Bayesian Logistic Regression via Gibbs Sampler & Applications to Analysis of Baseball Performance

Ryan Anderson
raanderson@g.ucla.edu, UID: 306076860

June 17, 2023

Abstract

We present a description of Bayesian logistic regression and the application of MCMC techniques to obtain estimates for parameters with intractable conditional distributions. We also describe a model for evaluating hitting performance in baseball structured as a Bayesian logistic regression and provide an update on its predictions for the 2019 season.

1 Introduction

In-game outcomes in Major League Baseball, the highest level of professional baseball played in the US and Canada, seem *prima facie* to be simple things – a batter may succeed or fail to get on base in any at-bat, and runners may score at home plate. In fact, these outcomes are complexly determined, none less so than the performance of the pitcher and the impact he may have on in-game events.

In this paper, extending Jensen, McShane, and Wyner (2009), we seek to update their estimates of hitter performance in a full season by predicting the number of home runs made. We are also interested in estimating the transition probabilities of a hidden Markov model categorizing hitters as either elite or non-elite.

2 Literature Review

The application of techniques for statistical estimation to baseball data is one of the oldest stories in our field. When Joe DiMaggio was having his breakout seasons with the Yankees in the mid-1930s, the legend goes that his father, a poor illiterate fisherman who cast off from San Francisco's Fisherman's Wharf, taught himself division to follow his son's batting average.

Classically, batters were evaluated based on a few simple metrics, among which chiefly their batting average. This figure, the number of hits divided by the number of at-bats, varies across players in any given season between 0.150 and 0.400. Over an entire career, achieving a 0.300 batting average is a remarkable feat.

In the last quarter-century or so, commentators have invested energy into compiling further metrics derived from the simple metrics. Bill James, the sportswriter and statistician who spearheaded much of this push, was essentially a nobody writing into the wind all the way back in the 1970s. Then Aaron Sorkin wrote and Brad Pitt starred in a movie about the early-aughts Oakland Athletics' successful use of his philosophy, which privileges metrics like on-base percentage and batting average on balls in play over hits, runs and runs batted in.

Since then, the so-called sabermetrics community has come to dominate baseball discourse, seeking ever more accurate predictors of future success. The writer Dan Szymborski, who works at FanGraphs, is known for a model called ZiPS developed to assess the future impact of players and thereby the teams they play on, out to time horizons extending a decade or more.

Contemporaneously with Mr. James' efforts, Efron and Morris (1975) used baseball data to demonstrate the efficacy of shrinkage properties in the James-Stein estimator for estimating a vector of MLEs with $n > 3$. Even more recently, Bayesian techniques have come into vogue: Brown (2008) compared empirical Bayes methods including the James-Stein estimator but also methods which did not require equal variance assumptions among the players to estimate batter performance in the second half of a season given first half data.

In Brown (2008) it is mentioned that private correspondence with S. Jensen suggests that estimators could improve via adding another parameter for player position. Anecdotally, many baseball fans are aware of the fact that defensive contributions from an excellent shortstop or catcher may outweigh mediocre hitting, while first basemen (and especially designated hitters) who are expected to do very little fielding must be excellent hitters to earn a spot on the roster. Jensen et al. (2009) performs exactly the extension noted in the private correspondence between Jensen and Brown, estimating age, position, and ballpark parameters to inform the posterior distribution of hitting performance. Since this technique begins with a Bayesian logistic regression model, we walk through the regression model and the use of MCMC methods to simulate the posterior below.

3 Bayesian Logistic Regression via Gibbs Sampler

We follow Dogucu, Johnson, and Ott (2021) in the description of Bayesian logistic regression, and also build up the model for pitching performance by following the derivation.

Let Y_{ij} be the season total home runs by hitter j in year i. Home runs are modeled as binomial successes with the number of trials N_{ij} being given by the number of at-bats faced in a season, which we treat as exogenous, following Jensen et al. (2009). Then we have

$$Y_{ij} \sim Bin(n_{ij}, \theta_{ij})$$

with $E[Y_[ij]|\theta_{ij}] = N_{ij}\theta_{ij}$. Say we have player and year-level covariates $X_{ij}^{(k)}$. We use a logit link function to represent the relationship between θ_{ij} and covariates $X_{ij}^{(k)}$, and seek to estimate coefficients via regression in the following equation:

$$\log(\frac{\theta_{ij}}{1-\theta_{ij}}) = \beta_0 + \sum_k \beta_k X_{ij}^{(k)}$$

In Jensen et al. (2009), the exact covariates specified are position α_k, home ballpark β_k, and age, whereby they fit a smooth trajectory for each position using splines.

