12/18/2021

Analysis of MLB Game log Data Using Various Modelling Approaches

Sean Chapman

0959653

Data 6100

Abstract:

The game logs from each baseball game in each Major League Baseball season have been recorded and are available publicly. Each game log contains 161 covariates representing different variables from the game. Using this information, different modelling techniques can be used to answer scientific questions of interest. A generalized linear model trained on the data could be used to predict which team will win based on the game data. Additionally, if the salary of a team could be estimated using a multiple linear regression model, the parameters could indicate what statistics can be improved with salary. The generalized linear model trained on a subset of the data was able to predict the outcome of the game with 98% accuracy. There are several statistical features significantly related to salary, which include triples, walks, strikeouts and stolen bases. A non-parametric bootstrap was performed on a small sample of the data using the variables corelated to the home team salary. At bats have the largest estimated effect on the log odds of the home team winning.

Introduction:

The analysis of this paper was conducted on game log data from retrosheet as well as salary data from stevetheump. The game logs from the previous four seasons were downloaded from retrosheet and used for the purposes of this analysis. These game log data frames are posted for each season and contain information from Major League Baseball games. For each game, there are 161 covariates recorded. These covariates often contain little information which would be useful in determining the outcome of the game. For example, 53 of these covariates are the names, id and positions of players in the game included as factors. In a generalized linear model, this information provides minimal to no information which can be usefully interpreted. Additionally, these factors contain as many levels as there are players who have played that position within the last four years. In many cases this would be in the hundreds. Therefore, for the purposes of this analysis, only the offensive and defensive statistics of the game were used to train the models.

The purpose of the models is to answer a few scientific questions of interest. One such question is what are the variables which are important to winning a baseball game? This question could be answered by constructing a model which is able to predict the winning team of a baseball team and analysing the summary statistics to determine which covariates are most significantly related to a winning outcome. Another scientific question of interest is, can the salary of a team be predicted based on the performance of the team? This question could be answered by constructing a multiple linear regression model using the statistical covariates of the data set and analysing the summary to determine which covariates are significantly related to team salary. The covariates which are most significantly related to team salary would be those covariates which the various teams are valuing the most. Another question of interest is whether the team salary is related to wins. This would be determined by the outcome of the generalized linear model. The team salaries will be included in this model. The estimates of the coefficients of the team salaries effect on the log odds of winning will be determined by this model. This will provide an answer to the scientific question of whether team salary is correlated with winning. The final question of interest is whether the regression coefficients of the generalized linear model with winning as the response can be estimated using a non-parametric bootstrap of a sample taken from the data. If this can be done successfully, fewer game records are needed to determine the effect of a statistic. This could prove advantageous if a new advanced statistic is incorporated into game log data since few games including this statistic would be needed before incorporating it into new analysis models.

Background:

The statistical information being used in the analysis includes several covariates which record the instances of a game event in each MLB game. The covariates used in the analysis are at bats, singles, doubles, triples, home runs, RBIs, sacrifice hits, sacrifice flies, hit by pitch, walks, intentional walks, strikeouts, stolen bases, caught stealing, grounded into double play, catcher interference, left on base, pitchers used, individual earned runs, team earned runs, wild pitches, balks, putouts, assists, errors, passed balls, double plays and triple plays. All the variables are integers and are included for the visiting and home teams separately. An initial model trained using these covariates had an unusually high estimate for the influence of putouts on the log odds of the response. This is due to the bottom half of the final inning. If the home team is winning after nine defensive innings, they do not need to play an offensive ninth inning. Therefore, if the putouts of the visiting team are less than the putouts of the home team, it can be determined that the home team wins. For this reason, putouts were removed from the model.

While checking the data for missing values, several missing values were identified within the dataset. All covariates containing the missing values were removed from the data set prior to the construction of the model as they were factor variables with many levels that contained little information.

Table 1: The count of NA values for each column in the compiled data set. Most variables contain no missing values and are listed as “All Others”.

