Scott Waugh

Microeconomics Capstone

Dr. Neal Becker

29 April 2022

Regression Analysis of Wins in Major League Baseball

Abstract

Baseball analytics has increased in popularity ever since the early 2000s. As the payroll gap increases from the top to bottom team, it is crucial to get the best players possible with the amount of you have. Regression analysis can help teams find the most important stats that they should look at when deciding to keep and buy players for the team to help generate wins. I completed two different regression analyses of baseball data the first regression analysis from 1969-2021 excluding the 1981, 1994, 1995, and 2020 seasons and the second regression analysis from 2011-2021 excluding the 2020 season. For the regression analyses, I start with 31 variables that all effect the amount of wins a team gets. For my regression analysis of baseball since 1969, I discover that the variables most impacting wins are WHIP, SV, tSho, OBP, HR, CS, SO, BB, and FIP. Furthermore, I discover in my regression analysis of baseball since 2011, that the variables most impacting wins are WHIP, SV, tSho, CG, BBA, HRA, and RBI. In conclusion, my research shows that different statistics are more important in determining wins when just looking at baseball since 2011 than when looking at baseball since 1969.

Introduction

When you think of what America's pastime would be most people would think of the sport baseball. Baseball has been around a long time, it first gained popularity in North America during the 1850's. This led to the rise of Major League Baseball (MLB), which is the baseball league in the United States and Canada. Something that makes baseball so interesting is all the different statistics that come from the game. Things such as runs scored, batting average, and earned run average allow us to differentiate the best players from the worst. In the late 1900's, a way of analyzing baseball statistics more in depth was created and it is called sabermetrics. Since the creation of this idea, we have seen examples, such as Moneyball, of how this might be the future of how managers build sports teams. In this research paper, I want to show how the use of regression analysis leads me to find the most important variables that determine wins in the game of baseball.

Sports Context

Before I get into sabermetrics and baseball analytics, there is a lot of basic information that should be known about baseball. A baseball team consists of a 26-man roster that consists of pitchers, infielders, and outfielder. There are 9 players on the field at a time when the team is on defense, this includes the pitcher, catcher, first baseman, second baseman, shortstop, third baseman, right fielder, center fielder, and left fielder. Teams play 9 inning games and the team to score the most runs wins. In the MLB, teams play 162 game seasons, and the top teams make it to the playoffs in hopes to win the World Series. Baseball does not have a salary cap so there is a big discrepancy between the top team paying $279 million and the bottom team only paying $50 million. This makes it crucial for teams with lower payrolls to find a way to determine how to efficiently spend their money. A way they can do this is by finding out what stats matter the most by using statistical analysis.

Statistics for baseball games have been around since the creation of the box score in 1858. Sabermetrics research began in the mid 1900's with Earnshaw Cook. In 1964 his book *Percentage Baseball* was one of the first uses of sabermetrics (Wikipedia Contributors). His book was originally dismissed as meaningless by many baseball teams and professionals. It was not until Bill James published his book *Baseball Abstract* in 1982 that people started to believe sabermetrics was something worth looking into, though it was slow to find widespread acceptance of the idea. Sabermetrics was originally defined as the search for objective knowledge about baseball by Bill James ("A Guide to Sabermetric Research | Society for American Baseball Research"). Sabermetric researchers often use statistical analysis to question traditional measure of baseball such as batting average and pitcher wins. Sabermetrics allows statisticians or other data analysts to analyze baseball on a deeper level and find out the impact of each stat and see which statistics are most important.

Economic Context

By being able to tell what stats are most important through sabermetrics we can figure out what players we want in the future more easily since we know what to focus on. With sports being done publicly, it allows data to be easily collected and utilized in things like research. Player and organizations will change what they are doing based on analysis of this data. Moneyball is the most widely known example of this being done (*Moneyball*). The manager of the Oakland Athletics in the MLB, Billie Beane, used sabermetrics to identify unvalued players to put together a team on a low payroll. They were able to win a historic 20 games in a row that season and finish the season 103-59. After seeing the success other teams started to follow this strategy and this made statisticians and economists a hot commodity in sports organizations.

Data analysis in sports can be seen to maximize the money you are putting into your team. By using regression analysis, you can figure out the most important variables that lead to a certain outcome such as wins. With this knowledge that we gain from the regression analysis, teams can focus on signing player that are better than other players in these stats and get rid of players that are bad in these stats. They can incorporate money into these equations by seeing if they are paying their players the amount, that they should be paid based on how well they are in these specific areas that are most important to look at. If these stats are as important as our regression analysis says they are then we will be able to build a team that it economically efficient since we will not be wasting our money on players that seem good but are not as good where it truly matters.

Data

The dataset I made used the data from the website Baseball Reference, which has all the baseball data you would need (“MLB Stats, Scores, History, & Records | Baseball-Reference.com”). I made my dataset from looking at team batting stats, team pitching stats, and some general team stats from the from the years 1969-2021. I choose this group of years since it had all the variable’s I need from that time forward. I also only wanted to use seasons where they played 162 games, so I excluded the 1981, 1994, 1995, and 2020 seasons. I then compressed the data down into an excel spreadsheet, removing the extra variables I was not planning on using and getting rid of the years I did not want. I then imported this spreadsheet into R Studio where I planned on doing my regression analysis. The data on this website was very good in that there were no missing numbers, and everything was complete. It also had variables calculated for me, where other websites or R packages did not have very in-depth statistics. The variable I used were BatAge, RG, R, H, SB, TB, HR, RBI, SB, CS, BB, SO, BA, OBP, SLG, OPS, PAge, RAG, ERA, CG, tSho, cSho, SV, HA, RA, ER, HRA, BBA, SOA, FIP, WHIP as my independent variables and W as my dependent variable. Here is a description of what each variable means:

- BatAge – Average age of batters on a team weighted by at bats and games played.
- RG – Average of runs scored per game.
- R – The total amount of runs scored during the season.
- H – The total amount of hits during the season.
- SB – The total amount of doubles during the season.
- TB – The total amount of triples during the season.
- HR – The total number of home runs during the season.
- RBI – The total number of runs batted in during the season, this is when a ball is hit and a person on base scores from the hit.
- SB – The total amount of stolen bases during the season.
- CS – The total amount of times the runner is tagged out stealing bases during the season.
- BB – The total amount of walk during the season.
- SO – The total amount of strike out during a season.
- BA – Batting average during the season, calculated by taking hits divided by at bats.
- OBP – On base percentage calculated by adding hits, walks, and times hit by pitch and dividing it by the sum of at bats, walks, times hit by pitch, and sacrifice flies.
- SLG – Slugging Percentage, which is total number of bases divided by at bats
- OPS – On base percentage plus slugging percentage.
- PAge - Average age of pitchers on a team weighted by 3 times games started plus games plus games saved.
- RAG – Average of runs allowed per game.

