**Counting Balls in Nets: A Prediction of German Goals in the 2018 World Cup**

Final Project for Bayesian Method for Data Science (DATS 6450)

Project Group 4

Michael Siebel

Data Science, George Washington University

**Introduction**

The 2010 World Cup was one of the lowest goal scoring World Cups in history. At the time, journalists were predicting that low scoring World Cups would become the norm as defenses professionalize in tactics and technique. Then, to most spectators’ joy, the 2014 World Cup proved itself one of the highest goal scoring World Cups. In both previous World Cups, Germany scored more goals than any other team—with the last World Cup securing them as world champions. The 2018 World Cup is eleven days in, which offers me the opportunity to see if my goal predicting algorithm is having any success.

**Background**

Sports has had a short but enthusiastic history of applying pioneering analytical methods to predict match outcomes, player performances, and other athletic statistics. In 2003, Michal Lewis published *Moneyball: The Art of Winning an Unfair Game*, which brought the idea of using statistical techniques to achieve richer insights than coaches, scouts, journalists, and pundits where able to muster from individual experience and intuition. Simon Kuper and Stefan Szymanski brought this practice to soccer with their 2009 book S*occernomics*.

Despite having similar goals in mind, the authors of these two books were forced to use very different approaches due to the data abundance (in the case of baseball) or scarcity (in the case of soccer). Baseball is a mostly static game in which most action leads to an outcome such as a hit, RBI, or out can be counted and recorded. In contrast, soccer is a mostly fluid game in which most action does not directly lead to an outcome that can be counted and recorded. For example, a dribble is difficult to capture as a statistic and off-the-ball movement—vital in gameplay—can only be capture with heatmaps and not point estimates.

Given these constraints, new research is emerging on soccer data (see Jürisoo M., 2018), much of which use Bayesian analysis, which can take advantage of expert advice from coaches, scouts, journalists, and pundits, and incorporate it into the latest data science methods.

**Data**

Török (2018) published a dataset that includes FIFA rankings from 1993 to 2018, International Soccer matches from 1872 to 2018, and a list of upcoming FIFA World Cup 2018 match fixtures, webscrapping this information from Wikipedia, fifa.com, rsssf.com, and individual football associations' websites.

The main dataset includes international football matches, as the level of analysis, with home and away goals and location information for analysis. Supplementary datasets include FIFA rankings—a global ranking of best teams by the body overseeing international soccer—and group stage games of the 2018 World Cup.

After a long process of data cleaning and preparation, the final data was created to include a count of goals scored by Germany from 1950 to 2018.

**Methodology**

This research uses a Generalized Linear Model to predict a Poisson distribution (i.e., a Poisson regression) to predict the goals scored by Germany from 1950 to 2018. The resulting prediction will be used to generate predicted probabilities of goals scored by Germany in their game against Mexico (which occurred recently on June 17, 2018) and against Sweden (which occurred on June 23, 2018). These match results are not in the dataset and will not be used in generating the predicted probabilities.

Soccer goals follow Poisson distributions and/or negative binomial distributions. They are never negative, crowd early in the count such as around 0 or 1, and can extend, often in rare cases, to higher values such as 7. As such, a Poisson regression can be specified to accurately predict common, low-scoring games, while often struggling to accurately predict rarer, high-scoring games—a problem that is a component of overdispersion.

This research will first predict goals scored in these two matches using traditional, frequentist techniques. Then, this research will fit a Bayesian model using median goals scored by Germany across all teams between 1950 and 2018 on a set of standard deviation parameters as its prior. A Markov Chain Monte Carlo (MCMC) algorithm allows the estimation of regression parameters by their probability/frequency distribution, starting from the initial prior. Code will be re-formatted and re-specified from Kruschke (2015), borrowing from the R syntax file *Jags-Ycount-Xnom2fac-MpoissonExp*. In particular, it will be change the model design and alter chart output to work with a different type of parameters in order to predict count probabilities instead of predicted proportions. Finally, this research will evaluate which approach led to the closest predictions.

**Experimental Results and Analysis**

Win and lose ratios between Germany and its opponents were generated as leader predictor variables. These ratios were calculated as mean wins and mean loses for every two World Cups campaigns, an eight-year period. In other words, all matches played between Mexico and Germany between 2007 to 2014 were arithmetically averaged for that period’s win ratio. Eight-year periods were needed to measure wins and loses between each opponent as international soccer does not occur on a frequent basis. The most recent period was only four years, from 2015 to now, in order to keep the current campaign, and therefore current team form, separate from other periods. Win ratios were used in models to predict goals scored by Germany.

Other explanatory variables include Germany’s opponent in each match, a factor variable consisting of 86 teams, and whether the match was a “friendly” or a “competitive.” The prior variable is important in that Germany tend to play well against certain teams, e.g. England, and poorly against other teams, e.g. Italy. The latter variable is important, because performance differs by type of match. In particular, more goals tend to be scored in friendlies as teams are more relaxed and more experimental in their approach to the match.

Figure 1 shows Germany’s FIFA ranking from 1993 to 2018 compared to Brazil, often viewed as history’s greatest soccer team, and England, a mid-level team. A ranking of one is the highest ranking a team can hold. The figure shows that with the exception of around 2002, Germany have long been a high-ranking team, which should be expected to score lots of goals and concede few goals.

