SML 320

Bayesian Analysis

A Bayesian Approach to Projecting the MLB Aging Curve

Authors:

Tucker Saland tsaland@princeton.edu

Contents

1 Introduction

Baseball, colloquially referred to as America's pastime, has seen rapid global growth in past years, with the MLB experiencing its largest year-over-year increase in attendance in 2023 (Connon, 2023). Star players are paid upwards of $50 million per year, with Shohei Ohtani, a Japanese phenom, recently inking a deal for $700 million (Drellich, 2023). With such substantial sums at stake, now more than ever, understanding the factors affecting baseball players' career longevity and performance trajectory becomes even more crucial for teams looking to maximize their investments and for players aiming to navigate their careers effectively.

Traditionally, sabermetricians have used techniques such as delta-averaging and k-nearest neighbors to produce point estimates for player performance in future years (Frontin, 2019). While useful in certain contexts, these methods are not optimal when predicting the performance of any given player. In the case of delta-averaging, the aging pattern for the average player does not necessarily translate well for outliers, players such as Shohei Ohtani, who do not fit the mold of an average player, which is especially worrisome for teams since these players often comprise the vast majority of their financial investments. The k-nearest neighbors approach, which incorporates aging patterns of players that are "close" to a given player, is complicated by the issue of quantifying "closeness" between players across seasons and eras. Importantly, neither of these methods can quantify the uncertainty of a player's future performance and only provide a point estimate rather than a probability distribution. As a result, building on the work of Frontin (2019), I intend to (1) project the aging curve of MLB players over time and (2) determine whether there are certain play styles and characteristics that lend themselves to player longevity. More precisely, I want to project players' WAR (Wins Above Replacement), a composite measure of a given player's contribution to team performance, over time to construct probability distributions for player performance rather than the point estimate predictions that are commonly in use. Understanding these probability distributions can be immensely valuable to front offices when making free agency and trade decisions. Suppose, for instance, that the GM for the Yankees receives a call from the Twins in February 2022 that

they are willing to part with Josh Donaldson with two years left on his contract in exchange for a stable of prospects. While the consensus is that the prospects will likely help the team in the long run, a 3.0 WAR 2022 season from Donaldson would bring the Yankees from a projected wildcard team to legitimate World Series contenders. In this scenario, the probability distribution of future performance could serve as a crucial tool for the Yankees' front office. If, for example, Donaldson is projected to have at least a 3.0 WAR in 70% of scenarios, that would make the trade far more enticing than if the probability was merely 20%. In this paper, I hope to more accurately quantify these uncertainties and contribute to the growing body of sabermetrics literature.

2 Data

The first dataset I consider as part of my analysis is a list of all baseball games from 2016-2022 with aggregate box scores, location, odds, and team records scraped from ESPN (Mohr, 2022). In addition to basic game information, the data rudimentary counting stats for pitcher and hitter performance (e.g., walks, strikeouts, and total bases) contains 13,439 observations of 43 variables.

Game <dbl>	away <chr>	away-record <chr>	awayaway-record <chr>	home <chr>	home-record <chr>	homehome-record <chr>	away-score <dbl>	home-score <dbl>
360403123	STL	0-1	0-1 Away	PIT	1-0	1-0 Home	1	4
360403130	TOR	1-0	1-0 Away	TB	0-1	0-1 Home	5	3
360403107	NYM	0-1	0-1 Away	KC	1-0	1-0 Home	3	4
360404108	SF	1-0	1-0 Away	MIL	0-1	0-1 Home	12	3
360404101	MIN	0-1	0-1 Away	BAL	1-0	1-0 Home	2	3
360404113	SEA	0-1	0-1 Away	TEX	1-0	1-0 Home	2	3

Figure 1: A sample of the games dataset

In addition, I obtain hitter-specific game logs from 2016-2022 from Fangraphs scraped using the baseballr package (Petti & Gilani, 2024). This dataset includes traditional hitting metrics, along with advanced hitter statistics (e.g., exit velocity, launch angle, wRC+). The dataset contains 320,351 observations and 293 columns. From Fangraphs, I also obtain batting statistics from 2010-2022 aggregated at the season level, along with a given position player's WAR (with subdivisions for offense, defense, etc.), a dataset which contains 8,268 observations of 339

variables. An analogous dataset of 2016-2022 pitcher game logs is collected from Fangraphs, which includes advanced pitching metrics (e.g., FIP, WHIP, and pitch mix) along with traditional counting statistics. As above, I collect measures of pitching performance and WAR from 2010-2022, which is comprised of 9,708 of 409 observations.

PlayerName	playerid	Date	Te...	Opp	season	Age	BatOrder	Pos	G	AB	PA	H
Bryce Harper	11579	2022-10-04	PHI	@H...	2022	29	1	DH	1	4	4	1
Bryce Harper	11579	2022-10-03	PHI	@H...	2022	29	3	DH	1	4	4	1
Bryce Harper	11579	2022-10-02	PHI	@...	2022	29	3	DH	1	4	4	1
Bryce Harper	11579	2022-10-01	PHI	@...	2022	29	3	DH	1	4	5	0
Bryce Harper	11579	2022-10-01	PHI	@...	2022	29	3	DH	1	3	4	0
Bryce Harper	11579	2022-09-30	PHI	@...	2022	29	3	DH	1	4	5	1

PlayerName	playerid	Date	Opp	teamid	season	Te...	HomeA...	Age	W	L	ERA
Zack Wheeler	10310	2021-09-28	@ATL	26	2021	PHI	A	31	0	1	2.571429
Zack Wheeler	10310	2021-09-22	BAL	26	2021	PHI	H	31	0	0	1.500000
Zack Wheeler	10310	2021-09-17	@N...	26	2021	PHI	A	31	1	0	1.800000
Zack Wheeler	10310	2021-09-11	COL	26	2021	PHI	H	31	1	0	1.350000
Zack Wheeler	10310	2021-09-06	@MIL	26	2021	PHI	A	31	1	0	0.000000
Zack Wheeler	10310	2021-08-30	@W...	26	2021	PHI	A	31	1	0	6.000000

Season	team_name	PlayerName	playerid	Age	WAR
2010	TEX	Josh Hamilton	1875	29	8.4392
2010	TBR	Carl Crawford	1201	28	7.7354
2010	TBR	Evan Longoria	9368	24	7.5364
2010	CIN	Joey Votto	4314	26	6.8910
2010	STL	Albert Pujols	1177	30	6.7756
2010	ATL	Brian McCann	4810	26	6.6870

Season	team_name	PlayerName	playerid	Age	WAR
2010	- - -	Cliff Lee	1636	31	7.3268
2010	DET	Justin Verlander	8700	27	6.6966
2010	SEA	Félix Hernández	4772	24	6.6821
2010	PHI	Roy Halladay	1303	33	6.1569
2010	FLA	Josh Johnson	4567	26	6.0465
2010	COL	Ubaldo Jiménez	3374	26	5.9748

Figure 2: Sample data from Fangraphs

3 Exploratory Data Analysis

I begin my exploratory data analysis by first considering league-wide trends from game box scores (Mohr, 2022). As illustrated in Figure 3, the vast majority of baseball games in the

2016-2022 sample have between 5 and 10 total runs scored. Home teams tend to score more runs on average, which aligns with the documented "home-field advantage" in existing literature (Jamieson, 2010) and likely stems from a team's familiarity with their home field in conjunction with increased crowd support. Most starting pitchers tend to have ERAs (earned run averages) between 2 and 4, and although league-wide averages vary year over year, generally, an ERA below three is considered elite, while an ERA above five is often seen as below-average. In terms of strikeouts, starting pitchers tend to throw between 2 and 8 on average, and throwing more than ten is exceedingly rare.

