Bayesian Analysis (2009) 4, Number 4, pp. 653–660

Comment on Article by Jensen et al.

Jim Albert* and Phil Birnbaum†

1 Introduction

Prediction of future batting performance is an important problem in baseball. Due to trades and the free agent system, there is a good movement of players between teams in the "hot-stove league" (the baseball off-season) and teams will acquire new players with the hope that they will achieve particular performances in the following season. The authors propose a Bayesian hierarchical modeling framework for estimating home run hitting probabilities and making predictions of future home run hitting performance. Generally, this is an attractive methodology, especially when one is collecting data from many players who have similar home run hitting abilities. By use of hierarchical modeling, the estimates of the home run probabilities shrink or adjust the observed rates towards a combined regression estimate. One attractive feature of the Bayesian approach is that it is straightforward to obtain predictions from the posterior predictive distribution and the authors test the value of their method by comparing it with two alternative prediction systems MARCEL and PECOTA. It is straightforward to fit these hierarchical models by MCMC algorithms and the authors provide the details of this fitting algorithm.

Although we admire the authors' paper from a Bayesian modeling/computation perspective, it seems deficient from the application (baseball perspective). There is a substantial research on home run hitting and in the modeling of career trajectories of ballplayers and we believe this research should be helpful in defining relevant covariates and proposing realistic models for trajectories. In the following comments, we discuss several concerns with the basic modeling framework, focus on the choice of suitable adjustments and suggest a more flexible framework for modeling career trajectories.

2 Data

The authors use data from the Lahman database where the counts of home runs and at-bats are collected for each player for each season in the period 1990 and 2005. Although this is a rich dataset, we are puzzled that the authors did not use the more detailed play-by-play data available from the Retrosheet organization (`www.retrosheet.org`). This dataset is easy to access and manipulate. As will be seen shortly, this richer dataset would allow for the inclusion of suitable covariates in the adjustment of the home run rates.

*Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH, http://www-math.bgsu.edu/~albert/

†Society of American Baseball Research, http://philbirnbaum.com/

 DOI:10.1214/09-BA424A

3 Adjustments for Home Run Rates

In comparing baseball hitters across eras, Schell (2005) explains the importance of adjusting home run rates for the era of play, the distribution of league-wide talent, the ballpark effect, and a player's late-career decline. Adjustments for league-wide talent and the ballpark are also crucial in the modeling of a player's hitting trajectory and the prediction of future performance. There have been dramatic changes in home run hitting from 1990 to 2005. The overall major league home run rate increased by 26% between 1992 and 1993 and the rate has shown a 50% increase in this 15 year period. Schell documents the significant impact of ballparks on the pattern of home run hitting. In the current baseball season, it appears to be much easier to hit home runs in the new Yankee stadium in New York. The park factor for the new Yankee Stadium is currently 1.295, which means that the rate of home run hitting in the Yankee home games is about 30% higher than the rate of home run hitting in the Yankees away games.

One can understand changes in league-wide hitting talent by the fitting of a random effects model. For a given season, we observe the number of home runs and at-bats (y_i, n_i) for all batters. We assume that y_i is binomial(n_i, p_i) and then we assume the home run probabilities $\{p_i\}$ follow a beta distribution with shape parameters a and b. The fitted values $\hat{a}$ and $\hat{b}$ are informative about the location and shape of the home run abilities of the batters. This random effects model is fit separately for each season, obtaining estimates $\hat{a}_j$ and $\hat{b}_j$ for season j. The top graph in Figure 1 displays the median home run ability of the players for the seasons 1990 to 2005, and the bottom graph plots the interquartile spread of the home run ability distribution against season. This figure shows dramatic changes in the location and spread of talent of hitting home runs in this 15 year period. One way of adjusting a player's season home run rate compares his rate relative to the distribution of home run rates hit for that particular season. Specifically, one can compute a predictive standardized score as in Albert (2009) using the average and standard deviation of the predictive distribution.

This paper does include some adjustments in their regression model (2), specifically, covariates for the home ballpark and fielding position. As the authors explain, the data does not break down a player's home run data by home and away games and so the "home ballpark" covariate actually confounds two variables, the ballpark effect and the team hitting ability. One could define a true ballpark effect by using the Retrosheet data. We are puzzled by the inclusion of the fielding position covariate. Although there are some tendencies, for example, first-basemen tend to hit more home runs than second-basemen, modern hitters of all non-pitching positions are proficient in hitting home runs. Why do the authors believe that fielding position is an important covariate? More importantly, why do the authors believe that players of different positions have different home run trajectories?

Another possible regression adjustment is the number of opportunities AB. There is a general positive correlation between AB and home run rate – players with more at-bats tend to hit a higher rate of home runs. Also, if a young player has a limited number of AB one season, it is more likely that he will have a small number of home runs and be sent back to the minors the following season. Also the number of AB and

the player's career trajectory provides a good prediction of the player's AB in a future season. (The authors assume that the player's 2006 AB is the same as the AB in the previous season.)