- ERA – Earned run average, calculated by earned runs divided by innings pitched then multiplied by 9.
- CG – Number of games when a pitcher starts and finishes the game.
- tSho – Number of games when there is a shutout which is when the team doesn't allow any runs.
- cSho – Number of games when there is a complete game and a shutout.
- SV – Number of games when there is a save.
- HA – Hits allowed during the season
- RA – Runs allowed during the season
- ER – Earned runs allowed during the season, which is runs that are scored not on an error.
- HRA – Home runs allowed during the season.
- BBA – Walks allowed during the season.
- SOA – Strike outs by the pitcher during the season.
- FIP – Fielding Independent Pitcher, this stat measures a pitcher's effectiveness at preventing HR, BB, HBP, and SO. Calculated by (13*HR + 3*(BB+HBP) – 2*SO) / IP+Constant, where the constant is set so the league average FIP equals the league average ERA.
- WHIP – Walks and hits per innings pitched.
- W – Total number of wins during the season.

I selected these variables I felt they were very important variables to look at when determining wins. I also did not want to add too many variables, or it would have made the regression analysis harder to understand. Limiting the variables and selecting the variables based on what I thought was important could potentially lead to different findings if other people would use more, less, or different variables than I did. Also, you could get different results based on the years that are used, since I picked a certain range that seemed fit for getting good results.

Analysis – Complete Data

For my analysis of the data, I am going to be using multiple regression analysis techniques to find the best model possible to determine wins. At first, I am going to run the full model with all the original variables and see what we get at the beginning. The part we care about for this is the values under the Pr(>|t|) column. This is going to tell us whether a variable is significant or not, if the p value is less than 0.05 than it is significant. Looking at the output below, we see that many of the variables are not seen as significant in the model. Out of the 31 variables we see that only RG, R, H, CS, BB, SO, BA, OPS, CG, tSho, SV, HA, RA, HRA, BBA, and WHIP are seen as significant in the full model. A big reason for this is the fact that many of these variables explain the same things so that makes other similar variables less significant if they have already been explained by another variable before. Once we get rid of variables it will cause variables to have lower p values since nothing else is explaining the same thing. We also see that the model has an r-squared value of 0.929 meaning that 92.9% of wins can be explained by the above predictors. Also, another important thing to look at is the residual standard error of 3.107. This number is saying that if we use this model to predict wins, we will most of the time be in the range of 3.1 less or more wins than the actual number. This is good considering there are 162 games in a season. You might think that we should just keep this model but it is hard to find players that are good in all these stats so we want to simplify the model to figure out what stats are most important even if that would lower our r-squared or residual standard error.

```
Call:
lm(formula = W ~ BatAge + RG + R + H + SB + TB + HR + RBI + SB + 
    CS + BB + SO + BA + OBP + SLG + OPS + PAge + RAG + ERA + 
    CG + tSho + cSho + SV + HA + RA + ER + HRA + BBA + SOA + 
    FIP + WHIP, data = baseball)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.7991 -2.1342 -0.0171  2.0923 10.9872 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.052e+00  6.279e+00   0.645  0.51878    
BatAge       1.337e-01  8.149e-02   1.641  0.10101    
RG          -2.925e+01  1.441e+01  -2.030  0.04260 *  
R            2.612e-01  8.988e-02   2.905  0.00373 ** 
H           -1.005e-01  1.543e-02  -6.509 1.07e-10 ***
SB          -3.276e-02  2.437e-02  -1.344  0.17911    
TB          -6.653e-02  5.033e-02  -1.322  0.18640    
HR          -8.210e-02  7.255e-02  -1.132  0.25800    
RBI          3.800e-03  1.285e-02   0.296  0.76743    
CS          -4.009e-02  7.225e-03  -5.548 3.48e-08 ***
BB          -1.416e-02  6.424e-03  -2.204  0.02771 *  
SO           2.640e-03  1.027e-03   2.570  0.01028 *  
BA           5.011e+02  2.091e+02   2.396  0.01669 *  
OBP         -1.977e+02  1.425e+02  -1.387  0.16565    
SLG         -1.267e+02  1.694e+02  -0.748  0.45478    
OPS          2.852e+02  1.335e+02   2.136  0.03287 *  
PAge         4.387e-02  7.077e-02   0.620  0.53543    
RAG          2.301e+01  1.459e+01   1.577  0.11495    
ERA         -5.224e+00  1.131e+01  -0.462  0.64434    
CG           1.342e-01  1.438e-02   9.336  < 2e-16 ***
tSho         1.639e-01  3.646e-02   4.494 7.59e-06 ***
cSho         2.465e-02  5.250e-02   0.470  0.63872    
SV           3.274e-01  1.494e-02  21.922  < 2e-16 ***
HA           9.418e-02  2.310e-02   4.078 4.82e-05 ***
RA          -2.012e-01  9.026e-02  -2.229  0.02598 *  
ER           3.997e-02  7.093e-02   0.564  0.57313    
HRA         -2.117e-03  1.090e-02  -0.194  0.84601    
BBA          9.643e-02  2.329e-02   4.140 3.69e-05 ***
SOA          5.395e-04  1.462e-03   0.369  0.71213    
FIP         -8.661e-01  1.015e+00  -0.853  0.39359    
WHIP        -1.463e+02  3.330e+01  -4.395 1.20e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.107 on 1325 degrees of freedom
Multiple R-squared:  0.9291,	Adjusted R-squared:  0.9275 
F-statistic: 578.6 on 30 and 1325 DF,  p-value: < 2.2e-16
```

The main way I choose to reduce my model was to use stepwise selection. Stepwise selection is the combination of forward selection and backward elimination, where either the best variable is added, or the worst predictor is removed at each step. So basically, this technique continuously adds and removes variables until it finds the best variable. It does this by looking at the AIC variable and the model keeps going until it reaches the lowest AIC value possible. In the output below, I decided to just show the last step where it realizes that there is no lower AIC than the model that produces an AIC of 3092.58. So, stepwise selection lowers our model to the variables WHIP, R, RAG, SV, CG, tSho, OBP, BatAge, HA, BBA, HR, H, BA, CS, SO, BB, and FIP. This model it selected is down 14 of our original variables, but this is not our final model since we can simplify it some more.

