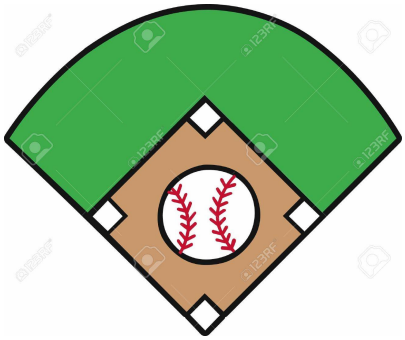
CAPSTONE PROJECT 1

BASEBALL PREDICTING WINS

Scott Lew

Sept , 2017



Milestone Report Capstone

1. Introduction

The goal of this project is to predict a baseball teams wins for a single season using a combination of batting and pitching statistics. It is assumed that some of these predictors have causation and could identify areas of a team's performance that needs improvement in order to win more games. For instance, if a team's batting average is determined to be a strong predictor of a team's wins, then batting average is an area that could be focused on by a coach or manager in order to improve performance.

Background

Baseball is considered America's pastime, and one of it's oldest games
During a Major League Baseball (MLB) season, a total of 162 games are played in the modern era of the game. Two opposing teams take turns playing offense, batting, and playing defense, pitching and fielding.

Client

The prospective clients are baseball coaches and managers in professional, collegiate, and amateur leagues who are interested in winning more games.

Approach

Batting and pitching statistics from 1876-2018 will be collected from the web and merged into a single table as a Pandas dataframe. Then, regression analysis will be applied to determine the features/variables that are most useful in predicting a team's wins for a season.

Deliverables

The code will be written in the form of Jupyter Notebooks will be displayed and shared on Github. In addition, slides and a written reports will also be available on Github.

2. Data Wrangling and Data Cleaning

Dataset

The data set consists of both batting and pitching statistics for all major league baseball (MLB) teams in the years 1876-2018. The data was scrapped from the website baseball-reference.com using the Beautiful Soup program. A batting statistics table and a pitching statistics table were merged for by joining on team and year columns. Then, tables for each season were concatenated to obtain all the MLB statistics dating from 1876 to 2018. 2019 statistics were not used because at the time of writing this report the 2019 season is not complete.

For this project, statistics were taken for all Major League Baseball (MLB) seasons from 1876 to 2018.

The batting statistics are made up of 28 features while the pitching statistics contain 35 features. These statistics will be described in greater detail in other sections of this report.

All batting statistics were downloaded and saved as csv files and then converted into pandas dataframes which were subsequently concatenated. This same approach was also used for the pitching statistics data. Then, the dataframes for pitching and batting were merged using an inner join on two columns: Team(Tm) and Year, which produced a dataset consisting of 2815 rows and 65 for the subsequent analysis.

For the linear regression analysis, it was necessary to remove rows from the data that had nan values, which left 1618 rows of data for regression models. Unfortunately, the regression analysis with these predictor variables required losing 1197 rows. However, as will be shown, the regression models were able to predict wins.

3. Exploratory Data Analysis

Data Features

Multiple features were used to predict wins. These predictor variables also known as regressors include the following:

#Bat: Number of players used in game

BatAge: Batters average age

G_x: Games played by team

PA: Plate appearances

AB: At bats

R_x: Runs scored offensively.

H_x: Hits produced by team's batting

2B: Double hits on which the batter reaches second base safely without the contribution of a fielding error.

HR_x: Home runs produced by team's batters.

BB_x: Bases on balls and walks

SO_x: The number of strikeouts made by team's batters

OPS+: A metric consisting of a team's on-base plus slugging percentage and normalizes the number across the entire league.

TB: Total bases made by teams batting, where singles counts as 1 base, doubles as 2 bases, triples as 3 bases and home runs as 4 bases.

#P: The number of pitchers used in games

BB_y: bases given up as walks or balls by team's pitching

BF: Batters faced

BK: Balks. A balk is an illegal act by the pitcher when one or more runners are on base.

CG: Complete games

ER: Earned runs allowed by pitching

G_y: Games pitched

GF: A relief pitcher is credited with a game finished (denoted by GF) if he is the last pitcher to pitch for his team in a game.

