MLB Hitter Analysis

Patrick Weatherford

Bellevue University - DSC 520 Statistics for Data Science

# Introduction

In the Major League Baseball association (MLB), there are many different statistics used to determine how well a player is batting, one of which is On-Base Plus Slugging Percent (OPS). This batting statistic takes into account On-Base Percent (OBP) and Slugging Percent (SLG) and is considered by many to be a good statistic for measuring batting performance. Although Batting Average is more commonly reported which is the total number of hits per at-bat, the statistic does not take into account walks, hit by pitch, or the average total number of bases reached per bat. For this reason, OPS is considered superior by many for it’s ability to report the total number of bases per plate appearance and not just hits per plate appearance.

For this analysis, variables will be assessed to determine if if they explain or predict the outcome variable On-Base Plus Slugging Percent (OPS). In order to assess and measure predictor variables with a potential effect on this outcome, various data sets will bee obtained and filtered prior to analysis.

# Hypothesis

Based on the initial data found, the below variables were chosen for testing and are assumed to be indicators which may potentially explain or predict OPS. Data analysis using a combination of linear vs. logistical regression was performed on the variables to help determine if the variables are likely correlated and if so, if they are statistically significant.

* Player Age
* Bats Left/Right/Switch
* BMI
* Swing %
* Outside of Zone Swing %
* Hard Hit %
* Combined Walks Plus Hits per Inning (WHIP) per League

# Data for the Analysis

Data for the analysis was found in multiple locations, the R CRAN package (Lahman) Dalzell C. (2021), and also online on FanGraphs.com, FanGraphs (2021). For the FanGraph data, there is a tool that allows you to select the statistic and year desired. The result was then exported and saved as a csv file. CSV files were exported for years 2010-2020 and then R was used to read and merge the files into a single data frame. Below is an example output of the data that will be used for this analysis.

Table 1

*Example Data Set*

| **Name** | **Year** | **Age** | **Bats** | **BMI** | **Team** | **League WHIP** | **OPS** | **Out of Zone Swing %** | **Swing %** | **Hard Hit %** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Bobby Abreu | 2,010 | 36 | L | 29.83 | LAA | 1.35 | 0.787 | 0.177 | 0.326 | 0.330 |
| Elvis Andrus | 2,010 | 21 | R | 28.48 | TEX | 1.35 | 0.643 | 0.210 | 0.364 | 0.230 |
| Erick Aybar | 2,010 | 26 | B | 27.98 | LAA | 1.35 | 0.636 | 0.338 | 0.468 | 0.212 |
| Jason Bartlett | 2,010 | 30 | R | 25.77 | TBA | 1.35 | 0.675 | 0.239 | 0.432 | 0.258 |
| Daric Barton | 2,010 | 24 | L | 29.16 | OAK | 1.35 | 0.798 | 0.152 | 0.345 | 0.242 |
| Jose Bautista | 2,010 | 29 | R | 27.80 | TOR | 1.35 | 0.995 | 0.233 | 0.409 | 0.396 |
| Adrian Beltre | 2,010 | 31 | R | 30.68 | BOS | 1.35 | 0.919 | 0.386 | 0.522 | 0.401 |
| Yuniesky Betancourt | 2,010 | 28 | R | 29.41 | KCA | 1.35 | 0.692 | 0.373 | 0.532 | 0.260 |
| Casey Blake | 2,010 | 36 | R | 26.32 | LAN | 1.35 | 0.727 | 0.264 | 0.413 | 0.300 |
| Brennan Boesch | 2,010 | 25 | L | 27.38 | DET | 1.35 | 0.736 | 0.398 | 0.562 | 0.321 |
| Michael Bourn | 2,010 | 27 | L | 26.50 | HOU | 1.35 | 0.686 | 0.233 | 0.403 | 0.240 |
| Ryan Braun | 2,010 | 26 | R | 26.32 | MIL | 1.35 | 0.866 | 0.308 | 0.456 | 0.344 |
| Jay Bruce | 2,010 | 23 | L | 28.74 | CIN | 1.35 | 0.846 | 0.289 | 0.466 | 0.373 |
| Billy Butler | 2,010 | 24 | R | 35.26 | KCA | 1.35 | 0.857 | 0.263 | 0.421 | 0.365 |
| Marlon Byrd | 2,010 | 32 | R | 33.22 | CHN | 1.35 | 0.775 | 0.345 | 0.510 | 0.329 |
| Chris Taylor | 2,020 | 29 | R | 25.86 | LAN | 1.33 | 0.842 | 0.213 | 0.444 | 0.366 |
| Mike Trout | 2,020 | 28 | R | 30.17 | LAA | 1.32 | 0.993 | 0.174 | 0.367 | 0.415 |
| Kyle Tucker | 2,020 | 23 | L | 24.22 | HOU | 1.32 | 0.837 | 0.313 | 0.505 | 0.378 |
| Trea Turner | 2,020 | 27 | R | 23.75 | WAS | 1.33 | 0.982 | 0.290 | 0.469 | 0.367 |
| Christian Vazquez | 2,020 | 29 | R | 30.27 | BOS | 1.32 | 0.801 | 0.338 | 0.478 | 0.308 |
| Alex Verdugo | 2,020 | 24 | L | 26.04 | BOS | 1.32 | 0.844 | 0.290 | 0.427 | 0.325 |
| Jonathan Villar | 2,020 | 29 | B | 31.60 | MIA | 1.33 | 0.593 | 0.368 | 0.513 | 0.261 |
| Jonathan Villar | 2,020 | 29 | B | 31.60 | TOR | 1.32 | 0.593 | 0.368 | 0.513 | 0.261 |
| Luke Voit | 2,020 | 29 | R | 31.87 | NYA | 1.32 | 0.948 | 0.333 | 0.521 | 0.419 |
| Joey Votto | 2,020 | 36 | L | 28.24 | CIN | 1.33 | 0.800 | 0.195 | 0.362 | 0.364 |
| Christian Walker | 2,020 | 29 | R | 28.48 | ARI | 1.33 | 0.792 | 0.355 | 0.518 | 0.444 |
| Evan White | 2,020 | 24 | R | 27.50 | SEA | 1.32 | 0.599 | 0.284 | 0.438 | 0.343 |
| Kolten Wong | 2,020 | 29 | L | 28.97 | SLN | 1.33 | 0.675 | 0.273 | 0.429 | 0.247 |
| Mike Yastrzemski | 2,020 | 29 | L | 25.54 | SFN | 1.33 | 0.968 | 0.234 | 0.389 | 0.372 |
| Christian Yelich | 2,020 | 28 | L | 24.37 | MIL | 1.33 | 0.786 | 0.203 | 0.346 | 0.411 |

