Prediction of future NBA games' point difference: A statistical modeling approach

Ruize Han (Jack), Tianrui Hu (Riley), Shunying Shi (Jenny), Sijian Tao (Raymond)

7/25/2021

Abstract

Nowadays, the popularity of the NBA along with the penetration of gambling ideas is rising all around the world. More and more fans prefer to bet on sports gambling. People want to predict the final score difference for each individual game. In addition, there are many either superficial or underlying factors that will affect the game's result. The central claim of this article is to establish a reliable model that could predict each game's result by extracting and analyzing previous games' outcomes. The major analysis is done by modeling and exploratory data analysis, and is composed of several graphical methods. For the modeling part, ideally the model could predict the games' score differences by analyzing previous dozens of games' data sets. The database consists of a single season 2016-2017 on Goldsheet website to explore more analysis. Then step by step checking and demonstrating to prove the feasibility and precision of the model. The model will collect and analyze the data information, like rebounds, teams, assists, turnovers, three points, free throws, blocks, and injury, from Season 2016-2017. The model will predict the future game result in Season 2016-2017 by using each team's previous game result in this season. This article indicates the idea of linear regression model to find the best fit by comparing the correlation strength of each variable, which include the home-field advantage, teams' technical statistics, and injury data, to the result of a game. The model final outcomes implies that the game's result has the most correlation with the home-field advantage and player injury; however, team's basic technical statistics, included the rebounds, blocks, turnovers, free throws, three points, and assists, have low correlation coefficient with the result to demonstrate that they are not significant to a game's result.

1. Introduction

Basketball, ranked at the third place on the world's ten most popular ball games, has around 2.2 billion fans all around the world. In all basketball leagues, the NBA, the National Basketball Association, is obviously the most popular basketball game in all countries. 30 teams participate in the NBA each year, and they are divided into two parts, named Western Conference and Eastern conference, and each of them contains 15 teams. Each season is also divided into two separate parts, the regular season, in which each team will play 82 games in total with others, and the playoff. The playoff only contains the best 16 teams, with eight teams each from the Eastern Conference and Western Conference, and they will compete for the final championship.

The NBA's audience is so large that the gambling industry is on the rise, and one of the gamble types is point spread betting. Point spread is a bet on the margin of victory in a game(1). People bet on whether the stronger team could win by point spread over the weaker team. Sometimes, winners' profits even reach up to $2,500(2). Thus, accurately predicting the outcome of a match can be very rewarding, that leads to more research nowadays, especially focusing on the aspect of sports gambling and point spreads prediction.

Common sense dictates that the ability of the players on the pitch can make a big difference to the outcome of a match. This research is dedicated to exploring factors that affect score difference per game, such as home-field advantages, turnovers, three points, injury and so on. Several models have been tried and compared in this study.

This paper is organized in five sections that describe the process and result of the research project. Section 2 describes the data, data cleaning process and merging; Section 3 presents the exploration and analysis; Section 4 summarizes the work that group did and the observations that were found; Finally, section 5 will give all the references that were used for this project.

2. Methods (Data Cleaning)

Data description:

In this study, the database for NBA season 2016-2017 is downloaded as an html document from Goldsheet webpage. Unfortunately, this html document includes a lot of messy code with dirty data. Thus, data cleaning is the first crucial step to address with. The first thing to do is to figure out a data frame containing all the information needed for the following research: date of games, names of team 1 and team 2, point spread, scores of team 1 and team 2, score differences(scores of team 1 minus that of team 2), and locations.

