THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF STATISTICS

PREDICTING MAJOR LEAGUE BASEBALL PLAYOFF PROBABILITIES USING LOGISTIC REGRESSION

EVAN J. BITTNER
FALL 2015

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Statistics
with honors in Statistics

Reviewed and approved* by the following:

Andrew Wiesner
Lecturer of Statistics
Thesis Supervisor

Murali Haran
Associate Professor of Statistics
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

Major League Baseball teams are constantly assessing whether or not they think their teams will make the playoffs. Many sources publish playoff probabilities or odds throughout the season using advanced statistical methods. These methods are somewhat secretive and typically advanced and difficult to understand. The goal of this work is to determine a way to calculate playoff probabilities midseason that can easily be understood and applied. The goal is to develop a method and compare its predictive accuracy to the current methods published by statistical baseball sources such as Baseball Prospectus and Fangraphs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

- To my parents who supported me throughout college and my entire academic career

- To Dr. Wiesner, for always being available to offer valuable advice

- To Dr. Haran, for giving me the opportunity to do research and being a mentor

- To St. Basil the Great, St. Gregory the Theologian, & St. John Chrysostom, patron saints of learning

Chapter 1

Introduction

Every Major League Baseball (MLB) team starts its season off with spring training in February. There is one ultimate goal – win the World Series. With winning the World Series as the final goal, there is one important prerequisite – making the playoffs. The 162 game marathon of the Major League Baseball season determines which select teams, no more than one third of organizations, have the opportunity to play playoff baseball and fight for a World Series victory. Everything that occurs in a Major League Baseball organization is with making the playoffs in mind. Managerial decisions such as resting players, making lineups, and organizing pitching rotations and management decisions such as call-ups, send downs, and trades are often based upon their team's chance of making the playoffs. An accurate and interpretable chance, or probability, of making the playoffs is essential to organizations' decision-making process.

How does one quantify the probability a team has of making the playoffs? Analysts and media members often reference how many games teams are behind division and wild card leaders, and whether or not they think it is possible or plausible a team will end up making the playoffs. These are educated guesses based off experience or prior knowledge. Popular media sources such as ESPN, Fangraphs, and Baseball Prospectus all publish playoff odds or probabilities. These numbers are calculated using advanced statistical methods and simulations taking into account a host of factors such as each team's year-to-date run differential, current roster composition, playing time projections and remaining schedule according to Baseball Prospectus. Fangraphs calculates its playoff probabilities in a similar fashion. While

these models use a lot of information and advanced statistical methods to predict the outcome of the current Major League Baseball season, the mechanics behind the calculation are proprietary. Since little is publicized about their methods, a typical fan is unlikely to understand from where the probabilities come. Statisticians often encounter a balancing act between parsimony and predictability. The goal of this research is to outline an analysis procedure that is effective in predicting Major League Baseball postseason probabilities and be easily understood and applied.

The rest of this work is organized as follows. Chapter 2 contains a literature review covering the history of the use of statistics in baseball along with how analysts and statisticians have used statistics for applications such as winning games and predicting playoff probabilities. Chapter 3 provides an overview of the variables used in the study and the methods used to analyze the data. Mixed multiple logistic regression models using a backwards elimination procedure were fit. Model evaluation criteria for assessing how well the models fit and performed when making predictions such as area under the curve and concordance were used. The models themselves along with evaluations and potential explanations are found in Chapter 4. Chapter 5 uses the models in Chapter 4 to make predictions for the 2013-2015 seasons. The predictions were compared to Baseball Prospectus's playoff odds and who actually made the playoffs in the 2013-2015 seasons. Chapter 6 explores conclusions, considerations, and points for future studies.

Chapter 2

Literature Review

History of Baseball Statistics

Baseball and numbers go hand in hand. It is rare to discuss baseball without the use of numbers and statistics. How many wins does the team have? Have many runs has the team scored? How many home runs did this player hit? What is this pitcher's ERA? The list of questions and answers involving numbers is seemingly endless. Some statistics such as home runs, batting average, and runs allowed have been used for as long as baseball has been played, and new statistics and metrics are being developed today. Bill James is often regarded as the father of baseball statistics, but as Ben Baumer notes, "*Bill James never claimed to invent baseball analytics and he didn't. Many of James's insights were developed years earlier by others. Nonetheless, James did suggest the term 'sabermetrics,' advance the sophistication of statistical analysis, and significantly helped spread its practice*" (Baumer). James defines sabermetrics as "the mathematical and statistical analysis of baseball records." (Palmer). In addition, Michael Lewis's book *Moneyball* featuring the Oakland Athletics is regarded in popular culture as revolutionary. While not to diminish its impact, there were other contributors long before, even decades before, who made significant and groundbreaking contributions to baseball research and statistics.

On October 25, 1845 the New York Herald printed the first box score from game a between the New York Ball Club and the Brooklyn Club. The box score only contained

"Hands Out" (at bats in which the batter did not end up scoring) and "Runs" for runs scored as it was based on cricket. Shortly after, the box score began to evolve to better explain the results of a baseball game, in large part due to Henry Chadwick (Palmer). Chadwick is also given credit for creating a way to measure a player's fielding ability which was a predecessor to Bill James' range factor. Early in the twentieth century F.C. Lane detailed the problems with batting average describing a hitter's ability and even called it "worse than worthless" pointing out it should also account for walks and extra base hits. This led to Lane devising the earlier linear weights system which assigned weights to each type of hit – single, double, triple, and homerun – to give the number of expected runs each hit resulted in paving the way for James' runs created metric later in the twentieth century (Baumer).

Significant sabermetric progress was made in the late nineteenth and early twentieth century, with new and improved ways to evaluate players continually being created. Yet nearing the middle of the twentieth century no statisticians, were employed by a professional baseball organization. That changed with the innovative baseball minded Branch Rickey when he hired Allan Roth in 1944 to be the first team statistician in baseball. Roth persuaded Rickey that on base percentage should be used as a "basis for evaluating a batter's talents" – long before *Moneyball* popularized it. In 1964, Earnshaw Cook published *Percentage Baseball* in which heavily cited statistics today such as on base percentage and slugging percentage were emphasized in his Scoring Index. While many of Cook's mathematically advanced methods were hard for many to understand, they provided a look forward to current mathematical and statistical methods. Statistics and sabermetrics also started to creep from management to the field in the middle of the twentieth century. Earl Weaver is often credited for pioneering the use of sabermetrics during in-game situations with opposing hitting and pitching matchup data. He also

did not like giving up outs with strategies such as bunting or being caught stealing bases. Influential people in the industry started taking notice of Weaver, and gradually MLB organizations started hiring people to conduct statistical analysis (Baumer).

