What Contributes to a Major League Baseball Team's Wins in a Season?

Selah Dean

STAT 325

Fall 2023

Abstract

The use of data in baseball has increased over the years as teams are trying to gain every advantage they can. This project looks at determining what aspect of a team performance contributes to a team's number of wins. The data came from Baseball-Reference.com and included basic team statistics for all 30 MLB teams in batting, pitching, and fielding for the 2000 through 2023 MLB seasons (excluding 2020 due to the shorten season). The final dataframe included 690 observations for 34 variables which consisted of 14 batting statistics, 12 pitching statistics, and 4 fielding statistics all of which were counting statistics. The remaining variables were team, year, number of wins, and winning percentage. Using this data, I created four different linear models with the 2000 through 2022 season data as training data and the 2023 season as testing data. I found that the best model included 11 predictors and had an adjusted R-squared of 0.978 for both the training and test data. These results are able to give insights into what makes a team successful and help them find areas for improvement.

Introduction

The goal of any Major League Baseball team is to win the World Series, but in order to do so they must do well in the regular season. All 30 teams are trying to find the secret formula to success which may be different for each team. However, there tends to be some commonalities between teams that find success. There are three main components that a team is trying to balance: batting, pitching, and fielding. Ideally a team would want to do well in three aspects, but there are times where they are forced to sacrifice one for another. The question is what matters the most when it comes to winning. Predicting wins is something that both teams themselves do as well as other individuals not associated with the front offices of teams.

In 2001 the concept of Moneyball greatly impacted the game of baseball as data and analytics began to play a bigger role in how baseball teams operated. Bill James was a pioneer in the sabermetric revolution and created many statistics that are used in the game today (*A Guide*). He created the Pythagorean Winning Percentage which predicts a team's actual winning percentage based on runs scored and runs allowed (Rothman, 2014). This is a relatively simple formula given by: $W\% = \frac{R^2}{R^2+RA^2}$ where R represents runs scored and RA represents runs allowed (Rothman, 2014). While this is an effective measure of win percentage, it doesn't explain the components that go into both scoring runs and preventing runs. Because of this, an area of research is considering different batting, pitching, and fielding statistics and their impact on a team's wins. The goal of this project is to determine what subset of statistics and what model best predicts a baseball team's wins.

Methods

The data used for this project came from Baseball-Reference.com which provides statistics for every player in Major League Baseball history. For this project I used four different

tables of team statistics for the years 2000 through 2023 (excluding 2020 since the season was shorten due to the COVID pandemic). The tables I used where the Team Standard Batting, Team Standard Pitching, Team Standard Fielding, and each division's standings' table. In R, I used web scraping to get all of the tables for all of the years. I added a column to each table that represented the year the table came from so I was able to join the tables together based on year and team. Since I scraped 6 different tables to get the standings for all 30 teams, I created one table for standings for each year by binding the rows of the separate tables.

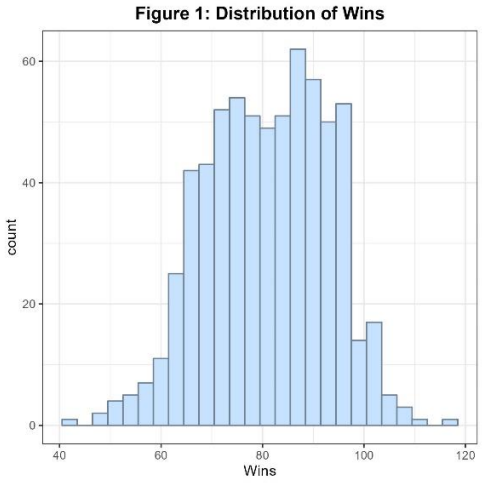
Before joining all of the tables together, I chose a subset of features for each table. My focus was on counting statistics only and no statistics that were previously calculated. This is the reason why I excluded the 2020 season since only 60 games were played instead of the traditional 162 game season. The original batting table had 30 features, but I reduced it to 16 features which included: plate appearances, runs scored, hits, doubles, triples, home runs, stolen bases, caught stealing, walks, strikeouts, double plays, hit by pitch, intentional walks, and runners left on base. The original pitching table had 37 features, but I reduced it to 14 which included: shutouts (no runs allowed in a game by one or more pitchers), hits allowed, runs allowed, earned runs allowed (runs scored without the aid of errors), home runs allowed, stolen bases allowed, walks allowed, strikeouts, double plays, hit batters, intentional walks issued, and runners left on base. The original fielding table had 20 features, but I reduced it to 6 features which included putouts (outs made by a fielder), assists (credited to every defensive player who fields or touches the ball prior to the recording of a putout), errors (when a fielder fails to make an out on a play that an average fielder would make), and double plays. Finally, the original standing table had 6 features, but I reduced it to 4 features which included: number of wins and

