.734976e-06
##       (Intercept) sweet_spot_percent
## 2015  0.011337160      -2.631315e-04
## 2016  0.006771742      -1.571533e-04
## 2017  0.007060536      -1.638755e-04
## 2018 -0.012673003       2.941418e-04
## 2019 -0.003802322       8.825224e-05
## 2021 -0.008694113       2.017663e-04
##
## with conditional variances for "year"
```

Fixed Effects of Model

```
##        (Intercept)       sprint_speed  exit_velocity_avg   launch_angle_avg
##       -0.230641332        0.007420760        0.002545839       -0.004521167
## sweet_spot_percent
##        0.004856970
```


BABIP vs. Sweet Spot Percentage
log(BABIP)
−1.0
−1.2
−1.4
3.0
3.2
3.4
3.6
3.8
log(Sweet Spot Percentage)
Year
2015
2016
2017
2018
2019
2021