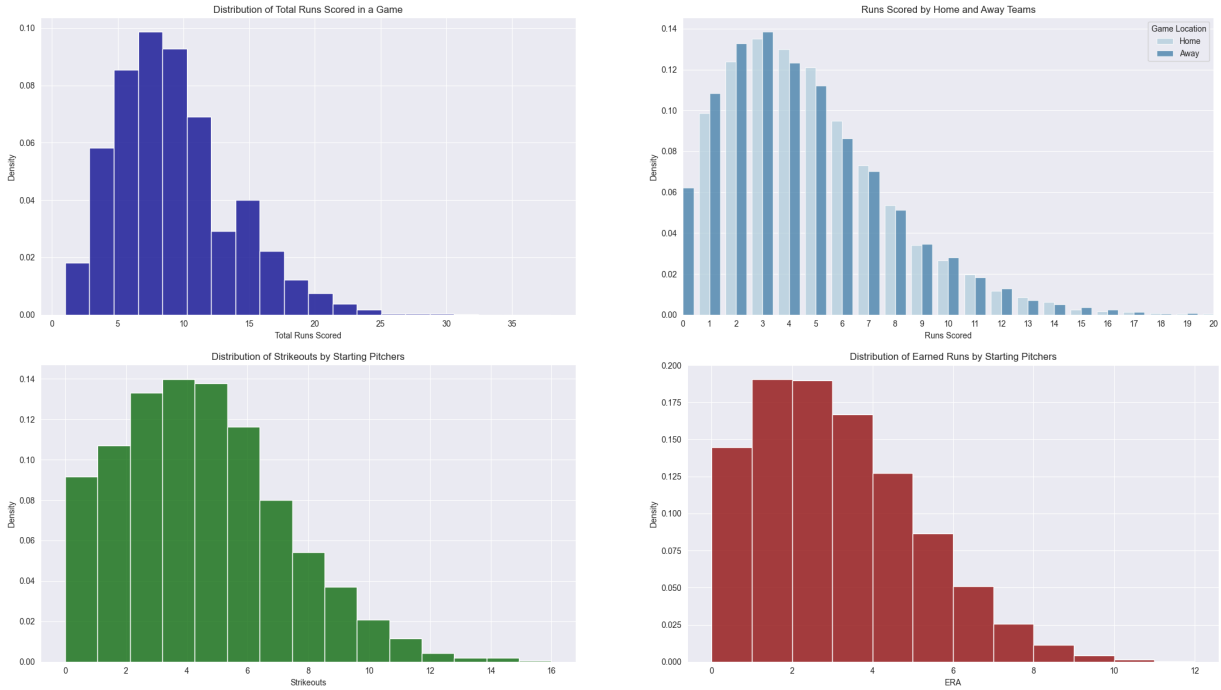

Figure 3: Distribution of in-game statistics

I consider league-wide batting trends over time among qualified batters (those with 3.1 PAs per team game), filtering out players with limited appearances and thus reducing the noise in the data that can arise from small sample sizes. As shown in Figure 4, slugging percentage (SLG), which measures the average number of bases gained per at-bat, On base percentage (OBP), how often a player gets on base accounting for walks, and batting average all steadily decreased over 2016-2022, with the trend being most pronounced for slugging. There has also been a marginal increase in strikeout rates over this span, peaking in 2020, although walk rates

have remained relatively constant. This pattern underscores a shift in the hitting environment of Major League Baseball, possibly due to the rampant use of foreign substances among pitchers during this time frame (Verducci, 2023). Prior to a crackdown in 2021, pitchers often applied sticky substances to the ball to increase spin rates, ultimately making the ball more difficult to hit, coinciding with the increased strikeout rates observed in the data.

Figure 4: Batting trends over time

I next examine the correlation between various hitting rate stats and how these correlations have evolved over time among hitters. I use wRC+ (Weighted Runs Created Plus) as a composite measure of a hitter's overall performance. wRC+ is a comprehensive metric designed to quantify the number of runs a player contributes to their team's offense, adjusted for ballpark effects and the era in which they play. Expressed relative to the league average, in which 100 denotes the league average, it enables the seamless comparison of players across teams and generations. As seen in Figure 5, wRC+ is most strongly correlated with OBP (with a correlation coefficient of 0.929) and SLG (with a correlation coefficient of 0.923), and this trend has remained constant over time. This finding aligns with my a priori expectations since their combined measure of on-base plus slugging (OPS) is frequently used as another metric to evaluate hitting performance (see Figure 6 and the correlation coefficient of 0.945). Strikeout rate (K%) has a moderately negative correlation with wRC+ (with a correlation coefficient of -0.392), indicating that as a player's strikeout rate increases, their overall offensive contribution tends to decrease, as measured by wRC+. This relationship underscores the intuitive notion that players who strike out less often contribute more to their team's offense, as they are more

likely to put the ball in play and create scoring opportunities. However, power hitters often accept a higher strikeout rate as a trade-off for the potential of hitting more home runs and extra-base hits, which could potentially be a worthwhile trade-off if the run potential from increased power hitting significantly contributes to the team's overall scoring, which could offer an explanation for the why the correlation between wRC+ and K% is not stronger.

