Formal Model

Andres Izquierdo

3/1/2022

1. The data and data generating process:

The raw data I will be analyzing in this project is batting statistics for baseball players collected from Major League Baseball (MLB) games. The baseball data set is Lahman's Baseball Database which has baseball statistics going all the way back to the 1800s I will mainly be focusing on batting statistics from the year 1955 to 2020. Out of the statistices in this dataset I will calculate Batting Average (number of hits is then divided by the number of times that the batter gets a chance to hit during an At Bat) and On-base-plus slugging (adds the hitter's on base percentage (number of times reached base—by any means—divided by total plate appearances) to their slugging percentage (total bases divided by at bats)) and use these two values as predictors to my model.

The batting data is generated by at bats taken by the batter and their performance during that at bat. Whether they strikeout, get on base, score, etc. The outcome of the batter performance also depends on the performance of the pitcher during that at bat, this interaction will be approximated as a stochastic process where we will be predicting the future batting averages and OPS of these players. OPS and batting average are the most useful statistics in determining the overall performance of a batter.

This analysis will be done on statistics of yearly frequencies to forecast batting statistics for future years.

No transformation has been done on the data yet.

2. Formal model of data-generating process:

The batting average and OPS statistic do both have significant autocorrelation and will be modeled to follow an AR(1) process.

The AR(1) process will

The AR(1) model will be formulated by: $y_t = \alpha y_{t-1} + \varepsilon$

There error is: $\varepsilon \sim N(0, \sigma^2)$

3. Discussion of the statistical model:

The model captures the data-generating process because it captures the information of the lagged predictors taking the information into account to make future predictions for future years. The error term will take all other factors that are not captured in the data such as pitching faced, etc.