Do characteristics that predict MLB regular season success translate in the post-season?*

An analysis of what physical characteristics can be used to predict a batters' success in the regular and post season

James Richards

27 April 2022

Abstract

Although baseball as a sport has over 100 years of statistical history and has already been extensively analyzed, I decided it would be interesting to see if a regression model measuring players batting in the regular season would match their playoff batting success when solely predicted through their physical characteristics (such as height, weight, handedness, etc.). This paper does an in-depth exploration of the physical characteristics of MLB players from the 1871 to 2021 and measures them against their on-base plus slugging percentage (OPS) in the regular season and playoffs through the use of multiple linear regression models with different predictors. Using ANOVA tables and likelihood ratio tests, predictors were gradually removed from the base models until models that best explained the relationship between a players physical characteristics and their batting prowess in the regular and post season was discovered. After this relationship was discovered, their implication was discussed.

Contents

*Code and data are available at: github.com/Konrad-99/Baseball.

List of Figures

Introduction

"A young man named Abner Doubleday invented the game known as baseball in Cooperstown, New York, during the summer of 1839. Doubleday then went on to become a Civil War hero, while baseball became America's beloved national pastime" (Staff 2013). Unfortunately, as good as this story may sound, it was later found to be untrue (Staff 2013). What is true, however, is that an Englishman by the name of Henry Chadwick was the first to start recording statistics of players' and teams' performance in 1859 with the introduction of box scores ("Henry Chadwick," n.d.). Since then, baseball has been at the forefront of sports analytics with teams using extensive statistics to determine who they want on their team and how much they are willing to pay them. This study will explore players' physical characteristics to determine what characteristics increase or decrease a players likelihood of being a good batter, which will be measured by OPS, the sum of a player's on-base and slugging percentage (MacLennan 2019). On-base percentage measures the frequency of a player reaching a base each time they go up to bat while slugging percentage measures the total number of bases a player records per at bat (MacLennan 2019). According to most players, OPS is the most important statistic when it comes to measuring a batters efficiency as it rewards players who get on base often as well as those who are sluggers (MacLennan 2019). In order to determine what characteristics are important for good batters in the regular and post season, a multiple linear regression model will be fitted using the data after which ANOVA tables and likelihood ratio tests will be used to determine if any variables can be removed from the model to better fit the data.

Data

In order to conduct this analysis, the R programming language (R Core Team 2020) was used, with the `tidyverse` (Wickham et al. 2019), `janitor` (Sam Firke 2021), `dplyr` (Wickham et al. 2021) and `reshape2` (Wickham 2007) packages being used for data cleaning and manipulation. In order to create and display graphs and tables, the `ggplot2` (Wickham 2016), `car` (Fox, Weisberg, and Price 2021), `broom` (Robinson, Hayes, and Couch 2022) and `kableExtra` (Zhu 2020) packages were used. Finally, the `knitr` (Xie 2021b) and `bookdown` (Xie 2021a) package was used to knit the markdown file and produce a pdf copy.

Initially, when searching for data to use in this paper, I attempted to use data from the MLB website as I believed they would have an extensive source of free statistics to use, however, they did not and I was forced to search elsewhere. The data used in this research paper is from Lahman's Baseball Database, a free database that contains pitching, hitting, and fielding stats for Major League Baseball from 1871 to 2021 (Friendly et al. 2022). Moreover, the database includes statistics from both the National and American league as well as any leagues formed before that (Friendly et al. 2022). The website was initially formed in 1994 by Sean Lahman in order to provide free baseball statistics to the public and currently they have a dedicated team that have no formed the most extensive source of baseball statistics available to the public (Friendly et al. 2022). The 2021 version of the dataset contains 28 tables measuring varying statistics, however, in this case only the batting and people tables were used for both regular and post season stats. The player table contains data measuring player's physical attributes as well as career information. In the case of the batting table, it measures each players' batting stats in every season played as well as the team they played for. To use this data, download the `Lahman` (Friendly et al. 2022) package. Unfortunately, however, it did not include OBP, SLG or OPS which resulted in myself calculating them. The following are the formulas used to calculate each stat:

- $OBP = \frac{H+BB+HBP}{AB+BB+HBP+SF}$
- $SLG = \frac{H+(2*2B)+(3*3B)+(4*HR)}{AB}$
- $OPS = OBP + SLG$

After calculating the OBP, SLG and OPS for each player, the two tables were combined using the left_join() function and any observations containing NA were removed. Additionally, when looking at the OPS and AB values, it became evident that the data was skewed as there were too many high OPS values due to low

at bats for the large majority of players. This led to the creation of a cutoff for measuring OPS in which a player had to have at bat at least 502 times, the same amount a player needs to compete for a hitting title (Writer 2020). For the post season, I removed all players that did not make the cut in the regular season. Once ready, any variables not being used were removed which left the following two datasets:

Table 1: The first 5 rows of the final OPS dataset for regular season

	Age	weight	height	bats	throws	OPS
110608	30	210	75	B	R	0.9445153
110609	31	210	75	B	R	0.9661626
110610	32	210	75	B	R	0.8396700
110611	33	210	75	B	R	0.8314875
110614	35	210	75	B	R	0.9308980

Table 2: The first 5 rows of the final OPS dataset for regular season

	Age	weight	height	bats	throws	OPS
110608	30	210	75	B	R	0.9445153
110609	31	210	75	B	R	0.9661626
110610	32	210	75	B	R	0.8396700
110611	33	210	75	B	R	0.8314875
110614	35	210	75	B	R	0.9308980

Model

For this study two models were created, one for the regular season data and one for the post season data. In each case, a multi-linear regression model will be fitted to the data. Firstly, I split each dataset into a testing and training dataset with a 50/50 split while performing an EDA on all the datasets. Following this, I checked whether the quantitative variables in both training datasets satisfied conditions I & II through residual scatterplots and qqplots. Once this was done, base models for both the post and regular season were created and are as follows:

- Regular Season Model
  - $OPS = \beta_0 + \beta_1 * Age + \beta_2 * Weight + \beta_3 * Height + \beta_4 * BatsL + \beta_5 * BatsR + \beta_6 * BatsB + \beta_7 * ThrowsL + \beta_8 * ThrowsR$
- Post Season Model
  - $OPS = \beta_0 + \beta_1 * Age + \beta_2 * Weight + \beta_3 * Height + \beta_4 * BatsL + \beta_5 * BatsR + \beta_6 * BatsB + \beta_7 * ThrowsL + \beta_8 * ThrowsR$

In order to check the Linearity, Uncorrelated Errors, Common Error Variance, and Normality of Errors assumptions, scatterplots, Q-Q and residual vs fitted plots were employed. In the case of the regular season data, it appeared as if the assumptions needed to create a model were met, however, for the post season data the qqplot indicated some of the quantitative variables may need to undergo transformations.

