Bayesian Modeling of MLB Data

Kevin Finn MSDS Candidate UVA

John Gallagher MSDS Candidate UVA

I. Abstract

The 2023 Major League Baseball (MLB) season marked a pivotal moment with strategic rule adjustments aimed at modernizing the sport, including base expansion, pitch clock implementation, and the prohibition of defensive shifts. This study systematically evaluates the implications of these changes on the determinants of season-long victories by conducting a comparative analysis of data from the 2022 and 2023 seasons. Utilizing a probabilistic modeling approach, the research investigates the evolving importance of predictor variables in influencing team success. The dataset, obtained from Baseball Reference, underwent thorough Exploratory Data Analysis (EDA), leading to a refined set of variables for modeling. The probability model, accounting for team performance variability, employed uninformed priors with Student-t and Gaussian distributions for beta and alpha parameters, respectively. Model checks, including trace plots and Arviz-assisted model comparison, identified optimal models for 2022 (pitching) and 2023 (batting). Stacked models introduced greater uncertainty, with pitching dominating in 2022 and hitting in 2023, while equal weighting resulted in too much uncertainty compared to weighted models. The study concludes with insights into the evolving landscape of baseball strategies post-2023 rule changes, highlighting the diminished dominance of power hitters and strikeout pitchers and the increased importance of 'small ball' tactics. The impact of rule changes on defensive shifts, pitch clocks, and base size is evident, emphasizing the adaptability required in the post-rule-change era and the absence of a singular winning formula.

II. Problem Description

The 2023 season ushered in a series of strategic rule modifications by Major League Baseball (MLB), reflecting a concerted effort to modernize the sport. Noteworthy alterations encompassed the expansion of bases, the introduction of a pitch clock, and the banning of defensive shifts marking unprecedented shifts in MLB's historical rule landscape. This project is designed to systematically assess the implications of these rule changes on the determinants of season-long victories. Through comparative analysis of data from the 2022 and 2023 seasons, the objective is to discern the evolving importance of predictors and investigate their potential impact on the predictive modeling of team success throughout a season. [2]

III. Data Description

The dataset for this project was sourced from the Baseball Reference website, renowned for its accessibility and comprehensive MLB data.[1] Leveraging the freely available and downloadable information on batting, fielding, and pitching for both the 2022 and 2023 seasons, a thorough Exploratory Data Analysis (EDA) was conducted to prepare the data for subsequent modeling. Given the extensive array of variables obtained, a judicious trimming process was undertaken, guided by the author's expert knowledge of baseball and its nuanced relationship to the recently implemented rule changes.

The refined dataset, detailed in Figure 1, comprises the final predictor variables for each of the three models. To enhance the robustness of the analyses, normalization techniques were applied to control for outliers in the data. The response variable for each model corresponds to the number of games won by each team. Post-EDA, six curated data frames were deemed ready for the subsequent stages of probabilistic modeling

IV. Probability Model

In this section, the intricacies of the probability model were thoroughly explored, with an examination of the representation of objectives within the established framework. Guided by domain knowledge, the expectation that games won should adhere to a Student-T distribution, given the inherent variability in team performances. This choice is motivated by the observed pattern where some teams excel significantly while others face substantial challenges, a phenomenon accentuated in the modeled years, with the 2022 Los Angeles Dodgers approaching the record for wins and the 2023 Oakland Athletics nearing a record for losses.

To address the uncertainty in the beta parameters, an uninformed prior distribution was selected—specifically, a Student-T distribution with a zero matrix as the mean (mu) and the identity matrix as the covariance matrix (sigma). The choice of a multivariate T distribution was justified by the need for wider tails to capture the variability of beta parameters within the analysis. Additionally, the prior distribution for alpha, representing the team's overall performance, was modeled as a Gaussian distribution with a mean of 81 wins and a wide standard deviation of 15.

Figure 2 provides a detailed illustration of the exact model employed for all six models, which, despite minor variations in the number of predictor variables, remained consistent in predicting wins. The only divergence in the models, not depicted in Figure 2, pertained to the betas, with pitching incorporating 11 betas and fielding comprising 4 betas.

V. Approach

The Approach section presents an elaborate account of the strategic methodologies deployed to address the identified problem. Following the construction of all six models, a series of checks were executed to ensure their accuracy. Trace plots and rank plots were instrumental in validating the models. Uniformity in rank plots, as well as, convergence and stability in trace plots was observed across all six models, with R-hat values indicating convergence.

The evaluation extended to prior predictive and posterior predictive plots, exemplified in Figure 3, offering a visual representation of the model's efficacy in capturing both prior and posterior predictive scenarios. The HDI plots further scrutinized uncertainty in mus and betas, critical for analyzing differences in predictor variables across the modeled years.

With six functional models, combining them into a stacked model for predicting wins necessitated careful consideration. Leveraging the Arviz package, a model comparison was conducted, evaluating their performance based on Leave One Out Cross Validation (LOO). As depicted in Figure 4 (2022) and Figure 5 (2023), pitching and batting emerged as the optimal models, respectively. The resultant weights from these evaluations were employed in creating the stacked model. To validate the accuracy of these weights, an alternative stacked model was generated with equal weighting, given the minimal disparity in LOO values.