Count of NA values	Column index
0	All Others
8171	Completion information
8129	Additional Information
899	Game Attendance
3327	Left Field Umpire ID
5756	Right Field Umpire ID
8187	Forfeit Information, Protest Information

Methods:

In order to predict the outcome of a given baseball game using the chosen covariates, a generalized linear regression (GLM) model was used. A GLM is used for the purpose of modelling a binary response variable. The model estimates the coefficients of each covariate in order to maximize the likelihood function for a given response. In order to interpret the coefficients, we use the log odds. We can exponentiate the estimated parameters to interpret the effect of the predictor variables on the response.

$$\log(\frac{\mathrm{p}}{1\text{-p}}) = \beta_0 + \beta_1 + \ldots + \beta_{\mathrm{p}}$$

For example, the unit increase in a predictor variable will increase the odds of the response by the exponentiated coefficient estimate. Since the outcome of a baseball game is binary, either the home team or the visiting team wins, then this can be modelled using a GLM.

In order to model the response of team salary against the covariates a multiple linear regression model is used. Team salary is a continuous variable and would not be modelled well using a GLM unless a logarithmic transformation was made on the data. A multiple linear regression model estimates the coefficients of each covariate in order to maximise the likelihood function. The interpretation of the coefficients is the change in the response variable estimated with a unit change of the predictor variable, assuming all other variables are held constant.

To assess the performance of the generalized linear model, cross validation is used. This technique involves sampling a portion of the data set and setting this portion aside. The remaining data is then used to train the model based on the predictor variables. The trained model is then used to predict the response variable for the sampled data using the predictor variables. The predicted response for the test set is then validated against the observed response in the test set. A more accurate model can more accurately classify the data points which were not used to train the model.

A non-parametric bootstrap is a resampling technique which can be used to sample additional data points using the empirical distribution of values for each covariate. The reason for using a non-parametric bootstrap would be in cases where there are few observations in the data. The purpose is to maximise the information available in the observed data. This is achieved through sampling the observed data with replacement many times. This provides many estimates of the coefficients of the model and allows a confidence interval to be created for these estimates.

Results:

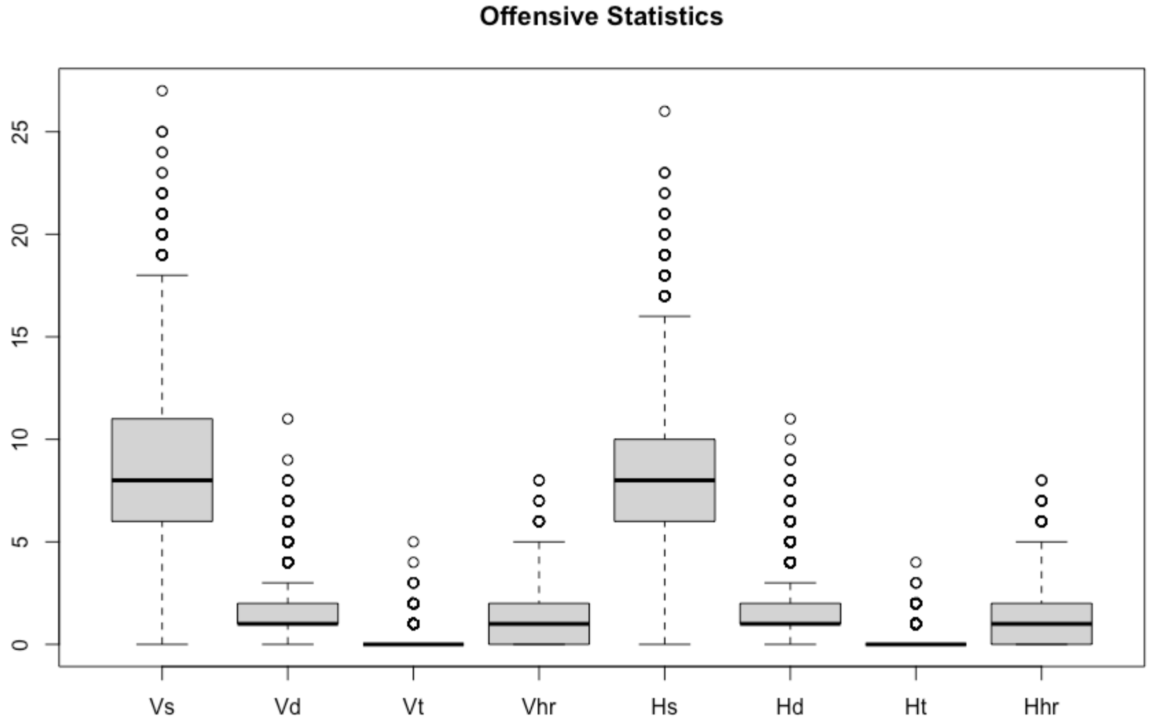

Figure 1: The distribution of offensive statistics including visiting singles (Vs), doubles (Vd), triples (Vt), home runs (Vhr) and home singles (Hs), doubles (Vd) triples (Ht) and home runs (Hhr).

Visualizing the distribution of offensive hitting statistics, we can see that singles are the most frequent offensive event in the game. The difference in the outcome of the game is not expected to be highly influenced by each single, although the total number of singles in the game could be corelated to the outcome. For example, if a team has greater than 12 hits in a game, that would be above the 75th percentile of hits in a game. Since hits are related to runs scored, we would expect a team with more than 12 hits in a game to have a high probability of a win. There are several outlier games which contain a large amount of hits. The median number of singles in a game for one team is around 7.

We can also see that home and visiting teams perform similarly for each offensive category. Home Runs occur around as frequently as doubles and are worth at least one run. Since home runs have a direct influence on the score, they would be expected to have a larger influence on the outcome of the game in the model. Triples occur much less frequently than the other offensive events. This could mean that each triple has a higher estimated coefficient in the model, especially if triples are more common among winning teams.

