Analyzing a Player's Performances in Consecutive Seasons in the MLB

Kevin Waite

Abstract

Every year there is a new MLB season. This report assesses the relationship between a player's performance one season and a player's performance the season after. This relationship will be assessed for hitters, starting pitchers, and relief pitchers. This relationship will be examined using various regression techniques, with 3 final regression models being built.

Table of Contents

I. Introduction

Baseball lends itself well to statistical analysis of singular players. In most other sports, there are many factors to consider for an individual's performance, so often times statistics alone are unable to capture performance effectively. For example, in football a receiver may have a bad season statically due to bad quarterback play, or vise versa. In basketball, there is a lot of teamwork involved, so a player may do poorly statistically due to bad teammates around them, or because he lacks chemistry with his teammates. In baseball, however, an at-bat, a single instance of a batter vs. a pitcher, can serve as a discrete trial upon which inference can be made. There are still certainly a few outside factors that may affect a player's performance. A pitcher could have a bad defense around him. A batter could play in a home park that has large dimensions, making it harder to hit home runs. Overall, however, outside factors like these are much less prevalent in baseball than in any other sport.

I.i Problem

Major League Baseball, like most all other sports leagues, has a new season every year. Some players from the season before are unable to play in this new season, whether it be due to injury, age, or a multitude of other reasons. On the other hand, some players enter a new season having not played in the previous season, whether they are rookies, they were injured the previous season, etc. However, many players entering a new season played the previous season, and are going to play in the new season, and these are the players of interest for this report. This work seeks to explore how much of an effect a player's performance from that previous season will have on how he will perform in the new season, and what statistics are most significant in this effect. There are 162 games played per season, which is much more than any other sport, giving each season's data a large sample size.

Much like in other sports, in baseball there is an offense and a defense. The offense consists of the batters that come up to the plate. The defense consists of the opposing teams batters now playing in the field, as well as the pitcher, who is the one throwing the ball to the batter who is attempting to hit the ball. While the players playing in the field certainly do matter, the primary defensive player is the pitcher. The vast majority of plays that fielders make are routine plays that would get made correctly by any fielder the vast majority of the time. The pitcher, who is in charge of making good pitches that are hard for the batter to hit, is by far the most important player on a team's defense.

There are two different categories of pitchers: starters and relievers. A starting pitcher is the first pitcher to pitch for a team in a given game, and usually pitches about 5-7 innings, depending on how he's performing and how many pitches he throws per inning. A reliever enters the game after a starter, and typically pitches 1-2 innings (sometimes less than an inning even). A starter typically pitches every 5 games, since they throw so many pitches when appear in a game and need time to rest their arm. Relievers, since they do not pitch as many innings at a time, can appear much more often, sometimes even on consecutive days. Overall, a starter playing a full season will appear in about 30-34 games and pitch about 180-220 innings. A reliever playing a full season will appear in about 60-70 games and pitch about 60-70 innings. While pitchers who are usually starters can sometimes appear in a relief role, and vice versa, managers tend to be consistent in a pitcher's role.

Overall, there are 3 different categories of players that will be analyzed separately in this report: batters, starters, and relievers. To analyze these categories separately, 3 different models will be made.

I.ii Data

The data used for this report is from the Sean Lehman Database. The data from this database contains data for both pitchers and batters dating back to 1873, and contains double digit statistics for both. To keep the data relevant to the MLB today, the report uses stats from 2010-

2019. 2020 was not included because the season was shortened to only 60 games due to COVID-19.

II. Batters

II.i Response Variable

The response variable used for batters is weighted on-base percentage (wOBA), widely regarded as the best "tell-all" statistic used in today's MLB. wOBA is calculated as follows:

$$wOBA = \frac{w_{uBB}uBB + w_{HBP}HBP + w_{1B}1B + w_{2B}2B + w_{3B}3B + w_{HR}HR}{AB + BB - IBB + SF + HBP}$$

wOBA Rules of Thumb	
Rating	wOBA
Excellent	.400
Great	.370
Above Average	.340
Average	.320
Below Average	.310
Poor	.300
Awful	.290

This statistic is essentially on base percentage, but with weights on the different ways you can get on base. For example, a double is much more valuable than a walk, even though they both count as one plate appearance where a batter gets on base. So, a double will have a larger weight than a walk. These weights change slightly year by year to adjust for how batters performed on average that year, but not by much. The equation for wOBA in 2019 was:

$$wOBA = \frac{.69 \times uBB + .72 \times HBP + .89 \times 1B + 1.27 \times 2B + 1.62 \times 3B + 2.10 \times HR}{AB + BB - IBB + SF + HBP}$$

II.ii Data Preparation

The data was in an excel spreadsheet and formatted in a way that made it very easy to upload into R. After it was uploaded, the data was filtered out to only include players with a minimum number of plate appearances (an explanation on how this number was decided on is included later in this report). Number of plate appearances is simply how many times a batter got to bat. This was done in order to eliminate players who did not play enough from the model, as including them may lead to inaccurate results.

The dataset used did not include wOBA, but it did include all the statistics needed to calculate it. It was a fairly complicated process to pair each players' statistics each year with their wOBA for the next year. The first step in accomplishing this was to separate the data into separate data sets for each year. The player's wOBA was then calculated from each of these data sets and placed into a separate data set along with the corresponding playerID. The initial data sets each year were then merged with the player's wOBA for the next year by the corresponding playerIDs. In doing this, the resulting data sets only included players who had the minimum number of plate appearances for that season as well as the season after. Finally, these 9 data sets were merged on top of each other, resulting in the final data set, which included players' stats for each season, as well as there wOBA for the season after. This final data set had 1716 observations.

