Major League Baseball Playoff Appearance Likelihood - Shannon Leiss

Introduction

Baseball is a game with many different variables and factors that contribute to a teams success that can be broadly grouped into offensive ability and defensive ability. Since the introduction of Sabermetrics and famous prediction formulas - made popular by Moneyball - many different attempts have been made to optimize a teams performance 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 explore in depth. The `Lahman` package in R provides many different historical Baseball data sets, including the `Teams` data set. The `Teams` data set has each teams overall performance in a wide variety of variables per season, along with information on playoff appearances, ball park factors, and attendance totals, a complete list of the variables in this data set is available in the Appendix.

One main area that analysts have tried to predict is a teams likelihood of making the playoffs. With the amount of data available, models can easily get over complicated and over fitted. Using the `Teams` data set for the seasons of 1998 - 2019, since 1998 was the last season additional teams were added to the major leagues and 2020 was a shortened season, offensive statistics will be used to model the likelihood of a team making the playoffs. Based off of these likelihood models, it will be investigated what “offensive threshold” a team will need to meet or exceed to have a 50% and 90% chance of making the playoffs. Finally, using the same offensive data, it will be compared which league - American or National - hits more home runs and which franchises historically hit the most home runs per season.

Predicting Probability of Playoff Appearance

Offensive Statistics Being Used: To model the likelihood of a team appearing the playoffs, the following offensive statistics were used in an initial model: Runs Scored, At Bats, Hits, Doubles, Triples, and Homeruns. Each of these variables are a teams total per season.

Along with the total offensive data per team per season, a secondary model with team batting average(Hits divided by At Bats) and a teams runs average(Runs divided by At Bats) will also be investigated. A few rows of the data set with the new offensive data are seen below:

R	H	AB	X2B	X3B	HR	RperAB	HperAB	Playoffs
787	1530	5630	314	27	147	0.1397869	0.2717584	0
665	1353	5491	235	46	159	0.1211073	0.2464032	0
826	1489	5484	297	26	215	0.1506200	0.2715171	1
817	1520	5565	303	11	214	0.1468104	0.2731357	0
876	1568	5601	338	35	205	0.1564006	0.2799500	1
861	1516	5585	291	38	198	0.1541629	0.2714414	0

Comparing Logistic Models: The above data shows that the `Playoffs` variable is a binary variable, with a 1 representing that that team made it to the playoffs that season and a 0 representing that the team did not make the playoffs that season. To model the likelihood of a team making the playoffs, logistic regression models were initially fit on the data. The basic formula of a binary regression using a logistic link function is:

$$log(\frac{p_i}{1-p_i}) = \alpha + \beta_i x_{ij}$$

where p_i is the probability that team i makes the playoffs.

Fitting an initial model with the basic effects of the offensive statistics listed above(R,H,AB,X2B,X3B,HR), the following model was produced:

$$log(\frac{p_i}{1-p_i}) = 25.306 + 0.0178(R_i) - 0.0068(AB_i) - 0.00015(H_i) - 0.0035(X2B_i) - 0.0112(X3B_i) - 0.003(HR_i)$$

This model had an AIC of 673.9, a residual deviance of 659.9 with 653 degrees of freedom, and only the intercept, Runs, and At Bats terms were significant. Since the residual deviance value divided by its degrees of freedom was 1.010562, there was no evidence of over dispersion.

From this initial model only having three significant terms, two versions of a reduced model were created - one with the effects of Runs Scored and At Bats and one that also included the interaction between the two predictors. These two models produced the following equations:

$$log(\frac{p_i}{1-p_i}) = 26.2955 + 0.0162(R_i) - 0.0071(AB_i)$$

$$log(\frac{p_i}{1-p_i}) = -115.292 + 0.1980(R_i) - 0.01837(AB_i) - 0.00003(R_i * AB_i)$$

