Bayesian Baseball- World Series 2018

Blake Shurtz[1]
[1]Cal State East Bay Department of Statistics, 25800 Carlos Bee Blvd, Hayward, CA 94542

Abstract
This poster presents a model for predicting the outcome of a baseball game with an application toward game 5 of the 2018 World Series between the Boston Red Sox (BOS) and the Los Angeles Dodgers (LAD).

Key Words: Bayesian statistics, baseball, multi-level models

1. The Model

Figure 1: Multilevel Regression Model

$$\mathit{rundiff}_{i,j} \sim N(\mu_{i,j},\sigma)$$
$$\mu_{i,j} = \mathit{team}_{i,j} + \mathit{batting}_{i,j} + \mathit{pitching}_{i,j} + \mathit{fielding}_{i}$$
$$\mathit{batting}_{i,j} = \mathit{singles}_{i,j} + \mathit{doubles}_{i,j} + \mathit{triples}_{i,j} + \mathit{home\ runs}_{i,j} + \mathit{walks}_{i,j}$$
$$\mathit{pitching}_{i,j} = \mathit{strikes}_{i,j} + \mathit{strike\text{-}outs}_{i,j} + \mathit{balls}_{i,j} + \mathit{hits\ allowed}_{i,j}$$
$$\mathit{fielding}_{i} = \mathit{putouts}_{i} + \mathit{assists}_{i} + \mathit{errors}_{i} + \mathit{double\ plays}_{i}$$

The model is an i=2 two-factor model that predicts the difference in runs (“rundiff”) between the home team and away team, whereby a positive value for rundiff indicates a win for the home team. The benefits of using a two-factor model include a consistently defined outcome variable for all possible games, thereby allowing all n=2362 games to be organized into a single data set.

There are j=2 two levels to the model. The game level contains statistical predictors in batting and pitching as well as the response variable. The team level contains varying intercepts for each team and performance statistics in fielding. The multi-level model allows for partial pooling whereby the predicted outcome for each game is balanced between the previous matches between the two teams and each team's overall performance for the season. The model is a varying intercepts / varying slopes model which allows for differences in offensive and defensive performance for all teams.

The model is analyzed in a Bayesian framework with a maximum entropy Gaussian likelihood function. Fielding variables are scaled with standard normal priors and all other predictors have adaptive priors that are themselves a function of the data. The priors for the variances have half-Cauchy distributions. The prior for the correlation matrix between intercepts and slopes is a LKJ “onion method” distribution.

1.1 Model Diagnostics

Both varying intercept (VI) and varying intercept & slope (VIVS) models have a far lower deviance (WAIC) compared to a standard regression. The VIVS has a higher out-of-sample deviance but the difference is not significant. Overall, the model favors home team success and has 53% accuracy without game-level predictors and 87% accuracy with game-level predictors.

Computational approximation of posterior distributions was executed using Hamiltonian Monte Carlo with a No-U-Turn sampler, executed in the software Stan. There was only 1 divergent transition and all parameters have an Rhat of 1.00 or 1.01, indicating precise estimation of all parameters.

2. World Series 2018

2.1 Posterior Predictive Simulation (Games 1-4)

Prior to the 2018 World Series, BOS and LAD had never played a match. Nevertheless, we can simulate games with team-level effects where all of the game-level predictors are zero. In other words, we can simulate the game up until the point that it starts. After each match, the model is updated with the results from the previous game.

2.2 Posterior Predictive Simulation Game 5

LAD hosts game 5 with a 1-1 record at home. The model predicts a 71% probability that LAD wins game 5. Nevertheless, LAD has only a 36% binomial probability of winning the necessary 3 remaining games.

2.2.1 Game 5: Parameterized Posterior Predictive Simulations

The model has been parameterized in order to simulate posterior distributions for team effects and batting, pitching and fielding performance. While BOS is a slightly better team overall than LAD, LAD has a higher mean point differential in batting (.33 pts), pitching (.37 pts) and fielding (.31 pts).

2.2.2 Play-by-Play Updates

Due to the game-level nature of the predictors, the outcome can be updated in real-time. The model begins with the prior 71% probability of success for LAD. Despite being behind by 1 point, the probability of a LAD win stays above p=50% due to LAD's higher batting average. However, Boston scores 3 consecutive runs in innings 6, 7, and 8, securing the pennant with near certainty by inning 8.