Statistics of Winning Games

The main determinant of whether or not a team makes the playoffs is the number of games the team wins. Much work has gone into evaluating and predicting wins. *Moneyball* states that Paul DePodesta estimated it would take 95 wins for the 2002 Oakland Athletics to make the playoffs, with a +135 run differential to win that many games with the state of the their team (Baumer). While *Moneyball* is unclear what drove DePodesta to those particular numbers, there has been research into winning and what characteristics cause a win. Pete Palmer attempted to find how runs and wins are related in his 1982 *"Runs and Wins"*. Palmer gave a rule of thumb that 10 runs scored or runs allowed roughly equates to a win. He also cited Earnshaw Cook's *Percentage Baseball* in which Cook estimated a team's winning percentage as .484*runs scored / runs allowed. Bill James's Pythagorean Record Formula is $(RS)^x/[(RS)^x+(RA)^x]$ which gives an expected winning percentage based on runs scored (RS) and runs allowed (RA). James used x = 2 mainly for simplicity but others have used x =1.81 or 1.83 as a more accurate parameter. James's Pythagorean Record formula is slightly flawed as it is impossible to predict a record above .500 with a negative run differential but it is still a useful tool. (Costa).

In addition to team totals such as runs scored and runs allowed, there have been many metrics developed over time to assess an individual player's contribution to his team winning a game. Today the most well-known statistic is wins above replacement (WAR). This statistic

estimates how many more wins a player contributes to his team than a league average replacement player does. There are several different ways of calculating WAR, and quantifying a "replacement player" is a topic of debate. Before WAR the focus was on runs because of the relationship between scoring and winning games. Bill James developed the earliest version the number of runs a player creates (called runs created), which has become more detailed over time. John Thorn and Pete Palmer came up with a pitching runs metric. This metric compares the number of runs a pitcher allows to the league average (Costa). Bill James also sought to improve on runs created by devising the win shares metric that can be read much like WAR is today. For example, an individual MVP caliber player equates to a wins share of 30, or above 6 or 8 on today's WAR (Costa).

Up until the second half of the twentieth century, computational methods were limited because of data storage and lack of computing power. Some of the first baseball researchers to rely heavily on computers were R.E. Truman and W.G. Brigss. Truman used a simulation model of 5000 thousand games in 1959, while Briggs used Monte Carlo simulation in 1960. Earnshaw Cook's mathematically rigorous methods were a glimpse into what could be studied or verified with the onset of computers. One of Cook's assertions was that a batting order should be created from best hitter to worst hitter instead of a more traditional lineup in which the fastest player hits first and power hitters hit third or fourth. Using a Monte Carlo simulation of 200,000 games, Allan Freeze was able to dispel Cook's ideas on the ideal batting order. In 1972 Harlan and Eldon Mills published *Player Win Averages: A computer Guide to Winning Baseball Players.* The authors gave a value of "win points" which took into account the outcome of a play and the how important this was for his team winning the game (Palmer). They also simulated the entire remaining games given a starting situation. Their work was based on play-by-play historical

data to determine how each play progressively changed the probability that the team would win the game (Mills). Now computer simulations are commonplace in predicting probabilities of winning games and making the playoffs.

Modern Methods for Predicting Playoff Probabilities

Baseball Prospectus calculates “each team's probability of winning the division or wild card, or any postseason berth. Probabilities are based on thousands of Monte Carlo simulations of the remaining season schedule incorporating each team's year-to-date run differential, current roster composition, playing time projections and remaining opponents.” (Standings) Baseball Prospectus also gives an Adjusted Playoff Percentage which gives the probability that a team will make it to the division series; hence it includes making a wildcard game and winning it (McQuown). Fangraphs also publishes its current year playoff probabilities while Coolstandings.com houses the historical playoff probabilities (Appelman). Coolstandings calculates the expected outcome of games by “estimating how many runs each team will score, on average, against every opponent on its schedule. Coolstandings takes into account factors such as home vs. away, team performance, remaining strength of schedule, run differential and league scoring averages. Previous season performance is also used, especially near the beginning of a new season. Coolstandings even takes into account all of the MLB tiebreaker scenarios when determining divisional and wild-card chances. Some factors that are not considered, include starting pitchers, trades, free-agent signings and injuries. Roster changes are indirectly taken into account because the model emphasizes recent team performance over older data. The estimated runs for each team are plugged into a modified version of Bill James' Pythagorean

Theorem to determine the odds of each team beating the other teams on its schedule. The data is regressed to the mean, so hot or cold starts do not weigh too heavily on the simulation. Random error is also added, since a .550 team may actually be an underachieving .600 team or an overachieving .500 team." (Agami). The current methods for predicting the probability of a team making the playoffs are rather complex for baseball officials or casual fans to understand. The following sections include a discussion of a new, user-friendly method of predicting playoff probabilities.

Chapter 3

Model Development

Logistic Regression Model

The type of statistical model used in this work to predict playoff probabilities is logistic regression. A summary of the variables under consideration is included below along with properties that are specific to logistic regression as opposed to linear regression. More details about the model and variables are provided as well as how the models were evaluated after they were fit.

Response

The response variable is a binary (0,1) variable where "0" represents a team that did not make the playoffs and a "1" represents a team that made the playoffs.

Predictors

X_1: Games behind the division leader (negative if team is in first place)

X_2: Number of teams ahead in the division (0 if team leads division)

X_3: Number of games remaining against teams ahead in the division (0 if team leads division)

X_4: Number of games behind the second wildcard leader (negative for second wildcard

leader or teams ahead of the second wildcard leader)

X_5: Number of teams ahead in the second wildcard race (0 if team leads second wildcard or is ahead of second wildcard leader i.e. division leader or first wildcard leader)

X_6: Number of games remaining against teams ahead in the second wildcard race (0 if team leads second wildcard or is ahead of second wildcard leader i.e. division leader or first wildcard leader)

X_7: Pythagorean Win Percentage

A random effect for year is also included in the model to control for variation due to the season.

