Using Regression Modeling to Predict Single Season Total Winning Percentage for Baseball Teams

Shannon Leiss

Introduction

Motivation/ Background

Baseball is a game that has been played professionally in the United States since 1869 and has many different elements that contribute to a team's overall success. These different elements can be grouped broadly into offensive elements and defensive elements. Since the introduction of Sabermetrics (Birnbaum) and famous prediction formulas - introduced by Bill James in 1980 - many different attempts have been made to optimize a team's performance in these different areas using statistical modeling. With the abundance of data available, dating back to the 1871 season, many different aspects of the game are able to be explored in depth. Unlike many other professional sports, baseball is not played in halves or quarters that are a set amount of time; rather a game of baseball consists of 9 innings with each inning consisting of two halves - where one team is given the opportunity to be on offense each half inning, with the other team being on defense. These half innings are ended when the defensive team has gotten three outs on the offensive team. Another unique trait that baseball has is each team plays 150+ games per season, which is almost double the number of games in a single season for any other professional sports league. This lengthy season causes the starting players for each team to constantly change each game, making the overall success of a baseball team hard to pinpoint on a single player.

The following analysis will be performed on a compilation of data from two different sources: RetroSheet and the `Lahman` package within R. Dating back to 1871, RetroSheet has per game data from every single game played for every single season, up to the 2016 season. These include many offensive and defensive statistics that will be defined below, along with additional information about each game such as the home and away team, the name of the ballpark, attendance numbers, the starting players, the managers, umpires, and any other oddities for each game. The `Teams` dataset from the `Lahman` package provides each team's overall performance in a wide variety of variables per season, along with information on playoff appearances, finish within league and division, and attendance totals. For this analysis, the variables of interest will be most offensive and defensive statistics - defined below - as well as the season, franchise, and league identifiers.

Goals of Project

With the plethora of different offensive and defensive statistics that are available for every single inning of baseball played since 1871, many different attempts have been made to predict a team's success based on the available data. The main measure of a team's success is their overall winning percentage for any given season. Currently, the industry standard for predicting a team's overall winning percentage is the Pythagorean Expectation that was created by Bill James in the mid 1990's ("Pythagorean Theorem of Baseball", 2021), which uses a team's total runs scored and runs allowed per season to predict the team's expected winning percentage. The formula takes the following form:

$$E(Win\%) = \frac{RS^2}{RS^2 + RA^2}$$

Where RS stands for a team's overall runs scored and RA stands for a team's overall runs allowed. This formula has been proven effective for predicting a team's overall winning percentage within a ± 3 game range and has since been applied to other sports as a general framework for predicting success (Albert, 2019). However, this model does not account for some major factors that impact a team's success - such as previous season winning percentage and the different offensive and defensive statistics that impact a team's run differential.

While the score of a baseball game is what determines which team wins and loses, it is the team's batting, pitching, and defensive skills that will determine which team scores the most runs. Since the team that has the most runs after 9 innings will win - no matter the score differential - predicting a team's expected winning percentage only based on runs scored and runs allowed will cause the Pythagorean model to predict poorly when a team wins all of their games by a wide margin and loses all games by a small margin. Using the different per game defensive and offensive statistics - broken into home games and away games for each given team - regression modeling will be used to create a model that predicts a team's winning percentage more accurately than the Pythagorean model along with a simplified model that incorporates a team's fielding percentage (FP) along with some offensive per game statistic that predicts better or similarly to the established Pythagorean method.

Data Cleaning

Since the two data sets are from different origins, for the data to be in a proper form for analysis, the sets were cleaned and then linked together through their season and team. The game-by-game data that was found in the RetroSheet set was first grouped by team and season - using their modern franchise identifier, as many teams have moved or changed names since 1871 - to find the total number of home game wins and away game wins, along with the number of losses at home and on the road. For these home and road splits, all offensive and defensive statistics were summed to find the total number of that particular statistic for the team at home and on the road. For example, the total number of runs scored by a single team was broken into the number of runs scored at home games and the number of runs scored during away games. Although modern baseball seasons consist of 162 games, Major League Baseball (MLB) has not always played that many games, so to account for the different number of total games across the years, per game statistics were then found for all of the above mentioned totals. This allowed for the values for teams in the 1920's to be comparable to those in the 2000's.

After the RetroSheet data were grouped into single seasons per team, the two data sets were able to be joined by the season and franchise identification. In total - using the 1920 season through the 2016 season - 2,112 seasonal statistics were used. Most of the data for the following analysis comes from the manipulation of the RetroSheet data, with a few exceptions of variables that were found in the `Teams` data set and will be defined below.

