Variables Associated With wRC+ for Qualified Hitters from 2017-2019

3/14/2023

Introduction

For decades, baseball has been a treasure trove of detailed statistical information. Statistics such as batting average, home runs, strikeouts, and many more have been used to evaluate player performance and inform managers and front offices on which players are most valuable to their teams. In recent years, however, advanced statistics have become more popular and teams are making important roster and lineup decisions after examining these statistics. One of these statistics is weighted runs created plus (wRC+), an all encompassing hitting statistic that takes into account many offensive statistics and standardizes the statistic to ensure that league average is 100. Each additional point of wRC+ higher than 100 means that a hitter is 1 percent higher than league average in wRC+, or in other words, they create 1 percent more runs than the average MLB hitter.

It is clear to see how powerful this metric can be in analyzing the productivity of a hitter, as we can simply look at one number and get a sense of how good they are compared to league average. And because this statistic takes into account so many other offensive statistics, we do not need to waste time looking at individual statistics such as batting average, home runs, or strikeouts. There have been various informal studies that have attempted to predict hitting data based on previous hitting data; however, all the studies that I was able to find only looked at hit probability, and did not look at wRC+. Still, it is likely that those with high hit probabilities have a high wRC+; therefore, these studies are still similar enough to the point where at least some of the factors they found contributed to hits can possibly help explain wRC+ as well. All of the studies that I was able to find listed launch angle (the vertical angle at which the ball leaves a player's bat after being struck) and exit velocity (the speed of the baseball as it comes off the bat, immediately after a batter makes contact) as the most important factors to the likelihood of a hit, with one article by Sports Illustrated mentioning that players themselves have been consciously thinking about their launch angle when training for the new season. Launch angle and exit velocity seem to be the consistent factors that people conclude are best at predicting the likelihood of a hit, so it is fair to hypothesize that launch angle and exit velocity will likely contribute heavily to any model that has wRC+ as the response variable.

In this paper, however, I want to examine other advanced offensive statistics (that are not included in the wRC+ calculation) such as swing rate on pitches in and outside of the strike zone, how clutch a player is, and how often a player pulls the ball in order to determine how wRC+ for qualified hitters changes between the 2017 and 2019. I also will examine how factors such as position and batting handedness are associated with wRC+ over this time period. While there has been many past research addressing the importance of wRC+ and the more simple factors associated with it, I hope to address this question differently by looking at much more niche offensive statistics and discovering the association between these statistics and wRC+.

Methods

The data set contains data on 171 different seasons from 57 different players from the 2017 to 2019 season. The 57 players were the only players to have 3 straight “qualified” hitting seasons, which means that they had at least 502 plate appearances in each season. Qualified hitters were selected to ensure a similar large sample size for all seasons in the data set. The response variable was the wRC+ of the season, and there were initially 9 potential explanatory variables. The variables that were evaluated at a season level were the outside swing percentage (how often the player swung at pitches outside the strike zone), zone swing percentage (how often the player swung at pitches inside the strike zone), the clutch rating of the player’s season (performance of hitter in “high leverage” situations), pull percentage (how often the player pulled the ball to their hitting side of the field), whether the player is younger or older than 28 in the current season, and a season variable (1 for 2017, 2 for 2018, and 3 for 2019). On a player level, the attributes include the position of the batter and the handedness of the batter. There are two position groups; infielders and outfielders, divided by the position the player played most in his career. There were originally 4 positions in the dataset, but there was only 1 catcher and 3 designated hitters; therefore, catchers were considered infielders and the designated hitters were assigned their secondary position (all outfielders). The final dataset is the result of data from three separate datasets. A majority of the data (wRC+, outside swing percentage, zone swing percentage, clutch rating, pull percentage, season, and age) was pulled from FanGraphs. The handedness of the player was from a dataset provided by Sean Lahman, while the position of the player was from a dataset provided by Robin Lock.

