Bayesian Baseball- World Series 2018

Blake Shurtz, Cal State East Bay

This poster presents and analyzes a model that predicts the outcome of baseball games with a focus on game 5 in the 2018 World Series between the Boston Red Sox (BOS) and the Los Angeles Dodgers (LAD).

The Model

$$rundiff_{i,j} \sim N(\mu_{i,j}, \sigma)$$

$$\mu_{i,j} = team_{i,j} + batting_{i,j} + pitching_{i,j} + fielding_{i}$$

$$batting_{i,j} = singles_{i,j} + doubles_{i,j} + triples_{i,j} + home\ runs_{i,j} + walks_{i,j}$$

$$pitching_{i,j} = strikes_{i,j} + strike\text{-}outs_{i,j} + balls_{i,j} + hits\ allowed_{i,j}$$

$$fielding_{i} = putouts_{i} + assists_{i} + errors_{i} + double\ plays_{i}$$

Two Factor Model 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.

Multi-Level Model 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.

Bayesian Framework 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.

World Series 2018: Posterior Predictions (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.

World Series 2018: Posterior Prediction Game 5 LAD hosts game 5 with a 1-1 record at home. **The model predicts a 65% probability that LAD wins game 5.** Nevertheless, LAD has only a 27% binomial probability of winning the necessary 3 remaining games.

Game 5: Parametric Analysis 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 (.14 pts), pitching (.12 pts) and fielding (.17 pts)**.

Game 5: Play-by-Play Probability Updates Due to the game-level nature of the response and the observability of the predictors, the game 5 outcome can be predicted in real-time. The model begins with the prior probability of a 65% chance of success for LAD. Both teams score early runs and the probability of a LAD win stays above p=50% due mainly to the higher batting average through inning 4. However, Boston scores 3 consecutive runs in innings 6, 7, and 8, securing the pennant with near certainty by inning 8.

Model Comparison 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, indicating a better fit for prediction. One explanation is the correlation between intercepts and slopes for the home team, suggesting home teams performance varies based on the visiting team.

Model Diagnostics 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.