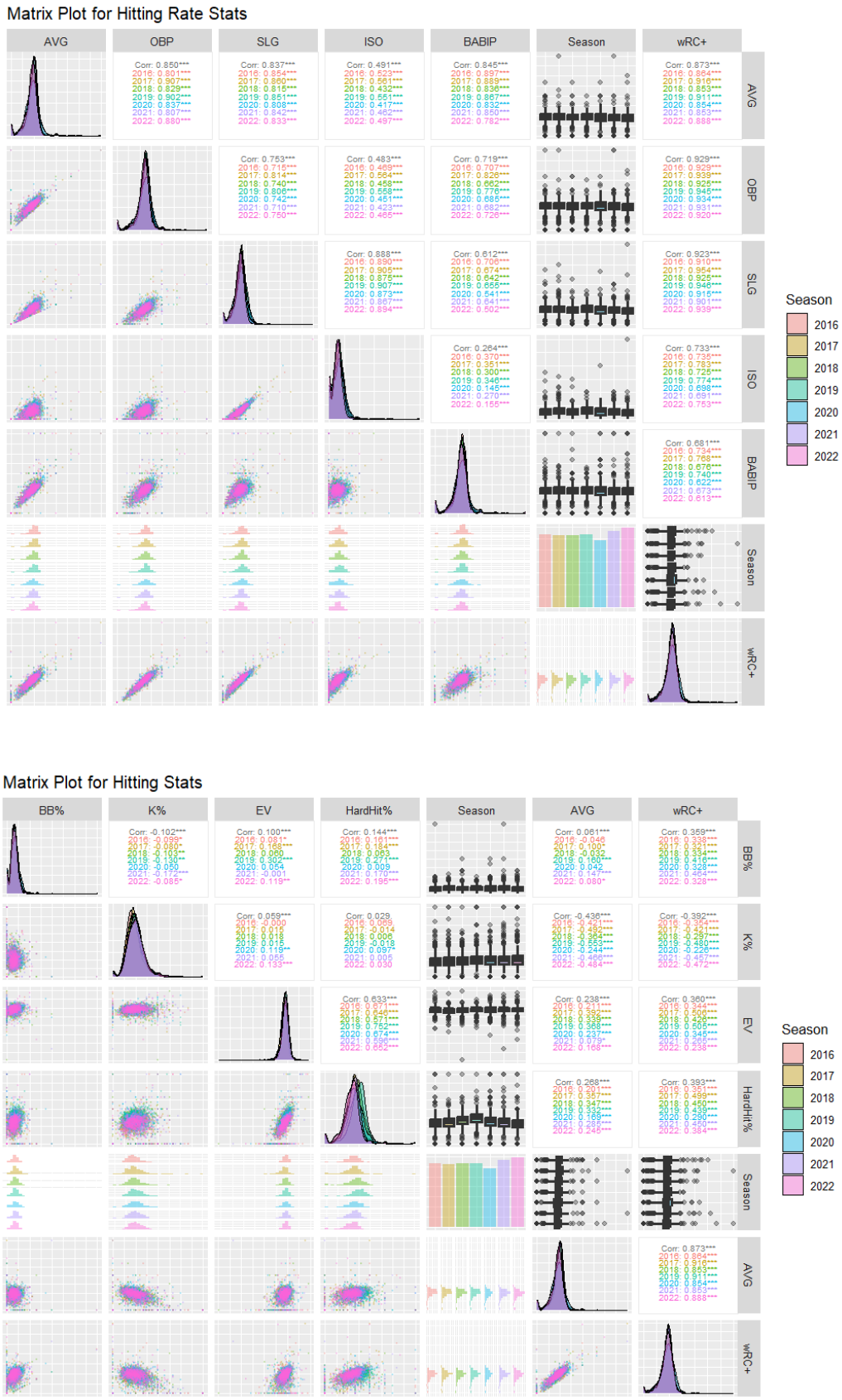

Figure 5: Hitting Matrix Plots

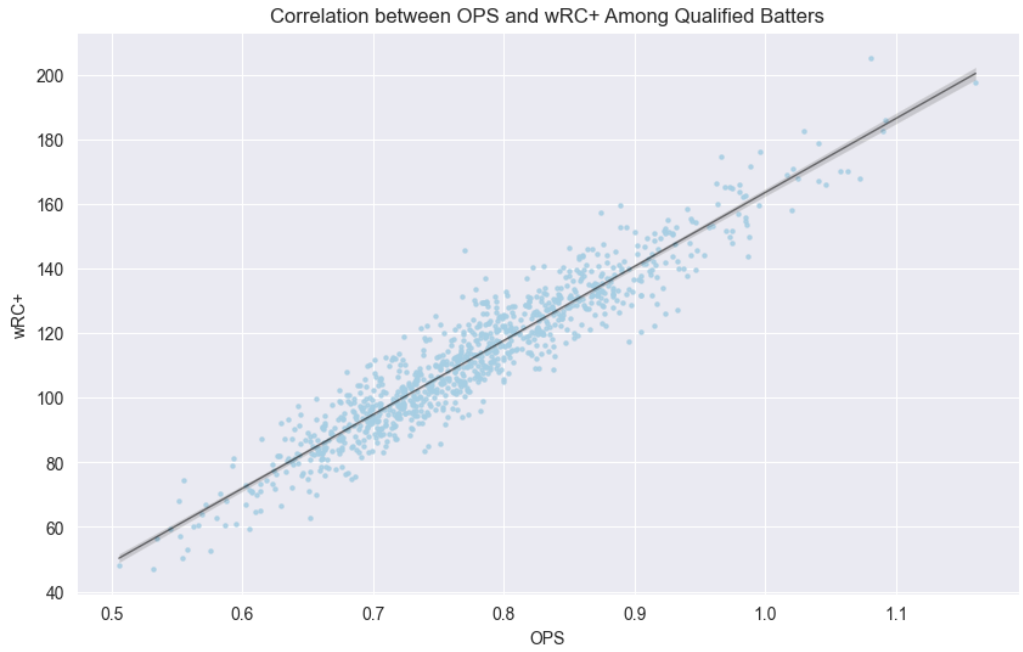

Figure 6: Relationship Between OPS and wRC+

Finally, I consider the persistence of WAR over time. My analysis excludes the 2020 COVID-19 shortened season since it had significantly fewer games than a full-length Major League Baseball (MLB) season and is not representative of the rest of the sample. In Table 3, I regress a given player's WAR on WAR lags over the previous five seasons, Age, Age^2, and a constant. WAR_{i-1}, WAR_{i-2}, and WAR_{i-3} all have positive coefficients that are significant at the 1% level, indicating that on average, a higher WAR in the recent past is associated with a higher WAR in the current season. This persistence is greatest for WAR_{i-1} (coefficient of 0.326) and decreases over time, which is consistent with the expectation that more recent performances are better predictors of current performance due to factors such as player development and injury history. The negative coefficient on Age and the positive coefficient on Age^2 (both significant at the 1% level) capture the non-linear relationship between age and performance, with player performance declining with Age at a decelerating rate.

VARIABLES	WAR_i
WAR_{i-1}	0.326***
	(0.0313)
WAR_{i-2}	0.182***
	(0.0303)
WAR_{i-3}	0.115***
	(0.0276)
WAR_{i-4}	0.0491*
	(0.0270)
WAR_{i-5}	0.00730
	(0.0253)
Age	-0.822***
	(0.201)
Age^2	0.0110***
	(0.00305)
Constant	14.78***
	(3.275)
Observations	1,120
R-squared	0.447

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

4 Bayesian Conjugate Priors

4.1 Zach Wheeler Strikeouts Per Game

Zach Wheeler is a starting pitcher for the Philadelphia Phillies who recently signed a contract extension worth $126 million for the 2025-2027 seasons (Stark, 2024). Known for his strikeout prowess, I was interested in exploring his strikeouts per game. Treating strikeouts as independent events, this question naturally lends itself to a Gamma-Poisson model. Let Y represent strikeouts in a given game, and $\lambda > 0$ be the rate at which these strikeouts occur. $Y_i|\lambda \overset{\text{ind}}{\sim} \text{Poisson}(\lambda)$, $\lambda \sim \text{Gamma}(s, r)$, and upon observing data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, the posterior model of λ is also a Gamma distribution with updated parameters: $\lambda|\mathbf{y} \sim \text{Gamma}\left(s + \sum y_i, r + n\right)$. I first tuned a $\text{Gamma}(3.89, 0.81)$ prior model using data from the COVID-19 shortened 2020 season (see Figure 7) and initially predicted Wheeler to have 4.81 strikeouts per game with a standard deviation of 2.44.