Both of these models had no evidence of over dispersion, with the none interaction model having an AIC of 667.51 and all three terms were significant. The interaction model had an AIC of 664.75 with the intercept term being significant on the $\alpha = .1$ level, the Runs and interaction term were significant on the $\alpha = 0.05$ level and the At Bats term was not significant. Since there was no evidence of over dispersion, a drop in deviance test using the likelihood ratio was able to be used to compare the reduced model without interaction to the initial model as well as comparing the reduced model with interaction to the reduced model without interactions. Performing a drop in deviance test that

$$H_0 : \beta_{Hits} = \beta_{Doubles} = \beta_{Triples} = \beta_{Homeruns} = 0 \text{ Against } H_a : \text{at least one } \beta_i \text{ does not eqaul } 0$$

A drop in deviance of 1.6161 was found with a p-value of 0.8059 - found since the difference in deviance T follows a χ^2_4 distribution - which allows for the conclusion that there is not enough evidence that at least one of the β_i values above is not equal to 0 and thus the reduced model using only Runs and At Bats to predict the likelihood of making the playoffs should be used for this data.

A similar drop in deviance test was performed to compare the reduced model with an interaction term and the reduced model without the interaction term. This drop in deviance of 4.763 was found to have a p-value of 0.02908, concluding that there is some evidence that the model with the interaction between Runs and At Bats per season fits the data better than the model without the interaction term.

A greatly reduced model, which only used the number of Runs scored per season to predict a teams likelihood of going to the playoffs was also investigated and compared to the above models. This model was found to have no evidence of over dispersion and using the drop in deviance tests, it was found that both models that use Runs and At Bats fit the data better than the model with only Runs.

It confirm that the model with the Runs per season, At Bats per season and the interaction term was the best model for the data, it was also compared to a model that included all initial terms and all possible interactions between these terms. Using the drop in deviance test, the difference in deviance of 61.683 was found to follow a χ^2_{60} distribution, producing a p-value of 0.4157, allowing for the conclusion that the reduced model should be used to predict the likelihood of a team making the playoffs.

Continuing to use logistic regression, an initial model including a teams batting average, run average, and the interaction between the two variables was fitted. This model produced the following equation:

$$log(\frac{p_i}{1-p_i}) = -41.92 + 330.99(RperAB_i) + 110.21(HperAB_i) - 922.93(RperAB_i * HperAB_i)$$

This model had an AIC of 674.71, a residual deviance value of 666.71 with 656 degrees of freedom and the runs average was the only term significant on the $\alpha = 0.05$ level, with the intercept term being significant on the $\alpha = .1$ level. Two additional reduced models were fit, one that had both run average and batting average of a team without the interaction between the two, and one model using only a teams run average. Since there was no evidence of over dispersion in any of the models, a drop in deviance test using the likelihood ratio was used to compare all three test to each other. Through these tests it was found that the model with run average and batting average without the interaction should be used over the model with the interaction. It was also found that the model with both predictors fit the data slightly better (p-value = 0.07576) than the model with only the run average of teams.

Finally, it was found that the best two models using logistic regression and offensive statistics for the likelihood of a team going to the playoffs was as follows:

$$(1)\ log(\frac{p_i}{1-p_i}) = -115.292 + 0.1980(R_i) - 0.01837(AB_i) - 0.00003(R_i * AB_i)$$
$$(2)\ log(\frac{p_i}{1-p_i}) = -7.808 + 87.704(RperAB_i) - 19.536(HperAB_i)$$

These are the two models there were used for the model checking and analysis in the following parts.

Comparing Different Link Functions on Models: Binary regression can also use many different links, including a probit link and c-log-log link. These links have the following properties:

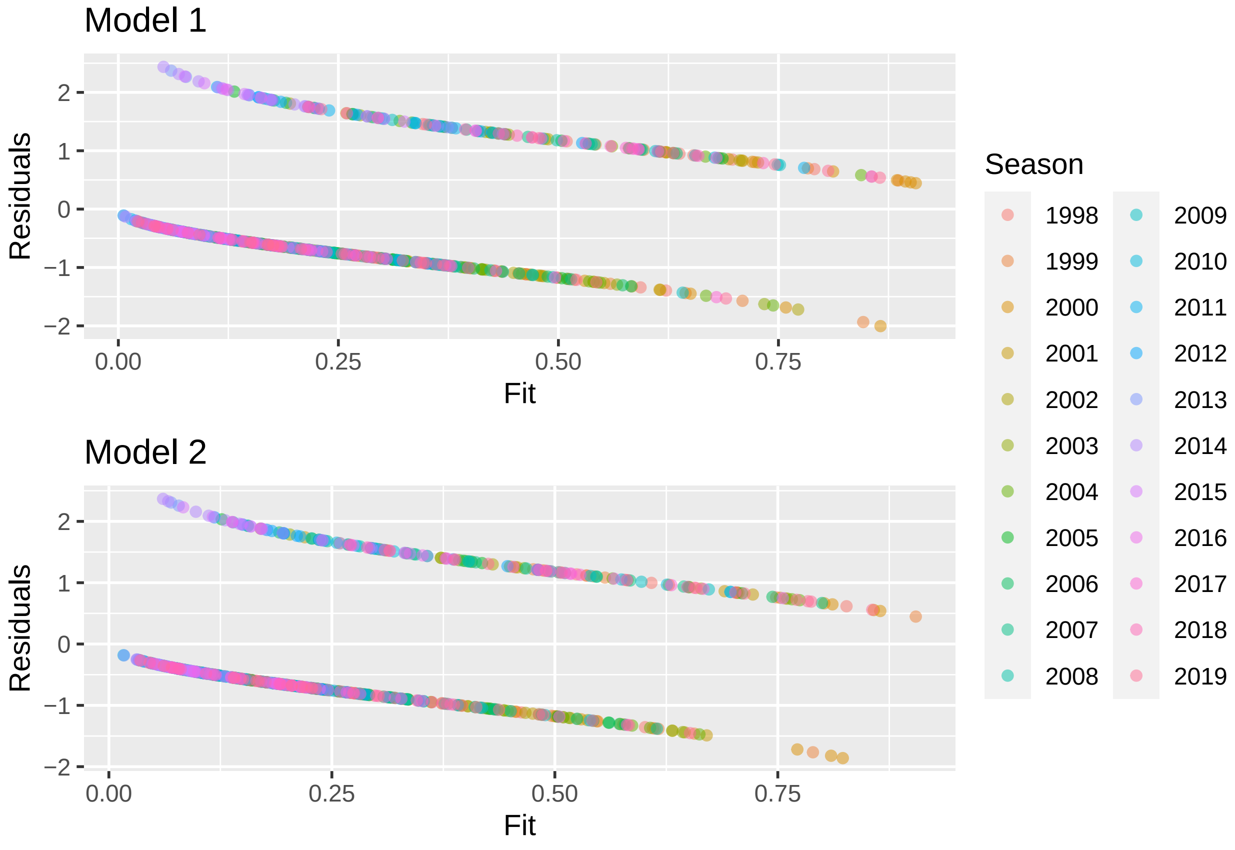
$$\text{Probit: } \Phi^{-1}(p_i) = \alpha + \beta_i x_{ij}$$
$$\text{C-Log-Log: } log(-log(1-p_i)) = \alpha + \beta_i x_{ij}$$

Both of these links were applied to the above "best" logistic models and slight variations of the models above - the form of formula (1) was also fit without the interaction on each link and the form of formula (2) was fit with the interaction on each additional link. The best model for each link was found using drop in deviance tests. Then the AIC values of each chosen model per link were compared to see if any other link fit the data better than the logistic models. For model (1) using the offensive data, the following AICs were found: **Logistic Link** had AIC of 664.7503, the **Probit** Link had an AIC of 664.6108, and the **C-Log-Log** Link had an AIC of 666.0074. It can be seen that for the first model using the outright Runs and At Bats with the interaction term - the probit model has the lowest AIC barely; however it is still high and is not a very large reduction to justify the computational complexity that comes with the probit model. Thus the logistic model for model (1) will continue to be used for further analysis.