Results

When taking a preliminary look at the data, below are the summary statistics for the quantitative variables:

Variable	Minimum	Mean	Maximum	Standard Deviation
wRC+	56.0	116.2	188.0	25.401
Outside Strike Zone Swing %	15.80	30.67	48.80	6.426
Inside Strike Zone Swing %	53.00	68.75	85.00	5.941
Clutch	-2.57	-0.038	1.780	0.905
Pull %	21.60	40.70	53.40	5.066

Table 1: Summary Statistics of Quantitative Variables

In looking at this table, there were a couple of things that stood out to us. First, we can see that the mean wRC+ for the players in the data set is 116.2, which is strange upon first glance, considering that wRC+ is standardized across the league to ensure that 100 is league average. However, the data set consists only of players who were qualified (had 490 at bats) for three seasons in a row. This means that these players are likely to be better players as they are consistently in the lineup, and therefore it makes sense that the average wRC+ of players in this data set is higher than league average. I also noticed that these qualified players are no more clutch than the average player, which confirms the widespread belief that there is no such thing as being an inherently clutch player. Finally, we see that the average pull percentage of players in this data is higher than 33%. This supports the argument for more teams shifting players to their pull side, and during this time period, many teams shifted for almost every batter. This will not be allowed during the 2023 season, and it is of extreme interest to see how that affects certain statistics such as wRC+.

Moving on to the categorical variables, position and hitting handednss, we observed the following summary statistics:

Infielders	Outfielders
34	23

Table 2: Distribution of Infielders and Outfielders

28 or Younger	Older Than 28
33	24

Table 3: Distribution of Age

Switch Hitters	Lefties	Righties
8	18	31

Table 4: Distribution of Hitting Handeness

We can see that there are slightly more infielders than outfielders in this data set, and this makes sense as there are 5 infield positions (including catcher) and 3 outfield positions (4 if we include the primary DH's who are secondary outfielders). Additionally, there are many more right handed hitters than there are left handed hitters or switch hitters. This is also consistent with what we would expect, as more people are taught to hit right handed at a young age.

Next, I examined some exploratory plots to view the relationship between these explanatory variables and wRC+.

It can clearly be seen in Figure 1 that there is a negative association between swing percentage on pitches outside the strike zone and wRC+ for the players in the data set. The best players by wRC+ tended to only swing at pitches outside of the strike zone between 15 and 20 percent of the time, while the worst hitters by wRC+ tended to swing at pitches outside of the strike zone between 40 and 50 percent of the time. This is a substantial difference in the percentage of the time different hitters swing at pitches outside of the strike zone, especially for hitters that are all consistent starters on their teams.

In examining Figure 2, the best players by wRC+ tended to swing at a fewer percentage of pitches inside the strike zone, but aside from these higher wRC+ observations, there appeared to be a slight positive association between the percentage of the time a player swings at a pitch inside the strike zone and wRC+. On first glance, this may appear to be contradictory, but we can attribute this phenomenon to the fact that the best hitters in the league do not swing at strikes that they know they cannot hit very well when they do not have two strikes. These pitches are often referred to as "pitchers pitches," and batters have a hard time hitting