Acknowledgements

Special thanks for all of my professors at Cal State East Bay, especially Dr. Ayona Chatterjee and Dr. Bruce Trumbo.

References

McElreath, R. (2016). Statistical Rethinking: A Bayesian Course with Examples in R and Stan. CRC Press. (book)

Gelman and Hill. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press. (book)

Figures

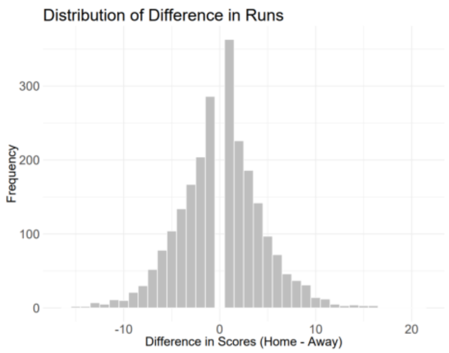

Figure 1: Distribution of sample data of difference in runs. There are almost no games that are ties, and we can see a home team advantage in the sample data.

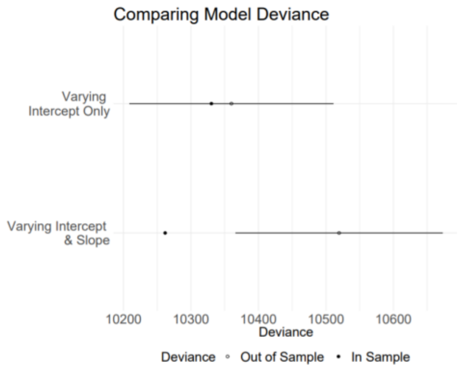

Figure 2: Model Comparison: Varying intercept vs. varying intercept and slope. Varying intercept has lower out of sample deviance (and is therefore better for out-of-sample prediction), but the difference is small.

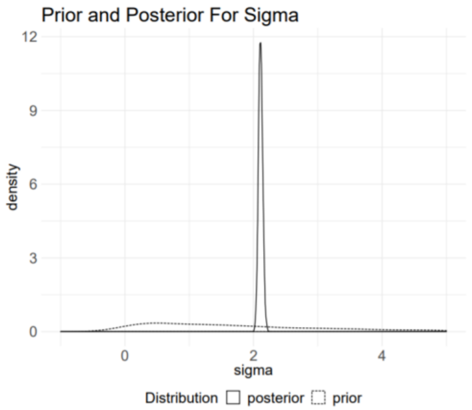

Figure 3: Prior and posterior distribution of variance parameter, the posterior variance is approximately a 2 run difference in runs between home and away teams.

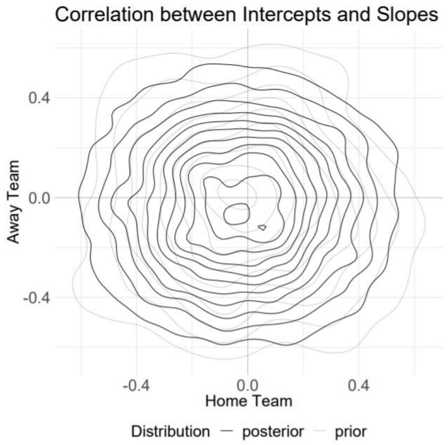

Figure 4: Prior and posterior distribution of correlation between intercepts and slopes for both home and away factors. The plot indicates very little difference in game-level predictors among all matchups.

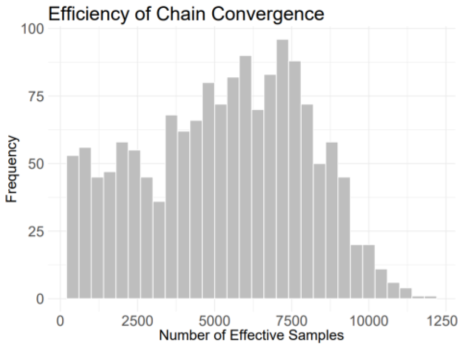

Figure 5: Efficiency of chain convergence for each variable, where the number of iterations are equal to 10,000.

