Cover Letter:

In Paper 3, Version 2, according to the comments from GSI and classmates, I improve my report by:

Firstly, I add explainations on specific variables will be used in this analysis.

Seconddly, I explain shortly on how each model I build works and how penalty model works in this study. Besides, in order to keep readabe for people who do not have a lot of background knowledge, I delect jargons and math elements.

Lastly, from the discussion part, discuss the “advantages of Bradley-Terry modeling over less sophisticated alternative measures such aswin-loss percentages or sum scores”, as required by the rubric.

Xiangsen Yang

Professor Hansen

Stats 485

11 April 2022

Discovering NCAA data by Bradley-Terry models

INTRODUCTION

Every year at the beginning of the spring season, one of the most exciting competitions in the United States–"March Madness"–the National Collegiate Athletic Association (NCAA) men's basketball tournament always ignites sports fans' blood after long cold wintertime. Recently, the 2021-22 men's college basketball season has finally been done here. During this season, on April 4, 2022, Kansas overcame 15 points halftime deficit against North Carolina to win the fourth national championship in its history. Beyond exciting games and the audience's passion, March Madness leaves a bitch of data, which is very valuable for us to discover more.

In this article, the main goal is to use the preferred statistical model to predict the rank of NCAA men's basketball teams. the original data for Division 1 includes the regular season and through the first 4 rounds of the postseason, up to but not including Final Four games. We try to apply the Bradley-Terry models to fit the data from ranked NCAA men's basketball teams for the 2021-22 season. After that, we need to use appropriate statistical calculations to summarize which model our study preferred. At the end of this paper, based on which model is chosen, we report the top 10 and bottom 5 teams on the rank of coefficients (larger coefficients correspond to better teams and vice versa). We also will try to answer a potential question: Do coefficients really reflect the information about the team's power?

METHODS

The data in this essay is from NCAA men's basketball teams for the 2021-22 season. The original data set includes 358 men's basketball teams' win-loss data for Division 1 contests during the regular season from the first 4 rounds of the postseason to the date 3/27/2022. Although this dataset does not include the whole season, it is still valuable for us to use this win-loss data to build a robust model. The data records 5345 games on the 2021-22 season. The available information includes the date of each game, the winning team, the losing team, and whether the winner is the home team. The Bradley-Terry model is used in this research to predict the outcomes of basketball games.

In this essay, the statistical model we used is the Bradley-Terry model, which allows us to predict the outcomes of basketball games. According to this model, we need to choose one team as the reference team, the team in our study is Villanova. By using Bradley-Terry model, we will get one coefficient which is a strength index, and higher coefficient means higher probability of winning the reference team. In our study, we randomly choose one team as the reference team: Villanova.

The first model is a standard Bradley-Terry model(plain-vanilla Bradley-Terry model) that is fitted by using maximum likelihood. The second model is a Bayesian generalized linear regression model, which is a penalized version of logistic regression. Penalized logistic regression means a penalty to the model for including "too many" variables (we have over 300 variables in this study). Penalization is able to reduce the influence of non-contributing variables and therefore improve model accuracy in some situations (The result is that the coefficients that were small in the "plain-vanilla" model will be 0). Villanova serves as the reference team in

both models, with each team's coefficient reflecting its relative strength in comparison to Villanova: a positive coefficient indicates that the team is better than Villanova, while a negative coefficient indicates that the team is worse.

However, these two models do not account for home advantage or the shifting strength of one team at different periods of the season. Therefore, the third and fourth models are modified versions of the first and second models, respectively, to account for home advantage and shifting team strengths. We consider Christmas break could help teams to change their strength. We will create a Christmas break as variable, which equals 1 if the game took place after Christmas break, otherwise equals 0. By including the indicator's interaction with the team variables, our models allow us to discover different coefficients (strength) for each team before and after the Christmas break.

In order to test the performance of each of our models, we choose two methods, mean squared error (MSE) and k-fold cross-validation. MSE allows us to drive the accuracy of our model by comparing the predicted values with the actual values in our dataset. Normally, we want our model to have a very low value on MSE. Moreover, we can perform k-fold cross-validation, in our study is 5-fold cross-validation, which is a method for comparing the performance of our model with existing data. This procedure is commonly used to avoid problems like overfitting or selection bias and to give an insight into how the model will generalize to an independent dataset (OARC Stats, 2016). For this method, we split the data randomly into 5 groups and use each group as a test dataset one at a time. The prediction model is trained with the training set, and then the verification set is used to test the trained model, which can be used as the performance index of the model.