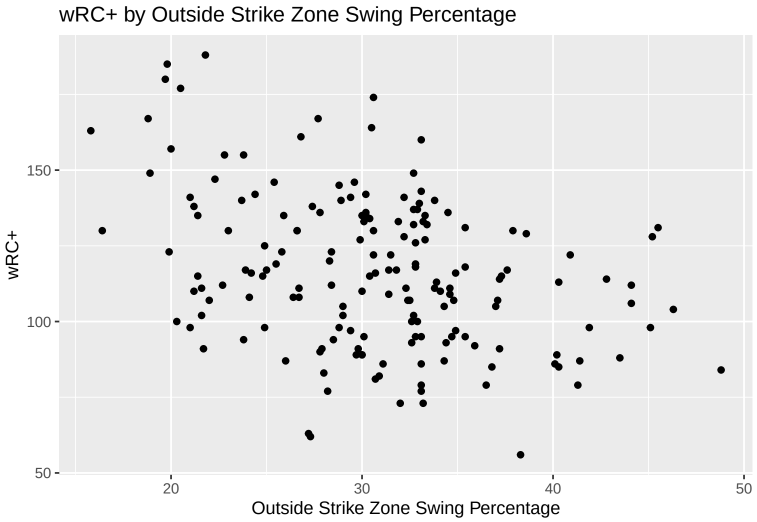


Figure 1: wRC+ by Outside Strike Zone Swing Percentage

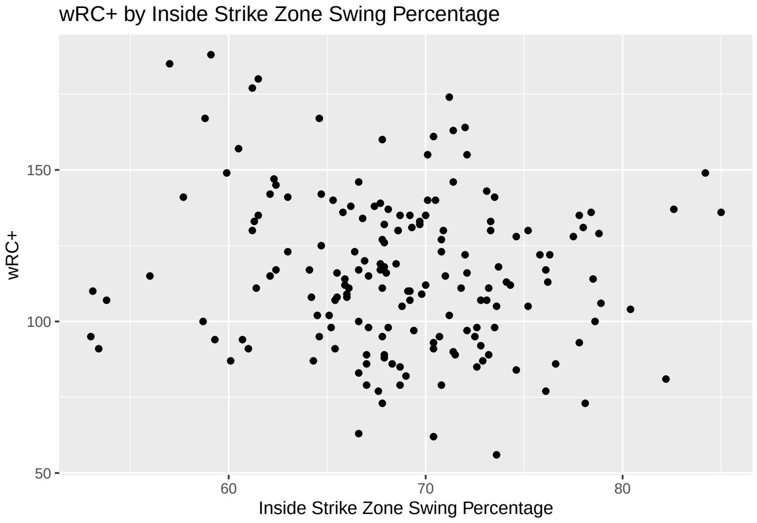


Figure 2: wRC+ by Inside Strike Zone Swing Percentage

these kinds of pitches hard. The best hitters are able to take these pitches, as well as pitches just outside the strike zone, and only swing when they know they will hit the ball hard. However, there are still many good hitters who swing at a higher percentage of strikes and are able to be successful, but maybe less successful than more selective hitters. This is the hypothesis as to why we see this slight positive trend when we ignore the best hitters by wRC+.

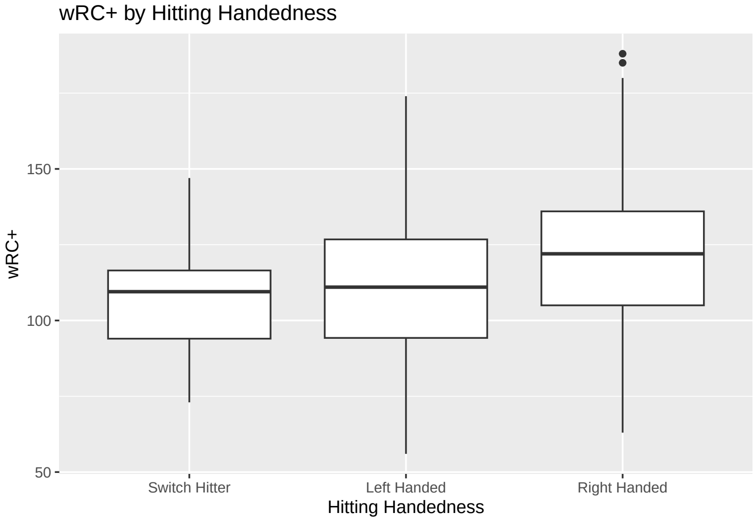


Figure 3: wRC+ by Hitting Handedness

In Figure 3, we saw an apparent relationship between the hitting handedness of a batter and their wRC+, with right handed hitters having a better average wRC+ than left handed hitters and switch hitters. I do not think that this has anything to do with right or left handed hitters being inherently better hitters, but rather due to the fact that some of the best hitters from 2017 through 2019 being right handed. It would be a great exercise to examine more historical data to see if right handed, left handed, or switch hitters had more success (a higher wRC+) in different eras of baseball.