For model (2), both the logistic and probit model chose the model without the interaction term between the predictors, while the c-log-log link chose the model with the interaction term. The AICs for each links chosen model are as follows: **Logistic Link** had AIC of 675.3243, the **Probit** Link had an AIC of 674.5555, and the **C-Log-Log** Link had an AIC of 674.1226. It can be seen that for the model with runs average and batting average has the best fit when using the c-log-log link and interactions are used. Again, for computational ease and consistency, the logistic model will still be used for any further analysis as the reduction in AIC is less than 1. It can also be noted that all possible links on the models using the runs and at bats per team will fit slightly better than all models using the runs average and batting average.

Model Checking: For the data being used, it can be noted that the hits, runs, and at bats of each team is independent of the hits, runs, and at bats of all other teams. They also assume that the seasons for each team are independent of each other, which is in reality not true, as a teams previous season performance can heavily influence the next seasons performance. This temporal correlation will be looked at through the below residual plots. These logistic models also assume that the base likelihood of a team is equal for all teams and that the probability of a team making the playoffs is independent of any previous likelihoods of that team making the playoffs. This assumption in reality is not true as each team has a different starting probability of making the playoffs based on the previous seasons performance and any new players acquired in the off season. While this would be an interesting thing to incorporate into the models, the information that this would require is not within the data set and would require extensive data cleaning and mutation.

Using the deviance residuals for both models, the following residual plots were found:



From the above residual plots it can be noted that both models have similar trends in the residuals. It appears that as the fitted values increase, the variation appears to be constant, but the residual values seem to be decreasing overall. It can also be noted from these trends that both models tend to underestimate the likelihood of going to the playoffs for more recent seasons and over estimate the likelihood of going to the playoffs for the 1990's and early 2000's seasons. It can also be noted that, when the residual plots were broken up by season, the plots still showed the same trend with under estimating for smaller fitted values and over estimating for larger fitted values. This is an interesting trend for the seasons as it can be noted that the game of baseball has become more offensively driven in the more recent seasons - with the average runs scored and at bats increasing for all teams. This residual trend could also be cause by the increase in number of teams that made the playoffs that occurred in the 2012 season. Before the 2012 season, only 8 teams would make the playoffs; however, in the 2012 season there were 10 teams that made the playoffs.

It is also worth noting that the probability that a team makes the playoffs does depend on the other teams within their division. There are 15 teams per league and 5 teams per division per league, making 6 divisional winners overall. Since only one team in each division will get a playoff spot, plus two additional playoff spot - known as wildcards that can be won by the teams with the best records that did not win their respective division, if another team in the division has already gotten the playoff spot, there is only two other possible playoff spot for a team to get. This will not be explored, but is worth remembering since it causes the observations on the response variable `Playoff` to be dependent of one another within a given season.

Applying Models to Data: Since the 2 models that have been fit are both logistic regression, the probability of each team making the playoffs can be found using the models and then back transforming the log-odds. The probability of a team making the playoffs can be found as follows:

$$(1)\ p_i = \frac{e^{-115.292+0.1980(R_i)-0.01837(AB_i)-0.00003(R_i*AB_i)}}{1+e^{-115.292+0.1980(R_i)-0.01837(AB_i)-0.00003(R_i*AB_i)}} \qquad (2)\ p_i = \frac{e^{-7.808+87.704(RperAB_i)-19.536(HperAB_i)}}{1+e^{-7.808+87.704(RperAB_i)-19.536(HperAB_i)}}$$

Using these equations and the two models to find the probability of each team making the playoffs in 2020 - if it were a regular season - using the 2019 season, it can be seen that the first few teams - the whole table can be found in the appendix - probabilities are:

	ARI	ATL	BAL	BOS	CHW
Model 1 Probability	0.3469	0.6563	0.1697	0.2806	0.1706
Model 2 Probability	0.4823	0.6565	0.2321	0.6516	0.1577

Using Models for Predictions

To find the "offensive threshold" a team needs to achieve to have a 50% and 90% chance of making the playoffs, the second model using the run average and batting average will be focused on as there is no interaction term.