H_y: Hits allowed by team's pitching

HR_y: Home runs allowed by team's pitching

IP: Innings pitched

PAge: The average age of the team's pitchers.

R_y: Runs allowed by pitching

SO_y: Strikeouts produced by team's pitching

SV: A save is awarded to the relief pitcher who finishes a game for the winning team, under certain circumstances. It is awarded if the relief pitcher maintains his team's leads and pitches at least 3 innings. It is a metric of how well the team's relief pitchers are performing.

WP: A wild pitch (WP) is made by a pitcher when his pitch is either too high, too short, or too wide of home plate for the catcher to control with ordinary effort, thereby allowing a baserunner, perhaps even the batter-runner on an uncaught third strike, to advance. It is another metric to evaluate the team's pitching ability.

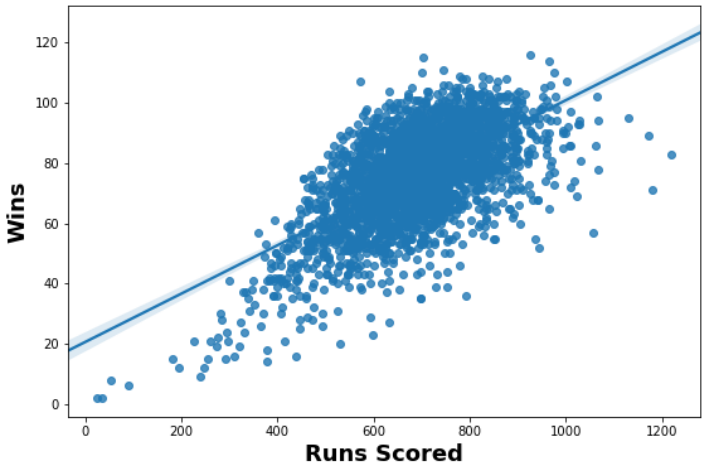
Correlation of Features to Wins

For a linear regression analysis to predict wins, it would be ideal to select those features that have a high correlation with wins.

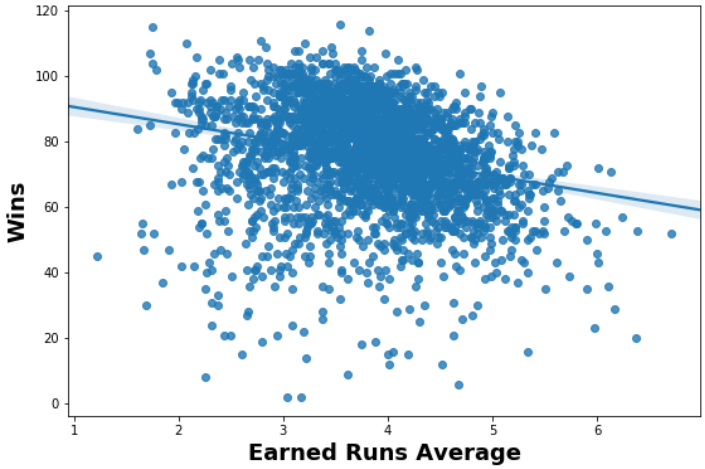
Scatter plots of several variables showing a positive or negative correlation with the target variable, wins are shown below.

Positive correlation of wins with hits.

As expected, there is a positive correlation of wins with runs produced by hitting, runs scored.



And, there is a negative correlation of Earned Runs Average (ERA) with wins.



Heatmap

Another way of visually displaying the correlation of features with one another is by using a heatmap. Features that have a strong positive correlation with wins are shaded more purple, while those features that have a negative correlation have a darker blue color.

