**Title**
Understanding Within-Season Changes in Major League Baseball Attendance



**Author**
Koscinski, Tanner Michael

**Publication Date**
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

# Understanding Within-Season Changes in Major League Baseball Attendance

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics

by

Tanner Michael Koscinski

2019

ABSTRACT OF THE THESIS

Understanding Within-Season Changes in Major League Baseball Attendance

by

Tanner Michael Koscinski
Master of Applied Statistics
University of California, Los Angeles, 2019
Professor Frederic R. Paik Schoenberg, Chair

Major League Baseball teams play considerably more games per season than the teams in any of the other major professional sports leagues in the United States, providing baseball fans with a large selection of games and frequently resulting in large fluctuations in attendance. In addition to studying the effect of the home and visiting teams' season win percentages on attendance, this paper examines various shorter-term metrics of performance to see which recent performance metric has the most significant impact on a home team's game-to-game attendance numbers. The number of wins in a team's previous 10 home games is shown to be statistically significant and the best recent performance metric to estimate attendance; however, the attendance prediction benefits are not practically much better than a simpler model that ignores such recent performance. This paper also discusses how the magnitude of the variation in a team's attendance numbers can differ from team to team and from season to season and suggests using a team-season standardized attendance response variable to eliminate this difference in variance issue between low average / low variance, high average / high variance, and high average / low variance teams. The models with the team-season standardized response variable consistently outperform models with the traditional attendance response variable.

The thesis of Tanner Michael Koscinski is approved.

Robert L. Gould

Vivian Lew

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# CHAPTER 1

# Introduction

A calculation of the average Major League Baseball game attendance per year reveals that attendance has been declining over the past few years. A decrease in attendance means not only a decrease in ticket sales revenue, but also smaller profits from concessions and parking. The league has its television deals as an additional source of income, but the fans at the stadium make up a vital part of the game. Fans amplify energy and excitement. The emotion created by a home run is more intense in a sold out stadium than in one that is half empty. The fans and the players can both feel the difference, and a large and passionate home crowd can be a great selling point to attract superstar athletes. A loud and supportive crowd can help provide the energy to keep up good momentum or revitalize a team and spark a comeback. This is why the crowd at a baseball game is sometimes collectively referred to as "the 10th man" of the baseball team.

Smith and Groetzinger investigated this attendance benefit and found that an increase in attendance was associated with a statistically significant increase in the number of hits, doubles, and home runs for the home team relative to the visiting team, even after they accounted for changes in attendance due to differences in the likelihood of a home team victory (2010). They also found that higher attendance increased the number of strikeouts thrown and decreased the earned run average for the home pitchers relative to the visiting pitchers. They estimate that an attendance increase of 25% of the stadium capacity results in an increase in the home team's probability of winning the game by approximately 5.5%. Silver estimates that one additional win in a baseball season results in $1.8 million in additional revenue for a team (2006), and Smith and Groetzinger hypothesize that it is actually a smart financial decision to decrease ticket sales in order to increase attendance to increase

the probability of winning more games for home teams that are seeing poor attendance and are able to accommodate the increase in attendance (2010).

Compared to the other three major professional sports leagues in the United States, Major League Baseball (MLB) sees higher variations in attendance. This is largely because MLB teams play 162 games a season, which is about twice as many games as National Basketball Association (NBA) and National Hockey League (NHL) teams (82 games each) and about 10 times as many games as National Football League (NFL) teams (16 games). With such a large supply of games to choose from, and because MLB stadiums are generally much larger than NBA or NHL arenas, MLB games are less likely to sell out. The large number of games and relatively large variation in attendance makes Major League Baseball attendance a good candidate for analysis.

This paper attempts to explore and understand the factors that affect Major League Baseball attendance within a team's season. With an increased understanding of what attracts fans to games, marketing teams will be able to create more effective advertising campaigns. More advanced ticket pricing methods could be developed to increase profits for games that are likely to sell out, and promotions, giveaways, and programs with schools or youth baseball teams can be scheduled on games that are likely to be relatively empty in an effort to maintain an exciting atmosphere and increase profits from concession sales. Games could also be scheduled at times that result in higher attendance, but scheduling to optimize overall television viewers would likely take precedence.

# CHAPTER 2

## Goals and Areas of Interest

There are two main ways to analyze Major League Baseball attendance: between-season (changes in the average attendance from season to season) or within-season (changes in attendance from game to game). The primary goal of this paper is to gain an overall understanding of the factors that affect within-season Major League Baseball attendance, and secondary goals include examining if the significant factors or their effects change when reducing the analysis to a single team and then further reducing the analysis to a single season of a single team. The average effects of game scheduling, opponent, and team performance on attendance figures within a season are explored.

Between-season changes in attendance are not analyzed in this paper, but some factors that may influence average season attendance that typically have no significant effect on game-to-game attendance include the state of the economy, general interest in baseball, and the arrival or departure of superstar athletes (unless there is a mid-season trade). It is also well known that a team's performance over the course of a season has a significant positive impact on the team's average attendance for the season. This paper examines the effect of home and visiting team win percentages on game-to-game attendance and also takes a look at some metrics only concerned with very recent performance.

The differences in the average attendance for each team is likely largely due to location and stadium; this is not explored here, but previous work by Denaux et al. has concluded that the per capita income of the home team's city has a positive effect on attendance (2011).

Regression and mixed-effects models are better suited for providing an understanding of what factors influence attendance figures than less interpretable machine learning methods,

so the better predictive capabilities of machine learning is sacrificed. Attendance prediction accuracy is still of interest, so the accuracy of the created models will be examined. LASSO (least absolute shrinkage and selection operator) is used to optimize variable selection for each case. Simpler models are then created to see if the statistically significant variables are also practically significant. If a simpler model is not practically significantly worse at predicting attendance, then it is favored since it allows for better interpretability.

# CHAPTER 3

# The Dataset

Due to low attendance at their home stadium, the Montreal Expos played over 20 of their "home" games in San Juan, Puerto Rico in both the 2003 and 2004 seasons. In 2005, the Montreal Expos moved to Washington, D.C. and became the Washington Nationals. Since then, several teams have moved to newer stadiums between seasons, but each new stadium has been relatively close to the old one, and no team has played more than three of their "home" games in a stadium other than their usual home stadium. For this reason, 2005 is a good starting point for analyzing attendance.

## 3.1 Obtaining the Data

The website Retrosheet[1] has game logs containing information from every Major League Baseball game played since 1871. All of the game logs from 2005 through 2018 are used for this analysis.

Variables in the dataset include attendance, baseball stadium, home and away team, date, day of week, and if the game took place during the day or at night. In 2015, a matchup between the Chicago White Sox and the Baltimore Orioles was closed off from the public due to riots in Baltimore and had an official attendance of zero; this game was removed from the dataset. Approximately two percent of all games played were part of a double-header. Because this is a relatively small percent, and double-header games are likely to see different or distorted attendance numbers, these games are not included in the analysis. As mentioned

[1]The information used here was obtained free of charge from and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at "www.retrosheet.org".

before, at most three "home" games per season per team were played in a stadium other than the team's usual home stadium; these games were also removed. Only regular season games are analyzed since postseason games naturally see higher attendance and usually sell out; the rare tie-breaker games at the end of the regular season are also not included. Finally, the first true home game of each season for each team was removed because these games typically have festivities and promotions that result in unusually high attendance. The final dataset contains 32,859 games over 14 seasons. The same 30 teams are in each season, and each team played between 68 and 80 home games after the removal of these unusual games.

## 3.2 Definition of Attendance

Major League Baseball requires teams to report attendance as the number of tickets sold rather than the number of people who actually attended the game; this is in part because of the way revenue is shared among teams. For this analysis, this way of reporting attendance is an advantage if the goal is to gain insight into how to increase ticket sales, but this definition is unfortunate when trying to understand how to increase the number of fans that actually show up to games. While there is surely not a perfect correlation between ticket sales and actual attendance, it is safe to assume this is a fairly strong positive correlation, and an increase in ticket sales is highly likely to lead to an increase in the actual attendance. Regardless, increasing ticket sales is very desirable.

## 3.3 Construction of New Variables

The Retrosheet game logs are similar to game box scores and mostly contain detailed information about each game and very little information about team performance prior to each game. Because attendance numbers are only affected by events that occur prior to the beginning of a game, a significant amount of work was put into the construction of new predictor variables. The following variables were constructed for each game:

- The home team's win percentage prior to the start of the game
- The visiting team's win percentage prior to the start of the game
- The home team's win percentage in home games prior to the start of the game
- The visiting team's win percentage in the previous season
- Series game number: the number of consecutive home games that the home team has played against the same visiting team including the current game
- If the home team and the visiting team are in the same league
- If the home team and the visiting team are in the same division
- The home team's number of wins in the previous n games for n = 1, 2, . . . , 10
- The home team's number of wins in the previous n home games for n = 1, 2, . . . , 10
- The home team's number of runs in the previous n games for n = 1, 2, . . . , 10
- The home team's number of runs in the previous n home games for n = 1, 2, . . . , 10
- The home team's number of home runs in the previous n games for n = 1, 2, . . . , 10
- The home team's number of home runs in the previous n home games for n = 1, 2, . . . , 10

Note that the last six items each represent 10 different variables. The 67 new explanatory variables are used to better understand and predict attendance. Many of these new variables are strongly correlated, so it is necessary to determine which are the best and which should be discarded.

# CHAPTER 4

# Initial Models

The effects of the home team (or the baseball stadium), the year, and the interaction between the home team and the year on attendance is not of interest, so they can be treated as random effects. One initial model that is created is a linear mixed-effects regression model with the interaction between the home team and the year as a random effect. There are 30 different teams with 14 seasons each, so the model has 30 x 14 = 420 random effect intercepts.

## 4.1 The Response Variable

It must be decided if the attendance response variable should be transformed or scaled. One might expect stadiums with higher average attendance numbers to have higher variance in attendance numbers; however, the Box-Cox transformation method suggests an untransformed response variable. To understand why higher average attendance does not necessarily correlate with higher attendance variation, imagine a hypothetical scenario. Stadium A has an average attendance of 10,000 people and Stadium B has an average attendance of 50,000 people. If both stadiums never reach maximum capacity, it would be reasonable to expect Stadium B to have higher attendance variation. Now imagine the capacity of Stadium A is 15,000 and the capacity of Stadium B is 51,000. Stadium A averages about 67% capacity while Stadium B averages more than 98% capacity. It is now more reasonable to expect Stadium A to have higher attendance variation. Average attendance has a positive correlation with the variance of attendance, but the percentage of sold out games has a negative correlation with the variance of attendance, resulting in no recommendation for a transformation of the response variable.