Keeping Team Batting Average Constant: For the seasons of 1998-2019, the mean team batting average was found to be .2603, with a low of .2265 and a high of .2940, with team batting average per season being about normally distributed. Assuming that a team is "average" - meaning that the teams batting average is .2603 - the run average that a team would need to achieve for a 50% and 90% chance at making the playoffs can be found by setting the p_i values to .5 and solving for run average, the steps for this and all other equations for predicted offensive values can be found in the appendix. It was found that for a team to have a 50% chance of making the playoffs, if they have an average team batting average, they would need to have a team run average of .147 - meaning that they would need to average 1.47 runs for every 10 at bats. For a team to have a 90% chance of making the playoffs, a team run average of 1.72 runs per 10 at bats would need to be obtained. Using the delta method to find the 95% confidence intervals for the 50% estimator - see code in attached file for steps - it was found that a team would need a run average between (0.0808,.2132). This confidence interval is not informative since the upper and lower bounds are both outside of the range of team run averages over the 1998-2019 seasons.

Keeping Team Runs Average Constant: A similar process occurred for finding the batting average a team would need to have a 50% and 90% chance of making the playoffs. Using the overall average run average for all teams of 0.1345, the batting average that a team would need to achieve would be 0.204 for a 50% chance and 0.3166 for a 90% chance of making the playoffs. The delta method can also be used to find the 95% confidence interval for the estimate of having a 50% chance of making the playoffs of (0.09,0.318). This confidence interval is also not informative since its bounds are outside of the range of team batting averages.

It can be noted that the number of at bats a team has within a season does not have a wide variety of values and can be set as the average at bats for all teams(5542.671) to find the number of runs that a team would need to score in a season for a 50% and 90% chance of making the playoffs. The equation for finding these estimates can be found in the appendix. Using model 1 and keeping the number of at bats constant, the number of runs a team would need to score for a 50% chance of making the playoffs is 806.4 and the number of runs needs to have a 90% chance of making the playoffs is 937.7. The confidence intervals for these estimates are also uninformative as the lower bounds are negative and the upper bounds are more than double the maximum amount of runs scored by any team in one season.

Home Run Rates

National vs American League: The Homerun variable in the data set is the number of Homeruns hit by each team per season. Since this variable is a count variable, poisson regression was used to see if the American League or the Nation League hit more homeruns. When the poisson model for homeruns was fit with the league identifier and using the log link for poisson, there was evidence of over dispersion - the residual deviance was 4976.2 on 658 degrees of freedom. When the quasipoisson model was fit, the following model was produced - with an over dispersion estimate of 7.516:

$$log(\lambda_i) = 5.19368(\mathbf{I}_{AL_i}) + 5.12345(\mathbf{I}_{NL_i})$$

where λ_i is the homerun rate for each league.

From the AIC of the original model(9582.9), the model does not do a good job at predicting the number of homeruns that each team will hit. One factor that could effect the amount of homeruns hit is the type of home stadium each team has. Since baseball fields are not standardized, every team has unique dimensions and thus some parks are easier hit homeruns in. This effect is represented in the data as the `BPF` or the ball park factor. Adding this `BPF` to the model as an offset allows for the model to account for more of the variability in the different homerun rates. The quasipoisson model that is produced from adding this offset is as follows:

$$log(\frac{\lambda_i}{BPF_i}) = 0.58768(\mathbf{I}_{AL_i}) + 0.51602(\mathbf{I}_{NL_i})$$

Through examining the residual plots for both models, it showed that the model with the offset term was better at accounting for the variability in the data than the model without the offset. Using this model with the offset, the predicted homerun rates for each league were as follows: American League had a rate of 1.799816 homeruns and the National League had a rate of 1.675343. From this is can be seen that the American League overall has a higher homerun rate than the National League.

