Executive Summary

I created a linear regression model to predict how much salary a Major League player will make in free agency. This model will help determine which free agents are affordable and fit into the team's yearly budget. I used a player's WAR values from the previous three seasons, All-Star and MVP status, position, and contract length to predict his salary. I found that players that sign one-year deals or deals that are seven years or longer have lower intercept values than players that sign for two to six seasons. Pitcher's also command higher salaries for the same performance compared to their position player counterparts.

Introduction

The purpose of this model is to predict the salary a Major League baseball player will earn in free agency. Major League teams work under a budget built on player salary and this model will help to identify which players can be signed to improve the team's chances of winning without going over their yearly budget. I will create a model that uses on-field performance as well as career awards and honors to predict the player's salary.

Data

The data were collected from ESPN's MLB Free Agent Tracker, FanGraphs.com and Baseball-Reference.com and it includes any player that signed a Major League contract from the 2006 to 2019 off-season. The model will be predicting the dollar amount the player signed for in present value divided by the length of the contract. Major League Baseball salaries have continued to increase every year and I believe that it would be easier to model the present value salary instead of introducing another variable into the model to control for the year the player signed his contract. I decided to use a 3% inflation rate and it seems to be a good facsimile for annual salary in 2020 terms.

I obtained the WAR values from the player's previous season and the two years prior from both FanGraphs and Baseball-Reference and averaged them together to create two new columns called MixedWAR_1 and MixedWAR_2. WAR is a comprehensive metric used to gauge the overall on-field value a player provided to his team. FanGraphs and Baseball-Reference have similar methodologies, but there are slight differences in the way the stat is calculated that can create a discrepancy for the overall value a player generated for his team. Therefore, I decided to average the two values together into one column. I decided to separate the most recent season from the other two seasons, because the most recent season is more predictive of a player's future performance than what a player produced three seasons ago. This could be due to either injury, aging or a change in skill level from one season to the next.

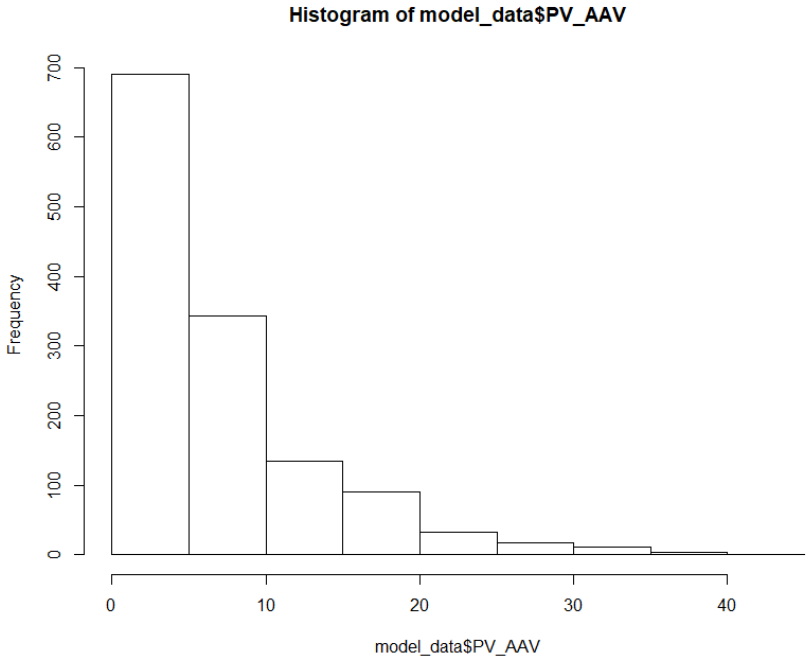
The next thing I wanted to account for other than on-field performance was a player's perceived value. A team will sometimes pay a player more for their past accomplishments and perceived upside than a player who has produced similar value over the last several years. I attempted to simulate this effect with two categorical variables called MVP_Candidate and All_Star. If a player ever played in an All-Star game or received an MVP vote, they were counted as a yes in these categories.

The last variable I decided to account for was if the player was a hitter or a pitcher. These two positions have different jobs, and it is quite likely that they are compensated differently. Pitchers are more prone

