A Hierarchical Bayesian Regression to Estimate the Posterior Probability that the Chicago Cubs Defeat the Cleveland Indians in the 2016 World Series

Blake Shurtz

August 31, 2018

Introduction

In 2016, the Chicago Cubs (CHC) won the World Series 4-3 against the Cleveland Indians (CLE). The results of the series are below.

Table 1: Runs Scored in 2016 Series

	1	2	3	4	5	6	7
Cubs	0	5	0	2	3	9	8
Indians	6	1	1	7	2	3	7

This clearly qualifies as a close series. The Indians won the first, third and fourth games, necessitating a 3-game winning streak by the Cubs. Furthermore, in the final game, the Cubs eked out a narrow victory by only one point. And that point was scored in the tenth inning!

It was not at all certain who would win the best out of seven games, especially considering that of the 161 games played by each team in the regular season, the Cubs and Indians did not play each other *once* prior to the World Series.

Goal

The goal of the paper is to describe and analyze a model that estimates the posterior probability that the Cubs defeat the Indians in the 2016 World Series.

Approach

A *multi-level model* was chosen for several reasons.

First, the hierarchical nature of a multi-level model fits the structure of a baseball season. For a given team, a baseball season is divided into multiple series, with each series typically involving 3 games against an opposing team. The "games within series" approach fits a multi-level model, which can include both game-level and series-level predictors. (For the model, the "games within series" approach is modified to "games within opposing teams.")

Second, the distribution of games against opposing teams is unbalanced. For example, the Cubs only played 20 of 29 opposing teams during the regular season. Of those 20 opposing teams, some teams were played for over a dozen games while other teams were played for only a single series. When the distribution of games is unbalanced, a multi-level model can pool among all of the games to prevent over-fitting when the sample size is small.

Finally, a multi-level model fits the problem in question. Given that the Cubs did not play against the Indians prior to the World Series, incorporating the overall performance of the Cubs relative to other teams using a multi-level model makes prediction against an unknown team possible.

The paper proceeds as follows. First, the data will be described. Second, two models will be analyzed. The first model, **the simple model**, is presented in order to calculate 95% Bayesian certainty intervals against the 20 teams that the Cubs played in the regular season. The second model, **the full model**, is presented in order to construct a posterior distribution of scores against the Cleveland Indians. Using this model, the probability of a Cubs victory is input into a binomial distribution to estimate the probability of winning the series. Finally, the full model is updated with data from each consecutive game in the World Series and the conditional probability of a Cubs victory in the World Series is estimated.

Data

The data contains each game played by the Cubs in the 2016 regular season. The Cubs played 161 games in the regular season, from April through early October.

Table 2: Sample 2016 Cubs Schedule with Predictors

Team	Win-Loss Ratio	Runs Scored	Runs Allowed	Score Difference	Pitcher	ERA
CIN	0.420	16	0	16	Arrieta	3.10
ATL	0.422	13	2	11	Lester	2.44
SEA	0.531	12	1	11	Lester	2.44
STL	0.531	13	2	11	Arrieta	3.10
PIT	0.484	12	2	10	Hendricks	2.13
LAA	0.457	9	0	9	Arrieta	3.10
CIN	0.420	9	0	9	Hammel	3.83
STL	0.531	12	3	9	Hammel	3.83

The outcome variable is the difference in scores between the Cubs and the opposing team. Positive values indicate a win and negative values indicate a loss. For example, if the score was 8-7 Cubs, then the outcome variable would be equal to 1.

The varying intercept/group-level predictor is a factor variable for the opposing team. (See model below.)

The data includes an individual-level predictor for the ERA (earned run average) of the starting pitcher for the Cubs. The ERA is the average number of runs allowed by the pitcher in a 9-inning game.

Finally, each opposing team's win-loss record (WL) at the close of the regular season is included as a fixed effect. Both the ERA and the WL were standardized according to their sample mean and variance.

Simple Model

The author chose to work with a Bayesian model written in R and with posterior sampling done in Stan, which uses Hamiltonian Monte Carlo (HMC) sampling in order to calculate log probability density functions for all variables in the model.