II.iii Predictor Variables

The table below includes all of the predictors used in the initial model. The data set had most of these variables as counting stats, so they were converted into rates by plate appearances. For example, HitRate = (# of Hits) / PA, i.e. the proportion of plate appearances that resulted in a hit.

Abbreviation	Term
PA	Plate Appearances
BBRate	Walk Rate
SORate	Strikeout Rate
HitRate	Hit Rate
DoubleRate	Double Rate
TripleRate	Triple Rate
HRRate	Home Run Rate
SBRate	Stolen Base Rate
IBBRate	Intentional Walk Rate
SFRate	Sacrifice Fly Rate
RunRate	Run Rate
HBPRate	Hit By Pitch Rate

III.iv Model

All three of the categories mentioned earlier will be analyzed using weighted multiple linear regression, with the weights being plate appearances for batters. Recall the form of a weighted linear regression model with response variable Y and predictors X_1, X_2, …., X_n:

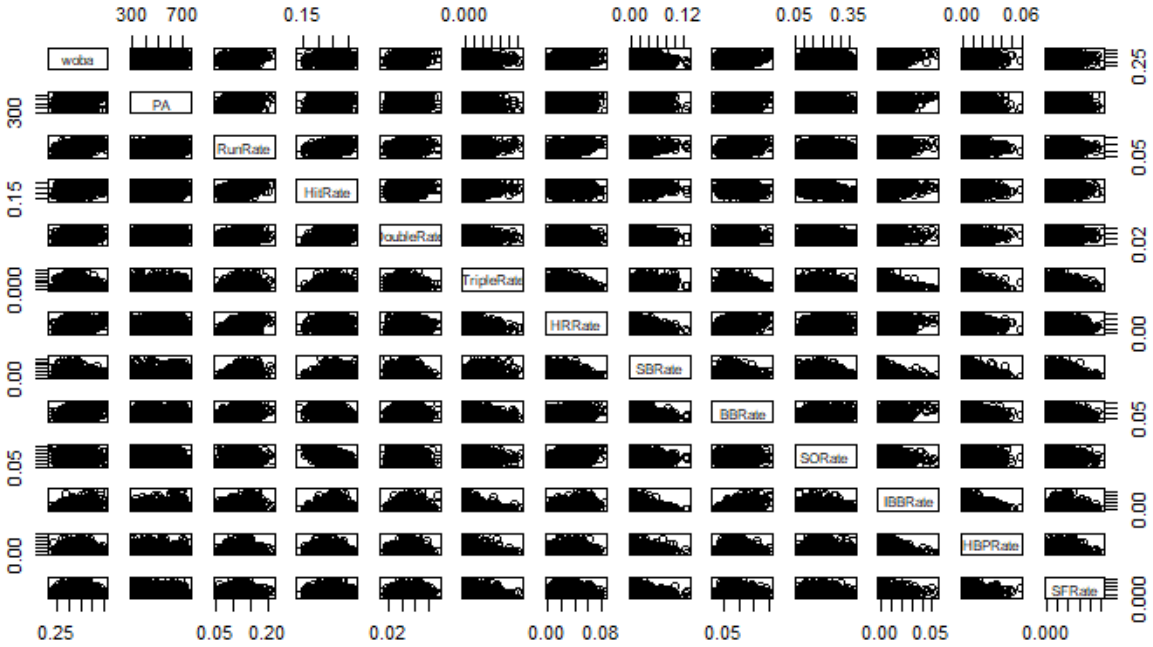
$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$$

$$Var(Y|X = x_i) = \frac{\sigma^2}{w_i}$$

where w_i is the weight of the ith observation.

III.v Exploratory Analysis

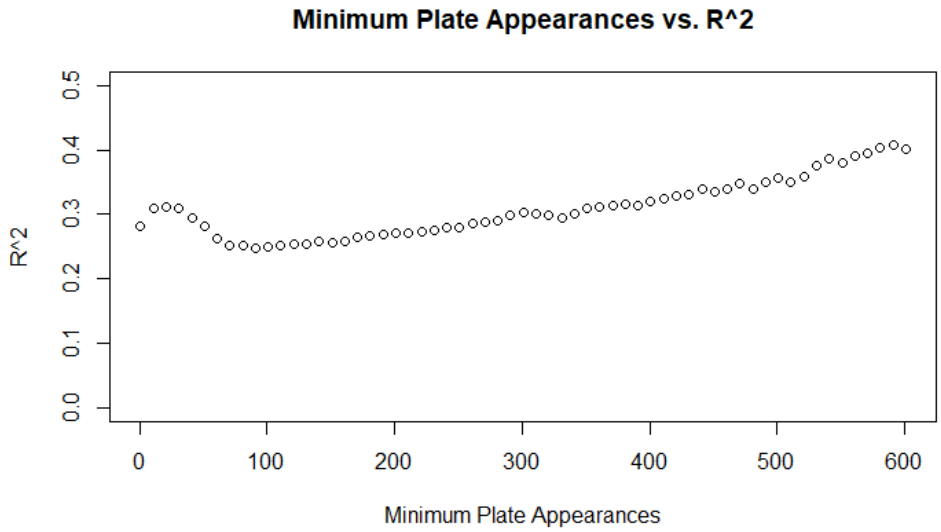
First order of business is to look at the scatterplot matrix of all the variables.