4 Elite/Non-Elite Players

The authors introduce a latent elite variable in their model with the justification that "that there exists a sub-group of elite home run hitters within each position that share a higher mean home run rate". The authors do not present any evidence in the paper that home run rates cluster in two groups of non-elite players and elite players. In our exploration of these data, there appears to be a continuum of home run ability that is right skewed with a few possible large outliers. It seems that the latent elite variable is introduced not because the data suggests the two clusters, but rather to induce some dependence in the home run rates for the same player. There is a more straightforward way to model this dependence, specifically to assume that each player has a unique trajectory, where the individual player regression coefficient vectors are assumed to follow a common distribution. This comment relates to the authors' approach for modeling trajectories which will be described next.

5 Modeling Career Trajectories

In the motivation for the career trajectories, the authors say that they "favor an approach that involves fewer parameters (to prevent over-fitting)". But they make the very restrictive assumption that players of a particular fielding position share the same career trajectory. This assumption does not reflect the variable trajectory patterns of home run hitting. To illustrate the variability in trajectories, consider the home run hitting patterns of the Hall of Fame players Mickey Mantle and Hank Aaron (both who played the same outfield position) who played in the same era. Figure 2 plots standardized home run rates for both players as a function of age, where the rates have been standardized using the predictive distribution as described above. Note that Mantle peaked in his late 20's and declined quickly until retirement. In contrast, Aaron peaked in home run hitting ability much later in his career and showed a more gradual decline towards the end of his career.

It can be difficult to estimate the player trajectories individually using regression models due to the high variability of the observed rates as shown in Figure 1. But one can obtain good smoothed estimates of the individual trajectories by use of a multilevel model. If the vector of regression coefficients for the ith player is represented by β_i, then one can assume that the $\{\beta_j\}$ are a random sample from a common normal distribution with mean vector β and variance-covariance matrix Σ, and the hyperparameters β, Σ are assigned a vague prior at the second state. The posterior estimates smooth the individual trajectory estimates towards a common trajectory. This multilevel model is shown to be successful in smoothing trajectories of WHIP (walk and hit) rates for pitchers in Albert (2009). We have also used it for estimating trajectories of batter on-

base percentages, and we would expect similar good results for estimating trajectories of home run rates. This analysis would lead to more realistic estimates of career trajectories and likely better predictions of future home run hitting. Certainly, one should make different predictions for the home run hitting for a 35-year old Mickey Mantle and a 35-year old Hank Aaron since their patterns of decline were very different.

6 A Sabermetrics Perspective

Sabermetrics is the scientific search for objective knowledge about baseball, and the search for better predictions of future performance is certainly something that sabermetricians – especially those who may be employed by major league clubs - are interested in. But they are concerned with more than just accurate predictions; they are concerned with what it is the projection reveals about players and changes in their performance.

Bill James, in a discussion about the existence of clutch hitting in James (1984), says "How is it that a player who possesses the reflexes and the batting stroke and the knowledge and the experience to be a .260 hitter in other circumstances magically becomes a .300 hitter when the game is on the line? How does that happen? What is the process? What are the effects? Until we can answer those questions, I see little point in talking about clutch ability." Likewise, sabermetricians are interested in the process that leads to a prediction of home run hitting.

Sabermetricians are unsatisfied with mere predictions, no matter how accurate. Given an accurate prediction of future performance, they ask, "what is it about that prediction that makes it accurate? What does it tell us about the relationship of past performance to future performance?"

One attractive feature of MARCEL is that it gives us clues to what might be going on. Tango (2004) gives the full MARCEL algorithm, in which we can see the assumptions that went into the formula. We see how it weights recent performance relative to more distant performance, how much one should regress to the mean, and how one adjusts the predictions to adjust to changes to league norms. These individual assumptions can be adjusted in order to minimize prediction error, and, in so doing, we would come closer to learning objective information about player hitting.

The Bayesian modeling approach presented in this paper, however, is more complex and opaque. It performs only marginally better than MARCEL, while using more information such as home team scoring and player position. It is uncertain what an experienced sabermetrician would learn from the Bayesian process, and it is uncertain whether the (marginally) improved predictions are the result of a better model, or simply the result of the additional information being used.

Further, while the Bayesian model has shown itself to be successful in predicting, certain of its assumptions are almost certain to be false. As has been noted, the classification of hitters into only two categories – elite and non-elite – is certainly false, as home-run-hitting ability appears to be a continuum; there is no evidence that the distribution of home run rates, even by position, is bimodal.