The simple model was used in order to construct 95% certainty intervals for the difference in scores for each opposing team (excluding teams that were not played by the Cubs in the regular season). The likelihood function is a varying-intercept regression which contains a different intercept for each opposing team.

$$\text{Score}_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \alpha_{\text{TEAM}[i]}$$
$$\alpha_{\text{TEAM}[i]} \sim \text{Normal}(0, 4)$$
$$\sigma \sim \text{HalfCauchy}(0, 2.5)$$

The assumptions of the model are simple enough. The difference in scores, $Score_i$, is normally distributed with mean μ_i and standard deviation σ.

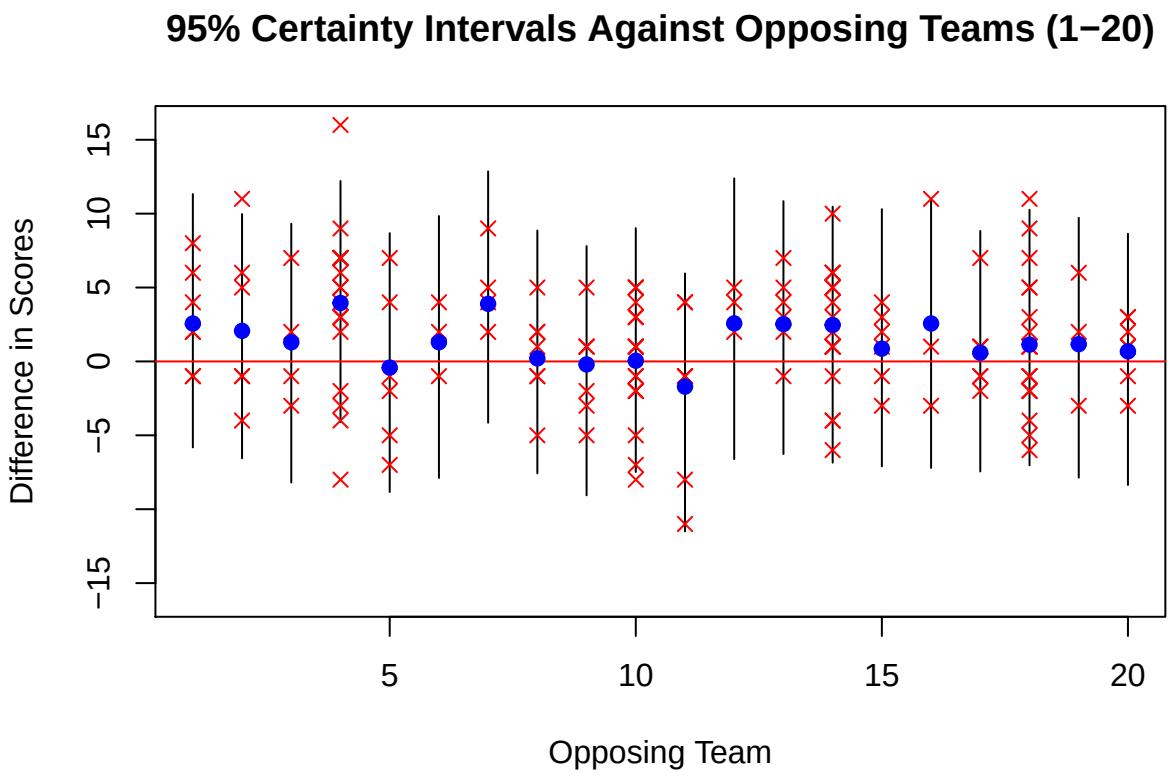
The prior distribution for the average difference in scores against an opposing team is normally distributed around 0 with a standard deviation of 4 points. In other words, prior to looking at the data, the model assumes that the Cubs have as many wins as losses against an unspecified opposing team.

The standard deviation in scores follows a half-Cauchy distribution with scale parameter of 2.5, which is a weakly-informing prior. (Gelman 2018)

Simple Model- Results

Four chains were run using 12,000 iterations each and a warm-up of 3,000 iterations. All variables converged to an R-hat of 1.00, which indicates that the posterior samples mixed well. (Trace plots are in the appendix.)