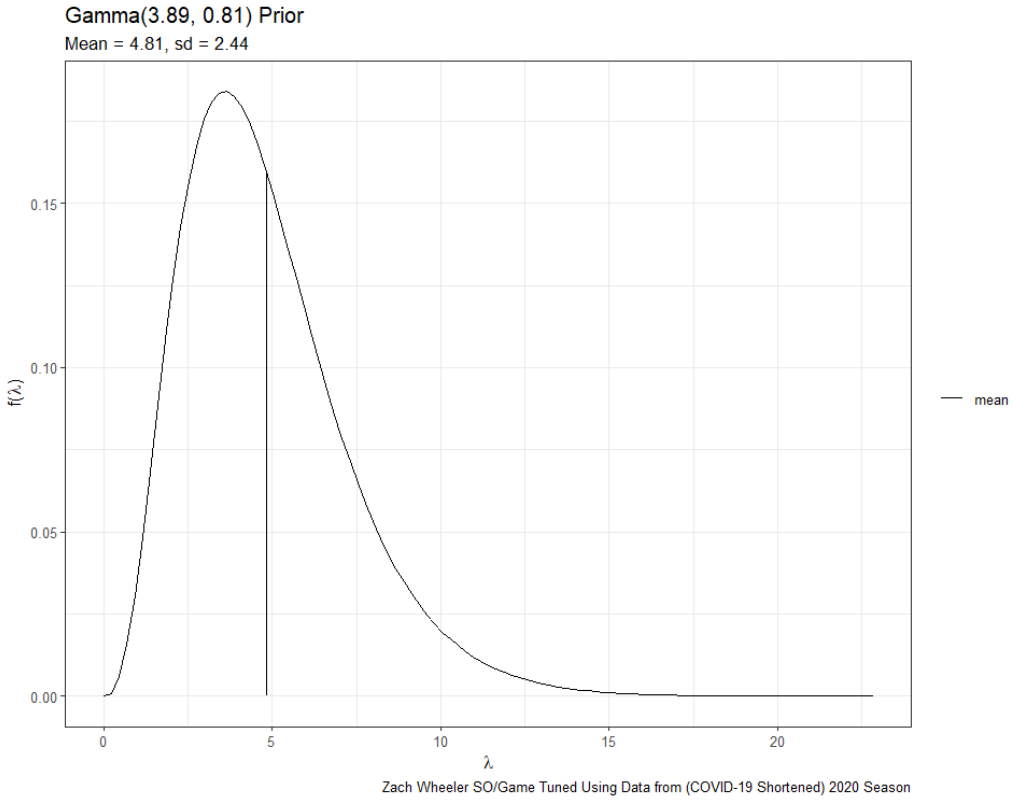

Figure 7: Tuned Prior Model

I then updated this prior using data from the 2021 season, in which Wheeler played 32 games and recorded 247 strikeouts. The mean strikeouts per game in this sample was 7.72, considerably higher than the mean of the prior. The resulting Gamma$(250.89, 32.81)$ posterior model had a mean of 7.65, a mode of 7.62, and a standard deviation of 0.48. Hence, as illustrated in Figure 8, the posterior is shifted to the right of the prior, closer to the likelihood, and there is much more confidence in the new estimates due to the increased sample size, as shown by the narrower spread.

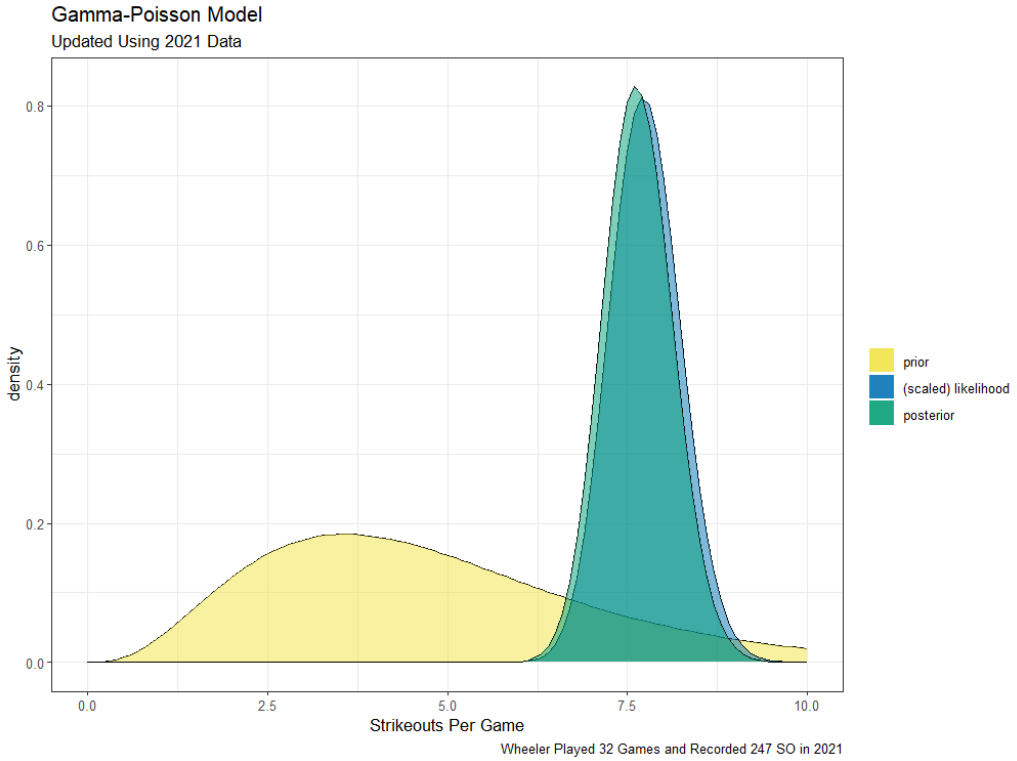

Figure 8: Posterior Model After 2021 Update

A similar process was performed with data from the 2022 season using the previous posterior as a prior. In the 2022 season, Zach Wheeler had a lower mean strikeouts per game of 6.63, recording 163 strikeouts in 26 games played. As a result, the Gamma$(413.89, 58.81)$ posterior model has a lower mean of 7.03, mode of 7.02, and standard deviation of 0.35. As shown in Figure 9, the mean of the posterior falls between the means of the prior and likelihood, and given the reduced variance, there is increased confidence that Wheeler's mean strikeout rate per game is approximately 7.

