Finding the Best Statistic that Leads to Wins in Baseball

Richard Kang
University of Colorado Boulder
richard.kang@colorado.edu

ABSTRACT

Baseball, played professionally as Major League Baseball (MLB) in the United States, is one of the most popular sports in the country and around the world. Therefore, there is good reason in researching how teams can improve the number of wins they can get by analyzing Sabermetrics, the mathematical and statistical analysis of baseball records. Teams want to get more wins in a season not only for their fans but to make more money. More wins lead to more tickets being sold and people watching the game on television. In this paper, I will discuss the relationship between the number of team wins to certain statistics such as Home Runs, Walks, ERA (Earned Run Average), and Strikeouts. These statistics will be from both the offensive and defensive sides of the game.

In order to compare these statistics, we will use a method called Linear Regression to solve for the coefficient of the independent variable for each statistic. These coefficients will give us a sense of how much a particular statistic affects the total number of wins in a season, on average. I will also calculate the correlation coefficient, which will also show how one statistic affects team wins. The data will come from Sean Lahman's website. Sean Lahman is a journalist, and he compiled all the statistics needed for this project in different file types. I will be using the .csv file he has available on his site. I will look at how effective the linear regression and correlation analysis is by dividing the data into testing and training sets. Efficiency will also be tested by looking at the run time for the methods. Key findings include ERA affecting wins more than HR (Home Runs), RA (Opponents Runs Scored) having the strongest effect out of all variables, SB (Stolen Bases) having the lowest, and patterns existing within the same component of the game, such as batting, but not to a strong extent.

1 INTRODUCTION

1.1 Problem

One of the biggest problems in baseball, for each club in a management standpoint is whether to pursue a player with higher statistics in one area over another. For example, should a baseball team go after a player with more home runs (HR) or runs batted in (RBI)? This is a critical decision since each club has a specific amount of money to spend. There is also a luxury tax, meaning if the team's entire payroll goes over a certain amount, they get taxed. If they stay below that amount, they do not get taxed. Clubs tend to prefer to stay under this amount as they believe going over is just basically a waste of money.

1.2 Importance of Problem

From a pitching standpoint, should teams go after someone with a lower earned run average (ERA) or a player with higher outs pitched (IPouts)? These types of questions are the issues facing baseball teams today which I will try to solve in this paper. They are important since baseball is watched by millions of people in the United States and around the world. MLB does not want their fans leaving due to their favorite teams not doing well. Teams want to make sure they receive as much money as possible through sales of merchandise and tickets, as well as the viewers watching each game.

1.3 Limitations of Existing Solutions/Potential Contribution

There are existing solutions online that deal with the same issues that will be discussed in this paper. However, I have not found any that deal with the depth of data that I will pursue in this report. Many of them deal with just a few years' worth of data. This paper will aim to look at all the information between 1962 and 2019. This will show us throughout the years which statistics provide the best addition to the amount of wins a team gets in a single season. Potentially, I will contribute information that should be helpful to baseball executives while they make decisions on who to sign during the baseball offseason. It will encompass a large amount of data which will make it more thorough as it will not just focus on a single or few years. Baseball has changed over the years, so I expect not one statistic to have an enormous effect with wins but there should be a statistic that is at least moderately effective with a team's success.

2 RELATED WORK

2.1 Completed Work #1

There are many related works that have already been published online. Mark Freker, a student, used the Pearson correlation coefficient method to compare statistics [1].
The Pearson correlation coefficient is found using the following formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

where r = correlation coefficient, x_i = values of the x-variable in a sample, $\bar{x}$ = mean of the values of the x-variable, y_i = values of the y-variable in a sample, and $\bar{y}$ = mean of the values of the y-variable.

Values will range between -1 and 1. A value close to negative 1 means a negative correlation (one variable goes up, while the other goes down), a value close to 1 is a positive correlation (one variable goes up, the other goes up), and a value close to 0 means there is no correlation. Freker found there to be a high correlation between Ks per Walk and Winning Percentage, a moderate correlation between Home Runs and Winning Percentage, and a medium correlation between On Base Percentage and Winning Percentage. In his report, not only did he explain through words how each statistic was correlated with wins, but he also visualized them with scatter plots. Visualization is key as it helps the reader see the relationship between one variable with the other.