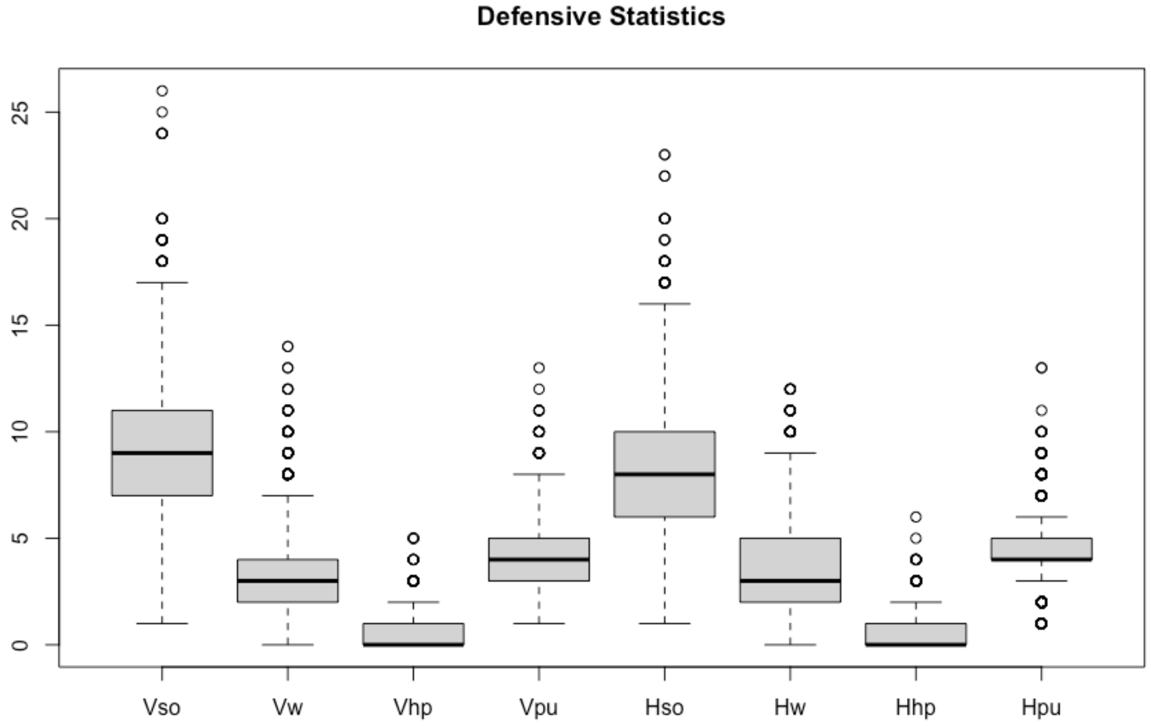

Figure 2: The distribution of defensive statistics including visiting strike outs (Vso), walks (Vw), hit-by-pitch (Vhp), pitchers used (Vpu) and home strikeouts (Hso), walks (Hw), hit-by-pitch (Hhp) and pitchers used (Hpu).

Visualizing the defensive statistics, we can see that the distributions for the different metrics are very similar for home and visiting teams across all games. Teams typically use a median of 4 pitchers per game. Most games, no players are hit by a pitch. The median strike outs per game are 9 per team, although this is skewed by outliers in the upper range. Strikeouts are the most common defensive event in the game, other than putouts which are not included in the model. Strikeouts typically do not have a large impact on the outcome of the game due to their frequency. Pitchers used is potentially an interesting covariate in the model. This is because traditionally more pitchers being used indicates the starting pitcher struggled early in the game and would most likely be positively corelated with a losing outcome. In more recent years, teams have been using a strategy where the starting pitcher will typically pitch one inning. This strategy has been dubbed the “opener” and the use of this strategy would likely change the dynamic between the number of pitchers per game and the outcome.

The number of walks yielded by a team would also be positively related with a losing outcome although weakly. This is because any hits with a runner on base is more likely to result in a run for the opposing team. This variable would not influence the outcome of the game as much as other, more common events like hits, or more impactful rare events like home runs.

Now that the distribution of several key variables is well understood, the full glm was constructed on the data using each of the covariates listed in the background section. The test data set for cross validation was created as a sample of 800 random observations from the 8187 game logs. This proportion is approximately 10% of the data. The remaining data was used to train the GLM. The ability of the full GLM to predict the outcomes of the test set is shown in table 2. The receiving operating characteristic (ROC) is plotted to assess the sensitivity and specificity of the model using different classification thresholds. The model performs best while using a classification threshold of 0.39. Observations where the probability of the home team winning is lower than this percentage are classified as visiting wins, while the other observations are classified as home team wins.