Analysis Procedure

Whether or not the team did make the playoffs is a binary response, and thus typical linear regression cannot be used. Analyzing the binary response data with linear regression would result in predicted values outside of the possible 0 or 1 range. Teams can either make the playoffs or not make the playoffs. There is nothing more or less and nothing in between making the playoffs as this work considers wildcard teams as full playoff participants. Binary responses also violate key assumptions of linear regression such as normality of the residuals and homogeneity of variance. The response distribution could follow a binomial distribution with a variance dependent on the probability of success. Since the response will change with a change in the predictors, the variance would not be constant across all levels of X. A more applicable approach to analyze binary response data is using a generalized linear model. The mean response of a logistic regression is the log-odds of a success. Generalized linear models, commonly

referred to as GLMs, utilize a link function so that the model is linear in the predictors. A major advantage of generalized linear models is that they can be reformed so that the log-odds are in terms of a probability.

Model Details

Several dates were chosen corresponding to major events in the Major League Baseball season. The dates included May 1st, June 1st, July 1st (roughly the halfway point of the season), July 30th (the day before the non-waiver trade deadline, and August 31st (the final day players must be on the forty man roster to be eligible for postseason play). Major League Baseball expanded its playoffs from having one wild card team to two wild card teams beginning in 2012. The entire data set includes data from 1998-2015. In order to have the models match the current five team per league playoff format for the years 1998-2011, the first team out of the playoffs or the would be second wildcard team was considered to be a playoff team and was coded "1." An initial analysis was conducted using data from 1998 through 2012. The years 2013-2015 were used to assess model predictability performance.

$$\text{Logit}(p) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6 + \beta_7X_7 + \beta_8X_{12} + \beta_9X_{13} + \beta_{10}X_{14} + \beta_{11}X_{15} + \beta_{12}X16 + \beta_{13}X_{17} + \beta_{14}X_{23} + \beta_{15}X_{24} + \beta_{16}X_{25} + \beta_{17}X_{26} + \beta_{18}X_{27} + \beta_{19}X_{34} + \beta_{20}X_{35} + \beta_{21}X_{36} + \beta_{22}X_{37} + \beta_{23}X_{45} + \beta_{24}X_{46} + \beta_{25}X_{47} + \beta_{26}X_{56} + \beta_{27}X_{57} + \beta_{28}X_{67} + \beta_{29}X_{123} + \beta_{30}X_{456} + \sigma^2_{\text{random}} + \varepsilon$$

Five models were run using R version 3.2.2 on the 1998 to 2012 data for the points of the season mentioned above. The seven predictors mentioned in the previous section were included. All twenty-one possible two-way interactions were included in the models along with two 3-way

interactions; one for all three variables involving divisional data and one for all three variables involving wildcard two data. A random effect for the year was also included in the model to account for year-to-year variation. The model is a mixed effect binary logistic regression model. Insignificant fixed effect terms were eliminated using a backward selection procedure screening out variables that had a z-value less than two which roughly corresponds to a p-value of .05. The result is a more parsimonious model for each of the time points with the response being the log odds of making the playoffs based on the values of the predictors. For years after 2012, the final year of data used in fitting the model, the values of the significant predictors can be supplied to predict the log-odds and subsequently the probability the team has of making the playoffs on or near the date of the respective model.

Model Performance

One way of looking at the effectiveness of a logistic regression is by using a receiver operating characteristic curve abbreviated ROC (Fawcett). The graph plots the true positive rate, correctly predicting whether a team makes the playoffs or not, or sensitivity against the false positive rate, the rate of incorrectly predicting teams to make the playoffs or not, or the complement of the specificity for all possible cutoff values. The common measure of accuracy associated with ROC curves is the area under the curve (AUC). An AUC of .5 means the model has no discriminating ability and is shown by the 45-degree line on each of the ROC plots. The closer the AUC is to 1 the more accurate a model is at making predictions. Visually this is seen as ROC curves closely following the upper left portion of the graph having higher AUC values.

A rating scale for AUC values is much like academic percent scores and can be found in Table 1 (Using The Receiver). A similar way of determining predictability of logistic regression models is to use concordance and discordance. For all possible 0,1 pairs of observations, the model is concordant if the fitted value of the observation with the response "1" is larger than the fitted value of the response "0". Concordance is the percent of all the pairs that were concordant[1]. Here it gives the percent of all pairs where the model gave higher log-odds of making the playoffs to the team that actually made the playoffs than the team who did not make the playoffs (Kutner). Concordance is related to AUC in that AUC = concordance + the percent tied, where the percent tied assigns the same probability to pairs with differing responses in actuality (Mainkar).

Table 1: AUC Rating Scale

AUC	Rating
0.90-1.0	Excellent
0.80-0.90	Good
0.70-0.80	Fair
0.60-0.70	Poor
0.50-0.60	Fail

[1] Concordance was calculated using the Association() R function written by Vaibhav Mainkar of StaTour

Chapter 4

Analysis

Major League Baseball standings and schedule data was taken for each of the following dates, May 1st, June 1st, July 1st, July 31st, and September 1st. The data from each date over the 1998-2012 seasons was fit using the model and procedure outlined in Chapter 3. Here the results of the model fits are displayed with possible explanations for why certain variables are significant predictors for making the playoffs on the dates included. The performance of each model is also evaluated.

