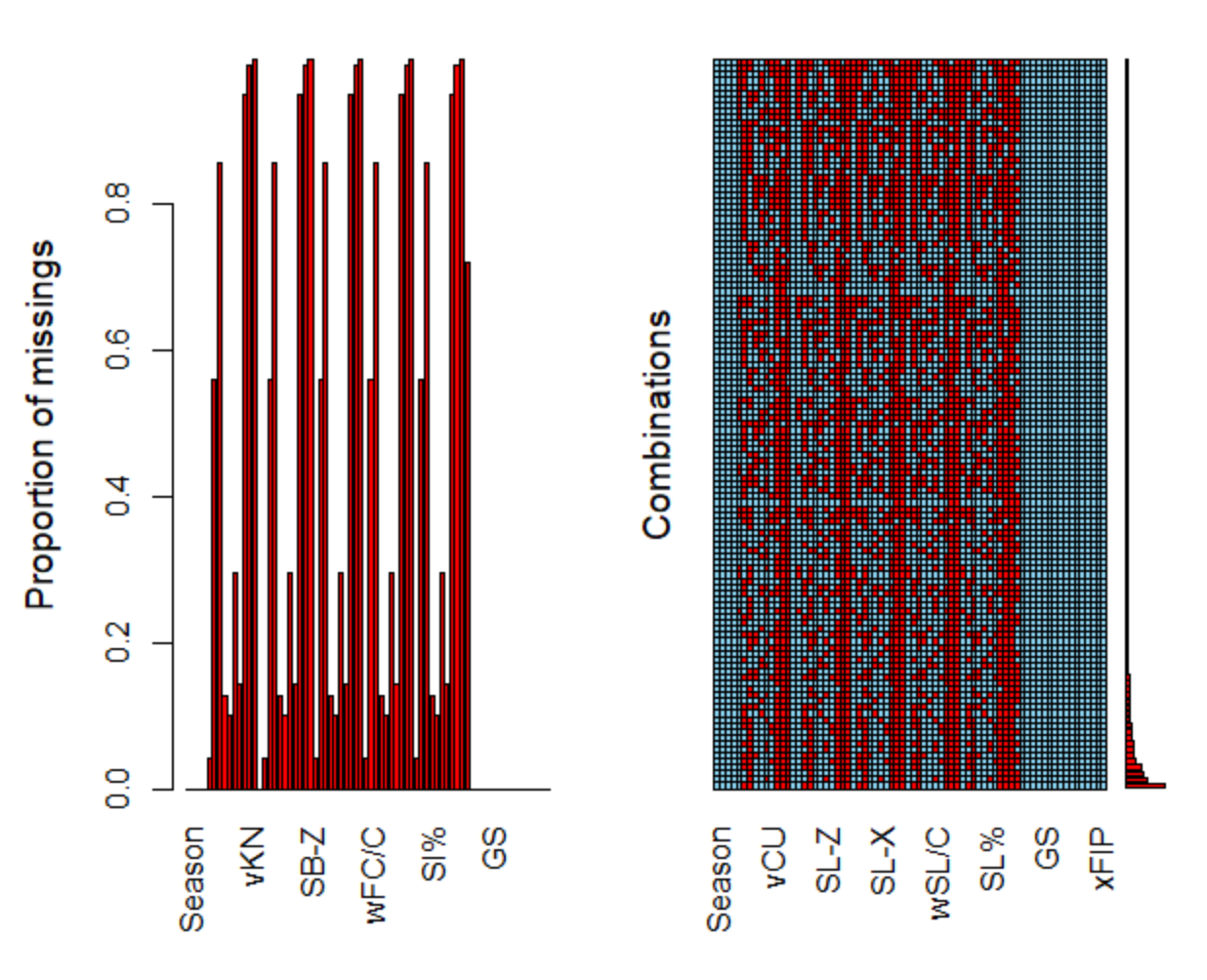
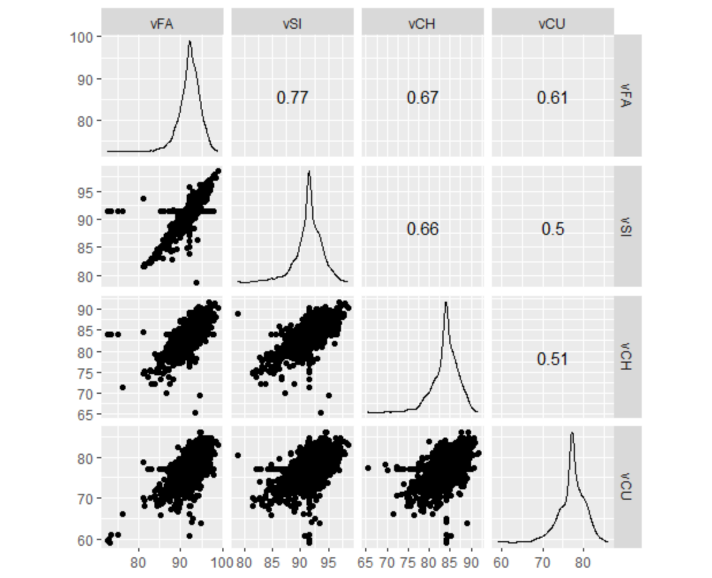
The driving force behind this project was my research question: “How can I predict pitcher success in the MLB?” The first question one might ask is, “What is ‘pitcher success’ and how is it measured?” Well, there are several things someone can look at to summarize how well a pitcher has performed in a game, season, or even career. One of the most basic statistics for a pitcher that summarizes their performance is called “earned run average” (ERA). This statistic summarizes, on average, how many runs that pitcher let score per nine innings. This is an excellent and easy-to-understand statistic that is widely used in pitcher analysis. However, one flaw with this statistic is it does not necessarily evaluate all players on an equal level. For example, relief pitchers, pitchers who only pitch a couple innings every couple of games, tend to have lower ERA’s than starting pitchers due to the fact that they are specialists in getting a small number of batters out. Furthermore, it is possible for a relief pitcher to have a lower ERA than a starting pitcher but clearly be less valuable. Therefore, I have decided to use a statistic called “Wins Above Replacement” (WAR) as a measure of pitcher success. WAR, as FanGraphs puts it, is: “an estimate to answer the question, ‘If this player got injured and their team had to replace them with a freely available minor leaguer or a AAAA player from their bench, how much value would the team be losing?’” And this is measured in wins, hence the name “wins above replacement.” Since one cannot simply look at one statistic to evaluate a pitcher’s “success”, we will take this into account when making conclusions.