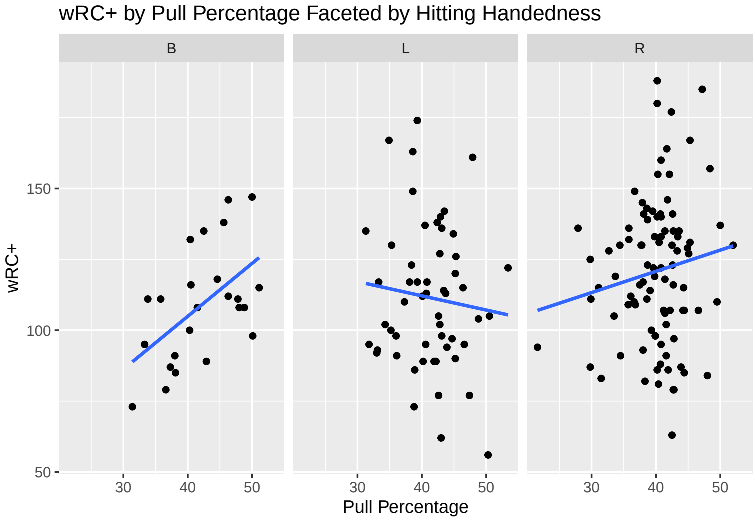


Figure 4: wRC+ by Pull Percentage Faceted by Hitting Handedness

Finally, in the exploratory analysis, I saw evidence for a potential interaction between the percentage of the time a batter pulls the ball and their hitting handedness when predicting wRC+. I noticed that for both switch hitters and right handed hitters, pull perecentage was associated with an increase in wRC+. This is something that I would expect, as players tend to hit the ball harder to their pull side, so players who are more consistently pulling the ball are more likely to be hitting more extra base hits and therefore are more likely to be more productive hitters. However, there was a negative association between pull percentage and wRC+ for left handed hitters. I believe that this is due to the fact that teams are able to shift the defense more for left handed hitters. Many lefties see defensive lineups that do not have a player to the left of the shortstop and instead put a player in shallow right field. This takes away lots of hits for lefties who pull the ball a higher percentage of the time, and therefore this negative association with wRC+ makes sense. Teams are not able to shift in the same way against right handed hitters because they must leave someone at first base, meaning we do not see this negative association for right handed hitters.

The first step in modeling the lesser thought of variables associated with wRC+ for a player between 2017 and 2019 was to start with an unconditional means model. This model showed that 55.73% of the variability in wRC+ was due to differences between players while 44.27% of the variability in wRC+ was due to variability within players. The next step in the modeling process was to determine if we needed a random slope for season. I found that a random slope was not needed in this model, so I moved on to the fixed effects for this model. I had seen evidence for potential interactions between the age of a player and the percentage of the time they pull the ball as well as evidence for potential interactions between position of a player and many of the other variables. However, after starting with a full model with all terms and these interactions, after conducting drop in deviance tests to ensure that we were keeping the statistically discernible variables, we were left with the following as the final model:

Level 1 (season i): $Y_{ij} = a_{ij} + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(0, \sigma^2)$

Level 2 (player j): $a_{ij} = \alpha_0 + \alpha_1 season_i + \alpha_2 outsideZoneSwingPct_i + \alpha_3 batsL_i + \alpha_4 batsR_i + \alpha_5 insideZoneSwingPct_i + u_i$ where $u_i \sim N(0, \sigma_1^2)$

Below is a table of parameter estimates and standard errors:

Parameter	Estimate	Standard Error	t-value
α_0	85.428	24.445	3.495
α_1	1.737	1.559	1.114
α_2	-2.252	0.419	-5.378
α_3	1.029	8.123	0.127
α_4	15.903	7.556	2.105
α_5	1.271	0.423	-3.003

Table 5: Parameter Estimates

After I finished this modeling process, I plotted standardized residuals against marginal means as well as conditional residuals against conditional means. These plots showed that the linearity assumption was met and that there were not any outliers that I needed to be worried about. I then checked the Cook's distance for this model to ensure that there weren't any observations that had too much influence on the model. I found that there were two observations that could be candidates for removal, so I removed them from the model and ran it again. The estimates, standard errors, and t-values did not change in a meaningful manner, and since we do not have that many observations to begin with, I did not remove these observations from the model. Therefore, the table above shows the estimates, standard errors, and t-values for the model run with all observations.