2.2 Completed Work #2

Connor Wolf from Samford University did a comparison between pitching, fielding and batting to see which of these three components of baseball best affects the total amount of wins a team gets in a season [2]. He used statistics from the 2019 baseball season to calculate his findings. First off, for each of these three parts of the game, Wolf found both the correlation coefficient and the equation of the linear regression line.
The Simple (one independent variable) Linear Regression line formula is as follows:

$$y = \beta_0 + \beta_1 * x + e$$

where y = Response/Dependent Variable, β_0 = y-intercept, β_1 = slope of independent variable x, x = independent variable, e = error term.

For batting, he compared batting average to winning percentage. There was a correlation coefficient of 0.716, which is high. Having a correlation coefficient of 0.716 means that if batting average goes up, winning percentage also goes up. The linear regression line was found to be y = 5.055x – 0.7669. The important thing to look at is the slope of this equation. It is 5.055. This means as you increase batting average by one percent, winning percent goes up by about five percent. He did these calculations for all three parts of the game. At the end, Wolf concluded that batting and pitching do more to affect total amount of wins than fielding. In his report, he not only gave readers the statistics and explanations, but also showed scatter plots with a linear regression line superimposed on them.

2.3 Building on Prior Work

My work will build upon these two related work and others by using more data to do my calculations. As mentioned earlier, most of the works that I found that were already done were based on a single season or a handful of seasons. This, in my opinion will not be enough data to claim which statistics are the best at affecting wins in baseball. My report should provide new knowledge, and even though the game has changed over the years, prioritizing pitching in some years and batting in others, will show statistics that went through all those phases of baseball and continued to affect win total. Hopefully, it will bring new insights to baseball fans and to executives, who choose which type of player to pursue and which statistics their team to value.

3 PROPOSED WORK

3.1 Dataset

The dataset I will be using for this report will be from Sean Lahman's website. Sean Lahman is a journalist and investigative reporter for the USA Today Network. He has a great site that provides baseball data to anyone for free. This data goes back all the way into the 1800s, when baseball was invented and became a professional sport. He also has the data in many different forms, such as .csv, .R, and even as a SQL file. This is all wonderful since it allows anyone to analyze and dabble with the enormous datasets he provides. Each dataset includes many essential statistics such as year, team, G (Games), W (Wins), SB (Stolen Bases), ERA (Earned Run Average), and FP (Fielding Percentage).

3.2 Main Tasks – Statistical Analysis/Visualization

Relevant statistics for this project include R, AB, H, 2B, 3B, HR, BB, SO, SB, CS, HBP, SF, RA, ER, ERA, CG, SHO, SV, IPouts, HA, HRA, BBA, SOA, E, DP, FP. They stand for Runs scored, At bats, Hits by batters, Doubles, Triples, Homeruns by batters, Walks by batters, Strikeouts by batters, Stolen bases, Caught stealing, Batters hit by pitch, Sacrifice flies, Opponents runs scored, Earned runs allowed, Earned run average, Complete games, Shutouts, Saves, Outs Pitched (innings pitched x 3), Hits allowed, Homeruns allowed, Walks allowed, Strikeouts by pitchers, Errors, Double Plays, Fielding percentage, respectively. These are all the relevant statistics that measure the offensive and defensive parts of the game.