Judging from these graphs, there is clearly multicollinearity between the predictors, as to be expected, particularly in RunRate vs HRRate and HitRate vs SORate. There are also a lot of funny shapes in the graphs. Both these problems are hoped to be eliminated through transformations and variable selection.

As mentioned earlier, a minimum number of plate appearances is set to filter out the players' stats for years where they hardly played. In doing this, there are two conflicting ideas to consider. On one hand, you want to the minimum low enough to where you get enough data points, as well as the fact that you want the model to be applicable to a large enough range of data points where players did not play a whole season, but still played enough to be able to reasonably deem their stats from that season significant and useful. On the other hand, you want it to be high enough to where, inversely as just mentioned, you do not include data points where players did not play enough to for their stats from that season to be significant and useful. Below

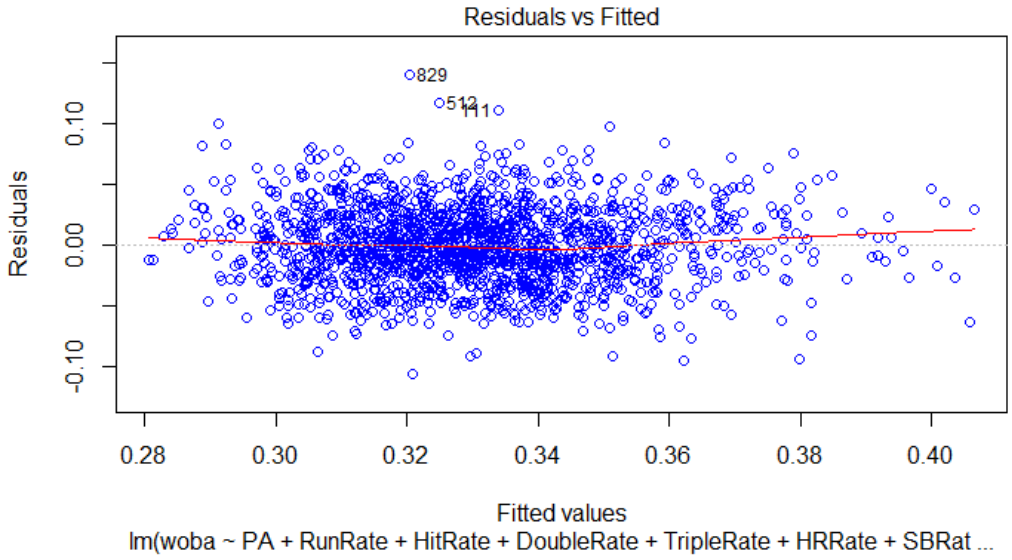
is a graph of the R^2 for the initial model vs. the minimum number of plate appearances used for the data.



The hope was that there would be some kind of "knee" in the graph that would make an obvious choice, but that does not appear to be so. There is a short spike at 300 PA, and that is about half a season's worth of plate appearances, so that is the number chosen.

II.vi Initial Model Fit

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.951e-01  1.175e-02  16.604  < 2e-16 ***
PA           2.465e-05  7.378e-06   3.340 0.000854 ***
RunRate      1.052e-01  5.452e-02   1.930 0.053806 .
HitRate      1.826e-01  4.355e-02   4.194 2.89e-05 ***
DoubleRate   3.204e-01  7.496e-02   4.274 2.03e-05 ***
TripleRate   3.637e-01  1.922e-01   1.892 0.058619 .
HRRate       6.196e-01  7.514e-02   8.246 3.24e-16 ***
SBRate      -4.481e-02  5.186e-02  -0.864 0.387675
BBRate       3.353e-01  3.480e-02   9.633  < 2e-16 ***
SORate      -2.927e-02  1.786e-02  -1.639 0.101474
IBBRate      3.557e-01  1.282e-01   2.775 0.005573 **
HBPRate      2.462e-01  1.062e-01   2.319 0.020538 *
SFRate       2.147e-01  1.994e-01   1.077 0.281835
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7119 on 1703 degrees of freedom
Multiple R-squared:  0.301,     Adjusted R-squared:  0.2961
F-statistic: 61.12 on 12 and 1703 DF,  p-value: < 2.2e-16
```



```
> ncvTest(batmodeltotal)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 17.59193, Df = 1, p = 2.7375e-05
> shapiro.test(batmodeltotal$residuals)

