145

UNITED STATES MILITARY ACADEMY

General standards 20/20

PROJECT TECHNICAL REPORT

MA376: APPLIED STATISTICS

SECTION G25

DR. ANNY-CLAUDE JOSEPH

By

CADET AUDREY HAMILTON ‘21, CO A1

CADET JOSEPH ZUCCARELLI ’21, CO F1

WEST POINT, NEW YORK

24 NOVEMBER 2020

JZ AH OUR DOCUMENTATION IDENTIFIES ALL SOURCES USED AND ASSISTANCE RECEIVED IN COMPLETING THIS ASSIGNMENT.

_______ WE DID NOT USE ANY SOURCES OR ASSISTANCE REQUIRING DOCUMENTATION IN COMPLETING THIS ASSIGNMENT.

SIGNATURE: Joseph Zuccarelli Audrey Hamilton

Are Sluggers More Likely to Strikeout?

Abstract

The purpose of this report is to analyze the relationship between strike outs and home runs for Major League Baseball (MLB) players. Previous studies indicate the presence of a positive linear relationship between player strike out and home run rates; however, these studies are relatively limited considering that they only analyze one potential explanatory variable. In this study we analyze players from the 2019 MLB season and build a linear regression model that predicts a player's seasonal home run rate based on strike out rate, batting average, and batting hand. We determined that this model is both valid and statistically significant. Overall, this model would perhaps be a useful tool for MLB analytics crews in evaluating player performance. Further studies concerning this topic should explore other player performance metrics to determine if they too are significant predictors of player strike out rates.

30/30

Introduction

In 2019, New York Mets first baseman Pete Alonso led Major League Baseball (MLB) with 53 home runs. However, Alonso was also atop a less desirable league leaderboard–strikeouts. Alonso struck out a total of 183 times during the 2019 season, just shy of the league leader Eugenio Suarez with 189 strikeouts on the season. For sluggers such as Alonso, is it safe to assume that they are swinging for the fences during most at-bats, and does this mean that they are more likely than other players to strike out? Albert (2006) first approached this question from a pitcher's perspective, as he developed a model to predict the strike out rate of pitchers based on walks, batters faced, and the rate at which they strike out batters.[3] Other researchers have attempted to answer this question from a batter's perspective. For example, Foot and Zaki-Asat (2020) conducted a study involving MLB players dating back to 1950 in which they analyzed the relationship between rate of strike outs and home runs for MLB players.[1] Although this study indicated the presence of a positive linear relationship between strikeouts and home runs, it is very limited considering that the researchers' failed to account for potential confounding variables such as batting hand or batting average. Therefore, the objective of the following study is to determine if there is a relationship between rate of strikeouts and home runs for MLB players after adjusting for batting hand and batting average. ✓

Methods

The observational units included in this study were MLB players who had over 100 at-bats during the 2019 season. There were 390 complete observational units in the data set. Each unit represented a player's total regular season numbers. The data set was created using the "Sean Lahman Baseball Database" in R (see R Code).[2] The variables included in the study were as follows: *bats*, *strike out rate*, *home run rate*, and *batting average*. *Bats* was a categorical variable that represented the player's batting hand. This variable had two levels: *L* for those who bat left-handed and *R* for those who bat right-handed. A value of zero was associated with left-handed batters and a value of one was associated with right-handed batters. Note that any switch hitters, or players that bat both left-handed and right-handed, were not included in the study. In this study, *bats* was a potential confounding variable. *Strike out rate* was a quantitative variable that was calculated by dividing the number of total strike outs by the total number of at-bats. In this study, *strike out rate* was the response variable. *Home run rate* was a quantitative variable that was calculated by dividing the number of total home runs by the total number of at-bats. In this study, *home run rate* was the explanatory variable of interest. *Batting average* was a quantitative variable that was calculated by dividing the number of total hits by the total number of at-bats. In this study, *batting average* was a potential confounding variable.