# Plots & Tables Needed

For analysis, the below output may be useful to help determine and assess any correlation effects and run diagnostics on the data.

* Scatter plot of predictor variable vs. outcome and residuals vs. outcome
* QQPlot of predictor variable data and residuals of model
* Histogram of predictor variable data and residuals of model
* Residuals vs. Fitted
* Scale-Location Plot
* Residuals vs. Leverage
* Outlier Table
* Combined Data Frame Table
* Linear Model Line for predictor variable vs. outcome

# Visualizing the Data

Figure 1

*Histogram of OPS*

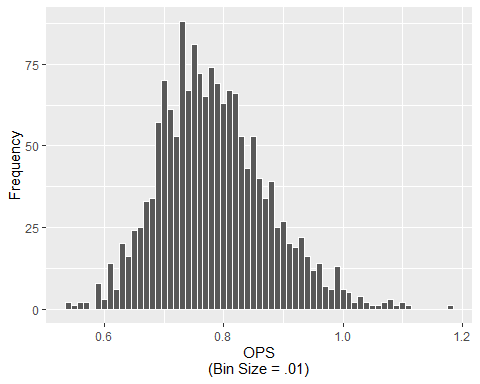


Figure 2

*Age vs. OPS*

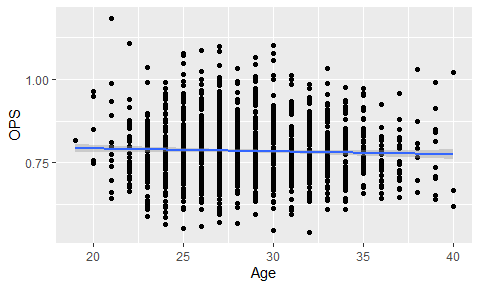


Figure 3

*BMI vs. OPS*

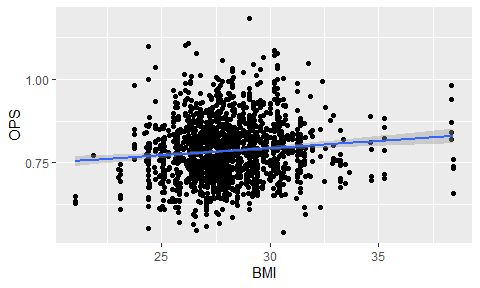


Figure 4

*League WHIP vs. OPS*

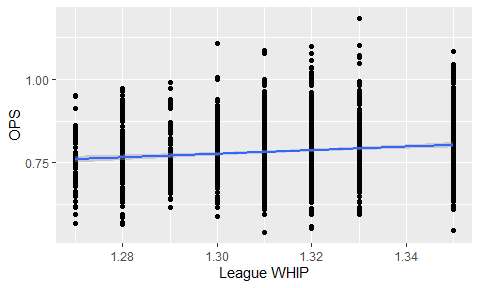


Figure 5

*Out of Zone Swing % vs. OPS*

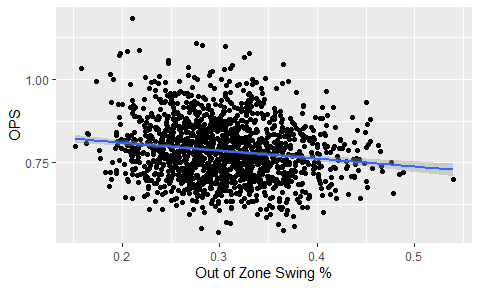


Figure 6

*Swing % vs. OPS*

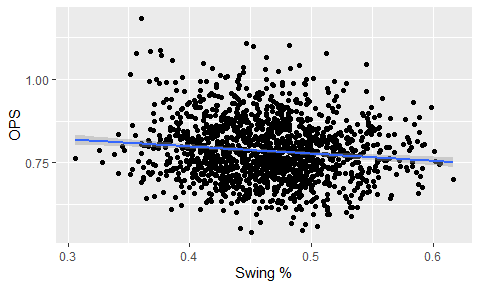


Figure 7

*Hard Hit % vs. OPS*

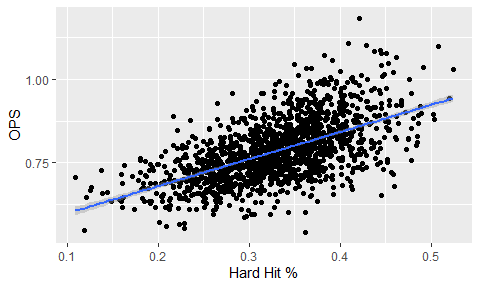
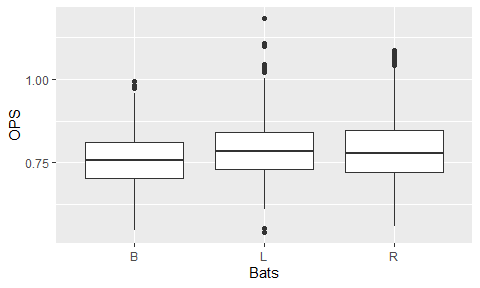


Figure 8

*Bats vs. OPS*



# Model Selection

Because data for OPS appears to be normally distributed and linear when compared to the predictor variables which are primarily continuous, I will perform a linear regression analysis. Also, because the variable ‘Hard Hit %’ appears to have a strong linear correlation with OPS and is continuous, this further strengthens the hypothesis that a linear regression model would be a good fit for this model.

# Linear Regression Analysis

## Regression with All Predictors (lm1)

* Age
* Bats
* BMI
* League WHIP
* Out of Zone Swing %
* Swing %
* Hard Hit %

| **call** | **fstat** | **adj.r.squared** |
| --- | --- | --- |
| lmOPS ~ . | 91.01969 | 0.372915 |

| **Variables** | **Estimate** | **P-value** |  |