The first observation is that there are more than two types of locations. Besides "H" which stands for "home game" and "V" which stands for "visitor game", numerical numbers appeared following them. Actually, these are the numbers of overtime periods, usually 5 minutes for each overtime period of play according to the official NBA website(https://official.nba.com/rule-no-5-scoring-and-timing/). In addition, the team name is also one checkpoint. In Goldsheet data, one individual team could be expressed differently. One example is that the team "LA Lakers" has been shown not only as "LA Lakers", but also as "LA Laker" as a mistake. Thus, addressing different formatting names and canceling different versions of one team name is necessary. Furthermore, the raw data presents names of team 1 and team 2 in different formats. For the convenience of future studying, names of team 2 are adjusted to be the same as team 1. Last but not least, the data is internally duplicated. The method used here to deal with this problem is to firstly select Team 1 alphabetically, and secondly Team 2. The result is shown below, which only relies on the arrangement of the team 1 names and team 2 names in alphabetical order. If the team 1 name is "ATLANTA HAWKS", team 2 will arrange the names alphabetically with "B", since there are no other teams starting with "A". If the team 1 name is "BOSTON CELTICS", which starts with letter "B", the matching team 2 will start with "B" if there is another team name starting with "B". Therefore, "ATLANTA HAWKS" will never appear in the list of Team 2 and "WASHINGTON WIZARDS", which is the last team alphabetically, will never be Team 1. Consequently, all duplicated rows that represent the same game will be abandoned. One of the advantages of this method is that the locations of games are randomly selected, which will be of critical significance in the following study of home-field advantages. The columns are listed in the following order: Date of the Game, Team1 name, Team2 name, Pointspread of the Game, Team1 Score, Team2 Score, score difference between Team1 and Team2, The Location of the Game

```
##    Date_of_game         Team1             Team2 Pointspread Score1 Score2
## 1    2017-01-13 ATLANTA HAWKS    BOSTON CELTICS        -3.5    101    103
## 2    2017-02-27 ATLANTA HAWKS    BOSTON CELTICS         4.5    114     98
## 3    2017-04-06 ATLANTA HAWKS    BOSTON CELTICS         1.5    123    116
## 4    2017-01-10 ATLANTA HAWKS     BROOKLYN NETS        -8.5    117     97
## 5    2017-03-08 ATLANTA HAWKS     BROOKLYN NETS        -9.5    110    105
## 6    2017-03-26 ATLANTA HAWKS     BROOKLYN NETS        -6.5     92    107
## 7    2017-04-02 ATLANTA HAWKS     BROOKLYN NETS        -3.0     82     91
## 8    2016-11-18 ATLANTA HAWKS CHARLOTTE HORNETS         2.0     96    100
## 9    2016-12-17 ATLANTA HAWKS CHARLOTTE HORNETS        -2.5     99    107
## 10   2017-03-20 ATLANTA HAWKS CHARLOTTE HORNETS         6.0     90    105
##    scorediff Location
## 1         -2        H
## 2         16        V
## 3          7        H
## 4         20        V
## 5          5        H
## 6        -15        H
```

```
## 7         -9        V
## 8         -4        V
## 9         -8        H
## 10       -15        V
```

