Introduction

The goal of this assignment was to predict full-season on-base percentage for the year 2019 based on the values of performance metrics from March and April of 2019. This assignment was broken into three primary pieces: Cleaning and Examining the Data, Feature Selection and Alternative Regression Methods, and OLS Model Creation. The purpose of each section and results will be summarized.

Cleaning and Examining the Data

Initial Data Cleaning and Formatting

In this section, the goal is to initially look at the data, clean and change the format of variables as necessary, and visualize the data using histograms, boxplots, and correlation plots. Percentage data that was initially read in character form was transformed into numerical data for future analysis.

This data did not contain significant outliers that needed to be imputed, as evidenced by the histograms and boxplots. From the correlation plots, it is apparent that most of the groups created, including box score counting statistics, base ratio statistics and plate discipline statistics, had clear correlation with full season on-base percentage. The exceptions to this are the batted ball type statistics, which is useful information for future model building.

Feature Selection and Alternative Linear Regression Methods

This section involves using Ridge and Lasso regression, as well as the Bootstrap method for resampling. The primary goal of this section is to narrow down which features should be included in the final regression models. Some preemptive feature selection was done to prevent multicollinearity (such as including plate appearances, but not including at bats). Predictions were generated using Ridge regression and then the Lasso was used to narrow the model to the most important components.

The Lasso narrowed the most important predictive variables down to March-April on-base percentage and March-April hits. This is an intuitive result, as it says that among statistics accumulated from March to April, hitters who get on-base most proficiently are more likely to continue to get on-base at a proficient rate and vice versa, and that hitters who accumulate large numbers of hits in March-April are more likely to have higher full-season on-base percentages and vice versa. This knowledge will be helpful later in devising a simple model with good predictive power for its simplicity.

Model Creation and Performance

This section is where the final models are created and evaluated. The stepwise model, which uses plate appearances, hits, home runs, walks, strikeouts, isolated power, on-base percentage, O-Swing% and Z.Contact% was the chosen champion model. The predictions of the stepwise model were exported as a CSV file along with player ID numbers, teams, and player names for viewing convenience. The stepwise model had the lowest mean squared error at 0.0007880487, the highest multiple R-squared value at 0.4417, and an Adjusted R-squared value of 0.431, which was close to the multiple R-squared value and suggests there were not many extra unnecessary variables in the model. March-April on-base percentage and walks had the highest coefficient magnitudes at 0.1795152 and 0.1380915 respectively.

The simple model also performed admirably, as it had an MSE of 0.0008641401, a multiple R-squared value of 0.3878 and an adjusted R-squared value of 0.384. This demonstrates that if necessary, at least 38% of the variation in season-long on-base percentage can be predicted by March-April on-base percentage and hit totals, which makes sense given that March-April on-base is a direct component of a player's season-long on-base percentage.

Concluding Thoughts and Future Extensions

While it is not prudent to read too heavily into small samples, such as statistics accumulated from March to April in a full season, this study demonstrates that it is possible to predict as much as 44% of the variation in full-season on-base percentage from statistics accumulated from March to April. This could be useful in making decisions about early season roster moves such as minor league demotions, trades, and waiver claims.

One interesting extension on this could be to include Statcast data, such as hard-hit rate, barrel rate and expected batting average to give this model further predictive power. This study demonstrated that power metrics such as HRs and ISO have predictive power in determining season-long on-base percentage, which motivates looking at base components of these power metrics that are less result-oriented for predicting season-long on-base percentage.

Using similar methodology for statistics accumulated over larger samples, such as from March-May would be interesting, as it would provide further context on when to buy into performance jumps and declines during a season. This type of analysis could also be conducted for other key performance metrics, such as wRC+ and wOBA.