Bayesian Analysis of Home-Run Hitting

Yingjie (Gary) Zhou

1 Introduction

Sluggers, baseball hitters known for their home run (HR) capabilities, are among the highest valued players in Major League Baseball (MLB). While much of baseball literature and research revolve around on-base-percentage, batting average, or other holistic measures of hitting prowess, I was interested in an analysis focused around home-run hitting. This report contains two parts: (1) a hierarchical Bayesian model with partial pooling to predict rest-of-season home run output and (2) an exploration of the distribution of home-run rate using a Gaussian mixture model. All code can be found through `https://github.com/yinggz/baseball_hr_bayes`.

2 Data

In the first model, we used 2022 batting data from Retrosplits, a Github repository for disaggregated MLB game log data. Since players miss time for injury and other reasons, their HR totals for the season may not reflect their true ability. Lead-off hitters typically get more opportunities to bat because of the structure of a 9-inning game, and hitters on high-scoring teams get more opportunities per game. As a result, their home run totals may be inflated. To account for these events, we created a new variable for home run as a proportion of plate appearance (PA). If players only stepped on the field for half of the season, but hit a high number of home runs relative to their batting opportunities, then they should still be considered a slugger. After transformation into proportions, this new variable, home-run rate, remained right skewed.

3 Bayesian Hierarchical Model with Partial Pooling

Given that we know the number of home runs a player has hit up to their 170th plate appearance, could we use a Bayesian Hierarchical Model to predict the rest of their season's home run performance? This question was inspired by a previous analysis on batting average [1]. Unlike, batting average, which stabilizes the quickest, home run rate takes more plate appearances to stabilize and comes with its own set of modeling challenges [2]. While home run rate is typically defined with at bats as the denominator, we wanted a stable cutoff point from which to divide our data into training and test set. Carleton defined stability as the number of PA needed for a given batting metric to reach the point where the correlation between that sample and another sample of the same size is 0.7 (i.e., R^2 of .49)[2]. For home runs, the number of PAs was approximately 170, so batters who did not reach that threshold were removed.

We treated plate appearances as repeated binary trials where each PA's chance of success (home run) was the same, satisfying the exchangeability requirement. There are Y_n home runs over K_n plate appearances for $n = 1, \dots, N$ players in a given season. A Bayesian hierarchical modeling approach was chosen to incorporate partial pooling, where each batter was assumed to have a different chance of hitting home runs, but the estimates were also informed by the overall league home-run rate.

We define the chance of hitting a home run in a given PA as θ, with a beta prior that matches its support of [0,1] and is conjugate to its binomial likelihood.

$$L(y|\theta,K) = \prod_{n=1}^{N} \text{Binomial}(y_n|\theta,K) \tag{1}$$

$$p(\theta|\alpha,\beta) = \prod_{n=1}^{N} \text{Beta}(\theta_n|\alpha,\beta) \tag{2}$$

$$\alpha,\beta > 0$$

Thus, the beta distribution's shape parameters, α and β, represent the prior number of successes + 1 and the prior number of failures + 1, respectively. They are reparametrized as $\alpha = \kappa\phi$ and $\beta = \kappa(1-\phi)$, and these new parameters are given hyperpriors.

$$p(\theta|\phi,\kappa) = \prod_{n=1}^{N} \text{Beta}(\theta_n|\kappa\phi,\kappa(1-\phi)) \tag{3}$$

$$p(\phi) = \text{Uniform}(\phi|0,1) \tag{4}$$

$$p(\kappa) = \text{Pareto}(\kappa|1,1.5) \propto \kappa^{-2.5} \tag{5}$$

$$\text{where } \phi \in [0,1] \text{ and } \kappa > 1$$

A flat prior is chosen for ϕ. In this context, ϕ becomes the population chance of success and is equal to the mean of a variable with a Beta(α,β) distribution: $\phi = \frac{\alpha}{\alpha+\beta}$. It serves as a direct representation of the average probability of home run success. Kappa (κ) becomes the concentration parameter (approximately inversely related to the variance) with a straightforward interpretation of $\kappa = \alpha + \beta$. In other words, κ is the prior number of successes and failures + 2. The heavy-tails characterized by a Pareto(1, 1.5) distribution make it a good choice of hyperprior because they can flexibly capture a wide range of prior beliefs about κ.