Once the model was built, 6000 games were simulated against each opposing team. Posterior distributions for the difference in scores were constructed and the mean and 95% certainty intervals were calculated for each opposing team.



The mean of each posterior distribution is a blue point. The mean is positive against most teams, indicating that the Cubs are likely to win against most of the teams given the data and prior assumptions. The interval against each team contains zero, suggesting that the either team can win a given game.

Full Model

The Cubs and Indians did not play a game prior to the World Series. Therefore, in order to predict the probability that the Cubs will win against the Indians in the World Series, an *a posteriori* simulation of a game against for an unknown team must be constructed. In order to do this, we use the full model.

$$
\begin{aligned}
\text{Score}_{\text{i}} &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \alpha_{\text{TEAM}[i]} + \beta_{\text{ERA}} + \beta_{\text{WL}} \\
\alpha_{\text{TEAM}[i]} &\sim \text{Normal}(\alpha_i, \sigma_i) \\
\alpha_i &\sim \text{Normal}(0, 1) \\
\sigma_i &\sim \text{HalfCauchy}(0, 2) \\
\beta_{\text{ERA}} &\sim \text{Normal}(0, 1) \\
\beta_{\text{WL}} &\sim \text{Normal}(0, 1) \\
\sigma &\sim \text{HalfCauchy}(0, 2.5)
\end{aligned}
$$

The varying-intercept term now has regularizing, adaptive priors α_i and α_{sigma}, which are known to improve model accuracy by allowing for the pooling of information among different teams. The author assumes that the prior distribution for the hyperparameters are standard normal. Standard normal priors are considered "generic weakly informative." (Gelman 2018)

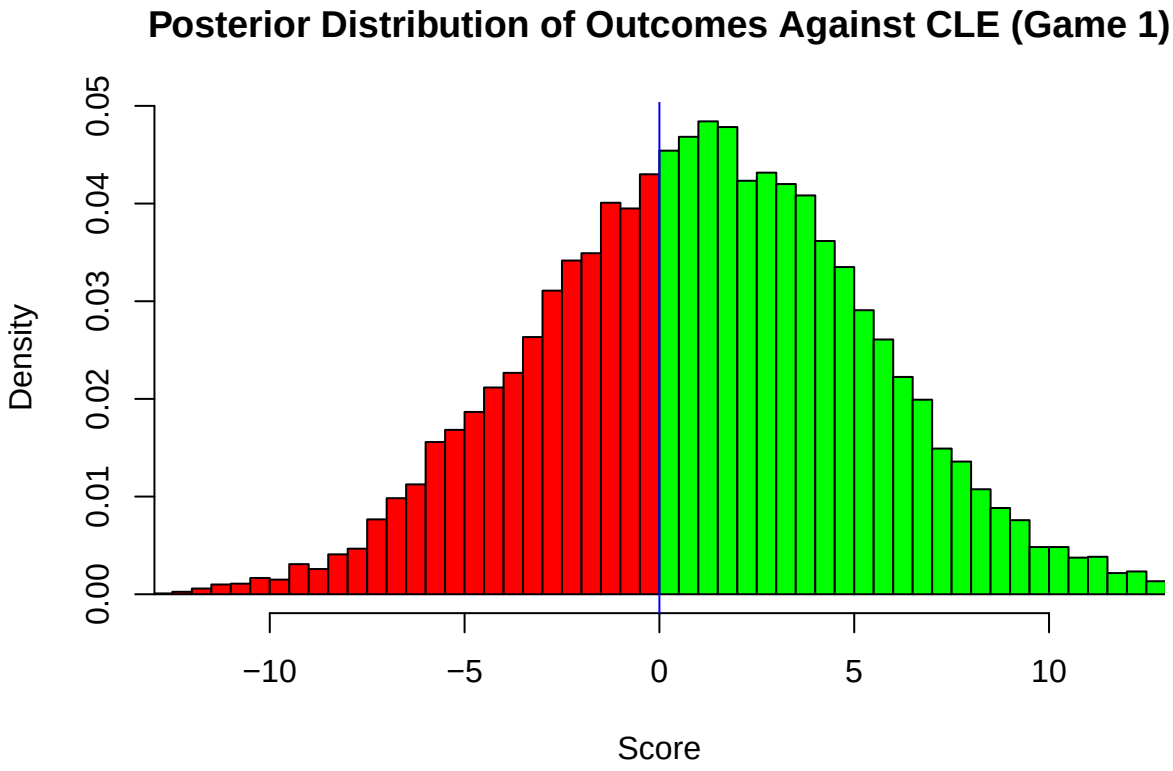
Also included in the regression function are the fixed effects for the Cubs' starting pitcher's ERA, β_{ERA}, and the opposing teams win-loss record, β_{WL}. These fixed effects have standard normal priors as well.