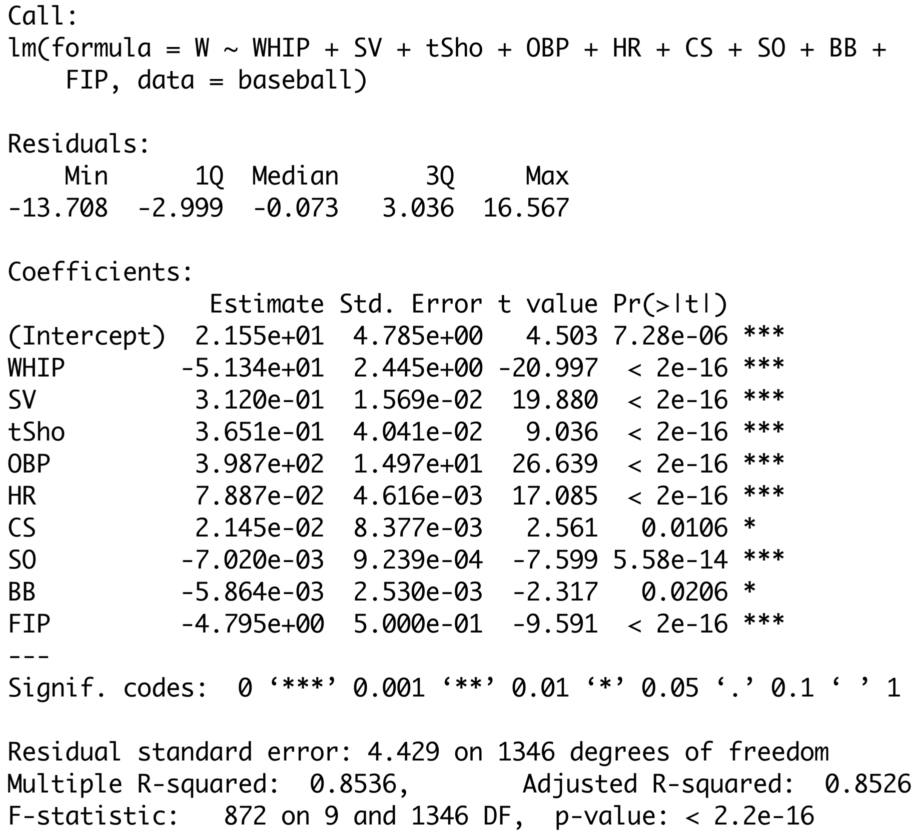
```
Step:  AIC=3092.58
W ~ WHIP + R + RAG + SV + CG + tSho + OBP + BatAge + HA + BBA +
    HR + H + BA + CS + SO + BB + FIP

         Df Sum of Sq   RSS    AIC
<none>                12918 3092.6
- HR      1      20.4 12939 3092.7
- OBP     1      23.3 12942 3093.0
+ SLG     1      10.5 12908 3093.5
+ ER      1       6.4 12912 3093.9
+ TB      1       6.3 12912 3093.9
+ OPS     1       5.9 12912 3094.0
+ SB      1       5.9 12912 3094.0
+ ERA     1       5.8 12913 3094.0
+ RA      1       4.8 12914 3094.1
+ PAge    1       3.8 12914 3094.2
+ SOA     1       3.7 12915 3094.2
+ HRA     1       1.2 12917 3094.4
+ cSho    1       1.2 12917 3094.4
+ RBI     1       0.3 12918 3094.5
+ RG      1       0.1 12918 3094.6
- BB      1      46.1 12964 3095.4
- FIP     1      54.9 12973 3096.3
- BatAge  1      67.6 12986 3097.7
- SO      1      95.8 13014 3100.6
- tSho    1     310.9 13229 3122.8
- CS      1     335.6 13254 3125.4
- BA      1     611.1 13530 3153.3
- H       1    1218.4 14137 3212.8
- CG      1    1554.6 14473 3244.7
- HA      1    1651.5 14570 3253.7
- BBA     1    1692.3 14611 3257.5
- WHIP    1    1825.7 14744 3269.8
- RAG     1    2228.6 15147 3306.4
- SV       1    4971.8 17890 3532.1
- R        1    5480.5 18399 3570.1
```

Another important thing that will help us simplify our model is looking at each variables variance inflation factor (VIF). A VIF of above 10 means that there is high correlation between other variables and that we need to get rid of this correlation. A VIF between 4-10 means there is correlation, but we do not have to fix it, though we most likely should if we can. An a VIF below 4 is where we want to be. As shown in the output below. WHIP, R, RAG, CG, OBP, HA, BBA, HR, H, BA, SO, BB, and FIP all have VIFs that should or must be fixed. The way that you can determine what should be fixed is by looking at a correlation matrix (Figure 1 in appendix). I looked for correlations between the independent variables that were positive or negative 0.700 and above since they indicate a strong correlation. I then decide to eliminate one of the variables with high correlation by looking then at the correlation between the competing variables and their correlation with wins. The variable that has the higher correlation with wins is the one I keep since it explains more about wins. When looking at these variables R, RAG, CG, HA, BBA, H, and BA had a lot of correlation with other variables and did not explain as much about wins so those are the ones I removed. After that I re-ran the model to find the VIF scores and now none of the variables have a VIF score higher than 4.

```
WHIP        R           RAG         SV          CG          tSho
157.956450  13.144543   13.928407   2.342199    4.182153    2.081504
OBP         BatAge      HA          BBA         HR          H
85.882095   1.370887    73.997832   38.774936   5.031586    112.132710
BA          CS          SO          BB          FIP
186.191068  1.983313    4.289048    25.542655   5.579742
```

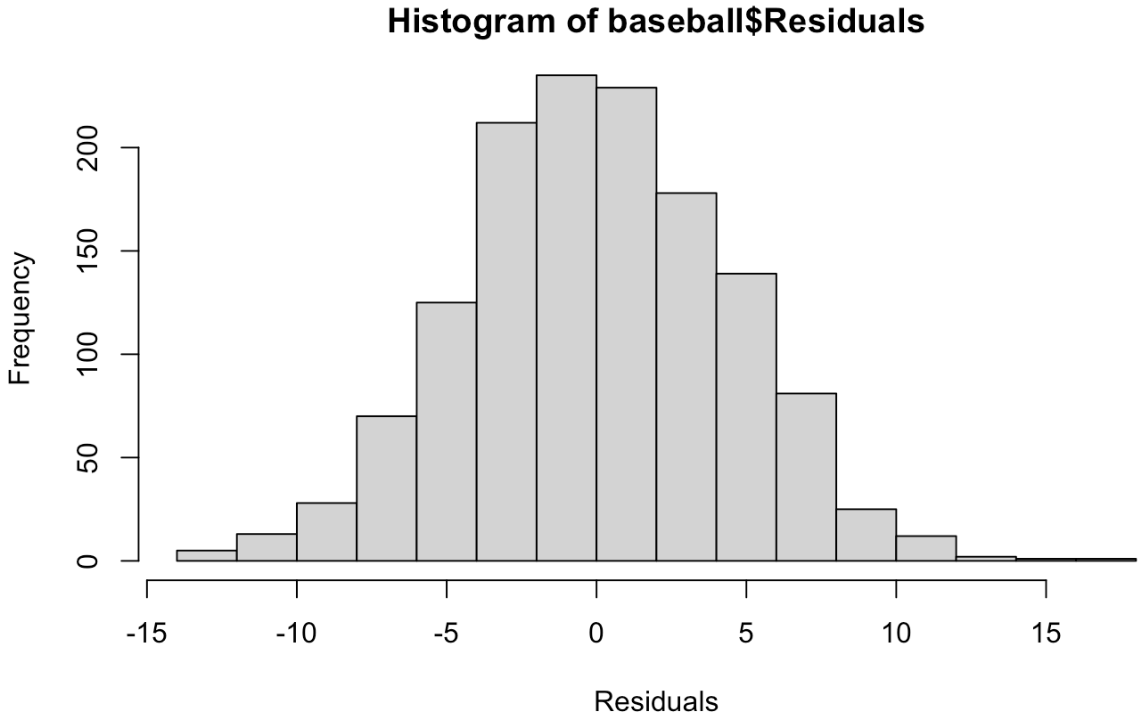
To finalize the model, I ran the model to see if all the variables were significant and I found that BatAge is not significant to the model. So, I took that variable out and our final model includes WHIP, SV, tSho, OBP, HR, CS, SO, BB, and FIP. This means that these variables are the most important variable to determine the amount wins a team gets when looking at data from 1969-