I specify half of our 5000 iterations for warm-up and the other half for sampling for each of the four chains. For each of our three parameters of interest, 10000 posterior samples are drawn from Stan's Hamiltonian Monte Carlo with No U-Turn Sampler (HMC-NUTS).

3.1 Convergence

Parameter	Mean	10%	90%	ESS	Rhat
Kappa	195.17	147.07	248.85	1333	1
Phi	0.03	0.03	0.03	6347	1

Table 1: Summary statistics from the posterior samples, with 80% credible intervals

Based on the effective sample size, Rhat value of 1, and the traceplots in Figure 1, κ appears to have converged, even if its chains didn't mix as well as those for the other parameters and there remains some autocorrelation, as shown in Figure 2. Still, there remains substantial variability for κ (close to the value of the prior count) as shown by the 80% credible interval of [147.07, 248.85]. All other parameters converged and showed no signs of autocorrelation. The mean value of ϕ, interpreted as the population home-run rate, is 0.03.

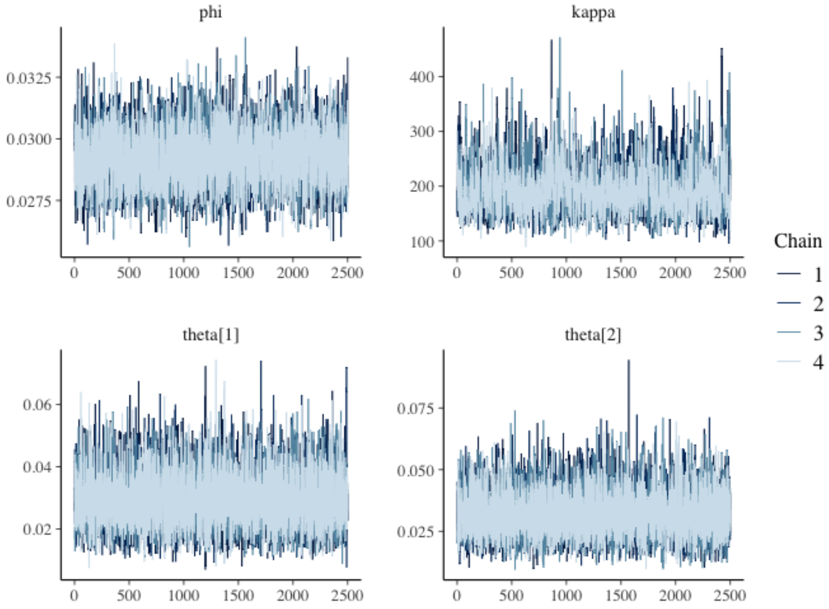

Figure 1: Traceplots from the Bayesian hierarchical model for phi, kappa, and the first two values of theta (out of 252)

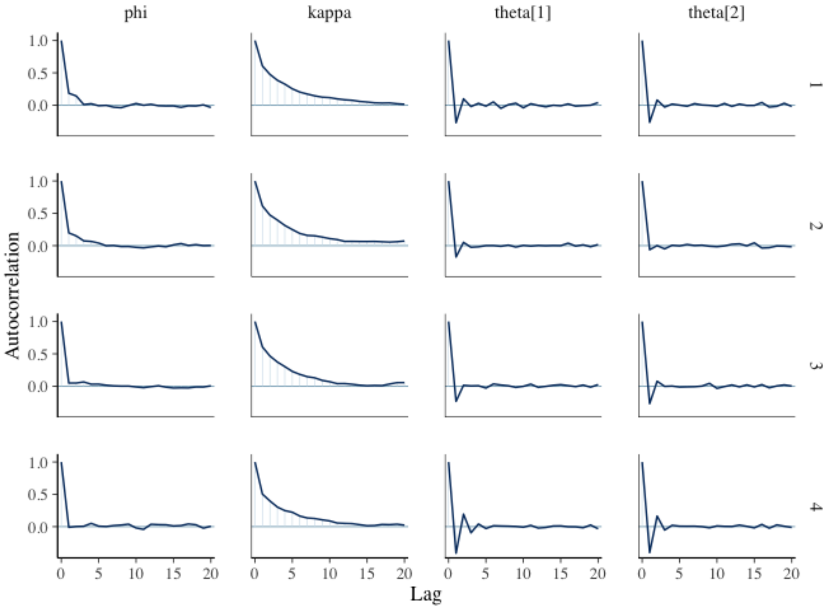

Figure 2: ACF plots from the Bayesian hierarchical model for phi, kappa, and the first two values of theta

3.2 Posterior Predictive Checks

Next, we draw from our posterior predictive distribution and conduct posterior predictive checks using the minimium, sample mean, sample standard deviation, and maximum test statistics.