Additionally, the data of three points, rebounds, free throws, assists, blocks, turnovers of each game in 2016-2017 is downloaded from the Basketball Reference website. On Basketball Reference website, the per game stats data are organized into separate csv files by team name and season. The key point of cleaning data of variables of games is to select exactly the same games as those that were chosen above. Accordingly, the cleaned data can be built into the model and makes sense. The columns are listed in the following order: Date of the game, Team1 name, Team2 name, Team1 score, Team2 score, Team1 Field Goals, Team1 Field Goal Attempts, Team1 Field Goal Percentage, Team1 3-Point Field Goals, Team1 3-Point Field Goal Attempts, Team1 3-Point Field Goal Percentage, Team1 Free Throws, Team1 Free Throw Attempts, Team1 Free Throw Percentage, Team1 Offensive Rebounds, Team1 Total Rebounds, Team1 Assists, Team1 Steals, Team1 Blocks, Team1 Turovers, Team1 Personal Fouls, Team2 Field Goals, Team2 Field Goal Attempts, Team2 Field Goal Percentage, Team2 3-Point Field Goals, Team2 3-Point Field Goal Attempts, Team2 3-Point Field Goal Percentage, Team2 Free Throws, Team2 Free Throw Attempts, Team2 Free Throw Percentage, Team2 Offensive Rebounds, Team2 Total Rebounds, Team2 Assists, Team2 Steals, Team2 Blocks, Team2 Turovers, Team2 Personal Fouls,

```
##          Date     Team1long          Team2long Score1 Score2 Team1FG Team1FGA
## 1  2017-01-13 ATLANTA HAWKS    BOSTON CELTICS    101    103      36       84
## 2  2017-02-27 ATLANTA HAWKS    BOSTON CELTICS    114     98      46       95
## 3  2017-04-06 ATLANTA HAWKS    BOSTON CELTICS    123    116      44       89
## 4  2017-01-10 ATLANTA HAWKS     BROOKLYN NETS    117     97      44       92
## 5  2017-03-08 ATLANTA HAWKS     BROOKLYN NETS    110    105      41       95
## 6  2017-03-26 ATLANTA HAWKS     BROOKLYN NETS     92    107      34       98
## 7  2017-04-02 ATLANTA HAWKS     BROOKLYN NETS     82     91      30       80
## 8  2016-11-18 ATLANTA HAWKS CHARLOTTE HORNETS     96    100      39       84
## 9  2016-12-17 ATLANTA HAWKS CHARLOTTE HORNETS     99    107      40       86
## 10 2017-03-20 ATLANTA HAWKS CHARLOTTE HORNETS     90    105      37       84
##    Team1FG. Team13P Team13PA Team13P. Team1FT Team1FTA Team1FT. Team1ORB
## 1     0.429      11       30    0.367      18       21    0.857       12
## 2     0.484       6       25    0.240      16       20    0.800       13
## 3     0.494      11       23    0.478      24       34    0.706       13
## 4     0.478       7       23    0.304      22       38    0.579       15
## 5     0.432       4       20    0.200      24       32    0.750       11
## 6     0.347       5       27    0.185      19       32    0.594       26
## 7     0.375       7       24    0.292      15       23    0.652        8
## 8     0.464       7       28    0.250      11       17    0.647        8
## 9     0.465      10       26    0.385       9       16    0.563        9
## 10    0.440      12       37    0.324       4       10    0.400       12
##    Team1TRB Team1AST Team1STL Team1BLK Team1TOV Team1PF Team2FG Team2FGA
## 1        43       22        5        5       12      17      36       83
## 2        55       22       10       10       14      19      34       87
## 3        52       26        7        4       17      30      37       90
## 4        53       25       11       12       12      17      35       87
## 5        47       19       12        7       15      22      37       80
## 6        58       16       10        3       20      19      38       86
## 7        44       19        9        4       19      19      32       82
## 8        49       25        9        6       16      18      37       90
## 9        47       26        4        2       12      18      43       88
## 10       45       26        3        5       15      17      40       80
##    Team2FG. Team23P Team23PA Team23P. Team2FT Team2FTA Team2FT. Team2ORB
## 1     0.434      17       44    0.386      14       19    0.737        9
```

```
## 2     0.391      10       34    0.294      20       25    0.800        9
## 3     0.411      15       43    0.349      27       34    0.794       10
## 4     0.402      10       29    0.345      17       23    0.739       14
## 5     0.463      11       33    0.333      20       25    0.800        5
## 6     0.442       6       22    0.273      25       30    0.833       12
## 7     0.390      10       33    0.303      17       20    0.850        8
## 8     0.411      10       29    0.345      16       24    0.667        8
## 9     0.489      12       27    0.444       9       13    0.692        7
## 10    0.500      12       30    0.400      13       20    0.650        5
##    Team2TRB Team2AST Team2STL Team2BLK Team2TOV Team2PF
## 1        43       22        7        3       12      16
## 2        40       21        9        7       18      20
## 3        38       28        9        4       12      29
## 4        48       22        5        4       17      24
## 5        40       28        7       11       22      25
## 6        50       21       13        7       17      25
## 7        51       21       11        6       19      22
## 8        44       25       11        4       11      16
## 9        43       26        8        3       12      17
## 10       33       28       12        4        7      13
```

