Course Number: MATH-570-01 *Instructor: Professor Osei Tweneboah*

A Statistical Approach to Understanding Baseball Attendance

Luis Amadis Madrigal [1], Paul Galli [2], Sebastian Pross [3], and Ian Hill [4]

[1] Affiliation 1; lamadism@ramapo.edu
[2] Affiliation 2; pgalli@ramapo.edu
[3] Affiliation 3; spross@ramapo.edu
[4] Affiliation 4; ihill@ramapo.edu

Abstract: Predicting home attendance in Major League Baseball (MLB) teams is essential for optimizing marketing and ticketing strategies. This study explores the relationship between team performance and home attendance for the 2016 MLB season by developing a multiple linear regression model using performance metrics such as wins, losses, and player statistics. Backwards K-stepwise regression and bootstrapping were applied to improve model accuracy and account for potential data variability. The model identified wins, losses, doubles, and triples as significant predictors of home attendance, with an average error rate of approximately 20%. Residual and Q-Q plots revealed issues with normality, suggesting the need for further refinement of model assumptions. While the model provided valuable insights into factors affecting home attendance, future improvements could include incorporating additional variables, using larger datasets, and exploring advanced machine learning techniques. This study has potential applications in optimizing fan engagement and revenue strategies in professional sports.

Keywords: Home attendance, Major League Baseball, multiple linear regression, team performance, wins, losses, player statistics, bootstrapping, stepwise regression, predictive modeling, sports analytics, marketing strategies, residual analysis, data visualization

---

1. Introduction

Major League Baseball is one of America's most culturally relevant and attended sports. Understanding what drives these attendance figures is crucial for the league's financial stability and growth. Determining what drives attendance can assist teams in optimizing their performance strategies and marketing efforts to attract fans. The study being conducted focuses on the 2016 MLB season, with the primary objective of determining the significant predictors of attendance using team performance metrics such as runs scored, wins, losses, batting averages, and more. League context such as league type and games played is also potentially significant. Additionally, we consider the impact of league-specific rules on attendance, such as the designated hitter rule in the American League. This rule, which allows teams to use a player solely as a batter without requiring them on the field, has drawn much interest and discussion in the world of American baseball.

Our research employs a multiple linear regression model to explore the relationships between various predictors and attendance figures. This model allows us to quantify the influence of performance metrics and league context on attendance, providing valuable insights for teams. Moreover, we conduct a two-sample-population hypothesis test to compare average runs scored between the American and National Leagues, offering a further understanding of league-specific attendance drivers. By employing these statistical methods, we can truly understand what attracts fans to attend MLB Games and find ways that teams can optimize their strategies and ultimately enhance the overall experience for their fans.

2. Materials and Methods

2.1 Dataset Acquisition and Data Cleaning

The analysis begins with the acquisition of the `attendance_2016.txt` dataset, which contains data for Major League Baseball (MLB) teams during the 2016 season. This dataset includes various team performance statistics such as wins, losses, runs, and batting averages, as well as attendance data for both home and road games.

To prepare the data for analysis, we perform data cleaning by removing columns that do not provide predictive value for the response variable, specifically, the `TEAM`, `League`, `Home.Attendance`, `Road.Attendance`, `Home.Games`, `Road.Games`, and `League` columns. These variables are excluded because they either represent categorical variables (which are not needed for regression analysis) or are totals of the response variables, which do not contribute to understanding the underlying factors that influence attendance.

2.2 Dataframe Creation and Renaming

After cleaning the data, we create three separate dataframes to focus on different aspects of attendance:

- Home Average Attendance (`Home.Avg.Att` as the response variable)
- Road Average Attendance (`Road.Avg.Att` as the response variable)
- Overall Average Attendance (`Overall.Avg.Att` as the response variable)

In each of these new dataframes, the respective attendance column is set as the dependent variable (`y`), and all other relevant team performance statistics are used as independent variables (`x`). To make the models and subsequent analysis easier to interpret, these columns are clearly labeled with the respective variable names.

2.3. Exploratory Data Analysis (EDA)

2.3.1 Normality Check

We begin the exploratory data analysis (EDA) by assessing the normality of the distribution of the response variables (home, road, and overall average attendance). This is done by plotting:

- Histogram plots: To visually check the distribution.
- Q-Q plots: Using `qqnorm()` to assess if the data follow a normal distribution. If the points on the plot closely follow a straight line, the data can be considered approximately normal. The `qqline()` function adds a reference line to this plot.

2.3.2 Correlation Matrix

To quantify the relationships between the predictor variables and the response variable, we compute the correlation matrix using `cor()`. This provides a numerical assessment of the strength and direction of linear relationships between the variables. A higher correlation value between a predictor and the response variable suggests a stronger influence.

2.4 Model Building and Selection

2.4.1 Multiple Linear Regression (MLR)

The next step involves building a multiple linear regression (LM) model using the `lm()` function. This model estimates the relationships between the predictor variables and the response variable. The goal is to understand how various team performance statistics influence the different attendance measures (home, road, overall).

2.4.2 Stepwise Forward and Backward Regression

To optimize the model, we perform stepwise forward regression. This method starts with a model that includes no predictors and progressively adds predictors based on their statistical significance (p-value). The process continues until no further significant predictors can be added. Stepwise regression is useful for:

- Model simplification: It reduces the model to include only the most important predictors, helping avoid overfitting.
- Predictor selection: It helps identify which predictor variables most significantly contribute to explaining the variance in the response variable.

2.4.3 Best Subsets Regression

We also apply best subsets regression to evaluate different combinations of predictor variables. This method tests all possible combinations of predictors and identifies the subset of variables that provides the best fit for the response variable. Best subsets regression is useful for:

- Identifying optimal predictors: It allows us to find the combination of variables that gives the best explanatory power.
- Model comparison: By comparing models with different subsets of predictors, we can select the most efficient and effective model.

We visualize the results of the stepwise regression and best subsets regression using plots to assess model fit and variable selection.
***Note:** Best Subsets Regression was unable to generate on any of the authors computing machines. The code will be provided within the R document.

