**Impact of Pitching on MLB Team Winning Percentage**

**By Colin McMorrow**

**Abstract**

The goal of this project was to create an explanatory regression model to identify which pitching statistics can be reliable indicators of increasing win percentage in Major League Baseball. A valid model could then be utilized to drive pitching departments to value certain statistics over others when making personnel decisions, which will directly lead to increased wins in the league.

**Design**

Pitching data was acquired via BeautifulSoup web scraping from baseball-reference.com. Initial least squares models were constructed using the most basic statistics such as hits and home runs. Feature engineering was implemented to incorporate these basic values into more complex measures such as WHIP and K/W ratio, and final features were selected. After obtaining a more acceptable metric value for the target variable, the model was assessed for multicollinearity and normality, and finally validated via 5-fold cross-validation.

**Data**

The data analyzed for this project included all team per season averages for the MLB from 1970-2021; 1444 data points were acquired. Each row of data contained cumulative and average statistics for an entire single season for a single MLB team. The key elements that were chosen were ones that were direct results of a pitcher’s play.

**Algorithms**

Web Scraping

* HTML parsing via BeautifulSoup
* Data manipulation using pandas

Linear Regression

* Least squares model building with statsmodels/scikit-learn
* Feature selection using regplots and other visualizations
* Assumption testing with visualizations
* Cross-validation with scikit-learn K-fold splits

**Tools**

* Requests and BeautifulSoup for web scraping
* Pandas and numpy for data manipulation
* Matplotlib and Seaborn for visualizations
* Statsmodels and scikit-learn for linear regression models

**Communication**



