# Time Series Intermediate Data

Danielle Sebring & Steven Barnett

4/5/2022

## 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. Therefore, we are going to investigate the evolution of MLB player, team, division, and league-wide statistics over time. We would like to identify certain events in history, outside influences, and changes to the game itself and how they alter player, team, division, and league-wide performance moving forward. We would also like to consider relationships between different statistics and correlations among their individual time series.

## Data Collection

Our initial data source, Baseball Reference via Bill Petti's baseballr package, has restricted access to the database. As a result, we searched for a database with similar content and found it in Lahman's Baseball Database. This database is much cleaner, easier to collect data from, and provides the potential for in-depth analysis down to an individual player-by-player level.

Initial data manipulation required aggregating the player level statistics to the league and team 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. After all data manipulation and cleanup, we have the following statistics for the years 1871-2021:

- **Batting**:

| Code | Definition |
|---|---|
| playerID | Player ID |
| yearID | Year |
| stint | Player Stint (Order of Appearance in a Season |
| teamID | Team ID |
| lgID | League ID |
| G | Games |
| AB | At Bats |
| R | Runs |
| H | Hits |
| X2B | Doubles |
| X3B | Triples |
| HR | Homeruns |
| RBI | Runs Batted In |
| SB | Stolen Bases |
| CS | Caught Stealing |
| BB | Base on Balls (Walks) |
| SO | Strikeouts |
| IBB | Intentional Base on Balls (Intentional Walks) |
| HBP | Hit By Pitch |
| SH | Sacrifice Hits |
| SF | Sacrifice Flies |
| GIDP | Grounded into Double Plays |
| BA | Batting Average |
| PA | Plate Appearances |
| X1B | Singles |
| OBP | On Base Percentage |
| TB | Total Bases |
| SLG | Slugging Percentage |
| OPS | On Base + Slugging Percentage |
| R.G | Runs per Game |

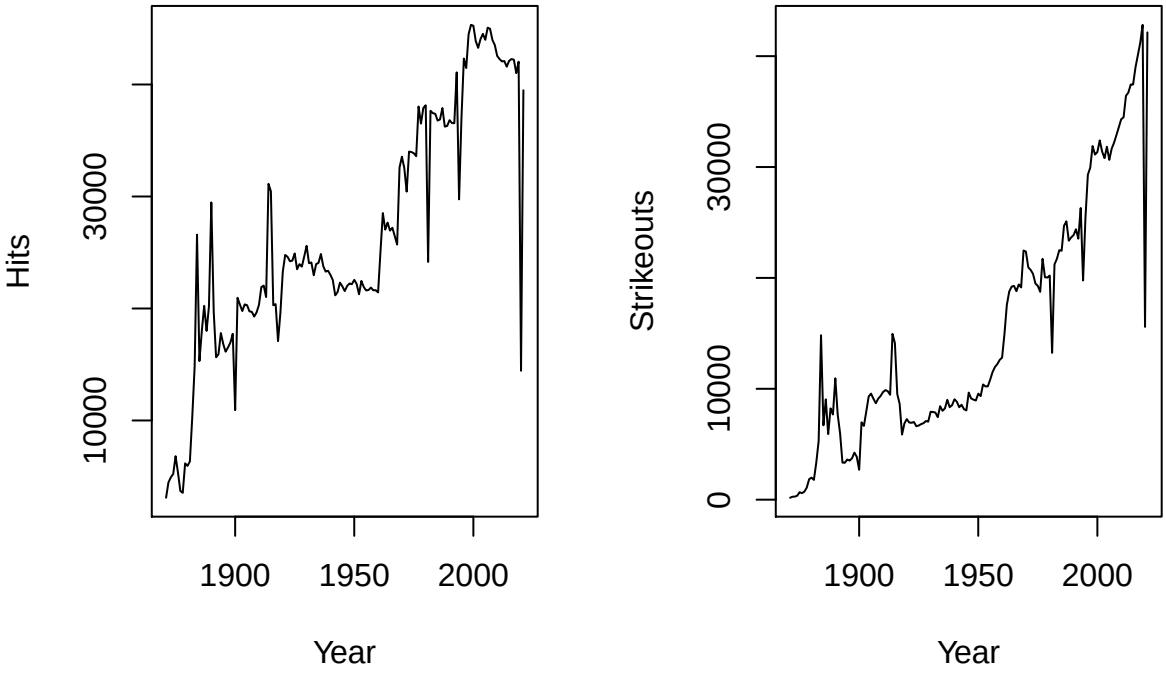
- **Pitching**:

| Code | Definition |
|---|---|
| playerID | Player ID |
| yearID | Year |
| stint | Player Stint (Order of Appearance in a Season |
| teamID | Team ID |
| lgID | League ID |
| W | Wins |
| L | Losses |
| G | Games |
| GS | Games Started |
| CG | Complete Games |
| SHO | Shutouts |
| SV | Saves |
| IPouts | Outs Pitched |
| H | Hits |
| ER | Earned Runs |
| HR | Homeruns |
| BB | Base on Balls (Walks) |
| SO | Strikeouts |
| BAOpp | Opponent's Batting Average |
| ERA | Earned Run Average |
| IBB | Intentional Base on Balls (Intentional Walks) |
| WP | Wild Pitches |
| HBP | Batters Hit by Pitch |
| BK | Balks |
| BFP | Batters Faced by Pitcher |
| GF | Games Finished |
| R | Runs Allowed |
| SH | Sacrifice Hits by Opposing Batters |
| SF | Sacrifice Flies by Opposing Batters |
| GIDP | Grounded into Double Plays by Opposing Batters |
| RA.G | Runs Allowed per Game |
| WLP | Win Loss Percentage |
| IP | Innings Pitched |
| WHIP | Walks and Hits per Innings Pitched |
| H9 | Hits per 9 Innings |
| HR9 | Homeruns per 9 Innings |
| BB9 | Base on Balls (Walks) per 9 Innings |
| SO9 | Strikeouts per 9 Innings |
| SO.W | Strikeout to Walk Ratio |

## Exploratory Data Analysis

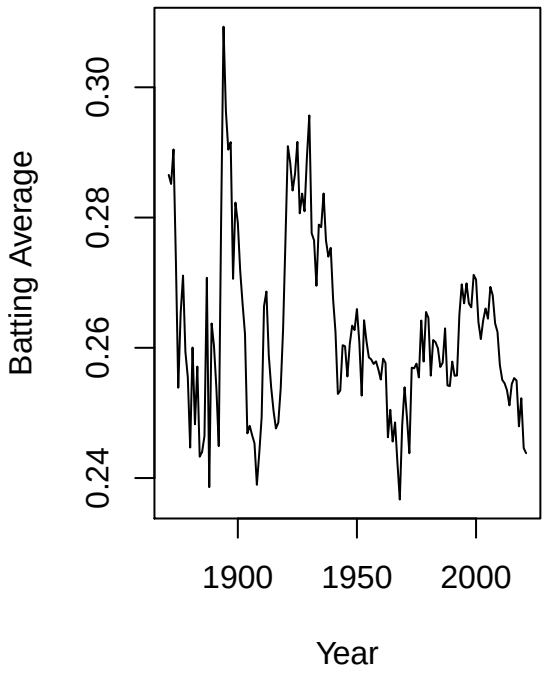
### Data Types

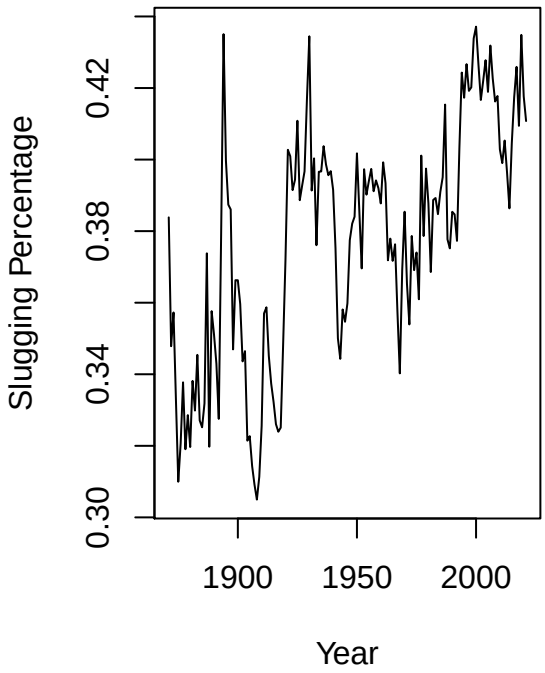
The statistics recorded in our dataset fall into a few different categories of data. For instance, we have many statistics that are count data (e.g. hits, strikeouts, walks, etc). Most of these count measurements lead to non-stationary time series, as shown below:



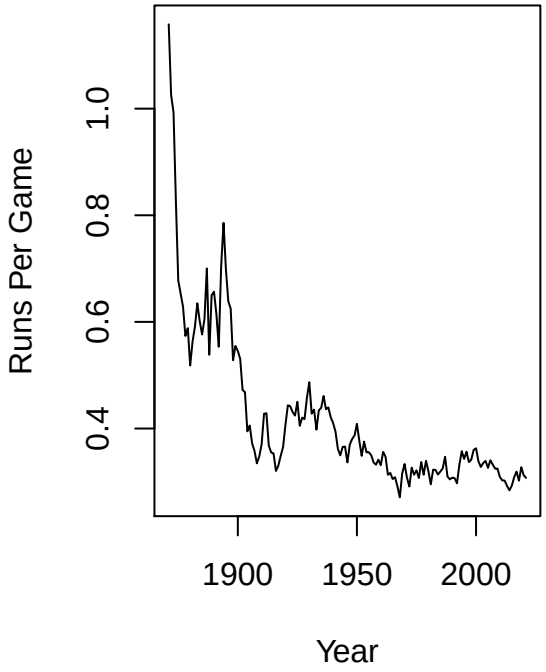

As we can see, these are clearly non-stationary. This is expected. Over the past 100-150 years, Major League Baseball (MLB) has grown, either to the expansion of the league to include more teams, or adding more games to the yearly schedule. As such, the number of hits, strikeouts, etc. will also increase. For these time series, we will considering either differencing or some type of model that accounts for the non-stationarity (e.g. MA, ARMA, ARIMA, etc).

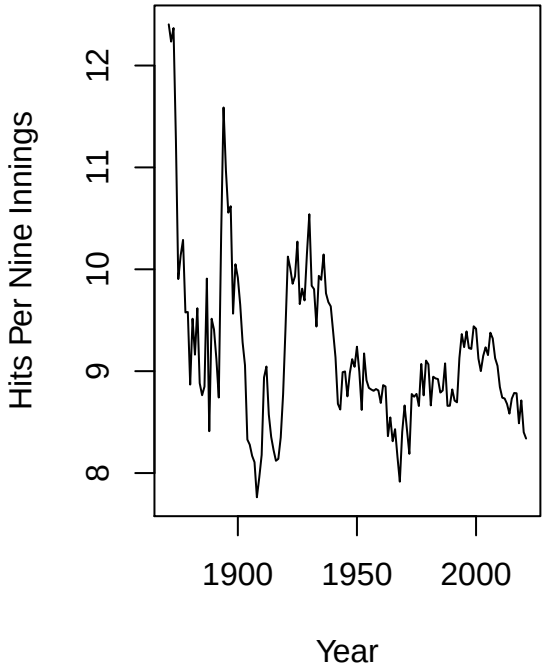
Another type of data we encounter in this data set is percentages (e.g. batting average, slugging percentage, on-base percentage. Each of these statistics fall between 0 and 1. Although this does not imply stationarity, these statistics cannot increase or decrease without bound.




We also have a couple fieds that are calculated averages. These fields don't have the same bounds that percentages do, but also do not increase without bound as the counts seem to. These fields include Runs Per Game, Hits Per Nine Innings, Earned Run Average, etc.