Based on this model, I can state the following: going from one season to the next (from 2017 to 2019) is associated with average wRC+ for qualified players increasing by 1.737 points. It would be interesting to apply an aging curve of some sort to this model; to see if this increase is due to the fact that there are younger players in the dataset. We are 95% confident that going from one season to the next is associated with an average change in wRC+ between -1.289 and 4.827 points.

We can also state that a unit increase in the percentage of swings on pitches outside the strike zone is associated with average wRC+ for qualified players decreasing by 2.252 points. We are 95% confident that a unit increase in the percentage of swings on pitches outside the strike zone is associated with average wRC+ for qualified players decreasing by between 1.439 and 3.055 points.

Conversely, a unit increase in the percentage of swings on pitches inside the strike zone is associated with average wRC+ for qualified players increasing by 1.271 points. We are 95% confident that a unit increase in the percentage of swings on pitches inside the strike zone is associated with average wRC+ for qualified players increasing by between 0.414 and 2.090 points. This shows that plate discipline is an extremely important metric for predicting productivity as a hitter, and it makes a lot of sense to see that players who swing at more strikes and don't swing at more balls are better hitters.

Finally, we see that being a left handed hitter compared to a switch hitter is associated with average wRC+ increasing by 1.029 points and being a right handed hitter compared to a switch hitter is associated with average wRC+ increasing by 15.903 points. We are 95% confident that being a left handed hitter compared to a switch hitter is associated with average wRC+ changing by between -14.557 and 16.603 points, and we are 95% confident that being a right handed hitter compared to a switch hitter is associated with average wRC+ changin by between 1.419 and 30.386 points. This change for right handed hitters compared to switch hitters appears to be drastic, but I hypothesize that this relatively large coefficient is due to the fact that there were a few dominant right handed hitters (Mike Trout and Mookie Betts) in the data set that caused this number to be inflated. However, it is worth exploring further why this is being observed.

Discussion

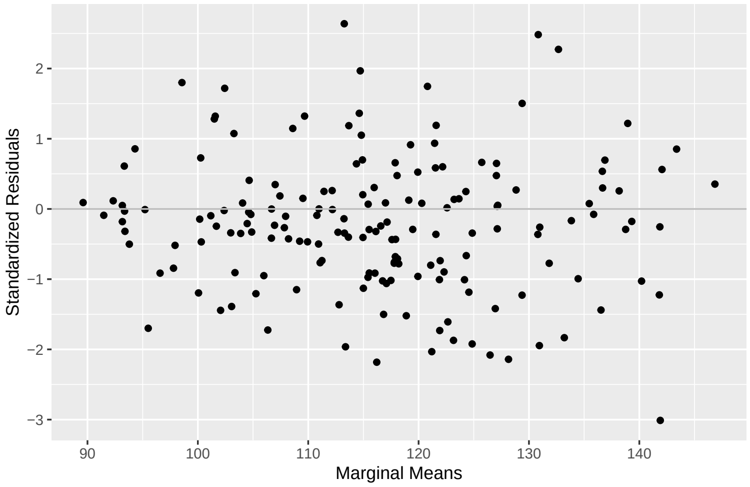
It is important to once again establish that I aimed to evaluate advanced statistics that have not been discussed as the more well-known metrics such as launch angle and exit velocity. Therefore, the model should not be determined as the most accurate model in predicting wRC+, but rather sheds light on the importance of not swinging at pitches outside the strike zone, swinging at pitches inside the strike zone, and the handedness of the batter.

A natural further study would be also including launch angle and exit velocity as potential explanatory variables, then cross-validating and determining if the "best" model includes more variables than just launch angle and exit velocity. If one were to find that a model that included more variables than launch angle and exit velocity is better at predicting wRC+, it can create a big advantage, especially when trading for a player. For example, let's say that the Minnesota Twins and Detroit Tigers are making a trade and each of the teams determines that the expected wRC+ is to be the best statistic to analyze player performance. However, the Minnesota Twins use a model only including exit velocity and launch angle, while the Detroit Tigers use a model that includes handedness of the batter, inside swing percentage, and outside zone swing percentage, along with launch angle and exit velocity. I will establish that the player the Minnesota Twins are trading will be player A and the player the Detroit Tigers are trading will be player B. The Minnesota Twins do past season analysis on each of these players, discovering that player A has an expected career wRC+ of 123, while player B has an expected career wRC+ of 128. Meanwhile, the Detroit Tigers also do past batted ball analysis on each of these players and discover that player A has an expected career wRC+ of 126, while player B has an expected wRC+ of 119. Given the results of their analysis, both teams are happy to make the trade; however, if one of the models that the teams used is not the best performing, then that slight change in expected wRC+ could mean that they actually traded a better player in exchange for a worse player.

