Baseball Player Performance and Its Factors

Vail Dorchester and Jackson Mediavilla

Project Statement/Motivation

We wish to examine the correlations between numerous factors and player performance in high stake situations. Data mining game events may reveal correlations that are not expected or easily identifiable. We will investigate which factors contribute most to player performance.

We intend to focus on "outside" factors such as weather, stadium attendance, and closeness of the game. We may or may not find interesting correlations between player performance and other factors. Based on the correlation results, we may be able to infer causation on player performance.

The results of this study may reveal interesting information regarding player performance. This information could be used for several reasons including player recruitment, predicting individual player performance given situational information, and predicting the outcome of a game given a current game state.

Literature survey

In his paper, *The Statistical Mirage of Clutch Hitting*, Harold Brooks discusses the statistics behind the phenomenon of clutch hitting, or more colloquially, when a player hits better when under pressure as opposed to normal circumstances. In this paper, Brooks defines being 'under pressure' as being tied or behind by four or fewer runs in the seventh inning, including some more detailed cases of when the bases are loaded. Brooks argues that in order for a model that predicts clutch hitting to be accurate, it must not only show strong correlations between performance and pressure, but it must be consistent throughout multiple seasons as well and not a single-season-anomaly. When following this model, Brooks argues that one can define successful models for clutch hitting in singular seasons, but that these models show little to no correlation with each other across multiple seasons and are therefore inconsistent and inaccurate. He also argues that it is necessary to first define how well the average player does in a clutch situation, and then compare if a certain player or group of players consistently plays above that average across multiple seasons. For this value to be significant, he argues that it must be greater than one standard deviation from the mean in either direction, and that certain players would fall into this statistically significant category season after season, or even that the clutch rating of players would be consistent throughout seasons. He then goes to show that neither of these cases are true; clutch rating tends to vary from season to season, and individual players tend to vary their performance and rarely consistently bat better under pressure. From this, Brooks argues that clutch hitting is irrelevant and "a mirage at best" (Brooks, H, 1988).

In another article, Scott Lindholm discusses the effect of temperature on batting performance. Interestingly, Lindholm notes that there is a strong positive correlation between temperature and batting performance, but he also notes that this could be due to many, many factors, and that temperature may not necessarily be a causal factor in increased batting performance, but merely a related factor. Still, though, this data would prove useful in building models to predict when a player would perform at their best (Lindholm, S, 2014).

Proposed Work

In our literature survey, we found that there has been a rather extensive amount of research over evaluating individual player, and this makes sense. Producing that data and selling it to recruiters or scouts proves to be quite valuable, and in fact, most professional teams employ a professional statistician to do reports and research in this exact field. However, less research has been conducted regarding general player performance (performance of all players, not just one), and the factors that go into that. We found an article on the topic of clutch hitting, and they discussed that a trend of correlations should hold across multiple seasons and players, and this is an insight we will most certainly make use of.

With all of this in mind, a plan of our proposed work begins to take shape. First, it will be necessary to download the data itself. Most of the data is split into categories such as All-Star game score, postseason games, and regular season events (meaning every single event in every single game of the regular season), which is one of the sets we are interested in. This data is then further divided into seasons or collections of seasons, so it will be necessary to join the datasets into one large set, and then to clean and normalize the data. This will likely include writing scripts to scrape file systems for data given that each decade contains upwards of 400 files. The cleaning will involve making sure the format is uniform across these files, and accounting for missing values or choosing which values to ignore. For example, one of the factors we were interested in was weather, yet many of the files list wind speed, precipitation, and sky conditions as unknown. Because of this, it will be necessary to decide if there exist enough points to make use of those attributes, if they should be discarded. There is also a rather significant amount of superfluous data that will be culled down such as who the umpires were, who reported the game, who kept score, who translated the game, and so on. Again, this will likely involve writing scripts given the massive amounts of data that would be required to sift through.

Once this has been done, we will still need to roll the data into more concise cubes of information; these cubes will likely vary but could contain dimensions such as season, team, and batting average for example. After the data has been properly imported, cleaned, and sorted, we will likely begin by calculating and examining summary statistics for the attributes that we are interested in — sky conditions, crowd turnout, temperature, home team vs away team, etc. Once some of the interesting attributes have been identified, will create or reuse some popular models for player performance including batting average, isolated power, slugging power, or other models of our choice. With these models, we will need to again calculate some summary statistics including the standard deviation, the mean, and the quartile ranges. With this, it will be possible to identify whether a player's performance is comparatively high enough to be statistically significant, i.e., one standard deviation above the mean. Once this has been done, we can calculate correlation coefficients for these performance models and some of the outside factors or combinations of factors. From this, we will gain some insight and knowledge into what factors tend to correlate to higher performance, and we can begin to group these factors and see how a combination of factors plays in to performance. From this, we hope to create a reasonably accurate model that can accurately predict under which conditions baseball players tend to perform better.

Data Set

We are using the Retrosheet event dataset, which can be downloaded from retrosheet.org/game.htm. The data is currently downloaded on Jackson's computer and will soon be downloaded on Vail's computer as well.

The dataset contains every play in the included games. Each entry in the dataset reports both the event and the game state after the play. There are more than 95 different fields which may apply to a single event. The fields provide situational information such as score, identity of batter, number of outs, handedness of batter and pitcher, etc.

Evaluation Methods

We can perform basic statistical analysis to discover relationships between a range of factors and player performance. This will require defining which situations qualify as "high stake" and a method to evaluate player performance. A possible method for evaluating player performance is to create a function that assigns each team a probability of winning given the current game state, and then analyze how player performance (i.e. specific events) change the probabilities.

Tools

- Python
- Numpy
- Pandas
- Jupyter Notebook
- Excel
- Git, GitHub

Milestones

We will have completed Part 3 - Project Progress Report by 11:00 a.m. on Tuesday April 10. The submission for this milestone will be an updated, extended version of this initial report. We will review the sections *Motivation, Proposed Work, Tools, Evaluation,* and *Milestones*. We will also remove the section *Summary of peer review session.* Two subsections will be added to *Milestones*: *Milestones Completed and Milestones To-do*. And finally, a new section, *Results so far*, will be added at the end.

We will have completed the remaining parts of the project by 11:00 a.m. on Tuesday May 1. Part 4 is the Project Final Report. The submission for this milestone will be an updated, extended version of the Part 3 Report. The sections to be included are *Abstract, Introduction, Related Work, Data Set, Main Techniques Applied, Key Results, Applications,* and possibly *Visualization*. Part 5 is the Project Code and Descriptions. The submissions for this milestone include all of the source code on GitHub, a five-minute video discussing the project and the results, and a README to display on the GitHub main page. Part 6 is the Project Presentation. On one of the last three days of classes, we will present our project to the class. We will also submit our slide deck to our GitHub. Part 7 is Peer Evaluation and Interview Questions. The submission for this part of the project will be the peer evaluation form submitted individually to Moodle.

Summary of peer review session

Because one of us (Jackson) is in Rhonda Hoenigman's Sabermetrics course, we will most likely pull information or techniques from that class. Anything taken from that course must be cited. If we coordinate with the other group using the Retrosheet dataset, we must also cite that information or code.

Citations

Brooks, H. (1988). The Statistical Mirage of Clutch Hitting. *Baseball Research Journal*. Retrieved from http://research.sabr.org/journals/the-statistical-mirage-of-clutch-hitting

Lindholm, S. (2014, March 27). Hitting and temperature. Retrieved from https://www.beyondtheboxscore.com/2014/3/27/5546952/hitting-temperature-baseball