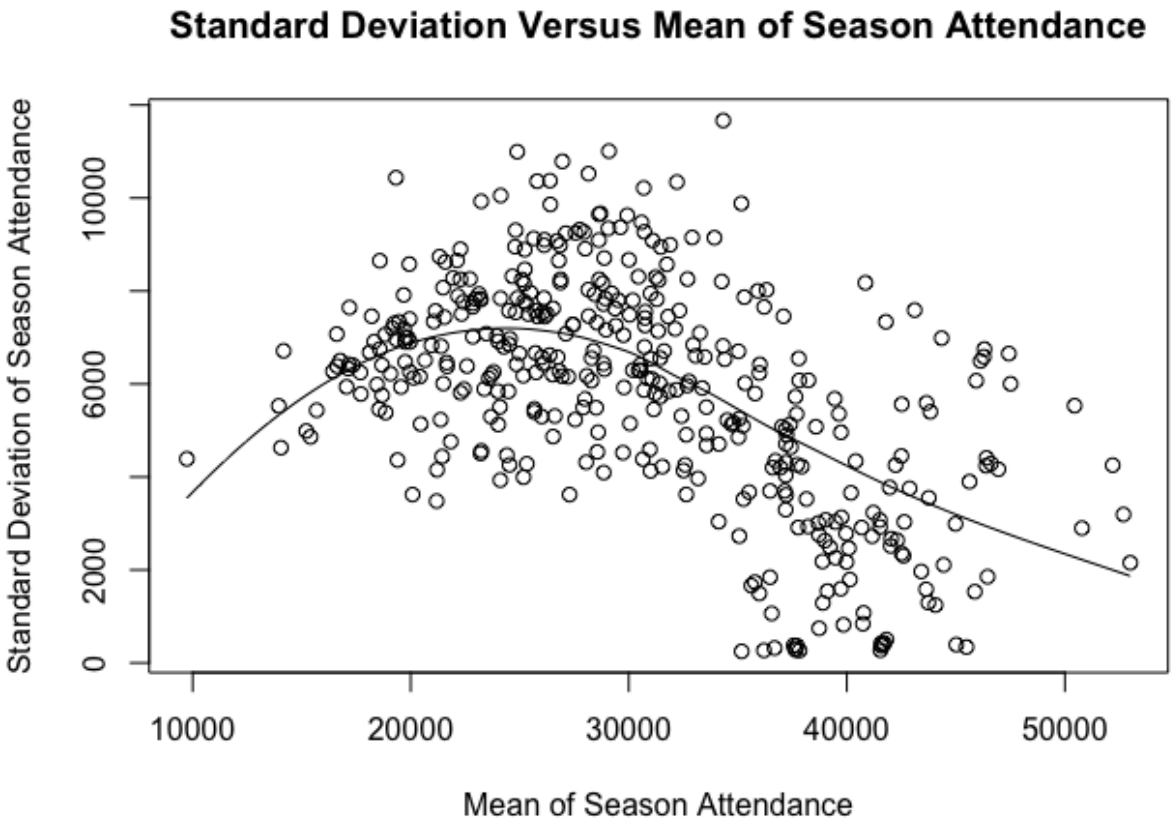


Figure 4.1: The standard deviation of the attendance for each of the 14 seasons of each of the 30 teams is plotted against the mean of the attendance.

Figure 4.1 shows how the standard deviation of season attendance seems to increase as the mean of season attendance increases until a mean of about 30,000 people at which point many of the stadiums begin reaching capacity and the standard deviation of season attendance decreases. Figure 4.2 makes it easier to see these trends by selecting eight of the stadiums studied and plotting them in different colors.

It now may seem reasonable to try to account for stadium capacity by changing the response variable to be attendance as a percent of stadium capacity; however, one can quickly see this is not helpful by looking at the same scenario as before. If a model predicts an attendance increase of 5% of stadium capacity, this is equivalent to an increase from 10,000 to 10,500 for Stadium A and an increase from 50,000 to 52,500 for Stadium B. This

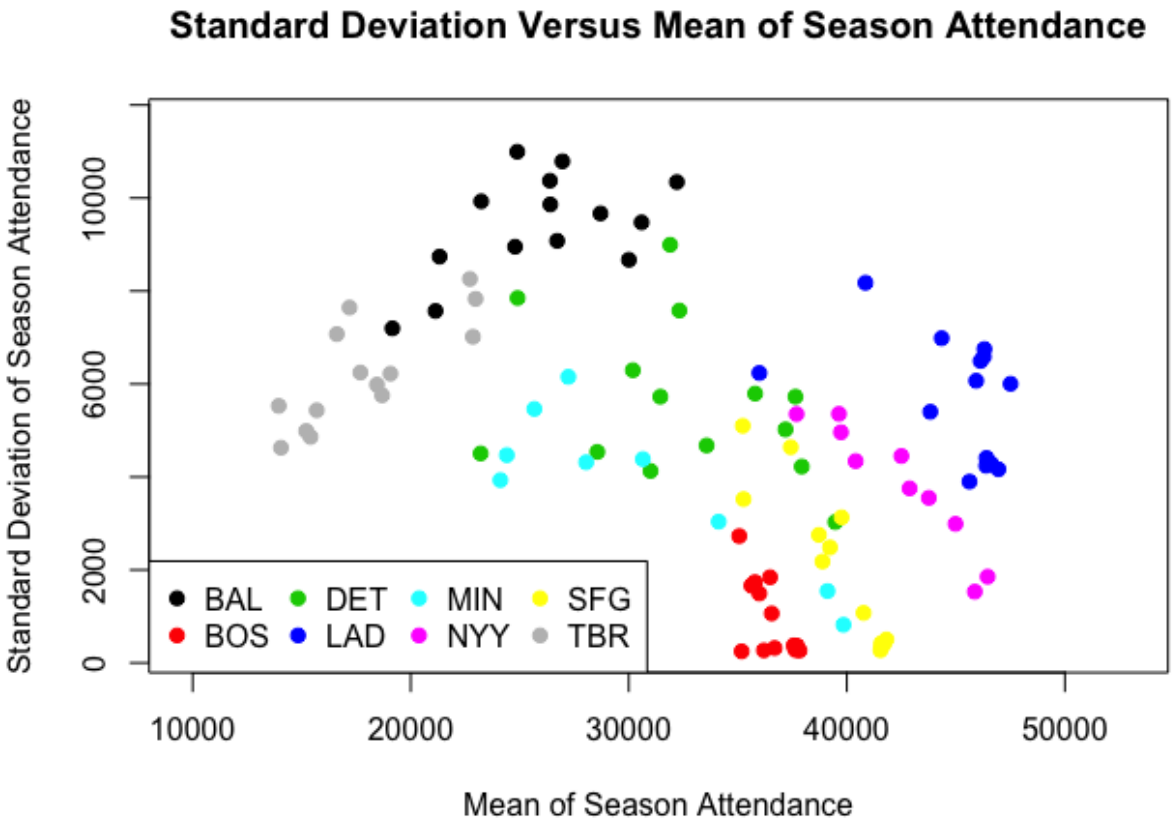


Figure 4.2: The standard deviation of the attendance for each season studied of eight selected stadiums is plotted against the mean of the attendance.

may not be a very significant increase in attendance for Stadium A, but it is five times larger of an increase for Stadium B, and it is 2.5 times the difference between Stadium B's capacity and its average attendance.

An additional issue associated with the use of attendance as a percent of stadium capacity is that it requires knowledge of the stadium capacity. This seems like it should not be an issue, but it turns out that the reported stadium capacity is not always correct. For example, the official Dodger Stadium capacity has always been 56,000, but renovations that occurred before the 2013 season reduced the true capacity to under 54,000, and the exact number is not available. To further complicate matters, sometimes teams sell standing room only tickets or uncover additional seats to go over their usual capacity. It also does not make much sense to consider stadium capacity for teams that never or almost never reach capacity. Changing the response variable to be attendance as a percent of stadium capacity would provide little to no advantage over the raw attendance number.

Most previous research regarding Major League Baseball attendance has decided to use either the raw attendance number, as in Davis (2009) and Denaux et al. (2011), attendance as a percent of stadium capacity, as in Smith and Groetzinger (2010), or the natural logarithm of attendance, as in Ormiston (2004). No method has consistently been used to solve the issue of different attendance variances for different team-seasons. Davis recognized the existence of a censoring issue but decided against trying to account for it and instead noticed it as the probable reason why some variables did not appear to be significant for certain teams (2009). Denaux et al. decided to eliminate all games with an attendance over the stadium's capacity (2011), but this method does not capture any of the many sold out games that do not actually meet the official capacity nor does it reduce the gap between low average / low variance teams and high average / high variance teams. Ormiston determined that treating a 95% capacity game as a sellout was still not capturing some true sellouts (2004), so he resorted to considering an at least 90% capacity game as a sellout and then ignored all team-seasons that were entirely sold out, thereby successfully removing most high average / low variance teams. He then took the natural logarithm of attendance to reduce the gap

between low average / low variance and high average / high variance teams. This is a good solution, but it can lead to a bit of an exaggerated interpretation of the model's coefficients. For example, let Team A and Team B both sell out Stadium 1, but Team A brings in twice the attendance to Stadium 2 as Team B. Ormiston's method makes it appear that we can expect to see, on average, double the attendance when Team A comes to town as opposed to Team B (which is true if neither team is ever expected to sell out a game), but in reality, sold out games will reduce this average attendance gap.

This paper uses a method that attempts to eliminate the difference in variance issue between low average / low variance, high average / high variance, and high average / low variance teams, that does not rely on unreliable stadium capacity numbers, and that does not sacrifice much interpretability.

## 4.2 Standardizing Attendance Within Each Group

An interesting possible solution to the problem of different attendance variances for different home teams and different seasons is to standardize attendance within each season of each home team. Not only does this make the distribution of the response variable much more normally distributed, but this also gives all 420 groups the same mean standardized attendance of zero, eliminating the need for random-effect intercepts. The response variable of this model is the z-score of the attendance of the game with respect to all of the home games in that team's season. This is actually quite easy to interpret and can easily be converted back to a raw attendance number by providing the home team's season's mean and standard deviation of attendance numbers from the training data. Since this model does not require specification of a home team or a season when predicting new data, it also has the benefit of being able to make predictions for arbitrary teams and seasons by producing a z-score that can be converted to raw attendance numbers by providing the theoretical mean and standard deviation of attendance. This method can also make accurate predictions for stadiums that sell out every game; it will produce a z-score that will be multiplied by the standard deviation of zero and then added to the mean attendance consequently predicting a sold-out

game. The mixed-effects model will not be able to handle this scenario nearly as well.

# CHAPTER 5

## Variable Selection

In addition to the variables already in Retrosheet's game logs, 67 predictor variables were created in hopes that some of them would be useful for explaining changes in attendance. Many of these new variables are very similar, and the goal is not to use all of them but rather to find the best small subset of them. In particular, most of the 60 variables of the form (wins / runs / home runs) in the past (1 / 2 / ... / 10) (games / home games) are very highly correlated with one another. These 60 variables will be referred to as "recent performance variables". Ten was chosen as the upper limit for the number of previous games to include because the goal of these variables is to evaluate recent performance, and games further than 10 games back start to represent season performance better than recent performance. The questions surrounding the recent performance variables are:

1. Is the number of wins, runs, or home runs the best predictor of attendance?

2. Is recent home game performance more or less significant in predicting attendance than recent overall performance?

3. What is the optimal number of recent games to consider when estimating the effect of recent performance on attendance?

The 60 recent performance variables add a unique element to this analysis of within-season attendance. The two most popular ways to evaluate team performance within a season have been by using either the team's win percentage or the number of games a team is above or below a .500 win-loss record. (For example, a team with a 22-24 record is two games behind.) The first method has higher variability in the beginning of the season while the

second method has higher variability at the end of the season. To account for the early season variability in win percentage, Meehan et al. recommended ignoring the first 10 games of the season (2007), but that is not very desirable, so this paper instead uses the team's previous season's win percentage as a substitute win percentage for the first 10 games. The variables that are considered that relate to win percentage are the home team's win percentage, the home team's win percentage at home, the visiting team's win percentage, and the visiting team's win percentage in the previous season.

The effect of game scheduling within the baseball season is predicted by the month, the game number of the season, or both of these variables.

The visiting team likely affects attendance, and interest lies in answering the following:

1. Which teams draw the biggest and smallest crowds on the road?

2. Is attendance similar or different when playing a team in the same division compared to a different division?

3. Is attendance similar or different when playing a team in the same league compared to the other league?

Other variables that are be tested for significance include the series game number, the day of the week, the day or night indicator, and the interaction between the day of the week and the day or night indicator.