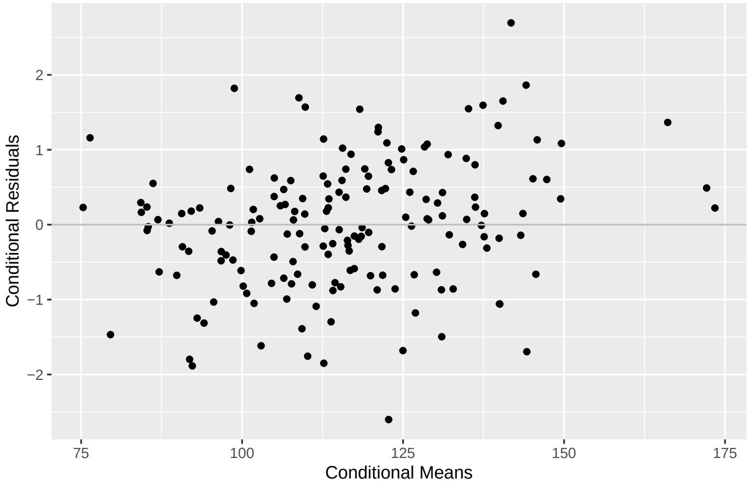
From the results, I briefly mentioned that the large increase in mean wRC+ for right handed hitters likely derived from a few dominant seasons from right handed hitters, particularly Mike Trout and Mookie Betts. After further investigation, I see that in 2018, Mookie Betts won MVP, while Mike Trout finished in second. Furthermore, in 2019, the top 9 AL MVP candidates were all right handed hitters, which further supports why there is such a large increase in mean wRC+ for right handed hitters.

Appendix

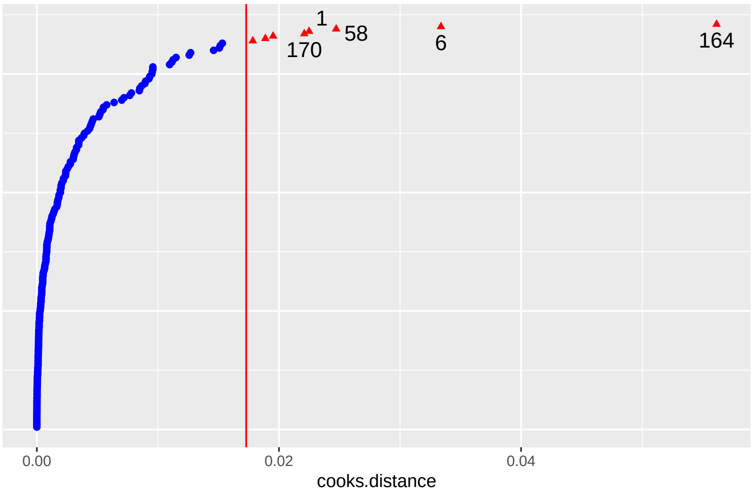
Appendix A: Standardized Residuals Against Marginal Means



Appendix B: Conditional Residuals Against Conditional Means



Appendix C: Cooks Distance



References

These data and studies were mentioned in the report

https://cran.r-project.org/web/packages/Lahman/index.html

https://www.si.com/mlb/2016/08/26/statcast-era-data-technology-statistics

https://cran.r-project.org/web/packages/Lock5Data/index.html

https://www.fangraphs.com/leaders.aspx?pos=all&stats=bat&lg=all&qual=y&type=8&season=2019&month=0&season1=2017&ind=0&team=0&rost=0&age=0&filter=&players=0&startdate=&enddate=