**Figure 1**

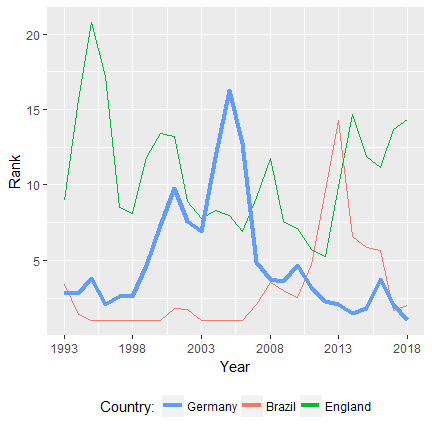


Figure 2 shows Germany’s goal differential since 1950. It averages 1.14 goals per game, although there are several outliers above 5 goals. Because the average is low but not too heavily concentrated around zero, Germany should be a good candidate for counting goals.

**Figure 2**

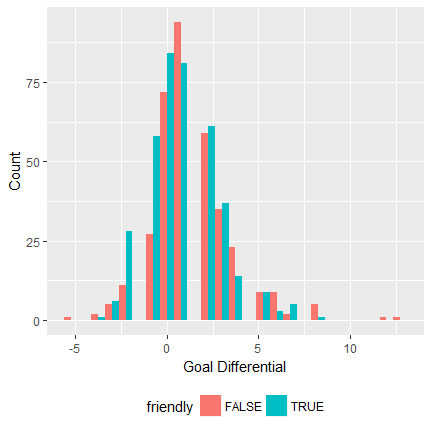
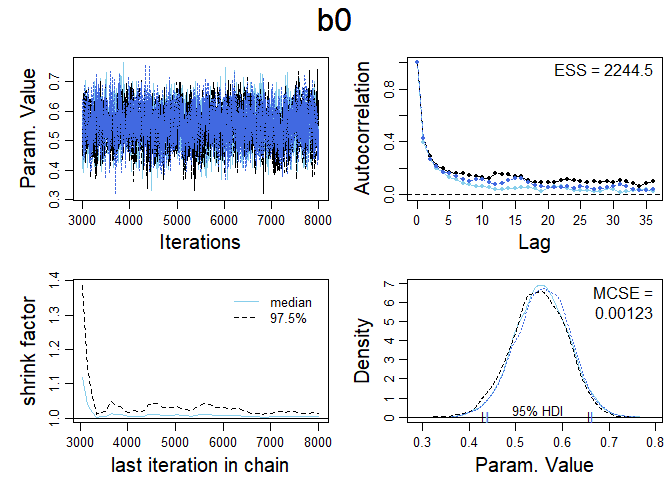


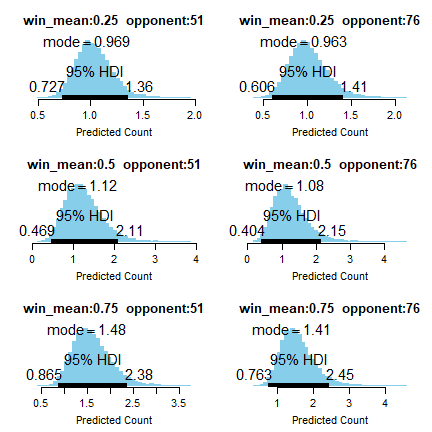
Figure 3 shows the MCMC diagnostics of the intercept parameter. The parameter value during each iteration shows no major divergent transitions. The EES value is rather low at 2,244.5, but the chain lags approach low autocorrelation rather quickly. The shrinkage factors approach 1 very quickly, indicating assurance of the representativeness of the samples from the posterior distribution. Other parameters followed similar patterns.

**Figure 3**



Finally, figure 4 displays the predicted probabilities of goals Germany would be expected to score against Mexico (opponent “51”) and Sweden (opponent “76”). Most recently, Germany had a win ratio of around 0.5 with both countries leading up to the tournament. However, all but one match with Mexico was a friendly. The predicted probabilities below are specified as competitive matches in the Poisson regression at various win ratios for comparison. Germany had predicted probability of 1.12 and 1.08 goals against Mexico and Sweden, respectively.

**Figure 4**



The frequentist model more aggressively predicted a probability of 2.12 and 1.95 goals against Mexico and Sweden, respectively. Overall, the Bayesian model has high credibility intervals (0.47-2.11 and 0.40-2.15) and the frequentist model featured high confidence intervals (1.69-2.56 and 1.64-2.26). One cannot compare credibility intervals to confidences intervals as the prior is probabilistic while the other refers to sampling methods. However, it is important to note that both would likely be considered large in their approaches given this data’s estimation.

**Conclusion**

On the actual match days, Germany scored no goals against Mexico and 2 against Sweden. While the frequentist approach was closer in the Sweden match, the Bayesian approach was close in the Mexican match. Arguably, the Bayesian approach was the safer model, as Germany as of now has not been scoring as freely as it had in previous tournaments. Because the Bayesian model made use of historic goals as its prior, perhaps it was less sensitive to recent performances by Germany. The frequentist model may have identified the high win ratios and high goal totals in the past eight years with Germany and therefore was more sensitive to recent performances by Germany.

**References**

Kruschke, J.. Doing Bayesian data analysis, 2nd Edition: A Tutorial with R, JAGS and Stan. 2015 Academic Press/Elsevier. ISBN:9780124058880

Jürisoo, M. (2018).  International football results from 1872 to 2018 [Data file and code book]. Kaggle. Retrieved from https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017

Török, A. (2018). Predicting the winner of the 2018 FIFA World Cup [Data file and code book]. Kaggle. Retrieved from https://www.kaggle.com/agostontorok/soccer-world-cup-2018-winner