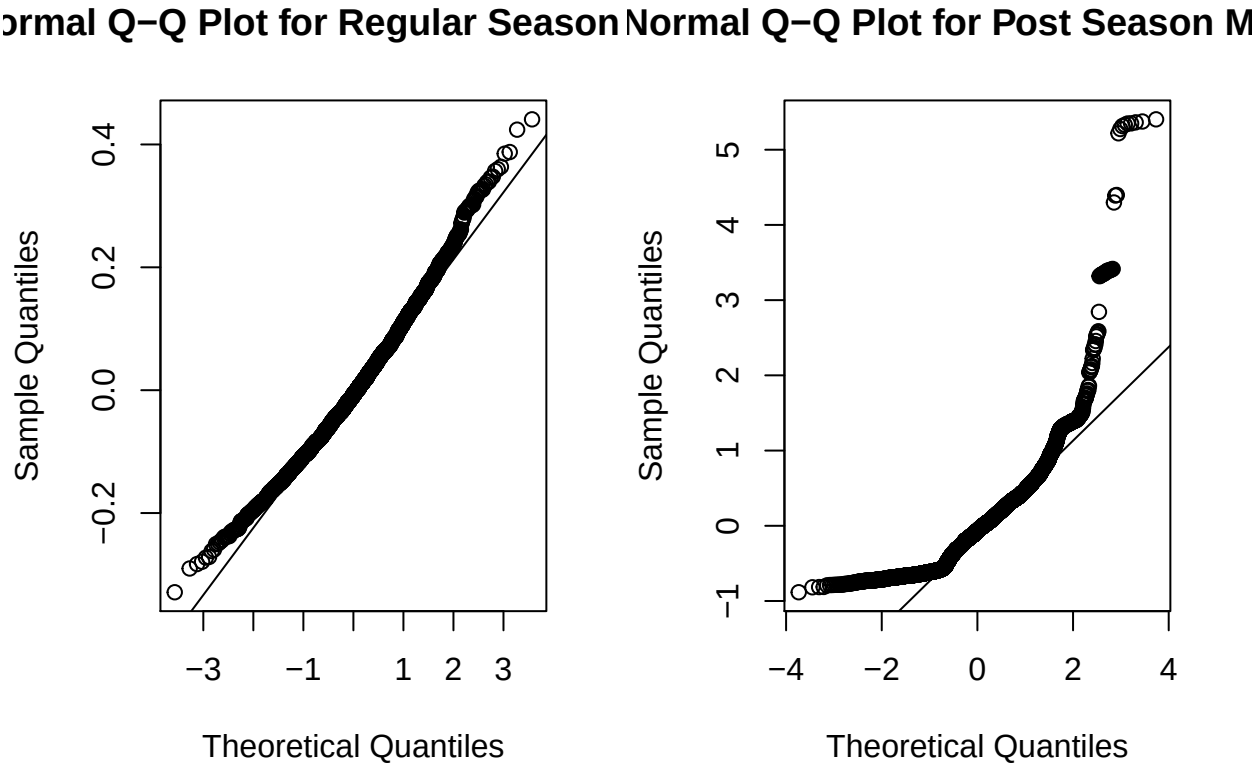

Figure 1: Fit of regular season model versus post season model on their respective data

To transform these variables, the box-cox function was utilized in which the user is told which variables should undergo what transformation, whether that be squaring or inverting the variable for example. For the regular and post season data, the following transformations were applied:

- Regular season:
  - $Age => Age^{-\frac{1}{2}}$
  - $Weight => Weight^{-1}$
- Post season:
  - $OPS => OPS^{\frac{1}{2}}$
  - $Weight => Weight^{-1}$
  - $Height => Height^{2}$

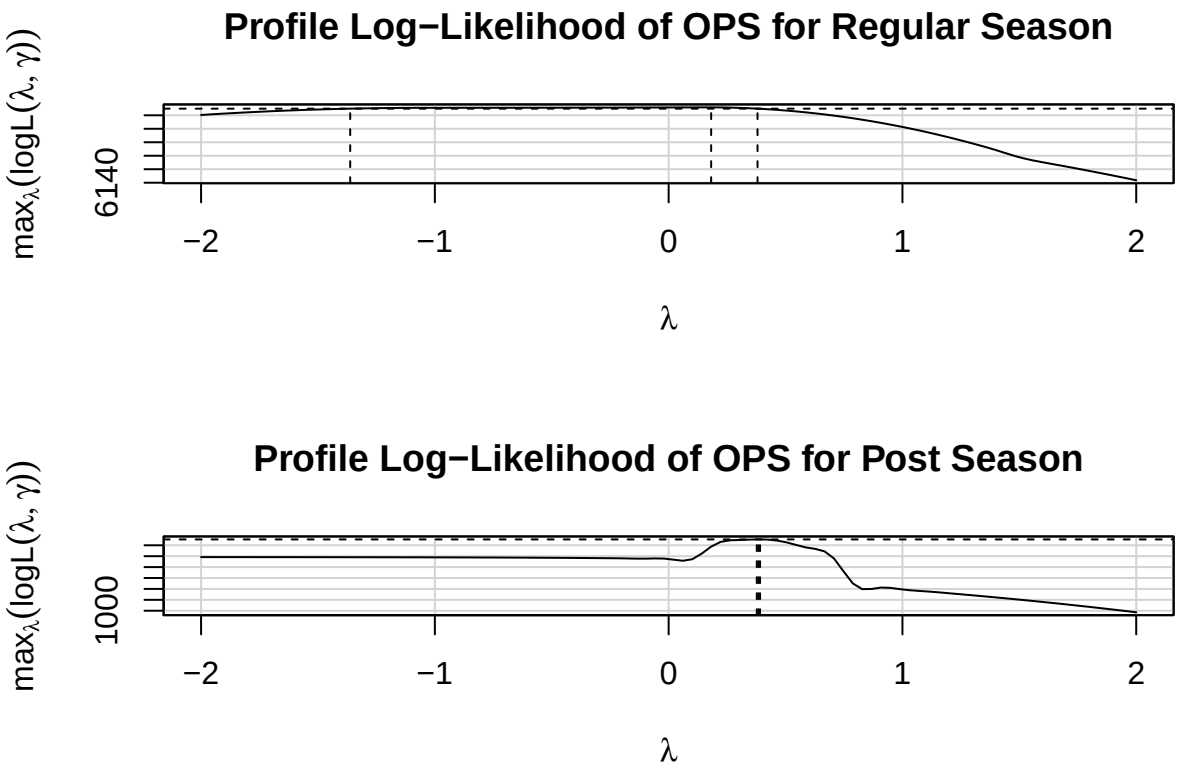

Figure 2: Box-Cox graph for OPS telling us what transformation to apply to OPS in both regular and post season models