Heatmap of correlation matrix of variables with wins

Independent Variables Correlation With Wins Table 1

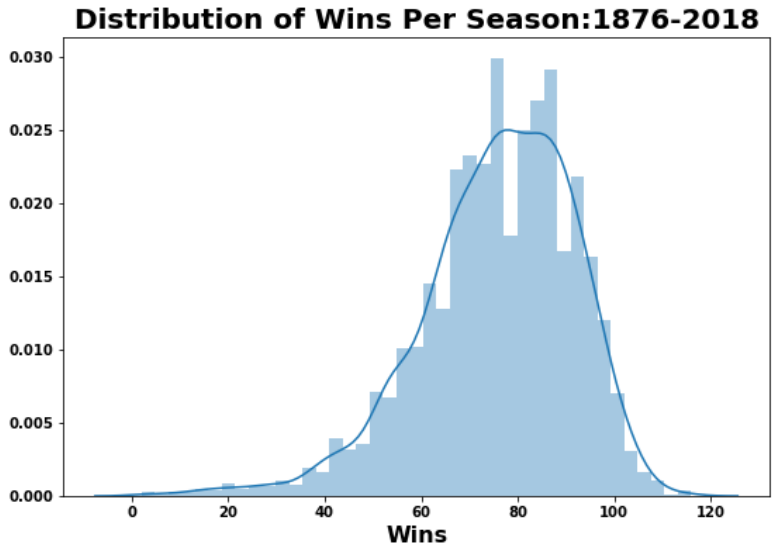
Feature	Correlation	Feature	Correlation
TB	0.649951	SLG	0.521056
H_x	0.649828	SV	0.482644
RBI	0.648062	LOB_x	0.359822
PA	0.629993	RA/G	-0.607445
R_x	0.626056	H9	-0.487337
IP	0.612138	WHIP	-0.336374
AB	0.599076	ERA	-0.242158
BB_x	0.586737	CG	-0.175771
G_y	0.582832	R_y	-0.162247
G_x	0.582830	WP	-0.107353
GS	0.582657	HBP_y	-0.092135
OPS	0.564913	IBB_y	-0.050684
ERA+	0.563912		
OPS+	0.549647		
OBP	0.533598		

As shown in the above table, Total bases(TB), batting hits (H_x), runs scored by batting (R_x), and runs batted in (RBI), innings pitched (IP) and saves (SV) all have a positive correlation with a team's wins. Total bases (TB) is the sum of all bases obtained from singles, doubles, triples, and home runs. RBI is defined as a statistic in baseball that measure the total number of runs a hitter generates off of their at-bats with exception to runs scored due to errors by the fielding team. Innings pitched (IP) are the number of innings pitched by a pitcher before being taken out of the game.

As shown in the above table, Runs Allowed Per Game(RA/G), and hits allowed per innings pitched (H9) have a negative correlation with a team's wins, which makes perfect sense.

Distribution of wins per season from 1876-2018

The distribution of team wins in a season has a relatively normal distribution that is skewed somewhat.



4. Machine Learning

Ordinary Least Squares (OLS) Regression Model

For Ordinary Least Square analysis, rows that had nan values in the feature columns were removed. The data was then split into training and test data sets.

In order to evaluate regression models, the coefficient of determination, R squared, which measures the amount of variance in the target variable that can be explained by the regressor variables, was calculated. In general, higher R squared values indicate the predictors have a strong relationship with the target variable.

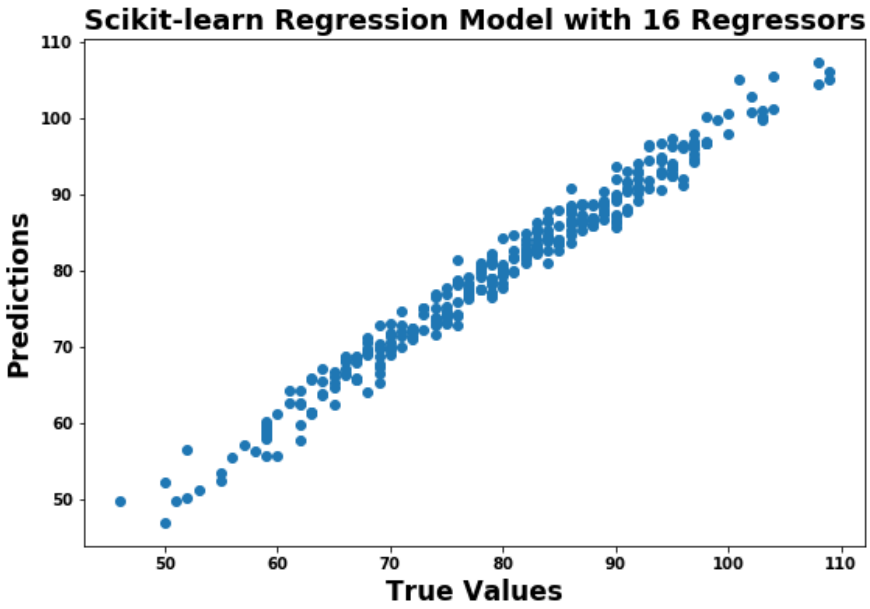
As a first step, 60 features were used to create an initial OLS regression model that had R squared of 0.98. Then, a backward stepwise regression method was utilized to select 33 of those original 60 features to create a second OLS model that achieved good results. In the stepwise regression approach, a model with many features is used in the beginning and each subsequent step removes some features in order to create a reduced model.