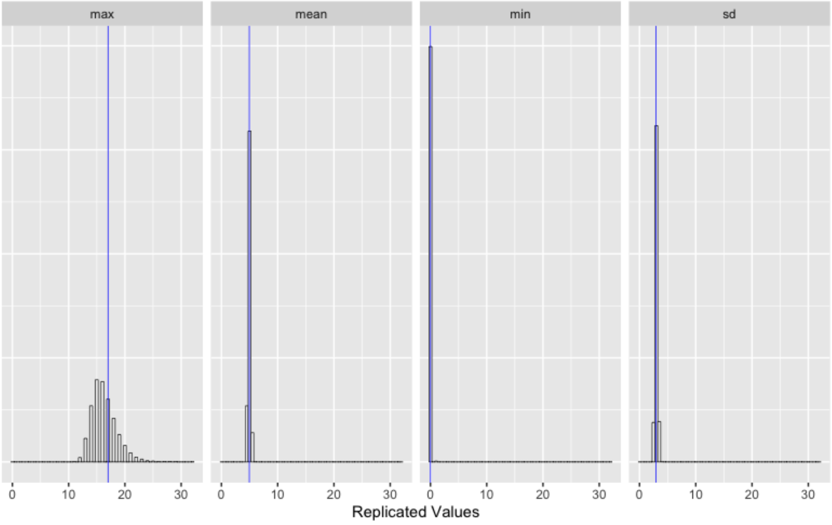

Figure 3: Posterior predictive checks for max, mean, min, sd are shown. The blue vertical lines are the observed values of the test statistics.

Figure 3 shows that for each of the test statistics, their observed values are contained within the posterior predictive distributions of the test statistics. Hence, the model has done a good job of capturing the trends of the data.

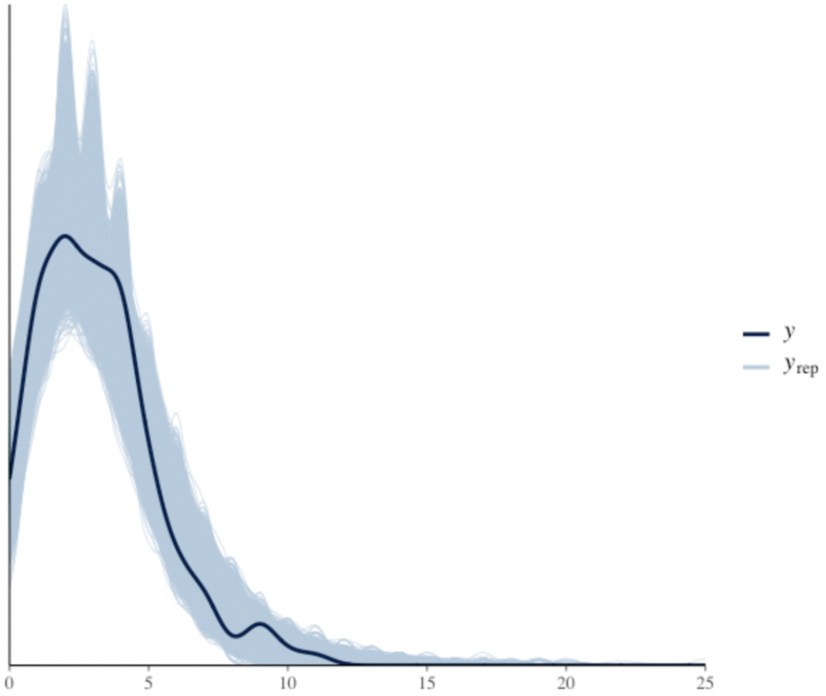

Figure 4: Plot of replicated home-run rates compared to the density curve of the observed home-run rates.

We compared the distribution of y to distributions of multiple simulated datasets from the posterior predictive distribution. Since the observed data line is within the range of the simulated data lines, the model appears to be capturing the data distribution well.

Using our posterior predictive distribution for new data y, we predict how many home runs our players will hit for the rest of the season. Then, we evaluate our predictive performance with the remaining home run and at bat data.