All the data I used in this project came from fangraphs.com. Fangraphs has lots of publicly available data gathered all the way from the 1871 season. There are many useful statistics that were used in this project; however, many are relatively new and therefore were only available in a smaller sample size, which is why I was only able to use the seasons 2007-2019 for this project. I scraped all relevant data I could find from Fangraphs.com and loaded it all into R and combined everything into one dataset. After doing this, I had one dataset with 1,788 observations and 72 variables. Each observation represents one season’s statistics for one player. Each variable represents a certain statistic that was recorded for that pitcher (e.g. WAR, ERA, Wins, as well as many unique statistics involving specific pitches that pitcher threw).

One of the first problems we run into when setting up a multiple linear regression model is the amount of missing values in our data. Model selection procedures are difficult to perform with any missing data. Since there are so many unique variables, especially involving specific types of pitches, there were several variables that had the majority of data missing, since most pitchers do not throw that type of pitch. I used the package “VIM” to visualize how much data from each variable was missing. It is impossible to distinguish which variables contain the large amounts of missing data; however, it does give us a good idea as to what the data looks like as a whole.

By looking at the plot on the left, you can see that there are quite a few variables that have more than half of their data missing. Before I proceeded, I decided variables that had more than 20% of their data missing did not have substantial data to be interpolated. So, I removed all variables that had more than 20% missing data. This brought the dataset down to 41 variables (we removed 31 variables!). Since it is extremely difficult to run a forward, backwards, or stepwise model selection procedure with any missing data at all, I had to interpolate the rest of the data. I used the package “mice” to perform this interpolation. Using this package, I replaced all missing values within each variable with the mean of that variable. This was easy since all the data I am using is numeric and both non-binary and non-categorical. The next thing that comes up is the evident multicollinearity between the variables in this dataset. Here is a small scatterplot matrix showing the multicollinearity between four variables: the average velocity of a pitcher’s fastball, sinker, changeup, and curveball.

This multicollinearity exists throughout the data and must be addressed. It is for this reason why my research question is centered around prediction; since multicollinearity does not really affect a model’s ability to predict.

I also added three variables to the data. Since I have variables that measure the vertical and horizontal movement on different pitches, I created variables that contain the differences between vertical, horizontal, and total movement between a pitcher’s fastball and curveball. I was interested to see if, not only does a pitcher have a nasty curveball, but does a pitcher also have an excellent fastball to complement that curveball, and does that affect the pitcher’s WAR?

