Proposal: Stand in the Sand

Sebastian Symula

Department of Statistics

University of Connecticut

February 14, 2025

Introduction If you've ever seen a baseball game where the starting pitcher is having an especially good performance, then you've probably seen those statistical graphics they put up on the broadcast. They often show the pitcher's ERA by times through the order and compare them to respective league averages, showing that the pitcher on the mound is clearly above average in WHIP and ERA more times through the order than average. Although (Moore, 2020) disagrees with this idea, it is a widely accepted trend in baseball. My goal is to confirm this trend and figure out why it happens.

Specific Aims If we could find a way to optimize pitcher performance, it would probably result in less runs allowed and less of a need to rely on the bullpen. One possible theory as to why some pitchers fair better later into the game is their pitch selection. Some pitchers use a wider variety of pitches than others, which might make a batter take longer to adjust. Additionally, some pitchers tire faster than others. This may lead to slower pitches and/or sloppy pitch placement. I want to investigate things like pitch types, pitch variation, and pitch speed to see if they correlate with pitcher performance each time through the order.

Data Description The data I will be using was collected from Baseball Savant and was obtained through the CSAS Data Challenge webpage. The data consists of every pitch from the 2024 MLB season. There are 701557 rows and 113 columns in this data set. This would fall under the category of an observational study as there were no treatments applied. The columns are a mix of categorical (ex: pitch type, description, player name) and continuous variables (ex: release speed, number of times through the order, bat speed). These will be some of the variables I will look at in searching for a correlation with pitcher longevity.

Research Design and Methods For my response variable, I plan to make my own column that denotes that pitcher's percent of plate appearances that end with a runner on base. I will make one of these columns for each time through the order. Additionally, I may drop rows from games where the starting pitcher was pulled before the third time through the order to avoid bias. To first answer the question of whether pitchers allow more runners on base the more times through the order they go, I can do an ANOVA test on the three columns to determine if the pitcher's performance differs for each time through the order. To determine what causes this assumed difference, I can make a separate dataset that shows the differences in independent variables (like average pitch speed, arm angle, and average unique pitch types) with total pitches and most common pitch type and also age. Also including dummy variables for at home vs away and most common pitch. Then I would run regression analysis to search for correlations with change in on base percentage from the first time through the order to the third time through the order.

Discussion I expect to see a gradual decline in pitcher performance through each time through the order. Whether that's actually due to times through the order or pitch count, is a point of much debate. For example, Lichtman (2013) suggests "it would behoove managers and pitching coaches to be much more mindful of a starter's 'times through the order' than his pitch count.", which is a stronger stance on the matter than some others are willing to take. There is also evidence that some pitchers seem to decline based on times through

the order, while some are much more dependent on pitch count (Brill et al., 2023). As for the cause of these declines. Handedness and home field advantage are both factors that are considered in this analysis. I expect to see a mixture of these results. It makes sense that both pitch count and times through the order may directly affect pitcher performance, but I also want to see what separates the longer lasting pitchers from the shorter term pitchers at an individual level. It is very possible that the results are intangible. Some pitchers may just be better than others without a solid statistical explanation. Additionally, The results may show that the only significant predictors are pitch count and time through order. A lot of the variables contained in this dataset are more batter-centric, therefore I may not have enough information to properly determine what sets some pitchers apart.

Core elements of Data Science I will most likely use packages like Pandas,SKlearn, ISLP, and Scipy for programming. Libraries like seaborn,matplotlib, and bokeh will be used for visualizations. Data management will be tricky due to how large the dataset is. I will have to avoid using expensive techniques or at least limit them to small subsets of the data. I will probably end up making multiple smaller datasets and groupings based off of certain pitcher characteristics. There is quite a bit of missing data in this dataset, but it makes sense in the context of the data (e.g. missing value for bat speed when the batter doesn't swing at a pitch), some of which can be handled as binary for my purposes. As mentioned previously I will probably add a few columns that may be useful for my goals. For my actual analysis I plan to run many different linear and nonlinear regression models, as well as potentially doing cluster analysis to set apart the good pitchers from the less good ones. I plan to include many visualizations, consisting of things like box plots, histograms, scatterplots, and regressions. The scatterplots could be used to demonstrate the groups that are formed by pitcher quality over time and averages by time through the order. A bar plot could show the difference in batting average when pitcher and batter have the same dominant hand vs when opposite. The regressions could show their fit on the data and what that says about

the correlations. This data was all collected ethically during the 2024 MLB season and is public data, so any models made will be transparent in that the data is available. This data also does not contain any protected or sensitive characteristics of an individual, therefore any bias in the model should not negatively affect any individuals.

Conclusion I will use a dataset that contains every pitch from the 2024 MLB season to assess pitcher tendencies. Through methods like linear regression and clustering, I plan to investigate the major factors that lead to a pitcher's inevitable decline during a starting appearance in an MLB game. There are currently a few different stances that people have on the topic and I want to assess these current understandings and potentially challenge them. I hope my findings can provide fresh insight into how to make it farther into the game as a pitcher.

References

Brill, R., Deshpande, S., and Wyner, A. (2023), "A Bayesian analysis of the time through the order penalty in baseball," , 1–36.

Lichtman, M. (2013), "Baseball ProGUESTus: Everything You Always Wanted to Know About the Times Through the Order Penalty," .

Moore, E. (2020), "Pitch Quality 3: Times Through the Order," *Medium*.