        Shapiro-Wilk normality test

data:  batmodeltotal$residuals
W = 0.99743, p-value = 0.006781
```

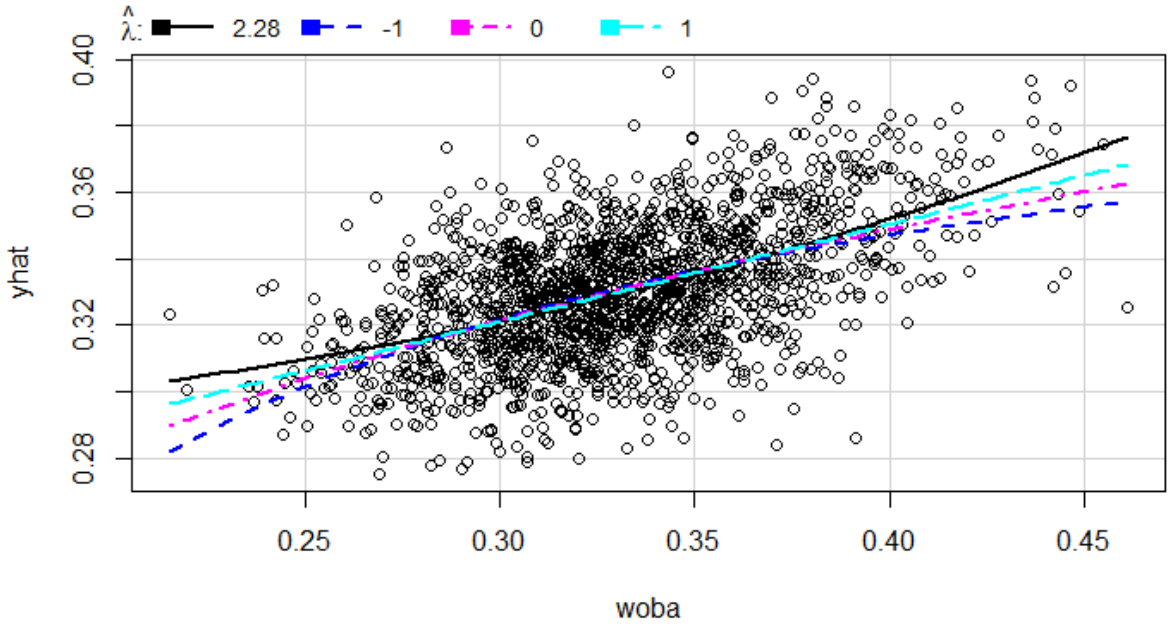
From the model summary for our initial model fit, there are several significant predictors, and a not good but decent R^2 of around 0.3, which is about what was expected. Also as expected from the scatterplot matrix, the constant variance and residual normality assumptions are unreasonable. In an attempt to fix these problems and eliminate some of the insignificant predictors, the next step is transformations and variable selection.

II.vii Transformations and Variable Selection

Before doing any transformations, a small constant needed to be added to any variables that have observations with a value of 0, since log and inverse transforms are undefined for those observations. Using the which function in R, it was found that TripleRate, HRRate, SBRate, IBBRate, HBPRate, and SFRate all had 0's in them, so 0.001 was added to all these variables. The rounded powers were then found using the powerTransform function in R, which gave the following:

Predictor	**Rounded Power**	**Transformation**
PA	1	None
BBRate	0.5	Square Root
SORate	0.5	Square Root
HitRate	1	None
DoubleRate	1	None
TripleRate	0	Log
HRRate	0.5	Square Root
SBRate	1	None
IBBRate	0	Log
SFRate	0.5	Square Root
RunRate	1	None
HBPRate	0.5	Square Root

Next, response transformations were explored.




```
> invResPlot(batmodeltransform)
     lambda       RSS
1  2.284057 0.4952514
2 -1.000000 0.5085552
3  0.000000 0.5017315
4  1.000000 0.4973036
```

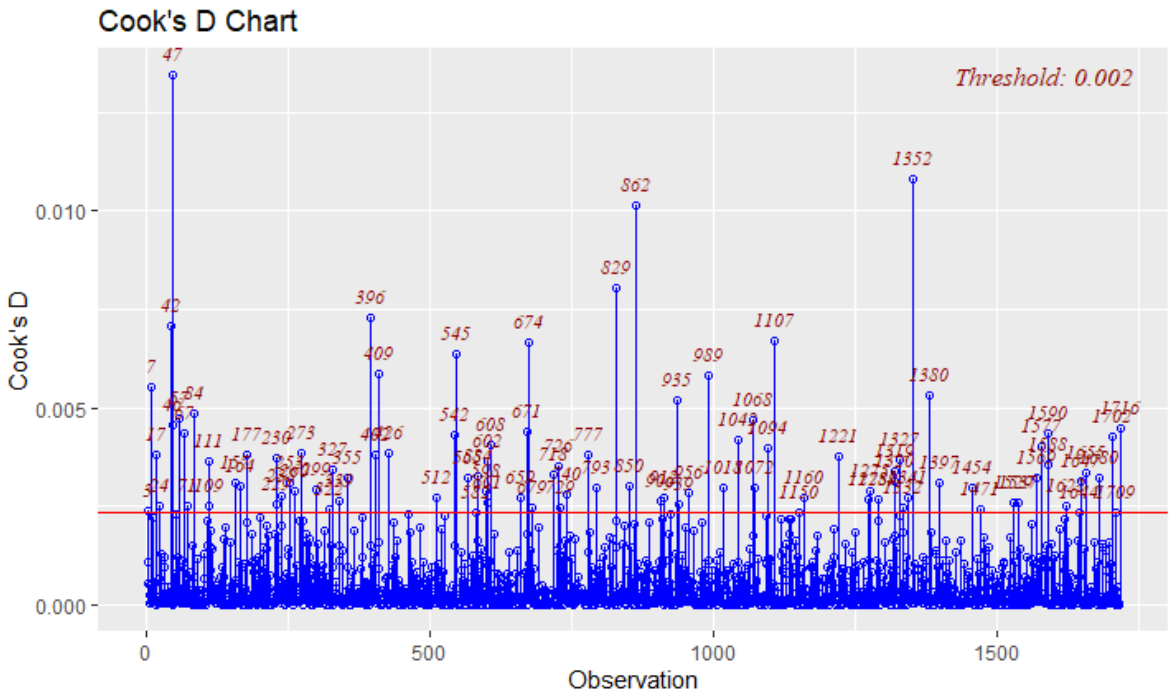
As you can see above, the response looks linear, and none of the response transformations significantly reduce the residual sum of squares. Thus, no transformation was made on the response.

Next, backwards stepwise regression was done using the AIC criterion. This eliminated SFRate, RunRate, and TripleRate. The new model was then examined.

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.678e-01  1.683e-02   9.974  < 2e-16 ***
PA                2.170e-05  7.426e-06   2.922  0.00352 **
HitRate           2.387e-01  3.849e-02   6.201 7.01e-10 ***
DoubleRate        2.895e-01  7.326e-02   3.951 8.09e-05 ***
I(sqrt(HRRate))   2.456e-01  2.062e-02  11.911  < 2e-16 ***
I(sqrt(BBRate))   2.173e-01  1.741e-02  12.480  < 2e-16 ***
I(sqrt(SORate))  -2.558e-02  1.477e-02  -1.732  0.08340 .
I(log(SBRate))    1.930e-03  6.905e-04   2.795  0.00524 **
I(log(IBBRate))   2.050e-03  8.837e-04   2.319  0.02049 *
I(sqrt(HBPRate))  5.436e-02  2.242e-02   2.424  0.01545 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.712 on 1706 degrees of freedom
Multiple R-squared:  0.2996,    Adjusted R-squared:  0.2959
F-statistic: 81.07 on 9 and 1706 DF,  p-value: < 2.2e-16
```

```
> ncvTest(batmodelfinal)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 20.66275, Df = 1, p = 5.4771e-06
> shapiro.test(batmodelfinal$residuals)

        Shapiro-Wilk normality test

data:  batmodelfinal$residuals
W = 0.99809, p-value = 0.04276
```

