In 2003, Michael Lewis released a book that discussed Billy Beane, the general manager for the Oakland A’s, and his new and interesting approach on running a major league baseball (MLB) team. Mr. Beane was utilizing an analytics software known as *sabermetrics* to predict player’s talent based on their previous stats. In fact, the data-driven method was adopted by the Boston Red Sox who credited the technique as being a main reason for their World Series victory in 2004; a win which broke the long standing curse of the Great Bambino. The success of the Red Sox helped propel data science into the main stream of the MLB as it is today. Due to my love for the game (and especially the Red Sox), I found it fitting to apply my newly learned statistical modeling skills to an MLB batting dataset.

Baseball is truly a unique sport that favors endurance and consistency. Every team plays a 162 game season (not including post-season games). The games themselves are quite long with a minimum of nine innings being played out in a 9-on-9 fashion. The core rules of the game allow for many different stats to be recorded. Since baseball is played in a discretized pitch-by-pitch manner, it is easy to collect quantitative data on each player’s performance. In fact, the dataset chosen for this work only incorporates batting statistics. This excludes fielding, pitching, and even overall team stats that could also be analyzed. With so many stats being generated by the game, it is easy to see why the MLB is an excellent source to gather and study data.

The overall dataset consists of batting statistics ranging from 1955 to 2016 for both the American and National Leagues in the MLB. The data was refined such that only MVP-qualifying seasons were looked at. The criteria to qualify for MVP is having an average plate appearance (PA) no lower than 3.1. PA can be described as:

.

Once the given criterion was applied, the dataset was found to have 6863 rows.

**statmod package**

As I do love baseball, I also have a strong interest in coding. Learning how to use a “black box” statistical modeling package can absolutely be done. However, I instead decided to spend the time to code my own modeling tools and apply them to the chosen dataset. With this approach, one can achieve a more fundamental understanding of not only how statistical modeling techniques work, but also why they are effective.

* Introduction
  + Reasons for choosing dataset
    - Billy Beane of Oakland A’s using sabermetrics
    - Boston Red Sox winning the World Series in 2004 by applying similar approach
    - There is a relatively large amount of data compared to other sports
      * 162 games per year
      * 9 batters per team
      * Rules allow for many different stats to be recorded
  + About the dataset
    - Mostly discontinuous data
    - Batting stats of all MLB players (AL & NL) ranging from 1955 to 2016
    - Only included MVP-eligible players
      * PA >= 3.1
    - Different ways to view the data
      * Player’s stats over their career
      * One large dataset of stats
  + Building a stat modeling toolset
    - Incorporate my strong interest in coding into the project
    - Achieve fundamental understanding of the modeling approaches
* Statistical modeling approaches used
  + OLS
    - Searching for high correlation coefficients with multivariate OLS
      * Iterate between all combinations of both two and three inputs with one output
  + Regularized regression
    - Use Elastic net to further analyze best fits from OLS
    - Attempt to make educated guesses on some non-linear equations
  + Kriging (or maybe neural network?)
    - Predict players cumulative rate of failure
* Results