MLB Team Statistics Between 2000 to 2010 Build Baysian Model that Predicts Current Standing in the 2024 Season[1]

Tracy Yang

April 21, 2024

Abstract Using Major League Baseball key team statistics between 2000 and 2010, this paper investigates the outcome of the 2024 season using each team's current statistics on predicting their total wins out of 162 games. We investigate the common conception that baseball is a numbers game by evaluating the predictability of wins that a team would obtain during a seasion from looking at their batting average (BA), defensive efficiency ratio (DefEff), and earned run average (ERA). Despite rules changes in 2023 that may cause inaccuracies in the model, the prediction for current standings from a multiple linear regression model is quite accurate. The findings show that baseball indeed should consider statistics in assessing the competitiveness of a team.

Table of contents

[1]Code and data are available at: https://github.com/YcartXin/MLB_Team_Stats

1 Introduction

Behind the entertaining personalities and eye-catching traditions of baseball is a game where teams and are meticulously crafted with calculation. Baseball's fascination with numbers persisted since its creation [1]. Runs scored have always been recorded to evaluate performance, while other statistics developed overtime into the complicated net of data that now reside upon the MLB official site [2] The term "sabermetrics" even developed to describe in specific baseball statistics [3] However,in Major League Baseball, a team's performance can differ greatly year to year. A team's last year wins are not often indicative of the coming season. Many factors play into the seasonal performance of a team including more unpredictable variables such as injuries and luck. However, with the common narrative of baseball being a numbers game, we aim to estimate the impact of important statistics of a team on its games won per season in MLB, more specifically, the progressing 2024 season.

Important statistics for a team in this paper is identified as the team's batting average (BA), defensive efficiency ratio (DefEff), and earned run average (ERA). These three statistics each refer to an important aspect of a strong team: hitting, pitching, and fielding prospectively. A baseball team cannot succeed with a weak end in any of the three categories. Therefore, these are important indicators on the estimand - the wins a team would achieve by the end of the 2024 seasons among 162 games. The specifics of these indicators will be elaborated on in Section 2. Among present research, 2024 season predictions exist, however, they mainly focus on the win for a specific game or series for betting purposes, focusing on the player data [3]. This paper is unique in its approach to predict and investigates the predictability of overall seasonal wins for baseball teams.

The rest of the paper is structured as follows. Section 2 (Data) explains the source of the data used, its collection process, and the process of cleaning that led to the final data sets used. The data section also explores the potential biases the collected and final data may possess. Section 3 (Model) explains and justifies the model used, then sets up the model using the 2000 - 2010 data. Section 4 (Results) will show the estimated values for the variables of interest and predict wins for the on-going season with the 2024 data. Lastly, Section 5 (Discussion) discusses the model coefficients, prediction results, weaknesses in the paper, as well as further possible exploration in the topic.

The overall analysis shows that team statistics in MLB is often indicative of the strength of the team and some top contenders in the league can be estimated by team statistics. Furthermore, BA seems to have the most effect on the number of wins a team should accumulate throughout a season among the three predictor variables. However, this investigation cannot validate the ability for current team statistics to predict future success in the same season.

2 Data

The data used in this analysis of different statistics on MLB Teams are sourced from Sports Reference, more specifically, Baseball Reference [4]. The data sets used for building the model in this investigation focus on the team statistics of the 30 teams in MLB and span from 2000 to 2010. The data sets were downloaded separately for each year among the batting, pitching, and fielding data. They were then combined into the final data set during the cleaning process. The prediction data of the 2024 season is downloaded in the same way and combined as a separate data set. The analysis in this paper will be carried out in **R** [5] using packages **tidyverse** [6], **modelsummary** [7], **ggplot2** [8], **knitr** [9], **here** [10], **arrow** [11], **rstanarm** [12].