May 1st Model

Model 1: $\text{Logit}(p) = -2.9835 - 0.1748X_4 + 4.9271X_7$

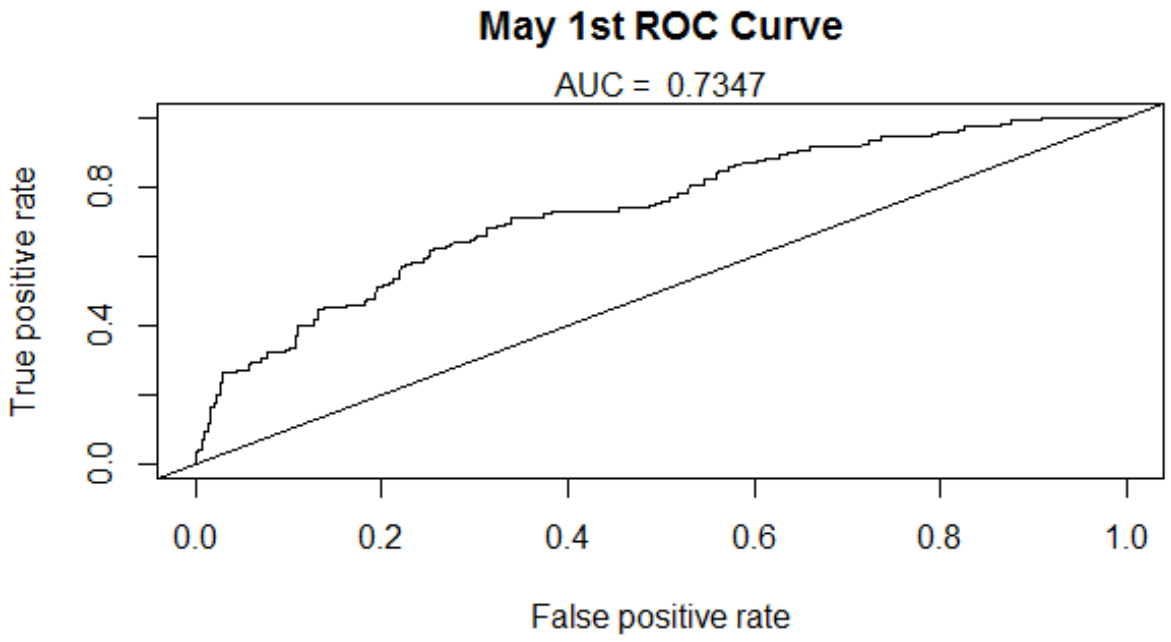
Just a month into the season the only significant predictors are number of games behind the second wildcard leader and Pythagorean win percentage (Table 2). Accurate predictions one month into the season are not expected as many teams initially play far above or below their true potential. It takes some time for teams to converge back to their true identity as more games are played. This is potentially why Pythagorean winning percentage is significant at this juncture – because it aims to avoid flukes by using runs scored and runs allowed which are generally reasonable indicators for how good or bad a team really is and what their true record should be. Games behind the second wildcard leader could be attributed to not falling too far behind to make up a difference later in the season. Only a month into the season wildcard pictures are still volatile and unclear, but if a team falls too far behind too early it can be tough to make up. The number of games behind the second wildcard leader explains the largest portion of deviance in

the model at 53.93%. The coefficient for the X_4 terms indicates that holding Pythagorean win percentage constant for each additional game behind the second wildcard leader the odds of making the playoffs decrease by 16.03% echoing the importance of not falling too far behind too quickly. Figure 1 shows the ROC curve for the May 1st model. The curve falls somewhat close to the 45-degree line and with an AUC of .7347 and a concordance of 73.47% indicating fair accuracy which is expected for predicting playoff probabilities this early in the season.

Table 2: May 1st Model Summary (Model 1)

Effect	Estimate	z-value	p-value	Deviance %
(Intercept)	-2.9835	-2.86	0.004	
GB_WC2	-0.1748	-2.90	0.004	53.93
Pythag	4.9274	2.59	0.010	6.63

Figure 1: ROC Curve for May 1st (Model 1)



June 1st Model

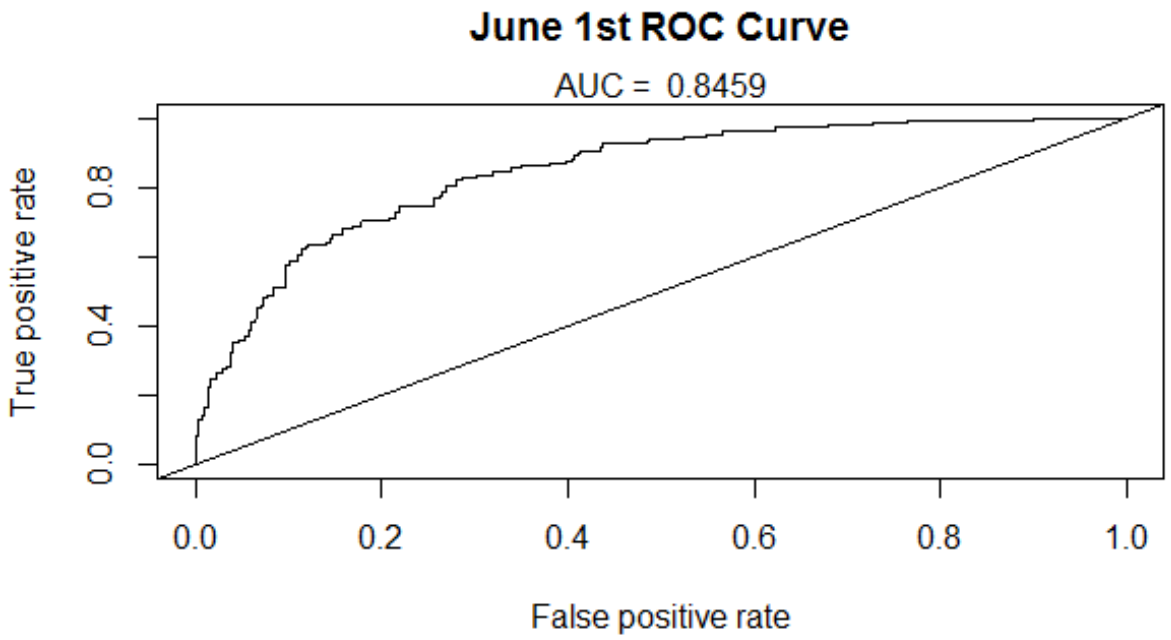
Model 2: $\text{Logit}(p) = -3.9678 - 0.3135X_2 - 0.1561X_4 + 7.8482X_7$

June 1st is an important point in the season. A third of the way through is generally when teams start to evaluate whether they will be contenders or not. This can drive important personnel decisions, particularly trades. At this point in the season the number of teams ahead in the division is significant along with games behind the second wildcard leader and Pythagorean win percentage. Both number of teams ahead in the division and number of games behind the second wildcard leader have a negative relationship with making the playoffs as shown in Model 2. Again, the theme is that teams do not want to fall far behind in the standings early on in the season and not be able to overcome the deficit. Number of teams behind in the division accounts for a large 71.18% of the deviance in the model (Table 3) and each additional team behind in the division decreases the odds of making the playoffs by 26.91% holding number of games behind the second wildcard leader and Pythagorean winning percentage constant. Even a third of the way into the season the standings can vary significantly, and some teams have not converged to the level of play measured in wins and losses as they should. Again, a significant Pythagorean win percentage is not surprising. The accuracy of the model for June is notably better for the model in May as the ROC curve (Figure 2) is farther away from the 45 degree line. With an AUC of .8459 and concordance 84.59%, this model displayed a good range for accuracy which is useful still early in the season.

