MLB Team Statistics between 2000 - 2010 on Predicting Games Won per Season[1]

Tracy Yang

April 20, 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 also 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 (AVG), defensive efficiency ratio (DefEff), and earned run average (ERA).

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 (Swhwarz, 2004). 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 (XX). The term "sabermetrics" even developed to describe in specific baseball statistics (XX) 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 defined as the team's batting average (AVG), 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. 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) will explain the source of the data used, its collection process, and the process of cleaning that led to the final data used. The data section will also explore the potential biases the collected and final data may possess. Section 3…

The overall analysis shows…

2 Data

The data used in this analysis of different statistics on MLB Teams are sourced from Sports Reference, more specifically, Baseball Reference (). The data sets used 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 analysis in this paper will be carried out in **R** [1] using packages **tidyverse** [2], **dplyr** , **ggplot2** [3], **knitr** [4], **here** [5], **arrow** [6] XX.

2.1 Measurement and Collection

The data available on Baseball Reference is gathered by many contributors through a variety of sources such as the Lahmen database, and much of older data is accessed from third-party websites such as Retrosheet (). Baseball Reference stores and manipulate the data gathered using MySQL (). 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 33 raw data sets of MLB team statistics in the raw data folder for constructing the model, these are split into 11 years and 3 categories (batting, pitching, fielding) for each team of every year from 2000 to 2010. 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 (**AVG**). AVG 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 (). AVG 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 (). 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 (). 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

The cleaned data now contains 330 observations of 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
CWS	0.286	0.685	4.66	95
CIN	0.274	0.702	4.33	85

We can understand this data 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

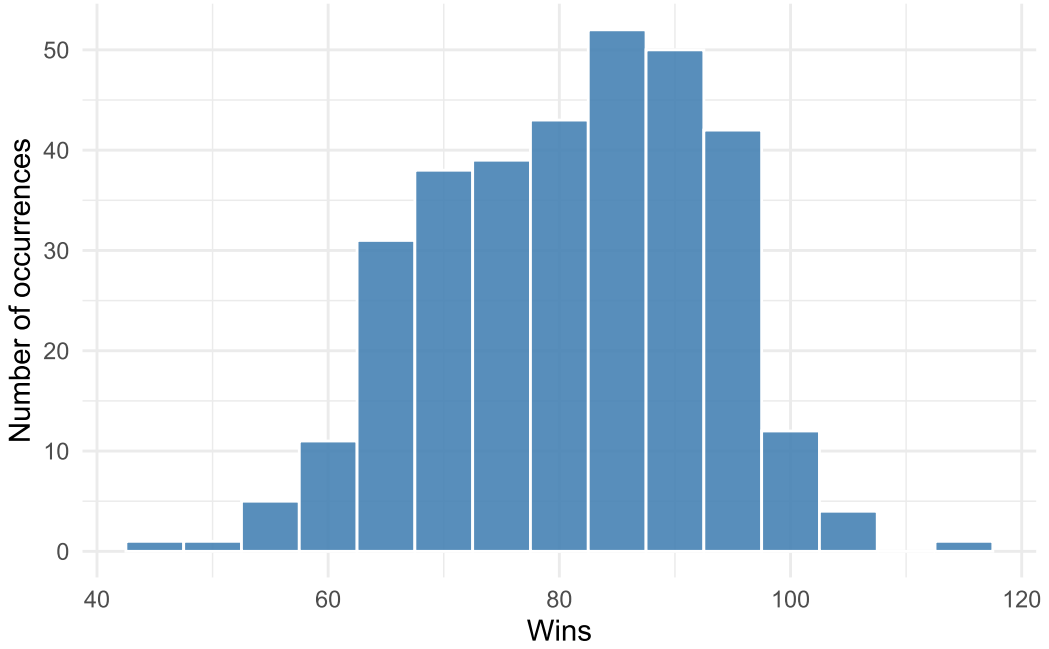

Figure 1: Distribution of Wins of each Team in a MLB Season between 2000 - 2010

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.

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

A positive 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 (XX). The priors are assumed as rstanarm defaults. However, they are auto scaled based on the available 2000 - 2010 data.

4 Results

The coefficients specified in the multiple regression model in Section 3.1 are shown in Table 3 below. Our results are summarized in .

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

	auto-scaling priors
(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

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Weaknesses and next steps

Weaknesses and next steps should also be included. Weaknesses: irregular start for some teams, pitching clock

6 Model details

6.1 Posterior predictive check

In 1 we implement a posterior predictive check. This shows…

In 2 we compare the posterior with the prior. This shows…

6.2 Diagnostics

is a trace plot. It shows… This suggests…

is a Rhat plot. It shows… This suggests…

7 References

Bibliography

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

[2] 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.

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

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

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

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