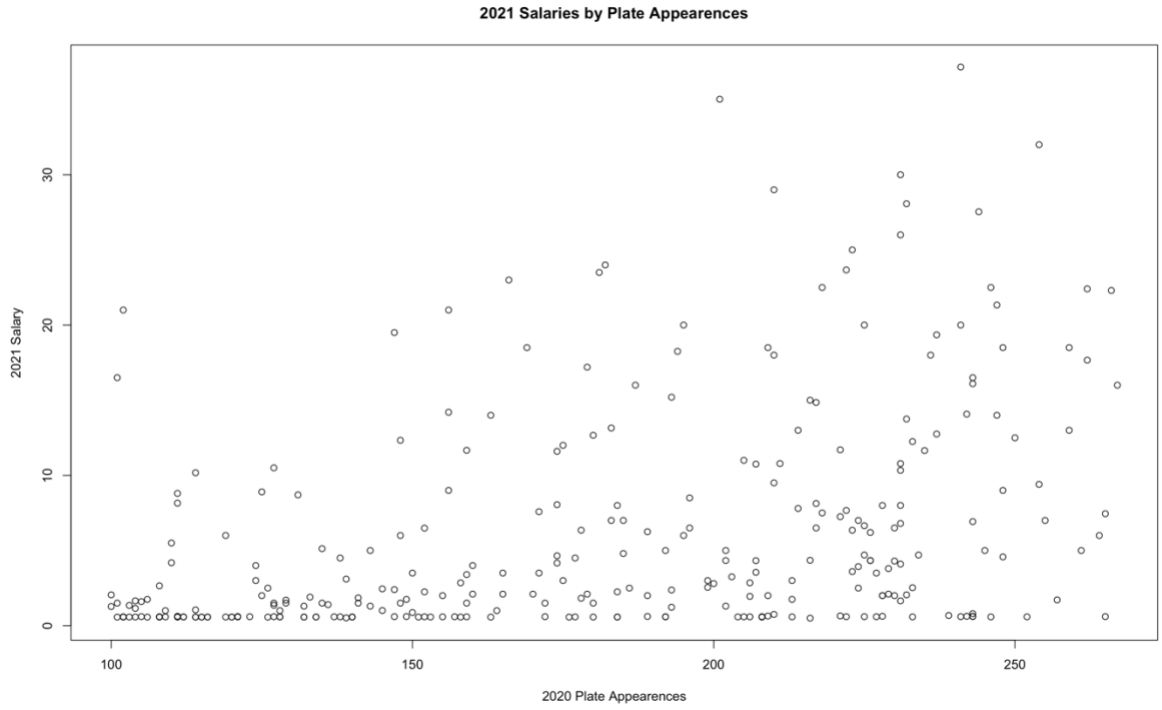
Predicting Baseball Players' Salaries Using the Past Year's Statistics

Introduction

In all professional sports the front office's job is to find players, assess their value in the larger market and pay them accordingly. Baseball has the strangest system of all sports, where control over a young player lasts the longest. Without there being a limit to what one team can spend on their roster, smaller market teams often lose their best players to the highest bidder after the rookie deals expire. Due to the long rookie deals, some of the best players in the league for a time have some of the lowest salaries. In this experiment, it will be determined if the batting statistics from the prior year are quality predictors of the next year's salary.

The data for this experiment will be from the 2020 shortened 60 game season of all batters with more than 100 Plate Appearances(baseball-reference.com) and the salaries will be from the 2021 season (spotrac). There are 295 players used in this study. Below are the salaries in millions of all the players included against their Plate Appearances.



This graph is in the introduction of this paper to display the uneven pay of the players in the MLB. As can be seen players who were everyday starters on the far right of the graph can be making league minimum. It can be seen from this that Plate Appearances are not going to be a very good predictor of salaries and my general null hypothesis and assumption going into this is that statistics are not good predictors of next year's pay; however, an attempt will be made to prove the null false and produce a model that can efficiently predict how much money a player will make in their next year of play.

The salaries were added to the batting statistics for each player and analysis will be done based off the following:

- Age - The age of the player in question. This is not technically a performance-based statistic and will be removed later in the study but is interesting to see its effect due to the nature of baseball.
- G - The number of games played by the player.
- PA - Number of Plate Appearances. Every time the player gets in the batter's box it counts as a plate appearance.
- R - Number of Runs scored.
- H - Number of hits, could be single, double, triple or homerun.
- 2B - Number of doubles.
- 3B - Number of triples.
- HR - Number of Homeruns.
- RBI - Number of runs batted in.
- SB - Number of stolen bases.
- CS - Number of times caught stealing, unsuccessful stolen bases.