Players per game stats (field goals, three points, free throws, assists, blocks, turnovers, minute played, turnover, plus/minus) in 2016 -2017 is also downloaded from the Basketball Reference website. On the website, each csv contains the per game stats for one player in one season. Concatenating all csv files together to construct a single dataframe that contains all per game player stats for the 2016-2017 season. Some rows contain "Did Not Dress", "Not With Team", "Inactive", "Player Suspended", "Did Not Play", which means the player did not play that game. Removing these rows gives clean data. The columns are listed in the following order: Date of the game, the Player name, Team Name of the player, Minutes Played, Field Goals, Field Goal Attempts, Field Goal Percentage, 3-Point Field Goals, 3-Point Field Goal Attempts, Free Throws, Free Throw Attempts, Free Throw Percentage, Offensive Rebounds, Defensive Rebounds, Total Rebounds, Assists, Steals, Blocks, Turnovers, Personal Fouls, Points, Game Score, Plus/Minus.

```
##         Date       player  Tm            MP            FG           FGA
## 1 2016-10-26 Álex Abrines OKC         13:24             1             2
## 2 2016-10-28 Álex Abrines OKC Did Not Play Did Not Play Did Not Play
## 3 2016-10-30 Álex Abrines OKC          1:49             1             1
## 4 2016-11-02 Álex Abrines OKC          7:54             2             2
## 5 2016-11-03 Álex Abrines OKC         19:20             1             6
## 6 2016-11-05 Álex Abrines OKC         17:16             0             1
##            FG.          X3P         X3PA         X3P.           FT          FTA
## 1         .500            1            1        1.000            0            0
## 2 Did Not Play Did Not Play Did Not Play Did Not Play Did Not Play Did Not Play
## 3        1.000            0            0                         3            3
## 4        1.000            2            2        1.000            0            0
## 5         .167            1            6         .167            2            2
## 6         .000            0            0                         1            1
##            FT.          ORB          DRB          TRB          AST          STL
## 1                           0            1            1            0            0
## 2 Did Not Play Did Not Play Did Not Play Did Not Play Did Not Play Did Not Play
## 3        1.000            0            0            0            0            1
## 4                           0            0            0            0            0
## 5        1.000            1            2            3            1            0
## 6        1.000            0            2            2            2            1
##            BLK          TOV           PF          PTS         GmSc         X...
## 1            0            2            3            3         -0.9           +3
```

```
## 2 Did Not Play Did Not Play Did Not Play Did Not Play Did Not Play Did Not Play
## 3            0            0            0            5          5.7            0
## 4            0            2            0            6          3.4           +4
## 5            0            0            0            5          3.2          -13
## 6            0            0            1            1          2.9           -4
```

The injury data is scrapped from Pro Sports Transactions Archive website. The data contains five columns: Date, Team, Acquired, Relinquished, and Note. Every row of the data indicates an event about a player.

```
##          X       Date  Team      Acquired  Relinquished
## 1167 16993 2016-12-28 Spurs                Kawhi Leonard
## 1234 17060 2017-01-01 Spurs Kawhi Leonard
## 1622 17448 2017-01-23 Spurs                Kawhi Leonard
## 1697 17523 2017-01-27 Spurs Kawhi Leonard
## 1866 17692 2017-02-06 Spurs                Kawhi Leonard
## 1898 17724 2017-02-08 Spurs Kawhi Leonard
## 2237 18063 2017-03-08 Spurs                Kawhi Leonard
## 2253 18079 2017-03-09 Spurs Kawhi Leonard
## 2267 18093 2017-03-10 Spurs                Kawhi Leonard
## 2284 18110 2017-03-11 Spurs                Kawhi Leonard
##                                Notes
## 1167          gastroenteritis (DTD)
## 1234             returned to lineup
## 1622           sore left hand (DTD)
## 1697             returned to lineup
## 1866 bruised right quadriceps (DTD)
## 1898             returned to lineup
## 2237          placed on IL for rest
## 2253              activated from IL
## 2267               concussion (DTD)
## 2284   placed on IL with concussion
```

If the a relinquishing of a player is immediately followed by the acquisition of this player, we regard such relinquishing-acquisition pair as a player injury-and-recovery event. All other rows are disregared as these rows are describing what happend for this player during the game, and we cannot have this infomation before the game happens. The player relinquishing date is the date he will not be able to play the game and the games after this date, whereas the player acquisition date is the date he will be able to play the game and he will be able to play the following games. In the example below, Kawhi Lenonard is not on the court from 2017-02-06 to 2017-02-07, and he will be avalible for the game that happens on 2017-02-08 or after this date.

```
##          X       Date  Team      Acquired  Relinquished
## 1866 17692 2017-02-06 Spurs                Kawhi Leonard
## 1898 17724 2017-02-08 Spurs Kawhi Leonard
##                                Notes
## 1866 bruised right quadriceps (DTD)
## 1898             returned to lineup
```

Pairing every player's Relinquishing and Acquisition produces a dataframe that every row is an injury event of a player including when he got injured and when he recovered (Team, Player, GoneDate, BackDate). the GoneDate and BackDate of a player are the relinquishing date and the a cquisition date for the player.

```
##       X      Team         Player   GoneDate   BackDate
## 1 15836   Hornets Frank Kaminsky 2016-06-29 2016-10-01
## 2 15922 Grizzlies     Marc Gasol 2016-10-15 2016-10-26
## 3 15926      Heat   Luke Babbitt 2016-10-15 2016-10-26
## 4 15937 Mavericks    Jose Barea  2016-10-24 2016-10-26
## 5 15945    Pacers     C.J. Miles 2016-10-24 2016-10-26
```

```
## 6 15948  Pelicans  Anthony Davis 2016-10-12 2016-10-26
```

Exploratory Data Analysis:

Home-Field Advantage :

Home-field advantage refers to the benefit given to the home team over the visiting team(https://en.wikipedia.org/wiki/Home_advantage). It is reasonable because the home game team should have some advantage on whether players are prepared mentally or they are more familiar with the home game field. The home team is inclined to have more fans to come to the court to support them, which will give them the strongest confidence and intense disturbance for the opponent team. In the case, if the team "LA Lakers" has an obvious higher score difference when they play at home than they play as a visitor on average, then the "home-field advantage" variable is assumed to play a role here.