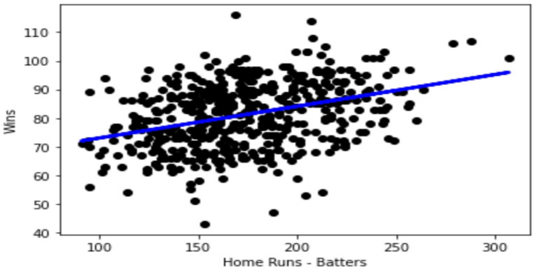
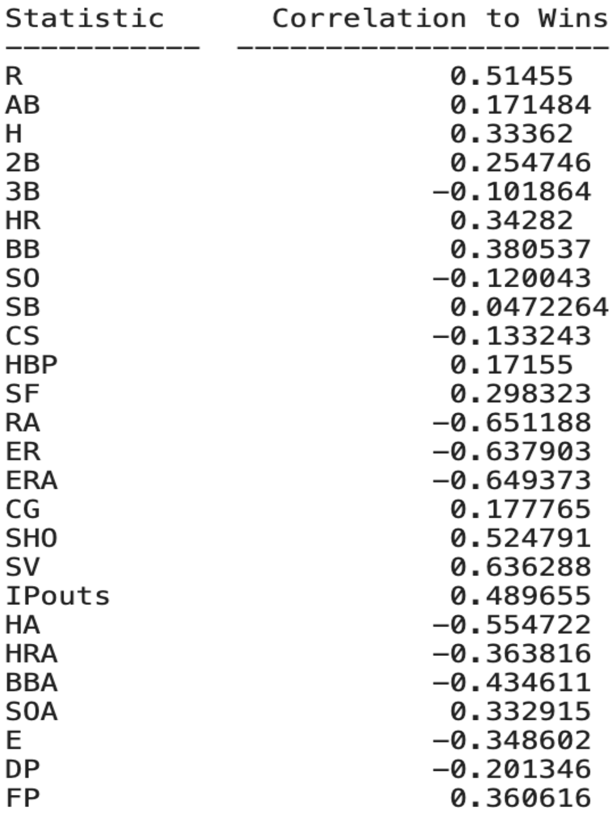


Figure 1: Plot of Home Runs (by batters) vs. Wins (Training)

First, let's discuss Linear Regression. The above plot represents just one of all the plots made which each compares statistic vs. wins for the training set. This plot compared home runs hit by batters to number of wins. Each point represents a specific team in history. The linear regression line was found to have a coefficient of 0.10777187. This means for every home run hit by a batter, it increases their wins by about 0.11 on average. Further analysis of the visualization shows that there is a huge blob in the middle of the plot. However, there are also outliers as you can see towards the edges of the scatter plot. These could indicate seasons in which one aspect of their game was excellent, but the other components of their game were very poor. For example, a team that hit a lot of home runs, but also gave up a lot which shows the lower win totals.

After doing all the calculations for Linear Regression, I realized that doing Correlation analysis would make it much easier for me to compare the statistics. This is because for Linear Regression, when you look at the coefficients of the independent variable (statistic), they mean different things. For example, a lower ERA (earned run average) is better than a higher one. However, for an offensive statistic like H (Hits), a higher total is better. So, when comparing these statistics, the coefficient can be both negative and positive and can each mean different things. Correlation analysis is the same however, at least the numbers can only be between -1 and 1. For Linear Regression, the coefficient can be anything, making it harder to compare.



```
Statistic      Correlation to Wins
-----------  ---------------------
R                          0.51455
AB                         0.171484
H                          0.33362
2B                         0.254746
3B                        -0.101864
HR                         0.34282
BB                         0.380537
SO                        -0.120043
SB                         0.0472264
CS                        -0.133243
HBP                        0.17155
SF                         0.298323
RA                        -0.651188
ER                        -0.637903
ERA                       -0.649373
CG                         0.177765
SHO                        0.524791
SV                         0.636288
IPouts                     0.489655
HA                        -0.554722
HRA                       -0.363816
BBA                       -0.434611
SOA                        0.332915
E                         -0.348602
DP                        -0.201346
FP                         0.360616
```

Table 1: Statistic vs. Correlation to Wins (Training)

The above table shows each statistic with their own correlation to wins. As you can see, the largest correlation belongs to RA (opponents runs scored). The correlation to wins for RA is -0.651188. This means as runs allowed goes up, wins goes down. This makes sense as if you allow more runs to score, the less chance you have of catching up to your opponent and winning the game. Just like the plot above for Linear Regression, all these correlations are from the training data.