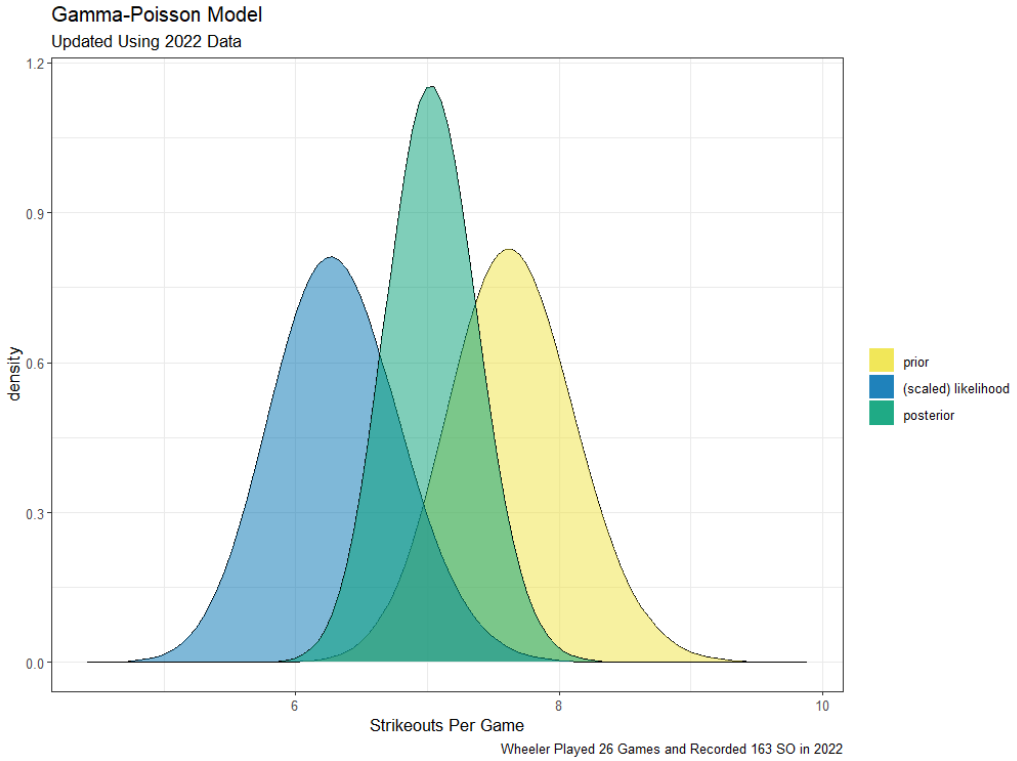

Figure 9: Posterior Model After 2022 Update

4.2 Bryce Harper Batting Average

Bryce Harper, similar to Zach Wheeler, is an all-star caliber player for the Phillies who signed a 13-year $330 million deal in 2019 (Freedman, 2023). Since the majority of Harper's value comes from his offensive production, I was interested in estimating his batting average, which can be captured through a Beta-Binomial model. Suppose there are Y hits Y in n at-bats. We can model batting average $\pi \sim \text{Beta}(\alpha, \beta)$, $Y|\pi \sim \text{Binomial}(n, \pi)$, and $\pi|(Y = y) \sim \text{Beta}(\alpha + y, \beta + n - y)$. I first tuned a $\text{Beta}(25.5, 69.5)$ prior model (see Figure 10) using hitting data from the 2020 season, in which Bryce Harper had 51 hits in 190 at-bats. I tuned α and β such that the expected value of the prior aligns with Harper's 2020 batting average (0.268) and the 95% confidence interval for his batting average is approximately 0.2 to 0.35, which is consistent with the historical range of batting averages among major league hitters (see Figure 11).

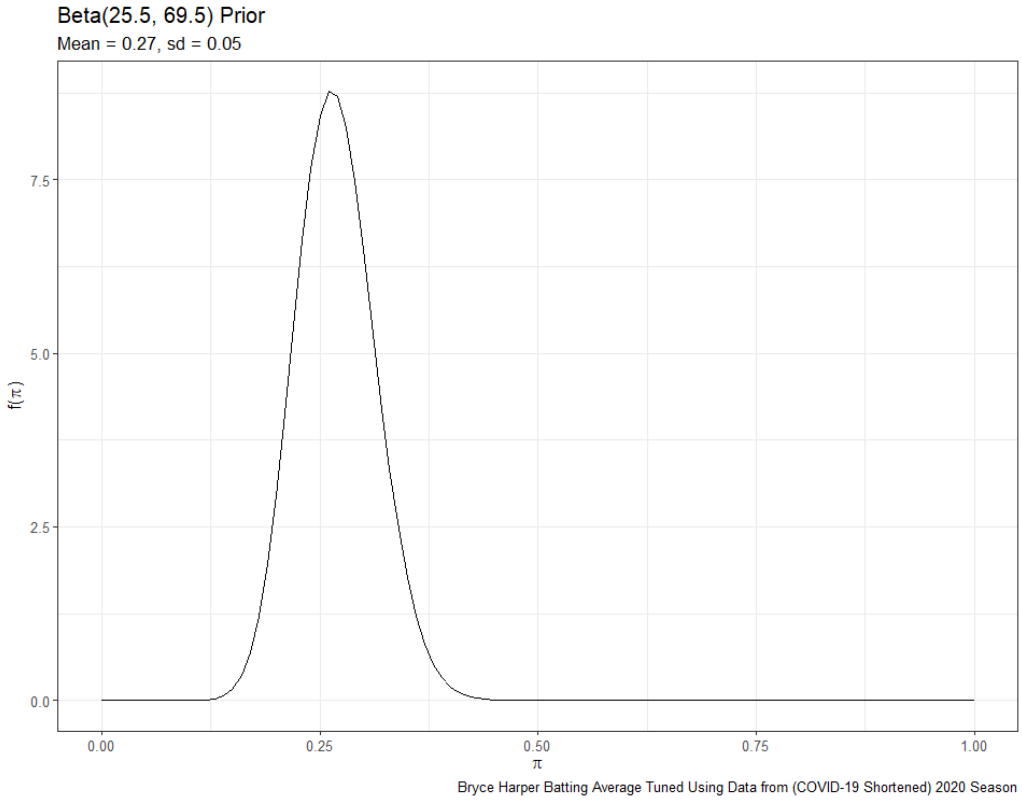

Figure 10: Tuned Prior Model

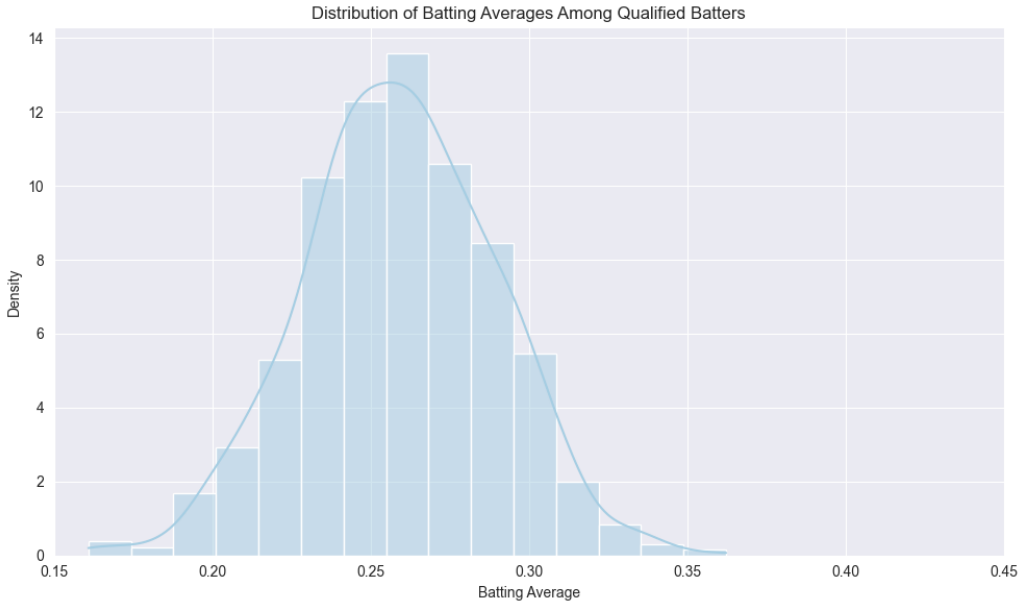

Figure 11: Batting Average Distribution

I updated this prior using data from the 2021 season, in which Harper had 151 hits in 488 at-bats (a 0.309 batting average). Hence, Harper had more hits in the 2021 season than expected from the prior, leading to a rightward shift in the batting average distribution as seen in Figure 12. The mean of the posterior distribution is 0.302, considerably higher than the prior, and there is a narrower spread and, thus, more confidence in the estimates stemming from the increased sample size.