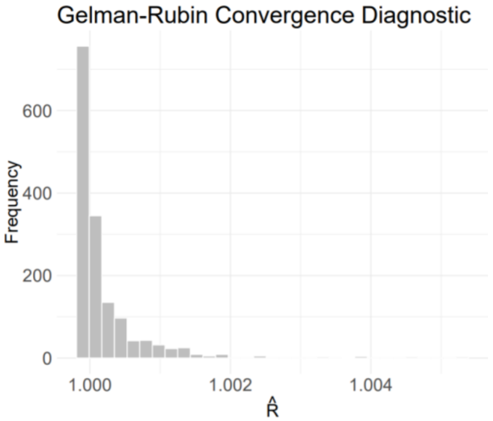

Figure 6: All variables converge with an R-hat of 1 (or slightly higher).

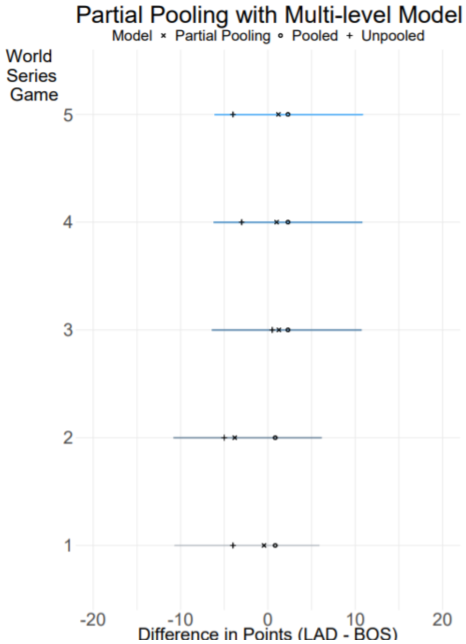

Figure 7: Partial pooling model estimates mean value between complete pooling and unpooled model.

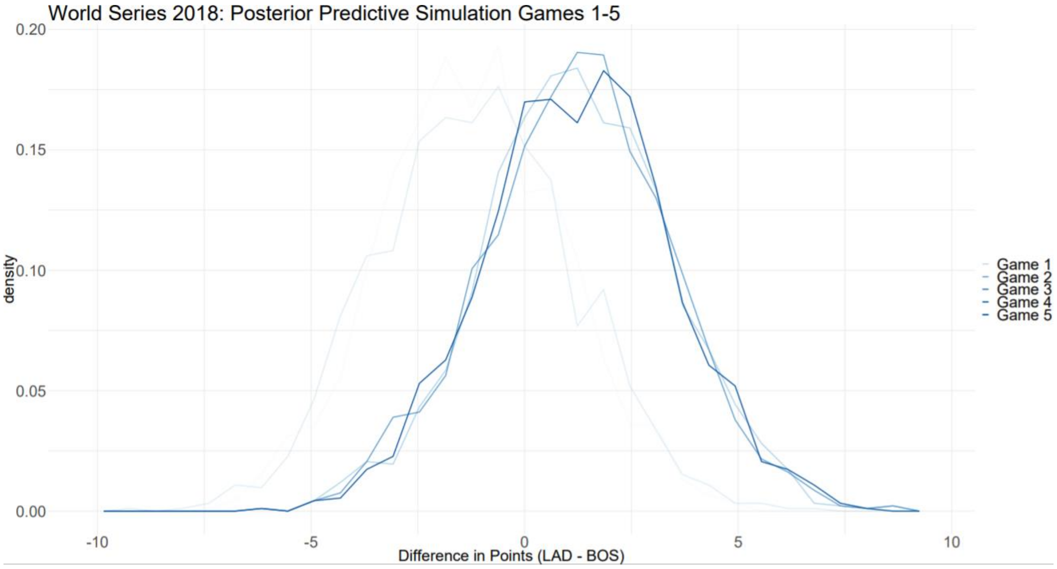

Figure 8: Posterior predictive simulation of games 1 through 5, simulated prior to the game. For games 3-5, with LAD at home, the probability of an LAD win settles around 71%.