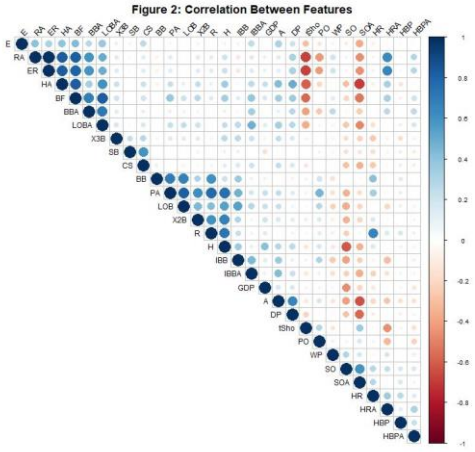
winning percentage. All four tables also included a column for the team and a column for the year.

After choosing the subset of features I performed an inner join by team and year on the four tables to create the dataframe that I used to create models to predict a team's number of wins. The final dataframe includes 34 features and 690 observations. The last step I did was to separate the dataframe into training and test data. I filtered all observations for years 2000 through 2022 to be included in the training data which gave 660 observations. The remaining 30 observations were from the 2023 season and made up my test set.

Results

This project is looking at predicting wins in a season using different team statistics. I created four different models which used data from the 2000 though the 2022 MLB seasons (excluding the 2020 season) as training data and the 2023 season as testing data. I evaluated the models on the data used to create the models as well as to predict the number of wins for the 2023 season. The number of wins in a season for each MLB team from 2000 through 2022 can be seen in the Figure 1.

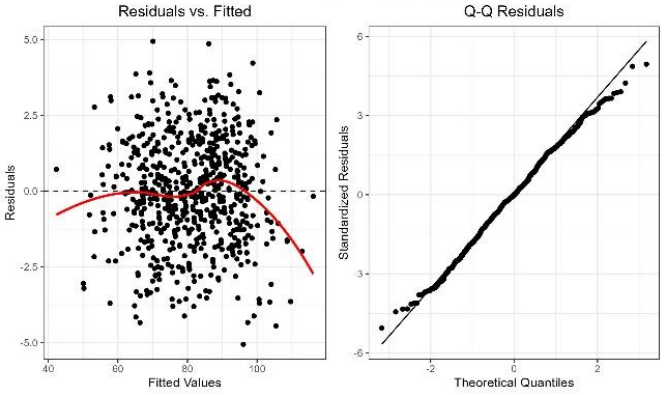




The histogram shows that the number of wins is approximately normally distributed meaning that a linear model may be able to effectively predict the number of wins. An issue that may arise is multicollinearity as seen in Figure 2 many of the statistics are highly correlated.

The first model that I fit was a multilinear regression model with number of wins as the response variables and the 30 different team batting, pitching, and fielding statistics as predictors. The p-value for the model utility test is approximately zero and the adjusted R-squared is 0.978. The summary of the model is displayed in Table 1 (see appendix). While the overall model is significant and the adjusted R-squared value suggests that the model is a good fit, Table 1 shows that a majority of the individual predictors have p-values greater than 0.05 suggesting that they are not significant. Figure 3 shows the Residual vs Fitted Plot and a QQ-Plot of the Residuals for this model, which show the residuals to be approximately normal and randomly distributed.