2.1 Measurement and Collection

The data available on Baseball Reference is provided by the official stats partner of the NBA, NHL, and MLB *Sportsradar* and gathered by many contributors through a variety of sources such as the Lahmen database [4]. Much of older data is accessed from third-party websites such as Retrosheet [4]. Baseball Reference stores and manipulate the data gathered using MySQL [4]. A potential weakness in the data collection includes missing data for certain teams, especially for early seasons in the 19th century when statistics were not kept in database and are easily lost. However, this analysis uses recent season data between 2000 and 2010 which mitigates this concern. Years 2000 to 2010 were chosen as this decade avoids the recent disturbances caused by COVID-19 yet is still immediate enough to be good training data for the model to predict the current MLB season.

There are in total 36 raw data sets of MLB team statistics in the raw data folder for constructing the model. The 33 for model building are split into 11 years and 3 categories (batting, pitching, fielding) for each team of every year from 2000 to 2010. The other three consist of the same categories but of 2024. Note that the statistics used for prediction in this paper is sourced on April 20th and is only representative of the statistics from games prior. Among each category, one representative indicator is chosen to demonstrate the overall strength of that category and be included as a variable for predicting wins over a season.

The batting data in a year contains 29 variables detailing the overall offensive strength of hitters in each team. These variables include average runs per game or 2nd base reached over a season. Among these variables, we are interested in the batting average (**BA**). BA is calculated by total hits divided by total at-bats. Hits is defined by the number of times a batter strikes the ball into fair territory and reaches base, in the absence of an error or a fielder's choice [2]. BA is a good indicator for offensive capabilities as it minimizes the impact of the pitcher by discounting walks and it is one of the universal tools to measure a hitter's success.

The pitching data in a year contains 36 variables for each team detailing the bullpen's strength in initiating outs and preventing runs. Among variables in the data, we are interested in the Earned Run Average (**ERA**) calculated by 9 x earned runs/innings pitched [2]. This number indicates the average earned runs allowed per 9 innings, which is better for a team if lower. This indicator tests the fundamental purpose of pitchers in the bullpen as their goal is to prevent runs of the opposing team and accurately portrays the pitching ability of a team.

The fielding data for each year contains 16 variables that details the defensive abilities of a team to get the hitter out after a ball is in play. The variable chosen to represent this ability is Defensive Efficiency (**DefEff**), calculated by 1 - (H[2] + ROE[3] - HR[4]) / (PA[5] - BB[6] - SO[7] - HBP[8] - HR), which is the percentage of balls in play converted into outs [2]. This indicator estimates the ability of the catcher and fielders to efficiently prevent runs after a ball allows hitters to reach base and home.

2.2 Cleaning and Exploration

There are two cleaned data sets. The data set for model building contains 330 observations while the one for predicting contains 30. They both include 5 variables:

- Team (Tm)
- Batting Average (BA) - hits over at bats
- Defensive Efficiency (DefEff) - out rate after a ball is in play
- Earned Run Average (ERA) - average earned runs allowed per none innings
- Wins (w)

There are thirty data points per year as there are thirty teams in MLB. The predictor variables are batting average, earned average, and defensive efficiency. Team names is necessary to identify teams throughout the years and the number of wins is imported in with the pitching data. Table 1 shows a sample of the the cleaned data with 5 variables.

Table 1: Sample of Cleaned MLB Prediction Data for Model Building

Tm	BA	DefEff	ERA	W
LAA	0.280	0.699	5.00	82
AZ	0.265	0.687	4.35	85
ATL	0.271	0.692	4.05	95
BAL	0.272	0.684	5.37	74
BOS	0.267	0.696	4.23	85
CHC	0.256	0.693	5.25	65

We can understand the model building data set better by understanding the distribution of some of the variables. For instance, we can understand the competitiveness of MLB by seeing the distribution of wins per season, whether the teams differ significantly by wins can help us analyze whether there is a significant disparity in the strengths of different teams. Figure 1 shows the distribution of the wins data of each team among 162 games in a season between 2000 to 2010.