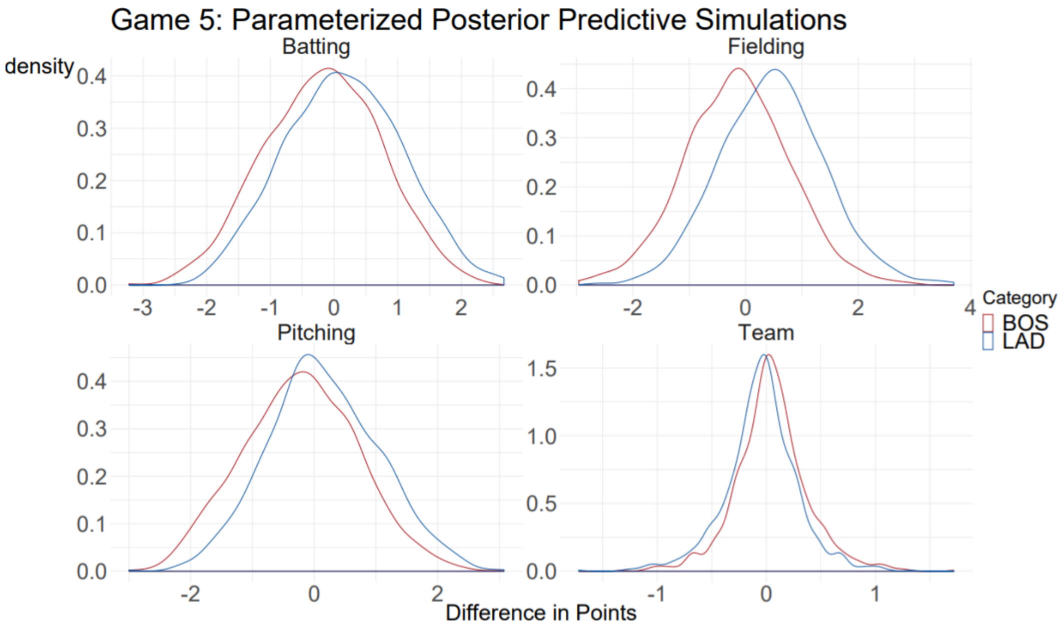

Figure 9: Parameterized posterior predictive simulations. Model was parameterized in order to compare different aspects of team performance. LAD has better batting, fielding and pitching when playing at home.

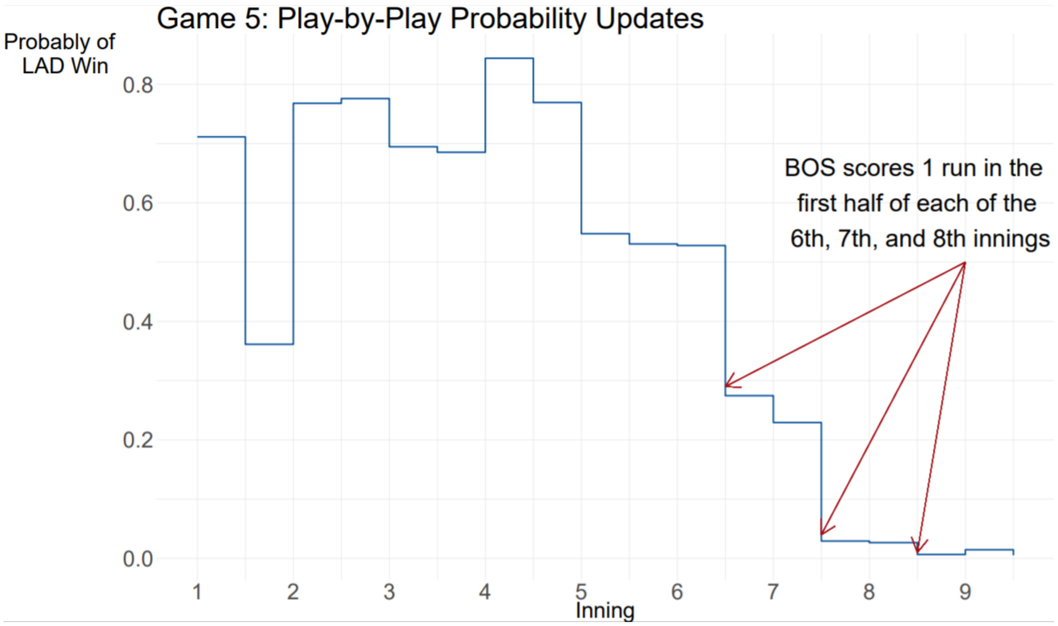

Figure 10: Play-by-play probability updates. After every half inning, simulation is re-run with updated game-level predictors.