NCAA March Madness basketball game prediction

Bin bin Wu, Le Yuan and Zheng ting Zhu

Introduction

Our team looked at a kaggle's competition called "Google Cloud & NCAA® ML Competition 2019-Men's". We are supposed to predict the outcomes of March Madness during this year's NCAA Division I Men's Basketball Championships. Our motivation is obvious that although people have tried and failed to forecast the perfect outcomes of March Madness during each year's NCAADivision I Men's Championships bracket and this problem is deemed too complicated, it can help us strengthen statistics, data modeling and cloud technology. We are provided with a large amount of historical data about college basketball games and teams, going back many years.

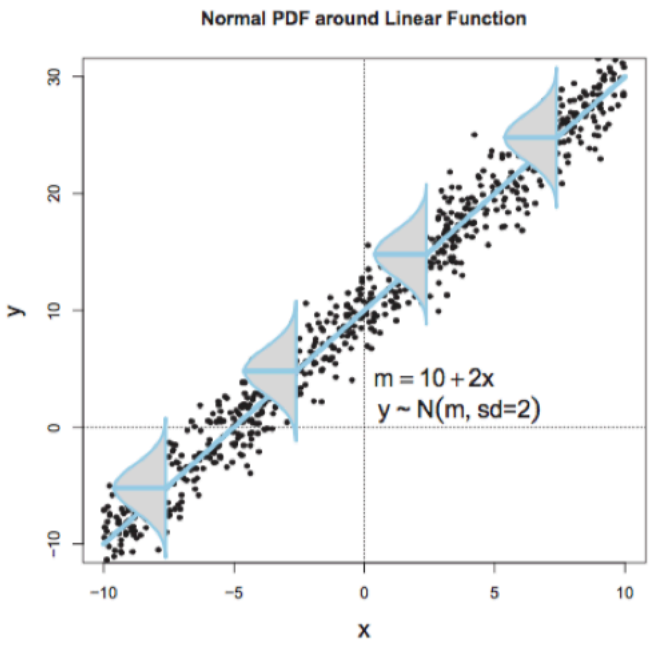
One of the biggest problems with predicting the outcomes are upsets which low seeded teams beating high seeded teams. We want to use Bayesian analysis and historical data to check how the seeding impact winning. We decided to extract data into 6 variables: TeamID, Seed, Season, WinR, LoseR and TotalG and use hierarchical model as well as generalized linear model to predict team with which seeds is more likely to win or lose in a game. Then I will explain dataset and models in details.

Chart below shows basic information of the extracted dataset.

1311 observations	6 variables					
TeamID	Int	1102	1102	1103	1103	…
Seed	num	11	13	12	13	…
Season	Int	2004	2006	2013	2009	…
WinR	num	0	0	0	3	…
LoseR	num	1	1	1	2	…
TotalG	num	1	1	1	5	…

Model— hierarchical model:is a model in which lower levels are sorted under a hierarchy of successively higher-level units. Data is grouped into clusters at one or more levels, and the influence of the clusters on the data points contained in them is taken account in any statistical analysis. I will specify hierarchical model we used from the bottom to the top. Y follows binomial distribution with parameters mu and N[i]; Mu follows beta distribution with parameters omega and kappa; Omega follows logistic distribution with parameter a0 and a[j]; a[j] follows a normal distribution.

