Bayesian Analysis (2009) **4**, Number 4, pp. 661–664

Comment on Article by Jensen et al.

Mark E. Glickman[*]

I offer my congratulations to Jensen, McShane and Wyner (hereafter JMW) on their paper modeling home run frequencies of Major League Baseball (MLB) players. It is always refreshing to read such a clearly written, well-organized paper on a topic of interest to a broad audience and one that illustrates cutting edge modeling and computational tools in Bayesian Statistics. It is also worth noting that the first author is becoming an accomplished researcher in quantitative aspects of baseball, most recently having developed complex statistical models for evaluating fielding (Jensen et al. 2009). The current paper adds to his accruing and impressive list of work on Statistics in sports.

In the current paper, the authors develop and investigate a model for home run frequencies for MLB seasons from 1990 through 2005 based on publicly available data. The data contains player performance information aggregated by season, so examining within-season variation is not possible. Home run frequencies for a player within a season are modeled as binomial counts (out of the total number of at-bats, appropriately defined), and the probability of a home run during a season is a function of the player's position, team, and age. The authors make some interesting specific assumptions that result in a unique model. First, they posit that the effect of age on the log-odds of the probability of a home run follows a cubic B-spline relationship for a given field position. Second, they assume a latent categorization of each player in a given season as elite versus non-elite, essentially treating a player's home run frequency as a mixture of two binomial components with different probabilities. Third, the latent elite status for each player is assumed to follow a Markov process with transition probabilities that are common for all players at the given field position. The authors also investigate a generalization of their basic model in which the transition probabilities can vary by player through model components specific to players at that position. The entire model is fit via MCMC simulation from the posterior distribution, and performance of their approach is evaluated through measures that compare model predictions in 2006 to observed home run frequencies. They conclude that their basic model fares well against existing competitor approaches that are not nearly as sophisticated. The authors deserve credit for constructing a model that is competitive with one that makes use of data obtained on a daily basis. It is also particularly impressive that their model predicts well given the paucity of covariate information.

One can raise minor quibbles with the authors' approach, but many of the concerns are an artifact of the constraints on the data available to them. For example, the ability to account for within-season variation strikes me as a clear deficiency in modeling home run probabilities. Given that players are generally improving from year to year in their twenties, it is not unreasonable to speculate that some of this improvement is occurring within a season rather than between seasons. Because the data JMW use is aggregated

[*]Boston University School of Public Health, Boston, MA, mailto:mg@bu.edu

DOI:10.1214/09-BA424B

by season, it is impossible to infer such changes. The authors also incorporate a team indicator in their model, which ostensibly is a proxy for playing half of the time in their own ballpark, though this does not account for minor artifacts such as within-season player trades. As JMW note, this team parameter may be difficult to interpret when it applies to a whole season of games. If individual game-specific data were available, then the impact of the actual ballpark could be incorporated into the model which may have a profound effect on inferences. My own bias is to wonder whether modeling and predicting home run frequencies is a question that baseball front office staff or other professionals really want answered. While forecasting home run probabilities seems like an interesting theoretical question, various metrics to measure hitting rates might be of greater practical utility. The authors do mention at the conclusion of the paper their interest in pursuing such activities. I also found curious that the expanded model involving Markov transition probabilities that varied by player produced worse predictions than the simpler model in which the transition probabilities were constrained to vary only by player-position. This may suggest some combination of a model not sufficiently capturing important features of the data, or an expanded model that is too highly parameterized.

To me, the most interesting aspect of the paper is the decision to incorporate a latent indicator of elite status into the model, and the accompanying stochastic process. On the one hand, JMW are able to account for variation in home run rates and improve predictions by introducing a 2-state hidden Markov model (HMM). One clear benefit of incorporating this model component is that it allows answering questions about when certain players can be considered elite versus non-elite. On the other hand, I wonder whether a 2-state Markov model is the most appropriate and most flexible for predicting home run frequencies. The authors consider a HMM in which players at the same position share the same transition probabilities, and another in which the transition probabilities vary by player but are centered at position-specific distributions. In both cases, the size of the effect of being elite for all players at the specified position is the same. I realize that JMW are focused on keeping the model as simply parameterized as possible, but the question arises whether accuracy (especially predictive accuracy, one of the main implied goals of the paper) is being sacrificed. Given that all the parameters of the HMM are integrated out of the posterior distribution in making predictions, it is the *structure* of the HMM that is most crucial, and not inferences about any of the HMM parameters.

