International Core Journal of Engineering Volume 7 Issue 1, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0049

Prediction of Women's College Basketball Team's Strengths and "Home Court Advantage" by Linear Model

Zheng Huo[1], Jingtian Xiao[2], Zijin Ye[3], Zihan Huang[4]

[1]Beijing Haidian Foreign Language Shiyan School, Beijing, Beijing 100097, China;

[2]Shenzhen Experimental School, Shenzhen, Guangdong 518055, China;

[3]Guangzhou Foreign Language School, Guangzhou, Guangdong 511455, China;

[4]Shenzhen Middle School, Shenzhen, Guangdong 518050, China.

Abstract

With the rise of popularity of basketball game, game result prediction become more popular and noticing as well. To predict the result of each Women College Basketball game, the prediction model is constructed. The prediction model is a linear model used to predict the strength of each basketball team played in Women's College Basketball games and the "home court advantage". This paper conducts the work in the R Language and Environment for Statistical Computing, and utilizes the data from the National Collegiate Athletic Association (NCAA) about these games in the past five seasons to fit the model and examine the accuracy of the model prediction.

Keywords

Basketball Game; Prediction; Linear Model; R Language.

1. Introduction

As one of the most important ball game worldwide, basketball plays an imperative role in most people's lives. Nowadays, specifically, it is a popular game among students, and most colleges take part in the College Women's Basketball League each year. While watching games, people always make some guesses about the game result and wonder which team would become the final winner this year based on their life experience. However, their assumptions would just be made based on their life experience without scientific evidence, which might make the prediction unreliable. Meantime, there are a large number of documents about analyzing baseball teams' performance based on statistical values, while significantly fewer researches have analyzed other sports such as basketball [1].

Thus, in this work, we construct a linear model and fit it with data about 2014-2015 to 2018-2019 Women's College Basketball Games for the prediction of game results and discuss the basic results of the model prediction and the out-of-sample prediction. The data were obtained from National Collegiate Atheletic Association (NCAA) website[2]. Our work was conducted in the R Language and Environment for Statistical Computing[3].

In this work, we discusses the process of obtaining, cleaning, and re-organizing the data, describes the process of building the linear model, and the basic results of the linear model and the out-of-sample prediction.

2. About the Data

2.1 General Cleaning Work

The raw HTML files contain total 351 women basketball teams' game information from season 2014-2015 to season 2018-2019. For each game, the game date, primary team, opponent team and game

ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0049

result are being recorded. Some games also have their tournament name being recorded. However, there are also some special HTML codes inside the raw files we do not need.

Therefore, In the data cleaning section, we aims to extract the data from the raw HTML files and process them to form a data frame x. The information about each game is corresponding to each row in x. There are 55169 rows and 11 column in x, which means that there are total 55169 games with their 11 types of information being recorded.

For example,

```
[572] "      <td>01/18/2019</td>"

[573] "      <td>"

[574] "        <a href=\"/teams/451888\"><img alt=\"Yale\" height=\"20px\" src=\"http://web2.ncaa.org/ncaa_style/img/Al
l_Logos/sm/813.gif\" width=\"30px\" /> Yale</a> "

[575] "      </td>"

[576] "      <td>"

[577] "          <a href=\"/contests/1627777/box_score\" class=\"skipMask\" target=\"BOX_SCORE_WINDOW\">W 86-71 </a>"
```

Figure 1. The raw HTML code containing information about the home game between Brown and Yale on Jan 18,2019 with other HTML codes we do not need.

```
           dates teamname    season  orgid opponents result myscores oppscores scorediff place
52793 01/18/2019    Brown 2018-2019 team80      Yale      W       86        71        15  home
                   gameid
52793 01/18/2019 Yale Brown
```

Figure 2. The data of this game after finishing cleaning process.