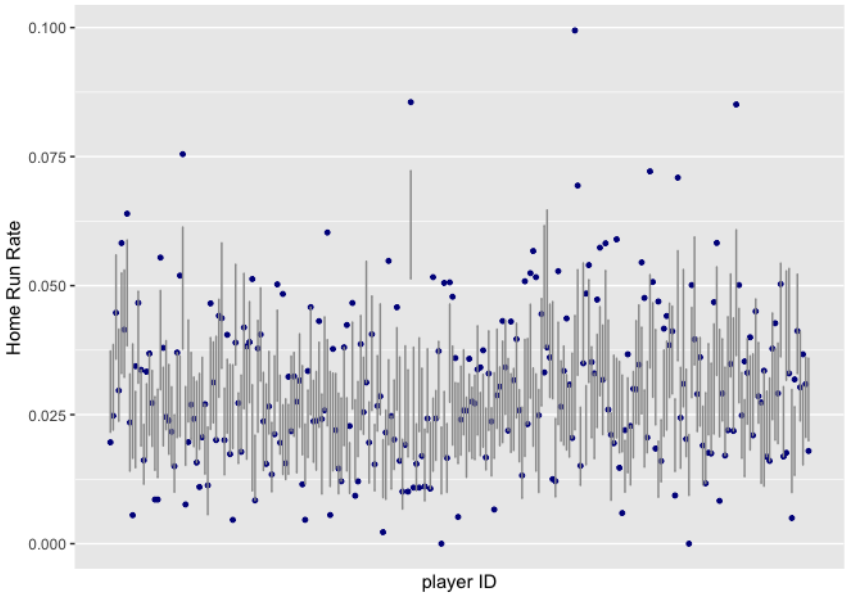

Figure 5: Posterior Predictions for Home Run Rate (Rest of Season)

Figure 5 shows the plot of each player's observed rest-of-season home run rate (in blue) and the corresponding 50% posterior predictive interval for their predicted rest-of-season home run rate. We see that while a large portion of the observed values are contained within the 50% interval, it does not contain many of the high home run rates, due to its pull towards the prior. In Table 2, the top 10 rest-of-season home run hitters, ranked by their posterior predictions, are shown.

Batter	HR at 170PA	Remaining PA	Predicted HR	Remaining HR
Aaron Judge	17	526	32.84	45
Jose Ramirez	11	515	23.73	18
Christian Walker	11	497	22.80	25
Pete Alonso	10	515	22.14	30
Kyle Schwarber	10	499	21.52	36
C.J. Cron	11	462	21.20	18
Willy Adames	11	447	20.49	20
Austin Riley	8	523	19.72	30
Yordan Alvarez	12	391	19.07	25
Mookie Betts	9	469	18.92	26

Table 2: Posterior predictions of remaining trials

Out of the top 8 hitters with the most home runs remaining, the model predicted 5 of them. The others in our projected top 10 finished 12th, 16th, 28th, 43rd, and 48th. The RMSE of the posterior predictions was 4.55, which is relatively good considering the scale of remaining home run totals ranges from [0, 45] and we are only using 170 PAs.

By specifying hyperpriors for the individual-level parameters and the group-level parameters to account for both levels of variation in the chance of hitting a home run at a PA, and allowing for pooling, the estimated parameters at each level borrowed strength across the different levels. Hence, we get a model that provides good predictive capability, especially for batters without outlier hitting ability.

4 Two-Component (Slugger vs. Non-Slugger) Gaussian Mixture Model of Player Seasons

In the previous section, the distribution of home-run rates appeared to be right skewed with a truncated left tail at 0. Based on subject matter knowledge, we hypothesize that there are different types of hitters: those who excel at hitting home runs (sluggers) and the rest. The sluggers may come from a different data generating process that is causing the long right tails. I proposed using a two-component Gaussian mixture model as a non-arbitrary method for categorizing sluggers, since it could incorporate these potentially different data generating processes within the same model. To test our assumption of a normal-normal mixture distribution, we relied on posterior predictive checks.

Using Lahman's data, which contains season-long batting statistics for each player from 1975 onwards, we fit this model to the distribution of home run proportion across player seasons with the goal of finding out the likelihood of each player season (and the proportion of player seasons) belonging to the sluggers group. I chose to use at bats (AB) instead of plate appearances (PA) as the denominator for home run rate for this problem, because AB represents the results when balls are in play. In the case of PA, if one takes a walk, it should have no bearing on their abilities as a home run hitter. We include only those hitters with at least 100 ABs.