The culmination involved the creation of stacked models, and HDI plots for each year and each weighted model to assess their predictive efficacy in determining wins.

VI. **Results**

Upon completion of the model evaluations, the discernment of results was contingent upon the chosen methodology. The overarching objectives revolved around the identification of variables influencing win prediction and the development of a model showing win uncertainty across respective years. A comprehensive analysis involved the comparison of beta coefficients across the models, showing variations in predictor variables.

The beta uncertainty plots underscored the significant impact of certain variables on the outcomes in 2022. Specifically, for batting, Runs Scored and On Base Percentage Plus Slugging (Slugging is the average amount of bases reached per at bat) Percentage Plus emerged as the most influential factors. In pitching, Home Runs Allowed and Walks Allowed took precedence, while fielding was notably affected by Total Zone Runs (the number of runs above or below average based on the number of plays made) and Chances (the number of opportunities he has to record an out).

In the transitional period from 2022 to 2023, a discernible shift in emphasis on predictor variables was observed, as indicated by the High-Density Interval (HDI) plots. The majority of beta coefficients gravitated towards zero, with only marginal deviations for those that exhibited more substantial changes. Noteworthy among these changes was the considerable reduction in beta values for Home Runs Hit and Strikeouts by a pitcher in 2023, suggesting a diminished significance of these skills following a rule change.

Upon stacking the models, both with uniform weights and weights derived from leave-one-out cross-validation performance, a heightened level of uncertainty was introduced. Notably, the stacked models exhibited greater uncertainty than any individual model. In 2022, the weighted model underscored the dominance of pitching, while in 2023, hitting assumed a predominant role in the weighted model. Furthermore, the equally weighted model manifested greater and too much uncertainty compared to its weighted counterpart in both years.

VII. **Conclusion**

In the comparative analysis of three pivotal facets within the game, before and after substantial rule changes that fundamentally altered gameplay dynamics, the study discerns a transformative shift in the trajectory of success. Notably, with numerous variables converging towards zero in 2023, it is conceivable that the era of dominance by power-hitting (i.e., home run) hitters and power (i.e., strikeout) pitchers has potentially reached its culmination.

The strategic implementation of defensive shifts poses a significant constraint on hitters, limiting their outcomes to strikeouts and home runs, which cannot be mitigated through defensive positioning. Simultaneously, the imposition of time restrictions on pitchers has curtailed the ability to consistently deliver triple-digit fastballs, necessitating an enhanced reliance on pitch movement to induce suboptimal contact rather than relying solely on swing-and-miss strikeouts.

The collective reduction of various variables in absolute value substantiates the assertion that there is no longer a singularly prescribed methodology for achieving success in baseball. In 2022, the prevailing strategy for victory involved assembling a lineup of home-run hitters complemented by a rotation and bullpen dominated by strikeout pitchers. However, the landscape altered markedly in 2023, where teams found success through a more diversified approach, incorporating doubles, triples, and stolen bases—an approach colloquially known as 'small ball.'

The increased significance of hitting in 2023 can be logically attributed to three pivotal rule changes. The prohibition of defensive shifts introduces defensive vulnerabilities, pitch clocks heighten the challenge of retiring hitters, and larger bases elevate the frequency of runners safely reaching the next base in closely contested plays. Consequently, the nuanced changes in gameplay dynamics underscore that the optimal strategy for success in baseball has evolved, rendering the landscape more diverse and multifaceted.

However, the work has its limitations. The availability of only one season of data with the new rule changes necessitates caution in drawing definitive conclusions; a more comprehensive analysis over subsequent seasons is warranted to identify overall trends. Additionally, the unequal distribution of variables among the three facets of the game, with fielding having fewer variables than batting and pitching, poses challenges in analysis. The difficulty in quantifying fielding, likely contributing to its lower weighting in the model comparison, is acknowledged. Finally, modeling the game of baseball proves challenging due to factors that are inherently hard to quantify, such as luck, weather, and the unique characteristics of each park. Given more time, a more nuanced consideration of these elements could have improved the predictive model.

Data used for analysis, code, and instructions on how to run code can be found on the authors' GitHub page. https://github.com/jjg5fg/Bayesian-Modeling-of-MLB-Data/

VIII. References

[1] Sports Reference LLC. Baseball-Reference.com - Major League Statistics and Information. https://www.baseball-reference.com/. 06 December 2023

[2] Rogers, J. (2023, March 29). 2023 MLB rule changes - Pitch clock, end of shift and more. ESPN. https://www.espn.com/mlb/story/_/id/35631564/2023-mlb-rule-changes-pitch-clock-end-shift-bigger-bases

IX. Appendix

Figure 1

Pitching

- ERA+
- Saves
- Hits allowed
- Runs allowed
- Home runs allowed
- Walks allowed
- Strikeouts
- Earned Runs
- Hit by Pitch

Hitting

- Hits
- Runs
- Doubles
- Triples
- Home runs
- Walks
- Strikeouts
- Stolen Bases
- Caught Stealing
- OPS+

