**STAT 684 Final Project**

**Predicting MLB Playoff Appearances**

**By: Matt Goldsmith**

**Section 1: Abstract**

This research paper demonstrates a statistical model that uses linear and logistic regression to predict, with ~90% accuracy if a major league baseball (MLB) team will make the playoffs based on hitting and pitching statistics. This research also determined that the most important statistics in determining the outcome of the season are On Base Percentage (OBP) for hitting and Earned Run Average (ERA) for pitching. This model builds on Bill James’s famous Pythagorean Expected Wins formula but replaces it with a linear regression model that utilizes Runs Scored and Runs Allowed to predict Total Wins. Using this model, any MLB general manager (GM) can gain an advantage in team building and player evaluating over their competitors.

**Section 2: Statement of Research**

This research aims to accurately predict which teams will make the playoffs based on hitting and pitching statistics as well as which of those statistics are most important in determining the outcome of the season. Sustained success for a professional sports team is the only assured way to continually grow the fan base, which in turn leads to consistent ticket and merchandise sales. MLB GMs, especially the GMs of losing teams, are under pressure by ownership to succeed and ultimately give the owner an increased return on their investment. Winning the championship is the pinnacle of success, but the team with the best overall record rarely takes home the World Series trophy. Only 12 teams who won the World Series in the past 50 years had the overall best season record (West); therefore, we can assume there is often luck involved in winning in the playoffs and winning the World Series. MLB GMs can still find success in making the playoffs every year, because it would take a lot for an owner to fire a GM that consistently gets the team there year-after-year. Obviously getting a lot of wins throughout the 162 game season is going to give a team a good shot at making it, but to get wins, the team needs to consistently score more runs than the teams they play. Any MLB GM would appreciate a tool that could tell them if their team is likely to make the playoffs as well as which statistics should be considered with the most weight when building their team.

There have been many articles and studies published attempting to tackle these questions, but probably the most famous statistician to do so is Bill James in his book titled *Historical Baseball Abstract.* In his book, he described a formula for calculating the expected number of wins using only total runs scored and total runs allowed. This formula, the Pythagorean Expected Wins, was very controversial when it was first introduced. The movie *Moneyball,* based off of a book with the same name written by Michael Lewis, dives into the story of the 2002 Oakland A’s attempt to use Bill James’s core ideas to find discount players that gave them the best chance of winning games. The 2002 Oakland team, noticeably short on cash, defied the odds and the classic ideas of how to build an MLB team and clinched their division to make the playoffs. Unfortunately, this team did not win it all in 2002, but the manger, Billy Beane, reportedly received an enormous contract offer from the Boston Red Sox. This story always resonated with me and was the core inspiration for this research paper. Ultimately, the goal of this research paper is to improve upon Bill James’s core ideas by using linear and logistic regression to predict wins and playoff appearances which would assist all MLB managers in building a playoff caliber team.

**Section 3: Study Design**

This study design is more of an observational study because I did not intervene in the data to study some effect, but instead used MLB teams statistics from 1962 to 2012 to build a model that can predict whether a team makes the playoffs or not. The data used in this study is a combination of 3 unique datasets. Two of the datasets were fairly easy to find on Kaggle.com with the titles: “Moneyball MLB Statistics 1962-2012” and “MLB dataset 1870s-2016”. These two data sets are the most popular on Kaggle.com using the search term “MLB.” The third dataset, mostly pitching data, was scraped from Baseball-Reference.com using the r-package “rvest.” Getting the third dataset took more work than the previous two because I had to figure out how to utilize the “rvest” functions to pull the desired data from a webpage.

The final merged dataset, had approximately 90 total variables; however, many of the variables were giving the exact same information, e.g. Wins, Games Played, etc. Only 25 of the 90 variables were considered in the initial predictive models and all of those 25 variables were numeric data. Of these 25 variables, there were no observations with any missing data.