LASSO (least absolute shrinkage and selection operator) was used for variable selection to determine which variables are significant predictors of attendance overall, within each stadium, and within each season of each team. All of the variables just described were available for selection, and the response variable was the standardized team-season attendance variable described in the previous section. Ten-fold cross validation was used to minimize the mean squared error, and variables were considered significant if they had a non-zero coefficient in the model resulting from choosing the largest value of lambda such that the error was within two standard errors of the minimum mean cross-validated error. Categor-

ical variables were considered significant if at least one of their levels was significant; for example, month is considered significant if June is significant even if May is not significant.

## 5.1 Variable Selection for One Comprehensive Model

Variable selection was performed on the full set of home games from all 30 teams and all 14 years to identify the factors that affect Major League Baseball attendance within a team's season regardless of the home team and season. The following variables were found to be significant predictors of attendance:

- Month
- Day of week
- Day or night
- The interaction between day of week and day or night
- Game number
- Visiting team
- If the teams are in the same division
- If the teams are in the same league
- Home team's win percentage
- Visiting team's win percentage
- Visiting team's win percentage in the previous season
- Home team's number of wins in the last nine home games
- Home team's number of wins in the last 10 home games

The questions surrounding the recent performance variables can now be answered. In general, it appears that the number of recent wins is a better predictor of attendance than the number of recent runs or home runs. Nine or 10 appears to be a good number of games to consider when evaluating a team's recent performance. Recent performance in home games appears to be a better predictor of attendance than recent performance in both home and away games, but the team's overall win percentage in both home and away games is more significant than its win percentage in home games only.

The visiting team significantly affects attendance, and it does matter if the visiting team is in the same division or the same league as the home team. The visiting team's win percentages in the current season and in the previous season both positively affect attendance. The New York Yankees as visitors had the largest positive effect on attendance followed by the Boston Red Sox, Chicago Cubs, and Los Angeles Dodgers. These four teams come from four of the most populated cities in the United States, so it is difficult to say how much of their popularity is due to their city and how much is due to their success. Game scheduling significantly affects attendance. The game number, month, day of week, time of day, and interaction between the day of week and time of day are all significant.

The series game number does not seem to impact attendance.

## 5.2 Variable Selection for Different Stadium Models

Between the end of the 2005 baseball season and the beginning of the 2018 baseball season, seven of the 30 teams in the MLB moved to a different stadium, so 37 different stadiums were used as the home baseball stadium for a team for at least one season over this 14-year span. It is possible that different stadiums see factors affect their attendance in different ways. In particular, it is likely that the effect of the month varies between stadiums because different parts of the country see different weather patterns over the course of the year. Also, the effect of the visiting team varies between stadium, especially because of rivalry games. While having one comprehensive model is better for providing an overall understanding of

attendance, a separate model for each stadium is more practically useful because individual stadiums are more concerned about how to increase their own attendance rather than MLB attendance as a whole. For this reason, it is interesting to perform a separate variable selection for a model of each stadium.

The stadium-specific variable selection results were similar to the variable selection results of the comprehensive model. Almost all of the 13 variables that were selected in the comprehensive model were part of the 12 most frequently included variables in the stadium models. The number of wins in the last eight home games appeared more frequently in the stadium models in place of the number of wins in the last nine home games, but the number of wins in the last 10 home games remained the most significant recent performance variable. The comprehensive model included the variable identifying if the two teams are in the same division, but the stadium models did not consider it significant because it no longer explains any information that is not already explained by the visiting team.

## 5.3 Variable Selection for Different Team-Season Models

It may not be very reliable to perform separate variable selection for each of the 420 combinations of home team and season because each dataset will only contain between 68 and 80 observations, but a quick examination revealed that the 10 most frequently selected variables remained the same as before, and the home team's number of wins in the previous 10 home games remained in the top two most frequently selected recent performance variables.

# CHAPTER 6

# The Regression Models

Based on the variable selection results, models were created with the following 12 variables:

- Month
- Day of week
- Day or night
- The interaction between day of week and day or night
- Game number
- Visiting team
- If the teams are in the same division
- If the teams are in the same league
- Home team's win percentage
- Visiting team's win percentage
- Visiting team's win percentage in the previous season
- Home team's number of wins in the last 10 home games

A linear regression model with the team-season standardized attendance response variable was created as well as a linear mixed-effects regression model with the raw attendance number as the response variable and the home team and season interaction as a random effect. The

MAPE (Mean Absolute Percentage Error) of each model was calculated, and the performance of the models were compared.

## 6.1 The Initial Models

The dataset was randomly split 80/20 into training and testing data. A linear regression model with the team-season standardized attendance as the response was fit on the training data with all 12 predictor variables. The observed MAPE on the testing data was 12.92%. A linear mixed-effects regression model with the raw attendance number as the response and the home team and year interaction as a random effect was fit on the same training data with the same 12 predictor variables, and the observed MAPE on the testing data was 14.05%. The linear regression model was able to outperform the mixed-effects model because the linear regression model accounted for the different variances in attendance for each team and each season.

## 6.2 Ignoring Wins in the Last 10 Home Games

A drawback of including the wins in the last 10 home games variable in the model is that one must know the result of the last 10 home games for the home team's season. This may not only be inconvenient to calculate, but it also imposes the requirement that there must be 10 previous home games in the season. Since the purpose of including this variable in a model is to gain insight into a team's recent performance, it would be misleading to include games from the previous season in the calculation of this variable. Consequently, the first 10 home games of each season for each team will not be able to be included in the models because these games do not have a value for this variable; this also means one cannot predict the attendance of the first 10 home games of each season. For this reason, it is very desirable to be able to drop any recent performance variable from a model if it does not significantly improve the predictive capabilities of the model. The ANOVA table comparing the linear regression models with and without the wins in the last 10 home games variable has a p-

Table 6.1: Prediction Accuracy With and Without Wins in Last 10 Home Games

| | **Standardized Attendance** | **Attendance** |
|---|---|---|
| **With Variable** | 12.92055% | 14.05396% |
| **Without Variable** | 12.92573% | 14.05457% |

value of nearly zero, indicating that this variable is highly statistically significant and should probably not be removed from the model; however, it is also important to consider practical significance. Table 6.1 shows the MAPE for the testing data for both the linear regression team-season standardized attendance model and the mixed-effects attendance model with and without this variable.

Both models saw a MAPE increase of less than 0.01%. This is not practically significant, so this variable can be removed from the models.

## 6.3 The Final Models

Because the wins in the last 10 home games variable was dropped from the models, the models no longer have to exclude the first 10 home games of the season, so the full dataset was split into 80% training data and 20% testing data, and the models were re-trained and re-tested on these more complete sets. Since the removal of the wins in the last 10 home games variable did not result in a practically significant decrease in the predictive capabilities of the models, other variables were tested to see if they provided practical significance to the models. In pursuit of the best balance of model simplicity and practically significant predictive capabilities, the following six variables were also removed from the model: the game number of the season, if the two teams are in the same division, if the two teams are in the same league, the visiting team's win percentage, the home team's win percentage, and the visiting team's win percentage in the previous season. This leaves a relatively simple model with only four categorical variables (month, day of week, day or night, and visiting team) plus one interaction term (day of week and day or night). Table 6.2 shows the MAPE for

Table 6.2: Prediction Accuracy With 11 and 5 Predictors

| | **Standardized Attendance** | **Attendance** |
|---|---|---|
| **With 11 Predictors** | 13.16% | 14.41% |
| **With 5 Predictors** | 13.58% | 14.72% |

the 11-predictor and the 5-predictor linear regression team-season standardized attendance model and the mixed-effects attendance model. As a baseline reference, if only the training data's average attendance for each team's season was used to predict the attendance of each game in the testing data, the MAPE would be 19.48%.

The linear regression model with the four categorical variables and one interaction term was able to provide a 30% reduction of that error. In both cases, dropping the six predictor variables from the model only increased the MAPE by less than half of a percent. To put these numbers into perspective, for a game with a true attendance of 30,000 people, a MAPE of 13.58% is equivalent to being off by 4,075 people, and an increase of less than half of a percent is equivalent to increasing that error by less than 150 people. This increase in error is statistically significant but not practically significant, so the simpler model is chosen as the final model.

## 6.4 Predicting Attendance for a New Season

An obvious practical use of these models would be to use the data from previous seasons to train the models and then use the models to predict attendance changes within future seasons. An analysis of the changes in average attendance from season to season is beyond the scope of this paper, so for a simple example here, the mean and standard deviation of the attendance in the future season will be assumed to be the same as the most recent available season. Using only the average attendance of each team in 2017 to predict their 2018 attendance gives a baseline MAPE of 28.34%; this is a pretty high baseline error. If the true 2018 mean attendance numbers are used to predict 2018 attendance numbers,

Table 6.3: Prediction Accuracy for 2018 With 2017 and 2018 Mean and Standard Deviation

| | **Standardized Attendance** | **Attendance** |
|---|---|---|
| **With 2017 Mean and SD** | 22.54% | 22.82% |
| **With 2018 Mean and SD** | 13.46% | ? |

the baseline MAPE decreases to 20.42%. This brings up another advantage of using the linear regression model with the team-season standardized attendance as the response over the mixed-effects model with the raw attendance number as the response. The mixed-effects model is not designed to predict attendance for seasons that were not included in the training dataset while the linear regression model does not require a season to be specified; therefore, if someone is able to supply an accurate estimate of the mean and standard deviation of a desired or theoretical season, the linear regression model will easily be able to produce an estimate while the mixed-effects model does not really allow for this information to be supplied. Table 6.3 shows the MAPE for the prediction of the 2018 attendance for each model.

Providing an accurate approximation for the true mean and standard deviation of attendance for the desired season can significantly improve the prediction accuracy of the model.

## 6.5 Predicting Attendance for the 2018 Dodgers Season

If a Major League Baseball team would like to predict the attendance of each of their home games in the upcoming season, it would be beneficial to fit a model using only data from that particular team's home games; this model will have more accurate estimates for how the month, day of week, time of day, and visiting team affect that home team's attendance. The Los Angeles Dodgers will be used as an example. The 2018 season is the most recent complete season available on Retrosheet at this time, so the 2005 through 2017 Dodgers seasons are used to fit the models, and the attendance of each game of the 2018 Dodgers season is predicted. The mean and standard deviation of the attendance of the 2018 season

is estimated by the 2017 season. The MAPE for the linear regression model is 6.26%, and the MAPE for the mixed-effects model is 6.55%. The error of the linear regression model is about 15% smaller than the 7.31% MAPE from using the average 2017 Dodgers attendance as the prediction for the attendance for each game of the 2018 Dodgers season. The average attendance at a Dodgers home game in 2018 was 47,042 people, so this prediction is equivalent to being off by an average of 2,945 fans per game.

# CHAPTER 7

# Results

Table 7.1 shows a summary of the linear regression model with the team-season standardized attendance as the response variable and the month, day of week, day or night indicator, and visiting team as predictors as well as an interaction term between the day of week and the day or night indicator.

The baseline levels are March (month), Monday (day of week), day (day or night), and Los Angeles Angels (visiting team), so the interpretation of the intercept is that, on average, a team can expect to see an attendance increase of 0.066 standard deviations above their average attendance for that season when playing a home game in March on a Monday during the day against the Los Angeles Angels. It is more insightful to look at the effect of one variable at a time.

