Single or Multiple Variable Regression: Picking the Best Fantasy Baseball Team

Chris Lynch

Due December 11th, 2020

Abstract

With technology advancing constantly more and more statistics are able to be found, but which are the most important? In this project I compare the differences in fantasy baseball teams decided from two different kinds of regressions: single variable and multiple variable. The main point is to see which regression method makes a team that produces the most fantasy points. The regression involving multiple variables proved to be a more accurate approximation as the R-squared values were larger. I arranged data frames according to the variables that proved to be most important. As expected, the multiple variable regression team produced 12,767 points, almost 300 points higher than the single regression team.

1 Introduction

ESPN and other sports related sites provide a fantasy aspect to real life sports. Fantasy baseball is one of the more common ones along with football and basketball. In these fantasy sports, you can join a league usually containing a minimum of 8 teams. Before the actual season starts, everyone in the league drafts a team picking players from any of the teams in the majors of that specific sport. In fantasy baseball, roster includes 25 players: a catcher, first baseman, second baseman, third baseman, shortstop, middle infielder (shortstop/second baseman), corner infielder (first/third baseman), 5 outfielders, 9 pitchers, and three bench players. There is a point system that is based on how the player performs in real life. Fantasy baseball games

last a week and the team with the most points at the end of the week wins. In this project I have come up with a method to construct the best fantasy baseball team from the 2019 season (I chose 2019 because last season was not a full 162-game season so the stats are not as good). This will be done by comparing single variable regression versus a multiple variable regression.

2 Data

I have 6 main data sets that I am working with. 3 mainly include average based stats, for example in set dealing with hitters it has batting average, slugging percentage, on-base percentage, etc. The other 3 include, what I refer to as, single-stat based statistics. Using the hitter example again, the stats include singles, doubles, triples, home runs, etc. There are 2 data sets for hitters, 2 for starting pitchers, and 2 for relief pitchers. There are only a select few of stats that produce points. So, I went into every data set and removed the columns that I knew I wasn't going to use. This saved space and made my data much more organized. I also rearranged all the data sets to be in alphabetical in order by name. This allowed the corresponding data sets to be lined up with one another as they were in very different orders. This made it easier to combine stats from, for example, from the Starter_Single_Stats data frame to the Starter_Averages data frame.

I noticed that the original data sets I chose did not provide positions for each hitter. To combat this, I found another one that included positions and player names (along with a ridiculous number of other unnecessary columns which I excluded). Using the left_join function I was able to associate the majority of the hitters with their positions. Lastly, before the determination of the best players, I found the fantasy points produced by each player. The hitters were straight forward, but the pitchers required a little more thought. In baseball, innings pitched are provided such that the number after the decimal point indicates how many outs there was when the pitcher was removed from the game. So, for example if pitcher came out of the game with 2 outs in the 6th inning, he would have pitched 5.2 innings. Pitchers get 3 points per inning pitched or one point per out. Because of that, I can't just multiply that 5.2 by 3 because that wouldn't be the actual points the pitcher got. So, I had to get the number after the decimal point, if there was one, and add to the product of 3 and the number before the decimal.

3 Model

The whole project is based off of linear regressions. I used two methods of regressions: single and multiple variable. I first start off with the single variable part. I break it into three sections: one for hitters, one for starting pitchers, and one for relievers. I then picked variables in the average data sets that I think could be good indicators of producing the most fantasy points. For hitters I looked at AVG (batting average), OBP (on-base percentage), SLG (slugging percentage), RBIs (Runs Batted In), Rs (Runs Scored), Hs (Hits), and HRs (Home Runs). For starting pitchers I chose Ws (Wins), IPs (innings pitched), SOs (Strike Outs), and ERA (Earned Run Average). Lastly for the relievers I chose the same as the starters but also added SVs (Saves).