2021. Looking at the output below, we see all under the Estimate column all the coefficients for each variable and under the std. error column we see how each coefficient varies by. In this model all our variables are quite significant besides BB and CS are just barely significant. The variables of this model can explain 85.4% of the wins when looking at our new r-squared value. This is around 7% lower than what our original model can explain, but we are at 9 variables instead of 31. So, this model will make it easier for teams to determine what players are best in getting wins since they have less stats to focus on. The only error I see in this model is that getting walked more seems like a bad thing since it's a negative value but in real life getting walked more is good since you are getting runners on base.



```
Call:
lm(formula = W ~ WHIP + SV + tSho + OBP + HR + CS + SO + BB +
    FIP, data = baseball)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.708  -2.999  -0.073   3.036  16.567 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.155e+01  4.785e+00   4.503 7.28e-06 ***
WHIP        -5.134e+01  2.445e+00 -20.997  < 2e-16 ***
SV           3.120e-01  1.569e-02  19.880  < 2e-16 ***
tSho         3.651e-01  4.041e-02   9.036  < 2e-16 ***
OBP          3.987e+02  1.497e+01  26.639  < 2e-16 ***
HR           7.887e-02  4.616e-03  17.085  < 2e-16 ***
CS           2.145e-02  8.377e-03   2.561   0.0106 *  
SO          -7.020e-03  9.239e-04  -7.599 5.58e-14 ***
BB          -5.864e-03  2.530e-03  -2.317   0.0206 *  
FIP         -4.795e+00  5.000e-01  -9.591  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.429 on 1346 degrees of freedom
Multiple R-squared:  0.8536,	Adjusted R-squared:  0.8526 
F-statistic:   872 on 9 and 1346 DF,  p-value: < 2.2e-16
```

Looking at the residual standard error from the previous output we got a 4.429 which is 1.3 higher than the full model but is still good with how many fewer variables we have. Looking at the graph below we do see that majority of the residual are between -5 and 5. This is very good in my opinion since we have only 9 variable predicting wins over a period of 52 years. During that time averages of these stats have probably changed a lot since the way we play baseball evolves when we figure out better ways to train and better strategies. So being able to predict wins with a -5 to 5 accuracy is good. Some examples of predictions the model made compared to the actual win amount are the 1973 Orioles won 97 games and the model predicted 94. Another example of the model getting close is the 1996 Pirates winning 73 games and the model predicting 74 wins. An example of the model predicting bad is the 2002 Red Sox winning 93 and the model overpredicted them to win 105 games. As we can see in these last 3 examples the model predicts quite well for the most part but there are obviously going to be some errors because it can only be so perfect. Looking at figure 2 in the appendix, we see that the model has a mostly straight line meaning that normality is not violated and that a regression model works well. Overall, we see that our final model of WHIP, SV, tSho, OBP, HR, CS, SO, BB, and FIP predicts wins quite well and that these stats are some of the most important stats too look at.



Analysis – Last 10 Years

For my second analysis of the data, I am going to be using the same regression strategy but looking at just the data from 2011-2021 to see if we get different results. First, I am going to run the full model with all the original variables and see what we get at the beginning. Looking at the output below, we see again that many of the variables are not seen as significant in the model. Out of the 31 variables we see that only H, RBI, CS, ERA, CG, tSho, SV, HA, ER, BBA, and WHIP are seen as significant in the full model. We already see a difference in the data since the full model has different variables that are significant at the beginning. We also see that the model has an r-squared value of 0.946 meaning that 94.6% of wins can be explained by the above predictors. We can see that with smaller number of years we are looking at the full model explains more wins than previously. Also, our residual standard error is 2.992 this time around. This number is saying that if we use this model to predict wins, we will most of the time be in the range of about 3 less or more wins than the actual number. The residual standard error is only slightly lower in this data than the other full data.

```
Call:
lm(formula = W ~ BatAge + RG + R + H + SB + TB + HR + RBI + SB +
    CS + BB + SO + BA + OBP + SLG + OPS + PAge + RAG + ERA +
    CG + tSho + cSho + SV + HA + RA + ER + HRA + BBA + SOA +
    FIP + WHIP, data = baseball10years)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8226 -1.8424  0.1842  1.8985  8.8683

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.105e+00  1.872e+01   0.326 0.744584
BatAge       1.043e-01  1.956e-01   0.533 0.594203
RG           2.745e+01  4.519e+01   0.607 0.544082
R           -2.040e-02  2.812e-01  -0.073 0.942218
H           -1.295e-01  3.905e-02  -3.317 0.001035 **
SB          -3.828e-02  5.549e-02  -0.690 0.490880
TB          -9.270e-02  1.143e-01  -0.811 0.418172
HR          -1.240e-01  1.655e-01  -0.749 0.454326
RBI         -5.394e-02  3.076e-02  -1.753 0.080680 .
CS          -9.528e-02  2.463e-02  -3.868 0.000138 ***
BB          -2.289e-02  1.362e-02  -1.681 0.093938 .
SO           1.717e-03  2.368e-03   0.725 0.468971
BA           6.063e+02  4.888e+02   1.240 0.215955
OBP         -6.145e+02  3.792e+02  -1.620 0.106360
SLG         -4.084e+02  4.289e+02  -0.952 0.341774
OPS          6.409e+02  3.695e+02   1.735 0.083931 .
PAge         9.886e-02  1.945e-01   0.508 0.611609
RAG         -2.486e+01  4.538e+01  -0.548 0.584189
ERA          9.682e+01  3.822e+01   2.533 0.011863 *
CG           4.150e-01  1.198e-01   3.465 0.000616 ***
tSho         1.840e-01  6.611e-02   2.783 0.005767 **
cSho        -1.885e-01  2.227e-01  -0.846 0.398248
SV           3.887e-01  3.568e-02  10.894  < 2e-16 ***
HA           3.470e-01  8.832e-02   3.929 0.000109 ***
RA           1.104e-01  2.807e-01   0.393 0.694438
ER          -5.951e-01  2.393e-01  -2.487 0.013502 *
HRA         -2.573e-02  4.381e-02  -0.587 0.557470
BBA          3.576e-01  8.876e-02   4.029 7.29e-05 ***
SOA          6.549e-04  5.849e-03   0.112 0.910927
FIP         -2.668e-01  4.499e+00  -0.059 0.952754
WHIP        -5.192e+02  1.280e+02  -4.056 6.54e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.992 on 269 degrees of freedom
Multiple R-squared:  0.9464,       Adjusted R-squared:  0.9405
```

```
F-statistic: 158.4 on 30 and 269 DF,  p-value: < 2.2e-16
```