Once transformations were applied to both datasets, multicollinearity was checked for using the variance inflation factor, however, no variables were removed from either dataset. Once ready, both models underwent a t-test test in order to determine which variables to remove. In this case, any variable whose p-value was greater than 0.05 was removed, leading to the following reduced models:

- Regular Season Model
  - $OPS = \beta_0 + \beta_1 * Age + \beta_2 * Weight + \beta_3 * Height + \beta_4 * BatsL + \beta_5 * BatsR$
- Post Season Model
  - $OPS = \beta_0 + \beta_1 * Weight + \beta_2 * Height$

After the removal of predictors from the models, they were compared using a partial f-test in which it was found that in both cases the reduced models explained the data better than the original models. Finally, the testing data was fit into each model to check for any bias in the data. No bias was discovered and the following qq plots indicate the fit of the testing and training models for both regular and post season.

Creating test models

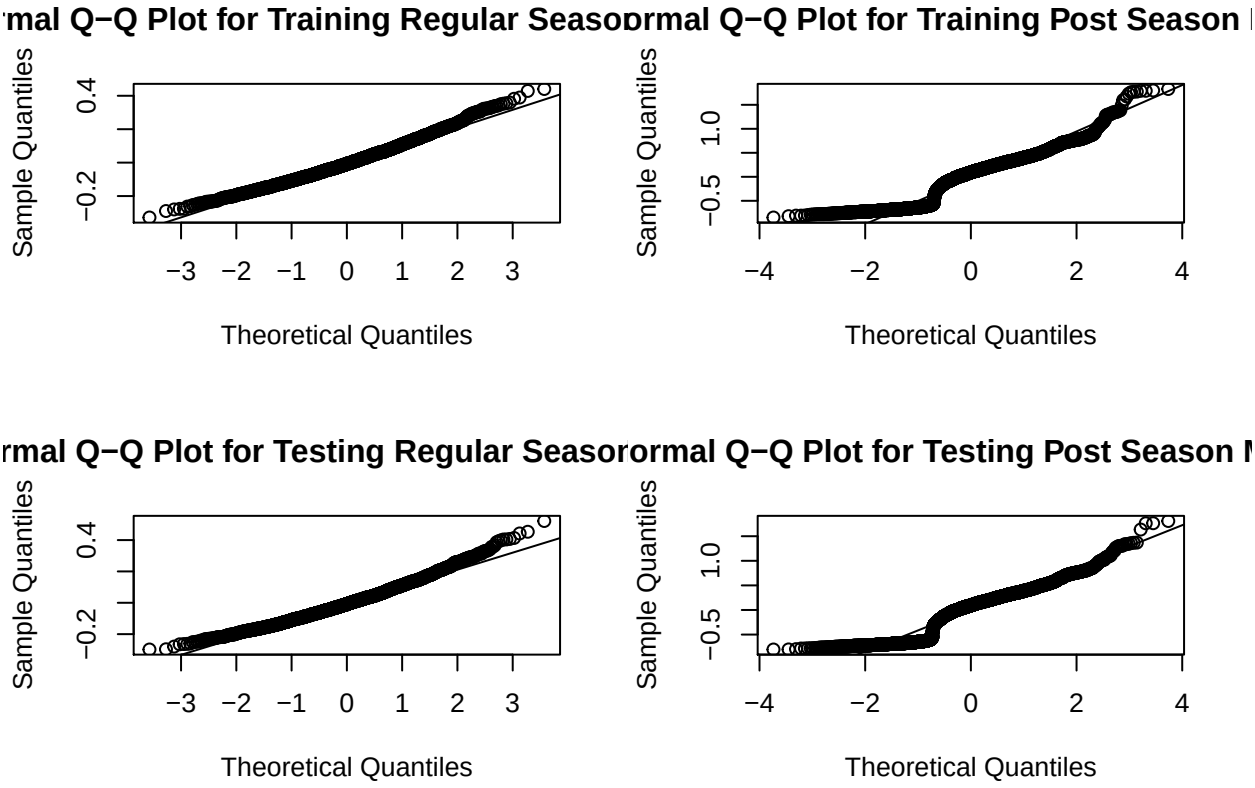

Figure 3: Q-Q plots comparing the training and testing models for both regular and post season