One of the biggest issues with the final dataset was that not all teams in all years played the MLB standard of 162 games. For example in 1962, the LA Dodgers and the SF Giants both played 165 regular season games because they were tied for first place at the end of the season and had to play a three game series to determine who would take home the National League Pennant. On the other end, the 1971 Baltimore Orioles only played 158 regular season games due to rained out games that were never made up. To combat this non-standard number of games played, I opted to only use statistics that were averaged per game (e.g. H9 = Hits per 9 innings pitched) or averaged per season (e.g. OBP = On Base Percentage). Some data needed to be altered to be either per game or per season. For example, Total Saves is one of the variables that was part of the dataset scrapped from Baseball-Reference.com but was converted to Saves per Game to account for the non-standard number of games played. Another example is Wins Adjusted, which was created by taking the teams Wins / Total Games (a.k.a. Win Percentage) and then multiplying the Win Percentage by 162.

Figure 3-1 below lays out all the variables in the final dataset and gives a brief description for each variable. Using the information in Figure 3-1, we can look at Figure 3-2, which provides a small sample of each variable, and understand better what each value represents.



Figure 3-1 Reference Table for Abbreviated Variable Labels

A picture containing text

Description automatically generated

Figure 3-2 Structure of Final Dataset

A lot of these variables were correlated, some were very highly correlated and thus some of the highly correlated variables were ultimately left out of the final models. The first level of models are two linear models: predicting Runs Allowed per Game and Runs Scored per Game. First let us look at the correlation among the initial predictor variables and response variables for both predictive models.

Figure 3-3 below shows the correlation between all the potential predictor variables and the response variable (RA.G). As we can see, ERA is highly correlated with the response variable but also with FIP, WHIP, and H9. Similarly, we can see that tSho.G and cSho.G are highly correlated with each other. Figure 3-4 below shows the correlation between all the potential predictor variables and the response variable (RS.G). Looking at Figure 3-4, we can see that OBP, SLG, and BA are all highly correlated with the response variables but are also highly correlated with each other.

Background pattern

Description automatically generated

Figure 3-3 Correlation Plot of Runs Allowed per Game and all the potential predictor variables.

Chart, bubble chart

Description automatically generated

Figure 3-4 Correlation Plot of Runs Scored per Game and all the potential predictor variables.

The initial models were fit with all the potential predictor variables listed above. Then a backward stepwise regression utilizing BIC as the selection criterion was run on both the Runs Allowed and Runs Scored linear models. BIC, Bayesian Information Criterion, penalizes a model based on the total number of predictor variables to avoid overfitting a model and a lower BIC score is considered a better model than a high BIC score. A backward stepwise function takes an initial model with all of the potential predictor variables and removes one variable at a time based on which variable would result in the lowest BIC. This process continues until removing any of the remaining variables would not improve the BIC score. Then to confirm there was not a collinearity issue, the VIF function was run on each model. VIF, Variance Inflation Factor, is a measure of collinearity; and collinearity is a measure of how correlated a variable is with the other variables in the model. If there were variables that had a VIF > 10, then the highest VIF variable was removed from the model first and then the VIF function was run again to get the updated collinearity values. This process was repeated until the only variables left in the model had a VIF < 10. Figure 3-5 and Figure 3-6 show the correlation plots of the final Runs Allowed and Runs Scored models, respectively.

Chart, bubble chart

Description automatically generated

Figure 3-5 Correlation Plot of Final Runs Allowed per Game model

Chart, bubble chart

Description automatically generated

Figure 3-6 Correlation Plot of Final Runs Scored per Game

**Section 4: Method**

The overall research question is can we accurately predict if an MLB team will make the playoffs using common baseball statistics and which baseball statistics are most important for a team to make the playoffs. The client for this study is an MLB organization looking to use a predictive model and understand how that model was developed.

The analysis performed in this study utilizes MLB team hitting and pitching data to first predict the expected Runs Scored per Game and expected Runs Allowed per Game, respectively. In this first level of predictive modeling, linear regression is utilized to predict expected Runs Scored per Game and expected Runs Allowed per Game with two separate models. Then the predicted Runs Scored and Runs Allowed per Game are used as the two predictor variables in the level two model, a linear regression model that predicts the expected Win Total. Lastly, the result of the level two model is used in the final model, a logistic regression model that uses the expected Win Total as a predictor for whether the team will make the playoffs. The first level models, Runs Scored per Game and Runs Allowed per Game, can also be used to determine which of the raw baseball hitting and pitching data are most significant in determining the Runs Scored per Game and Runs Allowed per Game.