2.4.4 Bootstrapping

Bootstrapping is a resampling technique used to assess the variability and stability of statistical estimates by repeatedly sampling from the observed data with replacement. In this study, bootstrapping was applied to the multiple linear regression model to generate a distribution of regression coefficients and to estimate the confidence intervals of the predictors. This approach helped mitigate the impact of small sample sizes and provided a more robust estimate of the model's stability. By resampling the data 1,000 times, bootstrapping enabled the evaluation of the precision of the coefficient estimates, offering insight into the reliability of the predictors in relation to home attendance.

2.5. Building Final Models

Based on the results of the stepwise forward regression and bootstrapping, we create final models using the most significant predictor variables. These models are expected to explain the variance in the response variables (attendance) more effectively. After building these models, we summarize the inferences drawn from the models, including the influence of key predictors on team attendance.

2.7. Model Evaluation and Final Inferences

Finally, the models are evaluated based on their goodness-of-fit measures (such as R-squared and adjusted R-squared), and we summarize the inferences from the regression models. We interpret the coefficients to understand the impact of each predictor variable (such as team performance metrics) on the attendance variables. These final models can be used to predict team attendance based on various performance metrics for future MLB seasons.

3. Results

3.1 Differences in League Runs and Home Attendance

The league rules on a designated hitter influenced the scope of this research. To understand the relationship this has between the American League and the National League and the number of runs scored and conversely the attendance for home games.

Independent t-tests were deployed between the two leagues and runs to understand if there was a difference in runs scored between the two leagues. The assumptions to be tested were the normality of the variables and the homogeneity of variances. In Table 1, using the Shapiro Wilks test of normality the runs scored in the AL computed a p-value of 0.1979 and the runs scored in the NL computed a p-value of 0.9732. Both pass the threshold of 0.05 significance and support the idea that the data are normally distributed.

Table 1. Shapiro Wilks test of Normality between Leagues for Runs scored

League	**American**	**National**
p-value	0.1979	0.9732

Next, the variance was computed between both leagues and the runs scored. A p-value of 0.65 was computed indicating that the variances between the data were approximately equal. Therefore, an independent t-test will be used.

After testing, a p-value of 0.5519, and confidence intervals [-0.197, 0.361] were computed. This p-value > 0.05 indicates that there is no difference between runs scored in the AL and NL. This is also supported by the confidence intervals containing 0. This means there is 0 difference between runs scored in both leagues.

To test whether or not home game attendance differed between both leagues. The same process was employed. Assumptions of normality and variance testing were deployed.

Table 2. Shapiro Wilks test of Normality between Leagues for Home Attendance

League	**American**	**National**
p-value	0.8938	0.1812

As seen in Table 2, a Shapiro-Wilks test computed that the p-values for the American League and National League home attendance normality hypothesis were both greater than 0.05 indicating that both sets of data were normally distributed.

The variance was computed with the Levene test once again and a p-value of 0.9295 was computed. This indicates that the data variances were equal. Because of these assumptions passing an independent t-test could be deployed once again.

After testing, a p-value of 0.3762, and confidence intervals of [-8360.64, 3529.44] were observed. Once again the p-value was greater than 0.05 supporting the idea that there is no difference between attendance in both leagues.

3.2 Assessing Normality and Correlations within the Data

Firstly, we are going to begin our Exploratory Data Analysis by assessing the normality of the distribution of our response variable, Home Average Attendance. How we first attempted to assess the normality by creating a Histogram plot and a QQ plot so that we could visually attempt

to assess normality. The histogram that we created, showed the frequency of distribution of the response variable.

3.2.1 Checking Normality

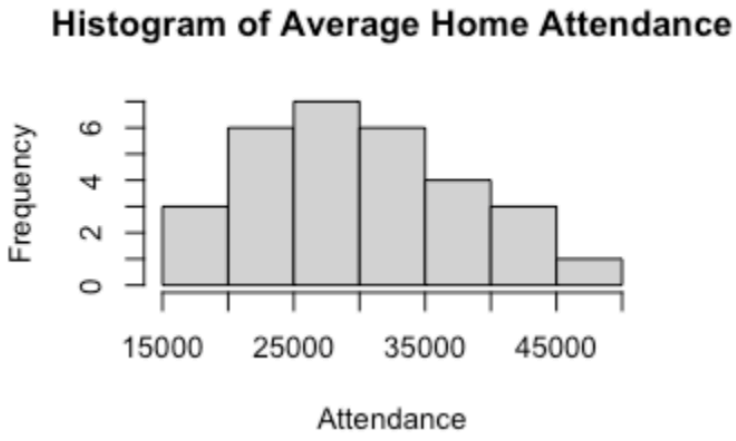

Figure 1. Histogram of Average Home Attendance

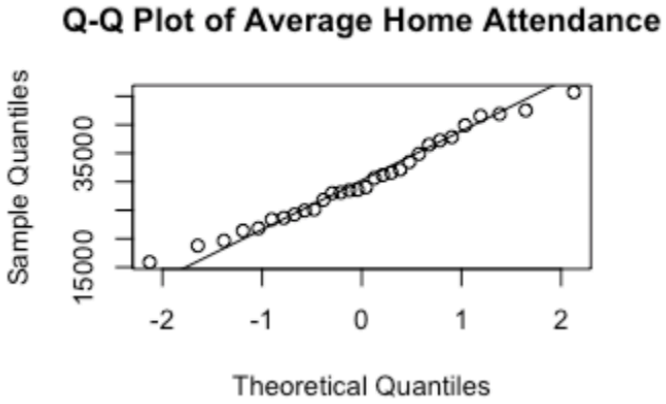

Figure 2. Quantile-Quantile Plot of Average Home Attendance

After visually viewing the histogram in Figure 1, it appeared as if it was approximately normal but it was slightly skewed leaning to the right. So afterward we created a QQ-Plot (Figure 2) that compared the sample quantiles of the data to the theoretical quantiles of the normal distribution. After viewing the plot, most of the points were lying near the line, suggesting that it was normal but on this plot, there were also slight deviations on the tails.

Table 3. Shapiro Wilks Normality Test for Average Home Attendance

Response Variable	**Average Home Attendance**
p-value	0.7578

We wanted to confirm that our response variable, Home Average Attendance, was following a normal distribution so we conducted a Shapiro-Wilk Normality test. If the p-value is greater than 0.05 then we fail to reject the null hypothesis and it means that the data is normally distributed.