July is the month with the highest coefficient of nearly 0.3, meaning that, on average, while holding all other factors constant, a game scheduled in July will see a 0.3 standard deviation higher attendance than games scheduled in March. The Major League Baseball regular season mostly runs from April through September with a very small percent of games at the end of March and beginning of October. With that in mind, attendance can be nicely summarized as having a peak in July and seeing a steady decrease when moving away from July until there is an increase in attendance for the first few and last few games of the season. Attendance in July is approximately 0.2 standard deviations higher than attendance in June and August, which is approximately 0.3 standard deviations higher than attendance in May and September, which is approximately 0.3 standard deviation higher than attendance in April. This makes sense because the weather is typically better during the summer in most

Table 7.1: Regression With Team-Season Standardized Attendance

| | Estimate | Std. Error | t value | Pr(> \|t\|) |
|---|---|---|---|---|
| (Intercept) | 0.0662 | 0.1845 | 0.36 | 0.7198 |
| April | -0.5067 | 0.1784 | -2.84 | 0.0045*** |
| May | -0.2162 | 0.1783 | -1.21 | 0.2251 |
| June | 0.1200 | 0.1783 | 0.67 | 0.5011 |
| July | 0.2995 | 0.1783 | 1.68 | 0.0931* |
| August | 0.1087 | 0.1783 | 0.61 | 0.5420 |
| September | -0.2205 | 0.1783 | -1.24 | 0.2162 |
| October | -0.0481 | 0.1852 | -0.26 | 0.7949 |
| Tuesday | -0.4994 | 0.0917 | -5.44 | 0.0000*** |
| Wednesday | -0.3729 | 0.0484 | -7.70 | 0.0000*** |
| Thursday | -0.3302 | 0.0486 | -6.80 | 0.0000*** |
| Friday | 0.0347 | 0.0737 | 0.47 | 0.6373 |
| Saturday | 0.5667 | 0.0466 | 12.16 | 0.0000*** |
| Sunday | 0.3190 | 0.0442 | 7.22 | 0.0000*** |
| Night | -0.6610 | 0.0453 | -14.60 | 0.0000*** |
| Tuesday Night | 0.6459 | 0.0940 | 6.87 | 0.0000*** |
| Wednesday Night | 0.5174 | 0.0529 | 9.77 | 0.0000*** |
| Thursday Night | 0.5345 | 0.0544 | 9.83 | 0.0000*** |
| Friday Night | 0.8605 | 0.0764 | 11.27 | 0.0000*** |
| Saturday Night | 0.9225 | 0.0515 | 17.91 | 0.0000*** |
| Sunday Night | 0.4487 | 0.0646 | 6.94 | 0.0000*** |
| Arizona | -0.1759 | 0.0369 | -4.76 | 0.0000*** |
| Atlanta | 0.0407 | 0.0375 | 1.08 | 0.2787 |
| Baltimore | -0.0981 | 0.0370 | -2.65 | 0.0081*** |
| Boston | 0.7051 | 0.0373 | 18.93 | 0.0000*** |
| Chicago Cubs | 0.4967 | 0.0372 | 13.36 | 0.0000*** |
| Chicago White Sox | -0.0050 | 0.0373 | -0.13 | 0.8937 |
| Cincinnati | -0.0369 | 0.0369 | -1.00 | 0.3177 |
| Cleveland | -0.0662 | 0.0371 | -1.78 | 0.0748* |
| Colorado | -0.1640 | 0.0371 | -4.42 | 0.0000*** |
| Detroit | 0.1243 | 0.0372 | 3.34 | 0.0008*** |
| Houston | -0.0913 | 0.0372 | -2.45 | 0.0142** |
| Kansas City | -0.1664 | 0.0370 | -4.49 | 0.0000*** |
| Los Angeles Dodgers | 0.3758 | 0.0371 | 10.13 | 0.0000*** |
| Miami | -0.2133 | 0.0374 | -5.70 | 0.0000*** |
| Milwaukee | -0.1292 | 0.0370 | -3.49 | 0.0005*** |
| Minnesota | -0.0571 | 0.0372 | -1.54 | 0.1247 |
| New York Mets | 0.2336 | 0.0370 | 6.31 | 0.0000*** |
| New York Yankees | 0.9298 | 0.0372 | 25.02 | 0.0000*** |
| Oakland | -0.0854 | 0.0371 | -2.30 | 0.0213** |
| Philadelphia | 0.0685 | 0.0371 | 1.85 | 0.0650* |
| Pittsburgh | -0.0676 | 0.0370 | -1.83 | 0.0679* |
| San Diego | -0.0824 | 0.0369 | -2.23 | 0.0255** |
| Seattle | -0.1247 | 0.0370 | -3.37 | 0.0008*** |
| San Francisco | 0.2239 | 0.0372 | 6.02 | 0.0000*** |
| St. Louis | 0.2389 | 0.0369 | 6.48 | 0.0000*** |
| Tampa Bay | -0.2000 | 0.0372 | -5.38 | 0.0000*** |
| Texas | -0.1150 | 0.0370 | -3.11 | 0.0019*** |
| Toronto | -0.1497 | 0.0371 | -4.04 | 0.0001*** |
| Washington | -0.0461 | 0.0371 | -1.24 | 0.2143 |

Table 7.2: Effect of Day of Week and Day or Night on Standardized Attendance

| | **Mon** | **Tue** | **Wed** | **Thu** | **Fri** | **Sat** | **Sun** |
|---|---|---|---|---|---|---|---|
| **Day** | 0 | -0.50 | -0.37 | -0.33 | 0.03 | 0.57 | 0.32 |
| **Night** | -0.66 | -0.02 | -0.14 | -0.13 | 0.20 | 0.26 | -0.21 |

parts of the United States. For perspective, a typical team has an average attendance of about 30,000 people and a standard deviation of about 6,000 people, so an increase of 0.1 standard deviations is equivalent to an increase of about 600 fans or roughly a 2% increase in attendance.

It is easiest to understand the effect of the day of the week and the time of the day when looking at the interaction of these two effects. The coefficients for day of week, day or night, and the interaction term have been combined and summarized in Table 7.2. The values in the table represent the standard deviation change in attendance relative to Monday games that take place during the day.

Saturday day games see the highest attendance numbers, followed by Sunday day games, Saturday night games, and Friday night games. Monday night games are the least popular, followed by Tuesday, Wednesday, and Thursday games during the day. It is interesting that Saturday day games see significantly higher attendance numbers than Saturday night games, but more Saturday games are scheduled at night rather than during the day; this may be due to television scheduling preferences.

The baseline level for the visiting team categorical variable is the Los Angeles Angels. Conveniently, they are a team that has a fairly average impact on attendance, so the coefficients of the levels of the visiting team variable can loosely be interpreted as the difference from an average team, even though the true meaning is the difference from the Los Angeles Angels. There are more large positive numbers of visiting team coefficients than large negative numbers, indicating that a visiting team can have more of a positive effect on attendance than a negative effect. The New York Yankees have by far the largest impact on attendance; teams see an average of nearly a full standard deviation increase in attendance when host-

ing the Yankees. Hosting the Yankees leads to an attendance increase of approximately 0.2 standard deviations more than hosting the Boston Red Sox, the second most popular team. The Boston Red Sox are about 0.2 standard deviations higher than the Chicago Cubs, who are about 0.1 standard deviations higher than the Los Angeles Dodgers, who are more than 0.1 standard deviations higher than every other team.

The R-squared value of this model is 0.3917, meaning that nearly 40% of the variation in team-season standardized attendance can be explained by this model. As mention earlier, the observed mean absolute percentage error of this model on testing data was 13.58%. This relatively simple model with only four predictors plus an interaction term does a good job at explaining variations in within-season attendance numbers, and it is not practically worse than more complex regression models with more variables.

Table 7.3 shows a summary of the linear mixed-effects regression model with the raw attendance number as the response variable and the same predictors variables as the previous model. This model is slightly less accurate than the previous model because it does not work as well for teams who have relatively small or large variance in their attendance numbers; however, the coefficients are easier to interpret because the response is the raw attendance number rather than the team-season standardized attendance variable. The conclusions drawn from this model are similar to those from the linear regression model.

Table 7.3: Mixed-Effects Regression Model Summary

| | Estimate | Std. Error | t value | Pr(> \|t\|) |
|---|---|---|---|---|
| (Intercept) | 30,651.110 | 1,259.899 | 24.328 | 0.000*** |
| April | −3,088.885 | 1,150.872 | −2.684 | 0.008*** |
| May | −1,400.612 | 1,150.275 | −1.218 | 0.224 |
| June | 681.831 | 1,150.329 | 0.593 | 0.554 |
| July | 1,629.262 | 1,150.498 | 1.416 | 0.157 |
| August | 624.637 | 1,150.354 | 0.543 | 0.588 |
| September | −1,234.973 | 1,150.452 | −1.073 | 0.284 |
| October | 102.266 | 1,196.205 | 0.085 | 0.932 |
| Tuesday | −2,405.690 | 589.040 | −4.084 | 0.000*** |
| Wednesday | −1,846.233 | 311.027 | −5.936 | 0.000*** |
| Thursday | −1,746.700 | 311.734 | −5.603 | 0.000*** |
| Friday | −337.272 | 493.618 | −0.683 | 0.495 |
| Saturday | 3,506.392 | 299.412 | 11.711 | 0.000*** |
| Sunday | 2,307.544 | 283.353 | 8.144 | 0.000*** |
| Night | −3,925.985 | 290.813 | −13.500 | 0.000*** |
| Tuesday Night | 3,269.309 | 603.287 | 5.419 | 0.000*** |
| Wednesday Night | 2,756.176 | 340.614 | 8.092 | 0.000*** |
| Thursday Night | 2,800.158 | 349.198 | 8.019 | 0.000*** |
| Friday Night | 6,077.077 | 512.006 | 11.869 | 0.000*** |
| Saturday Night | 6,096.092 | 331.816 | 18.372 | 0.000*** |
| Sunday Night | 2,060.323 | 416.172 | 4.951 | 0.000*** |
| Arizona | −683.035 | 248.156 | −2.752 | 0.006*** |
| Atlanta | 277.626 | 253.095 | 1.097 | 0.273 |
| Baltimore | −643.066 | 239.329 | −2.687 | 0.008*** |
| Boston | 4,819.635 | 240.752 | 20.019 | 0.000*** |
| Chicago Cubs | 3,210.163 | 250.351 | 12.823 | 0.000*** |
| Chicago White Sox | −291.442 | 240.966 | −1.209 | 0.227 |
| Cincinnati | 83.008 | 249.160 | 0.333 | 0.740 |
| Cleveland | −380.127 | 240.046 | −1.584 | 0.114 |
| Colorado | −815.265 | 249.468 | −3.268 | 0.002*** |
| Detroit | 832.875 | 240.374 | 3.465 | 0.001*** |
| Houston | −762.338 | 244.606 | −3.117 | 0.002*** |
| Kansas City | −1,131.021 | 239.651 | −4.719 | 0.000*** |
| Los Angeles Dodgers | 2,503.880 | 250.171 | 10.009 | 0.000*** |
| Miami | −1,147.966 | 251.566 | −4.563 | 0.000*** |
| Milwaukee | −544.812 | 249.615 | −2.183 | 0.030** |
| Minnesota | −419.420 | 240.403 | −1.745 | 0.082* |
| New York Mets | 1,633.412 | 249.601 | 6.544 | 0.000*** |
| New York Yankees | 6,694.227 | 240.223 | 27.867 | 0.000*** |
| Oakland | −724.498 | 239.018 | −3.031 | 0.003*** |
| Philadelphia | 683.604 | 249.584 | 2.739 | 0.007*** |
| Pittsburgh | −248.529 | 249.642 | −0.996 | 0.320 |
| San Diego | −344.562 | 248.410 | −1.387 | 0.166 |
| San Francisco | 1,658.441 | 250.395 | 6.623 | 0.000*** |
| Seattle | −1,070.141 | 238.586 | −4.485 | 0.000*** |
| St. Louis | 1,681.914 | 248.356 | 6.772 | 0.000*** |
| Tampa Bay | −1,358.580 | 240.302 | −5.654 | 0.000*** |
| Texas | −993.935 | 238.164 | −4.173 | 0.000*** |
| Toronto | −1,036.69 | 240.048 | −4.319 | 0.000*** |
| Washington | 17.495 | 249.811 | 0.070 | 0.945 |

# CHAPTER 8

# Conclusion

This paper attempted to explain and predict attendance by using a linear regression model with an attendance response variable that was standardized by each team-season. The model was compared to a linear mixed-effects regression model with each team-season as a random effect and the raw attendance number as the response variable, and the first model was found to consistently outperform the mixed-effects model. A team-season standardized attendance response variable may initially seem more difficult to interpret, but it is more informative than a raw attendance number. For example, an increase of 4,000 people is significantly different for a team with an average attendance of 10,000 and standard deviation of 2,000 than a team with an average attendance of 40,000 and a standard deviation of 8,000. A team-season standardized attendance response variable can better predict attendance for teams with relatively low or high variation in attendance.