In this study, I utilized linear regression and logistic regression and checked the assumptions that go with those models. For the linear regression models (Runs Allowed per Game, Runs Scored per Game, and Win Total), I plotted the residual and Q-Q plots to confirm that the assumptions of normality and homoscedasticity were met. Figure 4-1, Figure 4-2, and Figure 4-3 below show the residual and Q-Q plots for the Runs Allowed, Runs Scored, and Win Total models, respectively. Looking at the “Residuals vs Fitted” plots for all three models, we can conclude that there is adequate homoscedasticity because there does not appear to be a clear pattern in any of the plots. The “Normal Q-Q” plot is a gauge for the assumptions of normality. We can see that all the Q-Q plots below are fairly close to the diagonal line. The Runs Allowed per Game (Figure 4-1) has the most worrisome Q-Q plot, but ultimately it is sufficiently close to the line for the normality assumption to be met. As mentioned in Section 3, the VIF was used to test the collinearity in the model. Figure 4-4, Figure 4-5, and Figure 4-6 below show the VIF calculated for the Runs Allowed, Runs Scored, and Win Total models, respectively. As we can see, none of the variables have a VIF > 10 so we can conclude that the assumption of no collinearity is met for all three models. In all three models “Residuals Vs Leverage” plot, there does not appear to be any outliers in the models.

Chart, scatter chart

Description automatically generated

Figure 4-1 Runs Allowed per Game residual and Q-Q plots

Chart, scatter chart

Description automatically generated

Figure 4-2 Runs Scored per Game residual and Q-Q plots

Chart, scatter chart

Description automatically generated

Figure 4-3 Win Total residual and Q-Q plots



Figure 4-4 Runs Allowed per Game VIFs



Figure 4-5 Runs Scored per Game VIFs



Figure 4-6 Win Total VIFs

To check the assumption of a linear relationship between the predictor variables and the response variable, I plotted each predictor against the response. Figure 4-7 and 4-8 shows all 8 of the predictor variables against Runs Allowed per Game, Figure 4-9 shows the 2 predictor variables against Runs Scored per Game, and Figure 4-10 show Runs Allowed and Runs Scored against Win Total. For the Runs Allowed model, we can see that ERA has the strongest linear relationship with the response while Strikeouts per 9 Innings and Saves per Season appear to have a weak linear relationship with the response. In Figure 4-9, we can see that OBP has a very strong linear relationship with the Runs Scored while the Number of total Batters has a much weaker linear relationship with the response. In Figure 4-10, we can see that both Runs Scored and Runs Allowed have an equally strong linear relationship to Win Total. Although not all of the predictor variables show strong linear relationships, none of the variables show clear non-linear relationships, thus we can conclude that the linear relationship assumption is met.

Chart

Description automatically generated

Figure 4-7 Predictor Variables ERA, HR9, BB9, and SO9 against Runs Allowed per Game

A picture containing graphical user interface, chart

Description automatically generated

Figure 4-8 Predictor Variables ERAP, SV.G, CG.G, and tSho.G against Runs Allowed per Game

Chart, scatter chart

Description automatically generated

Figure 4-9 Predictor Variables OBP and Num.Bat against Runs Scored per Game

Chart, scatter chart

Description automatically generated

Figure 4-10 Predictor Variables RA.G and RS.G against Win Total

Now moving to the final model, a logistic regression model that used the results of the Win Total model as the predictor variable and the binary Playoff variable as the response. We need to check the assumptions of linearity and check if there are any influential points that could be outliers. Figure 4-11 shows the logit probabilities against the predictor variable Win Total. We can see that the linear assumption is met. In Figure 4-12, we can see the Cook’s distance and can confirm that there are no overly influential points. Cook’s distance is an estimate of how influential a data point is in the model.