Now we know that in order to draw samples from the full posterior distribution, we would need priors on β_i, but these are intractable – they are just regression coefficients. Therefore, we can use Metropolis-Hastings to generate proposals for β_i and target the full posterior which we know is binomial.

We follow the notes of James S Clark (2021) in describing the implementation of the Gibbs sampler and M-H algorithm. We begin by setting non-informative priors for the β_i, and given a current $\beta_i^{(m)}$ propose a new $\beta_i^{(m+1)}$ via sampling from a multivariate normal distribution with mean equal to $\beta_i^{(m)}$ and covariance matrix given by a small fraction of the sample covariance matrix:

$$\hat{\Sigma}^{(m+1)} = \gamma * \frac{1}{n}(X^t X)^{-1}$$

We accept the proposal according to ratio $\frac{[\beta^*]}{[\beta^{(m)}]}$, where $[\beta]$ are our posterior distributions, given by

$$[\beta^*] = \Pi_{i=1} Ber(y_i|\theta_i) N(\beta^*|\beta_i)$$

and

$$[\beta^*] = \Pi_{i=1} Ber(y_i|\theta_i) N(\beta^{(m)}|\beta_i)$$

We therefore can run M-H via the following algorithm

```
# a proposal
b_m+1 <- matrix( .rMVN(1, mu = b_m, sigma = gamma*(X^tX)^-1), ncol=1)
```

```
# link to theta
theta_m+1 <- invlogit(x%*%b_m+1)
theta_m <- invlogit(x%*%b_m)

# log likelihood + log prior
pnew <- dbinom(y, n, theta_m+1, log = T) +
    .dMVN(b_m+1, priorB, priorVB, log=T)
pnow <- dbinom(y, n, theta_m, log = T) +
    .dMVN(b_m, priorB, priorVB, log=T)

# accept proposal with probability min(a, 1)
a <- exp( sum(pnew - pnow) )
if( a > runif(1,0,1)) b_m <- b_m+1
```

4 Applications to Analysis of Pitching Performance

Below we resummarize the model in Jensen et al. (2009) and update data through 2019 to track its accuracy. We also rewrite the model slightly to predict pitching performance instead of batting performance. We thank the authors for providing their code base for our uses.

In Jensen et al., the goal is to perform Bayesian logistic regression according to the following equation

$$\log(\frac{\theta_{ij}}{1-\theta_{ij}} = \alpha_k + \beta_b + f_k(A_{ij})$$

with the following parameters: α are position-specific intercepts, $f_k(A_{ij})$ is a spline-smoothed function of player age which differs by position, and β are home ballpark effects.

We also include a hidden Markov model to account for differences in player ability at the position level. That is, we define a latent variable E_{ij} in each player-year and split α_k according to the value of E_{ij}. This setup allows for players to move in and out of elite status over the course of their career.

Our parameters have the following prior distributions: $\beta_b \sim N(0, \tau^2), \gamma_{kl} \sim N(0, \tau^2), \alpha_k = (\alpha_{k0}, \alpha_{k1}) \sim MVN(0, \tau^2 I_2)\mathbf{1}(\alpha_{k0} < \alpha_{k1})$. We also need to provide priors on the transition probabilities for E_{ij}, $\nu_{00k}, \nu_{01k}, \nu_{10k}, \nu_{11k} \sim Dirichlet(1, 1)$.

Then our posterior is given by

$$p(\alpha, \beta, \gamma, \nu, E|X) \propto \Pi_{i,j} p(Y_{ij}|N_{ij}\theta_{ij}) * p(\theta_{ij}|R_{ij}, B_{ij}, A_{ij}, E_{ij}, \alpha, \beta, \gamma) * p(E_{ij}|E_{i,j-1}, \nu) * p(\alpha, \beta, \gamma, \nu),$$

where we recall that R_{ij} is player-year position, B_{ij} is player-year home ballpark, A_{ij} is player-year age, and E_{ij} is player-year elite status.