Table 3: June 1st Model Summary (Model 2)

Effect	Estimate	z-value	p-value	Deviance %
(Intercept)	-3.9678	-2.33	0.020	
Div_Teams_Back	-0.3135	-2.07	0.039	71.18
GB_WC2	-0.1561	-2.66	0.008	21.19
Pythag	7.8482	2.54	0.011	6.37

Figure 2: ROC Curve for June 1st (Model 2)



July 1st Model

Model 3: $\text{Logit}(p) = -3.5062 - 0.0438X_3 + 2.1443X_4 - 3.8783X_5 + 7.2918X_7 - 0.0244X_{34} + 0.0208X_{35} - 4.0249X_{47} + 7.3310X_{57}$

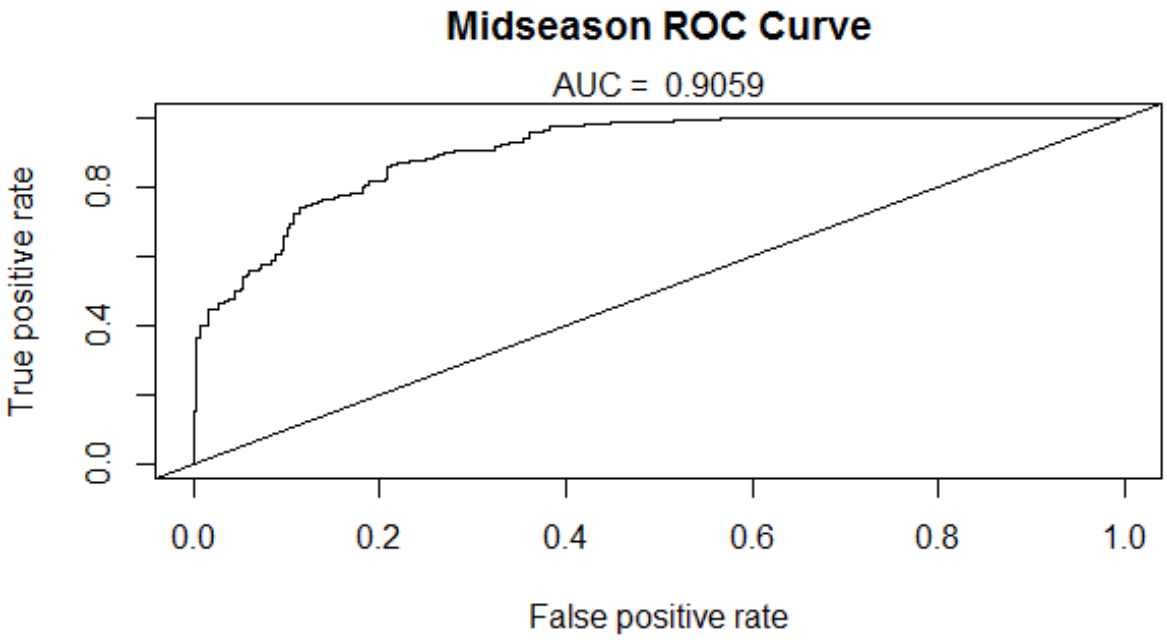
July 1st roughly represents the halfway point of the 162 game season and is a natural time to seriously consider playoff hopes as the trade market heats up. Front office officials need to

decide whether they will be buyers, sellers, or neither before the July 31st non-waiver trade deadline. Now many terms that are more significant enter the model. Model 3 contains main effects for games left against teams ahead in the division, games behind the second wildcard leader, games left against the teams ahead in the second wildcard race, and Pythagorean winning percentage. All five main effects appear in at least one two-way interaction. At about the midway point of the season the division and wildcard races start to take form. Given multiple avenues to qualify for the playoffs, it is not surprising that there are two interactions between divisional variables and wildcard variables. Second wildcard variables and Pythagorean winning percentage are featured in the other two interaction terms which appear to indicate a relationship between expected team performance and what the wildcard picture is in terms of games behind and teams to catch. The most deviance is being explained by games against teams ahead in the division and number of games behind the second wildcard leader (Table 4). While individual coefficients are difficult to interpret with interactions present, the overall accuracy of the model is excellent. The ROC curve (Figure 3) has an AUC of .9059 and a concordance of 90.59% which is desirable to be able to give teams advanced notice and time to make trades and acquisitions in the month before the non-waiver trade deadline.

Table 4: July 1st Model Summary (Model 3)

Effect	Estimate	z-value	p-value	Deviance %
(Intercept)	-3.5062	-1.15	0.249	
Div_Games_Left	-0.0438	-1.99	0.047	29.63
GB_WC2	2.1443	2.44	0.015	25.37
WC2_Teams_Back	-3.8783	-2.34	0.019	0.08
Pythag	7.2918	1.29	0.197	13.60
Div_Games_Left:GB_WC2	-0.0244	-3.08	0.002	4.07
Div_Games_Left:WC2_Teams_Back	0.0208	2.14	0.033	1.84
GB_WC2:Pythag	-4.0249	-2.52	0.012	1.22
WC2_Teams_Back:Pythag	7.3310	2.33	0.020	5.39

Figure 3: ROC Curve for July 1st (Model 3)



July 31st Trade Deadline Model

Model 4: $\text{Logit}(p) = -10.8813 + 2.0932X_4 + 0.0607X_6 + 21.1402X_7 - 0.0471X_{46} - 4.0743X_{47}$