The proportion of player seasons in the slugger group is given by α, while the proportion of non-slugger player seasons is its complement: $1 - \alpha$. This mixing proportion is modeled as a flat prior, using a symmetric beta distribution with shape parameters of 1 to assign equal probability to all values between 0 and 1. According to the model, the average home-run rate (μ_2) for player seasons in the sluggers group should be higher than the average home-run rate (μ_1) for those in the non-sluggers group. To avoid issues with non-convergence stemming from the fact that the mixture components are underlying exchangeable, we specified ordered priors ($\mu_2\, \mu_1$) [3]. These priors are normal distributions centered at 0.06 and 0.01, respectively, with a standard deviation of 1, to reflect the scale of home-run proportions. I set the standard deviations of the two components to both be inverse gamma distributions with shape parameter 2 and scale parameter 0.1 since I expected the variability in the data to be relatively small.

The full model architecture is provided here:

$$y_i \sim \begin{cases} \text{Normal}(\mu_2, \delta_2^2), & \text{with probability } \alpha \\ \text{Normal}(\mu_1, \delta_1^2), & \text{with probability } 1-\alpha \end{cases} \quad \text{for } i = 1, \ldots, \text{Number of Player Seasons}$$

Thus, the likelihood is represented by:

$$p(Y|\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2) = \prod_{i=1}^{n} [\alpha \cdot \mathcal{N}(Y_i|\mu_2, \sigma_2) + (1-\alpha) \cdot \mathcal{N}(Y_i|\mu_1, \sigma_1)]$$

The joint posterior distribution from which Stan derived samples using the Hamiltonian Monte Carlo (HMC) algorithm, is given by:

$$p(\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2|Y) \propto p(Y|\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2) \times p(\alpha) \times p(\mu_1) \times p(\mu_2) \times p(\sigma_1) \times p(\sigma_2)$$

$$\alpha \sim \text{Beta}(1,1) \tag{6}$$

$$\mu_1 \sim \text{Normal}(0.01, 1) \tag{7}$$

$$\mu_2 \sim \text{Normal}(0.06, 1) \tag{8}$$

$$\sigma_1 \sim \text{InverseGamma}(2, 0.1) \tag{9}$$

$$\sigma_2 \sim \text{InverseGamma}(2, 0.1) \tag{10}$$

In the Stan implementation, I transformed the likelihood to be defined in log-space. The burn-in (warmup) period consisted of the first 1000 (out of 2000) iterations for each of the 4 chains, leaving us with 4000 samples for each parameter.

Convergence

For each of the five parameters of interest, the trace plots in Figure 6 show good mixing, indicating convergence. Our Rhat values were all 1 or below, and the number of effective samples was sufficient. Figure 7 shows that there is minimal autocorrelation for each of the parameters.

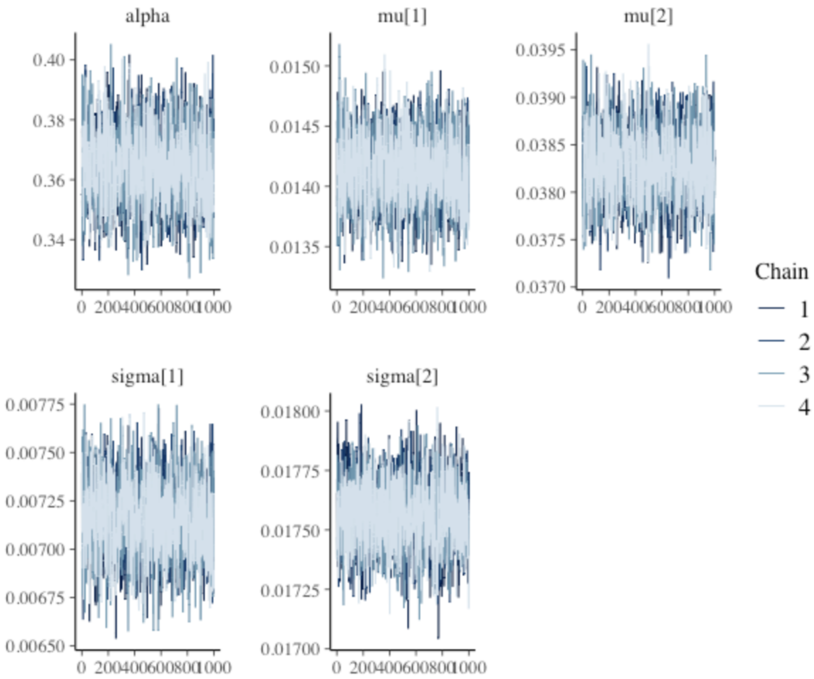

Figure 6: Traceplots from the two-component mixture model.