Definition of Baseball Terms

Throughout this paper, many different abbreviations of baseball statistics will be used, along with different mutations of these statistics. The following list will define these terms and abbreviations, along with any formula that was used to create these variables:

- **Total Win Percentage** is the number of games won divided by the number of games played in a given season for a given team
- **Home Win Percentage** is the number of games won at home divided by the total number of games played at home
- **Previous Season Win Percentage** is the team's previous season winning percentage
- **Previous Home Win Percentage** is the team's previous season winning percentage at home
- **FP** is the team's fielding percentage. This is the total number of putouts and assists by a team while in defense divided by the total number of opportunities - which is the sum of putouts, assists and errors by a team ("Fielding Percentage (FPCT): Glossary"). These values are for a whole team for a single season.
- All variables with the suffix "**per_HG**" are the average number of a specific offensive or defensive statistic that a team achieved during each home game. The variables with the suffix "**per_AG**" are the average number of a specific offensive or defensive statistic that a team achieved during each away game during a particular season.
- All variables can also be grouped into offensive and defensive statistics. These follow as:

Offensive Abbreviation	Description	Defensive Abbreviation	Description
H	Hits Batted in	**HA**	Hits Allowed
HR	Homeruns Batted in	**HRA**	Homeruns Allowed
RS	Runs Scored	**RA**	Runs Allowed
SO	Strikeouts By Batters	**SOP**	Strikeouts by Pitchers
BB	Walks By Batters	**BBP**	Walks allowed by Pitchers
RNS	Runners Stranded	**RNSP**	Runners Stranded by Pitchers
X3B	Triples Batted in	**X3P**	Triples Allowed by Pitchers

- The variable of **RNS** means the number of batters that reached base by getting walked or getting a hit and being left on base when the inning is ended. This variable can be used as a measure of possible runs that were not scored by a team. In contrast, **RNSP** means the number of batters that the pitcher allowed on base through walking them or giving up a hit, but leaving them on base and not allowing that runner to score.

- All of the above variables were also used to create differential statistics. These are the difference between the corresponding offensive and defensive statistics that match for a team. For example: **Hits_Dif_HG** is the difference between the average number of hits batted in and the average number of hits allowed by the pitcher per home game.

Methods

Overview of Procedure

To find the best prediction models, linear regression modeling will be used to create the possible models. Since the data set has over 2,000 observations, the models will be created and compared using the records of a single team. The basic matrix form of a linear model is:

$$\mathbf{Y} = \beta\mathbf{X} + \epsilon$$

Where $\mathbf{X}_{n\times p}$ is defined as the design matrix of predictors, $\beta_{p\times 1}$ is the matrix of parameters, $\epsilon_{n\times 1}$ is the error matrix which follows the properties of $\epsilon_i \sim^{iid} N(0, \sigma^2)$, and $\mathbf{Y}_{n\times 1}$ is the matrix of response variables. For the matrices, n is the number of observations in the dataset and p is the number of parameters in the model - with the matrices using only p since all models will not have an intercept. The linear least squares estimate of β can be defined as follows - given that $\mathbf{X}^T\mathbf{X}$ is invertible (Faraway, 2016):

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})$$

This model has assumptions that should be met to ensure the accuracy of the model. These assumptions, which will be assessed on the models chosen to investigate, are that the explanatory variables have a linear relationship with the response variable, all observations are independent of each other, the errors of the model are normally distributed, and that the errors have constant variance.

The structure of the data does cause a problem with multicollinearity which occurs when at least one of the explanatory variables is almost a linear combination of the other explanatory variables, meaning that it is highly correlated with the other explanatory variables. Multicollinearity can also cause the estimates of β to have higher standard error (Faraway, 2016). This is an issue with the data as all offensive statistics are highly correlated with one another - especially runs, hits, and homeruns - however; since the main point of this project is find the model that predicts the expected total winning percentage the best rather than inference on the models, this multicollinearity issue will be ignored as it does not explicitly violate the regression assumptions above. It can also be noted that the success of a team in a particular season can be associated with that team's success in the previous season. To account for this, the team's previous total winning percentage will be included in some models.

Once a group of linear models has been decided, bootstrapping will be used to find the optimal values of $\hat{\beta}_i$ terms in the models. The bootstrapping process will use a single team's records and create a training set and a test set of seasons - with 80% of the seasons occurring in the training set and 20% of the seasons occurring in the test set. The training set will then be used to fit a number of different linear models that were predetermined outside of the simulation. For each run of the simulation, the coefficient values and which of the terms in the model were significant on the $\alpha = 0.05$ level will be stored, along with the following model evaluation criteria: AIC, BIC, and adjusted R^2.

These criteria will be used to compare all the models simulated to see which of the models fit the data best for each training set of data. Along with the model coefficients and the model comparison criterion, the test set of seasons will also be used to create predictions with each model and calculate the absolute mean error (AME). The AME for any given model can be found as:

$$AME = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y_i}|$$

Where n is the number of test seasons, Y_i is the actual season total winning percentage and $\hat{Y_i}$ is the predicted winning percentage. AME was chosen for the data as the response variable ranges from 0 to 1 and the effects of outlier seasons can be reduced by not squaring the errors. The AME for each model being simulated, along with the Pythagorean Expectation model on the test set will be recorded and will also be used to compare the models. These AME comparisons will be used to decide on the final models used as the goal is to be able to predict better - have a smaller AME - than the Pythagorean method. Once all the above terms are found and computed, the bootstrapping process will re-sample the original data set to get a new training and test set and repeat the above process. Due to the size of the data, this simulation will be run 500 times.