The R^2 remained about the same, and we have more significant predictors now. However, the model assumptions still do not hold. In an attempt to help this problem, some outliers were removed.



The top 3 outliers were removed. Two of them were players who severely underperformed, and one player who overperformed. However, even after doing this, the assumptions still did not hold.

II.viii Final Model

In a further attempt to fix this problem, the response transformation was revisited. A log transform worked wonders in fixing the constant variance assumption, but the normality assumption was still questionable.

Finally, after removing 3 more outliers, the model assumptions held, and a final model was obtained.

```
> summary(batmodelfinal2)

Call:
lm(formula = I(log(woba)) ~ PA + HitRate + DoubleRate + I(sqrt(HRRate)) +
    I(sqrt(BBRate)) + I(sqrt(SORate)) + I(log(SBRate)) + I(log(IBBRate)) +
    I(sqrt(HBPRate)), data = battotalfinal, weights = PA)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-7.2544 -1.3168  0.0263  1.4473  6.4623

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.618e+00  5.013e-02 -32.269  < 2e-16 ***
PA                6.775e-05  2.214e-05   3.061 0.002242 **
HitRate           7.241e-01  1.148e-01   6.308 3.59e-10 ***
DoubleRate        8.425e-01  2.187e-01   3.852 0.000121 ***
I(sqrt(HRRate))   7.235e-01  6.155e-02  11.755  < 2e-16 ***
I(sqrt(BBRate))   6.597e-01  5.184e-02  12.725  < 2e-16 ***
I(sqrt(SORate))  -6.884e-02  4.416e-02  -1.559 0.119207
I(log(SBRate))    4.733e-03  2.056e-03   2.302 0.021445 *
I(log(IBBRate))   5.043e-03  2.632e-03   1.916 0.055542 .
I(sqrt(HBPRate))  1.681e-01  6.673e-02   2.519 0.011868 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.117 on 1700 degrees of freedom
Multiple R-squared:  0.2993,    Adjusted R-squared:  0.2956
F-statistic: 80.69 on 9 and 1700 DF,  p-value: < 2.2e-16
```

```
> ncvTest(batmodelfinal2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.6191831, Df = 1, p = 0.43135
> shapiro.test(batmodelfinal2$residuals)

        Shapiro-Wilk normality test

data:  batmodelfinal2$residuals
W = 0.99853, p-value = 0.1477
```

In the end, we get a model with 7 significant predictors, an R^2 of 0.2993, and reasonable model assumptions. All the predictors are positively correlated with the response except for strike out rate, which is to be expected. The two most significant predictors are home run rate and walk rate, which will be discussed later in the report.

III. Pitchers

In an attempt at concision in this report, the procedures in obtaining the final models for the pitchers will be explained in a much briefer manner. The procedures were essentially the same as the procedure for the batting model.

III.i Response Variable

The response variable for the pitchers is Earned Run Average (ERA), a standard "tell all" statistic for pitchers. Unlike wOBA for the batters, ERA is a much simpler calculation that stays constant each year:

$$\textbf{ERA} = \frac{[(Earned\ runs\ (ER) \times 9]}{[Innings\ pitched\ (IP)]}$$

An earned run is charged to a pitcher when a run is scored by the opposing team while he is pitching that was not made possible due to an error by a fielder. The earned runs are divided by innings pitched to turn it into a rate of earned runs allowed per inning pitched, and then multiplied by 9 because that is how many innings are in a standard baseball game. So essentially, ERA is the average number of earned runs a pitcher would give up if he was pitching an entire game.