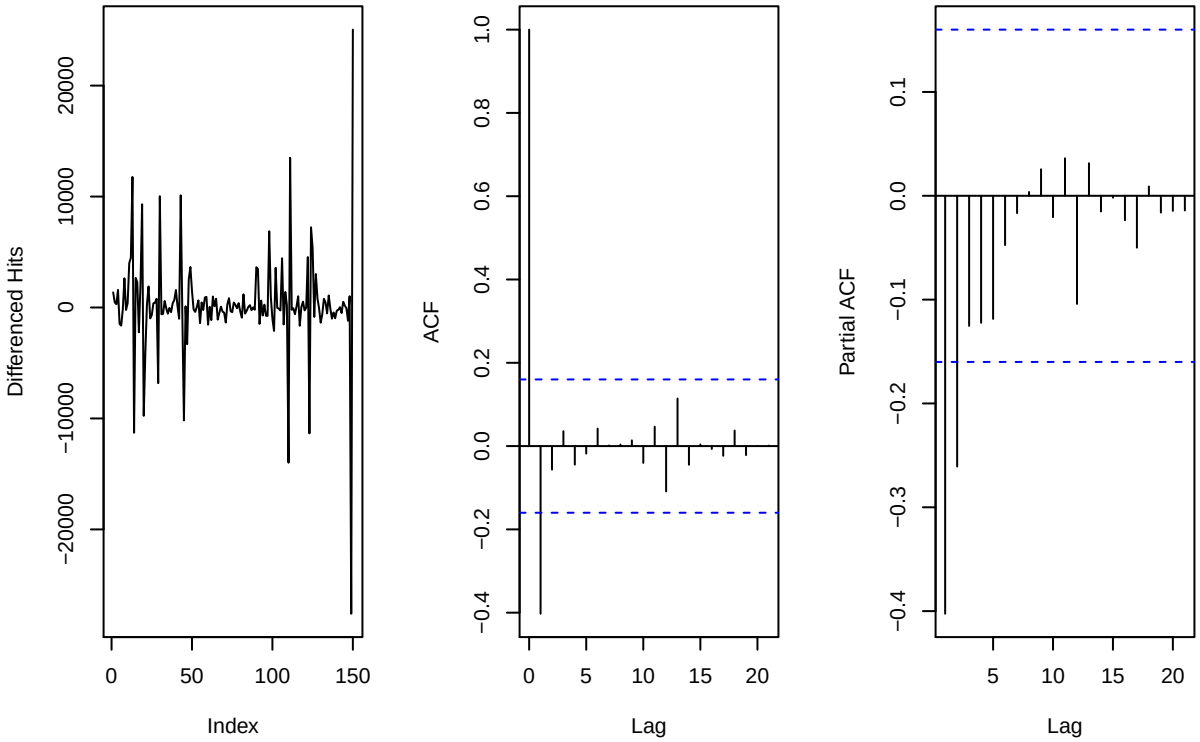



**Initial Models**

As shown above, our dataset has an extensive number of statistics observed through time. First, we'll look number of hits. This is a simple time series that can help us hone our analysis skills. We displayed a plot earlier that showed this time series is non-stationary. To make it stationary and allow for analysis using an AR(p) model, we will conduct differencing. There doesn't seem to be any sort of seasonality, so we'll calculate the differenced time series using a lag of one. We can see below that differencing is effective at

removing the non-stationarity. Also displayed below is the auto-correlation function (ACF) and the partial auto-correlation function (PACF).




Both the ACF and PACF above show that the time series may only depend on the first and second lags for a given time point. We will fit an AR(p) model and calculate the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) to see which order model fits best. To be safe, we will consider the models of up to order 5.

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Akaike Information Criterion | 37.77 | 6.07 | 0.00 | 0.98 | 1.49 | 1.92 |
| Bayesian Information Criterion | 40.79 | 12.10 | 9.05 | 13.05 | 16.57 | 20.02 |
| Approximate Posterior Probabilities | 0.00 | 0.16 | 0.72 | 0.10 | 0.02 | 0.00 |

Both the AIC and BIC agree that the AR(2) is the most appropriate. Additionally, we can see that the approximate posterior probability for the AR(2) model is over 0.7. Assuming that the true model is an AR(p) model of order 0-5, the AR(2) model is likely the best model. We cannot completely discount the AR(1) and AR(3) models, but it seems unlikely for them to be the correct model. Therefore, our fitted AR(2) model is as follows:

$$y_t - \mu = -0.499 \cdot (y_{t-1} - \mu) + \epsilon_t$$

Here we have only considered an AR(p) model. But we would also like to consider other models, such as the ARMA or ARIMA models. We fit ARIMA models for all combinations of auto regressive coefficients ($p$), moving average coefficients ($q$), and differencing coefficients ($d$). Below are our results:

| | q = 0, d = 0 | q = 1, d = 0 | q = 0, d = 1 | q = 1, d = 1 | q = 0, d = 2 | q = 1, d = 2 |
|---|---|---|---|---|---|---|
| p = 0 | 3243.54 | 2957.15 | 3130.29 | 2919.33 | 3089.31 | 2924.28 |
| p = 1 | 2983.64 | 2928.72 | 2950.74 | 2924.29 | 2955.63 | 2928.78 |
| p = 2 | 2958.71 | 2926.16 | 2955.64 | 2929.25 | 2960.33 | 2933.55 |

The best model, according to the BIC criterion, is the ARIMA(0, 1, 1) model. This is different than what we saw previously when only considering AR(p) models. The AR(2) model we fit earlier had a BIC of 2958.71,

so this ARIMA(0, 1, 1) model improves upon that. It is also a simpler model with one less coefficient.

$$y_t = 26555.05 + 0.7858\epsilon_{t-1} + \epsilon_t$$

**Next Steps**

We have only included one time series in this report, that of number of league-wide hits per MLB season. We have begun to explore other time series, but will not include them for brevity's sake. Our next steps will be to conduct forecasting for these time series and see how they perform against real data. Additionally, we want to look at how the individual time series relate to each other through cross-correlation functions. Lastly, as we learn more advanced topics in class, we plan to incorporate Time Varying Autoregressive (TVAR) models and Mixed Model Time Series.