Hitters Data Analysis

William Arnost, Daniel Crouthamel, Richard Palmer

# Introduction

The goal of this analysis is to determine which baseball performance metrics influence a player’s salary. We will conduct a modeling exercise to see which features are useful in predicting salary and test different modeling techniques. We will also perform a two way anova analysis to determine the effect of league and division on salary.

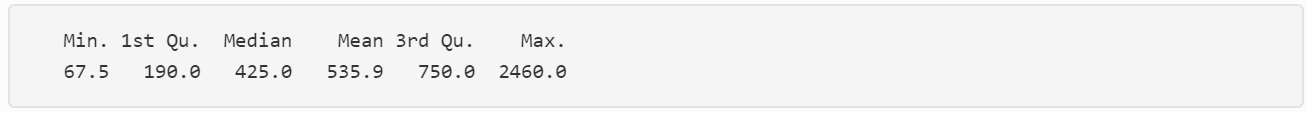
# Data Description

This data set contains salary information and career statistics for major league baseball players. The salary data was original from Sports Illustrated (April 20, 1987) while the career statistics are from the 1987 Baseball Encyclopedia Update published by Collier Books. We sourced the data from Kaggle (<https://www.kaggle.com/floser/hitters>) and it is also used in the ISLR R-package and is used in the textbook "An Introduction to Statistical Learning with applications in R" by G. James et al. (2013).

The data contains 322 observations, 1 per player, and 20 variables. The variables include salary and career statistics like Home Runs, times At Bat, Errors, and others.

# Exploratory Data Analysis

The salary variable will the response variable for modeling, so let’s look at that first. A quick summary shows that the minimum salary is $67.5K and the maximum is $2.46M. There are 59 NA values in the salary variable, so we will need to exclude those values for modeling (About 18% of the dataset).



The first two figures below are the histograms of Salary and logSalary respectively. We can see in the first histogram that Salary, untransformed, is very right skewed, with the maximum salary being almost 6x the median salary. We will need to transform the variable for modeling to deal with non-normality in the residuals.



In the second two graphs we compare the distribution of salary for two factors, League and Division. The distribution of salary in the American and National Leagues look similar, although the American league has a few outliers at the high end. There are more differences in the East and West divisions, with the east division having more salaries in the $1M+ range and the West division having more in the less than $250K range.

# Objective 1: Modeling

## Problem Statement

We would like to know which career statistics are indicative of a player’s salary. We will use a predictive model to assess which statistics, if any, are important. We will used variable selection techniques to help fit the best model.

## Model Selection

## Assumption Checking

## Compare Competing Models

## Parameter Interpretation

### Interpretation

### Confidence Intervals

## Final Conclusions for Objective 1

# Objective 2: Advanced Analysis

## State the method

We will conduct a Two Way ANOVA test to see if there is a difference in mean salary when considering the League and Division of a player.

## Main Analysis Content

## Conclusion / Discussion

# Appendix

## Well Documented Code

## Summary Graphics and Tables