After testing, the p-value seen in Table 3 came out to be 0.7578 so this confirmed that our response variable Home Average Attendance was normally distributed.

Now that we have confirmed that our response variable is normally distributed based on both visual assessment and the Shapiro-Wilk test. We can proceed with our objective for this project. We can continue with our multi-linear regression models and with our interpretations and inferences. Confirming that our data follows a normal distribution is a great foundation for our project and it sets the stage for us to produce reliable and valid results.

3.2 Correlation Matrix

Next, we are going to continue our Exploratory Data Analysis by quantifying the relationships between the predictor variables and the response variable. We are going to start by analyzing our first response variable, Home Average Attendance. This can be seen in Figure A1. Our response variable didn't have any significantly strong correlations.

The strongest correlations were pretty moderate. The strongest positive correlation and strongest correlation in general is wins(+.557).

This isn't anything out of the ordinary, fans want to see a winning product on the field. The more that a team wins, the more fans are excited and that equals an influx of fans wanting to attend games, especially home games. The other strong positive correlation variables were On Base Percentage(+.443), Runs Per Game(+.399), and RBI(+.391). These are all variables that pertain to a team's offense. With this information, we can conclude that fans want to see teams that hit well and score a lot of runs.

So it makes sense why teams that have high-powered offenses and their offensive statistics are high, are teams that people want to see. On the other hand, there was also a strong negative correlation and this was losses(-.546). This is also something that makes sense and shouldn't be a surprise. Fans don't want to see teams that are bad and are losing every game. That's why whenever you see that a team is struggling, there are always so many empty seats.

To summarize and interpret this correlation section, overall attendance increases with team success. The winning teams attract more fans both at home and also on the road. Teams that are also high-scoring are entertaining and also boost the overall attendance. The main strong negative correlation is losses, fans are not going to games of teams that are tanking or are losing almost every game they play. The more losses that a team has could reduce overall attendance but this correlation isn't as strong as the positive correlations of wins and offensive performance. Lastly, one of the more important things that we found out by assessing our Correlation Matrix is that wins and high offensive performance are more impactful than losses.

3.3 Model Generation for Predicting Home Attendance

A simple multiple linear regression model was created on trying to predict home attendance based on the predictor variables described earlier.

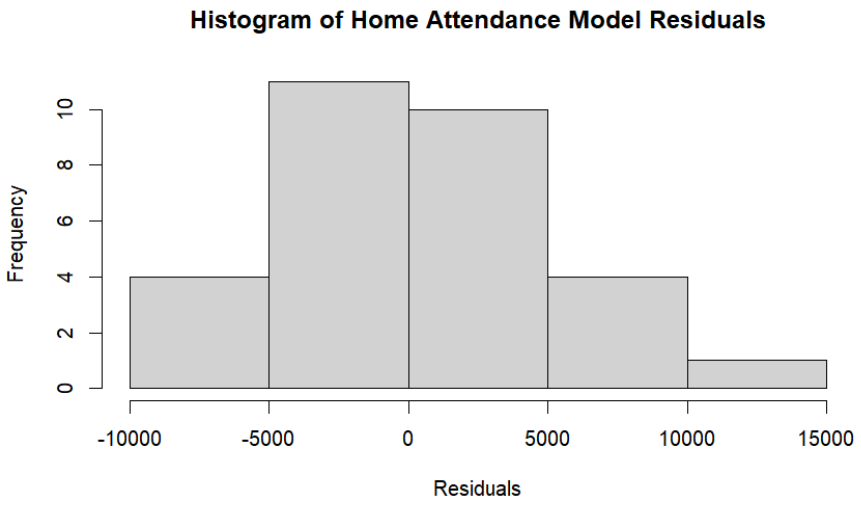

Figure 3. Distribution of first-generation model residuals

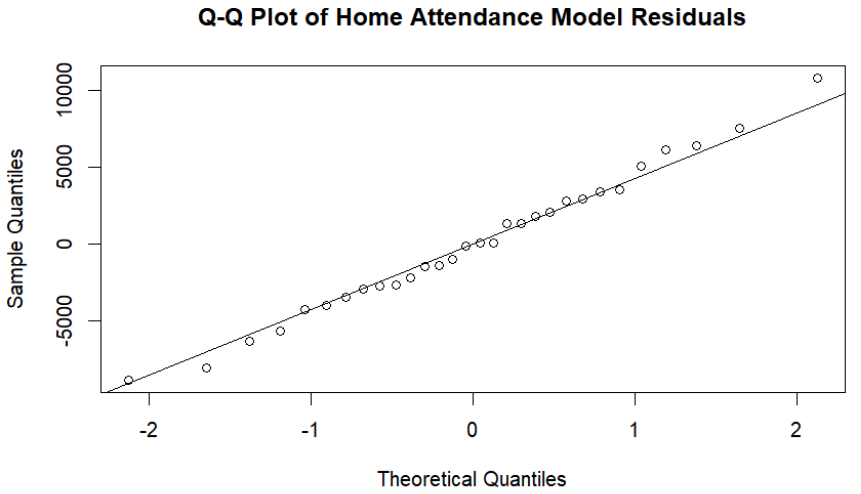

Figure 4. The quantile-quantile plot of first-generation model residuals

Upon first inspecting the histogram of the residuals in Figure 3 for this regression model the data appear to follow a normal distribution. The data points follow along the normal distribution line in Figure 4.

The adjusted R-square for this model came out to be 0.3509. This means that this model captured about 35.09% of the variability in the data. This is a poor model because it isn't able to differentiate between the changes in predictor variables to produce an accuracy response variable. As well as this the predictor variables of significance in this model contributing to predicting home attendance were won and lost with p-values lower than 0.05. These were the only variables to be significant in this bare of a model so they will be important to watch later.

3.3.1 K-stepwise Regression

To improve this model forward, backward and both k-stepwise regressions were used to determine the best model to use based on the first generation model.