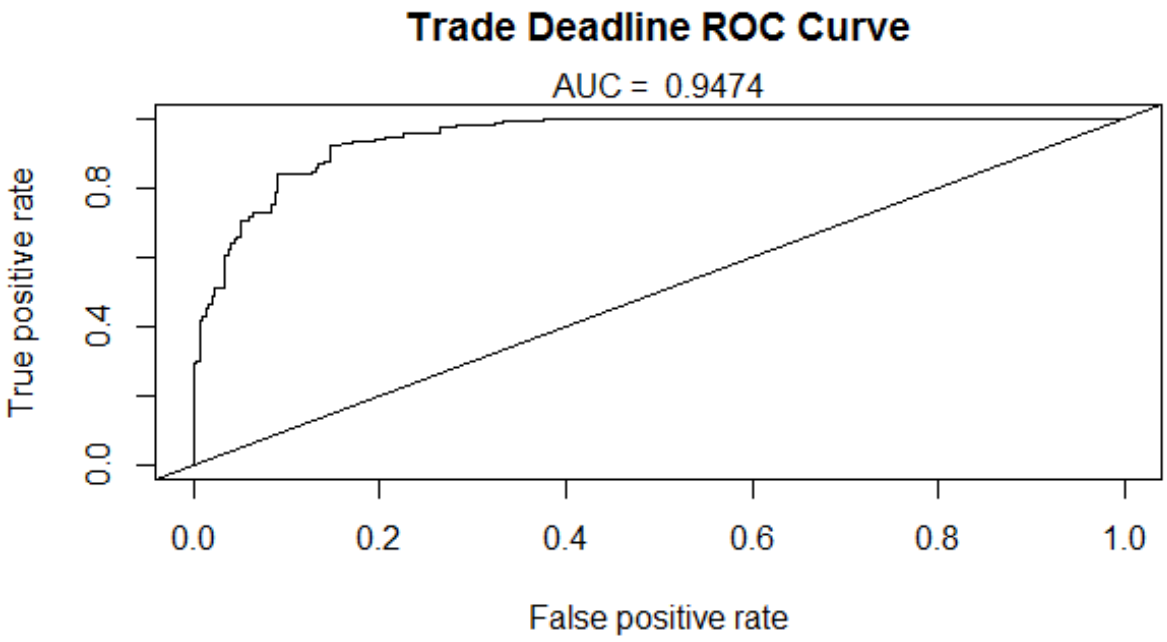
Perhaps the most important date in the baseball season is the July 31st non-waiver trade deadline. This date serves as the unofficial reality checkpoint where teams decide if they should make a run at the current postseason or essentially give up on the current year and try to become stronger for the next season via last minute deadline trades. The driver of Model 4 is number of games behind the second wildcard leader, the cutoff point for the playoffs. Games behind the second wildcard leader explains the most deviance of any of the other significant predictors at

37.14% and appears in both two way interactions; one with games left against teams ahead in the wildcard race and one with Pythagorean winning percentage (Table 5). Figure 4 shows the ROC curve closely following the upper left-hand portion of the graph with an excellent AUC value of .9474 and concordance of 94.74%. The higher accuracy of the model is essential as front offices decide on last minute deadline trades that cannot only affect the current season or the next season but also many seasons in the future.

Table 5: July 31st Trade Deadline Model Summary (Model 4)

Effect	Estimate	z-value	p-value	Deviance %
(Intercept)	-10.8813	-3.45	0.001	
GB_WC2	2.0932	2.52	0.012	37.14
WC2_Games_Left	0.0607	0.93	0.350	7.24
Pythag	21.1402	3.61	0.000	8.96
GB_WC2:WC2_Games_Left	-0.0471	-2.86	0.004	3.26
GB_WC2:Pythag	-4.0743	-2.65	0.008	6.98

Figure 4: ROC Curve for July 31st Non-Waiver Trade Deadline (Model 4)



September 1st Model

Model 5: $\text{Logit}(p) = -2.5708 - 0.4578X_1 + 0.9401X_2 + 0.3986X_3 - 0.3296X_4 - 13.2101X_5 - 0.4649X_6 + 4.8339X_7 + 0.3349X_{12} - 0.3423X_{23} + 0.2425X_{45} - 0.1875X_{16} - 0.8981X_{24} + 0.8772X_{26} + 0.1408X_{34} - 0.2693X_{35} + 27.1165X_{57}$

The last month of the season is the playoff push. Teams and fans alike want to know how likely it is their team will make the playoffs. To be eligible for the postseason roster a player must be on the forty-man roster before September 1st. This imposes a defacto waiver trade deadline (waiver trades can be made up to the end of the season) for acquiring potential postseason players. This date also serves as a deadline for placing minor league players already within the organization on the forty man roster (Taylor). With only a month to go in the season, corresponding to roughly 25-30 games left, the playoff picture is clear, except for a few teams, so models should be very accurate. The ROC curve in Figure 5 is almost perfect with an AUC of .9868 and concordance of 98.68%. The model is complex with all seven main effects included and nine interaction terms so interpretability is difficult but the model achieves its primary goal of being accurate (Table 6).

Figure 5: ROC Curve for September 1st (Model 5)