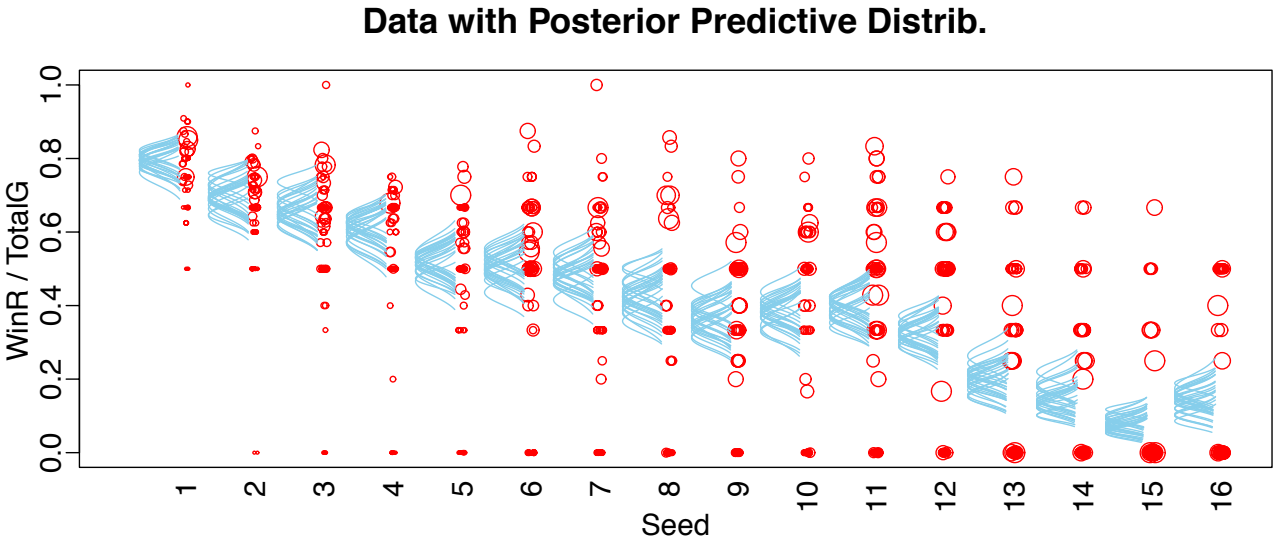
Model— generalized linear model: the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error



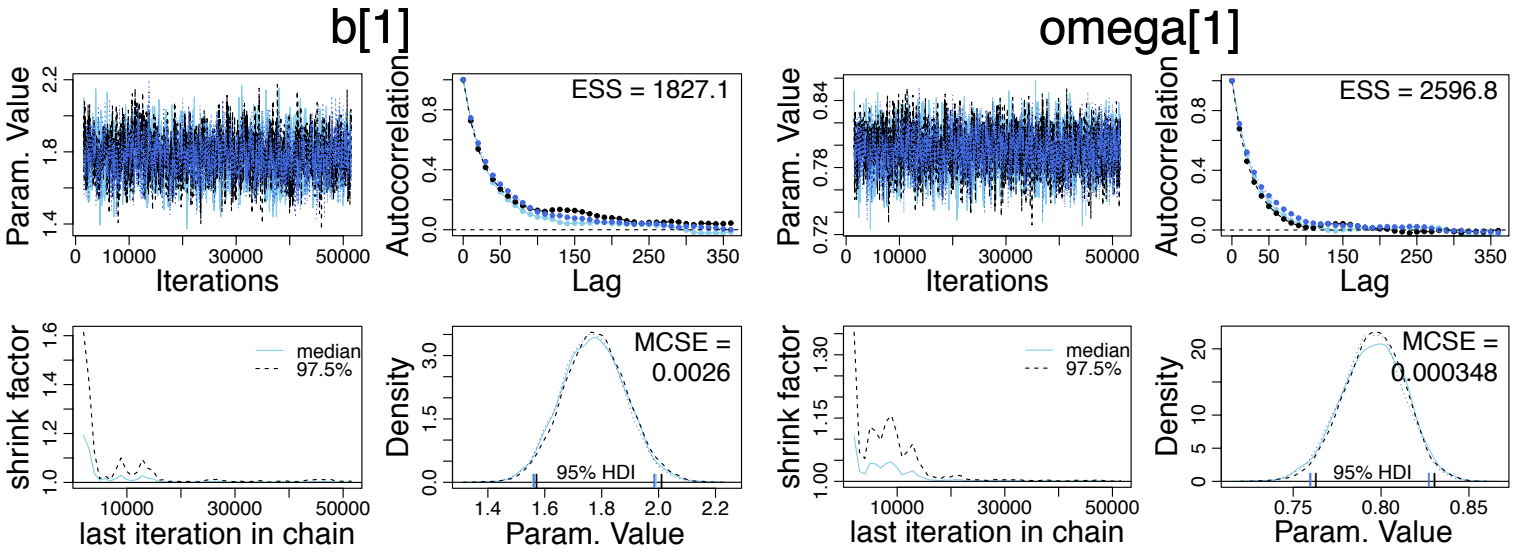
distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a *link function* and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. The formal definition of the GLM is μ = f (lin(x), [parameters]); y ~ pdf(μ, [parameters]). The graph below can show what GLM looks like and unlike linear regression, GLM will generate umbrellas to cover the points.