After that, I made three more regressions. This time, however, I included all of the variables for the respective positions into one big regression. After those were calculated I rearranged the data frames according to which variable proved to be the best indicator. I made one team using the best variables from the single variable regression section and another from the multiple variable one. I then compared the two teams to see which one was produced more points. I compared the R-squared values for each variable in each position (hitters, starters, and relievers) then rearranged the data frames according to the highest one. In case of a tie I added another variable that corresponded to the best slope. Then for the multiple variable regressions I chose the top 3 slopes between all the variables in each position and rearranged the data according to those.

4 Results

4.1 Single Variable Regression

After making all of the single variable regressions I looked at the summary of each one and chose the variable with the highest R-Squared term. This term shows how close all the points on the graph are to the generated line. It also determines the importance of the variable in predicting whatever is trying to be predicted, which in this case is fantasy points. So I arranged the average stat data frames in terms of the best R-squared value the the best slope. For the hitters I concluded that Rs had the highest R-Squared

at 0.9728 and RBIs had the best slope at 4.5. By best slope I mean it was the biggest slope that made the most sense. The slope produced by HR was around 12. The slope refers to how many more fantasy points the hitter will produce per one additional unit of the variable that is being evaluated. For a home run a hitter will get a minimum of 6 points and a max of 9. So this slope seems to be overestimating too much. The RBI slope makes more sense because if a hitter gets an RBI, it is likely that he got a hit which is also a point. So a single and RBI is 2 points, which is already roughly half way to the 4.5. In this day and age, most home run hitters will either hit a ball 500ft or strike out (the furthest fences are 420ft). And a strikeout is -1 points so even if a hitter does get a grand slam, a strikeout puts him at 8 points further away from the 12 slope.

The starting pitchers showed that strikeouts were the most important variable as the R-squared was 0.9109 while Ws had the best slope at 27. That slope is overestimated because the odds of a win resulting in 27 points is low. Starting pitchers in today's baseball don't last as long as they used to. I'd say the average innings pitched per game is around 6. So a win with 6 innings is 23 points. However, these won't be the only stats produced by the pitcher as there will probably be some strikeouts, hits, walks and earned runs factored in. The 27 points is definitely possible but the pitcher has to be really good on that day. Pitchers are bound to make mistakes so they definitely won't getting that 27 every game. I chose this over IPs because the slope for IPs was close to 2. If 1 IP corresponds to 3 fantasy points it is not a good indicator.

Similar to the starters, the best indicator of points for relievers was strikeouts having an R-squared value of 0.8014. The best slope turned out to be, as expected, SVs at 9.5. Saves are exclusive to relievers. They are only achieved if the pitcher's team has the lead in the last inning by 3 or less runs and he keeps the lead winning the game for his team. This slope is the most reasonable one because the fantasy points correlate with the slope much better. If a pitcher pitches the entire last inning, which is usually the case, he gets 5 points for the save and 3 for the inning pitched resulting in 8 total points. Only one away from the 9 all the reliever needs to do is record a strikeout, which is highly likely, and that 9 points is achieved.

4.2 Multiple Variable Regression

The multiple variable regressions laid out all the slopes for each variable involved and displayed one over arching R-squared value. The value for hitters was 0.9808, starters was 0.9473, and relievers was 0.9103. The slopes for hitters, in order of importance and how the data frames were arranged, was Rs, AVG, and RBIs. For the starters the slopes proved to be Ws, SOs, and IPs. Lastly for the relievers, the slopes turned out to be SVs, Ws, and SOs.

4.3 Comparison

After both sets of regressions were completed, I calculated the total points scored for each team. The team resulting from the single variable regression produced 12,476 points whereas the multiple variable regression team produced 12,767 points. This makes sense that the multiple variable one will have more points. The R-squared values produced by the multiple regressions was higher than any one of the other variables from the respective positions. For example, the single variable regressions for starting pitchers saw a max R-squared value 0.9109 while the multiple variable regression for starters was 0.9473. When working with regressions, the more variables included will usually result in a better approximation of the thing you are trying to predict. This is also reflected when comparing the corresponding slopes.

