A Bayesian Model for Evaluating MLB Hitting Performance

Carson Sievert

December 9, 2014

1 Introduction

Statistical methods have been used to understand and analyze the game of baseball for many years. An obvious and popular application is modeling and predicting player performance. Jensen [2009] proposes a bayesian model for predicting hitting performance at the season level. This model was shown to be effective at prediction - especially for players with little experience at the Major League level thanks to a shrinking towards the population mean component. In the discussion of that paper, reviewers call for a model at the game (rather than) level. The reasoning is to account for well-known factors on performance such as a “park effects” and “home field effects”. This paper explores a step in that direction by modeling home run performance at the game level and accounting for any “home field effect”.

2 Data Collection

The data used for this project was taken from the Major League Baseball Advanced Media (MLBAM) website `http://gd2.mlb.com/components/game/mlb/` using the `R` package `pitchRx` Sievert [2014]. The number of home runs and number of atbats in every game over the 2012 season was collected for the five players with the highest home run total in 2012.

3 Model Formulation

Since home runs are such a rare event, we may want some type of zero-inflated model if we want to model the number of home runs in a particular game. For that reason, consider a latent Bernoulli random variable X_{ij} that will govern whether player i in game j has “potential” to get at least one success in n_{ij} atbats. The Bernoulli parameter θ_{ij} could thought of as a players “potential” for a particular game.

$$f(X_{ij} = x_{ij}|\theta_{ij}) = \theta_{ij}^{x_{ij}}(1-\theta_{ij})^{(1-x_{ij})}, 0 < \theta_{ij} < 1$$

where

$$x_{ij} = \begin{cases} 0 & \text{no "potential success"} \\ 1 & \text{"potential success"} \end{cases}$$

To model the actual number of successes for player i in game j, we use a Binomial random variable Y_{ij} where the number of trials n_{ij} is considered fixed. In some sense, the success probability ψ_{ij} could thought of as a player’s “ability” to hit a home run in a particular game. To get the marginal distribution for Y_{ij}, we could “integrate out” the latent X_{ij}. For example,

$$\begin{aligned}
P(Y_{ij}=0) =& P(Y_{ij}=0|X_{ij}=0)P(X_{ij}=0)+P(Y_{ij}=0|X_{ij}=1)P(X_{ij}=1)=(1-\theta_{ij})+\theta_{ij}(1-\psi_{ij})^{n_{ij}} \\
P(Y_{ij}=1) =& P(Y_{ij}=1|X_{ij}=1)P(X_{ij}=1)=n_{ij}\psi_{ij}(1-\psi_{ij})^{n_{ij}-1}\theta_{ij} \\
P(Y_{ij}=2) =& P(Y_{ij}=2|X_{ij}=1)P(X_{ij}=1)=\binom{n_{ij}}{2}\psi_{ij}^2(1-\psi_{ij})^{n_{ij}-2}\theta_{ij} \\
&\vdots
\end{aligned}$$

We can describe this distribution in general by:

$$P(Y_{ij}=y_{ij})=\begin{cases}(1-\theta_{ij})+\theta_{ij}(1-\psi_{ij})^{n_{ij}} & \text{if } y_{ij}=0 \\ \binom{n_{ij}}{y_{ij}}\psi_{ij}^{y_{ij}}(1-\psi_{ij})^{(n_{ij}-y_{ij})}\theta_{ij} & \text{if } y_{ij}\in\{1,2,\ldots,n_{ij}\}\end{cases},0<\psi_{ij}<1$$

Now we consider the parameters θ_{ij} and ψ_{ij} to be unknown quantities and can be described through the deterministic relationships:

$$\log(\frac{\theta_{ij}}{1-\theta_{ij}})=\beta_i+\alpha_i H_{ij}$$

$$\log(\frac{\psi_{ij}}{1-\psi_{ij}})=\delta_i+\lambda_i H_{ij}$$

where H_{ij} is an indicator variable that is 1 if game j was a "home" game for player i and 0 otherwise. This leaves use to choose distributions for the (independent) set of random quantities $\beta_i, \delta_i, \alpha_i, \lambda_i$:

$$\begin{aligned}
\beta_i &\sim N(\beta_0,\sigma_\beta^2) \\
\alpha_i &\sim N(\alpha_0,\sigma_\alpha^2) \\
\delta_i &\sim N(\delta_0,\sigma_\delta^2) \\
\lambda_i &\sim N(\lambda_0,\sigma_\lambda^2)
\end{aligned}$$

where $\beta_0, \sigma_\beta, \alpha_0, \sigma_\alpha, \delta_0, \sigma_\delta, \lambda_0,$ and σ_λ are all "known" quantities that are chosen such that we have diffuse priors that reflects a lack of prior knowledge. In particular, the mean and standard deviation was set to 0 and 5, respectively.