Finally, we update these parameters via a Gibbs sampler:

$$\begin{aligned}
p(\alpha|\beta, \gamma, \nu, E, X) &= p(\alpha|\beta, \gamma, E, X) \\
p(\beta|\alpha, \gamma, \nu, E, X) &= p(\beta|\alpha, \gamma, E, X) \\
p(\gamma|\alpha, \beta, \nu, E, X) &= p(\gamma|\alpha, \beta, E, X) \\
p(\nu|\alpha, \gamma, \beta, E, X) &= p(\nu|E) \\
p(E|\alpha, \beta, \gamma, \nu, E, X)
\end{aligned}$$

Note that there is nothing in the above setup that constrains this model to predicting only batting performance, save for the fact that there will be less variance coming through the position parameters due to there being fewer pitcher positions.

We show below results from updating Jensen et al.'s model for the period 2010-2018 in order to predict 2018. Breaking down the predicted elite transition probabilities by position reveals in 1 that first basemen are most likely to become elite when non-elite.

Visualizations of draws from the parameter posteriors are instructive: 2 demonstrates clearly that elite status is important for predicting home run rate.

Finally, we can visualize the output – draws from the posterior of home run totals for selected players. 3 shows that Jorge Alfaro, at that time a catcher for the Miami Marlins, was predicted not to exceed 25 homers, while Jose Abreu, then a first baseman for the Chicago White Sox, had a chance at hitting 50.

We assess the efficacy of this model through computing the RMSE of predicted home run totals vs actual home run totals. For 2019 data predicted by training the model on 2010-2018 data, we get an RMSE of 10.49. State-of-the-art models cited in Jensen et al. (2009) were able to achieve RMSEs closer to 7 with a bevy of additional specifications, so this relatively parsimonious model performs not too shabbily.

4 displays prediction accuracy relative to the 45-degree line. Commentary in Jensen et al. (2009) notes that all hitting models perform worse on predictions of young players than old, and so here we denote by color an age split. Indeed, the model does seem to

underestimate young hitters' abilities.

5 Discussion

In this paper, we have reviewed preliminaries for using Bayesian logistic regression to estimate a model requiring a link function between observations and parameters. We moreover described the simulation of results from the posterior of both parameter and outcome distributions. We moreover updated Jensen et al. (2009), describing the mechanics of their model, and some limitations of latent-variable approaches.

In further work, we would like to see extensions to pitching performance – how do elite/non-elite transition probabilities look for pitching? How does home ballpark affect ability to get strikeouts?

Lastly, it would be useful to see if changes in rules in the intervening period since Jensen et al. wrote their paper influence the construction of the model. In particular, both the American and National Leagues use a designated hitter. Rules changes before the 2023 season were designed to reduce the ability of outfielders to play wherever in the field they desired, potentially increasing the number of home runs via fewer outfield saves. And anecdotal evidence suggests home ballpark effects may be changing – in the 2019 World Series, neither the eventual champion, the Washington Nationals, nor their opponent, the Houston Astros, won any of the games played at their home ballparks!

References

Brown, L. D. (2008, March). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *The Annals of Applied Statistics*, *2*(1). Retrieved 2023-06-17, from https://projecteuclid.org/journals/annals-of-applied-statistics/volume-2/issue-1/In-season-prediction-of-batting-averages--A-field-test/10.1214/07-AOAS138.full doi: 10.1214/07-AOAS138

Dogucu, M., Johnson, A., & Ott, M. (2021). bayesrules: Datasets and supplemental functions from bayes rules! book [Computer software manual]. Retrieved from https://github.com/bayes-rules/bayesrules (R package version 0.0.2.9000)

Efron, B., & Morris, C. (1975, June). Data Analysis Using Stein's Estimator and its Generalizations. *Journal of the American Statistical Association*, *70*(350), 311–319. Retrieved 2023-06-17, from http://www.tandfonline.com/doi/abs/10.1080/01621459.1975.10479864 doi: 10.1080/01621459.1975.10479864

James S Clark. (2021, March). *Markov chain Monte Carlo: Principles for Posterior Simulation* [Lecture]. Duke University ENV/BIO 665. Retrieved from https://rstudio-pubs-static.s3.amazonaws.com/733088_037a9e6d68b348bf899debbee8eec406.html