We are again going to use the stepwise model to start simplifying are model. In the output below, I decided to just show the last step again where it realizes that there is no lower AIC than the model that produces an AIC of 669.9. The AIC is a lot lower this time around because of how many less observations we have. Stepwise selection lowers our model to the variables WHIP, R, RAG, SV, tSho, CG, BBA, HA, CS, BB, H, BA, HRA, and RBI. This model it selected is down 17 of our original variables, which is different than the last model and we see variables such as HRA, and RBI are in this time and OBP is out.

```
Step:  AIC=669.9
W ~ WHIP + R + RAG + SV + tSho + CG + BBA + HA + CS + BB + H +
    BA + HRA + RBI

         Df Sum of Sq    RSS    AIC
<none>                2532.1 669.90
+ OBP     1      5.19 2526.9 671.29
+ cSho    1      4.83 2527.2 671.33
+ SO      1      4.59 2527.5 671.36
+ OPS     1      4.20 2527.9 671.40
- RBI     1     30.19 2562.3 671.46
+ BatAge  1      3.23 2528.8 671.52
+ SB      1      3.05 2529.0 671.54
+ ERA     1      2.77 2529.3 671.57
+ SOA     1      2.52 2529.5 671.60
+ PAge    1      2.34 2529.7 671.62
+ TB      1      1.96 2530.1 671.67
+ SLG     1      1.84 2530.2 671.68
+ RG      1      1.76 2530.3 671.69
+ ER      1      1.19 2530.9 671.76
+ HR      1      0.94 2531.1 671.79
+ FIP     1      0.92 2531.1 671.79
+ RA      1      0.55 2531.5 671.84
- HRA     1     41.82 2573.9 672.82
- tSho    1     63.93 2596.0 675.38
- BB      1    151.01 2683.1 685.28
- CG      1    155.76 2687.8 685.81
- CS      1    186.10 2718.2 689.18
- RAG     1    192.13 2724.2 689.84
- R       1    255.32 2787.4 696.72
- BA      1    286.12 2818.2 700.02
- H       1    317.13 2849.2 703.30
- HA      1    355.12 2887.2 707.28
- WHIP    1    415.83 2947.9 713.52
- BBA     1    416.36 2948.4 713.57
- SV      1   1369.55 3901.6 797.61
```

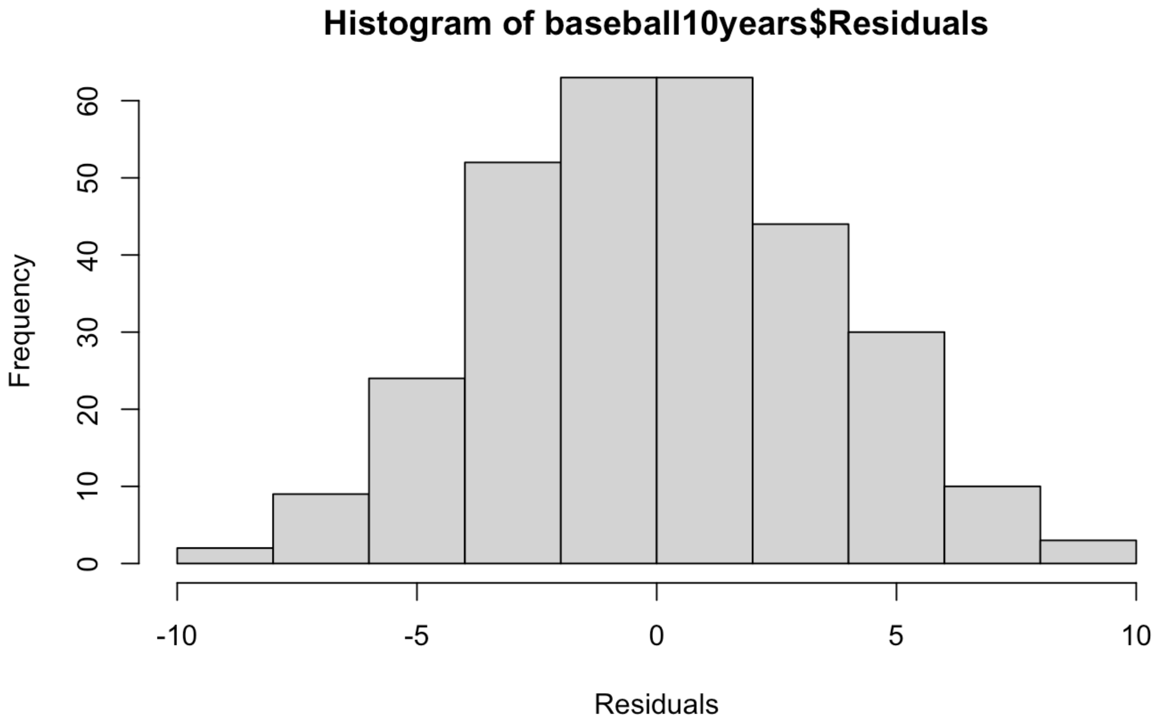
Let's look at the VIF scores again for each variable to figure out which variables we should get rid of. As shown in the output below. WHIP, R, RAG, BBA, HA, H, BA, HRA, and RBI all have VIFs that should or must be fixed. The way that you can determine what should be fixed is by looking at a correlation matrix (Figure 3 in appendix). When looking at these variables R, RAG, HA, and H had a lot of correlation with other variables and did not explain as much about wins so those are the ones I removed. I also had to remove BB in the end since it had close to 0.700 correlation with other variables and was holding me back from getting VIF scores of lower than 4. After that I re-ran the model to find the VIF scores and now none of the variables have a VIF score higher than 4.

```
WHIP         R            RAG          SV          tSho        CG
200.709177   158.086345   17.202079    2.062542    1.972563    1.586756
BBA          HA           CS           BB          H           BA
40.056851    99.271989    1.508850     2.589122    105.846748 98.086851
HRA          RBI
4.489701     156.159505
```

To finalize the model, I ran the model to see if all the variables were significant and I found that CS and BA is not significant to the model. So, I took that variable out and our final model includes WHIP, SV, tSho, CG, BBA, HRA, and RBI. This means that these variables are the most important variable to determine the amount wins a team gets when looking at data from 2011-2021. These variables are also quite different than the variables in the previous model and we also have 2 less variables this time in this model all our variables are quite significant besides tSho and CG are just barely significant. The variables of this model can explain 92.2% of the wins when looking at our new r-squared value. This is 7% better than what the previous final model got us with 2 less variable. It makes sense though since we are looking at a more condensed timeline that it would be different. So, this model will make it easier for teams to determine what players are best in getting wins when looking at the most important stats of a more recent time. The only error I see in this model is that allowing more walks seems like a good thing since it’s a positive value but in real life allowing more walks is bad since you are allowing runners on base.

```
Call:
lm(formula = W ~ WHIP + SV + tSho + CG + BBA + HRA + RBI, data =
baseball10years)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4650 -2.4005  0.0105  2.3523  9.0989

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  67.470780   6.216757  10.853  < 2e-16 ***
WHIP        -54.982403   4.131133 -13.309  < 2e-16 ***
SV            0.530086   0.034784  15.239  < 2e-16 ***
tSho          0.246410   0.068467   3.599 0.000375 ***
CG            0.342424   0.089804   3.813 0.000168 ***
BBA           0.020982   0.005315   3.948 9.90e-05 ***
HRA          -0.068951   0.008023  -8.594 5.16e-16 ***
RBI           0.090888   0.003011  30.186  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.472 on 292 degrees of freedom
Multiple R-squared:  0.9217,	Adjusted R-squared:  0.9198
F-statistic: 490.9 on 7 and 292 DF,  p-value: < 2.2e-16
```

Looking at the residual standard error from the previous output we got a 3.472 which is which is 1 lower than the model from the other regression. Looking at the graph below we do see that majority of the residual are between -5 and 5. This is still very good, and we also see in the graph that there are less farther out from -5 and 5 than in the previous model. Some examples of predictions the model made compared to the actual win amount are the 2011 Cardinals won 90 games and the model predicted 91. Another example of the model getting close is the 2017 Reds winning 68 games and the model predicting 69 wins. An example of the model predicting bad is the 2019 Reds winning 75 and the model overpredicted them to win 83 games. As we can see in these last 3 examples the model predicts quite well for the most part but there are obviously going to be some errors because it can only be so perfect. Looking at figure 4 in the appendix, we see that the model has a mostly straight line meaning that normality is not violated and that a regression model works well. Overall, we see that our final model of WHIP, SV, tSho, CG, BBA, HRA, and RBI predicts wins quite well from 2011-2021 and that these stats are some of the most important stats too look at in recent times.