4 Fitting the model

This section addresses how samples from the joint posterior distribution were obtained. Full conditional posterior distributions were derived for all quantities involved (besides the observed data). A Gibbs sampling algorithm was employed on these full conditionals to obtain draws from the joint distribution.

$$p(\beta_i,\delta_i,\alpha_i,\lambda_i|y_i)\propto p(y_i|\beta_i,\delta_i,\alpha_i,\lambda_i)p(\beta_i)p(\delta_i)p(\alpha_i)p(\lambda_i)$$

The notation $p(x|\cdot)$ is now used to represent the conditional density of X given all other quantities. Note that y_i is used to represent a vector whose length is equal to the number of games played by player i.

4.1 Full conditionals

1. The full conditional densities of β_i for $i = 1, \ldots, 5$ are:

$$p(\beta_i|\cdot) \propto p(y_i|\cdot)p(\beta_i) \propto p(y_i|\cdot)exp\left\{\frac{-(\beta_i - \beta_0)^2}{2\sigma_\beta^2}\right\}$$

   In order to sample from this form, an *adaptive* Metropolis-Hastings algorithm is used with a Gaussian proposal $N(\beta_i^{(k)}, (\tau_\beta^2)^{(k)})$ where $\beta_i^{(k)}$is the simulated value from the k^{th} iteration. The adaptive piece helps to obtain a reasonable acceptance rate by increasing or decreasing $(\tau_\beta^2)^{(k+1)}$ based on whether or not k^{th} proposal was accepted.

2. The full conditional densities of δ_i for $i = 1, \ldots, 5$ are:

$$p(\delta_i|\cdot) \propto p(y_i|\cdot)p(\delta_i) \propto p(y_i|\cdot)exp\left\{\frac{-(\delta_i - \delta_0)^2}{2\sigma_\delta^2}\right\}$$

   In order to sample from this form, an *adaptive* Metropolis-Hastings algorithm is used with a Gaussian proposal $N(\delta_i^{(k)}, (\tau_\delta^2)^{(k)})$ where $\delta_i^{(k)}$is the simulated value from the k^{th} iteration. The adaptive piece helps to obtain a reasonable acceptance rate by increasing or decreasing $(\tau_\delta^2)^{(k+1)}$ based on whether or not k^{th} proposal was accepted.

3. The full conditional densities of α_i for $i = 1, \ldots, 5$ are:

$$p(\alpha_i|\cdot) \propto p(y_i|\cdot)p(\alpha_i) \propto p(y_i|\cdot)exp\left\{\frac{-(\alpha_i - \alpha_0)^2}{2\sigma_\alpha^2}\right\}$$

   In order to sample from this form, an *adaptive* Metropolis-Hastings algorithm is used with a Gaussian proposal $N(\alpha_i^{(k)}, (\tau_\alpha^2)^{(k)})$ where $\alpha_i^{(k)}$is the simulated value from the k^{th} iteration. The adaptive piece helps to obtain a reasonable acceptance rate by increasing or decreasing $(\tau_\alpha^2)^{(k+1)}$ based on whether or not k^{th} proposal was accepted.

4. The full conditional densities of λ_i for $i = 1, \ldots, 5$ are:

$$p(\lambda_i|\cdot) \propto p(y_i|\cdot)p(\lambda_i) \propto p(y_i|\cdot)exp\left\{\frac{-(\lambda_i - \lambda_0)^2}{2\sigma_\lambda^2}\right\}$$

   In order to sample from this form, an *adaptive* Metropolis-Hastings algorithm is used with a Gaussian proposal $N(\lambda_i^{(k)}, (\tau_\lambda^2)^{(k)})$ where $\lambda_i^{(k)}$is the simulated value from the k^{th} iteration. The adaptive piece helps to obtain a reasonable acceptance rate by increasing or decreasing $(\tau_\lambda^2)^{(k+1)}$ based on whether or not k^{th} proposal was accepted.