Furthermore, the method I used to fit a multiple linear regression model so I can best predict a pitcher’s WAR was simple. First, I fit a model with all possible predictors. Then, I performed a forward, backward, and stepwise model selection and select the model with the lowest AIC. Lastly, I analyzed the final model. There were 32 total predictor variables that went through the model selection process. Knowing that observations represent averages for one pitcher in a single season, the possible predictors were: strikeouts(K), walks(BB), and home runs(HR) allowed per nine innings, ERA, FIP, xFIP, opponents’ batting average on balls in play (BABIP), HR/fly ball(FB) rate, left on base percentage, velocity of fastball, sinker, changeup, and curveball, vertical movement of fastball, sinker, changeup, and curveball, horizontal movement of fastball, sinker, changeup, and curveball, pitch values (look at reference on reference page) of fastball, sinker, changeup, and curveball, usage rate of fastball, sinker, changeup, and curveball, and lastly the three added variables I addressed earlier.

Here is the final model output:

Call:

lm(formula = WAR ~ vFA + vCU + FA.Z + SI.Z + SI.X + CH.X + wCH.C +

wCU.C + CH. + FA.CU.DIFF + K.9 + BB.9 + HR.9 + LOB. + HR.FB +

xFIP, data = all\_noNa)

Residuals:

Min 1Q Median 3Q Max

-3.1715 -0.5221 0.0010 0.4920 3.5196

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.072706 0.948275 -0.077 0.938893

vFA 0.024744 0.009910 2.497 0.012625 \*

vCU -0.012490 0.006364 -1.963 0.049858 \*

FA.Z 0.083851 0.021395 3.919 9.22e-05 \*\*\*

SI.Z -0.041834 0.015512 -2.697 0.007064 \*\*

SI.X 0.013717 0.005910 2.321 0.020406 \*

CH.X -0.013543 0.006172 -2.194 0.028346 \*

wCH.C 0.015268 0.007104 2.149 0.031754 \*

wCU.C 0.015926 0.009329 1.707 0.087974 .

CH. -0.591017 0.300317 -1.968 0.049226 \*

FA.CU.DIFF -0.028423 0.007844 -3.623 0.000299 \*\*\*

K.9 0.490004 0.020608 23.777 < 2e-16 \*\*\*

BB.9 -0.795971 0.036646 -21.721 < 2e-16 \*\*\*

HR.9 -2.120599 0.157405 -13.472 < 2e-16 \*\*\*

LOB. 2.366273 0.443679 5.333 1.09e-07 \*\*\*

HR.FB -5.034308 1.366399 -3.684 0.000236 \*\*\*

xFIP 0.212241 0.081723 2.597 0.009480 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.818 on 1771 degrees of freedom

Multiple R-squared: 0.7437, Adjusted R-squared: 0.7414

F-statistic: 321.2 on 16 and 1771 DF, p-value: < 2.2e-16

Although we started with 32 predictors, the model selection process eliminated exactly half of those to leave us with 16 predictors used to predict a pitcher’s WAR. This model gives us an adjusted r squared value of 0.7414, which is relatively high. This means that about 74% of the variation in WAR can be explained by the predictors that actually affect a pitcher’s WAR. Notice how close the adjusted r squared is to the multiple r squared, which tells us that most, if not all, of these predictors do affect WAR.

As can be seen in the model output and analysis, this model does a very good job at predicting a pitcher’s WAR for a single season, which is a variable we are using to equate to a pitcher’s “success”. Although this model is good at what its intended use is, predicting, it cannot be interpreted with anything more than that. Due to the very high multicollinearity, the coefficients, sum of squares, etc., might be skewed because of this property in our data.

All the time, in the industry of baseball, experts are trying to predict the futures of players by their performance right now. Oddly enough, many superstars in the MLB were not thought of as superstars when they were scouted and ultimately selected to join a team. Some argue that this is the most challenging part of professional baseball management. So many more factors come into play when deciding whether a player will be able to excel on **your** team or not. One must look beyond numbers to truly decide this. The analysis performed in this project only has so much power. First, it would be extremely difficult to use this model on players outside the MLB. Furthermore, while this model does have use, one must understand exactly what it can be used for in order to get the most out of it.

References

WAR definition: <https://library.fangraphs.com/misc/war/>

All data: <https://www.fangraphs.com/leaders.aspx?pos=all&stats=pit&lg=all&qual=0&type=8&season=2019&month=0&season1=2019&ind=0&team=0,ts&rost=0&age=0&filter=&players=0>

Pitch value definition:

<https://library.fangraphs.com/offense/pitch-type-linear-weights/>