To combat the multicollinearity issue, the same bootstrapping technique will also be employed for two different kinds of penalized regression: Ridge and Lasso.

- Ridge Regression standardizes the explanatory variables and adds λ to the diagonal of $\mathbf{X}^T\mathbf{X}$ to create a "ridge" to help remove multicollinearity and put everything on the same unit level. This is also useful when there is a large number of explanatory variables that all have some effect on the response variable. The estimates of the $\hat{\beta}$ terms differ from the equation for the $\hat{\beta_{LS}}$ used in the linear models, as it takes the following form:

$$\hat{\beta}_{ridge} = (X^TX + \lambda I)^{-1}X^Ty$$

Where ridge regression will have the same $\hat{\beta}$ terms as the linear least squares estimate when $\lambda = 0$ (Faraway, 2016).

- Lasso Regression is another form of penalized regression, but unlike ridge regression, it is best used when it is believed that only a small number of the predictors contribute to the response variable. While Ridge regression will not remove any predictor variables from the model, Lasso regression will zero out any predictor variables that are deemed to not contribute enough to the response variable.

These penalized regression models will go through a similar bootstrapping process as the linear regression models, with the training and test years being recorded. The model coefficients of both penalized regression will be recorded with their corresponding lambda values that were found through cross validation - which is a process that chooses the smallest value of λ that stabilizes the estimates of β - and the AME values on the test set of seasons for both the ridge and lasso models. The explanatory variables that will be used for these models will be the variables that were chosen using the simulation of linear models.

Once two finals models are found - a complex model and a simpler model - through comparison of the model outputs from the simulation, an additional simulation will be run with the two models to find the optimal coefficient values for each model. In this case, the optimal coefficients will be those that minimize the model's AME when fit on all teams for all seasons between 1920-2016. The performance of these models will be assessed through simulating each model 100 times using test and training sets - similar to the bootstrap process used to decide the models - with each simulated model fitted using the training sets to predict the expected total winning percentage of each individual team for all seasons. This will create 100 unique predicted values for a single team's single season total winning percentage for each model.

Procedure of Selecting/Fitting Models

The dataset being used to create the models for consideration are from the 1920 - 2016 seasons. These seasons were chosen as the 1920 season was the first season that was not significantly impacted by the World War I drafts that also has reliable records for all desired statistics. As this dataset consists of over 2,000 team/season records, the records for the Chicago Cubs will be used to build the models. The Chicago Cubs were chosen as the team to build the model off of as they have been in the same league and at the same ballpark for all seasons in the dataset.

To create a model that predicts better than the Pythagorean model, multiple models were created that included all available per game statistics split by home and away games, models focused on only home or away games, models focused on defensive/offensive statistics, along with a few models that used the differences between offensive and defensive statistics split by home and away.

To decide on the explanatory variables to be used in different "simpler" models, the correlation between Total Win Percentage and all possible variables was examined. Looking at the variables that had the highest correlation with Total Win Percentage, a few different versions of models were created to be ran through the bootstrap simulation as well. In total, 15 different models were simulated through the bootstrap process 500 times. Out of these 15 models, the following 5 models were considered for final model comparisons:

- **Model 1:** Predicting Total Win Percentage using the team's previous season total win percentage, the team's home win percentage, fielding percentage, along with the team's run, hit, homerun, strikeout, walk, triple, and runners stranded differentials per game - split into home game differentials and away game differentials. This model contained 18 explanatory variables.
- **Model 2:** Predicting Total Win Percentage using the team's previous season total win percentage, the team's home win percentage, the team's previous season home win percentage, fielding percentage, along with the teams run, hit, homerun, strikeout, walk, triple, and runners stranded differentials difference between home and away games.
- **Model 3:** Predicting Total Win Percentage using a team's overall fielding percentage and the team's run differential per game. This model contained 2 explanatory variables.
- **Model 4:** Predicting Total Win Percentage using a team's previous season total win percentage, fielding percentage, and the team's run differential during away games. This model contained 3 explanatory variables.
- **Model 5:** Predicting Total Win Percentage using a team's previous season total win percentage, fielding percentage, and the team's run differential during home games. This model contained 3 explanatory variables.

Comparison of Models

Using the five models that were defined above, an additional simulation was run to determine which of the more complicated models - Model 1 or Model 2 - was going to be used as the model that predicted consistently better than the Pythagorean and which of the simpler models - Model 3, Model 4, or Model 5 - would predict better or similar to the Pythagorean model. The first few rows of the model comparison criterion values from the simulation comparison tables are found below:

Table 2: AIC values for all 5 models fitted on the Cubs training seasons

Simulation	Model 1	Model 2	Model 3	Model 4	Model 5
1	**-396.078**	-376.184	-345.567	-282.172	-276.980
2	**-391.812**	-376.543	-346.906	-292.621	-286.723
3	**-388.984**	-361.243	-353.764	-288.818	-267.238
4	**-385.018**	-358.397	-346.548	-285.928	-279.726
5	**-386.840**	-365.945	-344.097	-292.095	-276.385

Table 3: BIC values for all 5 models fitted on the Cubs training seasons

Simulation	Model 1	Model 2	Model 3	Model 4	Model 5
1	**-351.546**	-348.058	-338.536	-272.797	-267.605
2	-347.280	**-348.417**	-339.875	-283.246	-277.348

Simulation	Model 1	Model 2	Model 3	Model 4	Model 5
3	-344.452	-333.117	**-346.733**	-279.442	-257.863
4	**-340.486**	-330.272	-339.517	-276.553	-270.350
5	**-342.308**	-337.819	-337.066	-282.720	-267.010

Below is a table of summary statistics from the model simulation that contains the AME for all 5 models and the Pythagorean model, along with the percentage of times - out of the 500 simulation runs - that each model was found to have the lowest AME, lowest AIC and lowest BIC of all models being compared.