Divison Comparison: A similar process of finding best models for modeling homeruns hit was done to compare the three divisions that are in each league. It was found that the quasipoisson model that involved the interaction between league and division with the ball park factor offset was best for finding the homerun rate for each division in both leagues. The code in the attached file gives more details on the process taken to find this model, which was found using similar steps as the above model. The model is as follows:

$$log(\frac{\lambda_i}{BPF_i}) = 0.50845(\mathbf{I}_{AC_i}) + 0.64126(\mathbf{I}_{AE_i}) + 0.61208(\mathbf{I}_{AW_i}) + 0.55202(\mathbf{I}_{NC_i}) + 0.50342(\mathbf{I}_{NE_i}) + 0.48658(\mathbf{I}_{NW_i})$$

From this model, it was found that the homerun rates for each division were:

	AL Central	AL East	AL West	NL Central	NL East	NL West
Rates	1.6627	1.8989	1.8443	1.7368	1.6544	1.6267

It can be seen that the top two divisions for homerun rates are the American league East and Central divisions.

Team Home Run Rates: A more useful model for looking at homerun rates is for individual teams, as even grouping teams into their respective divisions is too general. Again, a quasipoisson model was fit for the homeruns predicted by team and offset by the ball park factor. The code in the attached file gives more details on the process taken to find this model, which was found using similar steps as the two above models. Using the model, the homerun rates for all teams were found. Below is the 5 teams with the highest homerun rates.

NYY	TOR	TEX	CHW	BAL
2.125224	1.953083	1.924496	1.890085	1.888022

It can be seen that the New York Yankees have the highest homerun rate of all franchises, and the top five franchises are all in American League - specifically the AL East or AL Central. When the full table of homerun rates is inspected, it can be seen that there are only 5 National League teams in the top 16 teams. The full table can be found in the appendix below.