Full Model- Results

Four chains of 12,000 iterations were run with a warm-up of 3000 iterations. All variables converged with an R-hat of 1.00 or 1.01. A model comparison shows a lower WAIC for the full model. (See appendix for model comparison and table of coefficients.)

In order to construct a posterior distribution for the difference in scores between the Cubs and the Indians, new data is fed into the likelihood function. This data includes the normalized ERA for the Cubs' starting pitcher for game 1 (Jon Lester, normalized ERA of -.63) and the normalized win-loss record for the Indians (1.29).

12,000 "Game 1" games against the Indians were simulated and the mean and a 95% certainty interval for the posterior distribution were calculated.



Like for the other opposing teams, the model indicates that either team has a chance at victory. However, given pooling of the multi-level model, as well as the value of the predictors, the model favors a Cubs win (the green area above) with a probability of 60%.

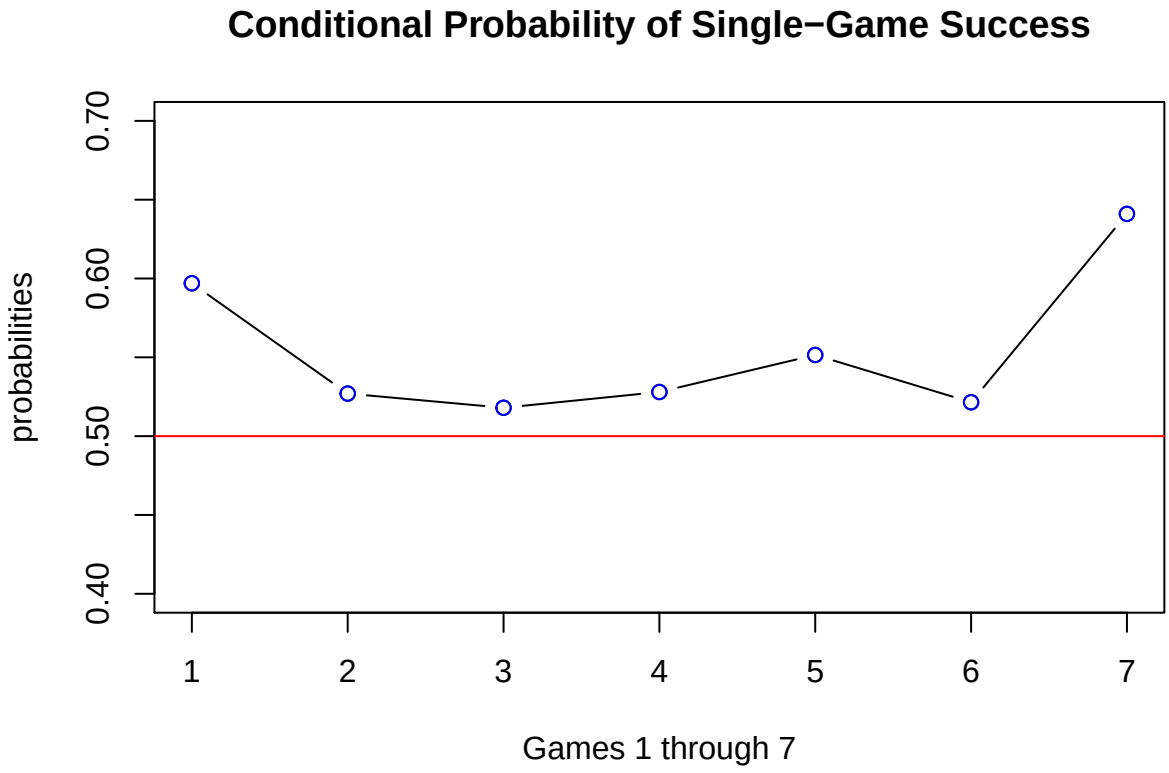
However, the World Series is a set of repeated matches against the two teams, best out of seven. Therefore, the probability that the Cubs win follows a binomial distribution, with a probability of success of 60% and a required number of successes of four or more.