Figure 3: Full Model Residuals

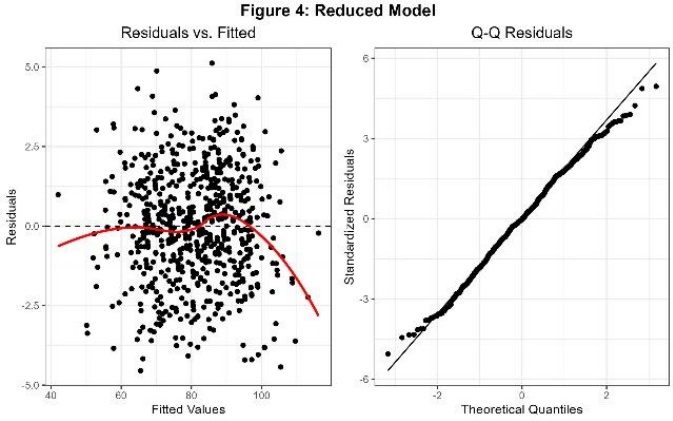


This suggests that the model is a good fit for the data. However, the model is complex which could lead to overfitting especially with the presence of multicollinearity in the data.

I decided to use stepwise selection with Akaike information criterion (AIC) to address the issue of multicollinearity. I started with the full model and used forward and backwards selection to get a model with only 11 predictors which were plate appearances, runs scored, doubles,

triples, runners left on base (by hitters), team shutouts, runs allowed, batters faced, runners left on base (by pitchers), and double plays (by fielders). The p-value for this model is approximately zero and the adjusted R-squared is 0.978. Figure 4 shows the Residual vs Fitted Plot and QQ-Plot of the Residuals for this reduced model which shows approximately normally distributed residuals, and no patterns presents which suggests that the reduced model is a good fit for the data.

Figure 4: Reduced Model

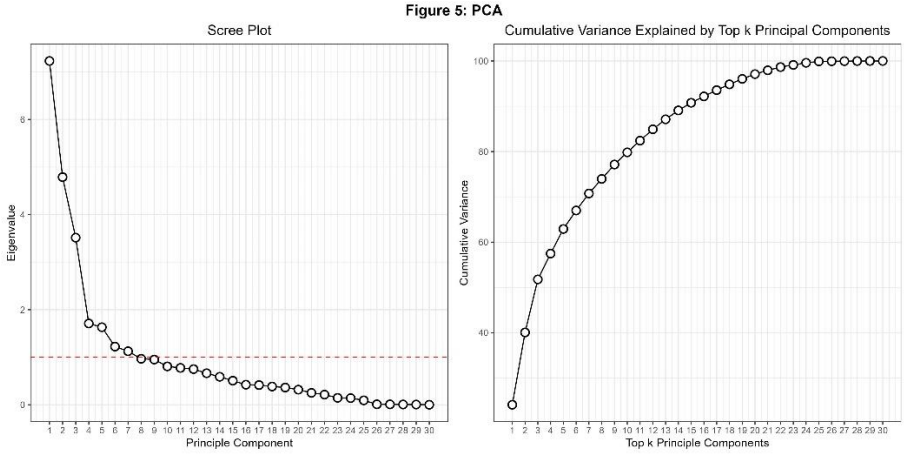


The summary of the reduced model is given by Table 2 (see appendix), and it can be seen that all but two of the individual predictors are significant at a significance level of 0.05.

An ANOVA test between the full model and this reduced model gives a p-value of 0.9729 which means that the full model does not provide more information and the reduced model is adequate at predicting the number of wins. The equation for wins given by the reduced model is as follows $\mathrm{W} = 74.903 - 0.307 * \mathrm{PA} + 0.317 * \mathrm{R} - 0.007 * 2\mathrm{B} - 0.011 * 3\mathrm{B} + 0.308 * \mathrm{LOB} + 0.057 * \mathrm{tSho} + 0.005 * \mathrm{HA} - 0.319 * \mathrm{RA} + 0.308 * \mathrm{BF} - 0.311 * \mathrm{LOBA} - 0.008 * \mathrm{DP}.$

Another technique I used to address the issue of multicollinearity was by applying Principal Component Analysis to the features before fitting the model. Figure 5 shows the Scree Plot as well as a plot of the Cumulative Explained Variance for each principal component.