Empirical results

After running the model, we have got lots of useful results and challenges to tackle. I will start with results. The graph below exhibits scatter plot with 16 seeds and the value of WinR/TotalG for each seed. Also, this graph shows the model we generated to cover each

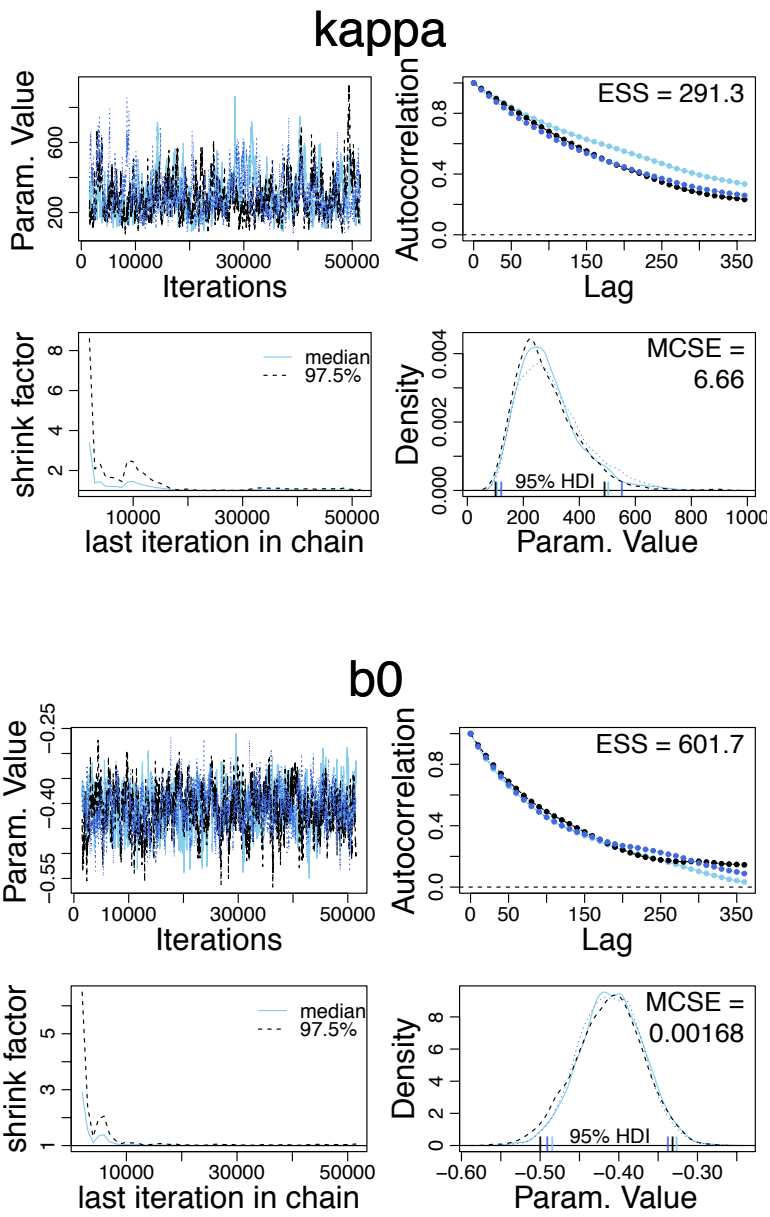


seed's winning percentage. It is obvious to note that the general trend of seed from 1 to 16, it follows a negative correlation which means high seeds does help to win more games. However, we will find many surprises if we look at the posterior of each seed generated by MCMC. First, the winning percentage is around 50% and the difference between seed 5, seed 6 and seed 7 is not obvious which means it is hard to make a conclusion when it comes to which team will win. During the first round of March Madness, picking seed 5 or seed 6 or seed 7 to win is not as safe as people may think. Second, from seed 9, seed 10 and seed 11, the trend follows a positive correlation, which means a team with seed 11 is more likely to win games compare seed 9. Finally, the difference between seed 13 and seed 14 is quite subtle, and seed 15 actually has a worse winning percentage than seed 16. The model give us a clearer perspective to see the tournament results. For example, when it comes to team A with seed 6 and team B with seed 11, team A is more likely to win team B, but the upset would highly likely happen. Besides the posterior distribution graph, we got 4 graphs regarding parameters beta1, omega, beta0 and kappa.



Two graphs above show beat1 and omega. Not only the representativeness but also the accuracy and efficiency, these two parameters look pretty good. As for the panel on the left top, beat1 convergence quite well since for each iterations, it moves at the same pack and overlap each other. As for the panel on the right top, lines also converge very well and it is worth noting that the effective sample size is as large as 1827.1. As for the panel on the left, lines overlap well and the value of shrink factor is 1 in the 5000 iteration, which means difference between chains is quite equal to the variance within chains. The last panel on the

right bottom shows density overlap well and the value of MCSE is pretty low. When it comes to the graph of omega, four panels are quite the same as that of beta1.