The second OLS model with 33 features had a coefficient of determination, R squared, of .98-.99 indicating that these 33 features can predict 98-99% of the variance of the target variable, wins. To further evaluate the regression model, the Root Mean Square Error (RMSE) was calculated in order to measure how close the observed data points are to the model's predicted values. RMSE is defined as the square root of the variance of the residuals, the difference between predicted and observed values. Lower values of RMSE indicate a better regression fit of the data. In this study, the second OLS model had an RMSE for the test data of 1.90.

As will be discussed, a third OLS regression model using only 16 features as regressors was created that had a comparable RMSE for test data of 1.90.

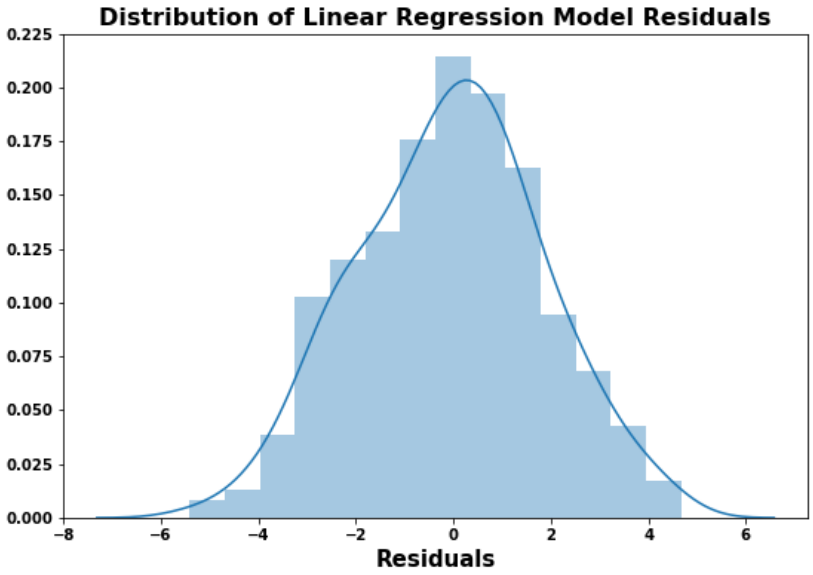
Predicted Vs. Actual Wins

A scatterplot of predicted wins by the OLS model and the observed wins is linear.