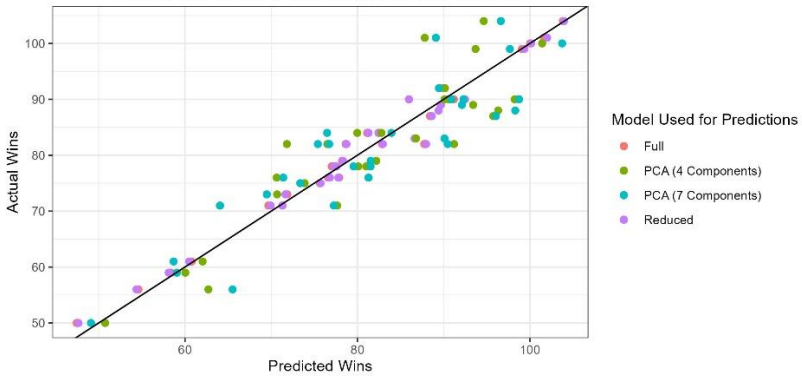
Figure 5: PCA



In order to select the number of principal components to use when fitting the model, I used two different techniques. First, I used the elbow method on the Scree Plot. There appears to be an elbow in the plot after four components, so I created a model using four components. This model has a p-value of approximately zero and an adjusted R-squared of 0.780 . The second technique I used to select the number of principal components was the Kaiser Criterion which selects any component which has an eigenvalue greater than 1. This technique selects seven components and a model fitted with seven components has a p-value of approximately zero and an adjusted R-squared of 0.804.

I applied each of the four different models to the 2023 season data. Figure 6 shows the predicted number of wins for each team in the 2023 season versus their actual win total.

Figure 6: Actual vs Predicted Number of Wins



The adjusted R-squared the 2023 season for the full model is 0.978, for the reduced model it is 0.978, with four principal components it is 0.799, and with seven principal components it is 0.811.

Discussion and Conclusion

The goal of this project was to determine what subset of team statistics and what model best predicted a baseball team's number of wins in a season. I created four different models to try and answer this question. All four of the models were linear models and I evaluated their accuracy with their adjusted R-squared for both the data used to create the model which came from the 2000 though 2022 season (excluding 2020) and on the held-out test data which was the 2023 season. The first model I created used all 30 predictors that I originally selected. These predictors included 14 team batting statistics, 12 team pitching statistics, and 4 team fielding statistics. A linear model with all of the predictors included was able to predict a team's number of wins in a season with a high accuracy. The adjusted R-squared for the model was 0.978 and the adjusted R-squared for the test data was 0.978 as well. Despite multicollinearity being present among the predictors, this model did not overfit on the testing data.

However, the model was complex with all 30 predictors, so I wanted to create another model that used stepwise selection to select a subset of the predictors. This reduced model included 11 predictors with 5 batting statistics, 5 pitching statistics, and 1 fielding statistic. The reduced model was also had a high accuracy when it came to predicting a team's number of wins as the adjusted R-squared was 0.9729 for the 2000 through 2022 seasons and 0.978 for the 2023 season. The adjusted R-squared for the reduced model was approximately the same as the full model but it included almost a third of the predictors.

Both models included fewer fielding statistics but were able to accurately predict a team's number of wins this suggests that better batting and pitching may be more significant in a team winning games. However, there are fewer standard fielding statistics collected so advanced statistics for fielding measures may suggest that good fielding has more significance in winning games.

In the reduced model, two of the predictors with the largest slopes where runs scored and runs allowed which is similar to the results found in the more complex model of the Pythagorean Winning Percentage. However, the reduced model is able to give more information about what a team needs to succeed other than scoring more runs and allowing fewer runs.

The final two models I created used Principal Component Analysis on the features. I chose to use PCA since there was multicollinearity present in many of the predictors and I wanted to reduce the effects. The two models used two different numbers of principal components but both models had much lower accuracies than the first two models created. The adjusted R-squared for the 2000 through 2022 season data was 0.78 and 0.804 with four and seven principal components respectively, and the adjusted R-squared for the 2023 season was 0.799 and 0.811. While the principal component analysis was used to create a model where the features were not correlated with each other, the resulting models were worse than using linear models with correlated features.

