Data Mining Major League Baseball's Pace of Play Problem

Aaron Crowley, Zhuolin He, and Rachael Hageman Blair

Department of Biostatistics, State University of New York at Buffalo

University at Buffalo
The State University of New York

Introduction

Baseball has long been recognized as America's national pastime. Unfortunately, America's fondness for baseball has faded over the years, shifting towards football.

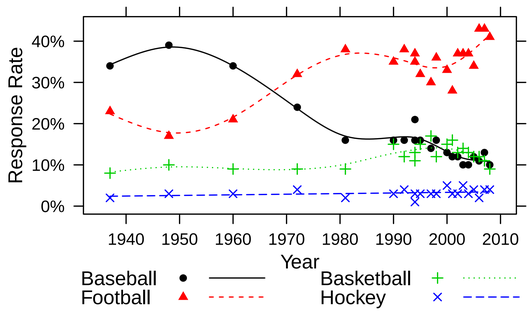

Figure 1: Results of Gallup polls asking "*What is your favorite sport to watch?*" [1].

At the same time, the length of Major League Baseball (MLB) games have steadily grown.

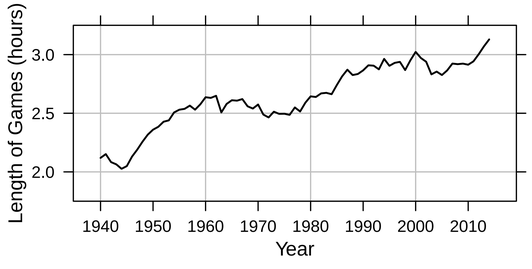

Figure 2: The average length of regular season MLB games from 1940 to 2014.

In the 2015 season, MLB implemented new rules to address the pace of play issue [2].

- Batters must keep one foot in the batter's box during an at bat.
- A timer is used to limit the time between each half-inning and pitching change.
- Players who do not comply with the new rules will be fined.

Objective

- Identify a robust predictive model for the length of MLB games to understand it's relationship with batting, pitching, and fielding performance.

Materials and Methods

Data

Yearly game logs are downloaded from the Retrosheet public website [3].

- The length of games is the response variable.
- Home and away team statistics are aggregated together for use as predictor variables.
- Nine inning games played between the 2000 and 2014 regular season are included in the analysis.
- Analysis dataset contains 33,208 observations and 26 variables without any missing data.

Statistical Models

Two statistical data mining techniques are utilized to extract parsimonious models for inference [4].

- *Lasso Regression:* a regularization method that supports feature selection.
- *Principal Component Regression:* a dimension reduction method

All data was processed and analyzed using R version 3.2.2 [5, 6, 7].

Results

Lasso Regression

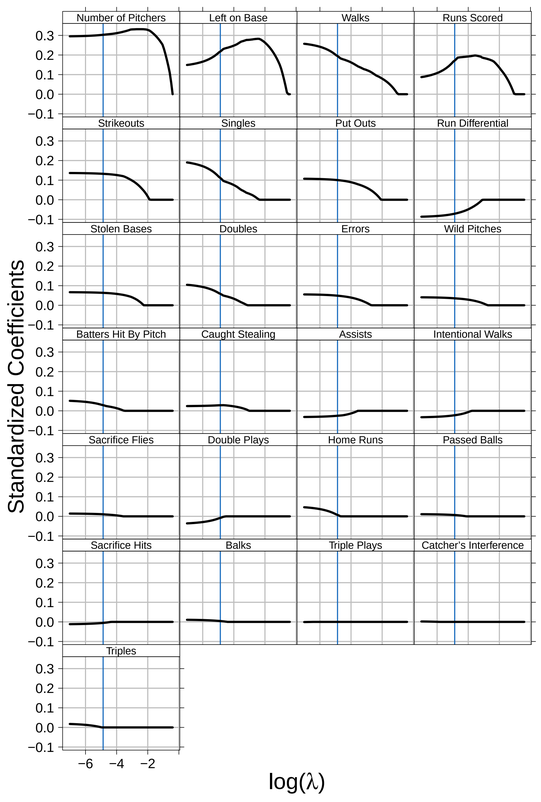

Figure 3: Profiles of the estimated standardized coefficients for the lasso regression models. Each curve represents a coefficient as a function of the complexity parameter λ. The blue vertical lines represent the best model determined by ten-fold cross-validation. Variables are displayed in ascending order by the absolute value of their coefficient in the best model.

Principal Component Regression

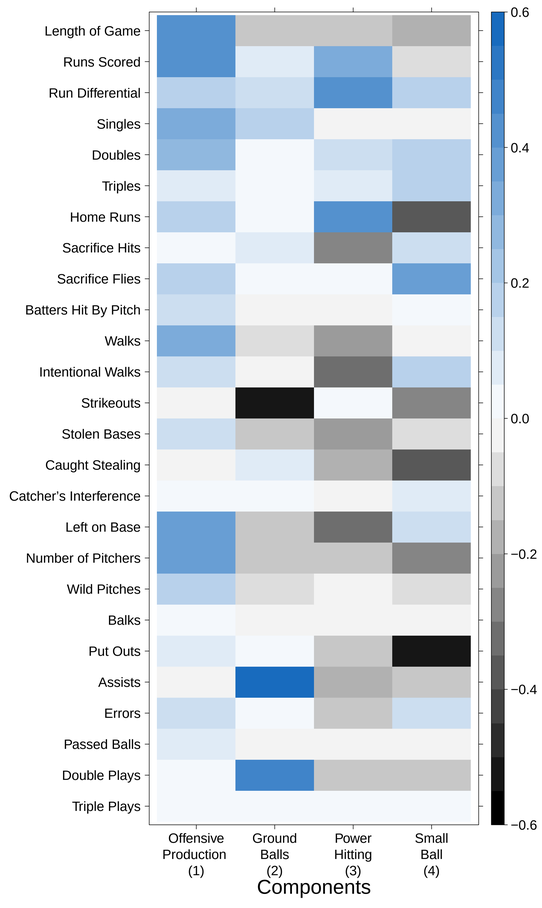

Figure 4: A heatmap of the principal component regression loadings. The loadings are correlations between each variable and the principal components in the model.

Discussion

- Offensive production is most predictive of the length of games.
- The number of pitchers and walks were found to prolong games the most.
- Close games take more time to complete than blow outs.

Future Work

- Evaluate the impact of the new pace of play rules after the completion of the 2015 MLB season.

Contact Information

Presented by Aaron Crowley

- acc34@buffalo.edu
- `https://aaronccrowley.wordpress.com/`

References

[1] Gallup. Baseball. URL `http://www.gallup.com/poll/1696/baseball.aspx`.

[2] MLBPA and MLB. MLBPA, MLB announce pace-of-game initiatives, replay modifications. MLBPA/MLB News Release, 2015.

[3] David Smith et al. Retrosheet website. URL `http://www.retrosheet.org/`.

[4] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. The elements of statistical learning, volume 2. Springer, 2009.

[5] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

[6] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1):1, 2010.

[7] Bjørn-Helge Mevik, Ron Wehrens, and Kristian Hovde Liland. pls: Partial least squares and principal component regression. R package version, 2:3, 2011.

Acknowledgements

The information used here was obtained free of charge from and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at 20 Sunset Rd., Newark, DE 19711.