Linear Regression Project

Major League Baseball Team Season Winning Percentage Breakdown

By William (Wei-Lun) Mai

Abstract

The project aims to see if a Major League Baseball (MLB) team's season winning percentage can be fit by a linear regression model by using its batting, pitching, and fielding statistics of the season. Furthermore, we would like to identify key statistics which could be used to explain the winning percentage of a team over a given season. MLB teams could use the result to analyze their win rate and potentially improving it by focusing on the underperforming statistics.

Design

First of all, 12 out of 81 features were selected for fitting an Ordinary Least Square (OLS) linear regression model. Both pair plots and correlation plot were generated for preliminary analysis. Multicollinearity was discovered in the 12-feature model. After cutting down to 7 features, OLS, ridge, and lasso regression were conducted, and cross-validation showed that OLS and ridge regression both have the best scores. However, multicollinearity still existed. After further trimming down to 4 features, multicollinearity finally no longer existed, and an OLS and ridge regression model were fit on the dataset. Since both showed similar good train-test scores, I settle with the OLS model over simplicity and for easier interpretation.

Data

The dataset is web scraped from Baseball Reference for all teams for the seasons from 1980 to 2021. Each row of data consists of a team's batting, pitching, and fielding statistics of a certain year. We have a total of around 1200 rows in our dataset. There are 81 statistics recorded for each team each year. 12 features are selected for further analysis. During residual analysis, outliers were discovered and they were due to shorter seasons resulted

from strikes or covid. As a result, data from these four seasons, 1981, 1994, 1995, and 2020, are removed from the analysis.

Algorithms

Final target and features:

Winning percentage: wins / (wins + losses) of a given year

Runs per game: runs scored per game as the team on offense

Earned run average: 9 x earned runs / innings pitched as the team on defense

Saves: saves as the team on defense

Errors: errors committed as the team on defense

Linear regression models: OLS, Ridge, Lasso, and Elastic Net are used during analysis.

Summary of cross-validation scores:

	12 features OLS	7 features OLS	7 features Ridge	7 features Lasso	4 features OLS	4 features Ridge
Train Score	0.913	0.912	0.910	0.853	0.909	0.909
Test Score			0.922	0.858	0.922	0.922

Tools

- BeautifulSoup for web scraping data from Baseball Reference
- Numpy and Pandas for data manipulation
- Statsmodels and Sklearn for fitting different models and cross-validation
- Matplotlib and Seaborn for plotting

Communications

Presentation, write-up, and Jupyter Notebook files can be found on the following GitHub page:
https://github.com/zyzzyva1423/LR_Project