Table 4: Summary table comparing all 5 models and the Pythagorean model

	AME	Lowest AME	Lowest AIC	Lowest BIC
Model 1	0.0156513	79.2	100	93
Model 2	0.0180447	16.0	0	5.2
Model 3	0.0213956	3.4	0	1.8
Model 4	0.0302909	0.0	0	0
Model 5	0.0308722	0.0	0	0
Pythagorean	0.0220278	1.4		

It can be noted that Model 1 has the lowest AME overall and had the highest percentage of times that it was noted to have the lowest AME, BIC, and AIC of all the models when fit on the training/test seasons for the Cubs within the simulation. Out of the five models, Model 1, Model 2, and Model 3 all have lower overall AME values than the Pythagorean Expectation when fit on the same seasons. The AME values for each simulation run can be seen in comparison to the other models and the Pythagorean model in the following scatterplot.

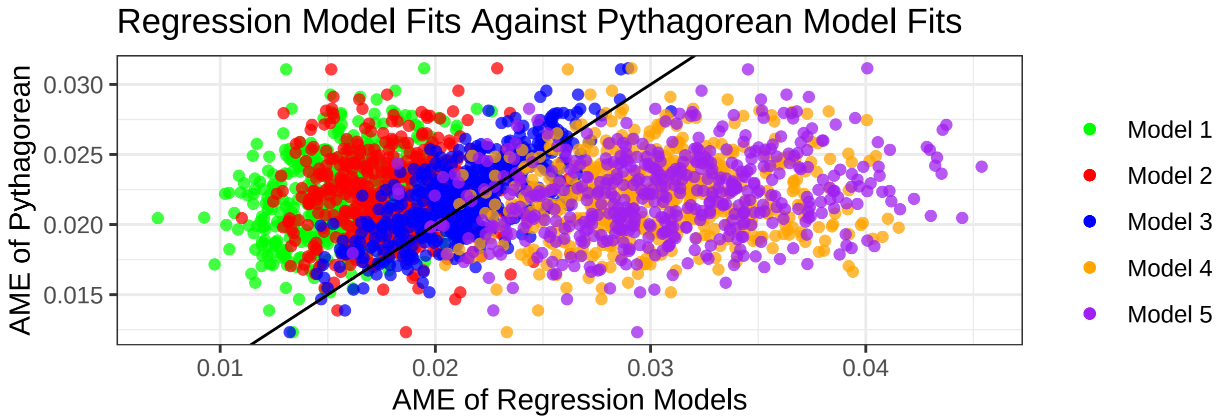

Figure 1: Linear Model AME values on Cubs test seasons compared to the Pythagorean Expectation AME for the test seasons

From the above graph, it is seen that Model 1 consistently has a predicted AME value less than the Pythagorean Expectation for the same test seasons. It also appears that Model 3 predicts better than the Pythagorean model for about 50% of the test seasons, while both Model 4 and Model 5 rarely predict better than the Pythagorean model. These observations are supported by the fact that for all 500 test sets, Model 1 had an AME lower than the Pythagorean method 97% of the time while Model 3 had an AME lower than the Pythagorean AME 63.2% of the time. It can also be noted that Model 2 had a lower AME than the Pythagorean method 86.6% of the time, while Models 4 and 5 both had AME values less than the Pythagorean method for only around 4% of the test season splits.

Final Models

From the above comparisons it was found that the more complex model that fit the data the best was Model 1 - which consisted of using a team's previous season winning percentage, fielding percentage, and per game differentials split between home and away games - and the simple model that predicted about the same or better than the Pythagorean model was Model 3 - which used a team's overall fielding percentage and their run differential per game to predict total winning percentage. To further examine these models and to find the optimal coefficient values that minimized the AME of predictions, a new simulation was run with these two models. Training and test seasons were used to examine the possible variability in the coefficient values, depending on the chosen seasons, as well as to combat any possible over parameterization. This new simulation broke the Cubs seasons into 100 training and tests sets and used these training sets to fit each model. These models were then used to find the predicted winning percentage for all teams in all seasons for 1920-2016, recording the coefficient values that were used to find these predicted values. From this simulation, each season for each team was predicted using 100 different versions of Model 1 and Model 3. For example, the 1935 New York Yankee's total winning percentage was predicted using 100 different variations of Model 1 and 100 different variations of Model 3.