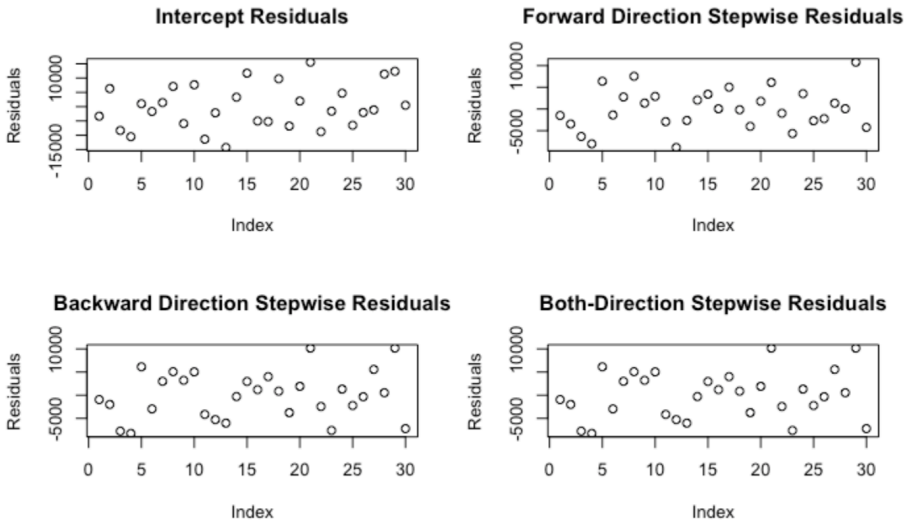

Figure 5. K-Stepwise Regression Residual graphs for intercept, forward, backward and both regression models. **(a)** Intercept residual graph. (b) Forward regression residuals graph. (c) Backward regression residuals graph. (d) Both regression residuals graph

Visualizing the residual graphs in Figure 5, graph (c) and (d) appear to be the same. This interpretation declares that when both stepwise regression was taking place, backwards k-stepwise regression produced the better performing model.

In Table 4 this is supported by the AIC values obtained from each model. An AIC or (Akaike Information Criterion value is an estimate for prediction error of a model. A lower AIC indicates a model with less prediction error and thus a better performing model. According to Table 4, backwards stepwise regression produced the best performing model at an AIC of 613.65.

Table 4. AIC values for home attendance k-stepwise regression models.

K-stepwise Dir.	**Forward**	**Backward**	**Both**
AIC	620.58	613.65	613.65

The model obtained from backwards stepwise regression came out to be:

Formula 1. home_avg_att = won + lost + x2b + x3b + hr + avg + ops

Using this model structure a new multiple linear regression model was generated.

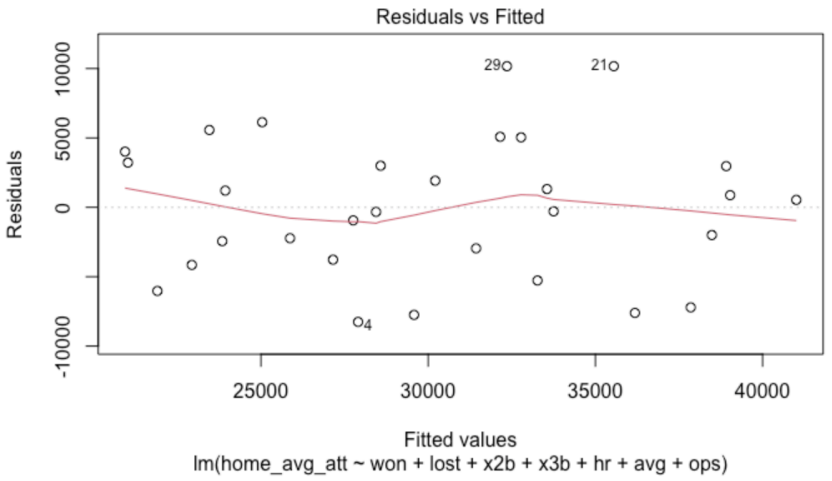

Figure 6. Residuals vs. Fitted graph of backwards stepwise regression model structure.

The residuals of this model appear to follow the normal distribution line in Figure 6. According to the summary of this model. It computed an adjusted R-square of 0.4409. This indicates that this model captures 44.09% of the variability in the data. This is better than the previous model as it captures more differentiation in the data.

The significant predictor variables in this model computed to be won and ops for on base plus slugging percentage. These had p-values less than 0.05 indicating their significance in predicting home attendance based on a backwards stepwise regression model.

3.3.2 Assumptions of the Model

The assumptions for a multiple linear regression model are normality, homoscedasticity, independence and linearity. Based on Figure 6, the data appear to be scattered randomly around 0 indicating linearity in the backwards regression model.

Table 5. Model assumption values for backwards stepwise regression model of predicting home attendance.

Assumption Test	**Shapiro-Wilks**	**Breusch-Pagan**	**Durbin-Watson**
p-value	0.5846	0.833	0.4213
Test value	W = 0.972	BP = 3.521	DW = 1.8716

According to the statistical tests used to check these assumptions all were met. A p-value of 0.5846 in the Shapiro-Wilks test is greater than 0.05 indicating the data are approximately normal. In the Breusch-Pagan test for homoscedasticity, a p-value of 0.833 is also greater than 0.05 indicating homoscedasticity. A p-value in the Durbin-Watson test for independence of 0.4213 is also greater than 0.05 showing that the model observations were independent.

3.3.3 Bootstrapping for Small Sample Size

Because of such a small sample size at only 30 observations, bootstrapping was incorporated into the backwards stepwise model in hopes to produce a more reliable and consistent model Along with this cross-validation was used to assess prediction accuracy of this model afterwards.

After 1000 resamples of the backwards stepwise model the following table was produced which gave the average model statistics for each coefficient during the 100 resamples as well as the standard error of these measures. These errors could be used to compute the confidence intervals of the coefficients. Any confidence intervals containing 0 would be considered not significant

Table 7. Bootstrapping table for backwards stepwise regression model

Coefficients	Original	Bias	Std. Error	**Conf. Intervals**
Intercept	-1803353.26	-87766.53	662679.03	**[-2.46603229*^6, -1.14067423*^6]**
won	10878.74	531.24	3964.97	**[6913.77, 14843.71]**
lost	10578.9	536.26	3954.52	**[6624.38, 14533.42]**
x2b	-133.99	-1.37	107.02	**[-241.01, -26.97]**
x3b	-195.29	-25.55	161.34	**[-356.63, -33.95]**
hr	-361.93	-21.11	139.72	**[-501.65, -222.21]**
avg	-764463.24	-77659.7	324803.82	**[-1.08926706*^6, -439659.42]**
ops	545522.91	36831.35	235125.62	**[310397.29, 780648.53]**

According to Table 7, all of the coefficients during bootstrapping did not contain 0 after average values were +/- with the standard error for each coefficient. This indicates that all of the coefficients were significant in the bootstrapping process. Below in Figure 7 the distribution of the coefficient values can be seen to appear normally distributed. This is also supported by the Q-Q plot of the bootstrapping model in Figure 8 as well the data appears to follow the normal distribution line closely. This indicates a stronger model than before due to the use of samples of the dataset.

