**Predicting the Long Ball**

By: William Paulson

**Abstract**

The goal of this project was to perform linear regression to create a model for predicting a season’s home run total for a player. Predicting home runs will help teams make better decisions about acquiring new players. I worked with baseball statistics scraped from [Baseball Reference (Links to an external site)](https://www.baseball-reference.com/).This dataset allowed me to have the stats for each player broken down by season. The Lasso model I created had a R-squared score of .937 on the test set and a R-squared score of .935 on the validation set. RBIs, Rs, SOs, and BBs exhibited the highest correlation to the independent variable.

**Design**

This project originates from the Oakland Athletics. The batting data from [Baseball Reference (Links to an external site)](https://www.baseball-reference.com/) allows for modeling of home run production based on other batting statistics. Creating a model for predicting home runs would allow the Oakland A’s to better identify new players to acquire to bolster their offensive production and excite the fans.

**Data**

The Baseball Reference dataset I scraped contained 3,556 unique data points with 20 initial features. This data encompassed over 500 different players from each team. The most important dependent variables included RBIs, Rs, Ws, and SOs.

**Algorithms**

**Web Scraping and Linear Regression**

1. Scrape 2021 team web pages to grab links of all players who had an at-bat.
2. Scrape over 1200 player pages to determine if they are positional players
3. Scrape of remaining 500+ pages for batting statistics broken down by season
4. Clean the data of missing data, poor features, and poor formatting
5. Perform linear regression modelling to fit a training data set
6. Curate features that negatively impact modeling results, and use regularization to standardize dependent features.
7. Test different linear regression models such as Lasso, Ridge, and Elastic Net to find best fitting model.
8. Visualize data using tools like Seaborn, matplotlib, and SKlearn.

I chose to focus the analysis on positional players to avoid issues issues of poor sample size of at bats by pitchers.

**Tools**

* Beautiful Soup and Requests for web scraping
* Pandas and Numpy for data frame manipulation
* Statsmodels, Seaborn, Matplotlib, and SKlearn for linear regression

**Communication**

In addition to the slides and visuals presented, [Baseball Reference (Links to an external site)](https://www.baseball-reference.com/) will be embedded on my personal website and blog.