By plugging these variables into the binomial formula, I find that **the probability that the Cubs win 4 (or more) out of 7 games is 70%.**

Conditional Model

If one is to assume a Bayesian worldview in which the probability of a Cubs victory is a random variable that can be updated with new data as the World Series progresses, then a further analysis using the model in this paper can be performed. This analysis will be of two different flavors.

First, the model can be updated with new data containing the outcome and predictors for the previous games against the Indians. Then, the model (the full model) can be re-run and the probability of a success within a single game can trend up or down based on the new data. The table below plots the probabilities of a Cubs victory for a given game after factoring in the previous games in the series.

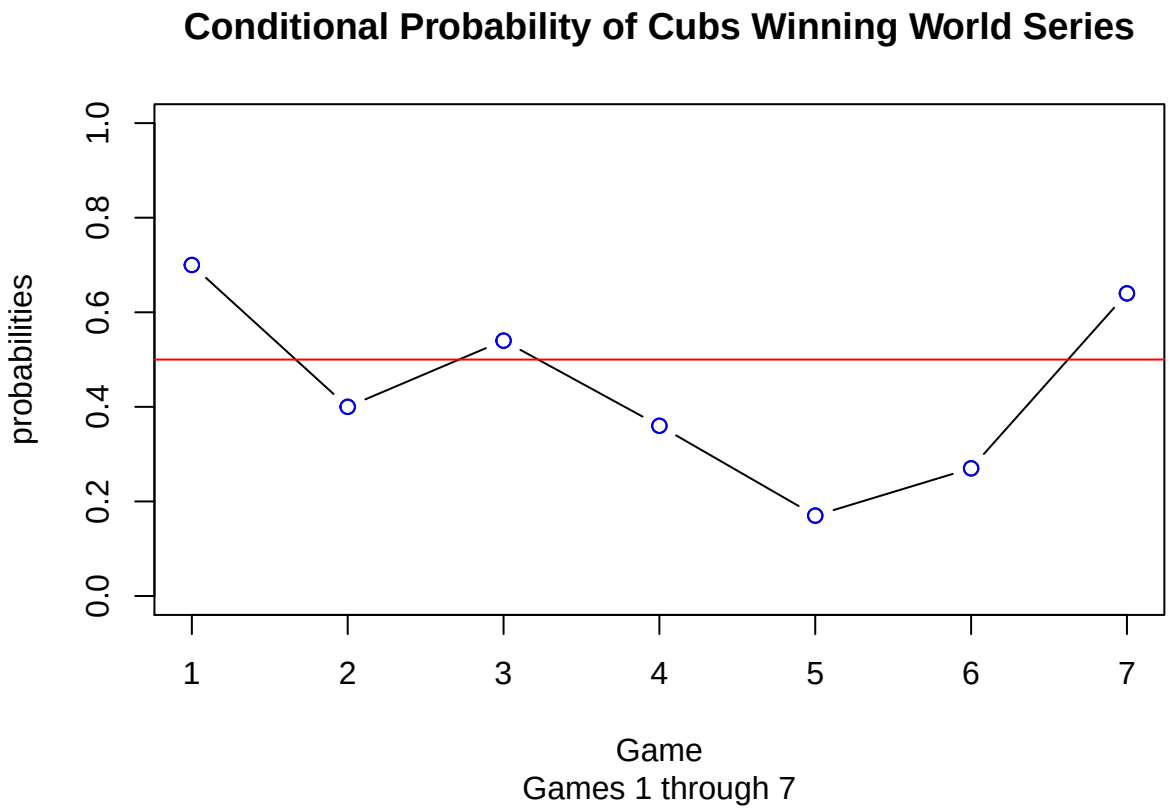


As the graph indicates, the probability of a single-game success declines from about 60% to about 52% during the first few games when the Cubs perform poorly. Even though the Cubs won game 5, the probability of winning in game 6 dips slightly. This may be due to the fact that a below average ERA pitcher (John Lester) is pitching in game 6.

Conditional Model (cont.)

Second, one can update the conditional binomial probability of winning the series given the updated single-game probability of success as well as the the number of remaining games in the series.

For example, the 2016 World Series got to the point where the Cubs needed to win the last three games in order to succeed. Is it really the case that the Cubs are still the expected winner when they are in that position?



This model requires careful interpretation. It is true, according to the model, that there is a 70% chance of victory prior to game 1. It is also true, according to the model, that there is only a 20% chance after game 4 and prior to game 5. These contrasting perspectives highlight the importance of understanding different interpretations of a model, in particular a self-updating Bayesian model.

It is noteworthy but expected that the conditional probability of winning the final game of the series under the binomial model converges to the probability of winning a single game under the single-game model.

Appendix

Shiny App

For trace plots, posterior distributions for all variables and more, check out the ShinyStan app at https://blakeobeans.shinyapps.io/BayesianBaseball