The number of wins in a team's previous 10 home games was shown to be a statistically significant predictor of attendance, and it performed better than similar recent performance variables that involved a fewer number of games, both home and away games, or runs or home runs in place of wins; however, this variable was not shown to have any practically significant effect on attendance. Additionally, it was shown that the predictive performance of the models did not have much of a practical decrease (although the decrease was statistically significant) when simplifying the models even further. In the end, it was determined that the visiting team, day of week, time of day, and month can be used to provide a reasonable estimate of the attendance of a Major League Baseball game.

# CHAPTER 9

# Limitations and Future Work

This paper explored the reasons behind changes in within-season attendance. Most of the factors that were examined in this paper were related to the opponent, the scheduling of the games, and the performance of both the home and visiting teams. There are many other factors that likely have both practically and statistically significant effects on attendance but were not available in the dataset used in this paper. Smith and Groetzinger showed that temperature not surprisingly had a positive effect on attendance (2010). They also attempted to estimate the effect of promotions on attendance, but despite seeing large attendance increases for games with promotions, they were unable to produce significant results due to the lack of availability of sufficient data.

A seemingly obvious predictor of attendance is the price of the tickets. Unfortunately, there is no perfect way to estimate this variable. Ticket prices vary from seat to seat, and close seats can cost several times more than seats in the higher sections. To further complicate matters, ticket prices for the same seat can fluctuate from day to day, tickets can be bought from multiple different platforms, re-sale tickets would need to be accounted for in some way, tickets can frequently be earned through promotions, and sometimes tickets are given away for free to schools or youth baseball teams. Denaux et al. were able to include a weighted average ticket price into their model but actually found it to be insignificant (2011), so it seems ticket prices may not currently vary enough to noticeably impact attendance, indicating that teams may need to decrease ticket prices more dramatically if they desire to increase attendance.

Ormiston studied the effect of both the home and visiting starting pitchers on attendance

and found statistically and practically significant results (2014). He determined that top tier starting pitchers increased attendance by an average of 8-9% compared to an average pitcher, and that the starting visiting pitcher can be more significant that the starting home pitcher. This paper decided against including starting pitcher or other player information because of the difficulty of identifying the star power of players and the significant increase in model complexity.

Other suggestions for factors to explore include the weather of the game day, if the stadium has protection from bad weather, if the game is scheduled on a holiday, if and where the game is televised, and if a superstar player has been added to the roster or became injured and cannot play.

It would also be interesting to look into actual attendance versus the number of tickets sold and factors that influence changes in this ratio, but actual attendance data is not available, so studies have had to use the official reported attendance. Bad weather likely increases the gap between actual attendance and tickets sold, but high average ticket prices probably reduce this gap.

This paper attempted to summarize the factors that influence changes in attendance within a season in general, but it would interesting to compare the effects of these factors from team to team. Southern California teams are likely less impacted by the month due to less dramatic changes in weather, and the day of week and time of day likely have slightly different effects in different regions of the United States. The effect of the visiting team surely varies from team to team considering rivalry games see increased attendance in all sports.

The most practically useful analysis of attendance would be an analysis of a single team done by someone with access to promotional data and detailed ticket pricing information. These two variables are of significant interest because they can easily be controlled by the team. An analysis of a single team would also allow for a more specific understanding of the effects of the month, day of week, time of day, and particular visiting team on that home team's attendance; a team specific analysis would also be better suited to include information

on the starting home pitcher and other player details.

Finally, linear regression and mixed-effects regression models were used in favor of machine learning methods because an understanding of how attendance varies was the primary motivation, but machine learning methods can be used to more accurately predict attendance if that is the desired goal.