[2]H = Hit
[3]ROE = Reached on Error, when a batter reaches base due to a defensive error
[4]HR = Home Run
[5]PA = Plate Appearance
[6]BB = Base on Balls, better known as a walk which happens when the pitcher receives 4 balls from the umpire
[7]SO = Strike Out, when the batter is out of play from three strikes from the umpire
[8]HBP = Hit By Pitch, when a batter is struck directly by a pitch without swinging and is awarded a walk

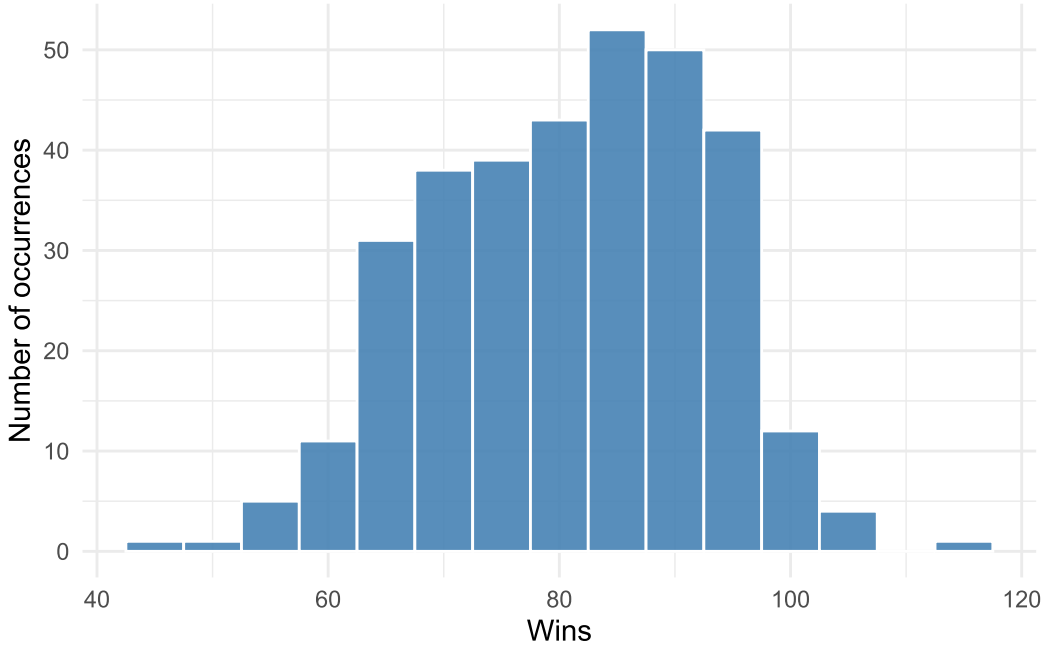


Figure 1: Distribution of Wins of each Team in a MLB Season between 2000 - 2010

With 162 games per season, 50% win rate would count to 81 wins. However, Figure 1 shows that the highest frequencies lie around 85 to 90. This means that the distribution of wins over the 11 years is slightly skewed to the left.The teams in the league are generally quite competitive and close in wins, though, there are some weaker teams that stray further from 50% win rate.

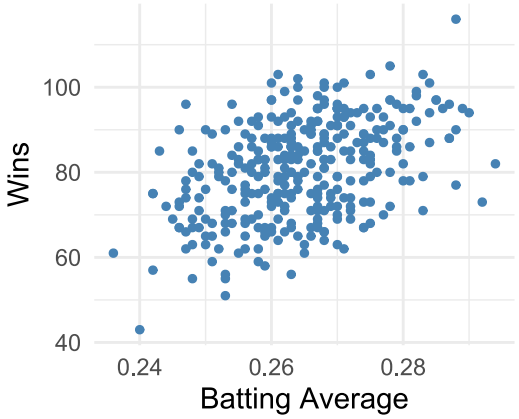
The data can be further explored with summary statistics. We can get a sense of the average strength of the a MLB team in terms of the three predictors by looking at their summary data. This is shown in Table 2.