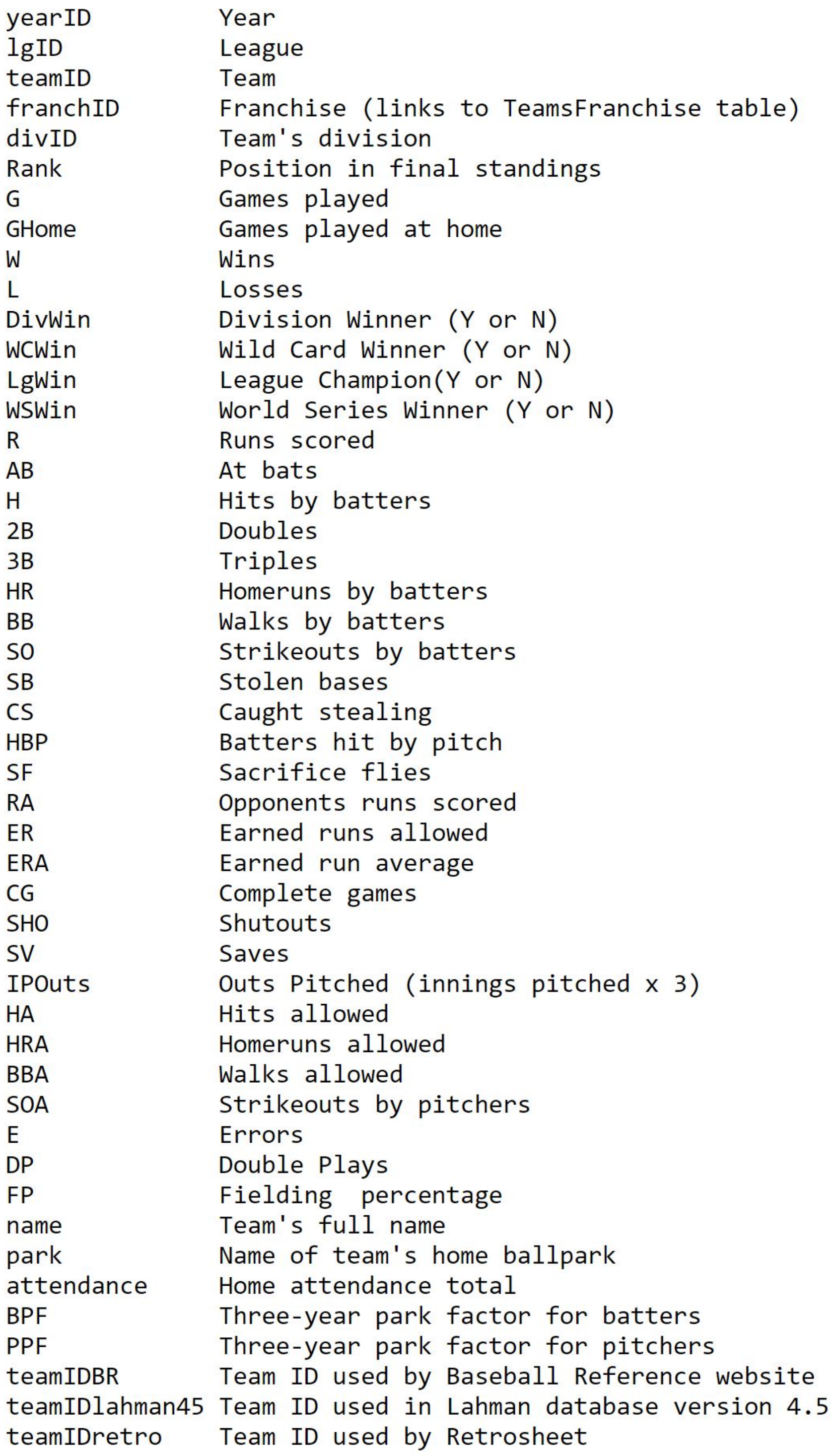
Conclusion

The main factor that currently distinguished the American League from the National League is the use of the Designated Hitter in the American League which is an additional player that does not play in the field and instead will bat in place of the pitcher during the game. The argument for keeping the Designated Hitter is that it creates more runs. Through the above quasipoisson models it was shown that American League teams will hit homeruns at a higher rate than National League teams, supporting the claim that Designated Hitter creates higher scoring games.

In a game with an abundance of factors that effect how a team will perform, the offensive statistics that matter the most for predicting the likelihood of a team making the playoffs are runs scored, at bats per season, team batting average, and team runs average. The models that were found using logistic regression could be used as a basis for further studying the effects of a teams previous season performance on the likelihood that that team will make the playoffs in the following season. These models also found that for a team to have a 50% chance of making the playoffs, they will need to score about 1.47 runs per every 10 at bats, or hit 806 hits in a season, which is about 5 hits per game.

Appendix

Code Book



yearID	Year
lgID	League
teamID	Team
franchID	Franchise (links to TeamsFranchise table)
divID	Team's division
Rank	Position in final standings
G	Games played
GHome	Games played at home
W	Wins
L	Losses
DivWin	Division Winner (Y or N)
WCWin	Wild Card Winner (Y or N)
LgWin	League Champion(Y or N)
WSWin	World Series Winner (Y or N)
R	Runs scored
AB	At bats
H	Hits by batters
2B	Doubles
3B	Triples
HR	Homeruns by batters
BB	Walks by batters
SO	Strikeouts by batters
SB	Stolen bases
CS	Caught stealing
HBP	Batters hit by pitch
SF	Sacrifice flies
RA	Opponents runs scored
ER	Earned runs allowed
ERA	Earned run average
CG	Complete games
SHO	Shutouts
SV	Saves
IPOuts	Outs Pitched (innings pitched x 3)
HA	Hits allowed
HRA	Homeruns allowed
BBA	Walks allowed
SOA	Strikeouts by pitchers
E	Errors
DP	Double Plays
FP	Fielding percentage
name	Team's full name
park	Name of team's home ballpark
attendance	Home attendance total
BPF	Three-year park factor for batters
PPF	Three-year park factor for pitchers
teamIDBR	Team ID used by Baseball Reference website
teamIDlahman45	Team ID used in Lahman database version 4.5
teamIDretro	Team ID used by Retrosheet