# CHAPTER 10

# Code Appendix

## 10.1 Dataset Creation

The following is the R code that was used to prepare the dataset for this analysis.

```
# Import datasets and combine.
first_season = 2004
last_season = 2018
data = data.frame()
for (i in first_season:last_season) {
  data = rbind(data, read.csv(paste("gl1871_2018/GL", i, ".TXT", sep = ""),
                              header = F, stringsAsFactors = F))
}

# Florida Marlins became Miami Marlins.
data[data$V4 == "FLO", "V4"] = "MIA"
data[data$V7 == "FLO", "V7"] = "MIA"

# Montreal Expos became Washingtion Nationals.
data[data$V4 == "MON", "V4"] = "WAS"
data[data$V7 == "MON", "V7"] = "WAS"

# Add missing attendance number.
data[data$V1 == 20170621 & data$V7 == "NYA", "V18"] = 39911
```

```
# Create date variables.
data$V1 = as.Date(as.character(data$V1), tryFormats = "%Y%m%d")
data$year = as.numeric(format(data$V1, "%Y"))
data$month = as.numeric(format(data$V1, "%m"))

# Calculate game at home number.
data$id = as.numeric(rownames(data))
data$game_at_home_number = NA
teams = unique(data[, c("V7", "year")])
for (i in 1:nrow(teams)) {
  ids = data[data$V7 == teams[i, "V7"] & data$year == teams[i, "year"], "id"]
  for (j in 1:length(ids)) {
    data[ids[j], "game_at_home_number"] = j
  }
}

# Determine winner of each game.
data$winner = NA
for (i in 1:nrow(data)) {
  if (data[i, "V11"] > data[i, "V10"]) {
    data[i, "winner"] = "H"
  } else if (data[i, "V10"] > data[i, "V11"]) {
    data[i, "winner"] = "V"
  } else {
    data[i, "winner"] = "T"
  }
}

# Calculate win percentages.
data$visitor_wins = NA
```

```
data$visitor_losses = NA
data$visitor_wins_last_season = NA
data$visitor_losses_last_season = NA
data$home_wins = NA
data$home_losses = NA
data$home_wins_last_season = NA
data$home_losses_last_season = NA
data$wins_at_home = NA
data$losses_at_home = NA
data$wins_at_home_last_season = NA
data$losses_at_home_last_season = NA
for (i in first_season:last_season) {
  teams = unique(data[data$year == i, "V4"])
  for (j in 1:length(teams)) {
    ids = data[data$year == i & (data$V4 == teams[j] | data$V7 == teams[j]),
               "id"]
    wins = 0
    losses = 0
    wins_at_home = 0
    losses_at_home = 0
    for (k in 1:length(ids)) {
      if (data[ids[k], "V4"] == teams[j]) {
        data[ids[k], "visitor_wins"] = wins
        data[ids[k], "visitor_losses"] = losses
        if (data[ids[k], "winner"] == "V") {
          wins = wins + 1
        } else if (data[ids[k], "winner"] == "H") {
          losses = losses + 1
        }
      } else {
```

```
        data[ids[k], "home_wins"] = wins
        data[ids[k], "home_losses"] = losses
        data[ids[k], "wins_at_home"] = wins_at_home
        data[ids[k], "losses_at_home"] = losses_at_home
        if (data[ids[k], "winner"] == "H") {
          wins = wins + 1
          wins_at_home = wins_at_home + 1
        } else if (data[ids[k], "winner"] == "V") {
          losses = losses + 1
          losses_at_home = losses_at_home + 1
        }
      }
    }
    data[data$year == i + 1 & data$V4 == teams[j],
         "visitor_wins_last_season"] = wins
    data[data$year == i + 1 & data$V4 == teams[j],
         "visitor_losses_last_season"] = losses
    data[data$year == i + 1 & data$V7 == teams[j],
         "home_wins_last_season"] = wins
    data[data$year == i + 1 & data$V7 == teams[j],
         "home_losses_last_season"] = losses
    data[data$year == i + 1 & data$V7 == teams[j],
         "wins_at_home_last_season"] = wins_at_home
    data[data$year == i + 1 & data$V7 == teams[j],
         "losses_at_home_last_season"] = losses_at_home
  }
}
data$visitor_win_percent = data$visitor_wins / (data$visitor_wins +
                                                data$visitor_losses)
data$home_win_percent = data$home_wins / (data$home_wins +
```

```
                                          data$home_losses)
data$win_at_home_percent = data$wins_at_home / (data$wins_at_home +
                                                  data$losses_at_home)
data$visitor_win_percent_last_season = data$visitor_wins_last_season /
  (data$visitor_wins_last_season + data$visitor_losses_last_season)
data$home_win_percent_last_season = data$home_wins_last_season /
  (data$home_wins_last_season + data$home_losses_last_season)
data$win_at_home_percent_last_season = data$wins_at_home_last_season /
  (data$wins_at_home_last_season + data$losses_at_home_last_season)

# Calculate game number in the series.
data$series = NA
teams = unique(data[, c("V7", "year")])
for (i in 1:nrow(teams)) {
  ids = data[data$V7 == teams[i, "V7"] & data$year == teams[i, "year"], "id"]
  data[ids[1], "series"] = 1
  for (j in 2:length(ids)) {
    if (data[ids[j], "V4"] == data[ids[j - 1], "V4"]) {
      data[ids[j], "series"] = data[ids[j - 1], "series"] + 1
    } else {
      data[ids[j], "series"] = 1
    }
  }
}

# Calculate home team wins in last n games.
data$wins_in_last_1 = NA
data$wins_in_last_2 = NA
data$wins_in_last_3 = NA
data$wins_in_last_4 = NA
```

```
data$wins_in_last_5 = NA
data$wins_in_last_6 = NA
data$wins_in_last_7 = NA
data$wins_in_last_8 = NA
data$wins_in_last_9 = NA
data$wins_in_last_10 = NA
var_names = c("wins_in_last_1", "wins_in_last_2", "wins_in_last_3",
              "wins_in_last_4", "wins_in_last_5", "wins_in_last_6",
              "wins_in_last_7", "wins_in_last_8", "wins_in_last_9",
              "wins_in_last_10")
for (i in 1:nrow(teams)) {
  ids = data[(data$V4 == teams[i, "V7"] | data$V7 == teams[i, "V7"])
             & data$year == teams[i, "year"] & data$winner != "T", "id"]
  for (j in 2:length(ids)) {
    if (data[ids[j], "V7"] == teams[i, "V7"]) {
      wins = 0
      games = 0
      for (k in (j - 1):(j - 10)) {
        if (k > 0) {
          games = games + 1
          if (data[ids[k], "V4"] == teams[i, "V7"] & data[ids[k],
                                                          "winner"] == "V"
              | data[ids[k], "V7"] == teams[i, "V7"] & data[ids[k],
                                                            "winner"] == "H") {
            wins = wins + 1
          }
          data[ids[j], var_names[games]] = wins
        }
      }
    }
```

```
  }
}

# Calculate home team wins in last n home games.
data$wins_in_last_1_home = NA
data$wins_in_last_2_home = NA
data$wins_in_last_3_home = NA
data$wins_in_last_4_home = NA
data$wins_in_last_5_home = NA
data$wins_in_last_6_home = NA
data$wins_in_last_7_home = NA
data$wins_in_last_8_home = NA
data$wins_in_last_9_home = NA
data$wins_in_last_10_home = NA
var_names = c("wins_in_last_1_home", "wins_in_last_2_home",
              "wins_in_last_3_home", "wins_in_last_4_home",
              "wins_in_last_5_home", "wins_in_last_6_home",
              "wins_in_last_7_home", "wins_in_last_8_home",
              "wins_in_last_9_home", "wins_in_last_10_home")
for (i in 1:nrow(teams)) {
  ids = data[data$V7 == teams[i, "V7"] & data$year == teams[i, "year"]
             & data$winner != "T", "id"]
  for (j in 2:length(ids)) {
    wins = 0
    games = 0
    for (k in (j - 1):(j - 10)) {
      if (k > 0) {
        games = games + 1
        if (data[ids[k], "winner"] == "H") {
          wins = wins + 1
```

```
      }
      data[ids[j], var_names[games]] = wins
    }
  }
 }
}

# Calculate home team runs in last n games.
data$runs_in_last_1 = NA
data$runs_in_last_2 = NA
data$runs_in_last_3 = NA
data$runs_in_last_4 = NA
data$runs_in_last_5 = NA
data$runs_in_last_6 = NA
data$runs_in_last_7 = NA
data$runs_in_last_8 = NA
data$runs_in_last_9 = NA
data$runs_in_last_10 = NA
var_names = c("runs_in_last_1", "runs_in_last_2", "runs_in_last_3",
              "runs_in_last_4", "runs_in_last_5", "runs_in_last_6",
              "runs_in_last_7", "runs_in_last_8", "runs_in_last_9",
              "runs_in_last_10")
for (i in 1:nrow(teams)) {
  ids = data[(data$V4 == teams[i, "V7"] | data$V7 == teams[i, "V7"])
             & data$year == teams[i, "year"], "id"]
  for (j in 2:length(ids)) {
    if (data[ids[j], "V7"] == teams[i, "V7"]) {
      runs = 0
      games = 0
      for (k in (j - 1):(j - 10)) {
```

```
        if (k > 0) {
          games = games + 1
          if (data[ids[k], "V4"] == teams[i, "V7"]) {
            runs = runs + data[ids[k], "V10"]
          } else {
            runs = runs + data[ids[k], "V11"]
          }
          data[ids[j], var_names[games]] = runs
        }
      }
    }
  }
}

# Calculate home team runs in last n home games.
data$runs_in_last_1_home = NA
data$runs_in_last_2_home = NA
data$runs_in_last_3_home = NA
data$runs_in_last_4_home = NA
data$runs_in_last_5_home = NA
data$runs_in_last_6_home = NA
data$runs_in_last_7_home = NA
data$runs_in_last_8_home = NA
data$runs_in_last_9_home = NA
data$runs_in_last_10_home = NA
var_names = c("runs_in_last_1_home", "runs_in_last_2_home",
              "runs_in_last_3_home", "runs_in_last_4_home",
              "runs_in_last_5_home", "runs_in_last_6_home",
              "runs_in_last_7_home", "runs_in_last_8_home",
              "runs_in_last_9_home", "runs_in_last_10_home")
```

```
for (i in 1:nrow(teams)) {
  ids = data[data$V7 == teams[i, "V7"] & data$year == teams[i, "year"], "id"]
  for (j in 2:length(ids)) {
    runs = 0
    games = 0
    for (k in (j - 1):(j - 10)) {
      if (k > 0) {
        games = games + 1
        runs = runs + data[ids[k], "V11"]
        data[ids[j], var_names[games]] = runs
      }
    }
  }
}

# Calculate home team home runs in last n games.
data$home_runs_in_last_1 = NA
data$home_runs_in_last_2 = NA
data$home_runs_in_last_3 = NA
data$home_runs_in_last_4 = NA
data$home_runs_in_last_5 = NA
data$home_runs_in_last_6 = NA
data$home_runs_in_last_7 = NA
data$home_runs_in_last_8 = NA
data$home_runs_in_last_9 = NA
data$home_runs_in_last_10 = NA
var_names = c("home_runs_in_last_1", "home_runs_in_last_2",
              "home_runs_in_last_3", "home_runs_in_last_4",
              "home_runs_in_last_5", "home_runs_in_last_6",
              "home_runs_in_last_7", "home_runs_in_last_8",
```

```
            "home_runs_in_last_9", "home_runs_in_last_10")
for (i in 1:nrow(teams)) {
  ids = data[(data$V4 == teams[i, "V7"] | data$V7 == teams[i, "V7"])
           & data$year == teams[i, "year"], "id"]
  for (j in 2:length(ids)) {
    if (data[ids[j], "V7"] == teams[i, "V7"]) {
      hr = 0
      games = 0
      for (k in (j - 1):(j - 10)) {
        if (k > 0) {
          games = games + 1
          if (data[ids[k], "V4"] == teams[i, "V7"]) {
            hr = hr + data[ids[k], "V26"]
          } else {
            hr = hr + data[ids[k], "V54"]
          }
          data[ids[j], var_names[games]] = hr
        }
      }
    }
  }
}

# Calculate home team home runs in last n home games.
data$home_runs_in_last_1_home = NA
data$home_runs_in_last_2_home = NA
data$home_runs_in_last_3_home = NA
data$home_runs_in_last_4_home = NA
data$home_runs_in_last_5_home = NA
data$home_runs_in_last_6_home = NA
```

```
data$home_runs_in_last_7_home = NA
data$home_runs_in_last_8_home = NA
data$home_runs_in_last_9_home = NA
data$home_runs_in_last_10_home = NA
var_names = c("home_runs_in_last_1_home", "home_runs_in_last_2_home",
              "home_runs_in_last_3_home", "home_runs_in_last_4_home",
              "home_runs_in_last_5_home", "home_runs_in_last_6_home",
              "home_runs_in_last_7_home", "home_runs_in_last_8_home",
              "home_runs_in_last_9_home", "home_runs_in_last_10_home")
for (i in 1:nrow(teams)) {
  ids = data[data$V7 == teams[i, "V7"] & data$year == teams[i, "year"], "id"]
  for (j in 2:length(ids)) {
    hr = 0
    games = 0
    for (k in (j - 1):(j - 10)) {
      if (k > 0) {
        games = games + 1
        hr = hr + data[ids[k], "V54"]
        data[ids[j], var_names[games]] = hr
      }
    }
  }
}

# Calculate wins_in_last_n and wins_in_last_n_home variables for tie games.
ties = data[data$winner == "T", c("V7", "year", "V9", "game_at_home_number",
                                  "id")]
var_names = c("wins_in_last_1", "wins_in_last_2", "wins_in_last_3",
              "wins_in_last_4", "wins_in_last_5", "wins_in_last_6",
              "wins_in_last_7", "wins_in_last_8", "wins_in_last_9",
```

```
            "wins_in_last_10")
for (i in 1:nrow(ties)) {
  ids = data[(data$V4 == ties[i, "V7"] | data$V7 == ties[i, "V7"])
           & data$year == ties[i, "year"], "id"]
  for (j in 1:10) {
    wins = 0
    for (k in (ties[i, "V9"] - j):(ties[i, "V9"] - 1)) {
      if (data[ids[k], "V4"] == ties[i, "V7"] & data[ids[k], "winner"] == "V"
          | data[ids[k], "V7"] == ties[i, "V7"] & data[ids[k],
                                                  "winner"] == "H") {
        wins = wins + 1
      }
    }
    data[ties[i, "id"], var_names[j]] = wins
  }
}
var_names = c("wins_in_last_1_home", "wins_in_last_2_home",
              "wins_in_last_3_home", "wins_in_last_4_home",
              "wins_in_last_5_home", "wins_in_last_6_home",
              "wins_in_last_7_home", "wins_in_last_8_home",
              "wins_in_last_9_home", "wins_in_last_10_home")
for (i in 1:nrow(ties)) {
  ids = data[data$V7 == ties[i, "V7"] & data$year == ties[i, "year"], "id"]
  for (j in 1:10) {
    wins = 0
    for (k in (ties[i, "game_at_home_number"] - j):
         (ties[i, "game_at_home_number"] - 1)) {
      if (data[ids[k], "winner"] == "H") {
        wins = wins + 1
      }
```

```
    }
    data[ties[i, "id"], var_names[j]] = wins
  }
}

# Create Divisions.
data[data$V4 %in% c("ARI", "COL", "LAN", "SDN", "SFN"), "visiting_division"] =
  "NW"
data[data$V7 %in% c("ARI", "COL", "LAN", "SDN", "SFN"), "home_division"] = "NW"
data[data$V4 %in% c("CHN", "CIN", "MIL", "PIT", "SLN"), "visiting_division"] =
  "NC"
data[data$V7 %in% c("CHN", "CIN", "MIL", "PIT", "SLN"), "home_division"] = "NC"
data[data$V4 %in% c("ATL", "MIA", "NYN", "PHI", "WAS"), "visiting_division"] =
  "NE"
data[data$V7 %in% c("ATL", "MIA", "NYN", "PHI", "WAS"), "home_division"] = "NE"
data[data$V4 %in% c("ANA", "HOU", "OAK", "SEA", "TEX"), "visiting_division"] =
  "AW"
data[data$V7 %in% c("ANA", "HOU", "OAK", "SEA", "TEX"), "home_division"] = "AW"
data[data$V4 %in% c("CHA", "CLE", "DET", "KCA", "MIN"), "visiting_division"] =
  "AC"
data[data$V7 %in% c("CHA", "CLE", "DET", "KCA", "MIN"), "home_division"] = "AC"
data[data$V4 %in% c("BAL", "BOS", "NYA", "TBA", "TOR"), "visiting_division"] =
  "AE"
data[data$V7 %in% c("BAL", "BOS", "NYA", "TBA", "TOR"), "home_division"] = "AE"
data[data$V4 == "HOU" & data$year <= 2012, "visiting_division"] = "NC"
data[data$V7 == "HOU" & data$year <= 2012, "home_division"] = "NC"
data$same_league = data$V5 == data$V8
data$same_division = data$visiting_division == data$home_division

# Save the full dataset.
```

```
#write.csv(data, "gl2004_2018_full.csv", row.names = F)
#data = read.csv("gl2004_2018_full.csv", stringsAsFactors = F)

# Remove unnecessary variables.
data = data[, c(1:5, 7:9, 13, 17, 18, 102, 104, 162:163, 165, 179:249)]

# Name unnamed variables.
names(data)[c(1:13)] = c("date", "game_of_day", "day_of_week", "visiting_team",
                         "visiting_league", "home_team", "home_league",
                         "home_game_number", "day_or_night", "park_id",
                         "attendance", "vp_id", "hp_id")

# Ignore the game that was closed to the public.
data = data[!is.na(data$attendance), ]

# Ignore games before the 2005 season.
data = data[data$year >= 2005, ]

# Ignore the home games played in a different stadium.
data = data[!(data$park_id %in% c("FTB01", "LBV01", "MNT01", "SJU01", "SYD01",
                                  "TOK01", "WIL02")), ]
data = data[!(data$home_team == "BAL" & data$park_id == "STP01"), ]
data = data[!(data$home_team == "CIN" & data$park_id == "SFO03"), ]
data = data[!(data$home_team == "CLE" & data$park_id == "MIL06"), ]
data = data[!(data$home_team == "HOU" & data$park_id %in% c("MIL06", "STP01")),
            ]
data = data[!(data$home_team == "MIA" & data$park_id %in% c("SEA03", "MIL06")),
            ]
data = data[!(data$home_team == "TBA" & data$park_id == "NYC20"), ]
data = data[!(data$home_team == "TOR" & data$park_id == "PHI13"), ]
```

```
# Ignore double-headers.
data = data[data$game_of_day == 0, ]

# Ignore tie-breaker games.
data = data[data$game_at_home_number != 82 | data$home_game_number != 163, ]

# Ignore the first true home game of the season.
data = data[data$game_at_home_number != 1, ]
data = data[data$home_team != "OAK" | data$year != 2008 |
            data$game_at_home_number != 3, ]
data = data[data$home_team != "OAK" | data$year != 2012 |
            data$game_at_home_number != 3, ]
data = data[data$home_team != "ARI" | data$year != 2014 |
            data$game_at_home_number != 3, ]

# Set the max of series to be 4 since few series are longer than 4 games.
data[data$series > 4, "series"] = 4

# Set win percents to percents last season for first 10 games.
data[data$home_game_number <= 10, "visitor_win_percent"] =
  data[data$home_game_number <= 10, "visitor_win_percent_last_season"]
data[data$home_game_number <= 10, "home_win_percent"] =
  data[data$home_game_number <= 10, "home_win_percent_last_season"]
data[data$game_at_home_number <= 10, "win_at_home_percent"] =
  data[data$game_at_home_number <= 10, "win_at_home_percent_last_season"]

# Standardize attendance within team-year.
d1 = tapply(data$attendance, list(data$home_team, data$year), mean)
d2 = tapply(data$attendance, list(data$home_team, data$year), sd)
```

```
for (i in 1:nrow(d1)) {
  for (j in 1:ncol(d1)) {
    data[data$home_team == rownames(d1)[i] & data$year ==
          as.numeric(colnames(d1)[j]), "mean_attendance"] = d1[i, j]
    data[data$home_team == rownames(d2)[i] & data$year ==
          as.numeric(colnames(d2)[j]), "sd_attendance"] = d2[i, j]
  }
}
data$z_attendance = (data$attendance - data$mean_attendance) /
  data$sd_attendance

# Re-order the variables.
data = data[, c(11, 6, 10, 14:15, 3, 9, 8, 16, 4, 84, 87, 5, 86, 85, 7, 13, 12,
              23, 1, 17:22, 24:83, 88:90)]

# Save the dataset.
write.csv(data, "gl2005_2018.csv", row.names = F)
```