III.ii Predictor Variables

Like the batting predictors, a lot of these stats were counting stats, and consequently were turned into rates by batters faced, which is essentially how many plate appearances batters had against the pitcher for that season. Batters faced is also the weight used in the regression model, as pitchers who faced more batters are more important to the model.

Abbreviation	**Term**
ERA	Earned Run Average
BAOpp	Opponent Batting Average
BFP	Batters Faced
HRRate	Home Run Rate
BBRate	Walk Rate
SORate	Strike Out Rate
IPOBF	Outs Per Batter Faced
IBBRate	Intentional Walk Rate
WPRate	Wild Pitch Rate
HBPRate	Hit By Pitch Rate
GIDPRate	Double Play Rate
WPG	Wins Per Game
LPG	Losses Per Game
IPOGS (for starters)	Outs Per Games Started
SVPG (for relievers)	Saves Per Game

ERA means the ERA for the pitcher during the predictor season. NextERA is the name of the response variable in the models, meaning the pitcher's ERA the following season.

Some of these variables, such as HRRate, were used in the batting model, but they are not the same. For the batting model, HRRate rate was how many home runs the batter hit per plate appearance. In this model, it essentially means the home rate of all batters that season when said pitcher was pitching.

Some of these variables were handmade in R using the data and may be difficult to understand. Outs Per Batter Faced means the proportion of batters faced that the pitcher gets out. Outs per game started means the average number of outs the pitcher records in games that he starts. This is only relevant for starting pitchers for obvious reasons. Saves per game means how often a pitcher records the last out in a game they did not start when their team is winning and by no more than 3 runs. This stat could be important since managers tend to use their better pitchers

towards the end of tighter ball games. This stat, conversely, is only relevant for relief pitchers for obvious reasons.

III.iii Data Preparation

Preparing the pitching data was a bit different than the batting data, since it first needed to be split into the data for starters and the data for relievers. The criteria for a starter was G-GS<=3, meaning they did not appear in more than 3 games where they were not the starting pitcher. The criteria for relievers was GS<=2, meaning they did not appear in more than 2 games as the starting pitcher.

The rest of the data preparation was done in the same fashion as the batting data. The criteria for starters was a minimum of 18 games started, and the criteria for relievers was a minimum of 40 games appeared in. The final result was 2 data sets including the pitching data for each season as well as the ERA for the next. All the pitchers met the minimum requirement for both the predictor year as well as the response year. This left 691 observations for the starting data and 902 observations for the relief data.

III.iv Starting Pitcher Model

```
> summary(startmodelfinal2)

Call:
lm(formula = I(sqrt(NextERA)) ~ ERA + I(sqrt(HRRate)) + BBRate +
    I(sqrt(SORate)) + I(sqrt(HBPRate)) + I(log(IPOGS)), data = starttotalfinal)

Residuals:
     Min       1Q   Median       3Q      Max
-0.65515 -0.15023 -0.01055  0.13659  0.64881

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.37693    0.47425   9.229  < 2e-16 ***
ERA              -0.02374    0.01525  -1.557  0.11989
I(sqrt(HRRate))   0.98852    0.37044   2.669  0.00780 **
BBRate            1.29278    0.48745   2.652  0.00818 **
I(sqrt(SORate))  -1.73108    0.18283  -9.468  < 2e-16 ***
I(sqrt(HBPRate))  0.66215    0.32368   2.046  0.04117 *
I(log(IPOGS))    -0.63521    0.14050  -4.521 7.26e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.202 on 681 degrees of freedom
Multiple R-squared:  0.2329,    Adjusted R-squared:  0.2261
F-statistic: 34.46 on 6 and 681 DF,  p-value: < 2.2e-16
```

```
> ncvTest(startmodelfinal2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.01087215, Df = 1, p = 0.91696
> shapiro.test(startmodelfinal2$residuals)

        Shapiro-Wilk normality test

data:  startmodelfinal2$residuals
W = 0.99682, p-value = 0.1922
```

Though this is probability due to leftover multicollinearity that still exists between the predictors, it is quite humorous that ERA was not a significant predictor in predicting next season's ERA. The final model has 5 significant predictors, an R^2 of 0.2329, and the constant variance and normality assumptions are reasonable. One must keep in mind that for the batting data, a positive coefficient meant that an increase in that variable lead to a better performance by the batter, since a higher wOBA meant a better performance. That is not the case for ERA, as a higher ERA meant a worse performance the pitcher, since the pitcher wants to give up as few runs as possible. HRRate, BBRate, and HBP had positive coefficients, and SORate and IPOGS had negative coefficients, as to be expected. However, ERA has a negative coefficient, which is not to be expected. This is almost certainly due to leftover multicollinearity in the model. When taking ERA out, the adjusted R^2 and the assumption p-values all got worse, so ERA was not

taken out of the model. The most significant predictor by far is strikeout rate, which will be discussed later in the report.