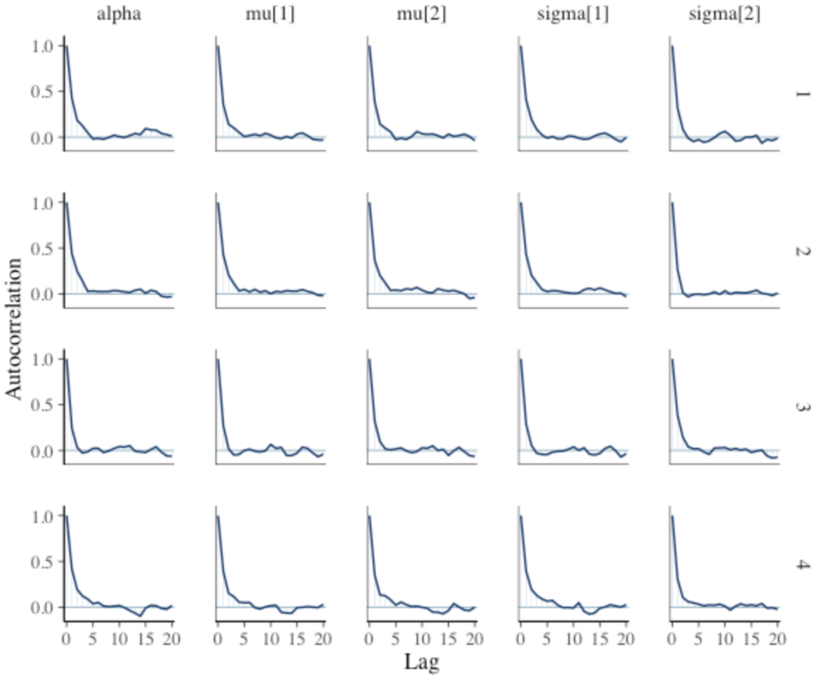

Figure 7: ACF plots from the two-component mixture model.

Posterior Predictive Checks

We generate a sample of data from the posterior predictive distribution and compare its test statistics to those from the observed data.