Results

Based on the model, physical characteristics that influence batting prowess in the regular season are age, height, weight and what side you bat on. Below are plots showing the relationship between continuous predictors and OPS for the regular season:

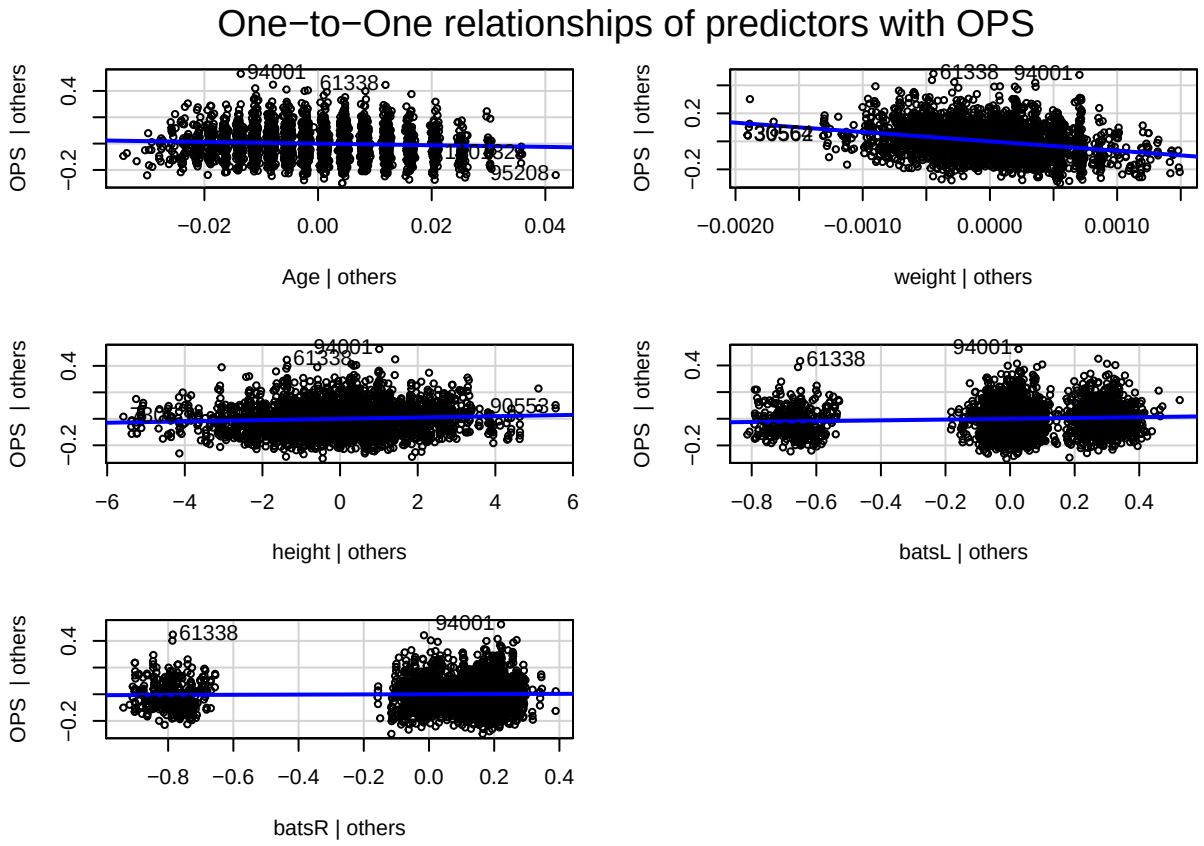

Figure 4: Added Variable plots indicating the one-to-one relationship of predictors with OPS in the regular season model