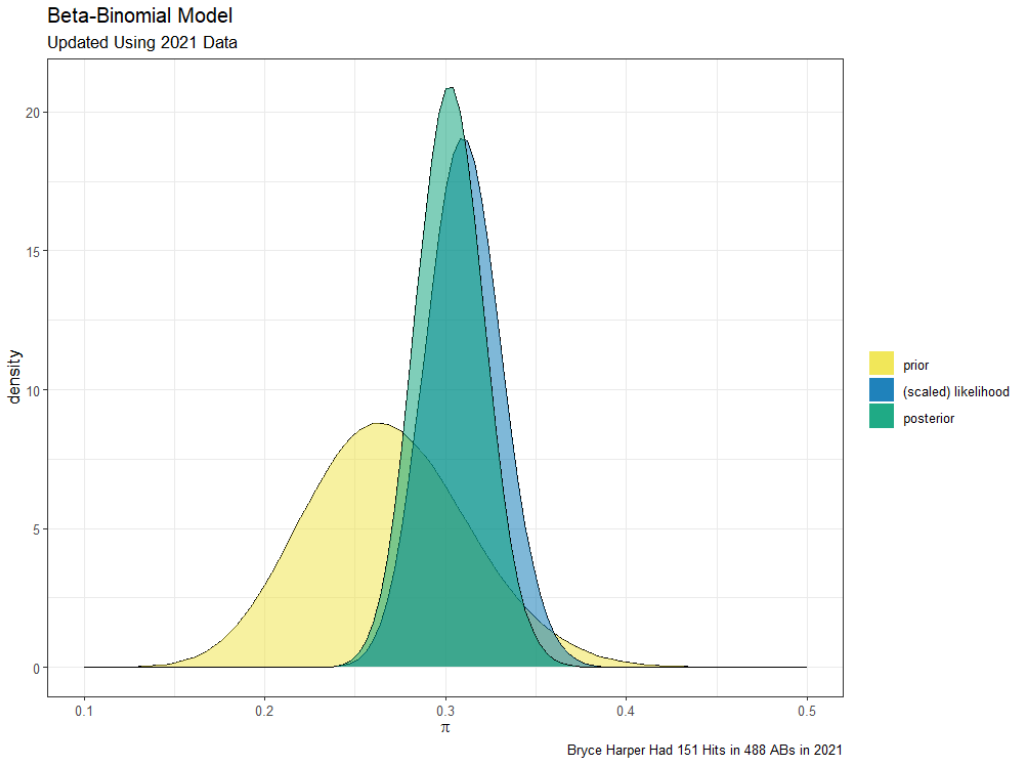

Figure 12: Posterior Model After 2021 Update

This process was repeated using the previous posterior as the prior for the 2022 season (see Figure 13), in which Harper had 106 hits in 370 at-bats (a 0.286 batting average). Since the mean of the likelihood is less than the prior, the distribution shifts slightly to the left with a mean of 0.296, and there is increased confidence that Bryce Harper's batting average is around 0.300.

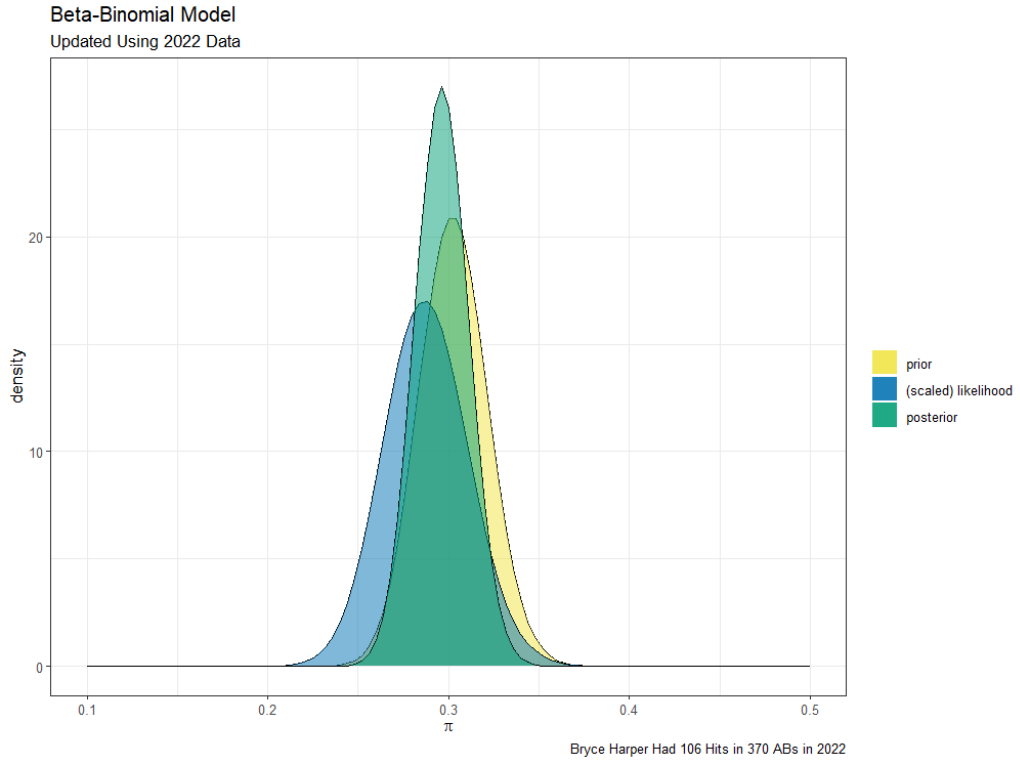

Figure 13: Posterior Model After 2022 Update

4.3 MLB Mean WAR: MCMC Simulation

Given the ultimate goal of this project is to project the aging curve for MLB players, I want to first estimate the player WAR from 2010-2019 for starting-caliber players (those who have played in more than half of team games) and test the hypothesis that the average WAR among starters is less than 2. I assume that a given position player's WAR in a particular season follows a Gaussian distribution with unknown mean μ and known σ, and thus $Y_i|\mu \overset{\text{ind}}{\sim} \text{Normal}(\mu, \sigma^2)$. I tune the prior such that $\mu \sim \text{Normal}(1.84, 2.01^2)$ based on data from the 2010-2014 seasons (see Figure 14) and assume a constant standard deviation σ from the prior. I then run an MCMC simulation using observed data from the 2015-2019 seasons to approximate the posterior distribution (see Figure 15). The posterior distribution has a mean of 1.80 and a standard deviation of 0.10 with a 95% from 1.6 to 2, illustrating that the observed data decreased μ and increased the confidence in the estimates. The resulting Bayes Factor is 34.37, and so, given the

observed data, there is some evidence that it is more plausible that the average WAR among position players is less than 2.