III.v Relief Pitcher Model

```
> summary(reliefmodelfinal2)

Call:
lm(formula = I((NextERA)^(1/2)) ~ I(sqrt(ERA)) + I(log(HRRate)) +
    I(SORate^(1/3)) + I(WPRate^(1/2)), data = relieftotalfinal2,
    weights = BFP)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-15.0033  -2.8063   0.0594   2.8779  16.6854

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.22757    0.19947  11.168  < 2e-16 ***
I(sqrt(ERA))     0.24583    0.03962   6.205 8.39e-10 ***
I(log(HRRate))   0.05971    0.01965   3.039  0.00244 **
I(SORate^(1/3)) -1.06569    0.19459  -5.477 5.64e-08 ***
I(WPRate^(1/2))  0.64448    0.23654   2.725  0.00656 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.422 on 889 degrees of freedom
Multiple R-squared:  0.1523,    Adjusted R-squared:  0.1485
F-statistic: 39.94 on 4 and 889 DF,  p-value: < 2.2e-16
```

```
> ncvTest(reliefmodelfinal2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.474297, Df = 1, p = 0.49102
> shapiro.test(reliefmodelfinal2$residuals)

        Shapiro-Wilk normality test

data:  reliefmodelfinal2$residuals
W = 0.99734, p-value = 0.152
```

The relief model has an R^2 of 0.1523. Only 4 variables remained after the variable selection, and all 4 of these variables are significant. ERA, HRRate, and WPRate all have positive coefficients, and SORate has a negative coefficient. For this model, ERA is the most significant predictor, and has a positive sign, which is much more logical. The change in significance and practicality for

ERA from the starter model is almost certainly due to having less predictors, and therefore less multicollinearity.

IV. Results and Conclusions

IV.i Model Effectiveness

	Batters	Starters	Relievers
R_2	0.2993	0.2329	0.1523

The R^2 was significantly greater for batters than it was for pitchers, whether they were a starter or a reliever. This can be attributed to a couple of factors. The most obvious one is that the batting data had many more observations. Second, pitchers tend to get injured more (part of the reason for less observations), making in harder for them to keep a routine which is crucial in consistent results. But, perhaps the most important factor is that, while pitchers do not get a knock on their ERA when a fielder makes an error, they are certainly affected by the performance of the defense behind them. There are many cases where a fielder will make a fantastic play that most fielders could not make. Contrarily, there are many cases where a play perhaps should have been made and was not, but it was not enough of a miscue to be deemed an error. Both of these are out of a pitcher's control. While batters are subject to this as well, they play against many different defenses every season, while a pitcher has the same rotation of defenders around them all season.

Looking at the 2 categories of pitchers, the R^2 was much higher for starters than it was for relievers, even though the relief data had significantly more observations. This can be attributed to the fact that starters tend to face many more batters in a season than relievers.

```
> mean(relieftotal$BFP)
[1] 244.0266
> mean(starttotal$BFP)
[1] 734.4168
```

As you can see above, starters faced about 3 times as many betters per season on average than relievers, making their season statistics much less variable.

Overall, none of these R^2 were particularly high, but they all certainly were significant. A player's statistics the season before has a lot to do with how a player will perform in a given season, but there many other extraneous factors that affect his performance as well.

IV.ii Final Predictors and Their Significance

Batters:

PA	6.775e-05	2.214e-05	3.061	0.002242 **
HitRate	7.241e-01	1.148e-01	6.308	3.59e-10 ***
DoubleRate	8.425e-01	2.187e-01	3.852	0.000121 ***
I(sqrt(HRRate))	7.235e-01	6.155e-02	11.755	< 2e-16 ***
I(sqrt(BBRate))	6.597e-01	5.184e-02	12.725	< 2e-16 ***
I(sqrt(SORate))	-6.884e-02	4.416e-02	-1.559	0.119207
I(log(SBRate))	4.733e-03	2.056e-03	2.302	0.021445 *
I(log(IBBRate))	5.043e-03	2.632e-03	1.916	0.055542 .
I(sqrt(HBPRate))	1.681e-01	6.673e-02	2.519	0.011868 *

Starters:

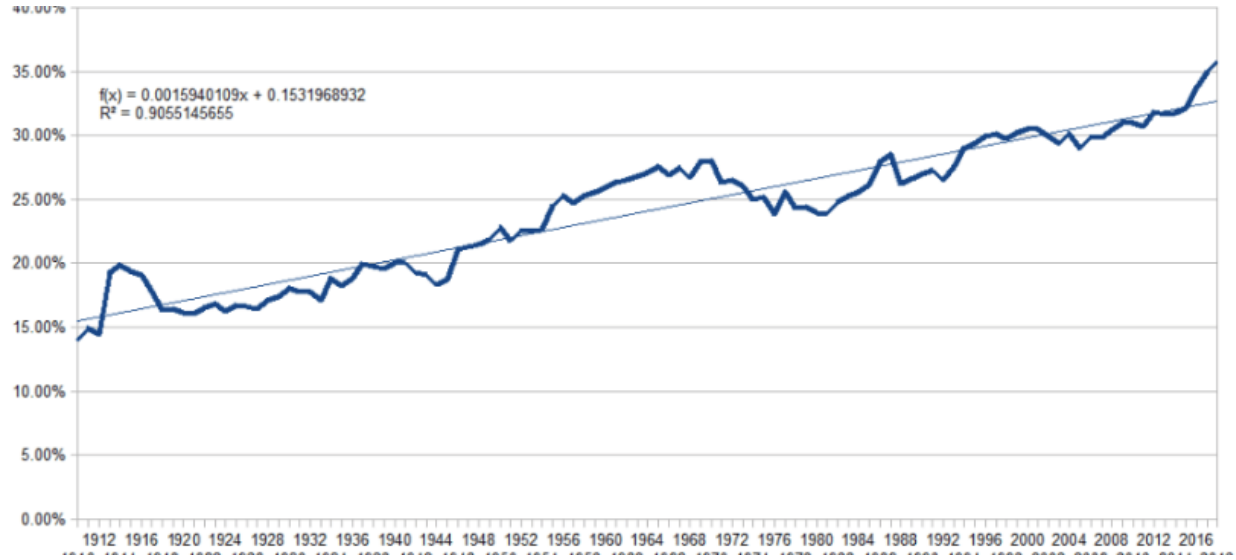
ERA	-0.02374	0.01525	-1.557	0.11989
I(sqrt(HRRate))	0.98852	0.37044	2.669	0.00780 **
BBRate	1.29278	0.48745	2.652	0.00818 **
I(sqrt(SORate))	-1.73108	0.18283	-9.468	< 2e-16 ***
I(sqrt(HBPRate))	0.66215	0.32368	2.046	0.04117 *
I(log(IPOGS))	-0.63521	0.14050	-4.521	7.26e-06 ***