Table 2: Summary of Cleaned MLB Prediction Data for Model Building

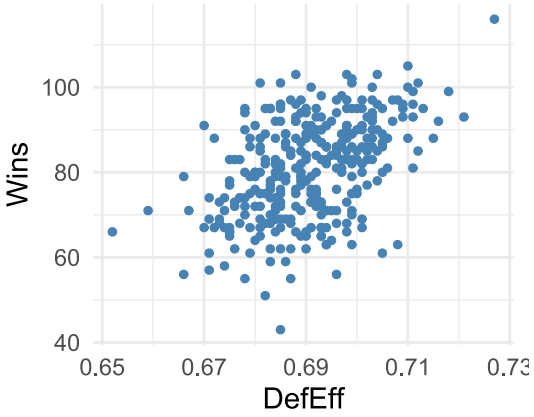
Statistic	BA	DefEff	ERA
Mean	0.2645848	0.6906242	4.388576
Median	0.2640000	0.6900000	4.350000

The mean and median of batting average are very close at around 0.264. This means that on average between 2000 to 2010, a MLB batter has 26.4% probability of getting a hit when at bat. The defensive efficiency is once again very similar between mean and median at around 0.690. MLB teams' defense against against a ball in play is quite strong, with around 69% of batters end up as outs when a ball is in play. However, the mean of earned run average is slightly higher than the median at around 4.39 compared to 4.35. This shows that there are some pitchers with very low ERAs that skew the mean to be lower. An average MLB pitcher between 2000 and 2010 overall allows around 4.39 runs per 9 innings.

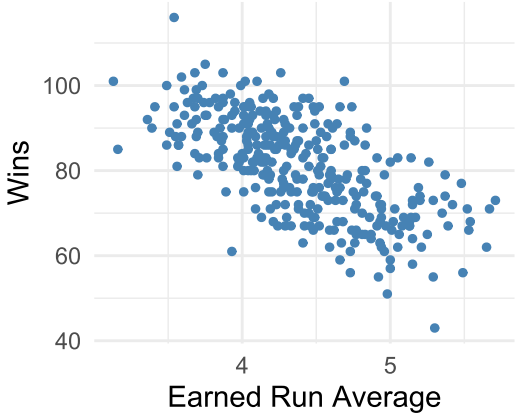
Furthermore, we can see the relationship between the predictor variables by regressing the number of wins on each as shown in Figure 2.



(a): Wins on BA



(b): Wins on DefEff



(c): Wins on ERA

Figure 2: Regressions of Number of Wins on each Predictor Variable

From Figure 2, we can see that both batting average and defensive efficiency have positive relationships with the number of wins for a team. The directions for both graphs seem to be similar, even though DefEff has a more middle-concentrated data spread. This may suggest a stronger relationship for BA than DefEff. For earned run average, there is a clear negative trend which suggests that pitchers are of high importance in ensuring wins in baseball. However, this relationship is also expected as a high ERA for the team would mean that many points were scored against the team which is a more direct indication of losses than runs or good defensive plays.

3 Model

There are two steps to the use of modeling in this paper. This paper will be utilizing a multiple linear regression containing three predictor variables to model their effects on the number of wins out of 162 for a MLB team. Subsequently, the model will be used to inference the team with the most wins in MLB in the 2024 season with their current team stats on batting average, defensive efficiency, and earned run average.

Section 3.1 Section 3.1.a will expand on the Bayesian analysis model and details of the multiple linear regression.