- BB - Number of Bases on Balls, or walks.
- SO - Number of strike outs.
- BA - Batting average. Found as hits/at-bat.
- OBP - On-base percentage. Number of times a base is reached over the number of plate appearances.
- SLG - Slugging; A weighted average for batting production. Total Bases are divided by the number of at-bats.
- OPS - On-base plus Slugging. Combination of on-base percentage and slugging percentage. A tell-tale statistic of the efficiency of a batter.
- OPS+ - Listed in the code as "OPS.", the weighted average of OPS. Also considers the ballpark the player played in. 100 is league average, the higher the number the better.
- TB - Total bases the player reached by hits weighted by how many they got with that single hit. For example, 1(50)+2(25)+3(8)+4(20) = 50 singles, 25 doubles, 8 triples and 20 homeruns.
- HBP - Number of times hit-by-pitch.
- IBB - Number of times a player intentionally walked.
- X21.Pay - Players 2021 salary in millions of dollars.

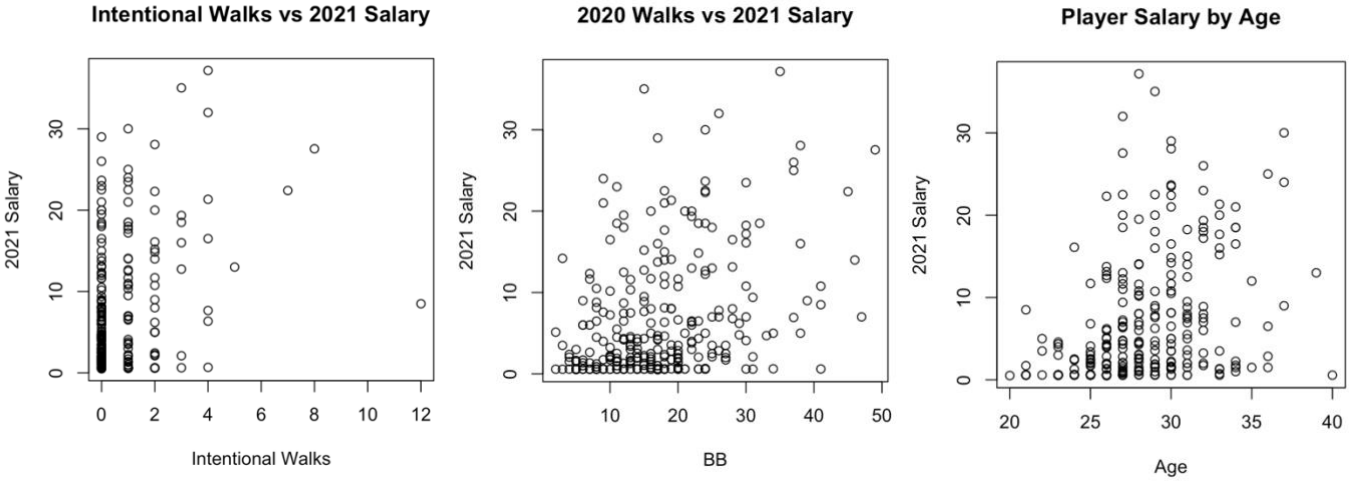
Methods

The first step was to randomly split players into a training and test set to conduct simulated predictions on. A multiple linear regression was employed, as well as a lasso, Principle Component Regression (PCR) and a PCS test all in that order. All these processes were used with the goal of comparing test errors and R-squared values to find the best possible model to

predict salaries. With variables like OPS, OBP, SLG and OPS+ different combinations were used to avoid influence of one on another's effectiveness. Another variable that was toyed with is player Age. This will be talked about in more detail in the results and discussion portion of the paper.

Results

To start, five multiple linear regressions were run to get a baseline on the effectiveness of each variable. After the initial full dataset was run as a model, there was an attempt to repeat the results with different combinations of variables. The variables that were repeatedly the most significant according to p-value were age, walks (BB) and sometimes IBB. Graphs of these variables against following year's salary will be listed below.