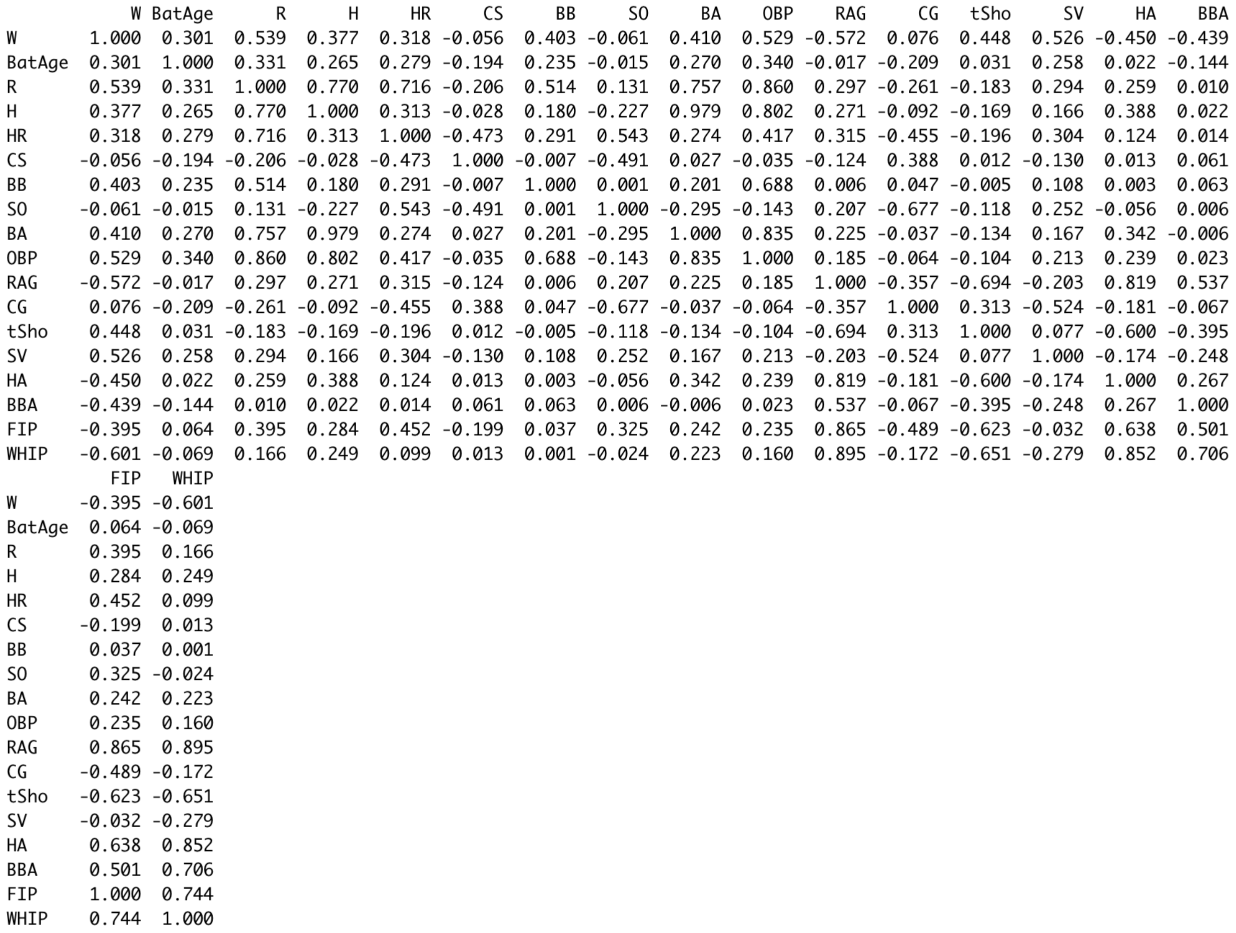


Conclusion

The regression analysis for the full data had a different result with WHIP, SV, tSho, OBP, HR, CS, SO, BB, and FIP being the variables that explain wins the best vs the model for the last 10 years with WHIP, SV, tSho, CG, BBA, HRA, and RBI as the variables. From the full data we see 4 pitching statistics and 5 batting statistics and from the condense period we see 6 pitching statistics and 1 batting statistic. This could mean that pitching matters more in the last 10 years and over the last 52 years batting and pitching were more equal in deciding wins. With all research there could be some different results based on how other people conduct their research. Some people might not use stepwise selection and that could give them different results. Also, some people might choose to remove different variables during the VIF process since there are technically different variables you can get rid of to get the remaining variables a VIF of under 4, but the way I used is the most common way. The whole goal of this research was to find a more simplified model for teams to use to build the best and most cost-efficient team. I ended up producing two models that could possibly be able to do that with the second regression model being more accurate to current baseball. In the end, both the models predict wins quite accurately with a lot less variables than the full model.