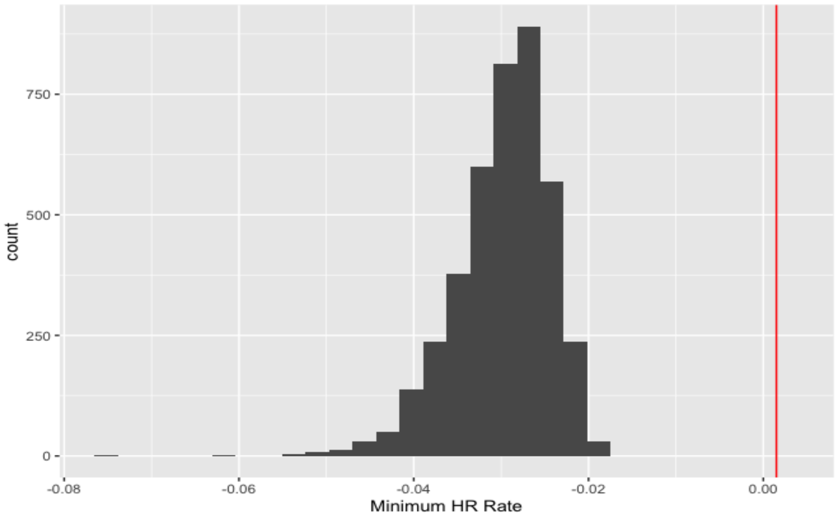

Figure 8: Posterior Predictive Check of Min

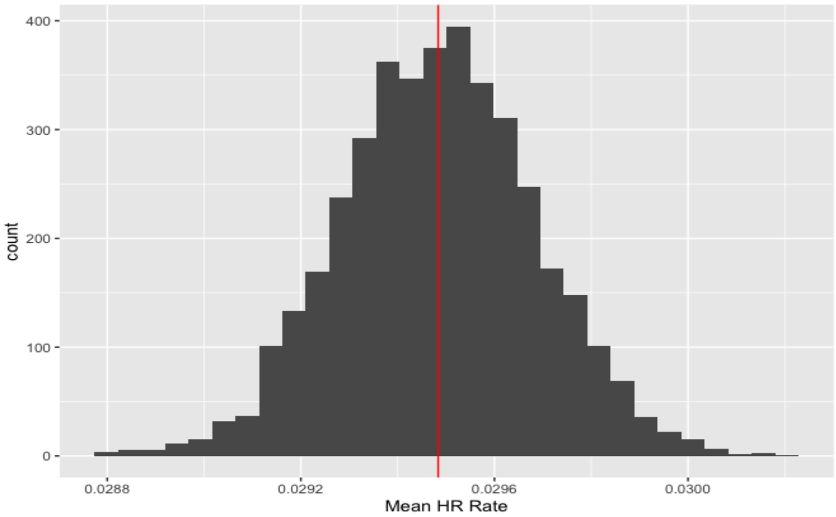

Figure 9: Posterior Predictive Check of Mean

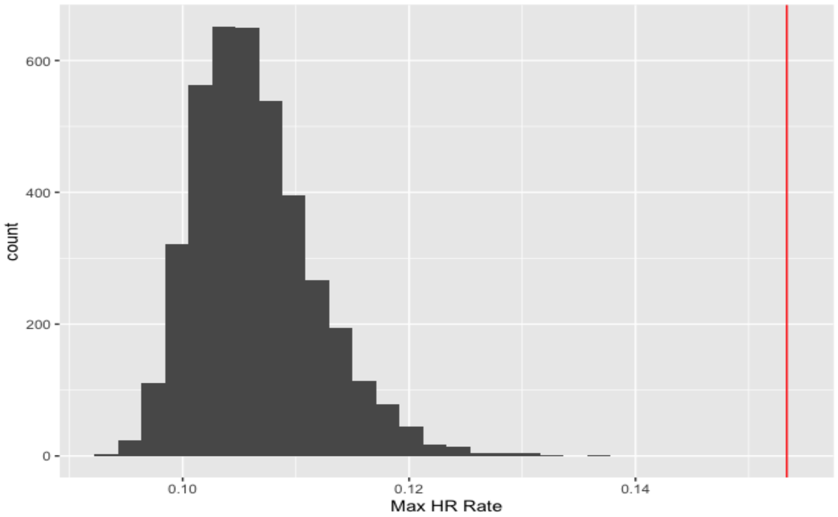

Figure 10: Posterior Predictive Check of Max

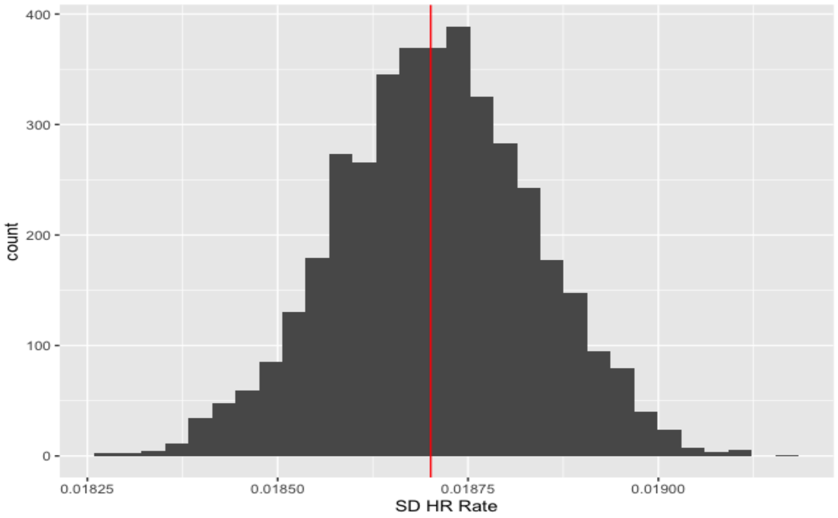

Figure 11: Posterior Predictive Check of SD

Figure 9 and Figure 11 show that the mean and standard deviation of the posterior predictive samples match those from the original dataset. However, the observed maximum and minimum, shown in Figure 10 and Figure 8, are not within the distribution of the posterior predictive maximum and minimum, meaning that we did not fully capture the patterns at the extremes. While the model's predictions might be off, they are still relatively close in value for all the test statistics, capturing the general trend of the observed data reasonably well.