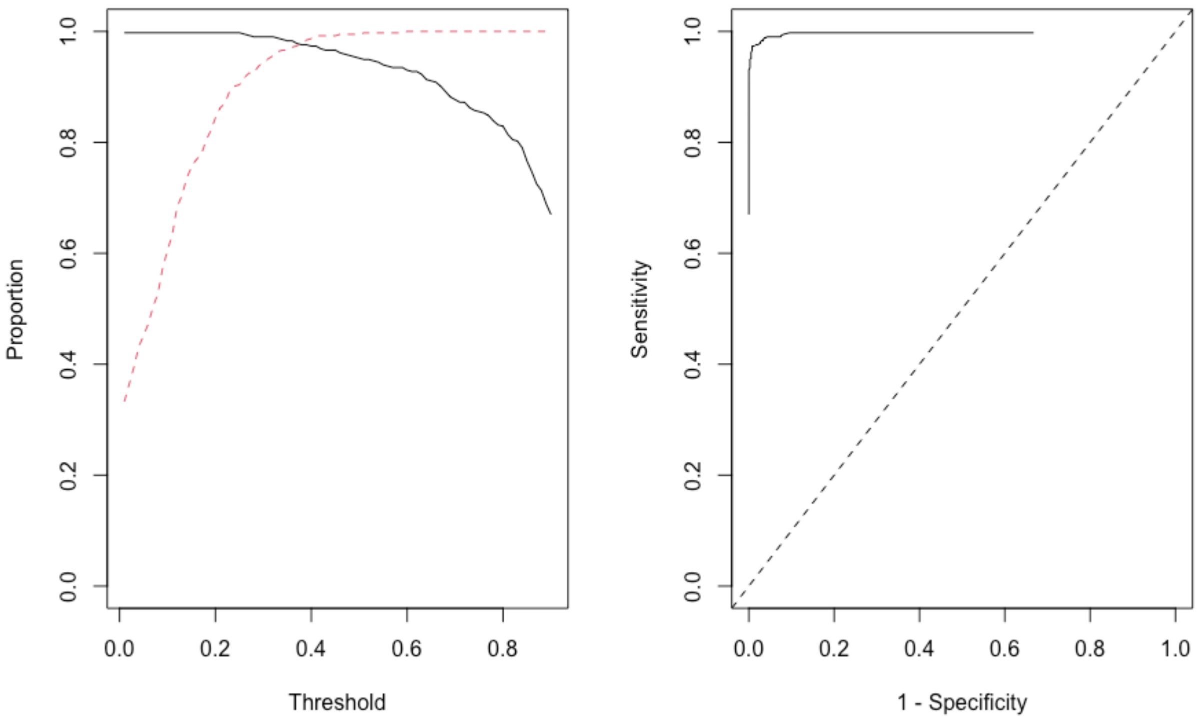

Figure 3: Sensitivity and Specificity of the full GLM predicting test results plotted as a function of the probability function on the left and as an ROC curve on the right. A classification threshold of 0.39 was chosen to maximize both Sensitivity and Specificity of the model.

Table 2: Confusion matrix for the full GLM. The full GLM has a misclassification rate of 2%.

Confusion Matrix			
		Predicted	
		Negative	Positive
Observed	Negative	379	6
	Positive	10	405

The full model was able to predict the outcome of the response for the test set with a 98% accuracy. However, this model uses many parameters and is difficult to interpret. Finding a reduced model which is provides similar predictive ability and allows greater interpretation is preferred. In order to do this, a stepwise regression model using AIC was created to reduce the number of parameters included in the model. The reduced model uses a classification threshold of 0.41 to classify the response of the test set. The model can predict the response of the test set with an accuracy of 98.25%. This model is not only more interpretable than the full GLM, but also predicts the response of the test set with greater accuracy.