On the next step, based on what model we choose, we present top 10 rank and bottom 5 rank teams based on their coefficient. For proving there is the uncertainty of our model, we use one-tailed hypothesis test to check the null hypothesis: the true coefficient of North Carolina and Duke are actually lower than Kansas's. We will set a 5% level confidence interval for one-tailed hypothesis test. If zero is included in confidence interval, we should reject the null hypothesis and conclude that we don't have enough evidence to claim that the strength of North Carolina or Duke is less than the strength of Kansas, and vice versa.

RESULTS

As a result, we chose the penalized model based on host advantage and season, because In-sample MSE cross-validation shows that this model has highest accuracy in predicting game outcomes(see Figure 1).

Model	In-Sample MSE	5-fold cross validation
Plain-Vanilla Bradley-Terry model	5.12	6.27
Penalized Plain-Vanilla Bradley-Terry model	3.10	4.15
Model accounting for game location and season	4.99	6.12
Penalized Model accounting for game location and season	2.97	4.00

Table 1: models' assessment (Villanova as reference team)

Based on the model we preferred, we are able to generate the rank of best NCAA men's basketball team for season 2021-2022. Fortunately, for this year, the final 4 are North Carolina, Kansas, Duke, and Villanova, three of these teams show up on my top 10 rank list. However, this does not mean the model we apply in this study is accurate. If one team has a bigger coefficient, this numeral value could not guarantee this team must be more powerful than teams with lower coefficients.

University	Top 10 Rank	Coefficient
Kansas	1	1.23
Gonzaga	2	1.16
Arizona	3	1.13
Baylor	4	1.01
Texas.Tech	5	0.95
Auburn	6	0.98
Duke	7	0.93
Tennessee	8	0.77
UCLA	9	0.76
Kentucky	10	0.69

Table 2: preferred model rank the teams in the top 10 for the 2021-2022 season (for coefficients, Villanova as reference team)

University	Bottom 5 Rank	Coefficient
Delaware.St.	1	-10.50
Lamar	2	-10.00
IUPUI	3	-7.86
Miss.Valley.St.	4	-6.51
Green.Bay	5	0.95

Table 3: preferred model rank the teams in the bottom 5 for the 2021-2022 season(for coefficients, Villanova as reference team)

Test	95% family-wise confidence level
Kansas - North.Carolina	$(-\infty,1.92)$
Kansas - Duke	$(-\infty,1.49)$

Table 4: the result of hypothesis testing

As both confidence intervals contain 0, we should reject the null hypothesis and claim that we do not have sufficient evidence to support that North Carolina and Duke is weak than Kansas. In addition, because of the hypothesis test results, we could realize that the real outcomesare inconsistent with the model results. Therefore, the uncertainty of our estimation model actually exists.

DISCUSSION

As previously stated, the Bradley-Terry model is a subset of logistic regression. This is a model that may be expanded upon by adding new features. Using Bayesian approaches, we'll be able to improve the model. By fitting the existing data, we now have a preliminary understanding of the level of each in NCAA. Our initial look at defining and quantifying competition comes from the estimated levels of the teams. We can further mine the characteristics of data and research the aspects that affect evaluation results based on the current work to get a better prediction effect. Indeed, comparing less sophisticated alternative measures such as win-loss percentages or sum scores, Bradley-Terry model could just base on one reference team for directly generating other teams' strength easily. Moreover, Bradley-Terry model could show the team strength's changes during the different period of season. Therefore, Bradley-Terry model is able to discover team' strengths from higher dimension, rather than just focusing on win-loss.

In addition, another important issue we also need to address in further research is discovering whether the choose a "strong" reference team will cause bias in the calculation of coefficient. First, we imagine a situation that if we coincidentally choose a strong basketball team i as a reference team, which means this strong team has ability to win almost matches.

Unfortunately, there also is another strong team j. If both team i and j beats other teams, the win-loss between team i and team j is becoming important when we select team i or j as a reference team at the same time. Anyways, if we just use win-loss data, rather than score data, the rank we have generated might has some bias. Back to our research question: Do coefficients really reflect the information about team's power? The answer is not. When we want to interpret the coefficients on each team based on B.T model, coefficients could not absolutely prove which team is stronger.

Works Cited

"FAQ: How are the likelihood ratio wald and LaGrange multiplier score tests different and/or similar." *OARC Stats*, 2016, https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/.