Table of Coefficients for Full Model (With Interpretation)

Estimates for individual parameters are excluded.

Table 3: Estimates for Prediction Coefficients

variable	mean	sd	2.5%	97.5%	n_eff	Rhat
sigma	4.19	0.24	3.75	4.68	2333.09	1.00
a	1.36	1.01	-0.68	3.33	1374.07	1.00
ai	0.18	0.94	-1.64	2.08	1556.57	1.00
as	0.73	0.43	0.11	1.70	456.11	1.01
b	-0.55	0.31	-1.15	0.07	6058.81	1.00
c	-0.64	0.36	-1.33	0.08	2878.77	1.00

The group-level standard deviation has a mean of 4.19, suggesting that there is a lot of unexplained variance in scores, no matter who the opposing team is. It is this variance that contributes to the "any team can win" interpretation of the model.

The mean score for the Cubs, without factoring in the varying intercept term, is 1.36. This is expected given that the Cubs win more frequently than they lose.

The mean for the starting pitcher's ERA is b=-0.55, suggesting that an increase in the ERA by one standard deviation reduces the Cubs expected lead by half a point.

The mean for the opposing team's WL record is c=-0.64, suggesting that a one standard deviation increase in the opposing team's WL record reduces the Cubs expected lead by 2/3 of a point the difference in scores declines by 1/2.

Both of these terms narrowly include 0 at a 95% probability interval and have logical *a priori* interpretations. Taken together, these two terms can reduce the expected difference in scores almost entirely to 0.

Comparison of Full and Simple Model

Table 4: Model Comparison

	WAIC	pWAIC	dWAIC	weight	SE	dSE
fullmodel	925.44	7.17	0.00	1	19.70	NA
simplemodel	938.90	15.64	13.46	0	18.83	6.83

The full model has a lower WAIC, indicating a better estimated out-of-sample deviance. (Rethinking 2015) The full model also has fewer effective parameters, despite the addition of fixed effects. This can be attributed to the use of adaptive priors.

Works Cited

I'd like to thank Richard McElreath and his wonderful book, R package and series of Youtube lectures, "Statistical Rethinking." I consider this the most valuable stepping stone from a basic knowledge of regression into full-blown Bayesian statistics with Stan.

http://xcelab.net/rm/statistical-rethinking/

Gelman, Andrew. "Prior Choice Recommendations." Last Updated 2018. https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations.