Analytics for Local Collegiate Baseball League: Improved Statistics and Favorable Factors

Researchers: Hunter Hartke and Brett Schulte, with Faculty advisor: Dr. Jessica Kraker
University of Wisconsin – Eau Claire, Mathematics Department

Abstract

This project focuses on analytics methods based on traditional, historic statistics gathered for baseball players, as well as team win-loss records within a defined competitor framework. Methodologies for both team-level and player-level analyses were adjusted for the Northwoods League, including the local team Eau Claire Express, using historical data. We hope to be able to provide value to the local community by sharing some of the insights gained.

Assessments of individual player batting and pitching strengths were computed, based on statistics developed recently within Major League Baseball; explanation of these metrics are available on sites such as FanGraphs. Comparisons of these newer metrics are made to historical assessment measures.

Summaries of team records were gathered through the most recent four seasons, for 18-20 teams in the league. Various recursive-record-updating methods were considered for predictive purposes. The current analysis examines summary statistics values that appear to be most associated with streaks of wins or losses. Methods for modeling streaks by incorporating team statistics and other metrics are examined.

Goals and Problem

This outline to the right tracks the research, data management, and programming used to gather, clean, compute, organize, and evaluate the team and individual player data for the Northwoods League.

Data Compilation

All individual and team statistics were compiled directly from the Northwoods League website for each season. For individual players that played on multiple teams during a given season, his full season’s data was collected for use. Data was parsed to include only useful statistics. New statistics were created from this data modeled off of MLB statistics. Two metrics used frequently, FIP and wOBA, are relatively new statistics in the MLB to better represent the value of pitchers and hitters, respectively. Following the formula outlined on FanGraphs, we created a modified statistic for the Northwoods League players and teams.

Methods

For team data, each team’s game schedule was copied and evaluated in order to look at the team’s overall progression of their record throughout the season. A table was put together for each team of a cumulative record for any given team - the organization and computation of the cumulative records was done with for loops in R. The team’s record could be obtained for any given point of the season. The win-loss records across all teams were further used to calculate Bradley-Terry ratings. These ratings can be used as a way to calculate the probability of each team winning a head-to-head game, based on all the past games they have played (strength of record).

We were interested to identify reasons why teams performed well or poorly during specific parts of the season. Initially, we wanted to look at winning and losing streaks of teams during the 2018 season, but we soon realized that this might not capture the whole picture. Teams can win five games in a row, lose one, and then win the next four games, but looking at streaks only would not capture this 16-game span. Thus we developed the idea of a chunk. A chunk is a number of games (we looked at 15 and 20 game chunks), in which a team had a winning or losing percentage above a minimum value (in our case, we used .8). From this we performed exploratory analyses on one 20 game winning chunk, one 20 game losing chunk, and one 15 game losing chunk. Unsurprisingly, the 20 game winning chunk came from the team with the best cumulative record for any given team in our case, we used .8). From this we performed exploratory analyses on one 20 game winning chunk, one 20 game losing chunk, and one 15 game losing chunk. Unsurprisingly, the 20 game winning chunk came from the team with the best cumulative record for any given team in our case, we used .8). From this we performed exploratory analyses on one 20 game winning chunk, one 20 game losing chunk, and one 15 game losing chunk. Unsurprisingly, the 20 game winning chunk came from the team with the best cumulative record for any given team. Madison, and the 20 game losing chunk came from the team with the second worst overall record, Thunder Bay.

Our exploratory analysis consisted of looking at various individual and team metrics during the chunk and comparing that to their season averages. For the team statistics, we made a game-by-game comparison. For the individual statistics, we compared values from the chunk with the season, weighted by playing time. To evaluate a team’s offense, we used wOBA, OBP, SLG, extra base percentage, and team efficiency. To evaluate a team’s pitching, we used FIP, ERA, K, BB, K/BB, and percentage of unearned runs give up. Due to the limited defensive data available, no conclusions can be made about the effect of defense during a chunk.

Another initial hypothesis was about the variation of certain metrics. FIP (Fielding Independent Pitching), a relatively new statistic developed for Major League Baseball, is thought to provide a better (compared to the classic statistic ERA) idea of how good a pitcher truly is because it only factors in results that the pitcher has control over. Because FIP takes away the variability of the defense for a pitcher, it is thought to have less overall variability between seasons for the same pitcher. So, we decided to test the hypothesis that this variable statistic is better than ERA when evaluated across seasons. Pitching data was gathered for the 2015-2018 seasons, considering only those pitchers who pitched in multiple seasons. Each statistic, FIP and ERA, was computed for these pitchers; the standard deviation, $s = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$, as well as the simple range, was the statistic computed across the repeated seasons. In total, $x_{FIP}$ and $x_{ERA}$ were compared for 258 players (218 were evaluated across two seasons, 37 across three seasons, and 3 across four seasons).

The final methodology involves a rating computed from team records, to allow us to predict future game results. The model for Bradley-Terry Ratings is assumed to have a set number of items, K. We compare items i to j assuming one wins and the other loses. A rating $x$ for each item can be calculated with each iteration with the equation \[ x_{i,j}^{n+1} = \frac{x_{i,j}^n}{1 + \frac{x_{i,j}^n}{x_{j,i}^n}} \] where $n$ can be calculated for period $n$ by using the $x_{i,j}$ from n to calculate it in the iterative update. Note that $W_i, W_j, and p_{i,j}$ are the total number of comparison wins for team i, total number of wins for team j versus team i, and the total number of wins for team j versus team i. Thus $W_i = \sum_j W_{i,j}$ equals the total number of comparisons won by item i, and $p_{i,j} = W_j / (W_i + W_j)$, with $p_{i,j} = 0$ are the number of comparisons between i and j.

In each period n, the $x_{i,j}$ for each team is calculated with each game outcome impacting p calculations for all teams. Each team, at n=0, will start with $W_{i,j} = 1$ and $p_{i,j} = \frac{1}{2}$, as well as compared to an equal start for all teams, and we use the constraint $x_{i,j} + x_{j,i} = p_{i,j}$ for all $i,j$. This constraint implies that for each team is equal to .05 before the season starts because each team has the same chance of winning a game as another team, with K=20 representing the total number of teams playing. Finally, these are used to calculate $x_{i,j} = P_{i,j} / (1 + P_{i,j})$, which is the probability $p_{i,j}$, with n (between 0 and 1) reflecting the iteratively-updated rating incorporating the team’s record to date. These constants $x_{i,j}$ are computed for the 2018 regular season to integrate strength of teams’ season records into predicting playoff results.

Some Conclusions

Comparison of pitching stats (ERA vs FIP):

If there is a significant difference between the two metrics, then the histogram should be centered around 0. The histogram, however, is not centered around 0, instead around some value between 0 and 1.

From this, it is clear that ERA tends to be a much more variable statistic than FIP for pitchers in the 2015-2018 seasons.

If $\log(x_{ERA}) > \log(x_{FIP})$, then we run a simple sign test to determine if the log(ERA) tends to be larger than log(FIP). This test results in a P-value of 3.10^{-11}, meaning that there is extremely strong evidence that pitchers’ ERA tends to be more variables than their FIP. The values of $x_{ERA}$ tend to be more extreme than those for $x_{FIP}$, hence the need to examine the logarithms. We discuss possible extensions of this analysis in Future Work.

Predicting playoff results: The Bradley-Terry package (calculated with the Bradley-Terry package in R) across the full season were not accurate for predicting team performance in the playoffs. Teams that were expected to win in the first round of the playoffs but hurt them that wasn’t predicted to win the game. However, if the Bradley-Terry ratings from about 180 games before the playoffs (after the season break) are calculated, it’s more accurate than the overall season ratings. Going back 150 games before the playoffs gives an even more accurate prediction of how the playoffs will turn out.

While we could talk about the iterative ratings work better for the latter part of the season, the winning-percentages in the “chunks” directly prior to the playoffs do not provide meaningful insight. This may be due to several factors: teams may rest their better players in the last few regular-season games; other players play in the playoffs because of the playoffs returning to college; or, the elimination format of the playoffs may not give teams the same opportunity that a series would. More playoff games would give a better overall picture of which team is better, and further examination of individual-player effects is discussed below.

Future Work

To continue to test the variability of FIP and ERA, we would like to consider number of innings pitched as a weight in our test in order to factor in players that pitched more, and exclude those who pitched few innings. Future work also includes tests about advantages of new batting metrics compared to historic statistics.

The KRAECH rating is a poor job at predicting playoff results, which leads us to believe individual players have a large effect on the playoffs; especially because most rounds are single elimination. In order to incorporate individual players into a predictive model, we need a value to assign to each player. The MLB uses WAR, defined below, as a single number to give explain the amount of value each player provided to his team that season. We would like to pair WAR with our KRAECH ratings to predict the playoffs, but many pieces to WAR are difficult to identify at a collegiate level.

WAR = \frac{Battting Runs + Base Running Runs + Fielding Runs + Positional Adjustment + League Adjustment + Replacement Runs}{Runs Per Win}

Review of how WAR–“like” statistics are computed for players with only traditional statistics could allow us to devise a similar measure.

On why, we would like to continue our exploratory analysis of winning and losing chunks by analyzing more chunks to look for consistent factors that led to both success and failure over periods of time, and devise tests to assess the strength of those factors.

Acknowledgements

This research project was funded by the Office of Research and Sponsored Programs (ORSP) during the 2018-19 academic year.

Computing support and research space from the Mathematics Department.

Data

All data gathered during October 2018-April 2019 from https://northwoodsleague.com/eau-claire-express, open to public access.

References

For Bradley-Terry package in R: https://cran.r-project.org/web/packages/BradleyTerry2/vignettes/BradleyTerry.pdf

Definitions for various statistics available at: https://www.fangraphs.com/, or notes available upon request.

Computation: Tableau v. 10, R v. 3.5.0 to 3.5.3 (2018-2019 versions), and Excel.