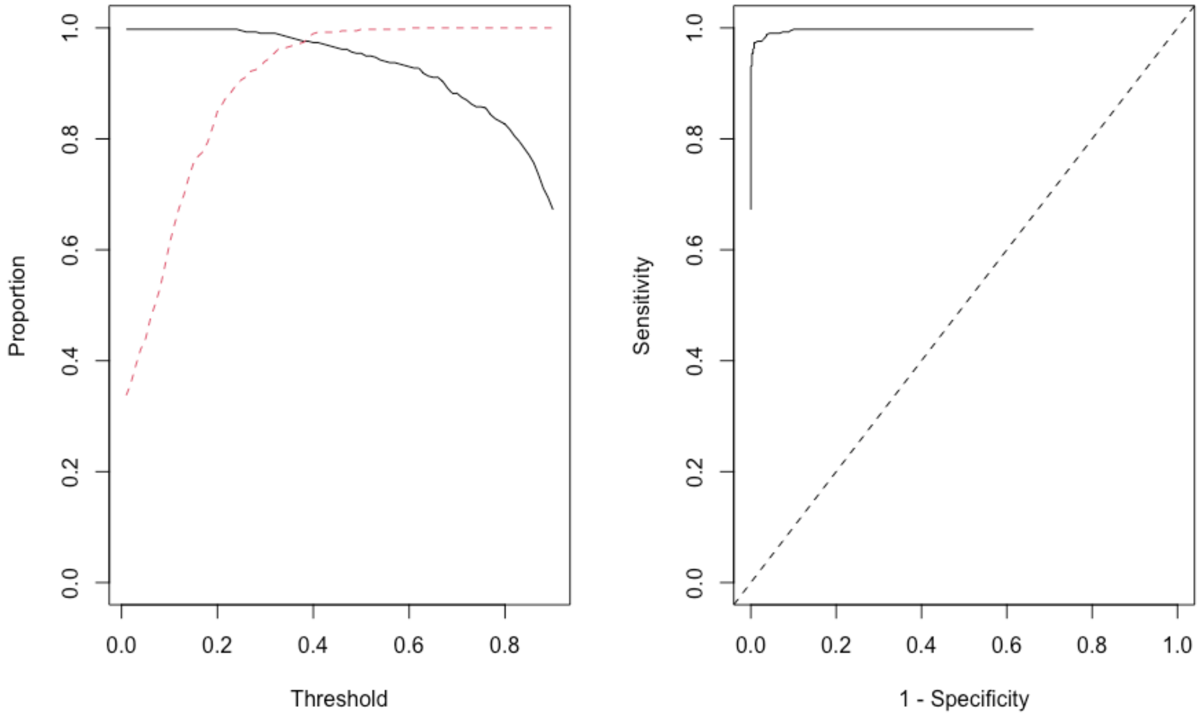

Figure 2: Sensitivity and Specificity of the full GLM predicting test results plotted as a function of the probability function on the left and as an ROC curve on the right. A classification threshold of 0.41 was chosen to maximize both Sensitivity and Specificity of the model.

Table 3: Confusion matrix for the reduced GLM. The full GLM has a misclassification rate of 1.75%.

Confusion Matrix			
		Predicted	
		Negative	Positive
Observed	Negative	382	3
	Positive	11	404

To answer the question of whether team salary is important to predict the outcome of a game, we can look at the coefficients included in the reduced model. Salary was not used in the reduced model, which suggests that salary is not significant in determining the wins. To verify this, the summary of the full GLM was observed to determine if the estimated effect of salary was significant. This estimate was not significant in determining wins of the full model. The visiting team salary had a p-value of 0.424 while the home team salary had a p-value of 0.783.

Variables identified as having a large influence on the outcome of the game included catcher interference. This event is very rare and was likely associated more strongly with one of the teams winning, creating a large, estimated coefficient in the outcome of the model. Hits, sacrifice flies and sacrifice hits all have a positive correlation with the winning team. These variables are traditionally associated with the strategy “small ball” which involves getting on base and advancing those base runners with sacrifice bunts, sacrifice flies and singles. Interestingly, home runs were not included in the reduced model.

To further investigate the relationship between team salary and the different variables, a multiple linear regression model was created using salary as the response. Two models were built separating visiting team and home team statistics. This was done to reduce the parameters of the model while not allowing the other team’s performance to influence the model. The initial models included all predictor variables. These models were reduced to create a stepwise regression model using AIC. The results of the reduced models are shown in tables 4 and 5.

Table 4: Coefficient estimates of multiple linear regression model predicting visiting salary using visiting statistics.

Parameter	Estimate	Std. Error	P-Value
(Intercept)	138553510	6328727	<2e-16
Triples	-6791602	4488219	0.1303
Sac Hits	-7681894	4430812	0.083
Sac Flies	5273767	3656634	0.1493
Walks	1828585	861368	0.0338
Strikeouts	-1084110	589765	0.0661
Stolen Bases	-5056438	2189941	0.021
Passed Balls	-11178733	6534430	0.0872

Table 5: Coefficient estimates of the multiple linear regression model predicting home salary using home statistics.