Table 1. 11 variables appeared in x and their meanings.

dates	Game date	myscores	Game score of primary team
teamname	Name of primary team	oppscores	Game score of opponent team
season	Game season	scorediff	Score differential between two teams
orgid	The team ID number	place	Game location (home / away / neutral)
opponent	Name of opponent team	gameid	Game identifier (Combination of game date, primary team name, and opponent team name)
result	Win (W) or Lose (L)		

2.2 Data Cleaning Challenges

2.2.1 Score of Each Game

In the raw HTML files, the score of each game is recorded with the pattern “primary team score - opponent team score ”. However, we need to get the score of these two teams separately in order to find the score differential. To solve the problem, we save the score in HTML files into a matrix and take out the list containing the two scores without "-" in the middle of them by using splitting elements of a character vector. Next, we convert the list to several characters and transform them into numerical data. At last, we make a two-column matrix with one column for primary teams' scores and the other one for opponents' scores.

For example,

ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0049

```
86-71
```

Figure 3. The score of one game in raw HTML file.

```
myscores oppscores
      86        71
```

Figure 4. The score of this game after cleaning.

2.2.2 Identifying Game Location

In the raw HTML files, the location of each game was not be given in words but with the appearance of "@" sign in each opponent team name or each tournament name. Thus, to find the location of each game, we search the location of "@" sign inside each opponent team name. The opponent name with no "@" sign indicates a home game for the primary team, while having "@" at the beginning of the name indicates an away game for the primary team. If the "@" sign appears at the beginning of the tournament name, the game would be a neutral site game.

For example,

```
[681] "        <a href=\"/teams/16696\"><img alt=\"Harvard\" height=\"20px\" src=\"http://web2.ncaa.org/ncaa_style/img/
All_Logos/sm/275.gif\" width=\"30px\" /> Harvard</a> "
```

Figure 5. The home game between Yale and Harvard with Yale as the primary team.

```
[698] "        @<a href=\"/teams/16919\"><img alt=\"Yale\" height=\"20px\" src=\"http://web2.ncaa.org/ncaa_style/img/Al
l_Logos/sm/813.gif\" width=\"30px\" /> Yale</a> "
```

Figure 6. The away game between Harvard and Yale with Harvard as the primary team.

```
[417] "        <a href=\"/teams/16615\"><img alt=\"Boise St.\" height=\"20px\" src=\"http://web2.ncaa.org/ncaa_style/im
g/All_Logos/sm/66.gif\" width=\"30px\" /> Boise St.</a> <br/>@Alaska Airlines Center - Anchorage, AK"
```

Figure 7. The neutral site game between Yale and Boise St.

2.2.3 Missing Games

There are some games with certain teams, which belong to the division 1 team list, were dropped after dropping the games with teams outside division 1 list. To figure out this problem, we do a detective work using gameid and look for the teams with some missing games. We found out three types of errors causing this problem:

As *Figure 8 and 9* shows, there are three teams being recorded as their whole names when they are in teamname, while being recorded as their abbreviate names in opponents, which result in the missing rows after we dropping the teams outside division 1 based on the name of teams. To fix the problem, we replace their abbreviate names with the matched whole names.

For example,

Team "Kansas City" is being recorded as the abbreviate name "UMKC" when it is not the primary team.

```
             dates     teamname    season    orgid opponents result myscores oppscores scorediff place
13017 01/07/2016 Kansas City 2015-2016 team2707 Seattle U      L       63        69        -6  away
```

Figure 8. Example of one game that "Kansas City" is recorded as its whole name when it is the primary team.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0049

	dates	teamname	season	orgid	opponents	result	myscores	oppscores	scorediff	place
3345	01/07/2016	Seattle U	2015-2016	team1356	UMKC	W	69	63	6	home

Figure 9. Example of this game that "Kansas City" is recorded as "UMKC" when it is the opponent team.

As Figure 10 and 11 shows, some opponent teams have their names being recorded as the combination of their own team names and the name of the tournaments in the raw HTML files. However, we just need the pure team name without the tournament name. To fix the problem, we use a special pattern to search these certain games with the special format of recorded name and only keep the pure team name in the data frame.