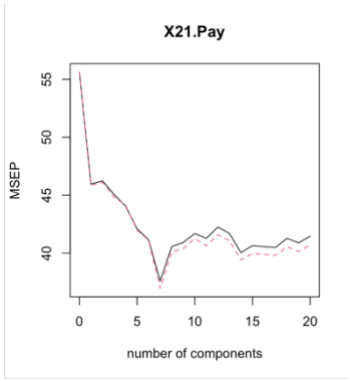


Based off the analysis with the linear models that have been conducted up to this point, it appears that the player age is doing most of the heavy lifting when it comes to predicting salary. This will be discussed in greater detail later, but when a model was run without player age it resulted in the worst R-squared at 36% and the worst test error rate of 45%. From the chart above, players follow consistent a nearly bell-shaped track when it comes to salary.

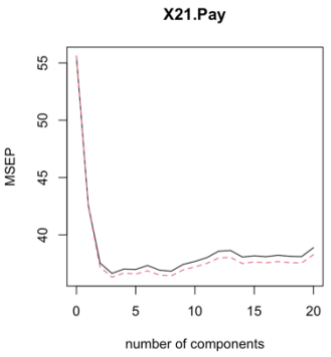
Another set of variables toyed with were the OPS, OBP and SLG variables, but switching them out ultimately had little effect with nearly identical R-squared and test-error rates.

Ultimately, the best model produced was the full model with every variable involved with an R-squared value of 48% and an error rate of about 40%. So far, the null hypothesis survives, and therefore switching to variable selection is necessary to find the optimal combination.

Next, a Lasso selection process was made to find another combination of variables that could possibly improve the model. A Lambda of 0.376 was chosen through a lambda.best selection as done in class and this resulted in the variables Age, PA, R, H, HR, BB, IBB being selected. A linear model with these variables was run and resulted in the best test error so far: 37%, but the R-squared value fell to 43%. This model was also repeated without player age which caused a significant increase in test error to 45%. From this point it still seems the model is still not great predictor of salaries, but there are two more methods to go through and select perhaps a new combination variable.



For the PCR, test 7 components were used based off the graph of predicted Mean Squared Errors (right). This process resulted in the best test error yet at 35%. This test error is still not of high enough accuracy to reject the null hypothesis and the following PLS test gave similar results. The PLS had 4 components and resulted in an error rate of 36%.



These results all suggest that the best this model can get to is in the 35-38% error rate range which is not low enough to be considered an effective model.

Discussion and Conclusion

There were a few different issues and limitations with this research. The biggest issue is that age is not a performance-based stat. Baseball players tend to follow similar career tracks and that is why the age vs salary graph appeared nearly normal. No one can work during the off

season to improve their age. Players play on their rookie deal and build up their portfolio, so they get signed to a bigger contract in their prime. Once those deals run out players get older, and the money gets smaller and smaller for the ones who manage to stick around. This trend is why Age will always be a good predictor of salary in an experiment like this because it is a big factor when it comes time to become free agents.

Which brings up the next big factor: free agency. Free agents are the only players who are getting truly performance-based pay for the next year. An interesting experiment would be to use a past year's statistics to predict the next year's salaries, but only for free agents. My hypothesis would be that there would be a very low test error rate. Sometimes guys do not live up to the salary they received or are being underpaid due to age or a recent improvement.

Baseball players are known to be streaky. If someone attended a baseball game for the first time ever and didn't know who any of the players are, statistics say there is a chance that person may see the greatest player ever go 0-4 with 4 strikeouts and they'd leave the ballpark thinking that guy stunk (Sidenote-This year I saw Mike Trout strike out 4 times in person and it was pretty neat). The streakiness of baseball players was on full display in the 2020 season because of the Coronavirus impacted 60 game schedules as opposed to the usual 162 games played by every team. This meant guys who were slumping who would normally come out of slumps ended up being in the slump for the whole season. In hindsight, data from the 2018 and 2019 seasons should have been used to have larger sample sizes as well as more applicable data to any other year of MLB history.

Based solely off the 2020 batter's statistics, salaries cannot be predicted effectively at a high rate. Perhaps with a different year selected and the incorporation of advanced metrics such as launch angle and bat-speed, a more accurate prediction could be made. Baseball is unlike any

other sport in the variety of data points and new knowledge that is being constantly created by very smart people, but at the end of the day investing in1 players is still gambling for an organization. It is still impossible to see the future and for the front office of a team to know if player will repeat past performances or fall back down to earth with the rest of us mortals.

References

Sports Reference. (n.d.). *2020 major League BASEBALL Standard batting*. Baseball Reference. https://www.baseball-reference.com/leagues/majors/2020-standard-batting.shtml#players_standard_batting.

Spotrac. (n.d.). *MLB Batter Salary rankings*. Spotrac.com. https://www.spotrac.com/mlb/rankings/salary/batters/.