Parameter	Estimate	Std. Error	P-Value
(Intercept)	84146621	17291923	1.16E-06
At Bats	1652710	770805	0.032052
Singles	2723140	1487354	0.067158
Doubles	7415973	1538848	1.47E-06
Home Runs	5692155	1980419	0.004061

RBI	-6208727	1521805	4.55E-05
Walks	4990371	1591714	0.001723
Strikeouts	-1099442	659292	0.095432
Stolen Bases	-4909718	2223230	0.027246
Caught Stealing	-6926180	4513644	0.124946
Left on Base	-5694459	1491387	0.000135
Pitchers Used	3198731	1428983	0.025218
Individual Earned Runs	-1064185	613189	0.082692

The covariates which are significantly related to team salary are underlined and include walks and stolen bases for the visiting team data set. The significant parameters include at bats, doubles, home runs, RBI, walks, stolen bases, left on base and pitchers used. The results of the home team salary are more intuitive. Players who can hit more home runs and singles are typically paid substantially higher salaries. A few of these such players on a team could increase the overall salary by a significant amount. Strikeouts are included in each model and are significant at the 10% level. This is negatively corelated to salary since teams which pay higher salaries can afford players who strike out less often.

The bootstrap was created from a sample of 1000 games using the covariates significant to determining home salary and sampling with replacement. The estimates of this parametric bootstrap are shown in Table 6.

Table 6: Bootstrap estimated coefficients bias and standard error.

Parameter	Original	Bias	Std. Error
At Bats	1.86052061	3.41E-04	0.13017309
Hits	-0.05938	1.09E-04	0.0053908
Doubles	0.03551425	-3.22E-04	0.00793114
Triples	-0.0052578	-6.37E-04	0.00857741
Home Runs	0.00678681	-4.91E-04	0.01103024
RBI	0.09784806	4.37E-04	0.00807452
Walks	-0.0240197	-1.95E-04	0.00864057
Strikeouts	-0.0011543	-9.01E-05	0.00356141
Stolen Bases	0.03125661	-2.33E-04	0.01285246
Left on Base	0.03895674	1.56E-04	0.00825759
Individual Earned Runs	-0.0702744	-2.39E-04	0.00387758

Conclusion:

The GLM constructed to predict the outcome of a given game based on the information from that game was successfully able to predict this outcome. The summary of the reduced model revealed that within a game, parameters associated with the more traditional “small ball” strategy are more conducive to winning. Results of the salary models show that hitters who do not strikeout often and can hit doubles and home runs are highly valued since these covariates are significantly related to salary. These two models suggest a discrepancy between what team's value and what in game events increase the likelihood of a positive result. Perhaps the ability of a hitter to produce a sacrifice fly in a crucial moment is more valuable than being a constant threat to hit a home run. A possible extension of this analysis would be to consider the temporal aspects of the data. Conducting a similar analysis using only information from previous games could create a model which predicts the outcome of future games. This would have more interesting outcomes such as preseason projections of the standings based on rostered players and schedules. Additionally, betting odds could be made to accurately present the probability of each team winning once the temporal aspects are incorporated.

Data Ethics Impact Statement:

There are potential ethical complications from the results of this analysis. There is a growing discussion of the role of analytics in sport. As mentioned earlier in this report, using more pitchers is a more common strategy. As the use of analytical information becomes more widespread within the sport, the resulting changes in strategy come with criticism. One of the criticisms of the game today is the length of games, of which pitching changes are a factor. If analytical analysis shows that more pitching changes are beneficial to a team, they are likely to incorporate this strategy. While this improves the chances of winning, it removes from the audience's enjoyment of the game, as games last much longer. These changes can cause many to view the growing incorporation of analytics to be a negative. It is therefore important to be transparent with why a certain analysis is being done and what the goal of the analysis is.

Additionally, the 2020 season was influenced by the COVID-19 pandemic, as fewer games were played under a different setting. Some players missed playing time throughout this season due to COVID-19 symptoms or vaccination. These factors may affect the results of this analysis or any other which uses data from the 2020 season or later.

References:

Game logs data retrieved from:
https://www.retrosheet.org/gamelogs/index.html

Team Salary information retrieved from:
http://www.stevetheump.com/Payrolls.htm