For example,

The game between CSUN and Washington St. was played in the tournament called "Warner Center Marriott Thanksgiving Basketball Classic", but what we want is just "Washing St." in the opponent.

```
[429] "        <a href=\"/teams/451871\"><img alt=\"Washington St.\" height=\"20px\" src=\"http://web2.ncaa.org/ncaa_style/img/All_Logos/s
m/754.gif\" width=\"30px\" /> Washington St.</a> <br/>Warner Center Marriott Thanksgiving Basketball Classic"
```

Figure 10. The information of a certain game in the raw HTML file.

295	11/23/2018	CSUN	2018-2019	Washington St.	W	65	52	13	home

Figure 11. The information of that game in x.

As *Figure 12* shows, there are 2 games with the opponent having two "@" signs in the team name in the HTEML files, which not only making us hard to clean but also causing confusion when determining the place based on the location of "@" in team name. To fix the problem, we choose to ignore these games and delete them since there are only a relatively small number of games containing the issue.

For example,

There are two "@" signs in the HTML code of the game between Iona and Siena.

```
[728] "        @<a href=\"/teams/451814\"><img alt=\"Siena\" height=\"20px\" src=\"http://web2.ncaa.org/ncaa_style/img/
All_Logos/sm/639.gif\" width=\"30px\" /> Siena</a> <br/>@Albany, NY (Metro Atlantic Conference Tournament)"
```

Figure 12. One example that contains two "@" signs in the HTML code of the game between Iona and Siena.

2.2.4 Duplicated Games

We realized that each game is recorded twice in x due to the switch of teamname and opponents like *Figure 13* shows.

For example,

52031	02/01/2019	Yale	2018-2019	Harvard	W	65	62	3	home	02/01/2019 Yale Harvard
13369	02/01/2019	Harvard	2018-2019	Yale	L	62	65	-3	away	02/01/2019 Yale Harvard

Figure 13. Two rows containing the information about one game.

However, we only need one of the two rows of each game to do the analysis and fit in the model later, avoiding using the score differences of each game twice. Thus, to drop the duplicated games, we search the rows with the same gameid and delete one of them.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0049

3. Linear Prediction Model

The past research utilized the multiple regression, discriminate analysis, or logistic regression analysis to predict the game result of possibility of winning games according to the strengths and scores of team members[4, 5, 6]. In this study, we employed a linear regression model to predict the game results based on both the strength of each whole team and the home court advantage.

3.1 Basic Model without Home Court Advantage

We will first describe the basic model, ignoring the home court advantage, which is using the linear model to find the coefficient of each team representing the strength of each team so that we could predict the result of each game with the score differential of it.

For example,

the score differential of the game between Yale and Harvard = the coefficient β of Yale - the coefficient β of Harvard + error term.

To build the basic model, we will firstly do the dummy-coding of indicator variables we need. Thus, we construct a model matrix, mm, with the number of games in one season as the number of rows of mm and the number of all teams in one season as the number of columns of mm, and we name each column with the name of all teams appeared in both teamname and opponents in that season. Then, fill teamname and opponents of each row with 1 and -1 respectively since we want to do the subtraction of score difference between each teamname and opponents in each game. After building the matrix, we choose the very first team "A&M-Corpus Christi" as the baseline of our model - in other words, let all other teams comparing their team strength with team "A&M-Corpus Christi" to show the difference in ability more obviously.

For example,

Table 2. The stimulated matrix of mm.

	Yale	Harvard	Cornell
1	1	-1	0
2	0	1	-1
3	-1	0	1

Table 3. The explanation of the model matrix.

Explanation of the model matrix
Yale and Harvard's game with Yale as the primary team
Harvard and Cornell's game with Harvard as the primary team
Yale and Cornell's game with Cornell as the primary team

The "0"s in *Table 2* indicate that those teams are not involved in the game at that row.

