## Bayesian Baseball- World Series 2018

### Blake Shurtz, blakeobeans@gmail.com

The poster describes and analyzes a multi-level regression model which attempts to predict the outcome of a baseball game with a focus on the 2018 World Series between the Boston Red Sox (BOS) and the LA Dodgers (LAD). The model uses data from the 2018 MLB season.

#### Two Factor Model

The model is a two-factor model. One factor is for the home team, and the other is for the away team. This provides three benefits: First, it adjusts for differences between home and away performance. Second, it allows all data to be organized in a single dataframe. Third, it allows the outcome variable to be defined consistently as the difference in scores between the home and away team.

#### Multi-Level Model

There are two levels to the multi-level model. The first is the game-level, which contains the response (difference in scores) as well as game-level predictors in batting and pitching. The second is the team-level, which contains overall team performance statistics in fielding.

The multi-level model is a varying-intercepts / varying slopes model. It allows unique differences in performance depending on the home and the away teams. For example, Boston's pitching may differ depending on whether LAD or Oakland A's (OAK) is batting.

One of the virtues of a multi-level model is that a given team's performance against another team is partially pooled between the observations for a given matchup and a team's overall performance. In other words, the model "remembers" past performance from other matchups against different teams. This is especially important in this context because, prior to the World Series, BOS and LAD never played a single match.

#### Bayesian Framework

The model is analyzed in a Bayesian framework with a likelihood function that is a maximum entropy Gaussian distribution.

Fielding variables are scaled with standard normal priors and all other variables have adaptive priors that are themselves a function of the data. The variance in factors and individual outcomes have a half-cauchy prior distribution. The prior for the correlation matrix between intercepts and slopes is a LKJ "onion method" correlation matrix distribution.

#### Model Diagnostics

Computational approximation of posterior distributions was executed using Hamiltonian Monte Carlo with a No-U-Turn sampler, executed in the software Stan. Stan has its basis in particle physics and simulates the movement of a particle along the log-posterior distribution. Above is a catepillar plot for sigma. There was only 1 divergent transition and all parameters have an Rhat of 1.00 or 1.01, indicating precise measurement.

#### Model Fit

Compared to a non-mlm and mlm for the intercept only, the varying intercept/slope model has higher mean WAIC, however, the differences overlap. Furthermore, as mscaled model fits has a lower in-sample deviance score. This is due to having more parameters in the model.

#### World Series 2018: Parametric Analysis

The likelihood function was re-parametized in order to compare BOS and LAD along offensive and defensive metrics.

#### World Series 2018: Prediction through Simulation

Prior to the World Series, Boston and LAD have never played. Nevertheless, we simulate games where Boston is the home team, LAD is the away team, and all of the game-level covariates are zero. (However, fielding/team-level covariates are included.) In other words, we can simulate the game up until the point that it starts.

After every game, the model is updated with data from the previous game. This allows us to make full use of Bayesian updating, as well as the partial pooling that makes multi-level models so useful.

Following game 4, the posterior distribution for game 5 is plotted whereby a score greater than 0 indicates a win for Boston. Although the posterior suggests that LAD has a higher chance of winning (approximately 60/40), the binomial probability of a necessary three-game winning streak is only 21%.

#### World Series 2018 Game 5: Real-Time Probability Updating

Due to the game-level nature of the response, Game 5 can be modeled in real-time. Furthermore, the model can assume the game starts with a 57% chance of success for LAD, which is consistent with all previous data.