## 10.2 Variable Selection

The following is the R code that was used to perform variable selection.

```
# Import the dataset.
data = read.csv("gl2005_2018.csv", stringsAsFactors = F)
data$year = as.character(data$year)
data$month = as.factor(data$month)
data$day_of_week = factor(data$day_of_week, levels
                          = c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun"))
data = na.omit(data)

# Standardize attendance within team-year.
d1 = tapply(data$attendance, list(data$home_team, data$year), mean)
d2 = tapply(data$attendance, list(data$home_team, data$year), sd)
for (i in 1:nrow(d1)) {
  for (j in 1:ncol(d1)) {
    data[data$home_team == rownames(d1)[i] & data$year ==
           as.numeric(colnames(d1)[j]), "mean_attendance"] = d1[i, j]
    data[data$home_team == rownames(d2)[i] & data$year ==
           as.numeric(colnames(d2)[j]), "sd_attendance"] = d2[i, j]
  }
}
data$z_attendance = (data$attendance - data$mean_attendance) /
  data$sd_attendance

# Look at which predictors are significant overall.
library(glmnet)
set.seed(1)
d = data[, c(89, 5:8, 10, 12, 14, 19, 21:24, 27:86)]
x = model.matrix(z_attendance ~ . + day_of_week * day_or_night - 1, d)
```

```
y = d$z_attendance
m = cv.glmnet(x, y)
coef(m, s = 2 * (m$lambda.1se - m$lambda.min) + m$lambda.min)
predictors = data.frame()
p = coef(m, s = 2 * (m$lambda.1se - m$lambda.min) + m$lambda.min)
for (j in 1:length(p@i)) {
  r = data.frame(predictor = p@Dimnames[[1]][p@i[j] + 1], beta = p@x[j],
                 stringsAsFactors = F)
  predictors = rbind(predictors, r)
}
predictors$variable = predictors$predictor
predictors[substr(predictors$variable, 16, 27) == "day_or_night",
           "variable"] = "day_or_night:day_of_week"
predictors[substr(predictors$variable, 1, 5) == "month", "variable"] = "month"
predictors[substr(predictors$variable, 1, 11) == "day_of_week",
           "variable"] = "day_of_week"
predictors[substr(predictors$variable, 1, 13) == "visiting_team",
           "variable"] = "visiting_team"
unique(predictors$variable)

# Look at which predictors are significant for each stadium.
parks = sort(unique(data$park_id))
park_predictors = data.frame()
set.seed(1)
for (i in 1:length(parks)) {
  d = data[data$park_id == parks[i], c(89, 5:8, 10, 12, 14, 19, 21:24, 27:86)]
  x = model.matrix(z_attendance ~ . + day_of_week * day_or_night - 1, d)
  y = d$z_attendance
  m = cv.glmnet(x, y)
  p = coef(m, s = 2 * (m$lambda.1se - m$lambda.min) + m$lambda.min)
```

```
  for (j in 1:length(p@i)) {
    r = data.frame(park_id = parks[i], predictor = p@Dimnames[[1]][p@i[j] + 1],
                   beta = p@x[j], stringsAsFactors = F)
    park_predictors = rbind(park_predictors, r)
  }
}
sort(table(park_predictors$predictor), decreasing = T)
round(sort(table(park_predictors$predictor) * 100 / length(parks),
           decreasing = T))
park_predictors$variable = park_predictors$predictor
park_predictors[substr(park_predictors$variable, 16, 27) == "day_or_night",
                "variable"] = "day_or_night:day_of_week"
park_predictors[substr(park_predictors$variable, 1, 5) == "month",
                "variable"] = "month"
park_predictors[substr(park_predictors$variable, 1, 11) == "day_of_week",
                "variable"] = "day_of_week"
park_predictors[substr(park_predictors$variable, 1, 13) == "visiting_team",
                "variable"] = "visiting_team"
d = unique(cbind(park_predictors$park_id, park_predictors$variable))
round(sort(table(d[, 2]) * 100 / length(parks), decreasing = T))

# Look at which predictors are significant for each year for each team.
seasons = unique(cbind(data$home_team, data$year))
season_predictors = data.frame()
set.seed(1)
for (i in 1:nrow(seasons)) {
  d = data[data$home_team == seasons[i, 1] & data$year == seasons[i, 2],
           c(89, 5:8, 10, 12, 14, 19, 21:24, 27:86)]
  x = model.matrix(z_attendance ~ . + day_of_week * day_or_night - 1, d)
  y = d$z_attendance
```

```
  m = cv.glmnet(x, y)
  p = coef(m, s = 2 * (m$lambda.1se - m$lambda.min) + m$lambda.min)
  for (j in 1:length(p@i)) {
    r = data.frame(home_team = seasons[i, 1], year = seasons[i, 2],
                   predictor = p@Dimnames[[1]][p@i[j] + 1], beta = p@x[j],
                   stringsAsFactors = F)
    season_predictors = rbind(season_predictors, r)
  }
}
sort(table(season_predictors$predictor), decreasing = T)
round(sort(table(season_predictors$predictor) * 100 / nrow(seasons),
          decreasing = T), 1)
season_predictors$variable = season_predictors$predictor
season_predictors[substr(season_predictors$variable, 16, 27) == "day_or_night",
                  "variable"] = "day_or_night:day_of_week"
season_predictors[substr(season_predictors$variable, 1, 5) == "month",
                  "variable"] = "month"
season_predictors[substr(season_predictors$variable, 1, 11) == "day_of_week",
                  "variable"] = "day_of_week"
season_predictors[substr(season_predictors$variable, 1, 13) == "visiting_team",
                  "variable"] = "visiting_team"
d = unique(cbind(season_predictors$home_team, season_predictors$year,
                 season_predictors$variable))
round(sort(table(d[, 3]) * 100 / nrow(seasons), decreasing = T), 1)
```