3.2 Final Model with Home Court Advantage

Other than the teams' own abilities, game location factors suggested that, generally, teams at home have more support from spectators than do visitors, and thus have a greater home court advantage that could influence the game result in some ways[7]. The home advantages as weak as 53% and as strong as 72% have been observed in professional baseball, football, basketball, ice hockey, soccer, and cricket[8]. Also, The home court advantage in baseball, football, and basketball has been presented at the collegiate competition[9, 10, 11]: Gayton, Mutrie, and Hearns found that a female basketball team had more victory records when competing at home than away in basketball(+ 13.8%)[12]. Thus, we want to make changes on the basic model and try to figure out how much extra advantage would the place bring. So, we now add to the basic model in order to considering the home court advantage. We add one additional column named "place" in mm and fill it with 1 for home games of the primary team, -1 for away game of the primary team, and 0 for neutral site games and use the coefficient of "place" as the home court advantage.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0049

For example,

the score differential of the home game between Yale and Harvard = β of Yale - β of Harvard + α (home court advantage) + error term

the score differential of the away game between Yale and Harvard = β of Yale - β of Harvard - α (home court advantage) + error term

the score differential of the neutral game between Yale and Harvard = β of Yale - β of Harvard + 0 (no home court advantage) + error term

3.3 Modeling Challenges

3.3.1 Identifiable Coefficient

When trying to estimate the strengths for all teams before choosing a baseline team, we failed to make the prediction since we could not uniquely estimate the coefficient of each team. Thus, the coefficient have to be identifiable, which requires choosing a baseline team and handling it specially - to make it a 0.

3.3.2 Symmetric Model

If we want to properly estimate the home court advantage, the predicting model needs to be symmetric. However, if we choose to use a standard analysis of variance approach with this categorical variable with three categories, the model would not be symmetric. This means that the points added to the home games would not be the same number as subtracted from the away games. Our model now achieves symmetric predictions by either adding or subtracting, as appropriate, the home court advantage represented by a single coefficient.

4. Results

4.1 Final model prediction of each woman basketball team's strength and the home court advantage

4.1.1 Basic prediction results

Due to a large number of games in all five seasons and the changes of each team's ability cross seasons, we decide to choose one season (2018-2019) as an example and fit the final model, lm.2, with the data of that season. Now, we could get both the predicted number of each team's strength and the estimated coefficient representing the home advantage.

```
      `Iowa St.`            Stanford           Louisville
           38.24               39.64               42.27
       Marquette              Oregon               UConn
           42.36               47.42               48.36
`Mississippi St.`       `Notre Dame`               Baylor
           48.77               50.42               51.40

 `Chicago St.`    `Ark.-Pine Bluff`              Alcorn
          -25.06              -19.34              -17.10
`LIU Brooklyn`      `Savannah St.`             McNeese
          -15.83              -15.65              -14.61
 `Florida A&M`     `Western Caro.`      `Saint Peter's`
          -14.51              -12.93              -12.83
```

Figure 14. The top 10 good teams and the last 10 bad teams in this season.

```
place
 2.95

Residual standard error: 11.16 on 5087 degrees of freedom
Multiple R-squared:  0.656,     Adjusted R-squared:  0.6322
F-statistic: 27.63 on 351 and 5087 DF,  p-value: < 2.2e-16
```

Figure 15. The home court advantage and the residual of the statistical summary of the prediction.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0049

We can tell from *Figure 15* that the predicted result might often be different from the real result by approximately one or two standard deviations, or 11 to 22 points.

We can predict the result of each single game between two teams with these number we got.

For example,

our estimated coefficient for Harvard is 20.17, and that for Yale is 11.23. Thus, we could make the prediction for the following games:

The score differential of the home game between Yale and Harvard = 11.23 - 20.17 + 2.95 + 11.16 = 5.17

The score differential of the away game between Yale and Harvard = 11.23 - 20.17 - 2.95 + 11.16 = - 0.73

The score differential of the neutral game between Yale and Harvard = 11.23 - 20.17 + 0 + 11.16 = 2.22

4.1.2 Distribution of residuals of the final model

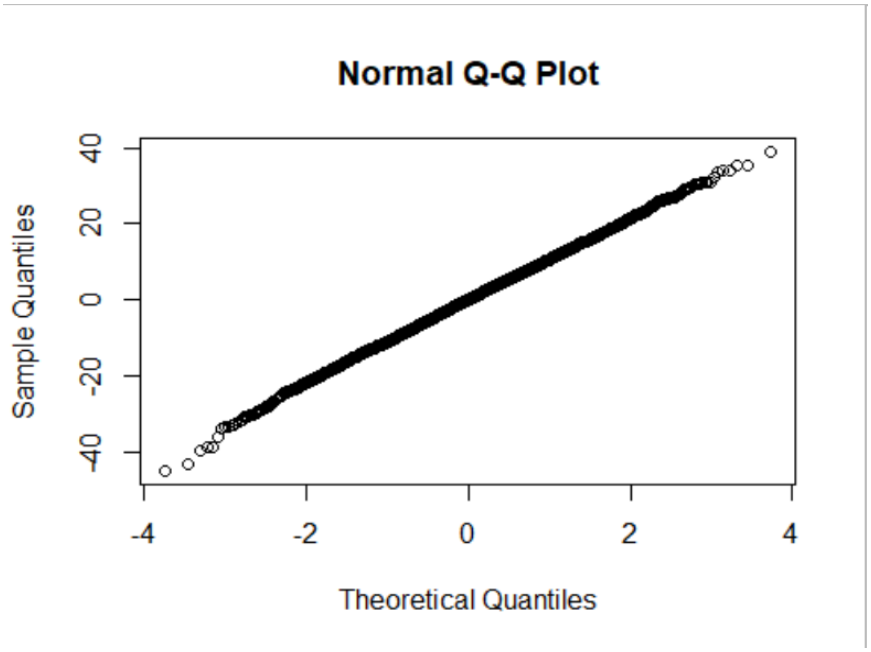

Figure 16. Normal Q-Q plot of the lm.2's residuals.

We can tell from the *Figure 16* that the model residuals are normally distributed since it is nearly a straight line, which indicates that our model is generally valid.

4.2 Out-of-sample Prediction of the results for games in March and April in season 2018-2019

Now, after getting each team's ability's predicted number and home advantage point from lm.2, we apply it into a more realistic situation: fit the model only using games through the end of February, and then use this model to predict game results for March and April in season 2018-2019. After getting the result of the model, we make a plot with the real and predicted score differences to indicate how well our model could predict the results of games.

```
Residual standard error: 11.8 on 794 degrees of freedom
Multiple R-squared:  0.5122,    Adjusted R-squared:  0.5116
F-statistic: 833.8 on 1 and 794 DF,  p-value: < 2.2e-16
```

Figure 17. Residual standard error and r-squared of the linear model prediction.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0049