This project showed that a simple linear model with standard statistics that are collected is able to predict a team's number of wins in a season with a high accuracy. This is helpful for teams as they can identify areas of weakness in their teams and work on improving them by either working with players or acquiring new players to hopefully increase their win total in a season. However, there exists more advanced statistics and further analysis could be done that

used these statistics instead to predict a team's season win total. Another area of research would be to look at a team's standard statistics year by year to determine if there are commonalities in the statistics throughout the years and the number of wins for each year. Every team is trying to find what makes their team and winning team, and this research was able to provide insight into what contributes to a team's success in a season.

Appendix

Table 1: Full Model Summary

		W	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	73.97	56.03 – 91.91	**<0.001**
PA	-0.31	-0.32 – -0.29	**<0.001**
R	0.32	0.30 – 0.33	**<0.001**
H	-0.00	-0.01 – 0.01	0.838
X2B	-0.01	-0.02 – -0.00	**0.014**
X3B	-0.01	-0.03 – 0.00	0.140
HR	-0.00	-0.01 – 0.01	0.708
SB	-0.00	-0.01 – 0.00	0.214
CS	0.01	-0.01 – 0.03	0.334
BB	-0.00	-0.01 – 0.01	0.689
SO	0.00	-0.00 – 0.00	0.969
GDP	-0.00	-0.02 – 0.02	0.921
HBP	0.00	-0.02 – 0.02	0.923
IBB	0.01	-0.00 – 0.02	0.194
LOB	0.31	0.29 – 0.33	**<0.001**
tSho	0.06	0.01 – 0.11	**0.024**
HA	0.01	-0.00 – 0.02	0.087
RA	-0.31	-0.35 – -0.28	**<0.001**
ER	-0.01	-0.02 – 0.01	0.443
HRA	-0.00	-0.01 – 0.01	0.425
BBA	0.01	-0.00 – 0.02	0.252
IBBA	-0.01	-0.02 – 0.00	0.174
SOA	0.00	-0.00 – 0.00	0.878
HBPA	0.01	-0.01 – 0.02	0.513
WP	-0.01	-0.02 – 0.01	0.256
BF	0.30	0.27 – 0.34	**<0.001**
LOBA	-0.31	-0.35 – -0.28	**<0.001**
PO	0.00	-0.03 – 0.04	0.896
A	-0.00	-0.00 – 0.00	0.602
E	-0.00	-0.02 – 0.01	0.889
DP	-0.01	-0.03 – 0.00	0.083
Observations	660		
R^2 / R^2 adjusted	0.979 / 0.978		

Table 2: Reduced Model Summary

		W	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	74.90	59.28 – 90.53	**<0.001**
PA	-0.31	-0.32 – -0.29	**<0.001**
R	0.32	0.31 – 0.33	**<0.001**
X2B	-0.01	-0.01 – -0.00	**0.030**
X3B	-0.01	-0.03 – 0.00	0.149
LOB	0.31	0.30 – 0.32	**<0.001**
tSho	0.06	0.01 – 0.11	**0.020**
HA	0.00	0.00 – 0.01	**0.004**
RA	-0.32	-0.33 – -0.31	**<0.001**
BF	0.31	0.30 – 0.32	**<0.001**
LOBA	-0.31	-0.32 – -0.30	**<0.001**
DP	-0.01	-0.02 – 0.00	0.084
Observations	660		
R^2 / R^2 adjusted	0.978 / 0.978		

Works Cited

A Guide to Sabermetric Research. Society for American Baseball Research. (n.d.). https://sabr.org/sabermetrics.

Rothman, S. (2014). A New Formula to Predict a Team's Winning Percentage. *The Baseball Research Journal*. https://sabr.org/journal/article/a-new-formula-to-predict-a-teams-winning-percentage/.

Sports Reference LLC. Baseball-Reference.com - Major League Statistics and Information. https://www.baseball-reference.com/.