3.1 Model set-up

Define y_i as the number of wins a team would have by the end of 162 games. Subsequently, β_0 is the coefficient for the intercept and $\beta_1, \beta_2, \beta_3$ are respectively defined as the effect that batting average, defensive efficiency, and earned run average has on y_i to be estimated.

$$
\begin{aligned}
y_i \mid \mu_i, \sigma &\sim \text{Normal } (\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 \text{BA}_i + \beta_2 \text{DefEff}_i + \beta_3 \text{ERA}_i \\
\beta_0 &\sim \text{Normal } (0, 2.5) \\
\beta_1 &\sim \text{Normal } (0, 2.5) \\
\beta_2 &\sim \text{Normal } (0, 2.5) \\
\beta_3 &\sim \text{Normal } (0, 2.5) \\
\sigma &\sim \text{Exponential } (1)
\end{aligned}
$$

3.1.a Model justification

From Figure 2, a positive linear relationship is expected between batting average and number of wins as a higher batting average for a team signals a stronger offense which leads to runs. We also expect a positive relationship between defensive efficiency as a higher percentage indicates a higher prevention of runs and should lead to more wins. Lastly, a negative relationship between earned run average and wins is expected because a lower ERA means that the pitchers effectively prevented the opposite team from scoring runs.

The model is ran using the **rstanarm** package. The priors are assumed as rstanarm defaults. However, they are auto scaled based on the available 2000 - 2010 data.

4 Results

The coefficient variables specified in the multiple regression model in Section 3.1 are shown in Table 3 below.

Table 3: Model of Team Wins per MLB Season based on Team Statistics

	posterior estimates
(Intercept)	−18.205
BA	586.508
DefEff	24.200
ERA	−16.557
Num.Obs.	330
R2	0.721
R2 Adj.	0.719
Log.Lik.	−1065.863
ELPD	−1069.8
ELPD s.e.	12.9
LOOIC	2139.7
LOOIC s.e.	25.8
WAIC	2139.7
RMSE	6.10

The intercept is estimated in the model as −18.205. This does not make much sense as it indicates that a team with BA, DefEff, and ERA of 0 would have −18.205 wins which is not possible.

The 586.508 for β_1 is a much higher than the other beta variables. The stronger relationship is also suggested by Figure 2 where there exists a more consistent positive correlation.

The coefficient variable for DefEff is 24.2. Although the positive value is expected, the large difference between this coefficient and that of BA may suggest that offensive strength is more important than defensive strength for a MLB team.

β_3 is estimated to be around −16.557. The negative relationship is once again expected. However, since earned run average has a direct relationship with runs instead of hits from BA and DefEff, it would make sense to anticipate a lower value.

Table 4: Top Five Teams for Predicted Wins of the 2024 Season

Tm	pred
ATL	97
MIL	97
CLE	96
NYM	95
KC	93

Using the model specified in Section 3.1, the current 2024 team statistics are used to predict the five teams with the highest number of wins in the league by the end of the season. Table 4 shows that the five teams are the Atlanta Braves, Milwaukee Brewers, Cleveland Guardians, New York Mets, and Kansas City Royals in this order.

We can verify this prediction with the current (as of April 20th) MLB standings [4].

- The Atlanta Braves are currently tied with the New York Yankees at second with 14 wins
- The Milwaukee Brewers is tied with Kansas City Royals and three other teams at third with 13 wins
- The Cleveland Guardians leads the league at first with 15 wins
- The New York Mets is tied with four other teams at fourth with 12 wins

5 Discussion

5.1 Coefficient estimates

From Table 3, we see that batting average has a much higher effect on the number of wins than defensive efficiency or earned run average. This can be interpreted as the importance of offense showing through the statistics. For both DefEff and ERA, when successful, would prevent the opposing team form scoring runs. However, baseball is a game for which the team with the higher number of runs win. This means that strong defense is ineffectual without offense, and that a strong offensive team can at times mitigate lacking in fielding or pitching and result in wins. This effect is captured by the high coefficient estimate for BA in the model. Furthermore, frequent hits and runs puts mental pressure on the other team and effect their performance, a lead can often drive a game in a certain direction.

On the other hand, as both DefEff and ERA captures the defensive abilities of a team, their coefficients would share the effects of keeping the other team at low or zero runs. These coefficients thus in comparison, more moderate than the batting average's.