Appendix

Figure 1. This figure below is a matrix of the correlations between each variable for the first regression analysis with all the years.



```
            W BatAge      R      H     HR     CS     BB     SO     BA    OBP    RAG     CG   tSho     SV     HA    BBA
W       1.000  0.301  0.539  0.377  0.318 -0.056  0.403 -0.061  0.410  0.529 -0.572  0.076  0.448  0.526 -0.450 -0.439
BatAge  0.301  1.000  0.331  0.265  0.279 -0.194  0.235 -0.015  0.270  0.340 -0.017 -0.209  0.031  0.258  0.022 -0.144
R       0.539  0.331  1.000  0.770  0.716 -0.206  0.514  0.131  0.757  0.860  0.297 -0.261 -0.183  0.294  0.259  0.010
H       0.377  0.265  0.770  1.000  0.313 -0.028  0.180 -0.227  0.979  0.802  0.271 -0.092 -0.169  0.166  0.388  0.022
HR      0.318  0.279  0.716  0.313  1.000 -0.473  0.291  0.543  0.274  0.417  0.315 -0.455 -0.196  0.304  0.124  0.014
CS     -0.056 -0.194 -0.206 -0.028 -0.473  1.000 -0.007 -0.491  0.027 -0.035 -0.124  0.388  0.012 -0.130  0.013  0.061
BB      0.403  0.235  0.514  0.180  0.291 -0.007  1.000  0.001  0.201  0.688  0.006  0.047 -0.005  0.108  0.003  0.063
SO     -0.061 -0.015  0.131 -0.227  0.543 -0.491  0.001  1.000 -0.295 -0.143  0.207 -0.677 -0.118  0.252 -0.056  0.006
BA      0.410  0.270  0.757  0.979  0.274  0.027  0.201 -0.295  1.000  0.835  0.225 -0.037 -0.134  0.167  0.342 -0.006
OBP     0.529  0.340  0.860  0.802  0.417 -0.035  0.688 -0.143  0.835  1.000  0.185 -0.064 -0.104  0.213  0.239  0.023
RAG    -0.572 -0.017  0.297  0.271  0.315 -0.124  0.006  0.207  0.225  0.185  1.000 -0.357 -0.694 -0.203  0.819  0.537
CG      0.076 -0.209 -0.261 -0.092 -0.455  0.388  0.047 -0.677 -0.037 -0.064 -0.357  1.000  0.313 -0.524 -0.181 -0.067
tSho    0.448  0.031 -0.183 -0.169 -0.196  0.012 -0.005 -0.118 -0.134 -0.104 -0.694  0.313  1.000  0.077 -0.600 -0.395
SV      0.526  0.258  0.294  0.166  0.304 -0.130  0.108  0.252  0.167  0.213 -0.203 -0.524  0.077  1.000 -0.174 -0.248
HA     -0.450  0.022  0.259  0.388  0.124  0.013  0.003 -0.056  0.342  0.239  0.819 -0.181 -0.600 -0.174  1.000  0.267
BBA    -0.439 -0.144  0.010  0.022  0.014  0.061  0.063  0.006 -0.006  0.023  0.537 -0.067 -0.395 -0.248  0.267  1.000
FIP    -0.395  0.064  0.395  0.284  0.452 -0.199  0.037  0.325  0.242  0.235  0.865 -0.489 -0.623 -0.032  0.638  0.501
WHIP   -0.601 -0.069  0.166  0.249  0.099  0.013  0.001 -0.024  0.223  0.160  0.895 -0.172 -0.651 -0.279  0.852  0.706
          FIP   WHIP
W      -0.395 -0.601
BatAge  0.064 -0.069
R       0.395  0.166
H       0.284  0.249
HR      0.452  0.099
CS     -0.199  0.013
BB      0.037  0.001
SO      0.325 -0.024
BA      0.242  0.223
OBP     0.235  0.160
RAG     0.865  0.895
CG     -0.489 -0.172
tSho   -0.623 -0.651
SV     -0.032 -0.279
HA      0.638  0.852
BBA     0.501  0.706
FIP     1.000  0.744
WHIP    0.744  1.000
```

Figure 2. This is a normality graph for the first regression analysis with all the years.

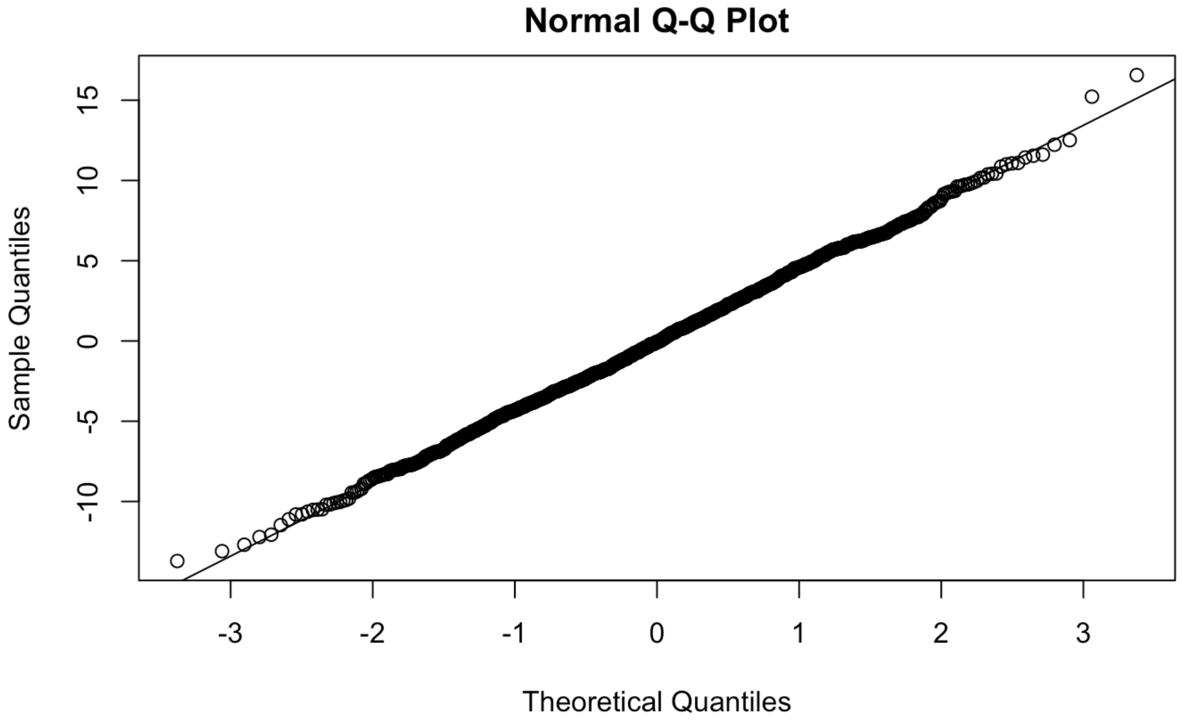
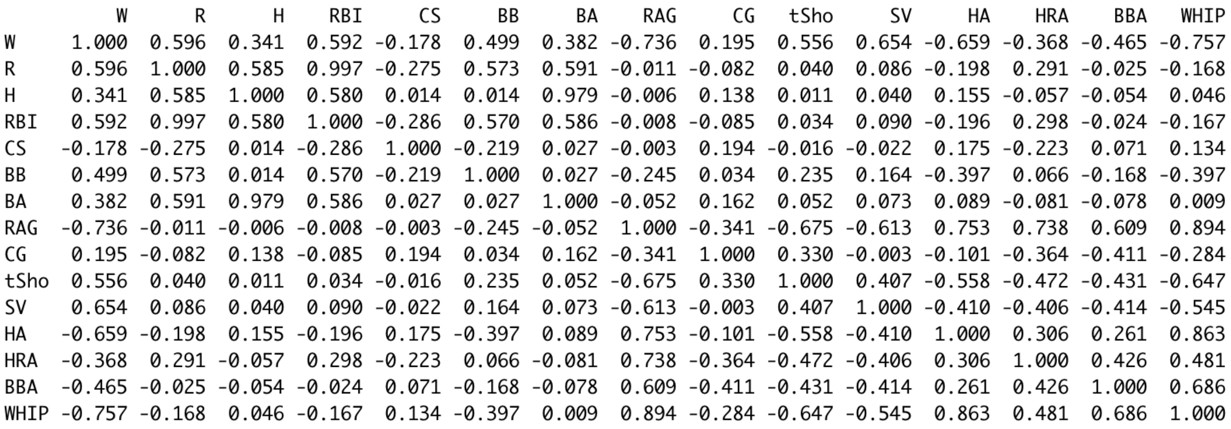
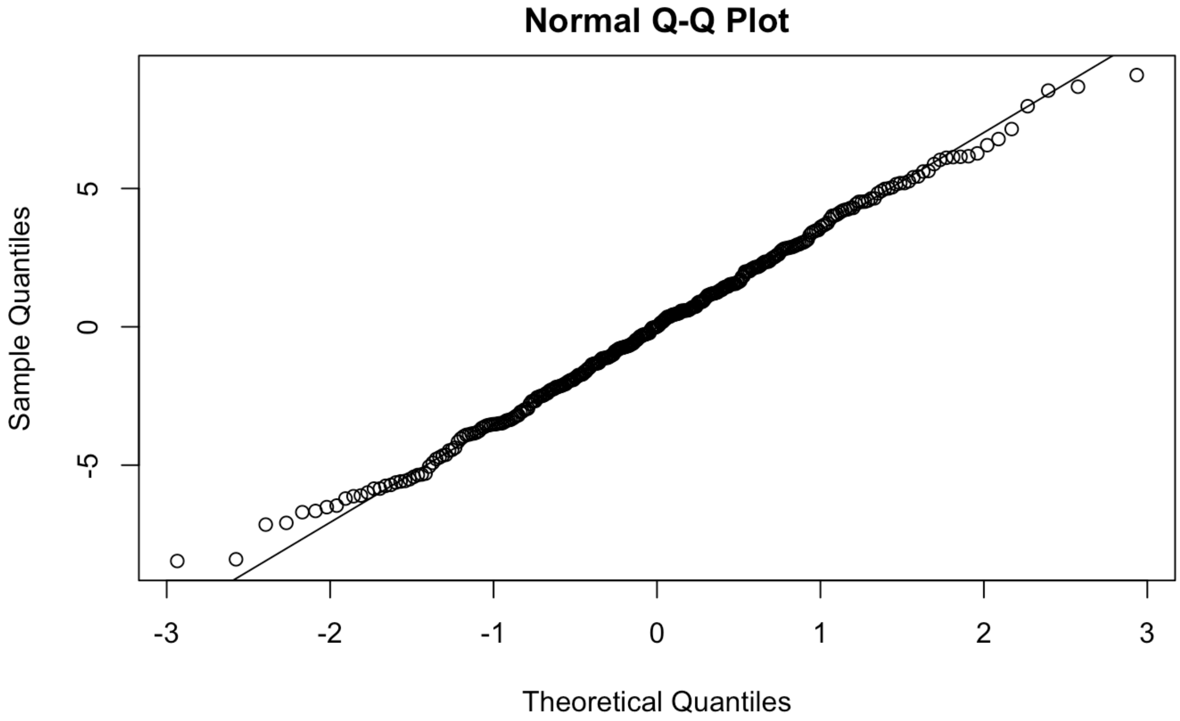