Error Analysis

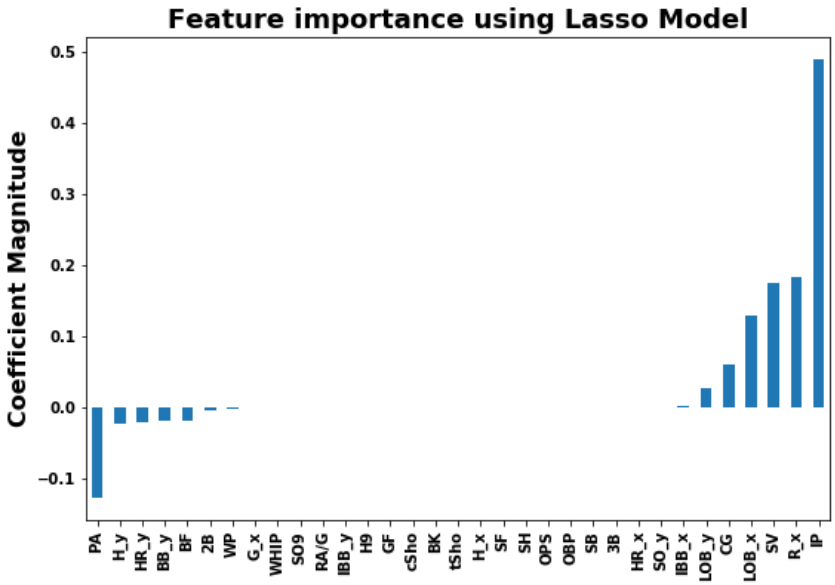
Normal distribution of errors is an assumption of linear regression. In this case, residuals, errors, have a normal distribution which is consistent with the assumption about the errors.



Feature Selection

In the initial analysis, 33 of 60 features in the dataset were used in a OLS regression model that had an RMSE of 1.90 for test data. For further optimization, the determination of which of these 33 features were most significant in predicting wins was also studied. A model with fewer features has the advantage of being computationally faster which is sometimes preferable.

LASSO (Least Absolute Shrinkage and Selection Operator) regularization was utilized for feature selection. If the feature is irrelevant, LASSO assigns a coefficient of 0 for the feature. So, features with non-zero coefficients are kept because they are considered important in predicting the dependent variable, and the features with coefficients of 0 are removed. Based on LASSO analysis, 16 of the original 33 regressors were selected as predictors for an optimized OLS regression model. This third model had a RMSE of 1.90 for test data and a R squared of 0.99.



LASSO Feature Selection

5. Conclusion

Key Findings

The original goal of this project is to see whether it is possible to predict a baseball team's wins with a regression model using a combination of batting and pitching statistics recorded over a period of more than a 100 years. In this study, a baseball team's wins can indeed be predicted by a combination of batting and pitching statistics.

In summary, a linear regression OLS model was constructed using16 features, regressors, that were considered significant by LASSO. And, this optimized OLS model was able to predict wins with an RMSE of ~1.90 for test data that is comparable to an earlier model with more features. This optimized model is also computationally faster than the model with more features. In this optimized model, IP,SV, PA, LOB_x, and R_x were among the 16 predictors used.

IP, innings pitched, is a metric that measures how many innings a pitcher remains in the game. Good pitchers have higher IP numbers than bad pitchers.

R_x, runs scored by hitting. As expected, the more runs a team scores the more games the team wins.

SV, save, is defined as the number of winning games credited to the team's relief pitchers who pitch at least three innings. When a team's relief pitchers perform well, a team wins more games. However, saves are not possible unless the team's starting and middle relief pitchers

also perform well in the game. Therefore, SV is a good barometer for a team's overall pitching performance.

LOB_x is a metric that measures the number of runners left on base when the team is hitting. Unexpectedly, the LOB_x has a weak positive correlation with wins. Naively, I assumed that having a high LOB_x would be negatively correlated with wins since leaving men on base would seem to prevent runs from being scored. However, on second thought, a higher LOB_x indicates that a team is at least able to get men on base which is often necessary to score runs. In contrast, a low LOB_x could mean that a team has difficulty getting men on base. It should be noted that the exclusion of LOB_x as a regressor resulted in a significant and large increase in the error, RMSE, of the regression model. Therefore, LOB_x should definitely be included as a regressor to predict wins.

PA, plate appearances, is the number of times a player completes a turn batting and has a weak positive correlation with team wins. In this study, PA is for the entire team.

Several of the features in the model are correlated with one another indicating multicollinearity. However, multicollinearity does not seem to affect the overall fit of the model as evidenced by the high R squared value of 99%, and it may not affect prediction of wins.

As expected, the ability to produce runs scored and good pitching are important for winning in baseball.

Next Steps

It would have been nice to see if factors such as team salary or average salary per player on a team are correlated with number of wins. In other words, do teams that spend more money for their players obtain more wins than teams who spend less? Unfortunately, such data was not easily found especially for baseball teams that played a 100 years ago.