The authors' HMM assumes that players at any given time are in one of two states, once accounting for age, position and team. However, it strikes me that player effects (beyond the effect of age, position and team) more justifiably fall on a continuum. A natural way to modify JMW's model is to assume

$$\operatorname{logit} \theta_{ijkb} = \alpha_k + \beta_b + f_k(A_{ij}) + \delta_{ijk} \tag{1}$$

where θ_{ijkb} is the home run probability for player i with home ballpark b in season j at position k; α_k, β_b and $f_k(A_{ij})$ are as defined in JMW; and δ_{ijk} is a player-specific effect following a stochastic process with a continuous state-space, such as

$$\delta_{ijk} \sim \mathrm{N}(\delta_{i,j-1,k},\ \psi^2), \tag{2}$$

where initial player effects may be assumed drawn from a common distribution centered at a position-specific model component,

$$\delta_{i1k} \sim \mathrm{N}(\eta_k,\ \phi^2) \tag{3}$$

with position-specific effects η_k. This model assumes that, beyond the effects of ballpark, position and age, an individual player effect in a given season is drawn from a distribution centered at last season's mean, thus inducing a time-correlation particular to that player. Such an approach can represent trajectories of not only elite players, but also better-than-average players as well as worse-than-average players. Similar models for binomial data in a game/sports context have been examined by Fahrmeir and Tutz (1994) and Glickman (1999), among others, though these approaches do not include an additive spline component for age. Various changes to the assumptions in (2) and (3) could be considered, such as having the innovation variance, ψ^2, depend on player position (that is, ψ_k^2), the transition model could be heavy-tailed, such as a t-distribution instead of normal (which would account for occasional bursts of improvement in home run probability), or having the prior variance, ϕ^2, depend on the player position (that is, ϕ_k^2).

An advantage to a continuous state-space compared to a 2-state system is that it recognizes varying degrees of improvement and worsening over time beyond what is captured by age-specific effects. Substituting the HMM in the authors' framework with that in (2) should involve straightforward modifications to the MCMC algorithm, so the computational details ought to involve tractable calculations. Again, because the parameters of a continuous state-space model are integrated out of the posterior distribution to obtain predictive inferences, or even age-curve estimates, the richer structure compared to the 2-state HMM may result in more reliable inferences. The richer structure may also more appropriately calibrate the levels of uncertainty in predictions which appear overly conservative as evidenced in Table 1 of their paper. Of course, one needs to fit such a model to the data to be convinced of such speculation.

Notwithstanding some of my suggestions for alternative directions the authors could take in further refining their model, I think that their approach makes an important contribution to a growing literature on sophisticated methods in analyzing sports data. Modeling the effect of age through a cubic B-spline is a nice feature of their approach, and accounting for time dependence in home run rates through a hidden Markov model is a novel addition, even though my feeling is that a continuous state-space Markov model may be more promising. I look forward to the continued success and insightful work from this productive group of researchers.

References

Fahrmeir, L. and Tutz, G. (1994). "Dynamic stochastic models for time-dependent ordered paired comparison systems." *Journal of the American Statistical Association*, 89: 1438–1449. 663

Glickman, M. E. (1999). "Parameter estimation in large dynamic paired comparison experiments." *Applied Statistics*, 48: 377–394. 663

Jensen, S. T., Shirley, K., and Wyner, A. J. (2009). "Bayesball: a Bayesian hierarchical model for evaluating fielding in major league baseball." *Annals of Applied Statistics*, 3: 491–520. 661