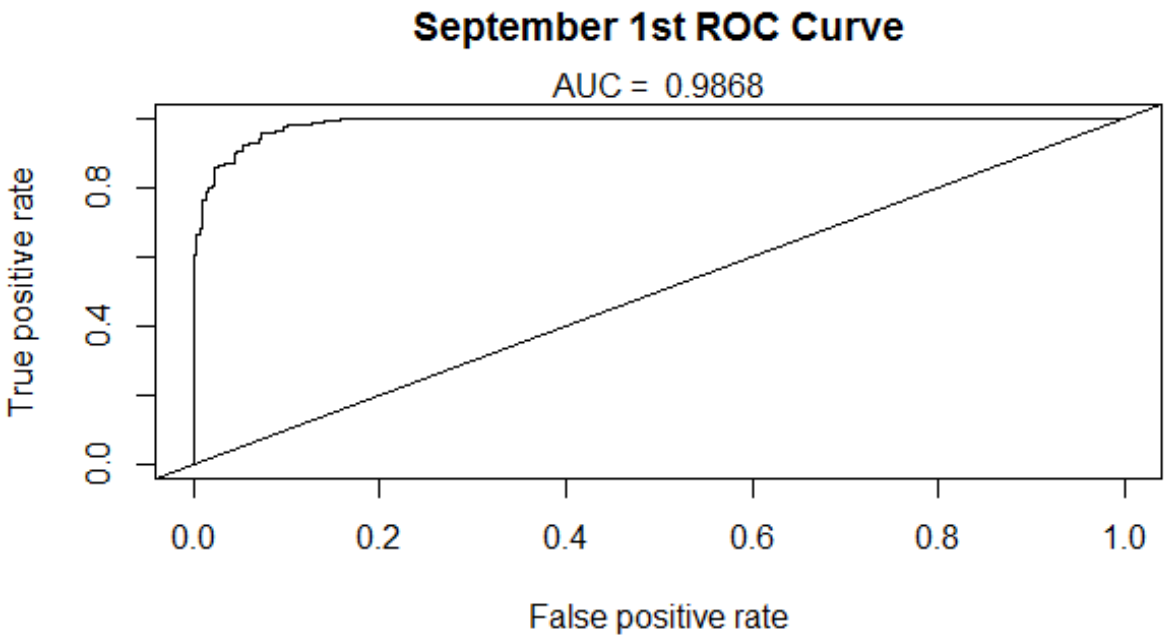


Table 6: September 1st Model Summary (Model 5)

Effect	Estimate	z-value	p-value	Deviance %
GB_Div	-0.4579	-3.54	0.000	4.54
Div_Teams_Back	0.9401	0.86	0.388	2.41
Div_Games_Left	0.3986	1.80	0.072	0.03
GB_WC2	-0.3296	-2.55	0.011	17.24
WC2_Teams_Back	13.2101	-2.03	0.042	0.02
WC2_Games_Left	-0.4649	-1.06	0.291	1.54
Pythag	4.8339	0.45	0.652	1.51
GB_Div:Div_Teams_Back	0.3349	2.99	0.003	0.62
Div_Teams_Back:Div_Games_Left	-0.3423	-2.25	0.024	0.28
GB_WC2:WC2_Teams_Back	0.2425	2.31	0.021	1.87
GB_Div:WC2_Games_Left	-0.1875	-2.19	0.028	1.20
Div_Teams_Back:GB_WC2	-0.8981	-2.41	0.016	0.01
Div_Teams_Back:WC2_Games_Left	0.8772	3.13	0.002	4.59
Div_Games_Left:GB_WC2	0.1408	2.30	0.021	1.49
Div_Games_Left:WC2_Teams_Back	-0.2639	-2.07	0.038	2.72
WC2_Teams_Back:Pythag	27.1165	2.15	0.031	4.59

Chapter 5

Prediction and Comparison to Other Methods

Every day of the baseball season Fangraphs and Baseball Prospectus publish playoff probabilities or playoff odds. As referenced earlier, these are highly computational and statistically advanced methods driven by a host of factors such as each team's year-to-date run differential, current roster composition, playing time projections and remaining schedule involving simulating the rest of the baseball season thousands of times. While the probabilities calculated by Fangraphs and Baseball Prospectus are well respected, specific details regarding their methods are not made public and their models are hard to understand. The models this work proposes are simpler and may work as well, or close to as well, making them favorable for interpretation. Models 1-5 fit on the 1998-2012 MLB data are used for prediction on the 2013-2015 MLB seasons. The resulting log-odds are transformed into probabilities of each making the playoffs at that particular point in the season. Playoff probabilities are compared between the models proposed in this work and Baseball Prospectus's methods. Fangraphs historical playoff probabilities are not available at the time of this work.

The playoff probabilities for 2013-2015 obtained from Models 1-5 were compared to Baseball Prospectus's playoff probabilities gathered from MLB.com. For each time point, year, and league the observations with the five highest playoff probabilities from Models 1-5 and Baseball Prospectus were saved giving thirty observations at each of the five time points. The percent of those thirty highest playoff probability observations that actually made the playoffs was then calculated and saved in Table 7. Baseball Prospectus achieves a higher

percentage of playoff teams for each time point and gets better as the season goes on. The logistic regression models only see about fifty percent of their top probability observations make the playoffs with a slight increasing trend as the season goes along.

Table 7: Percent of Five Highest Probabilities per League That Made Playoffs

Date	Logistc %	BP %
May 1st	40.00	60.00
June 1st	46.67	63.33
July 1st	53.33	73.33
July 3st	46.67	80.00
Sept 1st	60.00	96.67

Another way to compare the predictive power of the two methods is to do pairwise comparisons of the logistic regression and Baseball Prospectus's playoff probabilities for each team at each time point. For each observation in the 2013-2015 prediction dataset the logistic regression and Baseball Prospectus playoff probability was recorded. For each observation, the pair of probabilities is looked at to see which method was "more correct." If a team made the playoffs the method with the higher probability for that team and date "wins" and if a team did not make the playoffs the method with the lower probability "wins." The data is summarized in Table 8 which shows the percent of predictions for each method that was "more correct" than the other method. For teams that made the playoffs and teams that did not make the playoffs, Baseball Prospectus did a better job of predicting the outcome than the logistic regression models.

Table 8: Pairwise Comparison Percent Closest to Outcome

Outcome	Logistic %	BP %	Tie %
No Playoffs	35.59	57.63	6.78
Playoffs	29.68	69.03	1.29

Chapter 6

Conclusions and Considerations

Even though Baseball Prospectus's method did a better job overall of assessing playoff odds, there is still value to the logistic regression models presented in this work. The logistic regression modeling procedure is publicly available and can be extended upon by anyone. Baseball Prospectus, Fangraphs, and similar sources do not publicize how they determine their probabilities. Much more time has been devoted to developing the methods other sources use. With more time, it is reasonable to think that the logistic regression method here could be improved and be as accurate or close to as accurate as those resources. With the logistic regression method and the data it requires being readily available, anyone can apply the models or similar ones at any point in the future. If one would like to get an idea of the probability their team has of making the playoffs a week from the current date they would have to wait until that time when Baseball Prospectus or Fangraphs publishes them. With logistic regression models people can provide values for the predictors to see how their probability of making the playoffs changes. For example, if a team can make up any ground in the coming week they will have the ability to know where they will stand after that time. One of several major differences between methods is that the logistic regression models are heavily based off historical data whereas other methods heavily rely on current season data and projections for the rest of the season. The only variable pertaining to the composition of the current team in the logistic regression model is Pythagorean win percentage. It is possible other current year team statistics, possibly fielding independent pitching (FIP) or weight on base average, wOBA, could also be significant

predictors when it comes to the probability of making the playoffs. Further study investigating statistics such as those might be able to increase the accuracy of the logistic regression models without complicating them to the point where they would be much tougher to understand and apply.