30/30

The statistical analysis method employed in this study was a linear regression involving both quantitative and categorical variables. The null hypothesis for this analysis was that there did not exist a linear relationship between *strike out rate* and *home run rate*, after adjusting for batting average and batting hand. The alternative hypothesis for this analysis was that there did exist linear relationship between *strike out rate* and *home run rate*, after adjusting for batting average and batting hand. ~~There were two potential confounding variables, *bats* and *batting average*.~~ Using these variables, we fit the following model:

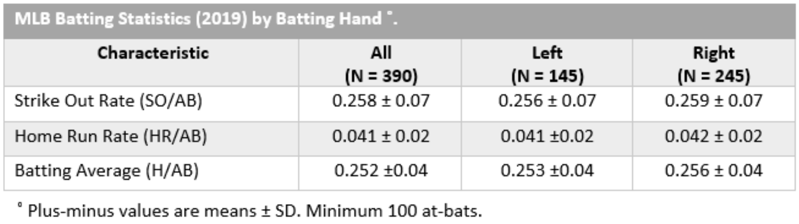
$$\text{StrikeOutRate}_i = \beta_0 + \beta_1 \text{HomeRunRate}_i + \beta_2 \text{BattingAverage}_i + \beta_3 \text{BattingHand}_i + \epsilon_i$$ ✓

$$\epsilon_i \sim \text{Normal}(0, \sigma^2)$$

Results

Table 1 summarizes the mean, standard deviation, and median for the variables *strike out rate*, *home run rate* and *batting average* based on the player's batting hand. ~~All of the values are rates, as each value is a decimal less than one~~. Left-handed batters represented about 37.18% percent of the data set, while right-handed batters represented about 62.18%. Based on the table, left-handed and right-handed batters shared very similar mean and standard deviation values for each variable.

ok.



MLB Batting Statistics (2019) by Batting Hand [*].

Characteristic	All (N = 390)	Left (N = 145)	Right (N = 245)
Strike Out Rate (SO/AB)	0.258 ± 0.07	0.256 ± 0.07	0.259 ± 0.07
Home Run Rate (HR/AB)	0.041 ± 0.02	0.041 ±0.02	0.042 ± 0.02
Batting Average (H/AB)	0.252 ±0.04	0.253 ±0.04	0.256 ± 0.04

[*] Plus-minus values are means ± SD. Minimum 100 at-bats.

Table 1: MLB Batting Statistics (2019) by Batting Hand

Figure 1 displays the distribution of strike out rates for players during the 2019 season. The plot of the distribution of strike out rates had one main peak (unimodal) at around 0.250 and was relatively symmetric. Strike out rates ranged between 0.0421 and 0.4760 with a mean of 0.258 and a median of 0.257.

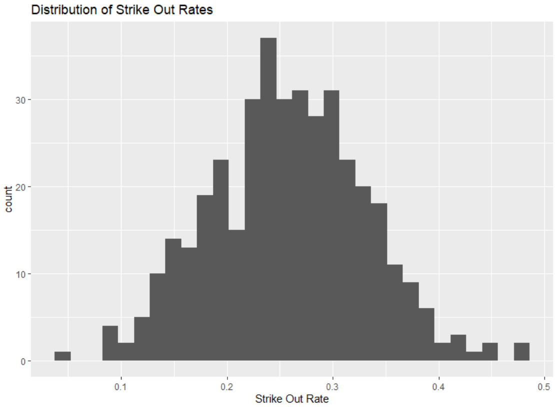


Figure 1: Distribution of Strike Out Rates

Figure 2.a displays the relationship between strike outs and home runs for players from the 2019 season. This plot was colored by batting hand, with left-handed hitters displayed in red and right-handed players displayed in blue. Based on the plot, there appeared to be a positive linear relationship between strike out rate and home run rate. The general trend displayed by the plot is that as a player's home run rate increases, their strike out rate increases as well. The relationship appears to be the same for both left-handed and right-handed batters, as the two regression lines were almost identical.

Figure 2.b displays the relationship between hits and strike outs for players from the 2019 season. This plot was also colored by batting hand, with left-handed hitters displayed in red and right-handed players displayed in blue. Based on the plot, there appeared to be a negative linear relationship between batting average and strike out rate. This linear relationship appeared to be a bit stronger than the linear relationship between strike out rate and home run rate indicated in **Figure 2**. The general trend displayed by the plot is that as a player's batting average increases, their strike out rate decreases. Note that batting average is simply a ratio of hits to at-bats, and thus it appears that the more a player made contact, the less they struck out. The relationship again appeared to be very similar for left-handed and right-handed batters based on the two regression lines.