Table 5: Example total winning percentage predictions from simulation, predicted using Model 1, Model 3, and the Pythagorean Expectation

Team	Season	Total Win %	Model 1 Fit	Model 3 Fit	Pythagorean Fit	Model 1 Residual	Model 3 Residual	Pythagorean Residual
SFG	2016	0.5370	0.5286	0.5492	0.5622	0.0084	-0.0122	-0.0251
STL	2016	0.5309	0.4977	0.5365	0.5448	0.0331	-0.0057	-0.0140
TBD	2016	0.4198	0.4500	0.4723	0.4704	-0.0303	-0.0526	-0.0507
TEX	2016	0.5864	0.5443	0.5017	0.5053	0.0421	0.0847	0.0812
TOR	2016	0.5494	0.5414	0.5536	0.5650	0.0080	-0.0042	-0.0156
WSN	2016	0.5864	0.5974	0.5894	0.6085	-0.0110	-0.0030	-0.0221

Grouping the predictions by simulation run, it was found that the only simulation run that had both Model 1 and Model 3 AME less than the Pythagorean method AME was simulation run 11, when looking at the predicted total winning percentage for all teams in all seasons between 1920 - 2016. This simulation run also had the lowest AME for both Model 1 and Model 3, with Model 1 predicting a team's total win percentage more accurately than the Pythagorean model 62.12% of the time and Model 3 predicting better than the Pythagorean model 49.76% of the time.

As these models are linear models, the assumptions of a linear model - outlined above - need to be checked for the chosen simulation run of the models. For Model 1, it was seen through examination of the pairs plot that all of the explanatory variables had a linear relationship with the total win percentage, the constant variance assumption was met through the plot of the residuals from the test sets plotted against their corresponding fitted values, the predicted total win percentage for a single team in a single season can be assumed to be independent of other teams, and the normal probability plot of the residuals shows that the residuals are approximately normal. Assessing the same conditions on Model 3 also found that all of these model assumptions were met.

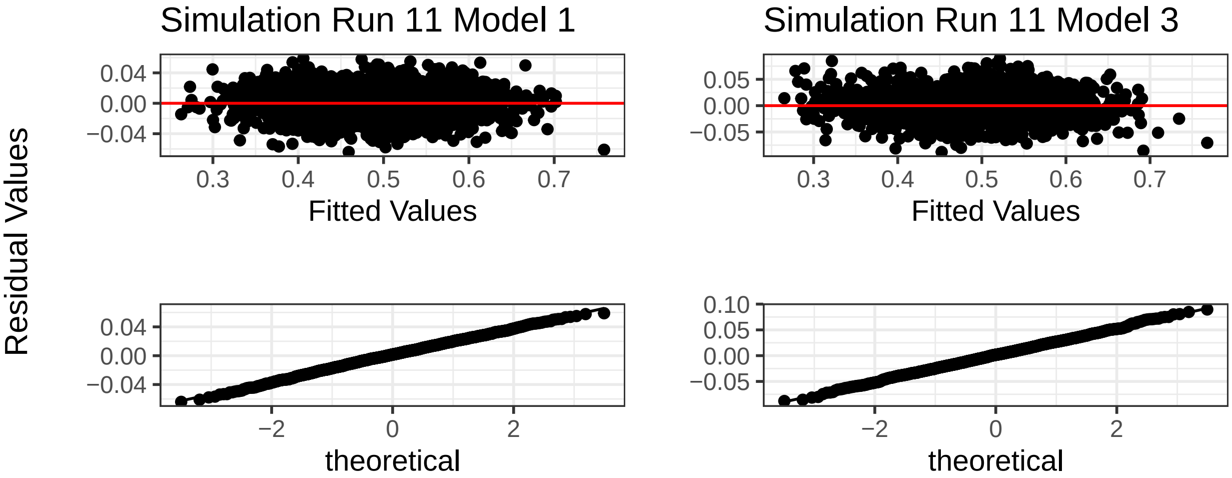

Figure 2: Residuals vs Fitted values plots and Normal QQ-plots for assessing model assumptions for Model 1 and Model 3

Even though the optimal coefficients for both models were produced in simulation run 11, each model's coefficient terms can also be looked at as the averaging of the coefficients over all 100 simulation runs. The following tables compare the coefficient values for Model 1 that were produced in simulation run 11 and the coefficient values found by averaging the value of the coefficients through all 100 simulation runs.

Table 6: Coefficient Values for Model 1

	Simulation 11	Average
Home_Win_Per	0.5853348	0.5693940
Previous_Season_Total_Win_Per	-0.0469708	-0.0111603
Previous_Home_Win_Per	0.0218164	-0.0235038

	Simulation 11	Average
FP	0.2100444	0.2244103
Run_Dif_HG	-0.0045270	-0.0009043
Run_Dif_AG	0.0417995	0.0426175
Hit_Dif_HG	-0.0040903	-0.0096557
Hit_Dif_AG	0.0050781	0.0051570
HR_Dif_HG	-0.0094223	-0.0047992
HR_Dif_AG	0.0139833	0.0005739
SO_Dif_HG	-0.0013648	-0.0006252
SO_Dif_AG	0.0053648	0.0056080
BB_Dif_HG	-0.0054235	-0.0082724
BB_Dif_AG	0.0054268	0.0094120
RNS_Dif_HG	-0.0006435	-0.0037466
RNS_Dif_AG	0.0030601	0.0063119
Trip_Dif_HG	0.0069812	0.0061852
Trip_Dif_AG	0.0179530	0.0186004