Chart, scatter chart

Description automatically generated

Figure 4-11 Logistic Model Linearity Assumption Check

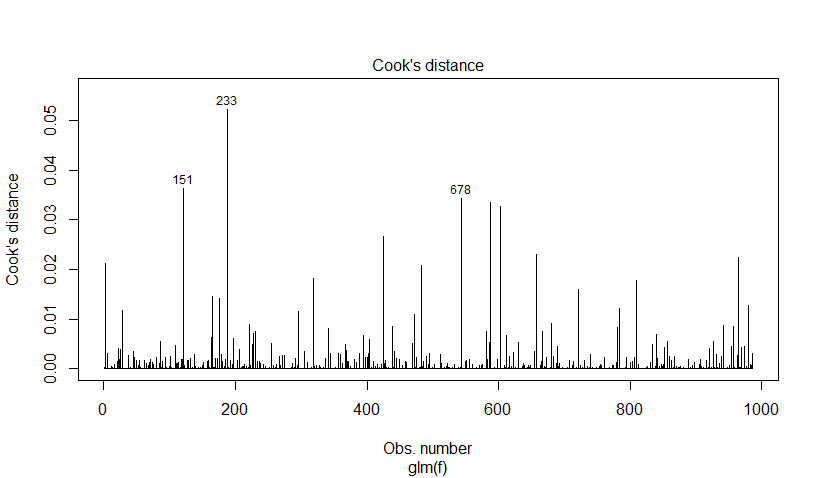


Figure 4-12 Cook’s Distance for Logistic Model

To confirm that these models were fit correctly, I used k-fold cross validation with k = 5. K-fold cross validation is where we create “K” different Train and Test set allocations of the data, and then train the model using the Train set and calculate the model’s accuracy using the Test set. We then take an average of the “K” accuracy percentages to get an overall accuracy percentage. By using cross validation, I was able to confirm that the models were fit well to the data and not overfit. The average accuracy rate {(True Yes + True No) / Total} comes out to 90.6% accuracy. One concern that someone might have about this method, is using the result of a regression model as a predictor variable for another regression model. If the first model is not fit correctly or has some inherent bias, it could negatively affect the subsequent models.

In the 1980s, Bill James published a simple formula, the Pythagorean Expected Wins, that could give an estimate to how many games a team should have won based solely on their total runs scored and total runs allowed (Caro & Machtmes 2013). This equation is . Interestingly, I decided to try running the k-fold cross validated model with this Pythagorean Expected Wins using the results from the Runs Allowed and Runs Scored models instead of the Win Total linear model. While the Pythagorean Expected Wins gives a win percentage, the Win Total gives a total win estimate out of 162 games, so to combat this difference I multiplied the Pythagorean Expected Wins by 162 before running it through the final logistic regression. While the Pythagorean Expected Wins is simple to calculate and understand, it under performs when compared to the linear regression model that regresses Runs Allowed and Runs Scored on the expected Win Total. Using the Pythagorean Expected Wins version resulted in an overall average accuracy of only 77.2% while the Win Total linear model gave us an overall average accuracy of 90.6%. For further information on the Pythagorean Expected Wins and its uses, I recommend reading Bill James’s *Historical Baseball Abstract*. Another good source is the *Testing The Utility Of The Pythagorean Expectation Formula On Division One College Football: An Examination And Comparison To The Morey Model* article by Caro and Machtmes that applies Bill James’ formula to division one college football.

**Section 5: Results**

If we look in detail at one of the five k-fold models, we can determine how important each individual statistic is in predicting if a team will make the playoffs. I am highly confident that the data used to build these models is accurate because the MLB has collected and recorded these statistics for well over one hundred years without any issues. Figure 5-1 is the summary of the Runs Allowed per Game model, Figure 5-2 is the summary of the Runs Scored per Game model, Figure 5-3 is the summary of the Win Total model, and Figure 5-4 is the summary of the Logistic Playoff model. As we can see in the 4 summary figures below, all of the predictor variables are statistically significant at an alpha=0.05; even more impressive is that all but 3 of the predictor variables are statistically significant at an alpha=0.001. The alpha value represents the probability that we will accidentally think a predictor variable is statistically significant when it really is not statistically significant (0.05 = 5%). The Runs Allowed linear model has a high adjusted r-squared of 0.9749. The Runs Scored linear model has a fairly high adjusted r-squared of 0.8122 and so does the Win Total linear model with an adjusted r-squared of 0.7683. The adjusted r-squared is a measure of how much of the data variation is explained in our model with a penalty term based on how many predictor variables were in the model. The residual standard error for the Runs Allowed and Runs Scored models are very small. The residual standard error for the Win Total model is significantly larger than the other two linear models; however, when you consider that the Win Total is predicting the total number of wins a team will get in a 162 game season, while the other two linear models are predicting how many runs will be scored by either team in a single game, the residual standard error of the Win Total model does not seem that large. The residual standard error is a measure of how close our actual data is to what the model predicts, a smaller residual standard error is better.

