STA631

Project Proposal

Rob Sanchez & Sydney Brougham

F2018

This project will explore historical baseball (MLB) statistics and produce analyses of *team winning ratio:* the number of wins divided by the number of games played in a season. Comparing a team’s winning ratio to its expected value can facilitate relative performance comparisons with other teams.

**Data**:

Lahman’s Baseball Database, complete batting and pitching statistics from 1871 to 2017, available from <http://www.seanlahman.com/baseball-archive/statistics/>.

**Goals:**

1. Linear regression

We will construct linear models predicting our response variable, *team* *winning ratio*. *Run differential* (runs scored – runs allowed) is already established in the field of baseball statistics[1] as an important component for the prediction of this variable, but we will examine adding others from the Lahman database that might improve model fit. Team overall pitching quality and batting performance are two candidate variables we aim to explore.

1. Pythagorean expectation formula

In the 1980s, Bill James empirically derived the so-called Pythagorean expectation formula[2] for team win ratio, based on the number of runs scored and allowed by a team throughout the course of a season. James’s original formula is specified as follows, where R = runs scored and RA = runs allowed:

More recent research[3] into this formula has found that an exponent providing a better fit resides somewhere between 1.7 and 1.95. We intend to estimate this exponent with a log ratio linear model for different eras of baseball history.

1. Weibull distribution fits

In his paper[3] on the theory behind the Pythagorean expectation formula, Steven Miller produced visualizations of runs-scored and runs-allowed histograms against Weibull distributions for all 2004 American League teams. These distributions are characterized by exponents derived from the Pythagorean expectation formula, and Miller left the National League team fits as an exercise for the reader. We aim to produce similar visualizations for the 2004 National League teams.

**References**

[1] Marchi, M., & Albert, J. (2014). *Analyzing Baseball Data with R*. CRC Press, Taylor & Francis Group: Boca Raton, FL.

[2] Pythagorean expectation. (n.d.). In *Wikipedia*. Retrieved November 1, 2018, from <https://en.wikipedia.org/wiki/Pythagorean_expectation>

[3] Miller, S. J. (2005). A Derivation of the Pythagorean Won-Loss Formula in Baseball. *ArXiv:Math/0509698*. Retrieved from <http://arxiv.org/abs/math/0509698>