Capstone Project 1: Baseball and Linear Regression

By Scott W. Lew

Capstone Mini-Project: Data Wrangling

Description of Dataset

The data set consists of both batting and pitching statistics for all major league baseball (MLB) teams in the years 1876-2018. The data was scrapped from the website baseball-reference.com using the Beautiful Soup program. A batting statistics table and a pitching statistics table were merged for by joining on team and year columns. Then, tables for each season were concatenated to obtain all the MLB statistics dating from 1876 to 2018. 2019 statistics were not used because at the time of writing this report the 2019 season is not complete.

For this project, statistics were taken for all Major League Baseball (MLB) seasons from 1876 to 2018.

The batting statistics are made up of 28 features while the pitching statistics contain 35 features. These statistics will be described in greater detail in other sections of this report.

Data Wrangling and Data Cleaning

All batting statistics were downloaded and saved as csv files and then converted into pandas dataframes which were subsequently concatenated. This same approach was also used for the pitching statistics data. Then, the dataframes for pitching and batting were merged using an inner join on two columns: Team(Tm) and Year, which produced the entire dataset used for the subsequent analysis. The entire dataset consisted of 2815 rows and 65 columns.

For the initial linear regression analysis using 60 numerical statistics, it was necessary to remove rows from the data that had nan values, which left 1618 rows of data for regression models. Unfortunately, the regression analysis with these predictor variables required losing 1197 rows. However, as will be shown, the regression models were able to predict wins.