Relievers:

I(sqrt(ERA))	0.24583	0.03962	6.205	8.39e-10 ***
I(log(HRRate))	0.05971	0.01965	3.039	0.00244 **
I(SORate^(1/3))	-1.06569	0.19459	-5.477	5.64e-08 ***
I(WPRate^(1/2))	0.64448	0.23654	2.725	0.00656 **

The 2 most significant predictors for batters were home runs and walks. The most significant predictor by far for starting pitchers was strikeouts. For relievers, strikeouts were almost as significant as ERA in predicting ERA. To quote the chicago.suntimes.com: "Baseball 2020 continues to be a three-true-outcomes game, with walks and strikeouts rising and home runs at

the third-highest per-game level in history." These aforementioned "three-true-outcomes" are walks, home runs, and strikeouts. They are considered the "true-outcomes" because they are the 3 end results of a plate appearance (outside of the rare hit by pitch) where the ball is not put in play. Below is a graph of the percentage of plate appearances resulting in a walk, home run, or strikeout since 1910.



This graph shows an obvious upward trend with an R^2 of 0.906, and it has been trending in an even more upward direction in the last 5 years.

In 2019, the Twins and the Yankees set the record for the most and second most, respectfully, home runs of any team in history. To make sure this is understood correctly, the Twins set the record, and if they had not, the Yankees would have.

In 2002, the General Manager of the Oakland Athletics defied baseball norms with the idea that a walk is as valuable as a hit. Rather than looking at a player's batting average (hits per at bat), he looked at a player's on base percentage (hits plus walks per plate appearance). This idea, which was greeted with much malignment by his peers, helped him propel a team with the lowest payroll in baseball to a then-record 20 game win streak.

The idea is simple, you want to hit a lot of home runs, and you want to have as many people on base as possible when you do. In the eyes of many in today's MLB, this is the most efficient and effective way to score runs.

For a pitcher, the strike out is so important for 2 reasons. The first is that when you strike a batter out, the batter is out. When a batter puts the ball in play, it is now out of the pitcher's control. The ball could be hit softly but happen to be hit in the perfect spot where no fielder is there. The fielder could have a shot to make the play but does not quite get to the ball in time. Whatever it may be, the pitcher has no control. To quote Dave Potts from rotogrinder.com: "The reason I look at strikeouts for my pitcher is that, besides being worth points (he's talking about points for fantasy baseball), they are the most predictable and consistent stat …. it so unpredictable what happens to a ball once it's hit into the field of play, you want to get as many balls in play as possible, giving yourself more chance for success."

One could then beg the question, why are strikeouts not a significant predictor then for batters? This is because the batter is the one putting the ball in play, not the pitcher. A batter could have a very fast swing which is prone to more strikeouts, but when he does hit the ball he will hit the ball much harder. The batter controls how fast he swings, how early/late he swings, and where he swings, not the pitcher. To quote Patrick Barron from M-SABR: "Much of this is due to the fact that a hitter has more consistency in and control over their individual batted ball statistics than pitchers do against opposing hitters."

The other, perhaps more important reason, is that pitchers who have more strikeouts tend to be better at pitching. One could say that an out is an out and it does not matter how you get it, and this is true in the short run, but not in the long run. The goal of almost every pitch for a pitcher is to have the batter swing and miss or look at a strike. When he is able to do that consistently, that means he is succeeding at what he is trying to do. Player's who pitch faster, who can put more spin on the ball, and who can locate the ball the best are the pitchers who get the most strikeouts. This is because these pitchers are able to make good pitches more consistently and successfully make the ball hard to hit for the batter.

While other stats are certainly still important, the three statistics: walks, home runs, and strikeouts, have taken over the way a lot of people analyze players in major league baseball, and the 3 models agree with that.

IV. References

- http://www.seanlahman.com/baseball-archive/statistics/
- https://www.fangraphs.com
- https://chicago.suntimes.com/2020/8/10/21362900/baseball-by-the-numbers-home-runs-strikeouts-and-walks-dominating-again
- https://rotogrinders.com/lessons/strikeouts-the-most-important-pitching-stat-736664
- https://msabr.com/2018/01/17/the-k-zone-why-strikeouts-affect-pitchers-more-than-hitters/