These graphs indicate the relationship of the respective predictor with OPS provided all other predictors remain constant. For example given all other predictors remain constant, a 1 unit increase in weight will result in a decrease in OPS. From these graphs we can see that in the regular season, the older and heavier you are, the worse your batting becomes while the taller you are, the better your batting. Unfortunately, when analyzing how what side you bat on it is difficult to see their relationship with OPS. According to the summary of the model, left handed batters have a slight advantage over right handed batters as there is a more significant increase in OPS given to a left handed batter given two players are otherwise identical. Interestingly, players that could bat on both sides appear to have a lower OPS on average. This can also be visualized in the following boxplot comparing left and right handed hitters' OPS:

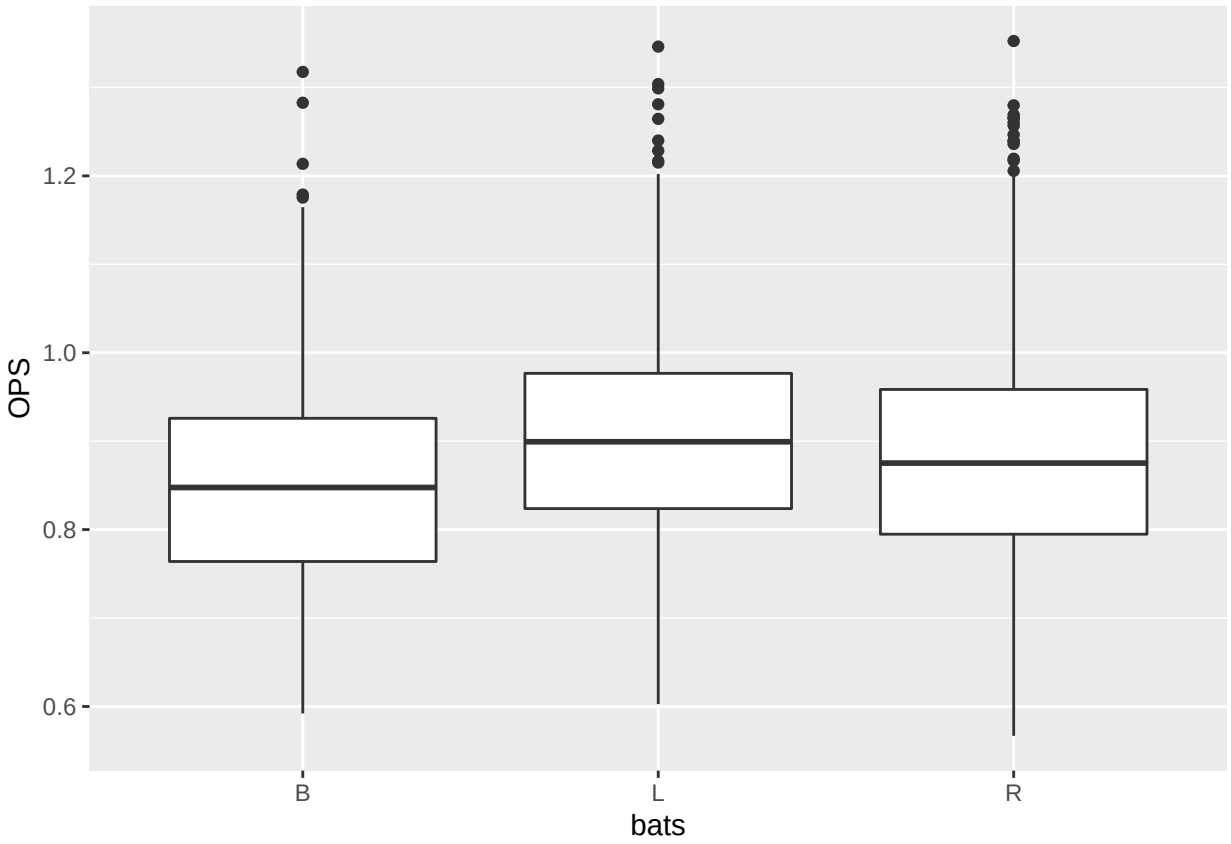

Figure 5: Boxplot of each batting side's OPS IQR and mean

When looking at what makes a good batter in the post season, there are decidely less factors as the only 2 predictors for OPS included in the model were weight and height. Their relationship can be visualized through the following plots