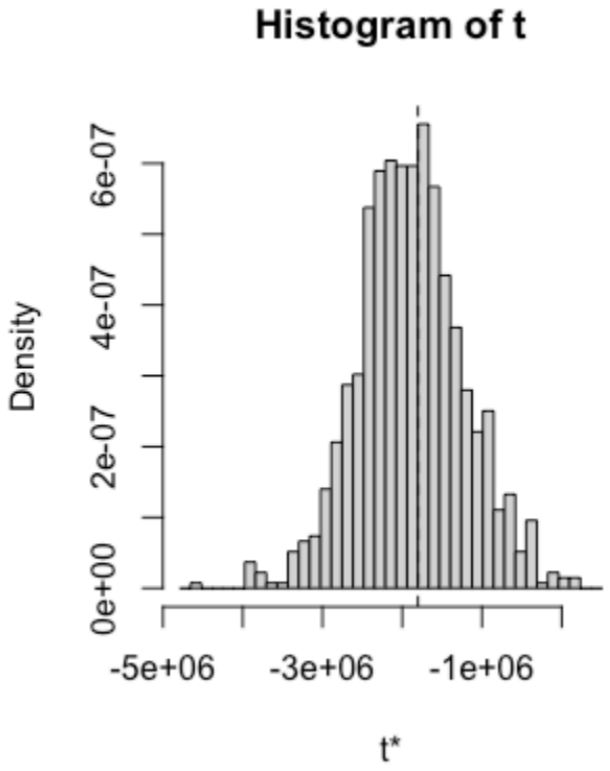

Figure 7. Histogram of coefficients after bootstrapping resampling

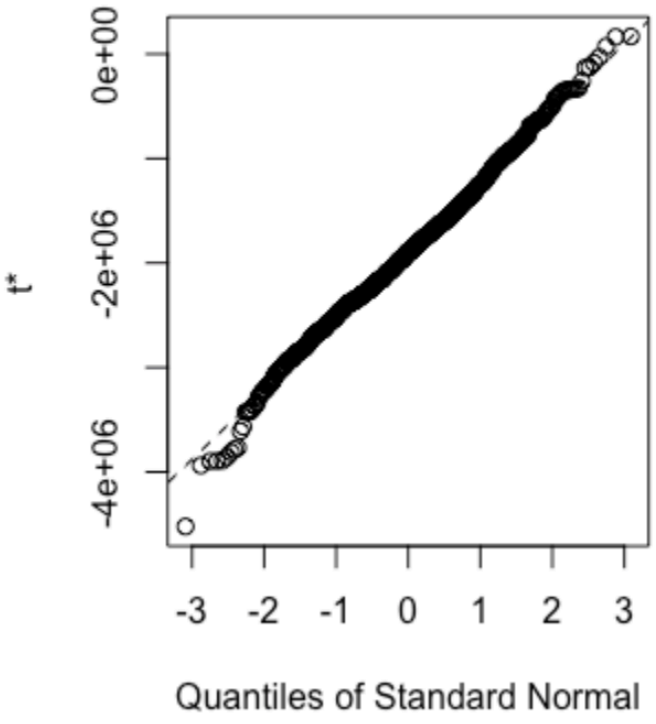

Figure 8. Quantile plot of the coefficients after bootstrap resampling showing a linear relationship between the model.

3.4 Prediction Accuracy

To assess the prediction accuracy of this new improved model, 10-fold cross validation was deployed to train the model on 9 folds of training data from the dataset to test on 1 fold from the dataset. To improve the accuracy 30 samples of cross validation were computed and the mean squared error was produced.

The RMSE or root mean squared error which is a measure that calculates the average error margin for predictions by the model from the correct home attendance line was calculated to be 4957.41. After collecting the range for home attendance which came roughly out to be ~24,000. The average error rate could be computed by:

Formula 2. 4957.41/24,000 = 0.207 x 100 = **20.7% average error rate**

This value of 20.7% means that rate of correct prediction would occur 20.7% of times within the dataset. A lower value is always sought after but for a small dataset size with limited predictor variables this is extremely improved from what the first generation model was.

This model may experience inefficiency from lack of data but with more seasons of data as well as more factors including weather, time of the data/season. The model accuracy may improve dramatically. Predicting average home attendance based on this model could give teams opportunities to plan accordingly during the season in terms of merchandise selling, staff allocation during games, etc.

3.5 What Contributes to Wins?

Based on the final model produced to predict home attendance new questions began to arise. First of all it seemed that wins appeared to be a dominant factor variable in determining home attendance. This makes sense. As a team tends to win more fans would tend to show up to home games in support.

3.5.1 Model Creation

A multiple linear regression model to determine wins was created to explore the relationship between wins and statistics that contribute to the game itself. A new model using wins as the response variable and all others variables besides lost as the predictor variables was created.

According to its output the predictor variables of significance from the first generation came out to be x2b (doubles): 0.0334 , x3b (triples): 0.0394 and slg (slugging percentage): 0.0445. These variables indicate that wins in this first model are most influenced by teams that can get on base rather than simply home runs.

At first glance this appears flawed as runs contribute to winning games directly. However, while runs score points, base hits allow teams the opportunity for more points to be scored during an inning. It keeps the offense on the field batting longer giving them more chances to score runs as well as allows the opposing team to fatigue.

Figure 9. Residuals histogram of first generation wins model.

According to Figure 9 the model data appears to skew slightly left. This may indicate outliers in the data and may violate normality of the data.