There are several additional drawbacks of the logistic regression models used in this work that could warrant further study and improvement. The models provided are specific to the dates from which data was used to fit the model. There is some leeway however, for example using the trade deadline model a week before the trade deadline should produce reasonable predictions, although using it a month in advance would likely not result in accurate predictions. A way to improve upon this would be to find one model that could be applied to any point in the season though a predictor such as total games played to that point. The models do not take into account how many games are left between each individual team, only the total games left against teams ahead. For example a third place team playing a second place team a dozen more times and the first place team zero more times has less opportunities to make up direct ground on the first place team than if the third place and first place team had a lot of games remaining. Likewise, strength of schedule or performance to date against specific teams is not currently accounted for. The models could be split up into a model for predicting whether a team will win the division and a second model for whether or not a team will win a wildcard berth. That could possibly increase accuracy and decrease complexity by eliminating hard to interpret interaction terms that are present in the current models. A better way of coding data for teams in first place, division or wildcard, could be investigated. For example teams in first place are assigned a value of "0" for X_2 and X_5 (depending on leading division, wildcard, or both), and a "0" for X_3 and X_6 in that case. An alternative coding or variable that can measure how beneficial a leading position

in a playoff race is, similar to being negative games behind when in first place, without being zero for significant interactions could help increase prediction accuracy.

Bibliography

Agami, Greg. "What are your team's playoff odds? Check the standings." *ESPN*. n.p. 5 July 2008. Web. 18 April 2015.

Appelman, David. "Fangraphs Now Featuring coolstandings.com Playoff Odds." *Fangraphs*. n.p. 28 August 2013. Web. 15 March 2015.

Baumer, Benjamin and Andrew Zimbalist. *The Sabermetric Revolution*. Philadelphia: The University of Pennsylvania Press, 2014. Print.

Costa, Gabriel, Michael Huber, and John Saccoman. *Understanding Sabermetrics: An Introduction to the Science of Baseball Statistics*. Jefferson and London: McFarland & Company Inc, 2008. Print.

Fawcett, Tom. "An introduction to ROC analysis." Pattern Recognition Letters - Special issue: ROC analysis in pattern recognition archive 27.8 (2006): 861-874. Web. 17 November 2015.

Kutner, Michael, et al. *Applied Linear Statistical Models*. 5th ed. New York: McGraw-Hill/Irwin, 2005. Print.

Lewis, Michael. Moneyball: *The Art of Winning an Unfair Game.* New York: W.W. Norton, 2003. Print.

Mainkar, Vaibhav. "Concordance and Discordance in Logistic Regression." *StaTour*. Blogspot. 3 December 2012. Web. 15 October 2015.

McQuown, Rob. "Feature Focus: Playoff Odds." *Baseball Prospectus*. n.p. 29 August 2014.

Web. 15 March 2015.

Mills, Eldon and Harlon. *Player Win Averages: A Computer Guide to Winning Baseball Players*. South Brunswick and New York: A.S. Barnes & Company, 1970. Print.

Palmer, Pete and John Thorn. *The Hidden Game of Baseball*. Garden City: Doubleday & Company Inc., 1984. Print.

"Standings." *Baseball Prospectus*. Major League Baseball. n.d. Web. 18 October 2015.

"Standings." *Major League Baseball*. n.p. n.d. Web. 15 March 2015.

Taylor, Brett. "Everything You Need to Know About Today's Deadline, Playoff Eligibility, and September Call-Ups." *Bleacher Nation*. Bleacher Nation. 31 August 2015. Web. 19 October 2015.

"Using the Receiver Operating Characteristic (ROC) curve to analyze a classification model." *Math.utah.edu*. University of Utah. N.d. Web. 15 October 2015.

ACADEMIC VITA

Academic Vita of Evan Bittner
ebittnermail@comcast.net

Education:

The Pennsylvania State University – The Graduate School **December 2015**
- Masters of Applied Statistics

The Pennsylvania State University – Schreyer Honors College **December 2015**
- Bachelors of Science in Statistics – Applied Option
- Minors in Mathematics and Economics

Experience:

Corning Incorporated – Quality and Statistical Engineering Intern **2015**
- Coordinated, executed, and analyzed a measurement system analysis to evaluate scratches on Gorilla® Glass and wrote an internal report detailing findings
- Conducted performance testing on laminated Willow® glass and analyzed results

Sustainable Climate Risk Management Network – Research Assistant **2013-2014**
- Compiled and analyzed cosmogenic nuclide exposure data to estimate thinning rates of the Antarctic Ice Sheet with applications to sea level rise
- Presented a poster, "Estimating long-term ice thinning rates from cosmogenic exposure dates" at the 2013 West Antarctic Ice Sheet Workshop (NSF funded)
- Continued to work on project as an independent study course starting Sept. 2014

United States Census Bureau – JPSM Junior Fellow Intern **2013**
- Tested the Survey of Income and Program Participation's survey instrument and summarized Field Representative's feedback to improve the instrument
- Debugged SAS code to analyze paradata including the interview times and audit trails from wave one of the 2014 Survey of Income and Program Participation

Activities:

Penn State Statistics Club **2013-Present**
- Elected President for Fall 2015
- Elected Webmaster for 2014-2015

Penn State Orthodox Christian Fellowship **2012-Present**
- Elected President for 2014-2015
- Elected Vice President for 2013-2014

Honors:

Donald L. and Ellen Eberly Endowed Scholarship in Science
Mu Sigma Rho – Statistics Honorary
President's Freshman Award

Skills:

R, SAS, Minitab, JMP, SQL, C++