There is still challenge. In other words, parameters kappa and beta0 are not representative and accurate as beta1 and omega. As for panel on the left top, trace plot of kappa does not converge well, since it is easy to see the gap between each iterations and lines do not move at the same pace. As for the panel on the right top, effective sample size is as low as 291.3 and the value of autocorrelation is larger than 0, which is a bad indicator since it means the value of kappa in each iteration is correlated. And these two problems are the same for beta0. The trace plot for beta0 does not overlap well(the left top panel) and the effectiveness sample size is as low as 607.1 and the value of autocorrelation is larger than 0. In order to improve the representativeness and accuracy of kappa and beta0, our team should increase the amount of the dataset by increasing the seeds to 64 and adding more historical data about college basketball games and teams.

Conclusion

Although corner cases exit and there are deficiency for beta0 and kappa, this model can predict team with which seeds is more likely to win or lose in a game and beta1 and omega is representative, accurate and efficient.

In the future, we will make full use of dataset provided in this competition and add more historical data about college basketball games and teams. After increasing the amount of the dataset, we could also modify the model to get better prior for kappa and b0 since the current model's prior has no preference and generic vague. After increasing the amount of dataset and using better prior, the result of team with which seeds is more likely to win or lose in a game can be accurately predicted and beta0 and kappa can become representative, accurate and efficient.

References

https://www.kaggle.com/c/womens-machine-learning-competition-2019/overview/description
https://blackboard.gwu.edu/bbcswebdav/pid-9395308-dt-content-rid-62825350_2/courses/45745_201901/Chap_17_Metric_Predicted_Variable_with_One_Metric_Predictor.pdf
https://blackboard.gwu.edu/bbcswebdav/pid-9301387-dt-content-rid-60980794_2/courses/45745_201901/Markov_Chain_Monte_Carlo.pdf
Kruschke, J. K. (2014). Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition. Academic Press / Elsevier