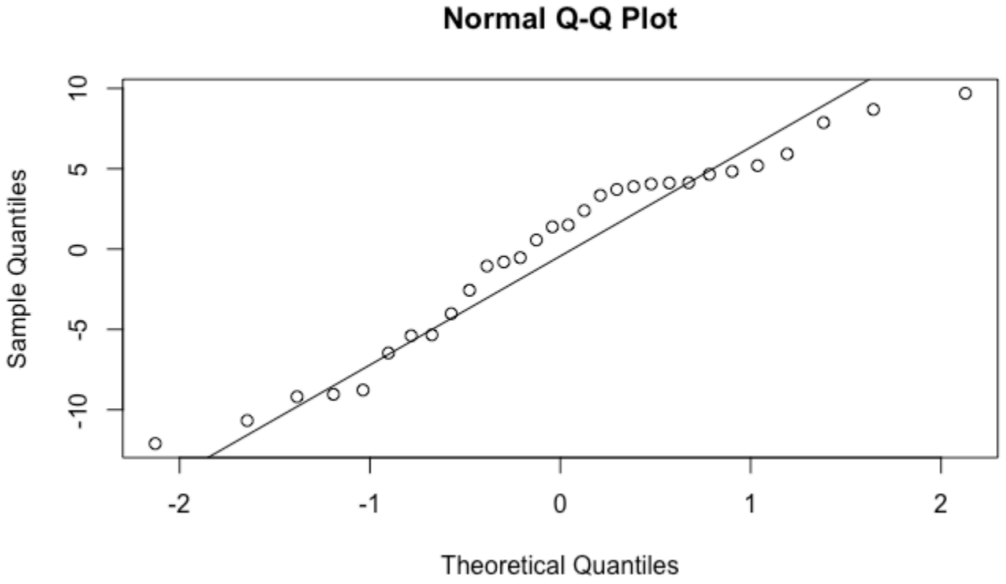

Figure 10. Quantile plot of residuals from first generation wins model.

According to the Q-Q plot in Figure 10, the assumption of normality appears to be violated. The data seems to not follow the normal distribution line as closely as one would hope and tend to deviate from the line at the tail ends indicating the model data is not normal.

3.5.1 K-Stepwise Regression

In hopes of producing a more appropriate model k-stepwise forward, backward and both regression was deployed for the wins model.

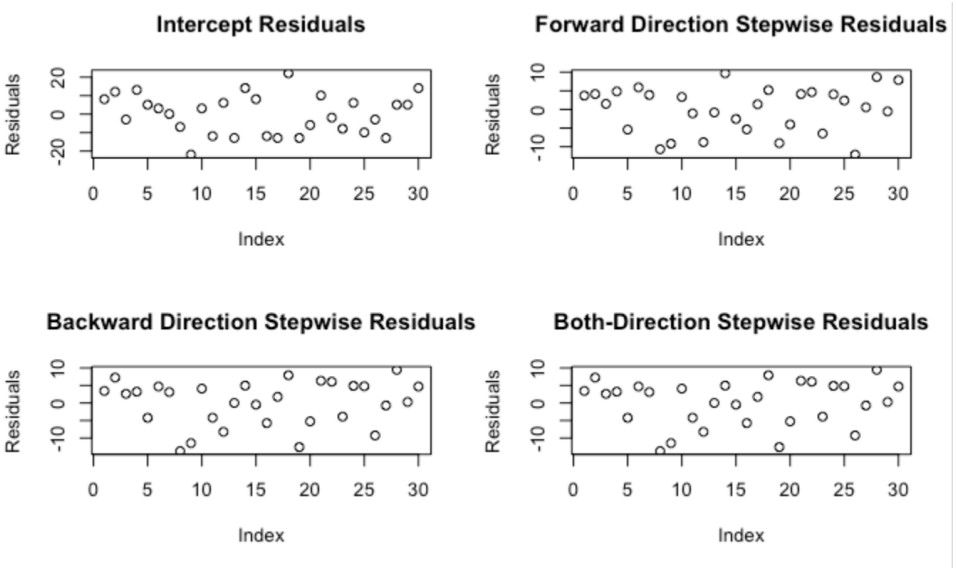

Figure 11. K-stepwise residual graphs for forwards, backwards and wins models. **(a)** Intercept residual graph. (b) Forward regression residuals graph. (c) Backward regression residuals graph. (d) Both regression residuals graph.

The residual graphs for the k-stepwise process can be observed in Figure 11. Once again the backwards and both direction stepwise graphs appear to be the same indicating that backwards direction stepwise produced the better performing model.

Table 8. AIC values for wins k-stepwise regression models.

K-stepwise Dir.	**Forward**	**Backward**	**Both**
AIC	218.28	213.79	213.79

In table 8, once again the AIC values were computed to determine the best model. In support of the interpretation of Figure 11, the backwards stepwise model produced the lower AIC value of 213.79 showing it was the better performing model.

The model obtained from backwards direction stepwise regression came out to be:

Formula 3. wins = ab + runs_per_game + h + x2b + x3b + hr + slg

This new model was used for multiple linear regression and the following residuals graph was produced.

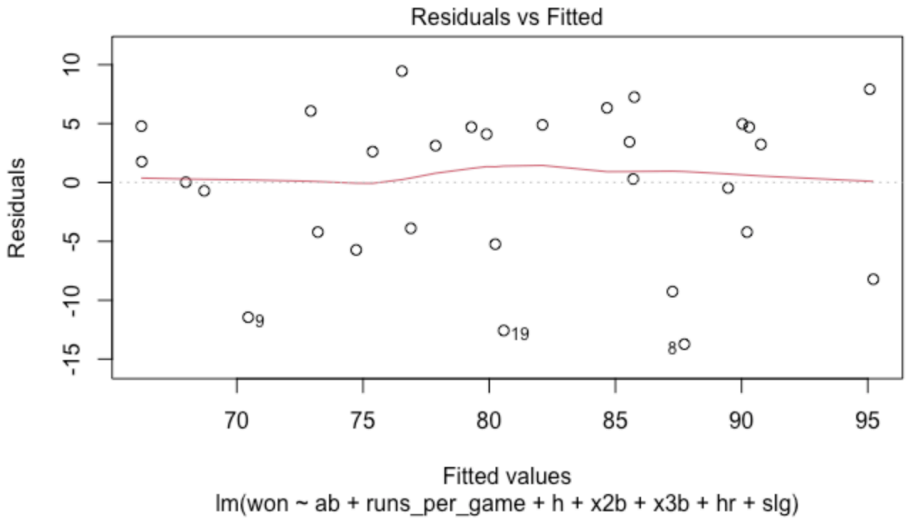

Figure 12. Residuals vs Fitted graph for k-stepwise regression model.