5.2 Is baseball predictable by numbers

The predictions made in Section 4 are not exactly the ranking of the MLB teams. However, the current league leader Cleveland Guardians is one of the predicted top fives total wins. The other teams predicted are also not far from the top. The least games won out of the five is New York Mets which has 12 wins and tied with four other teams at fourth. Considering that there are 30

teams and the rankings can drop quiet fast, the predicted teams clearly are comparatively strong teams in the league.

We can look at the distribution of the predicted wins in Figure 3 to further analyze the model.

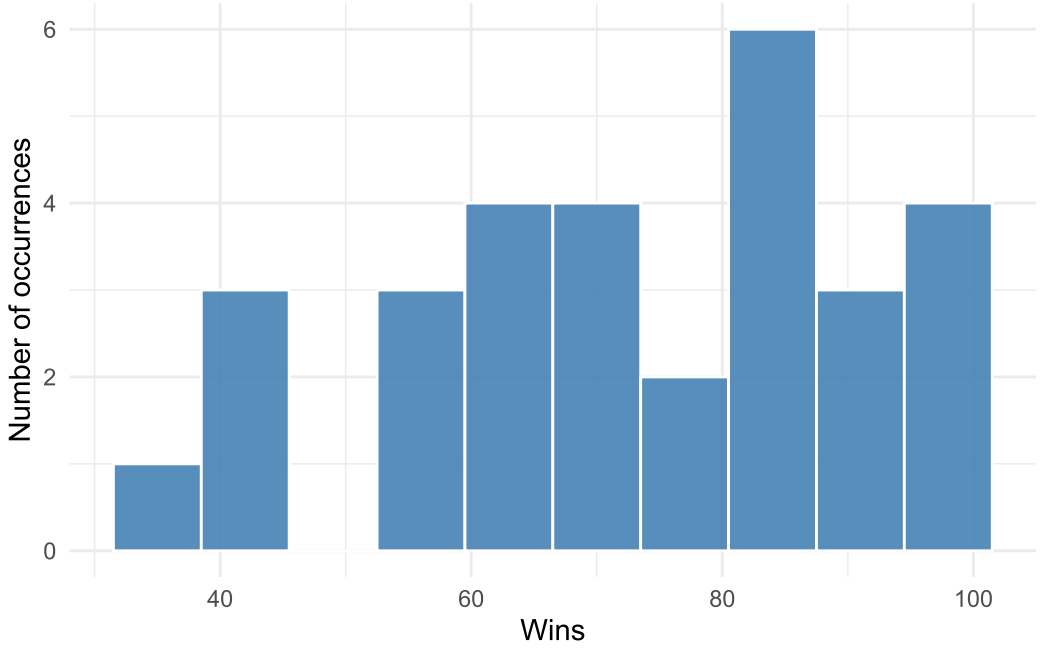


Figure 3: Distribution of Wins of each Team in a MLB 2024 Season

Figure 3 differs from the collected distribution data from 2000 to 2010 in Figure 1 with a higher frequency distribution on both ends in comparison to the middle. Figure 1 resembles more like a Normal distribution although slightly skewed. Figure 3 loses the resemblance to a Normal curve with a more even distribution. However, this is expected as there are only 30 data points in the predicted wins in comparison to the 330 that were collected.

We can further check the reliability of the model by comparing the overall predicted number of wins in the league in comparison to the number of wins there should be in Table 5.

Table 5: Total Wins Predicted VS. Real Total Wins per Season

total	real_total
2172	2430

Since whenever two teams play against each other there is always a winner, there should be a total number of wins a league should have by the end of the season. 162/2*30 = 2430 The total wins predicted in the 2024 season is 2172. The model under-predicted the number of wins there should be in the league. However, looking at Figure 3, we can see that the frequency of high wins is quite high. This means that the model severely under-estimated the wins of some teams that should be ranked more in the middle and slightly over-estimated some teams that should have

less wins. Other the other hand, it may just be the fact that team statistics have not yet stabilized at a reasonable value with the limited number of games in the season.