To determine which predictor variables are most important in the regression model, we should look at the absolute value of the “t value.” The “t value” is a measure of how significant a variable is in the prediction model, with a higher “t value” indicating higher variable importance. For the Runs Allowed model we can see that ERA has the highest “t value” by a significant margin and for the Runs Scored model we can see that OBP has the highest “t value” by a significant margin as well. In the Win Total model, both the Runs Scored per Game and Runs Allowed per Game have relatively close “t values,” but with Runs Allowed per Game being slightly larger. All three of these results agree with common sports knowledge about baseball. On offense, OBP is the most important statistic for scoring runs, while on defense ERA is the most important statistic for preventing runs. The phrase “Defense Wins Championships” is semi supported in the Win Total model because the Runs Allowed variable is slightly more important than the Runs Scored variable in predicting total wins.

Some of the coefficient signs (positive or negative) do not make intuitive sense in the Runs Allowed model. The HR9 variable has a negative coefficient which implies that as a team gives up more home runs per 9 innings, they will give up less runs per game. The other coefficient that does not make sense in Figure 5-1 is the CG.G, which has a positive coefficient. This implies that as a team has more pitchers throw complete games in a season, they will give up more runs per game. Logically this does not make sense because if a pitcher is giving up a lot of runs in a game, they are not likely going to pitch all 9 innings. Unsurprisingly, from Figure 4-4, the HR9 and CG.G variables have the second and third highest VIF value. We can then postulate that there is some collinearity affecting the coefficients of HR9 and CG.G.

Text

Description automatically generated

Figure 5-1 Runs Allowed per Game Linear Model Summary

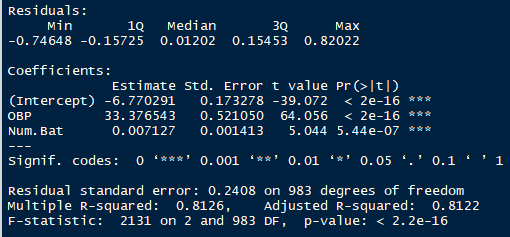


Figure 5-2 Runs Scored per Game Linear Model Summary

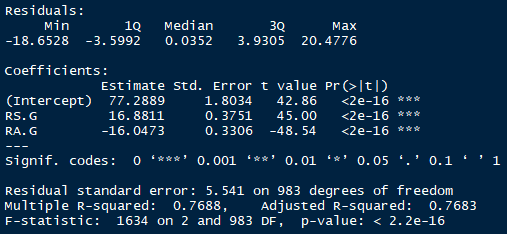


Figure 5-3 Win Total Linear Model Summary

Text

Description automatically generated

Figure 5-4 Playoffs Logistic Model Summary

Figure 5-5 below breaks out the accuracy {(True Yes + True No) / Total} percentages for each of the five k-fold models. The cutoff used for determining if the model predicted a playoff appearance from the logistic model was set at 0.5. The model summaries shown in Figure 5-1 through Figure 5-4 were created from the “Set 5” data. Taking a simple average of all five models’ performance, I got an overall accuracy of 90.6%.

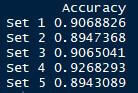


Figure 5-5 Overall Performance for five k-fold models

**Section 6: Conclusions**

Ultimately, this model shows that it is possible to predict with high accuracy if an MLB team will make the playoffs given common baseball statistics. The practical significance is that any MLB team who has fairly accurate estimates for what their team will produce on offense and defense, can use this model to correctly determine if their team is talented enough to make the playoffs 9 out of 10 times. Additionally, this model shows that the most important offensive and defensive statistic for predicting if a team will make the playoffs is OBP and ERA respectively. The next steps to improve this predictive process would be to create or improve existing models that aim to predict the individual statistics (e.g. OBP and ERA) based on previous performance by the team as a whole and the individuals that comprise the team.

**References**

1. West, Jenna. “How Many Teams Have Won the World Series After Having the Season’s Best Record?” *Sports Illustrated,* Meredith Corporation, 30 September 2018, https://www.si.com/mlb/2018/10/01/baseball-teams-best-record-win-world-series-red-sox.
2. Caro, Cary, and Machtmes, Ryan. “Testing The Utility Of The Pythagorean Expectation Formula On Division One College Football: An Examination And Comparison To The Morey Model.” *Journal of Business & Economics Research*, vol. 11, no. 12, 2013.