| --- | --- | --- | --- |
| Age | -0.0004931646 | 0.372 |  |
| BatsL | 0.0122681170 | 0.068 |  |
| BatsR | 0.0103286792 | 0.102 |  |
| BMI | 0.0010016413 | 0.333 |  |
| `League WHIP` | 0.2179348875 | 0.022 | \* |
| `Out of Zone Swing %` | -0.0903853261 | 0.317 |  |
| `Swing %` | -0.1374477065 | 0.182 |  |
| `Hard Hit %` | 0.8014519562 | 0.000 | \*\*\* |

After initial regression analysis with all variables, there are some that are not significant meaning that it is likely that there is no effect on OPS and gain confidence in accepting the NULL hypothesis that there is no correlation with OPS. Will remove these variables and re-run the linear regression.

## Regression with Below Predictors (lm2)

* League WHIP
* Hard Hit %

| **call** | **fstat** | **adj.r.squared** |
| --- | --- | --- |
| lmOPS ~ `League WHIP` + `Hard Hit %` | 339.9474 | 0.3588843 |

| **Variables** | **Estimate** | **P-value** |  |
| --- | --- | --- | --- |
| `League WHIP` | 0.2343487 | 0.015 | \* |
| `Hard Hit %` | 0.8199361 | 0.000 | \*\*\* |

## ANOVA

lm\_anova <- data.frame(anova(lm1, lm2))

## ANOVA: Is significant at p < .001

## AIC

##   
## Model selection based on AICc:  
##   
## K AICc Delta\_AICc AICcWt Cum.Wt LL  
## lm1 10 -2876.33 0.0 1 1 1448.26  
## lm2 4 -2855.64 20.7 0 1 1431.83

After running linear regression analysis using all predictor variables and again using only significant variables from the first analysis, the adjusted R squared value dropped by .02 which is very small however the F-statistic increased significantly from 91 to 340. After running ANOVA, it was found that the simpler model

# References

Dalzell C., Monkman M., Friendly M. 2021. “Sean ’Lahman’ Baseball Database.” <https://CRAN.R-project.org/package=Lahman>.

FanGraphs. 2021. “FanGraphs Baseball Custom Statistics.” 2021. <https://www.fangraphs.com>.