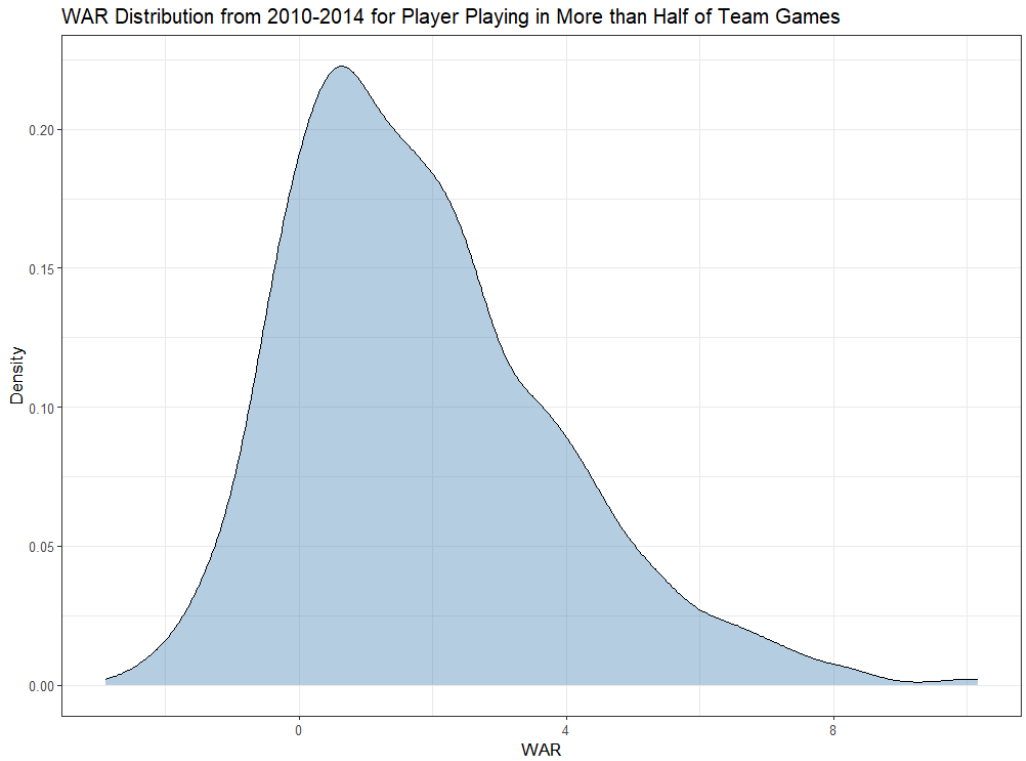

Figure 14: Density Plot of Position Player WAR

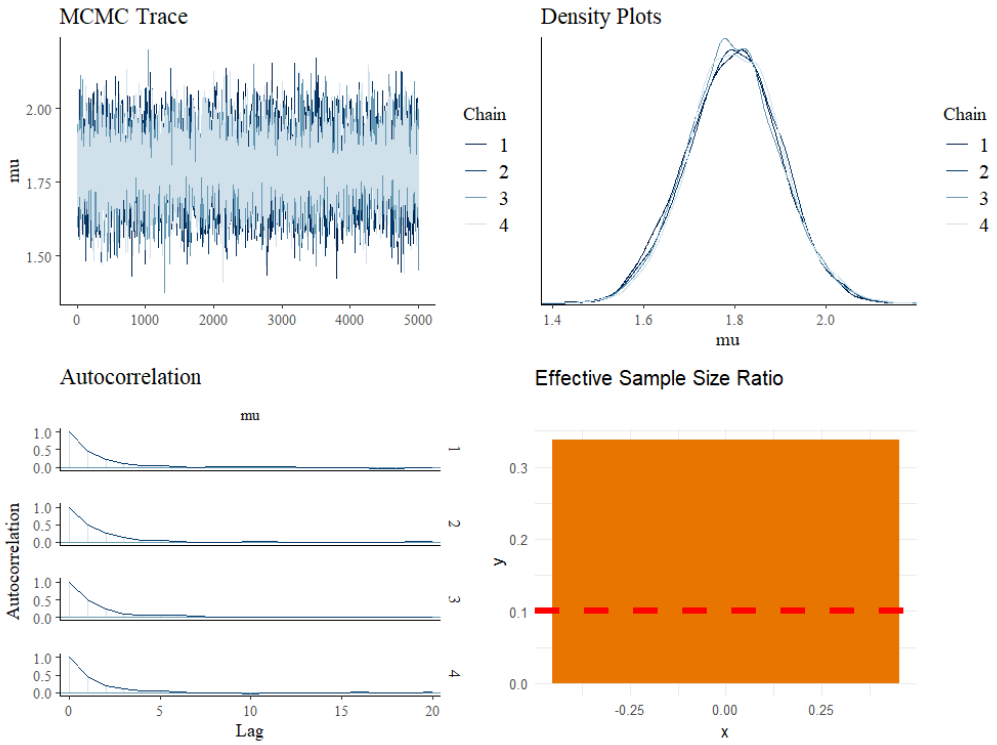

Figure 15: MCMC Metrics

5 A Simplistic Approach to Projecting WAR

Following Frontin (2019), I begin by assuming that the percent change in WAR for a given player is drawn from a Gaussian distribution, with mean and variance that are functions of a player's age and previous season's WAR. Thus, $\left(\frac{\text{WAR}_i - \text{WAR}_{i-1}}{\text{WAR}_{i-1}}\right)_i \sim \mathcal{N}(\mu, \sigma^2)$. To capture the idea that mean and variance are functions of age and previous WAR (see Figure 16), with variance decreasing as age and previous WAR increase (e.g., star players are less likely to have substantial swings in WAR, I adopt the following functional forms used by Frontin (2019):

$$\mu \approx \mu_{\text{model}} = \left(\alpha_0 + \exp(-\alpha_1 WAR_{i-1} - \alpha_2 WAR_{i-1}^2)\right) \times (\alpha_3 \text{age}^2 + \alpha_4 \text{age} + \alpha_5)$$

$$\sigma^2 \approx \sigma^2_{\text{model}} = \left(\beta_0 + \exp(-\beta_1 WAR_{i-1} - \beta_2 WAR_{i-1}^2)\right) \times (\beta_3 \text{age}^2 + \beta_4 \text{age} + \beta_5)$$

Model parameters were estimated by Frontin (2019) using MLE with data on qualified

hitters from 1955 to 2018. This framework allows for the estimation of future WAR given a draw from the approximate distribution and a player's previous season's WAR. In addition, it is possible to calculate the probability of a player achieving a specific target WAR in the following season by first calculating the required percentage change and evaluating the CDF for a Gaussian with μ_{model} and σ^2_{model}. This probabilistic approach is preferable to traditional point estimates as it incorporates the inherent variability and uncertainty in player performance from season to season and allows front offices to make more informed decisions.

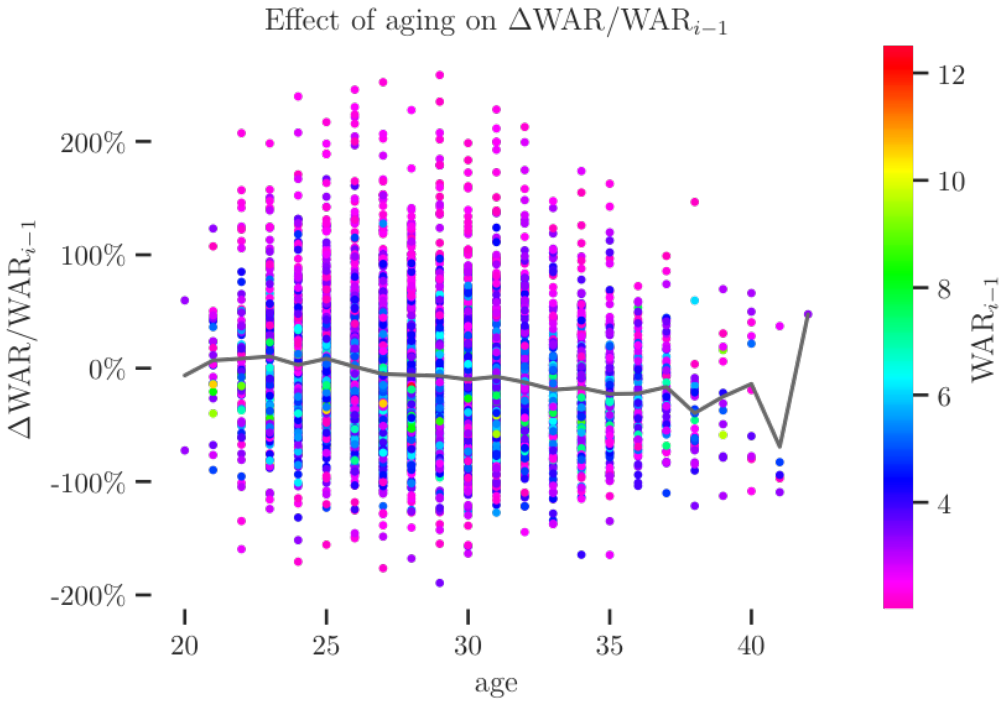

Figure 16: Effect of Aging on Percentage Change in WAR (Frontin, 2019)

In Figures 17 and 18, the points should be close to the 45-degree line if the model accurately predicts future WAR. The significant deviations for players with WAR above approximately four suggest that the model is not fully capturing the performance dynamics at the higher end of the WAR spectrum. While this modeling approach is far from perfect, I intend to use it as a starting point for my future research.