5.3 Weaknesses

There are a few weaknesses in the process of this investigation. Firstly, as we are using very early season data, there are many ties between teams in terms of number of wins and many teams have yet to display their true abilities. Each team has so far played around 20 games which is not enough to determine the direction for the rest of the season yet. Many players may have an irregular start in comparison to their last year statistics. The more data we have on the games each team plays, the team statistics would be updated and are more likely to be consistent with the accurate value.

Furthermore, this analysis matched the currently standings to the predictions made by the current statistics in the league. Therefore, the teams with better statistics should naturally have more wins. The predictions made by the model indicates that team statistics in many ways is representative of the strengths of the team. However, it does not validate the ability that current baseball statistics can predict for wins for the whole season.

Another weakness of this investigation is an important rule change implemented in MLB in 2023. The league added pitch timer that restricts the time that pitchers have in between pitches in order to increase the pace of play [2]. This means that pitchers do not have the liberty to relax and prepare as much as they did before. This change may cause inaccuracies in the model by increasing/decreasing the estimate of ERA. However, there is not enough data since the rule change to build a model on so far.

5.4 Next Steps

This analysis can be further developed by exploring the predictability of future wins based on a stable early season data. More specifically, analysis should be done by obtaining data after around 40 games for all teams and note down the predicted top winners in MLB at the end of the season to match the predictions made prior. By conducting such research over a few years, we would have a better idea of whether league winners show by statistics by a quarter of the season.

Another way to explore MLB team predictions in depth is to consider the individual players statistics instead of team statistics. The team statistics likely overlooks the importance of a few individual key players and favour teams with an overall more even and higher batting, fielding, or pitching statistics. However, baseball games can be determined by one hit or defensive catch and individual statistics would better capture the true abilities of the team.

Bibliography

[1] A. Schwarz, “A numbers revolution”. 2004. [Online]. Available: https://www.espn.com/mlb/columns/story?columnist=schwarz_alan&id=1835745

[2] M. L. Baseball, “STATS”. 2024. [Online]. Available: https://www.mlb.com/stats/

[3] T. Sun HC.and Lin and Y. Tsai, “Performance prediction in major league baseball by long short-term memory networks”, *Int J Data Sci Anal*, no. 15, pp. 93–104, 2023, doi: https://doi.org/10.1007/s41060-022-00313-4.

[4] S. R. LLC, “Baseball-Reference.com”. 2000. [Online]. Available: https://www.baseball-reference.com/

[5] R Core Team, “R: A Language and Environment for Statistical Computing”. 2023. [Online]. Available: https://www.r-project.org/

[6] H. Wickham *et al.*, “Welcome to the tidyverse”, *Journal of Open Source Software*, vol. 4, no. 43, p. 1686, 2019, doi: 10.21105/joss.01686.

[7] V. Arel-Bundock, “modelsummary: Data and Model Summaries in R”, *Journal of Statistical Software*, vol. 103, no. 1, pp. 1–23, 2022, doi: 10.18637/jss.v103.i01.

[8] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org/

[9] Y. Xie, “knitr: A Comprehensive Tool for Reproducible Research in R”, *Implementing Reproducible Computational Research*. Chapman, Hall/CRC, 2014.

[10] K. Müller, “here: A Simpler Way to Find Your Files”. 2020. [Online]. Available: https://cran.r-project.org/package=here

[11] N. Richardson *et al.*, “arrow: Integration to 'Apache' 'Arrow'”. 2024. [Online]. Available: https://cran.r-project.org/package=arrow

[12] S. Brilleman, M. Crowther, M. Moreno-Betancur, J. Buros Novik, and R. Wolfe, “Joint longitudinal and time-to-event models via Stan.”. [Online]. Available: https://github.com/stan-dev/stancon_talks/