It can be noted that the effect of previous season home win percentage is opposite for the simulation run 11 than it is for the averaging of all the coefficient terms. Other than that switch of signs, most of the predictors have similar effects for both. Producing a similar table for Model 3 returns the following:

Table 7: Coefficient Values for Model 3

	Simulation 11	Average
FP	0.5091103	0.5064585
Run Dif Per Game	0.0993291	0.0971794

The effects of predictor variables are similar for both the optimal model and the model created by averaging all models. The distribution of prediction accuracy of these versions of the two models are as follows:

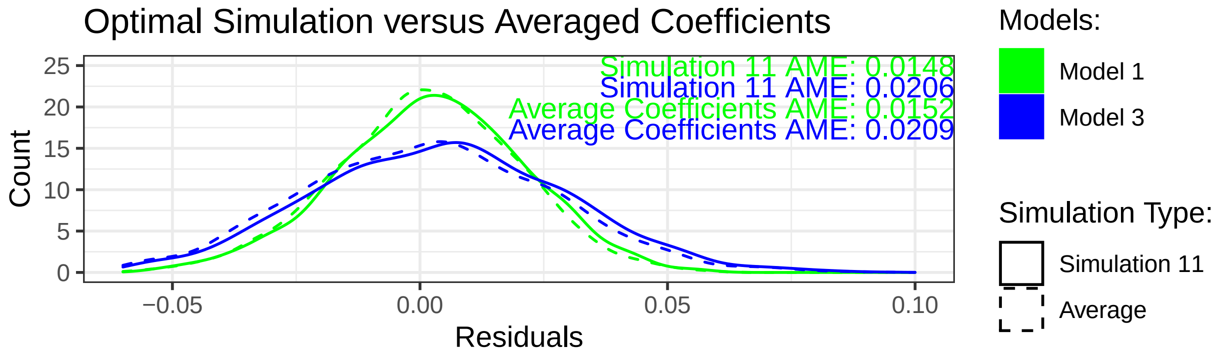

Figure 3: Distribution of residual values for the predicted total winning percentage of all teams from the 1920-2016 seasons for Optimal Simulation run and the Averaged Coefficient values

For both Model 1 and Model 3, the optimal coefficient values that were found during simulation run 11 were found to have a slightly lower AME, when used to predict the winning percentage of all teams over all seasons from 1920-2016. It follows that the complex model that is able to predict a team's expected winning percentage better than the Pythagorean model will have the coefficient values found from simulation run 11 - which can be seen in the above table for Model 1 - and the coefficient values for Model 3 that will predict the best are those produced by simulation run 11 and can also be found in the tables above.

Results

Ridge and Lasso Regression

Below is a small sample of the ridge regression and lasso regression's output that were created using the same bootstrapping method used for the linear models, with the AME values in this table being found by fitting the simulated models on the test seasons for the Chicago Cubs.

Table 8: Lambda values for Ridge and Lasso regression found using Cubs training seasons, along with Linear, Ridge and Lasso regression AME for predictions on Cubs test seasons compared to the Pythagorean Expectation AME

Model	Ridge Lambda	Lasso Lambda	Linear AME	Ridge AME	Lasso AME	Pythagorean AME
1	0.00614	0.00216	0.01544	0.01647	0.01441	0.02161
1	0.00584	0.00393	0.01636	0.01776	0.01347	0.02271
1	0.00571	0.00265	0.01333	0.01693	0.01293	0.02827
3	0.00685	0.00349	0.02162	0.02470	0.02405	0.02542
3	0.00645	0.00188	0.01956	0.01621	0.01465	0.01779
3	0.00654	0.00030	0.02045	0.02003	0.01994	0.02160

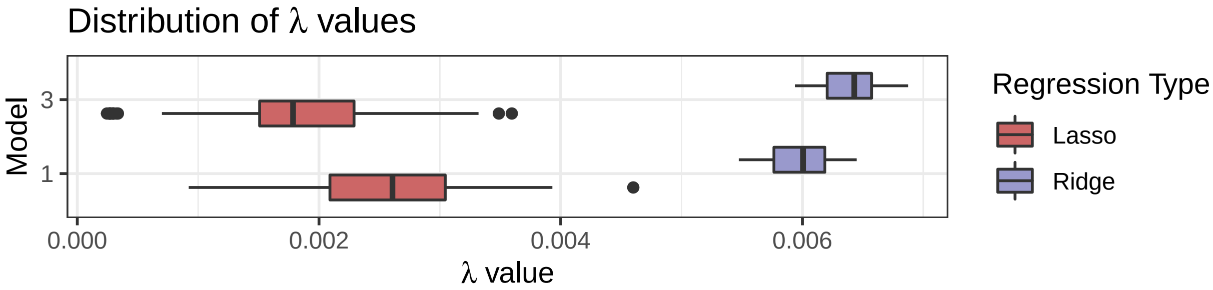

Figure 4: Distribution of lamba values for Ridge and Lasso regression

It can be noted from the graph above that all of the λ values are fairly small, near zero, and the ridge and lasso models are predicting at a very similar rate to the created linear models. This fact can be seen in the following table that compares the overall AME for ridge and lasso versions of both models - using the same predictors - to the linear model AME and the Pythagorean model AME.