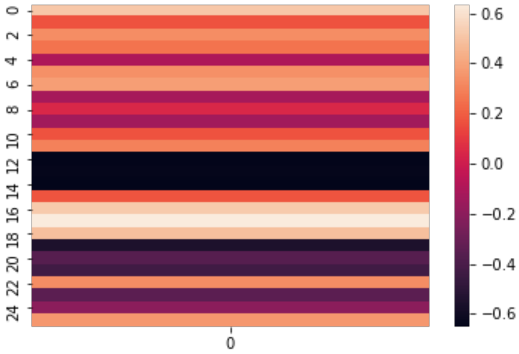


Figure 2: Heat map of correlation analysis (Training)

The above figure shows a heat map of all the correlations of the statistics in this project to wins. The right side shows the different colors as you go lower or higher in correlation. The left side shows each of the individual statistics by number. R is 0 and FP is 25. The black chunk in the middle represents the negative correlation for several pitching statistics which all require a lower number to mean success (RA, ER, and ERA).

3.3 Main Tasks – Data Preprocessing

Data preprocessing includes making sure to filter only for data points with 162 games (a standard season in MLB) in order to fairly compare the data points with each other. If one datapoint had 50 for games and another had 162, the analysis would not be fair or would it make any sense. The lower the number of games played, the lower the statistics such as home runs will be. Data has also been split 80/20 into training and testing sets. This is a standard split in the data science world. Finally, all the rows with NA values have been removed as they are not useful for our analysis. There are ways to replace these values such as using dummy variables of 1s and 0s, using the average of the column the data is missing, etc. However, I will just remove them for this project.

3.4 Main Tasks – Specific Questions/Patterns to Explore/Model

The first question I wanted to answer when doing my project is whether home runs or earned run average has a stronger effect on wins. The answer, by looking at the correlation coefficients, is ERA. ERA has a correlation coefficient of -0.649373 whereas HR has a correlation coefficient of 0.34282. This shows that ERA affects wins more than home runs. The strongest effect on wins I found earlier as RA. It has a correlation coefficient of -0.651188. The statistic with the lowest effect on wins is SB, stolen bases. It has a correlation coefficient of 0.0472264. That means there is no correlation between stolen bases and wins. For patterns within the same component of the game such as offense, we see a pattern in which the correlations are positive (for the most part). However, not all are correlated the same amount to wins. Some are in the 0.30's but SB as we mentioned, is at 0.04. So, to answer the last question, yes there seems to be a pattern but not to the extent that you would expect.

4 EVALUATION

4.1 Evaluation Metrics - Effectiveness/Efficiency

In order to see the effectiveness of the methods, I will be using the testing data to see if the model I produced with the training data is effective and produces similar results to the actual testing data. If it is effective, the model should produce similar values to the testing data. To measure efficiency, I will be looking at the run times of the methods. The lower the run time, the more efficient the process is. This is due to the process finishing faster, making it more efficient.

4.2 Experimental Setup

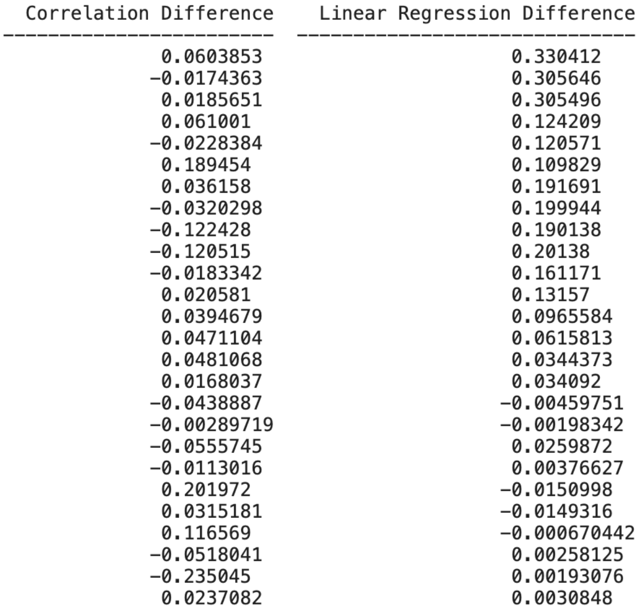
As mentioned earlier, I will split the data into two sets, a training set and a testing set. This is so I can ensure that I will have a fair evaluation of my model. If I used the same data for both training and testing, it would not be fair or reasonable to assume that in the real world will the model predict so accurately. Thus, splitting the data into two avoids this and creates a reasonable test for evaluation. The data will be split 80/20, meaning, 80% of the datapoints will be used to train, whereas 20% of the data will be used to test.