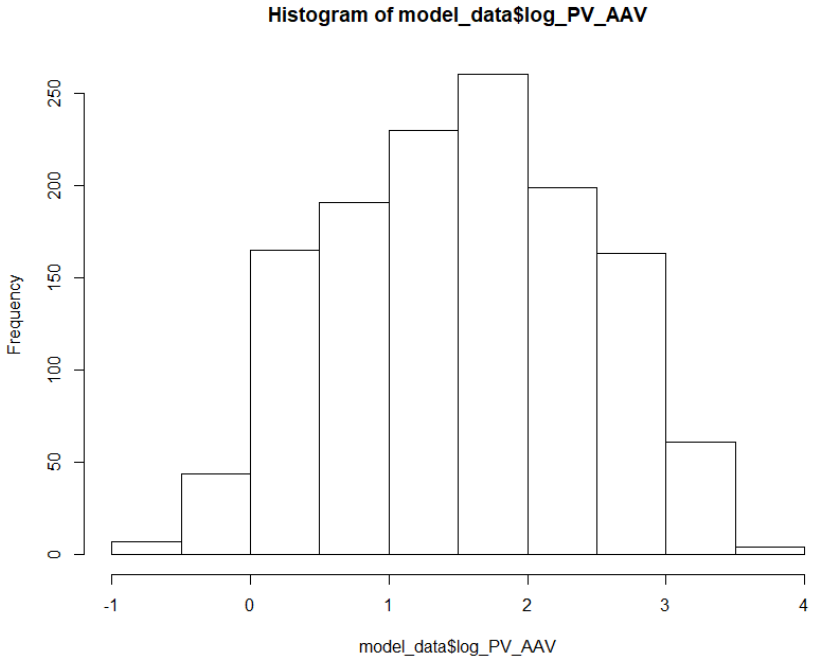
to injury, but they are also in higher demand, because they are easier to find playing time for than position players.

There were not too many challenges in collecting the data for this project since Major League Baseball is skilled at checking their data. The only thing I needed to do was omit any player that had signed to play in a different league or anyone that played in a different league but had not played in the Majors in the previous season. I excluded these players, because I do not think it would be fair to include a player's statistics from another league as Major League statistics or to completely ignore their on-field production in a less competitive league and give them a zero for the previous year's WAR value. If a team wants to sign a player that is coming from a different country's league, this model will not apply to them.

I plan on using a linear regression model, but first I need to see what distribution would be appropriate for my predicted values. Below is the histogram of a player's present value salary.



Clearly the data is right skewed, so using a standard normal distribution would not be an appropriate choice. Perhaps transforming the values using the logarithmic function will give a more normal distribution.



This is much better. Going forward I will be using this transformed column as my dependent variable.

Model

In the first model I decided to use the same non-informative prior that we used in the class example with all betas following a standard normal distribution (0, 1000000) and a precision that follows a gamma distribution (2.5,25) and a likelihood function of

$Y_i = \beta_1 + \beta_2$*MixedWAR_1$_i + \beta_3$*MixedWAR_2$_i + \beta_4$*MVP_Candidate$_i + \beta_5$*All_Star$_i + \beta_6$*Pitcher$_i$

that follows a standard normal distribution. I chose non-informative priors, because I did not have any preconceived notions of what the distribution should look like.

This model should be appropriate now that I have transformed my response variable to better resemble a standard normal curve and the WAR parameters approximately follow a standard normal curve and the other variables in the model are explanatory variables. Each of these parameters should help to identify how much of a raise to expect in salary when WAR is increased as well as how much value making an All-Star game or appearing on the MVP ballot is worth. This model accounts for on-field performance as well as career accomplishments and should be able to reasonably predict a player's salary for the upcoming season.

All the parameters in the model converge and there is minimal autocorrelation. The residual plot looks random and the normal QQ plot looks like a reasonably straight line, although the model does seem to have a little trouble with extreme outliers. The model does seem to have trouble overestimating the salary of high performing players. This may be because many top performers take longer term deals that artificially lower their salary but guarantees more money overall. The baseball industry usually makes these deals because it lowers the player's salary, and this allows the signing team to have more

flexibility to stay under the luxury tax threshold while the player gets greater security even when his performance starts to suffer. The DIC for the first model is 2461.

I will attempt to account for these longer-term deals by creating a hierarchical model that creates groups based on contract length. Group 1 will be for one-year deals, Group 2 will be for two-year deals, Group 3 will be for three to four-year deals, Group 4 will be for five to six-year deals and Group 5 will be for deals that are seven years or longer. All the parameters in the model converge and there is minimal autocorrelation. The residual plot looks random and the normal QQ plot looks like a reasonably straight line, however the model does seem to still have trouble with extreme outliers, but the DIC has decreased to 2219 which means that the newer model is superior.

Results

The mean coefficients for the model are as follows:

Intercept for 1-year deals = .45

Intercept for 2-year deals = .90

Intercept for 3-4-year deals = 1.12

Intercept for 5-6-year deals = .94

Intercept for 7 or more-year deals = .35

Coefficient for previous season's WAR = .20

Coefficient for previous two season's WAR = .09

Coefficient for MVP candidate = .03

Coefficient for All-Star appearance = .13

Coefficient for being a pitcher = .23

These coefficients show that pitchers receive a sizable bump in salary compared to position players, and surprisingly that an All-Star appearance is worth more than being an MVP candidate. The model also shows that length of the deal has a huge impact on a player's salary. As expected, longer term deals have a lower coefficient, but one-year deals are quite low as well. This is probably because many players that secure one-year deals are bench players that do not command as much salary and teams are not willing to commit to bench players for multiple years and this could be the reason why the intercept is so low. This probably means that the model will underestimate prominent older players who sign one-year deals not due to a decrease in performance, but because they are close to retirement and do not desire a long-term deal. If I could improve the model, I would probably exclude any deal over 7 years to try and eliminate some of the skew in the response variable.