## 10.3 Model Fitting and Prediction

The following is the R code that was used to fit the models and make predictions.

```
# Import the dataset.
data = read.csv("gl2005_2018.csv", stringsAsFactors = F)
data$year = as.factor(data$year)
data$month = as.factor(data$month)
data$day_of_week = factor(data$day_of_week, levels
                          = c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun"))

# Split dataset into training and testing data.
set.seed(1)
ids = sample(1:nrow(na.omit(data)), round(0.8 * nrow(na.omit(data))))
train = na.omit(data)[ids, ]

# Standardize attendance within team-year.
d1 = tapply(train$attendance, list(train$home_team, train$year), mean)
d2 = tapply(train$attendance, list(train$home_team, train$year), sd)
for (i in 1:nrow(d1)) {
  for (j in 1:ncol(d1)) {
    data[data$home_team == rownames(d1)[i] & data$year ==
           as.numeric(colnames(d1)[j]), "mean_attendance"] = d1[i, j]
    data[data$home_team == rownames(d2)[i] & data$year ==
           as.numeric(colnames(d2)[j]), "sd_attendance"] = d2[i, j]
  }
}
data$z_attendance = (data$attendance - data$mean_attendance) /
  data$sd_attendance
train = na.omit(data)[ids, ]
test = na.omit(data)[-ids, ]
```

```
# Calculate error for full model.
m = lm(z_attendance ~ . + day_of_week * day_or_night,
      train[, c(89, 5:8, 10, 12, 14, 21:22, 24, 46)])
test$predicted = predict(m, test)
test$predicted = test$mean_attendance + test$predicted * test$sd_attendance
mean(abs((test$attendance - test$predicted) / test$attendance))
# 0.1292055
library(lme4)
m = lmer(paste("attendance ~", paste(names(train[c(5:8, 10, 12, 14, 21:22, 24,
                                                   46)]), collapse = " + "),
             "+ day_of_week * day_or_night + (1 | home_team:year)"), train)
test$predicted = predict(m, test)
mean(abs((test$attendance - test$predicted) / test$attendance))
# 0.1405396

# Calculate error for model without recent performance variable.
m = lm(z_attendance ~ . + day_of_week * day_or_night,
      train[, c(89, 5:8, 10, 12, 14, 21:22, 24)])
test$predicted = predict(m, test)
test$predicted = test$mean_attendance + test$predicted * test$sd_attendance
mean(abs((test$attendance - test$predicted) / test$attendance))
# 0.1292573
m = lmer(paste("attendance ~", paste(names(train[c(5:8, 10, 12, 14, 21:22,
                                                   24)]), collapse = " + "),
             "+ day_of_week * day_or_night + (1 | home_team:year)"), train)
test$predicted = predict(m, test)
mean(abs((test$attendance - test$predicted) / test$attendance))
# 0.1405457
```

```
# Test for statistical significance.
anova(lm(z_attendance ~ . + day_of_week * day_or_night,
        train[, c(89, 5:8, 10, 12, 14, 21:22, 24)]),
     lm(z_attendance ~ . + day_of_week * day_or_night,
        train[, c(89, 5:8, 10, 12, 14, 21:22, 24, 46)]))

# Split full dataset into training and testing data.
set.seed(1)
ids = sample(1:nrow(data), round(0.8 * nrow(data)))
train = data[ids, ]

# Standardize attendance within team-year.
d1 = tapply(train$attendance, list(train$home_team, train$year), mean)
d2 = tapply(train$attendance, list(train$home_team, train$year), sd)
for (i in 1:nrow(d1)) {
  for (j in 1:ncol(d1)) {
    data[data$home_team == rownames(d1)[i] & data$year ==
          as.numeric(colnames(d1)[j]), "mean_attendance"] = d1[i, j]
    data[data$home_team == rownames(d2)[i] & data$year ==
          as.numeric(colnames(d2)[j]), "sd_attendance"] = d2[i, j]
  }
}
data$z_attendance = (data$attendance - data$mean_attendance) /
  data$sd_attendance
train = data[ids, ]
test = data[-ids, ]

# Calculate error for baseline.
mean(abs(test$attendance - test$mean_attendance) / test$attendance)
# 0.1948383
```

```
# Calculate error for full model without recent performance variable.
m = lm(z_attendance ~ . + day_of_week * day_or_night,
      train[, c(89, 5:8, 10, 12, 14, 21:22, 24)])
test$predicted = predict(m, test)
test$predicted = test$mean_attendance + test$predicted * test$sd_attendance
mean(abs((test$attendance - test$predicted) / test$attendance))
# 0.13155
m = lmer(paste("attendance ~", paste(names(train[c(5:8, 10, 12, 14, 21:22,
                                                   24)]), collapse = " + "),
             "+ day_of_week * day_or_night + (1 | home_team:year)"), train)
test$predicted = predict(m, test)
mean(abs((test$attendance - test$predicted) / test$attendance))
# 0.1440838

# Calculate error for final subset model.
m = lm(z_attendance ~ . + day_of_week * day_or_night, train[, c(89, 5:7, 10)])
test$predicted = predict(m, test)
test$predicted = test$mean_attendance + test$predicted * test$sd_attendance
mean(abs((test$attendance - test$predicted) / test$attendance))
# 0.1358207
m = lmer(attendance ~ visiting_team + month + day_of_week * day_or_night
        + (1 | home_team:year), train)
test$predicted = predict(m, test)
mean(abs((test$attendance - test$predicted) / test$attendance))
# 0.1471885

# Test for statistical significance.
anova(lm(z_attendance ~ . + day_of_week * day_or_night, train[, c(89, 5:7, 10)])
     , lm(z_attendance ~ . + day_of_week * day_or_night,
```

```
          train[, c(89, 5:8, 10, 12, 14, 21:22, 24)]))

# Use the most recent year as testing data.
train = data[data$year != 2018, ]

# Standardize attendance within team-year.
d1 = tapply(train$attendance, list(train$home_team, train$year), mean)
d1[, ncol(d1)] = d1[, ncol(d1) - 1]
d2 = tapply(train$attendance, list(train$home_team, train$year), sd)
d2[, ncol(d2)] = d2[, ncol(d2) - 1]
for (i in 1:nrow(d1)) {
  for (j in 1:ncol(d1)) {
    data[data$home_team == rownames(d1)[i] & data$year ==
          as.numeric(colnames(d1)[j]), "mean_attendance"] = d1[i, j]
    data[data$home_team == rownames(d2)[i] & data$year ==
          as.numeric(colnames(d2)[j]), "sd_attendance"] = d2[i, j]
  }
}
data$z_attendance = (data$attendance - data$mean_attendance) /
  data$sd_attendance
train = data[data$year != 2018, ]
test = data[data$year == 2018 & data$month != 3, ]

# Calculate error for baseline.
mean(abs(test$attendance - test$mean_attendance) / test$attendance)
# 0.2834097

# Calculate error for attendance of new year.
m = lm(z_attendance ~ . + day_of_week * day_or_night, train[, c(89, 5:7, 10)])
test$predicted = predict(m, test)
```

```
test$predicted = test$mean_attendance + test$predicted * test$sd_attendance
mean(abs((test$attendance - test$predicted) / test$attendance))
# 0.2253636
m = lmer(attendance ~ visiting_team + month + day_of_week * day_or_night
        + (1 | home_team:year), train)
test$year = 2017
test$predicted = predict(m, test)
mean(abs((test$attendance - test$predicted) / test$attendance))
# 0.2282362

# Use the true mean and standard deviation of test attendance for prediction.
d1 = tapply(data$attendance, list(data$home_team, data$year), mean)
d2 = tapply(data$attendance, list(data$home_team, data$year), sd)
for (i in 1:nrow(d1)) {
  for (j in 1:ncol(d1)) {
    data[data$home_team == rownames(d1)[i] & data$year ==
          as.numeric(colnames(d1)[j]), "mean_attendance"] = d1[i, j]
    data[data$home_team == rownames(d2)[i] & data$year ==
          as.numeric(colnames(d2)[j]), "sd_attendance"] = d2[i, j]
  }
}
data$z_attendance = (data$attendance - data$mean_attendance) /
  data$sd_attendance
train = data[data$year != 2018, ]
test = data[data$year == 2018 & data$month != 3, ]

# Calculate error for baseline.
mean(abs(test$attendance - test$mean_attendance) / test$attendance)
# 0.2042132
```

```
# Calculate error for attendance of new year.
m = lm(z_attendance ~ . + day_of_week * day_or_night, train[, c(89, 5:7, 10)])
test$predicted = predict(m, test)
test$predicted = test$mean_attendance + test$predicted * test$sd_attendance
mean(abs((test$attendance - test$predicted) / test$attendance))
# 0.1345861

# Only use Dodgers data.
data = data[data$home_team == "LAN", ]

# Split Dodgers dataset into training and testing data.
set.seed(1)
ids = sample(1:nrow(data), round(0.8 * nrow(data)))
train = data[ids, ]

# Standardize attendance within team-year.
d1 = tapply(train$attendance, list(train$home_team, train$year), mean)
d2 = tapply(train$attendance, list(train$home_team, train$year), sd)
for (i in 1:nrow(d1)) {
  for (j in 1:ncol(d1)) {
    data[data$home_team == rownames(d1)[i] & data$year ==
           as.numeric(colnames(d1)[j]), "mean_attendance"] = d1[i, j]
    data[data$home_team == rownames(d2)[i] & data$year ==
           as.numeric(colnames(d2)[j]), "sd_attendance"] = d2[i, j]
  }
}
data$z_attendance = (data$attendance - data$mean_attendance) /
  data$sd_attendance
train = data[ids, ]
test = data[-ids, ]
```

```
# Calculate error for Dodgers baseline.
mean(abs(test$attendance - test$mean_attendance) / test$attendance)
# 0.1066077

# Calculate error for Dodgers model.
m = lm(z_attendance ~ . + day_of_week * day_or_night, train[, c(89, 5:7, 10)])
test$predicted = predict(m, test)
test$predicted = test$mean_attendance + test$predicted * test$sd_attendance
mean(abs((test$attendance - test$predicted) / test$attendance))
# 0.09098148
m = lmer(attendance ~ visiting_team + month + day_of_week * day_or_night
        + (1 | year), train)
test$predicted = predict(m, test)
mean(abs((test$attendance - test$predicted) / test$attendance))
# 0.09194041

# Use the most recent year as testing data.
train = data[data$year != 2018, ]

# Standardize attendance within team-year.
d1 = tapply(train$attendance, list(train$home_team, train$year), mean)
d1[, ncol(d1)] = d1[, ncol(d1) - 1]
d2 = tapply(train$attendance, list(train$home_team, train$year), sd)
d2[, ncol(d2)] = d2[, ncol(d2) - 1]
for (i in 1:nrow(d1)) {
  for (j in 1:ncol(d1)) {
    data[data$home_team == rownames(d1)[i] & data$year ==
          as.numeric(colnames(d1)[j]), "mean_attendance"] = d1[i, j]
    data[data$home_team == rownames(d2)[i] & data$year ==
```

```
          as.numeric(colnames(d2)[j]), "sd_attendance"] = d2[i, j]
  }
}
data$z_attendance = (data$attendance - data$mean_attendance) /
  data$sd_attendance
train = data[data$year != 2018, ]
test = data[data$year == 2018 & data$month != 3, ]

# Calculate error for Dodgers baseline.
mean(abs(test$attendance - test$mean_attendance) / test$attendance)
# 0.07311782

# Calculate error for Dodgers attendance of new year.
m = lm(z_attendance ~ . + day_of_week * day_or_night, train[, c(89, 5:7, 10)])
test$predicted = predict(m, test)
test$predicted = test$mean_attendance + test$predicted * test$sd_attendance
mean(abs((test$attendance - test$predicted) / test$attendance))
# 0.06262847
m = lmer(attendance ~ visiting_team + month + day_of_week * day_or_night
        + (1 | year), train)
test$year = 2017
test$predicted = predict(m, test)
mean(abs((test$attendance - test$predicted) / test$attendance))
# 0.06547658

# Use the true mean and standard deviation of test attendance for prediction.
d1 = tapply(data$attendance, list(data$home_team, data$year), mean)
d2 = tapply(data$attendance, list(data$home_team, data$year), sd)
for (i in 1:nrow(d1)) {
  for (j in 1:ncol(d1)) {
```

```
    data[data$home_team == rownames(d1)[i] & data$year ==
          as.numeric(colnames(d1)[j]), "mean_attendance"] = d1[i, j]
    data[data$home_team == rownames(d2)[i] & data$year ==
          as.numeric(colnames(d2)[j]), "sd_attendance"] = d2[i, j]
  }
}
data$z_attendance = (data$attendance - data$mean_attendance) /
  data$sd_attendance
train = data[data$year != 2018, ]
test = data[data$year == 2018 & data$month != 3, ]

# Calculate error for Dodgers baseline.
mean(abs(test$attendance - test$mean_attendance) / test$attendance)
# 0.07415232

# Calculate error for Dodgers attendance of new year.
m = lm(z_attendance ~ . + day_of_week * day_or_night, train[, c(89, 5:7, 10)])
test$predicted = predict(m, test)
test$predicted = test$mean_attendance + test$predicted * test$sd_attendance
mean(abs((test$attendance - test$predicted) / test$attendance))
# 0.06181666
```

# BIBLIOGRAPHY

[1] Davis, M. C. (2009). Analyzing the relationship between team success and MLB attendance with GARCH effects. *Journal of Sports Economics*, 10(1):44–58.

[2] Denaux, Z. S., Denaux, D. A., and Yalcin, Y. (2011). Factors affecting attendance of Major League Baseball: Revisited. *Atlantic Economic Journal*, 39(2):117–127.

[3] Meehan, J. W. Jr., Nelson, R. A., and Richardson, T. V. (2007). Competitive balance and game attendance in Major League Baseball. *Journal of Sports Economics*, 8(6):563–580.

[4] Ormiston, R. (2014). Attendance effects of star pitchers in Major League Baseball. *Journal of Sports Economics*, 15(4):338–364.

[5] Silver, N. (2006). Is Alex Rodriguez Overpaid?. In J. Keri (Ed.), *Baseball by the Numbers* (174-98). New York: Basic Books.

[6] Smith, E. E. and Groetzinger, J. D. (2010). Do fans matter? The effect of attendance on the outcomes of Major League Baseball games. *Journal of Quantitative Analysis in Sports*, 6.