The slopes produced by the multiple variable regressions catered more towards the fantasy point system than their counterparts. We see this clearly looking at the two batting average slopes. The single regression of batting average produced a slope around 489 while the slope calculated by the multiple variable regression was near 70. The 489 slope says that a small change in batting average will incur a larger change in fantasy points. Batting averages in baseball only change in small increments, if at all. If a hitter increases their batting average a little that just means he gets a hit slightly more often than he did before. The hitter will produce more fantasy points on average but probably not the extent that the 489 slope states. The 70 slope is a better representation of the potential increase in fantasy points.

Another example can be seen with the Wins variable for starting pitchers. Going back to explanation of this variable in the single variable regression section, the 27 fantasy points is not guaranteed every start. Every start has to be near perfect to achieve such points. The slope for Ws in the

multiple regression calculator is approximately 15. This number is much more reasonable to accomplish. Pitchers, as I said before, do not last as long as they used to. Now 5 or 6 innings is normal. Hypothetically speaking, if a pitcher gets a win with 5 innings pitched that's 20 points. But those are not the only stats attributed to the pitcher. If anything, this is somewhat of underestimation, but not the same magnitude as the overestimation made by the single variable regression.

To further show why the multiple variable regression is better, I have calculated an average of, for both teams, strikeout and walk percentages for hitters and the SO/9 (strikeouts per 9 innings) and BB/9 (walks per 9 innings) for the pitchers. The strikeout and walk percentages for the hitters show how often a hitter records a strikeout or walk. The single variable regression team produced an average strikeout percentage of 0.18 and an average walk percentage of 0.12. The multiple variable regression team produced the same numbers. This means that the two regressions resulted in the hitters. I would have expected the strikeout percentage to be lower and walk percentage to be higher on the multiple regression team. The similarity is most likely due to the how both data frames were arranged after the regressions as the first variable for both was Rs.

Since the hitters are the same for both teams, the differences in points is seen more so in the pitchers. Combining the relievers and starters, the K/9 value for the single variable regression team was 13.22 and their BB/9 was 2.56. The multiple variable regression team produced a 12.13 K/9 value and a BB/9 value of 2.3. It is interesting that the the single variable regression team averaged one more strikeout than their counterparts, but, as expected, averaged slightly more walks than the multiple variable regression team.

5 Ablation

The more interesting side of this is that in the single variable regression, OBP had worse slope than AVG. Additionally, OBP's slope went negative once the multiple variable regression was calculated. This does not align with the strategy imposed by Moneyball. The movie Moneyball shows how former general manager of the Oakland A's, Billy Beane, devised a brand new strategy on how to build a team. He preached that to win games, you have to score runs and to scores runs you must get people on base. Well, OBP explicitly shows how often a batter gets on base. So, Beane went out and

got players that were cheap and had a relatively high OBP. He was offered a job in the Red Sox organization after the 2002 season, but turned it down. However, the Red Sox used his new method to construct their 2004 World-Series winning team breaking the Curse of the Bambino. So, it would have made a lot more sense had OBP been the best indicator of fantasy points.

6 Literature Review

There is not much in the way in depth analysis from my sources. I looked up player statistics from the 2019 season and found those 6 data sets. There is a link that takes you to the FanGraphs website. I also needed to go elsewhere to get positions as those original 6 didn't include positions for the hitters. On an ESPN article about fantasy baseball I found a layout of what the roster of a fantasy baseball team consists of. Lastly, from a SportingNews article I obtained the point breakdown for the hitters and pitchers as I forgot them from when I played fantasy baseball. All the references to the actual sport is from my knowledge of it. Links to all my sources are provided at the bottom of my code.

7 Conclusion

As suspected, the multiple variable regression proved to make a better team, around 300 points better. The R-squared values for the 3 multiple variable regressions were all higher than any of the single variable regressions for the corresponding positions. This means that the multiple variable regressions were more accurate in predicting fantasy points. This is the main reason why the point totals turned out the way they did. I used some of the other variables provided in the data frames to give another reason why the point totals favored the multiple variable regressions. These values, however, could be used as an argument against mine as one variable, K/9, was better for the single regressions. BB/9 was better in the multiple variables regressions but not by the same magnitude as K/9. In the end, the multiple variable regressions attested to my expectations as they produced a higher scoring team.