Jensen, S. T., McShane, B. B., & Wyner, A. J. (2009, December). Hierarchical Bayesian modeling of hitting performance in baseball. *Bayesian Analysis*, *4*(4). Retrieved 2023-05-08, from https://projecteuclid.org/journals/bayesian-analysis/volume-4/issue-4/Hierarchical-Bayesian-modeling-of-hitting-performance-in-baseball/10.1214/09-BA424.full doi: 10.1214/09-BA424

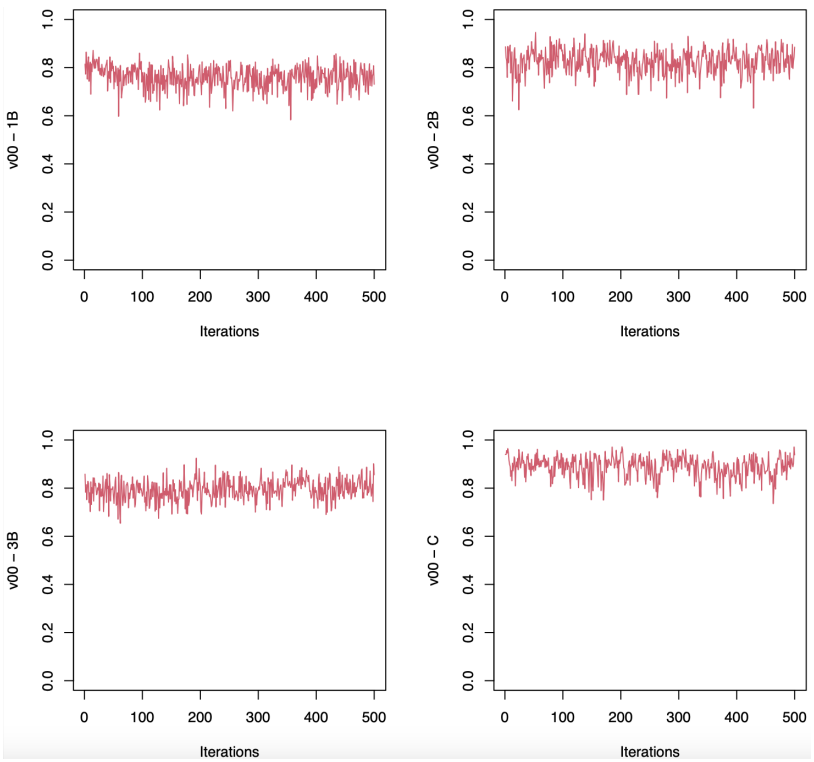

Figure 1: Draws from the posterior distribution of ν_{00k} for predicting the 2019 season

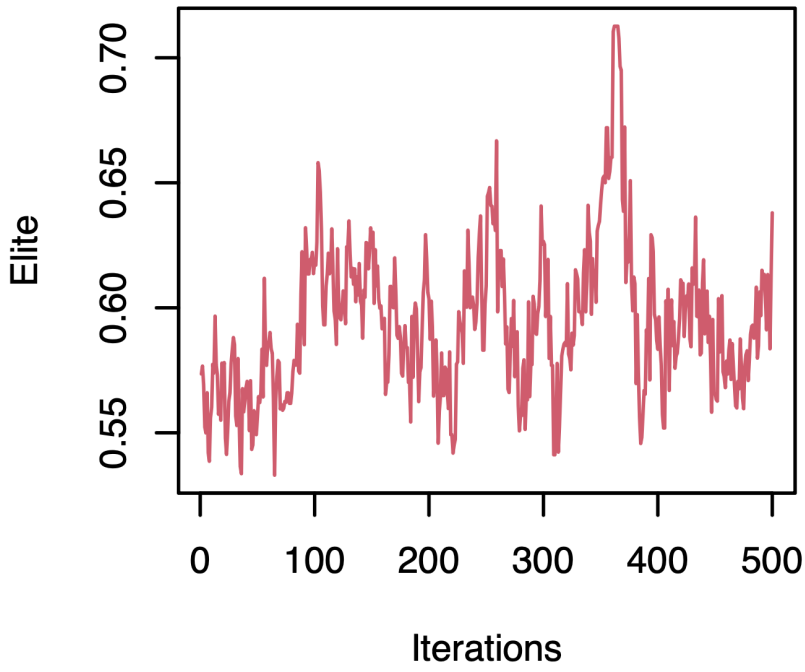

Figure 2: Draws from the posterior distribution of E_{ij} for predicting the 2019 season