The fact that the Bayesian model gives reasonable estimates cannot be taken as evidence that the assumptions are correct. For instance, a black-box model that predicts swine-flu infection rates is valuable, but, if the assumptions that went into the model are correct, this model is useful in predicting future outbreaks. If the assumptions are incorrect, the predictions based on the model may be inaccurate.

Sabermetricians would be very interested in the success of the Bayesian model in predicting home run rates for younger hitters; as Table 2 of the paper shows, the Bayesian algorithm beats MARCEL 62% of the time, and beats PECOTA 100% of the time. We note, however, that this is based only on a sample of eight players. Still, one could discover possible attributes of the prediction methodology by a case-by-case exploration. It would useful to see the full list of players and their estimates, along with a discussion of what kinds of players, such as power hitters or high-average players, are better estimated than others types of players. This would provide a useful comparison of the methods, and provide a direction for future research to improve the knowledge that the field of sabermetrics has compiled about the aging process.

As it stands now, the Bayesian method has made sabermetrics aware that slight improvements over MARCEL are possible, but, without further exploration, we are left with little understanding of where the improvements came from, where MARCEL is weak, what assumptions need to be refined, or, indeed, how the aging process in baseball can better be explained.

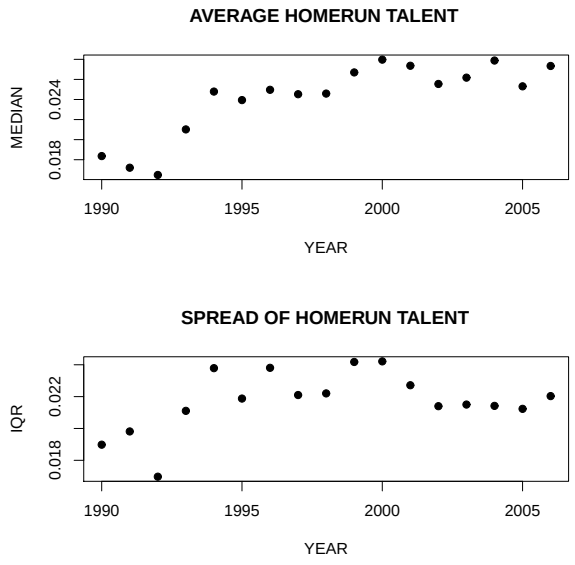

Figure 1: Fitted home run talent distributions for the seasons 1990 to 2005. The top graph displays the median home run ability and the bottom graph displays the interquartile range of the talent distribution.

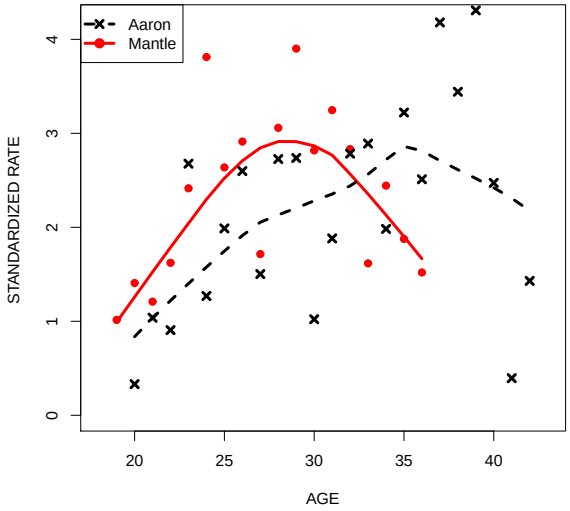

Figure 2: Standardized home run rates for Mickey Mantle and Hank Aaron ploted as a function of age. The lowess smooths show that the home run trajectories of the two players were significantly different.

7 Summing Up

The authors have proposed a useful hierarchical modeling framework and illustrated the potential benefits of Bayesian modeling in predicting future home run counts. But we believe the methods could be substantially improved by the proper adjustment of the home run rates, the inclusion of useful covariates, and more realistic modeling of the career trajectories. From the viewpoint of a baseball general manager, the prediction of a particular player's future performance is very important and it seems that this prediction has to allow for the player's unique career trajectory pattern. For the problem of individual predictions, we don't believe this methodology will be very helpful, since all players of a particular fielding position are assumed to have the same trajectory and lumped into the broad elite/non-elite classes. But we do believe that this general approach, with the changes described above, can be used to make helpful predictions of offensive performance.

References

Albert, J. (2009). "Is Roger Clemens' WHIP Trajectory Unusual." *Chance*, 22: 8–22.

James, B. (1984). *The 1984 Baseball Abstract*. Ballentine Books.

Schell, M. (2005). *Baseball's All-Time Best Sluggers: Adjusted Batting Performance*

from Strikeouts to Home Runs. Princeton University Press.

Tango, T. (2004). "Marcel the Monkey Forecasting System." Http://www.tangotiger.net/archives/stud0346.shtml.