4.3 Methods to Compare

I will be also comparing both methods of correlation analysis and linear regression to see which is more effective and efficient. This will help me see which method I should value more in my analysis. If one is higher than the other in efficiency whereas the other is higher in effectiveness, I may be able to choose either model to claim as better. Having a clear-cut winner is preferable, meaning they beat the other method in both effectiveness and efficiency. However, I know in the real world, this does not always occur.

4.4 Key Results

First, I tested the effectiveness of the Linear Regression method. The way I did this was to use the trained model to predict the test values for W. If the training model was effective, the difference between predicted values and actual values would be small. I took the average of the predicted values for each variable and found the difference with the actual values. You will see them in the chart below. For correlation, I simply took the difference between the correlations for testing variables and training variables. Same as Linear Regression, if the correlations of the training set are close to the values for the testing set, the model is effective.



```
  Correlation Difference    Linear Regression Difference
------------------------  ------------------------------
             0.0603853                        0.330412
            -0.0174363                        0.305646
             0.0185651                        0.305496
             0.061001                         0.124209
            -0.0228384                        0.120571
             0.189454                         0.109829
             0.036158                         0.191691
            -0.0320298                        0.199944
            -0.122428                         0.190138
            -0.120515                         0.20138
            -0.0183342                        0.161171
             0.020581                         0.13157
             0.0394679                        0.0965584
             0.0471104                        0.0615813
             0.0481068                        0.0344373
             0.0168037                        0.034092
            -0.0438887                       -0.00459751
            -0.00289719                      -0.00198342
            -0.0555745                        0.0259872
            -0.0113016                        0.00376627
             0.201972                        -0.0150998
             0.0315181                       -0.0149316
             0.116569                        -0.000670442
            -0.0518041                        0.00258125
            -0.235045                         0.00193076
             0.0237082                        0.0030848
```

Table 2: Correlation Difference vs. Linear Regression Difference

This table shows the differences for Correlation analysis on the left and Linear Regression on the right. After solving for the mean of each column, correlation has a mean difference of 0.006819480504763156 and Linear Regression has a mean difference of 0.0999535555398691. This shows that both methods are effective, however, Correlation analysis is more effective than Linear Regression. In terms of efficiency, I found the run times of each iteration of each method to be extremely small and negligible, so it is basically a tie in efficiency. If I had to pick a method, it would be Correlation analysis due to it being more effective than Linear Regression and it being easier to compare as its values are only between -1 and 1.

5 DISCUSSION

5.1 Timeline

In terms of project timeline, I have already finished Week 2 – Project Proposal and Week 3 – Project Checkpoint. I will aim to finish this final report within the next few days. I will then work on my slides and submit both sometime this week. I am currently working on this report, and I am not going to work on both report and slides simultaneously. After both are submitted and I have reviewed my peers' submissions, I will move on to my final week, Week 5. I will create a video in which I will present my final slides to my classmates.

5.2 Potential Challenges/Backup Plan

There were potential challenges as I could have made errors in my code. I made sure that my code is correct, to the best of my ability, in order to create accurate analysis and conclusions. I could have also found conflicting conclusions online from prior works, but this is to be expected as those works mostly only deal with a single season or few seasons. If the dataset I was using for this project did not work well for my analysis for some reason, I would have scoured the internet for additional datasets that I could use for my project. That was my backup plan, if it was needed. Having a backup plan or alternative approaches is key because not always will you find a situation where everything works out. You should always have a backup plan/alternative approach in order to continue working on your report and not hit a roadblock.