Table 9: Overall AME for Cubs test seasons using Ridge, Lasso, and Linear regression compared to the overall AME for the Pythagorean Expectation

Model	Ridge	Lasso	Linear	Pythagorean
1	0.01643	0.01423	0.01569	0.02186
3	0.02189	0.02120	0.02135	0.02186

Since the values for λ are so close to zero, the $\hat{\beta}$ terms produced by the Ridge and Lasso regression do not differ substantially from the least squares estimates produced from the linear models. Along with small λ values, it makes sense that both Ridge and Lasso regression tend to predict similarly to the linear models - with Lasso regression predicting slightly better and Ridge regression predicting slightly worse than the linear regression. Due to this similarity and the fact that the $\hat{\beta}$ terms produced using Ridge and Lasso regression are biased, it can be concluded that the linear models should be used for predicting a team's total winning percentage.

Model Predictions and Pythagorean Comparison

Using all 211,200 unique predictions for a team's expected total winning percentage produced for each model, it was seen that the AME for Model 1 is 0.01575, the AME for Model 3 is 0.02099, and the AME for the Pythagorean Expectation is 0.02067. For these models, it was also seen that Model 1 predicted for an individual team and season combination better than the Pythagorean Expectation (Model 1's AME was lower than Pythagorean) 59.216% of the time, while Model 3 predicted better than the Pythagorean Expectation 48.244% of the time.

While the overall performance of the models shows that Model 1 has a lower AME than the Pythagorean method and Model 3 has a similar AME to the Pythagorean method, the behavior of the models changes when looked at on different levels of the data.

- Model Performance for **Individual Teams**: For the 30 current teams in Major League Baseball, Model 1 predicted the Arizona Diamondbacks the best - with an AME rate of only 1.23% compared to the Pythagorean AME rate of 2.2%. The AME rates for all teams individually using Model 1 ranged from 1.23% - 1.89%. For the simpler Model 3, the Tampa Bay Rays had the lowest AME rate at 1.75%; however this was higher than the teams Pythagorean AME rate of 1.74%. Out of the 30 teams, 11 of the teams had Model 1 and Model 3 AMEs lower than the Pythagorean AME along with only 6 of these teams also having Model 1 and Model 3 predicting better than the Pythagorean rate for individual seasons more than 50% of the time. The only team that appears to have both Model 1 and Model 3 predicting at a worse rate than the Pythagorean model is the Miami Marlins.

- Model Performance for Specific **Decades**: For the 10 decades that were included in this data set, Model 1 and Model 3 were found to fit modern decades the best - with 2010's, 2000's, and 1990's having the lowest AMEs.

- Model Performance between **Leagues**: While Model 1 predicts the National League slightly better than the American League, Model 3 for both leagues predicts worse than the Pythagorean model. When the leagues are broken into decades as well, it can be seen that the league decade grouping that was predicted the best by Model 3 was the 1920's American League, followed by the 1960's American League. This is in contrast to the modern era having the best fitting Model 3 when the decades were examined overall. It can also be seen that Model 1 predicts best for the 2010's and 2000's National League, which corresponds to the findings for decades overall. In 7 of the league decade combinations, Model 1 and Model 3 predict better overall than the Pythagorean model. These are the 1920's National League, the 1940's - 1970's American League, the 1970's National League, and the 2010's American League.