We can see that the data for this model appear to be slightly not normally distributed along the normal residuals line. This may still indicate outliers. Validation of the model assumptions will be needed to verify this.

The predictor variables of significance from this were runs_per_game: 0.0001, x2b (doubles): 0.028 and slg (slugging percentage): 0.011. The variables seem to contribute heavily to wins in this small dataset sample.

3.5.2 Assumptions of Wins Model

Once again the assumptions for a multiple linear regression model are normality, homoscedasticity, independence and linearity. Based on Figure 12, the data appear to be scattered randomly around 0 indicating linearity in the backwards regression model.

Table 9. Model assumption values for backwards stepwise regression model of predicting wins

Assumption Test	**Shapiro-Wilks**	**Breusch-Pagan**	**Durbin-Watson**
p-value	0.03691	0.5598	0.4923
Test value	W = 0.925	BP = 5.8292	DW = 1.9143

According to the statistical tests used to check these assumptions, previous thoughts were supported. A p-value of 0.03691 in the Shapiro-Wilks test is less than 0.05 indicating the data are not approximately normal. This may be due to small sample size or outliers in the data. Bootstrapping will be employed to see if the model performance can be increased. In the Breusch-Pagan test for homoscedasticity, a p-value of 0.5598 is also greater than 0.05 indicating homoscedasticity. A p-value in the Durbin-Watson test for independence of 0.4923 is also greater than 0.05 showing that the model observations were independent.

3.5.3 Bootstrapping for Small Sample Size

Because of small sample size of the data and because the assumption of normality was not met, bootstrapping will be deployed on this model to try to increase performance.

After 1000 resamples of the backwards stepwise model the following table was produced which gave the average model statistics for each coefficient during the 100 resamples as well as the standard error of these measures. These errors could be used to compute the confidence intervals of the coefficients. Any confidence intervals containing 0 would be considered not significant

Table 10. Bootstrapping table for backwards stepwise regression model

Coefficients	Original	Bias	Std. Error	**Conf. Intervals**
Intercept	-3934.20	-156.940	1769.44	**[-5703.64, -2164.76]**
ab	0.7096	0.030	0.320	**[0.3896, 1.0296]**
runs_per_game	29.16	0.890	9.333	**[19.827, 38.493]**
h	-1.7	-0.071	0.765	**[-2.465, -0.935]**
x2b	-1.934	-0.094	0.809	**[-2.743, -1.125]**
x3b	-3.76	-0.094	1.574	**[-5.334, -2.186]**
hr	-5.261	-0.189	2.332	**[-7.593, -2.929]**
slg	9547.013	361.053	4266.34	**[5280.673, 13813.353]**

According to Table 10 all of the coefficients during bootstrapping did not contain 0 after average values were +/- with the standard error for each coefficient. This indicates that all of the coefficients were significant in the bootstrapping process. Below in Figure 13 the distribution of the coefficient values can be seen to appear normally distributed within the center of the data. However, the model appears to be skewed left and right which shows that the small sample size did in fact include not normal data. There appears to be inconsistent values within the data set that persist even after resampling This is also supported by the Q-Q plot of the bootstrapping model in Figure 14 as well the data appears to follow the normal distribution line closely but trails away from the line at both tails.

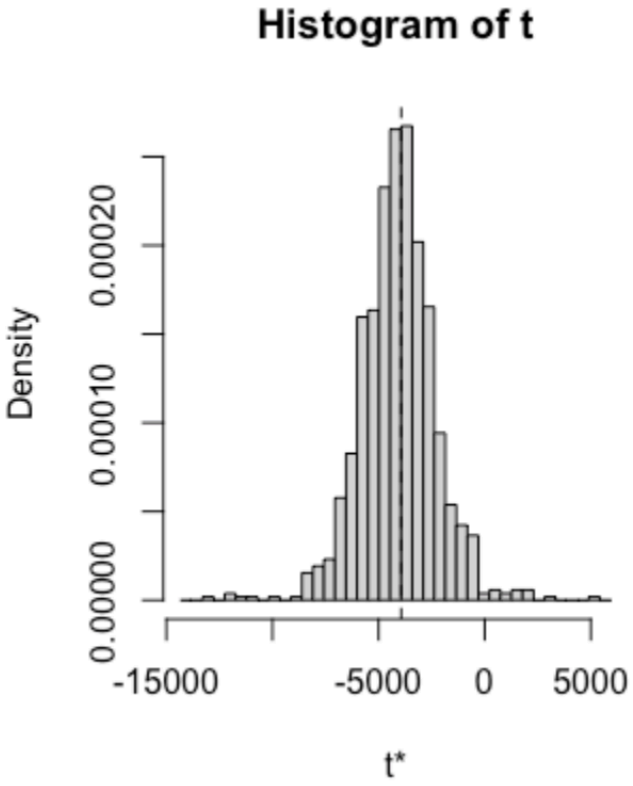

Figure 13. Histogram of coefficients after bootstrapping resampling of wins model

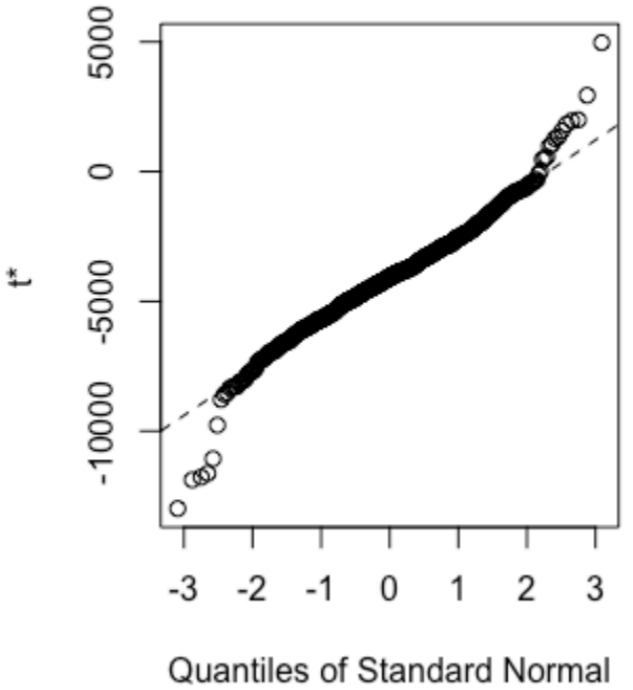

Figure 14. Quantile plot of coefficients after bootstrapping resampling of wins model

3.5.4 Prediction Accuracy

To assess prediction accuracy, 10-fold cross validation was deployed to train the model on 9 folds of training data from the dataset to test on 1 fold from the dataset. To improve the accuracy 30 samples of cross validation were computed and the mean squared error was produced.

