Time Series Final Project

Danielle Sebring & Steven Barnett

5/8/2022

Problem Statement

Major League Baseball (MLB) has collected data on its players and their performances since 1871. This rich supply of data allows one to investigate and learn about trends in performance over time. Additionally, the game of baseball itself has evolved over the years. This is due to a variety of factors, ranging from the introduction of new technology to changes in the rules established by MLB. For example, the advent of sports analytics in the early 2000s has had an enormous impact on the game. Recently there has been increased concern that with player improvements, particularly in the area of pitcher arm strength and pitch speed. This has the potential to make the game of baseball less entertaining and therefore cause fans to lose interest. As a result, MLB is concerned about major revenue losses from a decrease in in-person game attendance as well TV viewership. Therefore, we are going to investigate some of these trends that lead to a less strategic, nuanced, and entertaining game of baseball. We will look at several statistics over the last century that are indicative of these trends, find the best models that fit these time series, and perform forecasting to reveal what the future of MLB holds considering the current direction.

Data Collection

Our data source is Lahman's Baseball Database. This database has batting, pitching, and fielding statistics dating back to 1871. Lahman's data allows for a high-level view of trends across the league, as well as potential for in-depth analysis down to an individual team-by-team or player-by-player level. We also referred to the Baseball Reference via Bill Petti's baseballr package, although there is restricted access to the database. Baseball Reference provided some insight on statistics of interest that were not readily available in Lahman's Baseball Database, but that could be calculated through simple data wrangling procedures.

Initial data manipulation required aggregating the player level statistics to the league level. Some statistics present in the Baseball Reference database (e.g. batting average) were not present in Lahman's, so we had to do some calculations in order to maintain the consistency.

Our data has over 100 different statistics measured over the last century. An analysis of all of these would go beyond the scope of this project. Therefore, we identified six statistics that are indicative of the changes in baseball detrimental to the game's entertainment value: strikeouts, home runs, hits, wild pitches, batters hit by pitch, and sacrifices. Each of these has either increased or decreased over time due to the greater arm strength of pitchers and the resulting higher velocity of pitches. For example, as velocity of pitch has increased, the number of strikeouts naturally increases because batters have a harder time connecting with the ball. Additionally, home runs increase when pitch speed increases because the exit velocity of the ball from the bat is higher, making a home run more likely.

Theses six statistics are all tracked by count (e.g. number of hits in the league per season). However, it is common in sports analytics to look at count variables as averages or rates. This practice reduces extreme values (e.g. the 2020 MLB season was shortened due to COVID-19, resulting in a drastic drop in all statistics) and produces more well-behaved residuals in model fitting. Therefore, for each of these six statistics, we calculate the "per nine innings" rate.

We also found that data from the early years is highly variable. We attribute this to the rudimentary methods of data collection available in the late 1800s as well as the changes that often come as an organization begins.

The National League of MLB officially began in 1876. Due to this early variability, we only consider the last 100 years (1922-2021). Therefore, after all data manipulation and cleanup, we have the following statistics for the years 1922-2021:

Code	Definition
SO9	Strikeouts per Nine Innings
HR9	Home Runs per Nine Innings
H9	Hits per Nine Innings
WP9	Wild Pitches per Nine Innings
HBP9	Hit by Pitch per Nine Innings
SAC9	Sacrifices (bunts and fly outs) per Nine Innings

Model Fitting

Diagnostics

Forecasting

Discussion

Future Work