4.2 Monitoring Convergence

Three different chains with randomly dispersed starting values were each run for 10000 iterations with an adaptation period of 1000 and burnin of 5000. The Gelman and Rubin scale reduction factor for each parameter was computed and is presented below. Clearly, the factor value looks good (very close or equal to 1) for all α_i in table 1. Similarly, the factor value looks good for all β_i in table 2. The factor values for some δ_i in table 3 and λ_i in table 4 are a little higher than we would like to see, but they don't cause an overwhelming reason for worry.

To ensure the algorithm has explored the entire sample space and has also not over represented areas of high probability, we track the proportion of proposed jumps that are accepted for each parameter sampled via Metropolis-Hastings in table 5. Note that all of these proportions are between 0.31 and 0.61; thus, all these rates are inside the rule of thumb of 0.2 to 0.6.

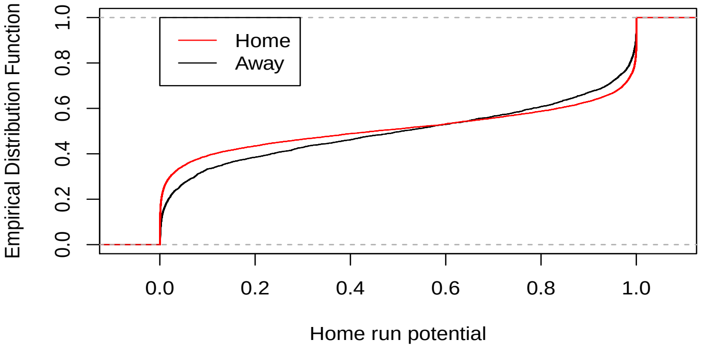

Figure 1: Empirical Distribution Function for Edwin Encarnacion's potential to hit home runs away (black) versus home (red).

4.3 Is there a home effect?

It's natural to think that players tend to perform better at their home stadium. However, according to the central 95% credible intervals for each α_i (in table 6), there obviously is not a significant home field effect on the "potential" to hit a home run for any of these players. Similarly, according to the central 95% credible intervals for each λ_i (in table 7), there obviously is not a significant home field effect on the "ability" to hit a home run for any of these players.

Although these effects are not significantly different, we can still investigate differences in the empirical distribution function for potential and ability given potential. Since Edwin Encarnacion is the player with the highest posterior mean for α_i among all the players, he is a good choice to demonstrate this difference for potential. In figure 1, the empirical distribution functions for Encarnacion clearly show that playing at home has a very small positive impact on home run potential.

In a similar manner, we can address the impact of playing at home given that a player has "potential". Since Josh Hamilton is the player with the highest posterior mean for λ_i among all the players, he is a good choice to demonstrate this difference in ability given potential. In figure 2, the empirical distribution functions clearly show that playing at home has a very large positive impact on this probability. Based on figure 2, we would conclude that, given Hamilton has potential in a game where he bats four times, the probability he hits at least one home run is 0.9 at home but only 0.2 away.

5 Model Assessment

To assess the model we address the adequacy of the "zero inflated-binomial" model for describing the data with respect to the number of games with no home runs and the range in the number of home runs per game. Using the procedure outlined on page 311 of the course notes, we obtain a posterior predictive p-value of 0.18 for the number of zeros and 0.97 for the range. The first p-value suggests that this model is adequate for describing the number of games without a home run. The second p-value suggests that this model lacks an ability to describe the spread in the number of home runs per game. Figure 3 shows the two different posterior predictive distributions and the corresponding actual values (vertical dashed line) computed from the observed data.

6 Conclusion

The "zero inflated-binomial model" for home run hitting at the game level allows for inference related to the impact that playing at home has on both a player's "potential" to hit home runs and "ability" to hit

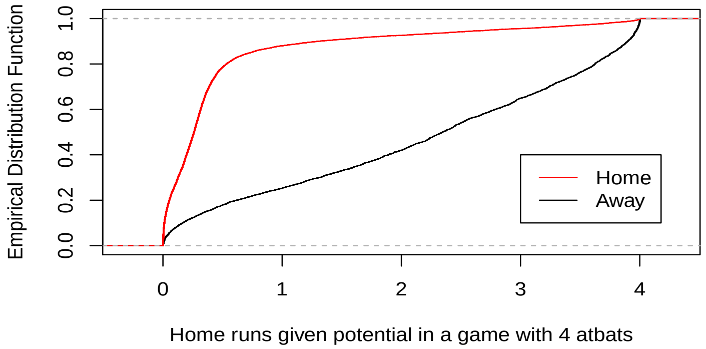

Figure 2: Empirical Distribution Function for Josh Hamilton's home run hitting ability given potential in a game with 4 atbats. The function for a home game is shown in red while away is in black.