Fielding

- Errors
- Double Plays
- Chances
- Rtot

Reponse : Wins

Figure 2

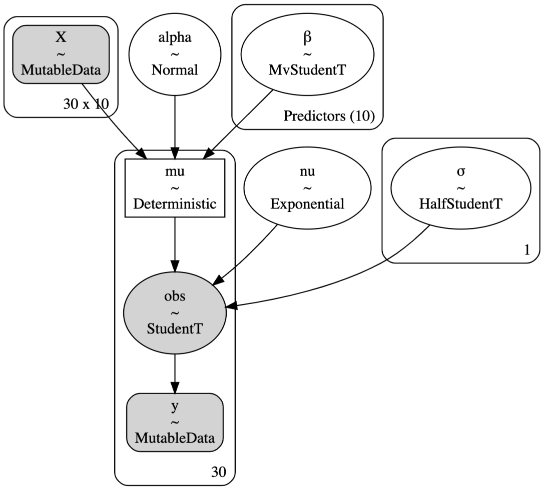


Figure 3 (Hitting 2022)

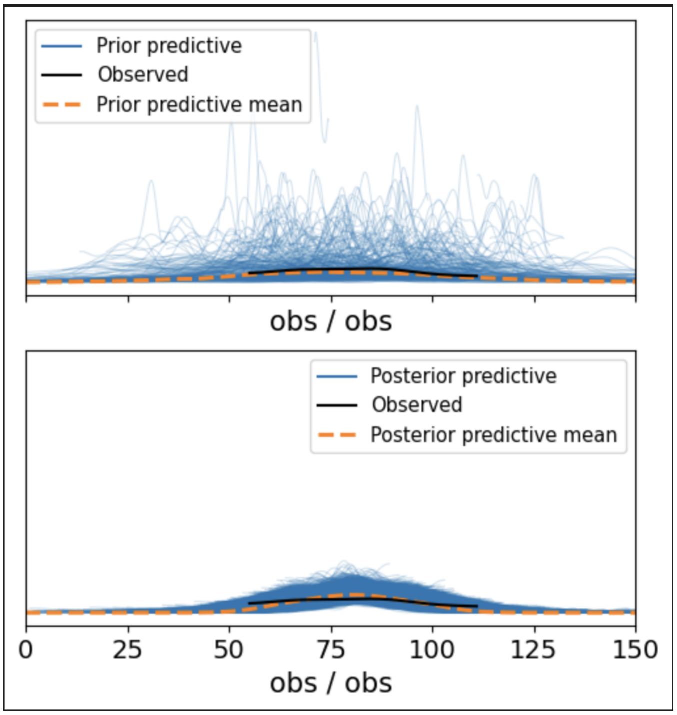
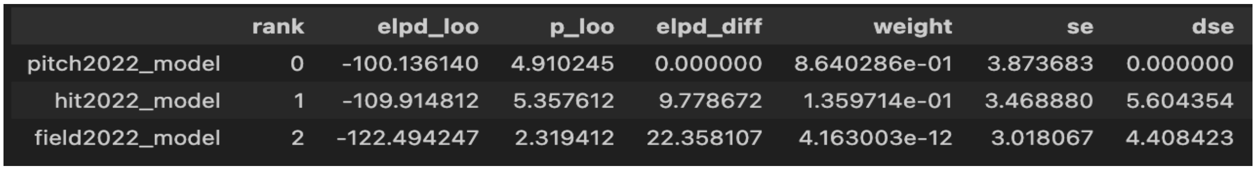
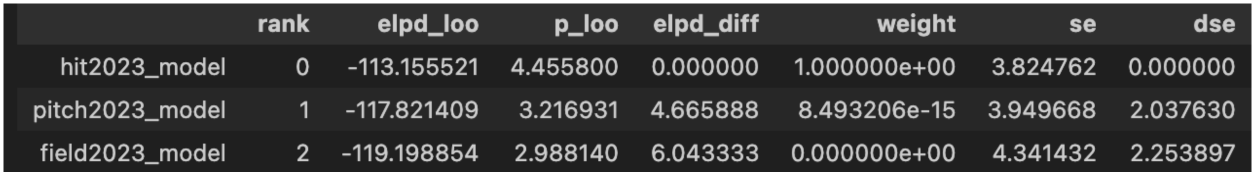


Figure 4



	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse
pitch2022_model	0	-100.136140	4.910245	0.000000	8.640286e-01	3.873683	0.000000
hit2022_model	1	-109.914812	5.357612	9.778672	1.359714e-01	3.468880	5.604354
field2022_model	2	-122.494247	2.319412	22.358107	4.163003e-12	3.018067	4.408423

Figure 5



	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse
hit2023_model	0	-113.155521	4.455800	0.000000	1.000000e+00	3.824762	0.000000
pitch2023_model	1	-117.821409	3.216931	4.665888	8.493206e-15	3.949668	2.037630
field2023_model	2	-119.198854	2.988140	6.043333	0.000000e+00	4.341432	2.253897