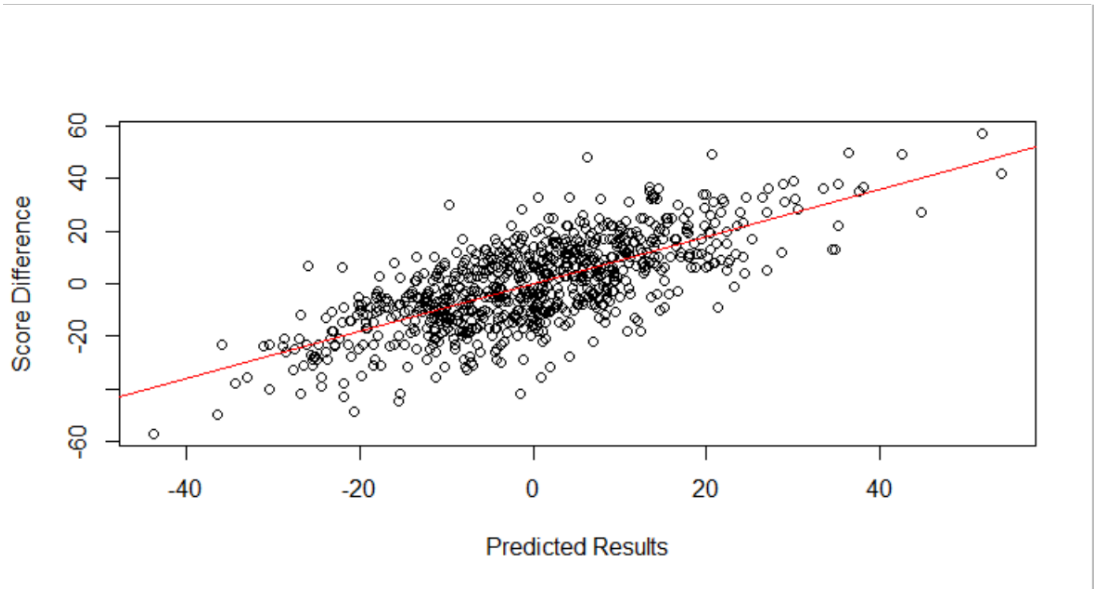

Figure 18. The relationship between predicted results and the real score differences.

We can tell from *Figure 17 and 18*, from both the summary of the model and the plot, that this model could explain about 51% of the variability of the score differences, and the predicted result might often be different from the real result by approximately one or two standard deviations, or 11 to 22 points, meaning the point on the plot could be 11 or 22 points above or below the red line.

For example,

03/16/2019	Cornell	Princeton	L	-21	-11.39

Figure 19. The result of game between Cornell and Princeton on Mar.16, 2019 including the real and prediction result.

As *Figure 19* shows, for the game between Cornell and Princeton on March 16, 2019, the real score differential is -21 points, while our predicted score differential is approximately -11.39 points. This predicted result is different from the real result by approximately 9.61 points, which is less than one standard deviation in size.

5. Conclusion

Our current model could predict the result of games in the same season as the sample games; however, if we try to use this model to predict the result of one season using the data from a previous season, the prediction might not be useful enough since the team ability could change a lot cross seasons. Therefore, to improve our model, we could find a team, which has similar strength in each season, as the baseline team.

Also, we assume the home court advantage is the same for all teams in this research. However, this might not be true. Thus, we could build a model and find the home court advantage for each team.

Furthermore, a team's ability might also change inside a season due to some accidental events, such as a player having injury or leaving the team. Therefore, we could look for some other data including more detail information do this investigation.

References

[1] Yuanhao (Stanley) Yang. (2015, May) "Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics." https://www.stat.berkeley.edu/~aldous/ Research/ Ugrad/ Stanley_Yang%20_Thesis.pdf.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0049

[2] NCAA.com. (2020, August 21). NCAA College Women's Basketball DI Stats. https://www.ncaa.com/stats/basketball-women/d1.

[3] RStudio Team (2019). RStudio: Integrated Development for R. RStudio, Inc, Boston, MA URL http://www.rstudio.com/.

[4] Haberman, S.J. Analysis of scores of Ivy League football games. In S. Ladany & R. Machol (Eds.), Optimal strategies in sports. Amsterdam: New Holland, 1977.

[5] Nie, N.H., Hull, C.H., Jenkins, J.G. Steinbrenner, K., & Bent, D.H.SPSS statistical package for the social sciences. New York: McGraw-Hill Book Co., 1975.

[6] Shanahan, Kathleen Jean. "A model for predicting the probability of a win in basketball." MA (Master of Arts) thesis, University of Iowa, 1984. https://doi.org/10.17077/etd.8wi48qk1.

[7] Carron, A., Loughhead, T., & Bray, S. (2005). The home advantage in sports competitions: Courneya and Carron's (1992) conceptual framework a decade later. Journal of Sports Sciences, 23(4), 395-407.(4).

[8] Courneya, K. S., & Carron, A. V. (1992). The home advantage in sport competitions: A literature review. Journal of Sport & Exercise Psychology, 14, 13-27.

[9] Courneya, K. S. (1990). Importance of game location and scoring first in college baseball. Perceptual and Motor Skills, 71, 624-626.

[10] Schwartz, B., & Barsky, S. E (1977). The home advantage. Social Forces, 55, 641-661.

[11] Silva, J. M., & Andrew, J. A. (1987). An analysis of game location and basketball performance in the Atlantic Coast Conference. International Journal of Sporf Psychology, 18, 188-204.

[12] Gayton, W. E, Mutrie, S. A., & Hearns, J. E (1987). Home advantage: Does it exist in women's sports? Perceptual and Motor Skills, 65, 653-654.