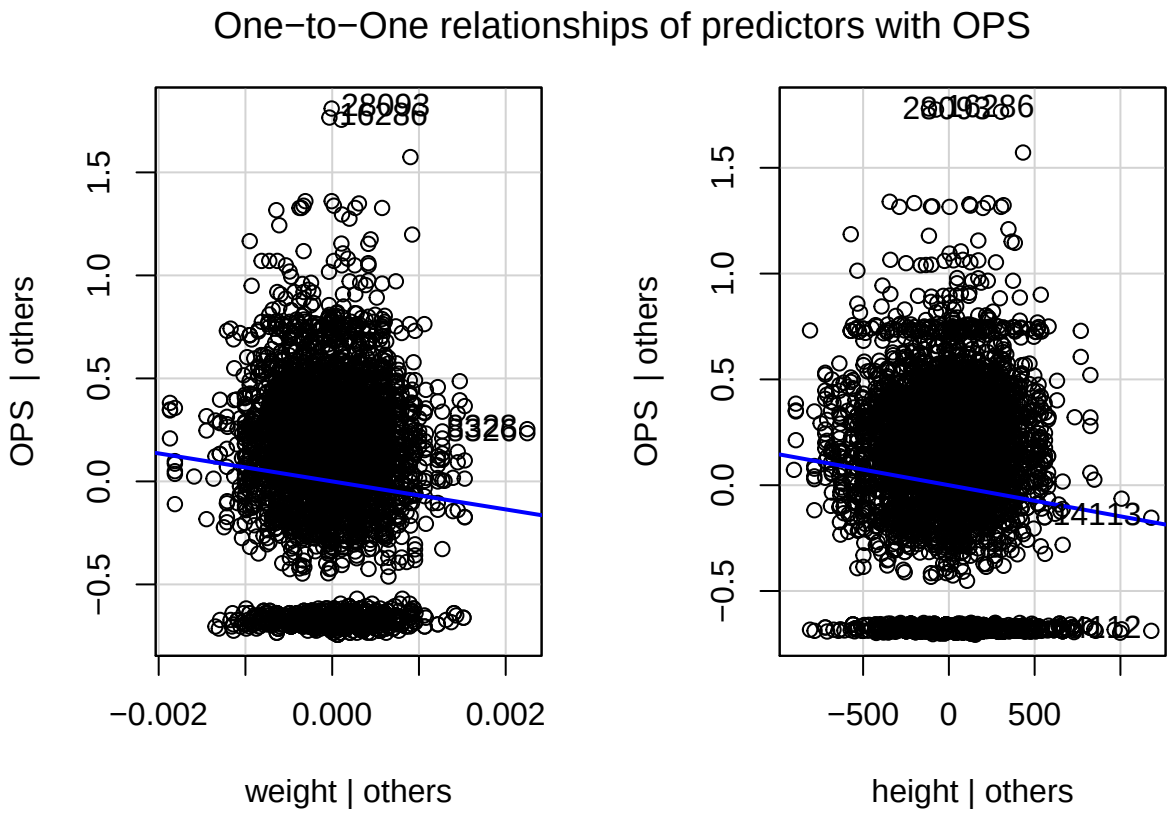

Figure 6: Added Variable plots indicating the one-to-one relationship of predictors with OPS in the post season model

These graphs once again illustrate the relationship between predictors and independent variables given all other predictors remain constant. This indicates that in the post season, the taller and heavier you are, the worse a player's OPS will be. For weight this follows the same trend as in the regular season, however, for height we see a negative relationship which is contrary to what we see in the regular season. It should be noted, the number of at bats a player has during the post season is significantly lower than during the regular season. This can lead to skewed batting stats as the law of large numbers is not able to average out players' inflated or deflated batting stats. Based on this, it is possible that the post season model is not a reliable predictor of how a player will perform in come the start of playoffs.

Discussion

Based on the determined models, it appears that physical characteristics that predict regular season batting success don't necessarily guarantee post season batting success. We can see that being tall, young and a left or right handed batter will increase your OPS while being heavier and older will decrease it. Interestingly, the common misconception of baseball players being overweight and out of shape (Conroy, Wolin, and Carnethon 2016) can be disproved in both models as the lower a player's weight, the higher their OPS. In the post season, however, we see that regular season predictors do not matter as age, and batting side do not significantly affect OPS while being taller and heavier decreases players' OPS. This could be explained by taller hitters having larger strike zones (Freiman 2018), giving more skilled pitchers (typically found in the post season) a larger strike zone to throw at. In an interesting study, it was found that taller players generally are struck out more by low balls than average height players (Freiman 2018). Could the fact that taller players face a higher quality of pitching along with a larger strike zone diminish their supposed batting advantage in the regular season? Moreover, the advantage of being a "lefty" in baseball has been well

documented as they are able to hit a ball a little more out in front than other batters (Brooks 2020) which allows for greater distance on their hits. When looking at the playoffs, however, this advantage seems to disappear. In order to better understand the change in batting prowess from regular to post season, further analysis on how pitching changes is needed as batting and pitching are linked to one another.