5.3 Accomplishments/Lessons

Accomplishments-wise, I have written a 5-page report about a topic I am highly interested in as I am a big baseball fan. I have analyzed data using a large dataset and Python. I have sharpened my toolkit when it comes to analyzing using statistical methods and programming. In terms of lessons, I have learned ways to visualize my data and to leverage what I have learned in this report and in my analysis. I have also learned that when splitting your data using the train_test_split function, you are going to get different sets each time you run the cell. So, in order to keep consistent data, make sure to save the training and testing sets to variables and use them from that point on. Another lesson I learned was to utilize more for loops. They help a lot in saving time. Without the loop, I was doing each iteration one by one which was annoying and time-consuming.

6 CONCLUSION

6.1 Project Summary/Key Findings

This project analyzed the relationships of various statistics in baseball to wins. These various statistics are relevant to the offensive and defensive parts of the game (batting, pitching, and fielding). It looked at Linear Regression and Correlation analysis to figure out how these statistics are related to wins. Correlation analysis was favored over Linear Regression due to its practical nature of comparison among variables. Some key findings include RA having the strongest effect on wins and SB having the weakest through correlation analysis. I also found ERA affects wins more than home runs. In terms of effectiveness, I found Correlation analysis to be more effective than Linear Regression. For efficiency, it was a tie as both completed their tasks in negligible times.

6.2 Future Work

Future work to be done that relates to this project includes using another dataset to see if you can get the same result. I only used one dataset for this project since it had all the variables I needed, and it was deep as it went back to the 1800s. However, confirming your analysis with another dataset is important if you have the time and resources, in order to verify that your initial conclusions of your analysis with your first dataset is correct. This can be done as soon as possible or in the next decade as we get more data from the seasons coming ahead. Having more data is better than having less data.

Some additional questions to explore include whether advanced metrics [5] like Defensive Efficiency Ratio (DER), Batting Average on Balls in Play (BABIP), and Adjusted Earned Run Average (ERA+) affect wins more strongly than the standard metrics used in this project. If so, we can search for a dataset that includes such metrics. Baseball today is incorporating more advanced Sabermetrics into the game so it would make sense to analyze with these kinds of data. Another question to ask is whether weather has any effect on wins. We can see whether teams tend to perform better on cloudy, windy, warm, cool, sunny, etc. days.

We can also use other methods to do our analysis. I took a more statistical approach in this project by doing Correlation analysis and Linear Regression. Using machine learning tools such as unsupervised or supervised learning, and methods such as clustering, classification, anomaly detection, etc. may help add additional insight to our analysis. For example, we can use clustering to see whether a certain amount of home runs is more likely to be added to a higher win cluster or a lower win cluster. We may also see what the anomalies in the data are using anomaly detection.

Finally, we can also conduct more experiments in order to see whether certain statistics affect win total. The more experiments we do, the better we can be at making conclusions. We can see how many experiments lean one way as opposed to the other way. If there is a consensus, that would be most desirable. A larger scale project may be useful in our analysis. I already used a large dataset with as many data points I could use, however, if there is a larger dataset out there with less NA values, it would be very useful for future work on this topic.

REFERENCES

[1] Frerker, M. (2013, March 15). Measuring Correlation between Statistics and Wins in Major League Baseball. Retrieved September 24, 2021, from https://courses.cs.washington.edu/courses/cse140/13wi/projects/frerkerm-report.pdf.

[2] Wolf, C. (2019, July 17). Batting, pitching, or fielding: What's most important in today's MLB? Samford University. Retrieved September 24, 2021, from https://www.samford.edu/sports-analytics/fans/2019/Batting-Pitching-or-Fielding-Whats-Most-Important-in-Todays-MLB.

[3] Lahman, S. (2021, April 9). Teams: Teams table in lahman: Sean 'lahman' baseball database. R Package Documentation. Retrieved September 28, 2021, from https://rdrr.io/cran/Lahman/man/Teams.html.

[4] Lahman, S. (2021, February 16). Download Lahman's baseball database. SeanLahman.com. Retrieved September 28, 2021, from http://www.seanlahman.com/baseball-archive/statistics/.

[5] Advanced stats: Glossary. MLB.com. (n.d.). Retrieved September 29, 2021, from https://www.mlb.com/glossary/advanced-stats.