Figure 3. This figure below is a matrix of the correlations between each variable for the second regression analysis with the last 10 years.



```
         W      R      H    RBI     CS     BB     BA    RAG     CG   tSho     SV     HA    HRA    BBA   WHIP
W    1.000  0.596  0.341  0.592 -0.178  0.499  0.382 -0.736  0.195  0.556  0.654 -0.659 -0.368 -0.465 -0.757
R    0.596  1.000  0.585  0.997 -0.275  0.573  0.591 -0.011 -0.082  0.040  0.086 -0.198  0.291 -0.025 -0.168
H    0.341  0.585  1.000  0.580  0.014  0.014  0.979 -0.006  0.138  0.011  0.040  0.155 -0.057 -0.054  0.046
RBI  0.592  0.997  0.580  1.000 -0.286  0.570  0.586 -0.008 -0.085  0.034  0.090 -0.196  0.298 -0.024 -0.167
CS  -0.178 -0.275  0.014 -0.286  1.000 -0.219  0.027 -0.003  0.194 -0.016 -0.022  0.175 -0.223  0.071  0.134
BB   0.499  0.573  0.014  0.570 -0.219  1.000  0.027 -0.245  0.034  0.235  0.164 -0.397  0.066 -0.168 -0.397
BA   0.382  0.591  0.979  0.586  0.027  0.027  1.000 -0.052  0.162  0.052  0.073  0.089 -0.081 -0.078  0.009
RAG -0.736 -0.011 -0.006 -0.008 -0.003 -0.245 -0.052  1.000 -0.341 -0.675 -0.613  0.753  0.738  0.609  0.894
CG   0.195 -0.082  0.138 -0.085  0.194  0.034  0.162 -0.341  1.000  0.330 -0.003 -0.101 -0.364 -0.411 -0.284
tSho 0.556  0.040  0.011  0.034 -0.016  0.235  0.052 -0.675  0.330  1.000  0.407 -0.558 -0.472 -0.431 -0.647
SV   0.654  0.086  0.040  0.090 -0.022  0.164  0.073 -0.613 -0.003  0.407  1.000 -0.410 -0.406 -0.414 -0.545
HA  -0.659 -0.198  0.155 -0.196  0.175 -0.397  0.089  0.753 -0.101 -0.558 -0.410  1.000  0.306  0.261  0.863
HRA -0.368  0.291 -0.057  0.298 -0.223  0.066 -0.081  0.738 -0.364 -0.472 -0.406  0.306  1.000  0.426  0.481
BBA -0.465 -0.025 -0.054 -0.024  0.071 -0.168 -0.078  0.609 -0.411 -0.431 -0.414  0.261  0.426  1.000  0.686
WHIP -0.757 -0.168  0.046 -0.167  0.134 -0.397  0.009  0.894 -0.284 -0.647 -0.545  0.863  0.481  0.686  1.000
```

Figure 4. This is a normality graph for the second regression analysis with the last 10 years.



Annotated Bibliography

"A Guide to Sabermetric Research | Society for American Baseball Research." *Sabr.org*, 2011, sabr.org/sabermetrics.

This website talks about the history of sabermetric research in baseball. It also goes on to talk about how you can conduct your own sabermetric research, like what questions to ask, how to find data, what to research, and what to watch out for when completing it.

Wikipedia Contributors. "Sabermetrics." *Wikipedia*, Wikimedia Foundation, 11 July 2018, en.wikipedia.org/wiki/Sabermetrics.

This source gives some more information about sabermetrics in baseball. I only used this information for some interesting insights on history before Bill James' *Baseball Abstracts* book.

Moneyball. Directed by Bennet Miller, Columbia Pictures, 2011.

One of my favorite movies and a big inspiration to get into data analytics in sports. It brings the book Moneyball alive and shows how influential data analytics can be and how backwards traditional scouting can be. It's a good place to start when thinking of ways to analysis a sport using data you have.

"MLB Stats, Scores, History, & Records | Baseball-Reference.com." *Baseball-Reference.com*, 2000, www.baseball-reference.com/.

Baseball Reference is where I got all my data from. This website has countless different statistics that you can choose to work with, and I never found a case of missing data when looking through team statistics. I would continue to use this in the future if I am ever looking at baseball analytics again.