Playoff Likelihood for 2020 all teams

	Model 1 Probability	Model 2 Probability
ARI	0.3469	0.4823
ATL	0.6563	0.6565
BAL	0.1697	0.2321
BOS	0.2806	0.6516
CHW	0.1706	0.1577
CHC	0.6904	0.5831
CIN	0.2074	0.2163
CLE	0.5431	0.4374
COL	0.3610	0.4867
DET	0.0228	0.0355
HOU	0.7458	0.7711
KCR	0.1502	0.1678
ANA	0.3495	0.3869
LAD	0.8652	0.7875
FLA	0.0411	0.0617
MIL	0.3495	0.3886
MIN	0.4703	0.7837
NYY	0.8560	0.8559
NYM	0.2956	0.3794
OAK	0.6168	0.6586
PHI	0.3232	0.3957
PIT	0.1758	0.2269
SDP	0.1849	0.2051
SEA	0.3692	0.4119
SFG	0.0922	0.1402
STL	0.4778	0.4254
TBD	0.2306	0.3149
TEX	0.5205	0.5424
TOR	0.2539	0.3024
WSN	0.8064	0.7127

Offensive Threshold Equations

For finding the run average needed for 50% chance of making the playoffs when the team has an average batting average:

$$\begin{aligned} log(\frac{.5}{.5}) &= -7.808 + 87.704(RperAB_i) - 19.536(.2603) \\ 0 &= -7.808 + 87.704(RperAB_i) - 19.536(.2603) \\ RperAB_i &= \frac{(7.808 + 19.536(.2603))}{87.704} \end{aligned}$$

For finding the run average needed for 90% chance of making the playoffs when the team has an average batting average:

$$\begin{aligned} log(\frac{.9}{.1}) &= -7.808 + 87.704(RperAB_i) - 19.536(.2603) \\ log(\frac{.9}{.1}) + 7.808 + 19.536(.2603) &= 87.704(RperAB_i) \\ RperAB_i &= \frac{(log(\frac{.9}{.1}) + 7.808 + 19.536(.2603))}{87.704} \end{aligned}$$

For finding the Runs needed for a 50% chance of making the playoffs when the number of at bats is equal to the average number of all at bats:

$$\begin{aligned}
log(\frac{.5}{.5}) &= -115.292 + 0.1980(R_i) - 0.01837(5542.671) - 0.00003(R_i * 5542.671) \\
115.292 + 0.01837(5542.671) &= 0.1980(R_i) - 0.00003(R_i * 5542.671) \\
115.292 + 0.01837(5542.671) &= R_i(0.1980 - 0.00003 * 5542.671) \\
R_i &= \frac{115.292 + 0.01837(5542.671)}{0.1980 - 0.00003(5542.671)}
\end{aligned}$$

For finding the Runs needed for a 90% chance of making the playoffs when the number of at bats is equal to the average number of all at bats:

$$\begin{aligned}
log(\frac{.9}{.1}) &= -115.292 + 0.1980(R_i) - 0.01837(5542.671) - 0.00003(R_i * 5542.671) \\
log(\frac{.9}{.1}) + 115.292 + 0.01837(5542.671) &= 0.1980(R_i) - 0.00003(R_i * 5542.671) \\
log(\frac{.9}{.1}) + 115.292 + 0.01837(5542.671) &= R_i(0.1980 - 0.00003 * 5542.671) \\
R_i &= \frac{log(\frac{.9}{.1}) + 115.292 + 0.01837(5542.671)}{0.1980 - 0.00003(5542.671)}
\end{aligned}$$

Homerun Rate per Team

Team	Rate
NYY	2.125224
TOR	1.953083
TEX	1.924496
CHW	1.890085
BAL	1.888022
OAK	1.847020
MIL	1.835753
CLE	1.827539
BOS	1.821537
HOU	1.808482
STL	1.807710
CIN	1.801697
CHC	1.797853
SEA	1.779580
LAD	1.745747
ANA	1.714816
ATL	1.712279
PHI	1.709444
DET	1.709207
TBD	1.700515
NYM	1.698495
ARI	1.635619
SDP	1.624877
WSN	1.616983
COL	1.573328
SFG	1.564645
MIN	1.538217
FLA	1.532114
PIT	1.497692
KCR	1.353072