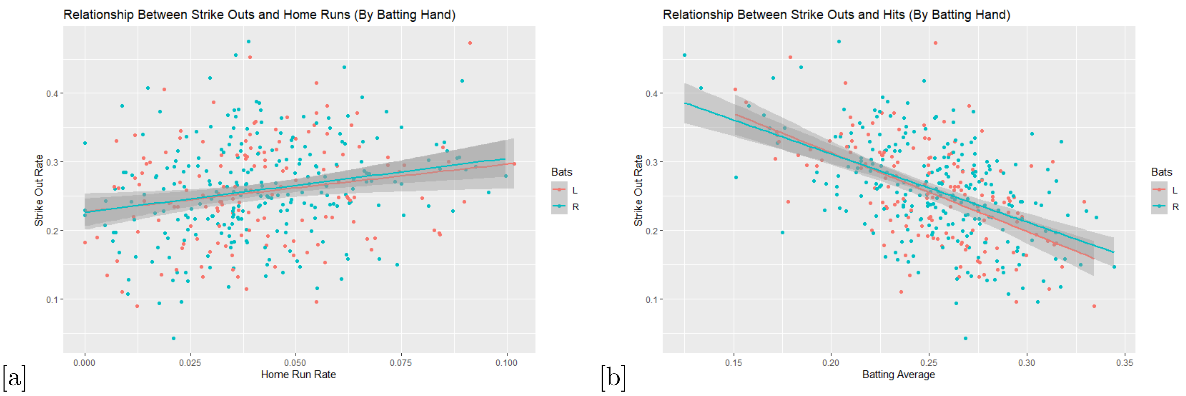
38/40

Table 2 includes the regression output from the model provided in the **Method** section. Overall, the residual standard error of the model was 0.05675 ~~on 386 degrees of freedom~~. The model explained approximately 39.54% of the variation in predicted home run rate. The p-value associated with the overall model was 2.00×10^{-16}, indicating that we could reject the null hypothesis that there did not exist a linear relationship between *strike out rate* and *home run rate*, after adjusting for batting average and batting hand. The p-values associated with the other variables in the model indicated that home run rate and

write p-value <0.0001

Also this p-value is not for the parameter of interest.

This is comparing your model to single mean model.

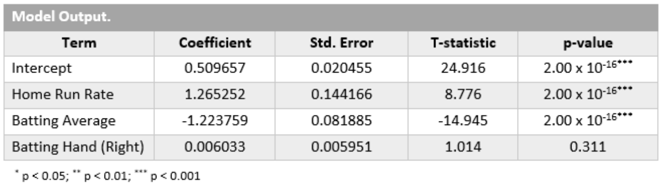


✓

Figure 2: Relationship Between a.) Home Runs b.) Hits and Strike Outs by Batting Hand

batting average were significant predictors of strike out rate, yet batting hand was not. The intercept of this model was not very meaningful, as it represented a left-handed batter whose home run rate and batting average are both zero. However, the other coefficients included in **Table 2** were a bit more insightful. For every one unit increase in home run rate, predicted strike out rate increased by 1.27 units. For every one unit increase in batting average, predicted strike out rate decreased by 1.22 units.

The model met all four validity conditions associated with a linear regression model. The observational units in this study could be considered independent of each other, as no player was included in the data set twice. The histogram of the residuals was approximately symmetric (see **Figure 3** in Appendix B). The residuals vs. predicted values graph did not show any strong evidence of any patterns and showed a relatively constant width (see **Figure 3** in Appendix B).



Model Output.				
Term	**Coefficient**	**Std. Error**	**T-statistic**	**p-value**
Intercept	0.509657	0.020455	24.916	2.00 x 10^{-16***}
Home Run Rate	1.265252	0.144166	8.776	2.00 x 10^{-16***}
Batting Average	-1.223759	0.081885	-14.945	2.00 x 10^{-16***}
Batting Hand (Right)	0.006033	0.005951	1.014	0.311

* p < 0.05; ** p < 0.01; *** p < 0.001

<0.0001

Table 2: Model Output

* You established a relationship. You did not build a predictive model.