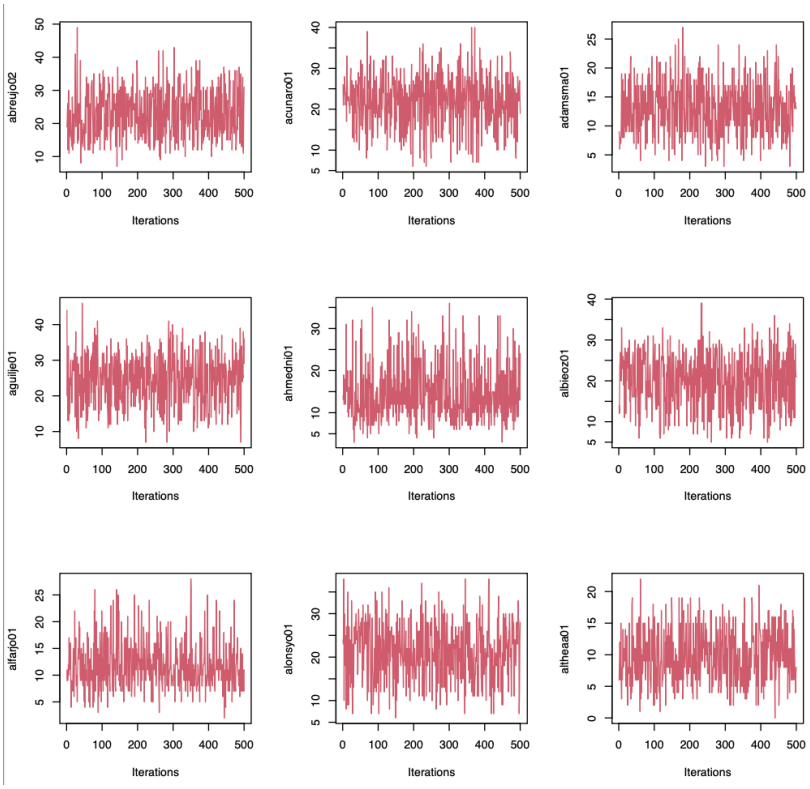

Figure 3: Draws from the posterior distribution of home run total for predicting the 2019 season

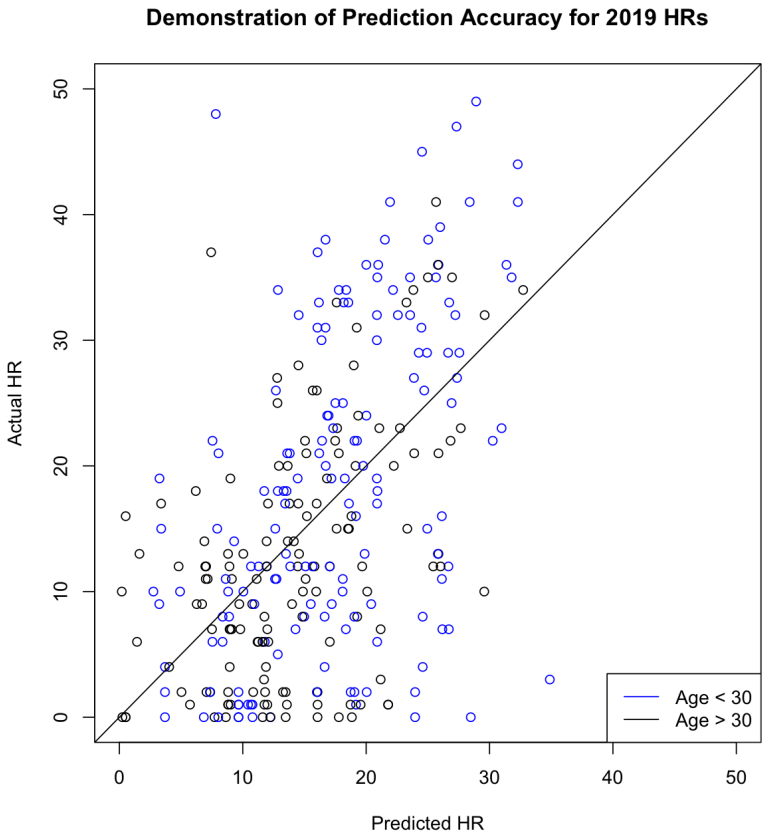

Figure 4: Predicted vs Actual HR Count by Age