Appendix

Key Words

- OPS: A measurement of how good a player is at slugging and getting on base
- OBS: A measure of a player reaching one of the 4 plates without getting out
- SLG: A measurement of how many bases a player records for every at-bat
- At Bat: when a player goes up to bat
- Bats: Side a player bats on (left, right or both)
- Throws: Arm a player throws with (left or right)
- Strike Zone: The zone that a player pitches at for a strike. Any area outside this zone does not count as a strike
- Regular Season: The portion of a baseball season where teams play games to determine place in standings
- Post Season: The portion of a baseball season where only the best teams remain and compete for the world series

Rules of Baseball

If interested or confused with how baseball works, the rules of the game can be found at the MLB official website (MLB 2021).

References

Brooks, Mark. 2020. "Pitchers Vs Batters: Left-Handed &Amp; Right-Handed Hitting Approaches." *Applied Vision Baseball*. https://appliedvisionbaseball.com/pitchers-vs-batters-left-handed-right-handed-hitting-approaches/#:~:text=Lefties%20have%20an%20advantage%20in,to%20higher%20levels%20in%20baseball.

Conroy, David E., Kathleen Y. Wolin, and Mercedes R. Carnethon. 2016. "Baseball Players Are Getting Fat Like Us: Column." *USA Today*. Gannett Satellite Information Network. https://www.usatoday.com/story/opinion/2016/10/04/baseball-mlb-playoffs-obesity-fat-column/91479384/.

Fox, John, Sanford Weisberg, and Brad Price. 2021. "Companion to Applied Regression [r Package Car Version 3.0-12]." *The Comprehensive R Archive Network*. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/car/index.html.

Freiman, Nate. 2018. "How Much Does Height Affect a Hitter's Zone?" *FanGraphs Baseball*. https://blogs.fangraphs.com/how-much-does-height-affect-a-hitters-zone/.

Friendly, Michael, Chris Dalzell, Martin Monkman, Dennis Murphy, Vanessa Foot, Justeena Zaki-Azat, and Sean Lahman. 2022. "Sean 'Lahman' Baseball Database [r Package Lahman Version 10.0-1]." *The Comprehensive R Archive Network*. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/Lahman/.

"Henry Chadwick." n.d. *Baseball Hall of Fame*. https://baseballhall.org/hall-of-famers/chadwick-henry#:~:text=In%201859%2C%20Chadwick%20formulated%20his,them%20with%20the%20letter%20'K.

MacLennan, Ashley. 2019. "A Complete Beginner's Guide to Baseball Stats: Batting Statistics, and What They Mean." *Bless You Boys*. https://www.blessyouboys.com/2019/1/8/18171919/baseball-stats-for-beginners-batting-average-on-base-percentage-explained.

MLB. 2021. "Official Baseball Rules 2021 Edition." *MLB*. MLB. https://img.mlbstatic.com/mlb-images/image/upload/mlb/atcjzj9j7wrgvsm8wnjq.pdf.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

Robinson, David, Alex Hayes, and Simon Couch. 2022. "Convert Statistical Objects into Tidy Tibbles [r Package Broom Version 0.8.0]." *The Comprehensive R Archive Network*. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/broom/index.html.

Sam Firke, Chris Haid, Bill Denney. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Staff, History. 2013. "Who Invented Baseball?" *History.com*. A&E Television Networks. https://www.history.com/news/who-invented-baseball#:~:text=You%20may%20have%20heard%20that,not%20even%20in%20the%20ballpark.

Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Writer, Staff. 2020. "How Many at Bats Are Needed to Qualify for the MLB Batting Title?" *Reference*. IAC Publishing. https://www.reference.com/world-view/many-bats-needed-qualify-mlb-batting-title-1c15c9ba52996866.

Xie, Yihui. 2021a. *Bookdown: Authoring Books and Technical Documents with r Markdown*. https://github.com/rstudio/bookdown.

———. 2021b. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2020. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.