Discussion

27/30

Given the findings included in the **Results** section, we have statistically significant evidence of a linear relationship between strike out rate and home run rate after adjusting for batting average and batting hand. Our results suggest that home run rate and batting average are significant predictors of a player's strike out rate, yet batting hand is not. Therefore, the model that * we fit in this report can perhaps be used by MLB analytics crews to predict player home run rates. This would certainly be a useful tool for those involved in the process of scouting and evaluating players. However, considering that this was an observational study, we can not draw any causal conclusions from our findings. Since the model was fit using only MLB players with over 100 at bats during the 2019 season, it would be reasonable to generalize our results to players with over 100 at bats during other seasons as well barring any significant changes in MLB rules or style of play.

In the future, this same analysis could be done on both a smaller or larger scale. For example, each MLB team could fit their own linear model to * predict strike out rate. This could also be done on a larger scale to include minor league and international teams. In order to enable our model to possess more predictive power, perhaps would could also include some other variables that relate to player performance such as age or ✓

some sort of binary variable that accounts for previous injuries. Looking back we also could have left out the variable *batting hand*, as it turned out that it was not a significant predictor in our model. Overall, despite these limitations, this study extends all previously published research concerning the relationship between strike outs and home runs by analyzing potentially confounding variables involved in this relationship. It is highly likely that MLB teams possess better models than ours that they are not willing to share in order to maintain a competitive advantage over others.

Author Contributions

1. CDT Zuccarelli completed all the R code. He developed and included all figures and tables included in the report. Additionally, CDT Zuccarelli completed the abstract, methods and results sections.

2. CDT Hamilton found the peer-reviewed articles, read through them, and incorporated them in the report. Additionally, CDT Hamilton completed the introduction and discussion sections.

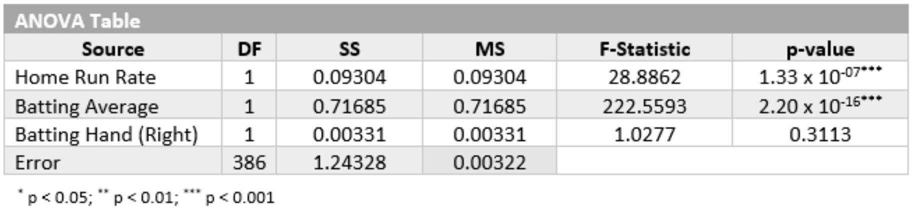
Works Cited

[1] Foot, Vanessa and Justeena Zaki-Azat. *Relationship Between Strikeouts and Home Runs.* The Comprehensive R Archive Network, 2020. https://cran.r-project.org/web/packages/Lahman/vignettes/strikeoutsandhr.html. Accessed 25 September 2020.

[2] Lahman, Sean. *Batting.* Lahman's Baseball Database, http://www.seanlahman.com/baseball-archive/statistics/. Accessed 25 September 2020.

[3] Albert, James. "Pitching Statistics, Talent and Luck, and the Best Strikeout Seasons of All-Time." Journal of Quantitative Analysis in Sports 2.1 (2006). https://doi.org/10.2202/1559-0410.1014.