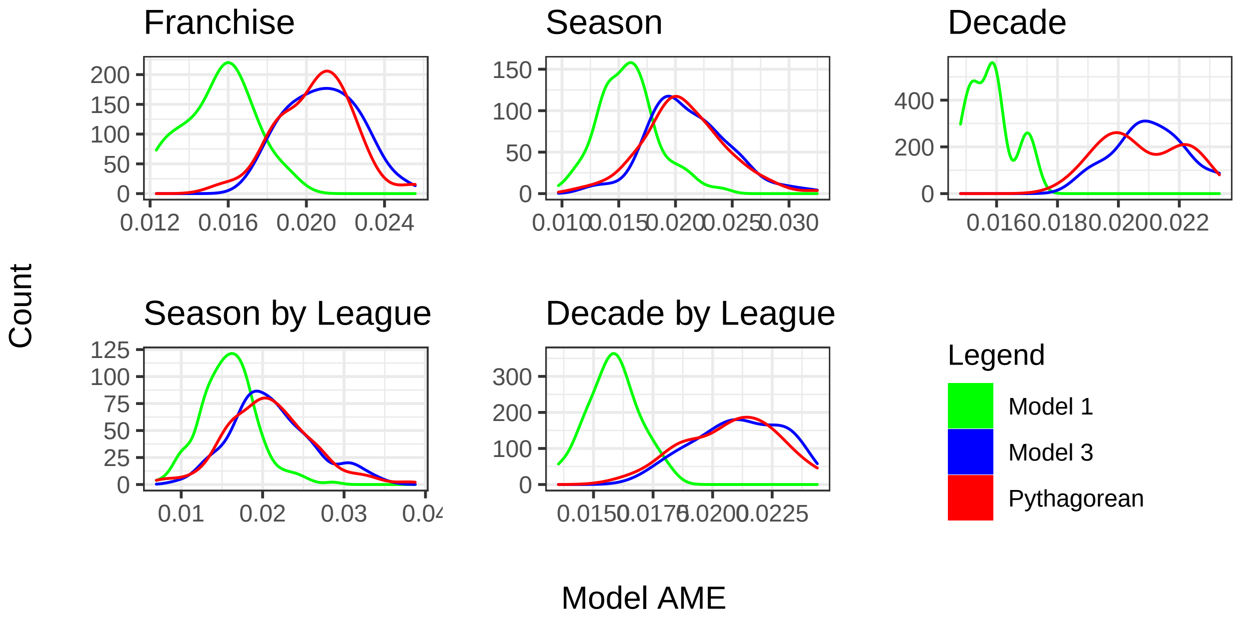

Figure 5: Distribution of AME values for predictions from 1920-2016 season, when predictions are grouped by specific criteria

Further Work and Applications

Baseball - as well as sports in general - have a large amount of randomness caused by the players that is hard to account for in any models that are created. However, further examination of the models that have been presented above could give useful insight into a team's expected winning percentage for a season, given some offensive and defensive statistics. These models could be useful for managers and executives to figure out what areas of the team need to be improved to reach a specific expected winning percentage. For example, if a team is half way through the season with a per game run differential of +0.78, they would be able to figure out how much they would need to improve their team's overall fielding percentage to have an expected winning percentage of 55% using Model 3. This breakdown of expected winning percentage into per game differentials - both offensive and defensive - is a major advantage to the models presented above over the Pythagorean Expectation.

While the above models have been shown to predict better than the industry standard model in many cases, there is still much work that could be done to improve these models and different variables that could be incorporated into the models that have not yet been examined. Further testing of these models could be done by making mid season predictions to test the accuracy compared to the team's actual winning percentage at the end of the season. This method could also be used to see how many games a team must play for the models to achieve an AME around or below 0.02, as a model that can accurately predict earlier in the season is a great advantage for teams to have.

It would also be interesting to build into the model a difficulty of schedule term. The difficulty of the season that each team plays depends on which division and league the team is in, as teams that are in the National League West will play more of their games against those in its division while possibly only playing teams outside of their leagues once or twice per season. This can majorly impact a team's winning percentage as their division may be more competitive than other divisions. This can normally be seen in the National League West and American League East as those divisions usually have power house teams with much more difficult schedules than some of the teams in the Central divisions. While both of the created models are baseball specific, these models could possibly be transferred into other sports, similarly to how the Pythagorean Expectation has different versions for different sports.

Originally, this project was intended to analyze if the sports assumption of 'Home Field Advantage' was present in baseball and if teams that performed better at home compared to away games would have a higher winning percentage compared to teams that performed at an average level for all games. However, due to the large amount of games played per season per team - 162 with an equal home and road split for all teams - the models that performed the best were those that incorporated both home and away statistics as well as the models that were based off of only per game statistics. This would be interesting to dive into further research on especially since - unlike other sports - baseball does not have a standard size for each field. Due to this, every single ballpark (30 in all) has different dimensions that cater to different skill sets. For example, the Colorado Rookies field is infamously a hitters park due to the distance from home plate to the outfield walls, along with the thinner air that is present in Denver, Colorado. On the other hand, Dodger stadium in Los Angeles is notoriously a pitchers park as the outfield walls are quite a ways from home plate. The only standard measurement that all ballparks have to follow is the distance between bases and the distance from the pitchers mound to home plate. This 'Home Field Advantage' would be an interesting predictor to explore and use in future work with this data.

References

[1] Albert, Jim, et al. *Analyzing Baseball Data with R.* CRC Press/Taylor and Francis Group, 2019.

[2] Birnbaum, Phil. "A Guide to Sabermetric Research." *Society for American Baseball Research*, https://sabr.org/sabermetrics.

[3] Faraway, Julian James. *Linear Models with R, Second Edition.* Taylor & Francis, 2016.

[4] "Fielding Percentage (FPCT): Glossary." *MLB.com*, Major League Baseball, https://www.mlb.com/glossary/standard-stats/fielding-percentage.

[5]Michael Friendly, Chris Dalzell, Martin Monkman and Dennis Murphy (2021). *Lahman: Sean 'Lahman' Baseball Database.* R package version 9.0-0. https://CRAN.R-project.org/package=Lahman

[6] "Pythagorean Theorem of Baseball." *Pythagorean Theorem of Baseball - BR Bullpen*, Baseball Reference, 4 June 2021, https://www.baseball-reference.com/bullpen/Pythagorean_Theorem_of_Baseball.

[7] RetroSheet. "Retrosheet's Datasets." *Data.world*, RetroSheet, 2 Nov. 2017, https://data.world/retrosheet.