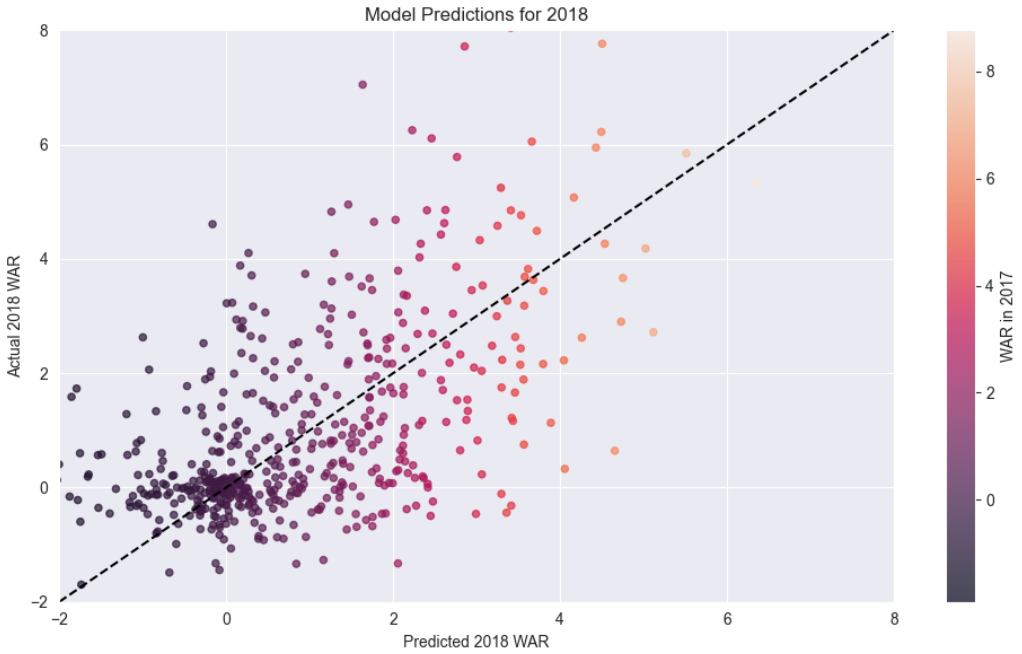

Figure 17: Model Predictions in 2018 using Data From Fangraphs

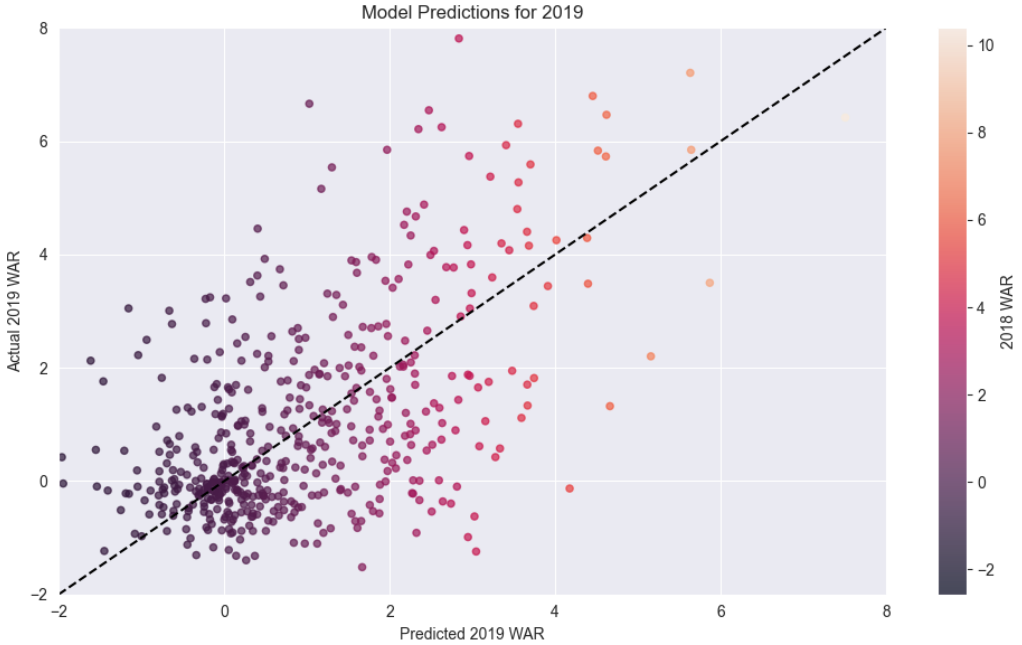

Figure 18: Model Predictions in 2019 using Data From Fangraphs

6 Future Directions

While the work of Frontin (2019) provides an invaluable starting point and could perhaps be used as a prior distribution in a Bayesian framework, I would like to further refine the WAR prediction model. I plan to start by first using Bayesian Neural Networks (BNNs) to project oWAR (offensive WAR) for position players over time, using historical player data (e.g., advanced hitting statistics and injury history) to derive probability distributions. I then hope to expand this model to project overall WAR, taking into account factors such as positional flexibility and defense. In order to determine whether there are certain characteristics that lend themselves to player longevity, I intend to use a hierarchical model or a Poisson regression to account for differences in injury rates due to factors such as player position, age, and playing style. Lastly, I would like to perform factor analysis to see if there are any latent factors driving performance.

References

Connon, S. (2023, October). MLB Announces Record Attendance Increase in 2023 as League Continues to Rebound. Retrieved March 5, 2024, from https://www.si.com/fannation/mlb/fastball/news/mlb-announces-record-attendance-increase-in-2023-as-league-continues-to-rebound

Drellich, E. (2023, December). How Shohei Ohtani's contract could save him $100 million in taxes — and start a Silicon Valley trend - The Athletic. Retrieved March 5, 2024, from https://theathletic.com/5153421/2023/12/21/shohei-ohtani-dodgers-contract-deferrals-silicon-valley/

Freedman, D. (2023, December). $330 Million Is Not Enough — Bryce Harper Wants A Contract Extension [Section: SportsMoney]. Retrieved March 7, 2024, from https://www.forbes.com/sites/danfreedman/2023/12/11/mlb-330m-is-not-enoughbryce-harper-wants-a-contract-extension/

Frontin, C. (2019, October). Baseball ProGUESTus: Casting Uncertainty on How Players Age (Part 2). Retrieved March 5, 2024, from https://www.baseballprospectus.com/news/article/54726/baseball-proguestus-casting-uncertainty-on-how-players-age-part-2/

Jamieson, J. P. (2010). The Home Field Advantage in Athletics: A Meta-Analysis. *Journal of Applied Social Psychology*, *40*(7), 1819–1848. https://doi.org/10.1111/j.1559-1816.2010.00641.x

Mohr, J. (2022). MLB Game Data. Retrieved March 6, 2024, from https://www.kaggle.com/datasets/josephvm/mlb-game-data

Petti, B., & Gilani, S. (2024). *Baseballr: Acquiring and Analyzing Baseball Data*. https://billpetti.github.io/baseballr/

Stark, J. (2024, March). Zack Wheeler's extension shows why his last deal is The Greatest Starting Pitcher Contract Ever. Retrieved March 7, 2024, from https://theathletic.com/5316386/2024/03/04/zack-wheeler-phillies-greatest-starting-pitcher-contracts/

Verducci, T. (2023, May). MLB Is Better and Fairer Because of League Crackdowns. Retrieved March 7, 2024, from https://www.si.com/mlb/2023/05/12/mlb-crackdown-on-pitchers-sticky-substances-making-the-game-fairer