Results

The HMC algorithm provided us with posterior samples for each of our five parameters. Figure 12 contains curves for the fitted densities of the hypothesized elite and non-elite groups, as well as the density of the mixture distribution. The model appears to capture the shape of the observed data relatively well, with two distinct normal distributions.

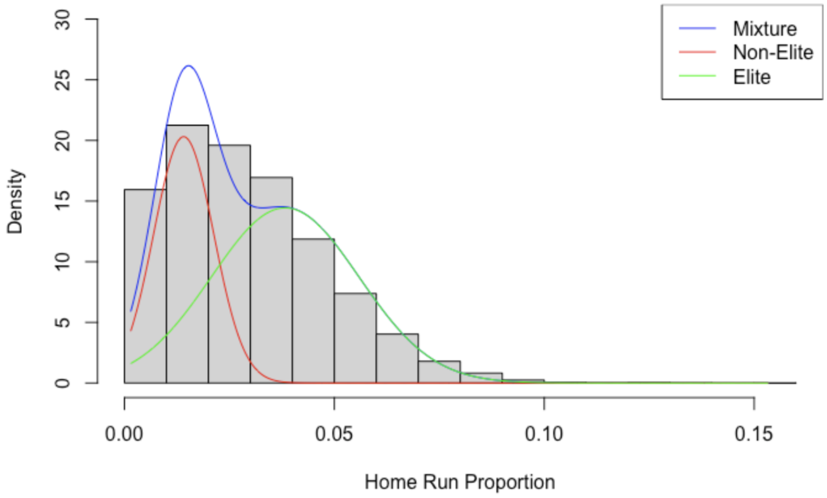

Figure 12: Histogram of observed home run proportions overlaid with the densities of the mixture distribution and the respective normal distributions of each component.

	Mean	10%	90%	ESS	Rhat
α	0.6357	0.6203	0.6509	1582	1
μ_1	0.0141	0.0138	0.0145	1729	1
μ_2	0.0383	0.0379	0.0387	1565	1
σ_1	0.0072	0.0069	0.0074	1700	1
σ_2	0.0176	0.0174	0.0177	2153	1

Table 3: Summary statistics for parameters of a two-component mixture model

Counter to our expectations, we see in Table 3 that the proportion of those who are in the elite hitting category (sluggers) is 64%, as opposed to 36% for the non-elite group (non-sluggers). This tells us that, rather than separate distributions for elite home-run hitters and the rest, we actually have below-average home-run hitters and the rest. The average home run proportion for those in the below-average group, which we coin as the "weak hitters" group, is 0.014, compared to 0.038 for those in the "non-weak hitters" group. Furthermore, the standard deviations for the weak hitters distribution is less than half that of the non-weak hitters group.

In addition to providing uncertainty measures for home run proportions, the posterior samples can help derive the probability of belonging to the weak hitters group. By using the posterior means of the parameter estimates, we calculate the expected value of belonging to the the weak hitters group as follows:

$$\text{P(Weak)} = (1-\alpha)\cdot\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(y_i-\mu_1)^2}{2\sigma_1^2}\right)$$

$$\text{P(Non-Weak)} = \alpha\cdot\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{(y_i-\mu_2)^2}{2\sigma_2^2}\right)$$

$$E_i = \frac{\text{P(Weak)}}{\text{P(Non-Weak) + P(Weak)}}$$

Using this expectation formula, we can classify those hitter seasons on the fringes (e.g, those with a greater than 50% chance of being in the weak group are considered weak hitters). We see that there were 7461 weak-hitting seasons from 1975-2021. We get a non-arbitrary cutoff threshold for home run proportion; hitters with a home run proportion of 0.022 or less were considered weak-hitting. This may help inform decisions makers on each team during contract season.

5 Limitations

The dataset contains only those observations with at least 100 at-bats or 1 HR (to eliminate pitchers who hit in the NL), meaning that a small portion of bench players or those without any home run hitting ability were excluded. Including them may see a slight change in the distribution of HR totals, possibly requiring us to adjust for zero-inflation. Attempts to refit the data with truncated normal and poisson distributions failed to converge.

References

[1] Bob Carpenter. Hierarchical partial pooling for repeated binary trials, 2016. [Online; accessed 15-May-2023].

[2] Piper Slowinski. Sample size, 2010. [Online; accessed 15-May-2023].

[3] Bruno Nicenboim, Daniel Schad, and Shravan Vasishth. An introduction to bayesian data analysis for cognitive science, 2023.