Appendix A



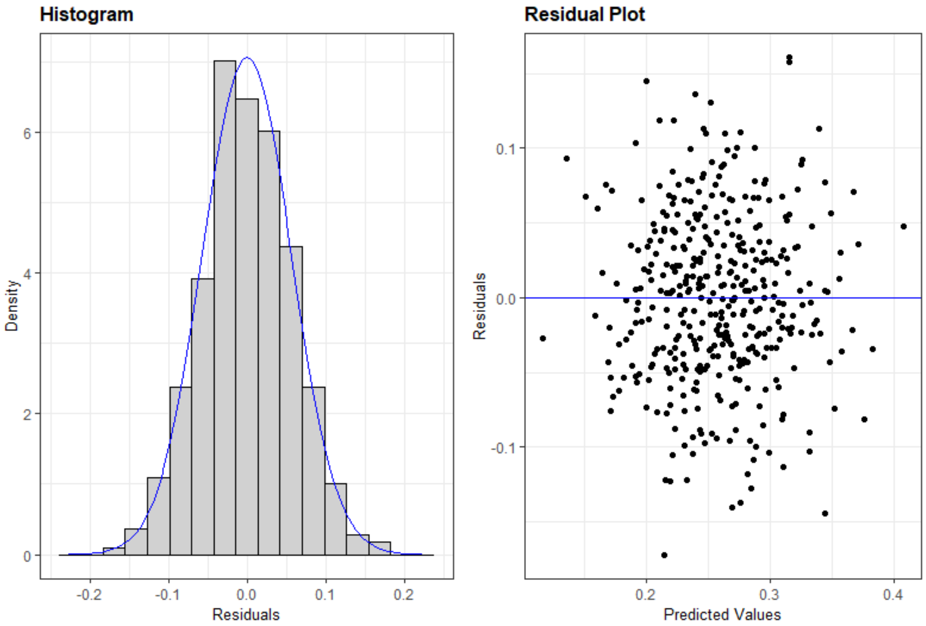
ANOVA Table					
Source	DF	SS	MS	F-Statistic	p-value
Home Run Rate	1	0.09304	0.09304	28.8862	$1.33 \times 10^{-07***}$
Batting Average	1	0.71685	0.71685	222.5593	$2.20 \times 10^{-16***}$
Batting Hand (Right)	1	0.00331	0.00331	1.0277	0.3113
Error	386	1.24328	0.00322		

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

✓

Table 3: ANOVA Table

Appendix B



✓

Figure 3: Plots for Validity Conditions

R Code

The following R Code was used to obtain the data set and produce all figures included in this report.

```
##Libraries Required to Run Code
library(Lahman)
library(tidyverse)
library(table1)
library(knitr)
library(ggResidpanel)


##Merge Lahman Data Frames to Obtain Player Name and Batting Hand
Master$name <- paste(Master$nameFirst, Master$nameLast, sep=" ")
batting <- merge(Batting,
                 Master[,c("playerID","name","bats")],
                 by="playerID", all.x=TRUE)


##Filter Batting Data Based on Inclusion Criteria
battingData <- batting %>%
  filter(yearID == 2019) %>%
  filter(AB >= 100) %>%
  filter(bats == "R" | bats == "L") %>%
  mutate(strikeoutRate = SO/AB,
         homerunRate = HR/AB,
         battingAverage = H/AB) %>%
  select(name,bats,strikeoutRate,homerunRate,battingAverage)


##Table 1 Output
table1(~strikeoutRate + homerunRate + battingAverage | as.factor(bats), data = battingData)


##Plot Distribution of Strike Out Rates
battingData %>%
  ggplot(aes(x= strikeoutRate))+
  geom_histogram()+
  labs(x = "Strike Out Rate") +
  ggtitle("Distribution of Strike Out Rates")

##Plot Relationship Between Strike Outs and Home Runs (By Batting Hand)
battingData %>%
  ggplot(aes(y= strikeoutRate, x= homerunRate, color= bats))+
  geom_point()+
  geom_smooth(method = "lm") +
  labs(y = "Strike Out Rate", x = "Home Run Rate", color = "Bats") +
  ggtitle("Relationship Between Strike Outs and Home Runs (By Batting Hand)")


##Plot Relationship Between Strike Outs and Hits (By Batting Hand)
battingData %>%
  ggplot(aes(x= battingAverage, y= strikeoutRate, color= bats))+
  geom_point()+
  geom_smooth(method = "lm") +
  labs(x = "Batting Average", y = "Strike Out Rate", color = "Bats") +
  ggtitle("Relationship Between Strike Outs and Hits (By Batting Hand)")

##Building a Linear Model
model1 <- lm(strikeoutRate ~ homerunRate + battingAverage + as.factor(bats), data = battingData)
summary(model1)

##ANOVA Table
anova(model1)


##Checking Validity Conditions
resid_panel(model1, plots = c(’hist’, ’resid’), bins = 17)
```