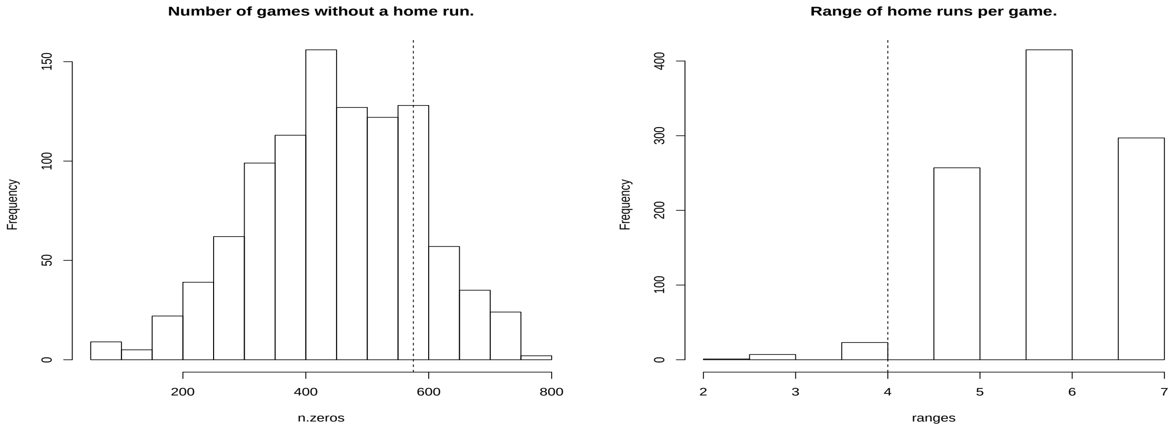

Figure 3: Posterior predictive distributions.

home runs. Although a significant difference in home effects was not shown for either, that would likely change if we were include more observations into the model. In particular, there is evidence to believe that playing at home has a positive effect "ability" to hit a home run (given "potential").

References

Wyner Jensen, McShane. Hierarchical bayesian modeling of hitting performance in baseball. *Bayesian Analysis*, 4(4):631–652, 2009.

Carson Sievert. Taming pitchf/x data with pitchRx and XML2R. *The R Journal*, 6/1:5–19, 2014. URL `http://journal.r-project.org/archive/2014-1/sievert.pdf`.

	Point est.	Upper C.I.
1	1.00	1.00
2	1.00	1.00
3	1.00	1.01
4	1.00	1.00
5	1.00	1.01

Table 1: The potential scale reduction factor for all alpha parameters

	Point est.	Upper C.I.
1	1.00	1.01
2	1.00	1.01
3	1.01	1.02
4	1.00	1.01
5	1.00	1.01

Table 2: The potential scale reduction factor for all beta parameters

	Point est.	Upper C.I.
1	1.02	1.03
2	1.03	1.06
3	1.03	1.04
4	1.05	1.06
5	1.07	1.14

Table 3: The potential scale reduction factor for all delta parameters

	Point est.	Upper C.I.
1	1.06	1.08
2	1.01	1.05
3	1.04	1.05
4	1.00	1.01
5	1.07	1.08

Table 4: The potential scale reduction factor for all lambda parameters

	alphas	betas	deltas	lambdas
1	0.36	0.50	0.48	0.31
2	0.41	0.40	0.47	0.41
3	0.45	0.55	0.40	0.33
4	0.36	0.40	0.52	0.52
5	0.45	0.56	0.59	0.61

Table 5: Acceptance Rates for model parameters sampled via Metropolis-Hastings

	2.5%	97.5%
1	-9.79	9.75
2	-9.93	9.65
3	-10.08	9.54
4	-9.89	10.42
5	-9.69	9.45

Table 6: Central 95 percent credible intervals for every alpha.

	2.5%	97.5%
1	-5.90	5.24
2	-5.89	6.26
3	-6.33	4.72
4	-4.82	3.92
5	-5.35	5.93

Table 7: Central 95 percent credible intervals for every lambda.