The RMSE or root mean squared error which is a measure that calculates the average error margin for predictions by the model from the correct home attendance line was calculated to be 6.323. After collecting the range for home attendance which came roughly out to be 35. The average error rate could be computed by:

Formula 2. 6.323/35 = 0.181 x 100 = **18.1% average error rate**

At an 18.1% average error rate this model appears to do slightly above average in predicting wins based on the metrics provided in Formula 3.

While this model experiences a lack of normality, more data along with the alleviation of outliers may be able to contribute to the model's increased performance. Wins can be attributed to several factors. However, in this model they seem to mostly be attributed to ab, hits and runs per game. This makes sense as winning games is due to the ability to score runs as well as having the ability to score runs in the form of hits and at bats.

4. Discussion

In this project, the goal was to discover if league rules determined whether or not teams scored more as well as gained more home attendance. Through statistical testing using independent t-tests it was observed that neither runs nor home attendance was different between the American League and the National League of the MLB in 2016.

To determine what contributes to home attendance the goal turned into building a reliable model for predicting home attendance for MLB teams in the 2016 season and beyond using a variety of performance metrics. Initially, a general linear regression model was applied to the dataset but its performance was sub-optimal. Despite meeting several key assumptions of multiple linear regression, the model appeared to be a poor predictor of home attendance.

To improve the model backwards k-stepwise regression and bootstrapping techniques were applied due to the model's small sample size. Stepwise regression helped identify the most influential predictors by iteratively adding and removing variables until it produced the model formula seen in Formula 1. These techniques helped identify wins and losses to be the most influential predictor variables when it comes to predicting home attendance. Although the model improved, the predictive error remained at about 20% which is reasonable for a small dataset. This suggests that the model has limitations due to the dataset's size, but it still provided valuable insights into factors influencing attendance.

Further examination of the regression residuals revealed that the wins were the main force in predicting home attendance. Multiple linear regression was deployed on wins to focus on predicting wins based on game statistics which in turn would predict home attendance. The wins model initially suffered from issues with normality as seen in Figure 9 and Figure 10. This was likely caused by outliers in the data as indicated by the plots. Despite this, applying backwards k-stepwise regression and bootstrapping helped slightly mitigate these issues and improve model performance. The wins model showed good prediction ability with an accuracy of approximately 18%, which again, is acceptable given the small sample size from the data.

The results suggest that both wins and losses, as well as team specific performance metrics like, hits, double, triples, runs per game and on base percentage, play a significant role in determining home attendance. However, the non-normality of the residuals and the potential

impact of outliers highlight the importance of carefully validating regression assumptions, especially in small datasets like this one.

While both model's prediction accuracy is reasonable given the dataset size, further improvements could be made by addressing outliers, using a larger dataset, and exploring more advanced modeling techniques, such as regularization or machine learning models, to enhance predictive performance.

5. Conclusion

This study aimed to predict Home Attendance for MLB teams during the 2016 season using various performance metrics. While the models built in this project were able to produce reasonable predictions, there are several opportunities for improvement. The relatively small dataset (30 observations) and the presence of outliers likely limited the model's performance, as residual and Q-Q plot analysis indicated non-normal distributions during win modeling. To enhance model accuracy in future iterations, incorporating additional variables that influence home attendance, such as team demographics, player popularity, stadium capacity, and external factors like ticket pricing and weather conditions, could provide a more comprehensive view of what drives fan attendance. Additionally, using a larger and more recent dataset, potentially spanning multiple seasons, would allow the model to account for evolving trends in fan behavior and team performance, improving the robustness and generalizability of the predictions.

Further improvements could also be made by addressing the data quality, such as outliers and missing values, through more rigorous data cleaning techniques or using robust regression methods. Additionally, employing advanced machine learning models like random forests or gradient boosting could capture more complex relationships between predictors and home attendance, improving prediction accuracy. In terms of application, the models developed in this study can help MLB teams and franchise managers in planning marketing strategies, optimizing ticket pricing, and improving fan engagement by predicting attendance and adjusting strategies accordingly. This study also offers a foundation for further research into fan behavior and could serve as a model for analyzing attendance patterns in other sports leagues. The insights derived from this study have the potential to guide decisions around scheduling, fan loyalty initiatives, and revenue maximization, benefiting both teams and sports organizations in general.

Author Contributions: L.A.M., P.G., S.P. and I.H conceptualized and designed the study. L.A.M. and P.G. wrote the initial introductory draft. I.H. wrote the methodology, provided the visualization, formal analysis and writing of R code and results of the manuscript. S.P. and P.G., L.A.M. and I.H. performed the investigation, writing-review and editing, writing-original draft preparation. All authors have read and agreed to the published version of the manuscript."

Data Availability Statement: The data that supports the findings of this study are available in the "Data and Story Library" at:

https://dasl.datadescription.com/datafile/attendance-2016/?_sfm_methods=Multiple+Regression&_sfm_cases=4+40

Appendix A

Table 1. Correlation between Home Average attendance and baseball metrics from the dataset.

Vars	home_avg_att	ab	won	lost	runs_per_game	h	x2b	x3b	hr	tb	rbi	avg	obp	slg	ops
h_a_a	1	0.021	0.557	-0.546	0.399	0.128	0.666	-0.130	0.183	0.192	0.391	0.129	0.444	0.217	0.328