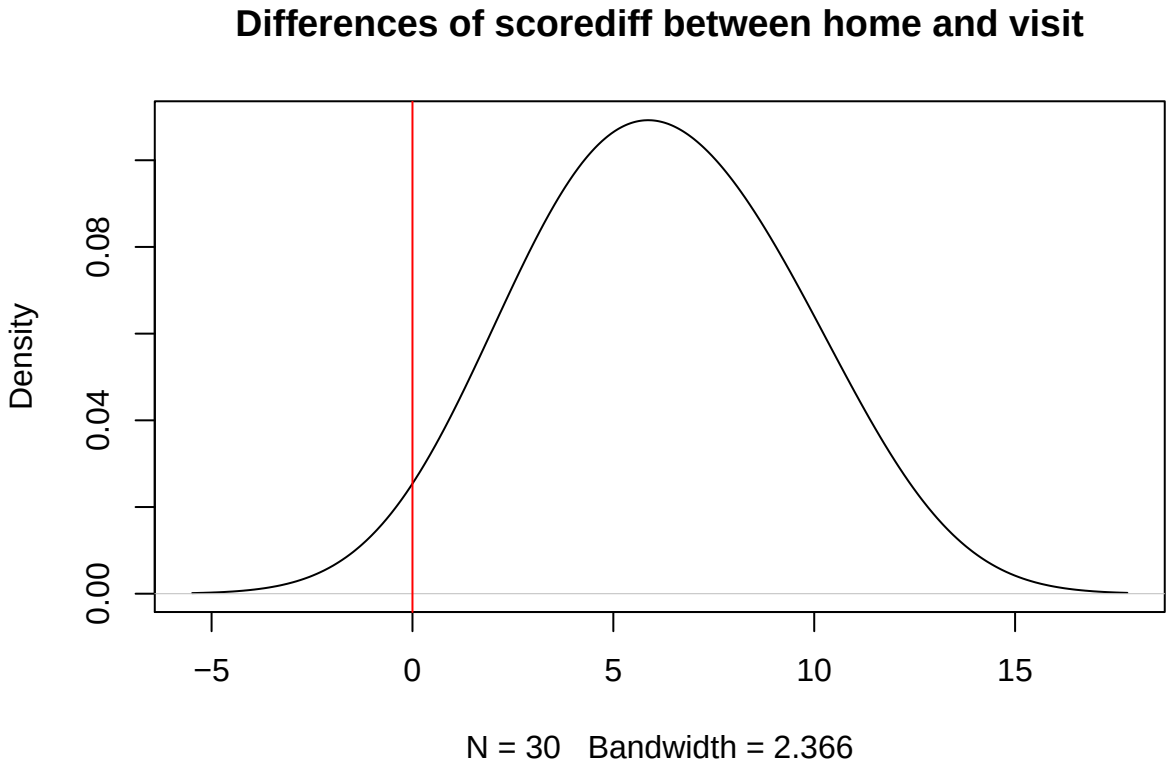
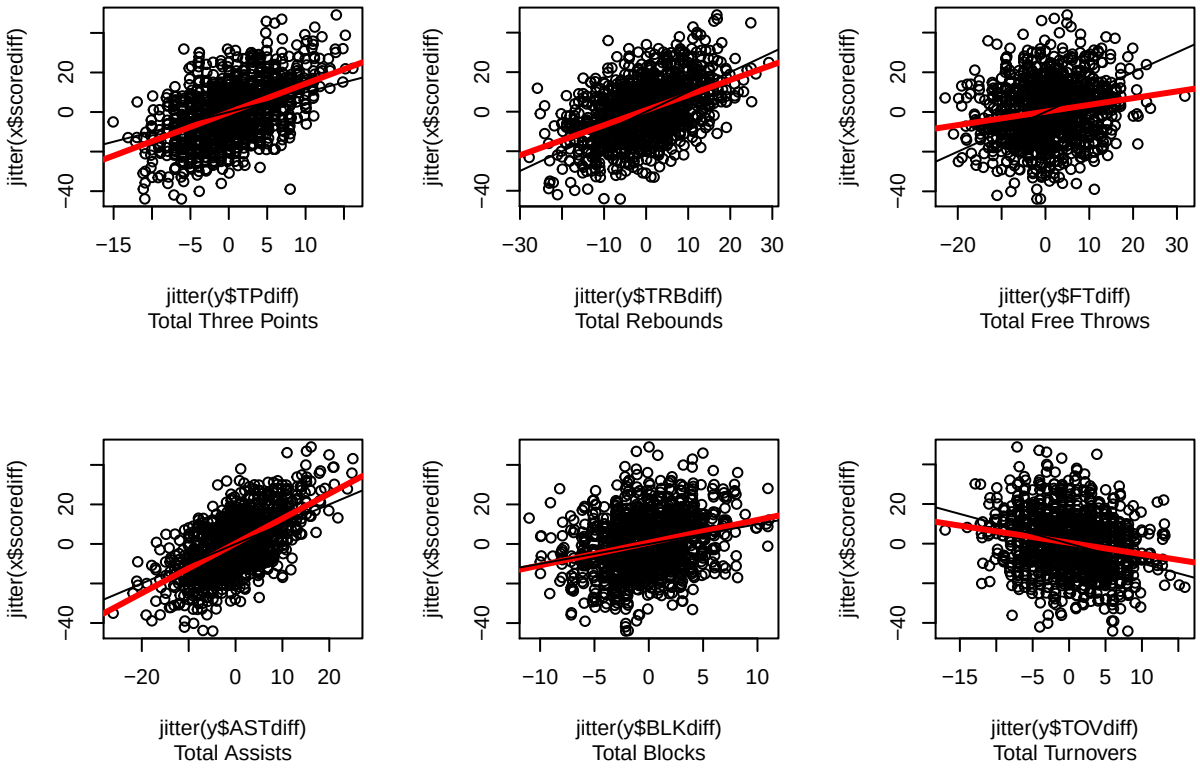


Figure 1: Comparison of the mean score difference based on game status

After the data cleaning process mentioned above, the total numbers of home games, visitor games and natural games are 644, 642 and 3 respectively. By plotting the density of differences of 30 teams' score differences between home games and visitor games, it can be easily seen that the mean is greater than 0. It is shown in Figure 1, which implies that teams tend to perform better at home than play as a visitor. To calculate the mean average score differences based on locations, it is interesting to notice that home-field advantage seems to give the home team higher scores by 3.07 points. According to Marshall(2007), he also reviewed the significance of home-field advantage, and calculated the average score difference between home and away at about 3.5 points, which is quite similar. He even finally concluded that home teams have 60% confidence to win the game.

Game Statistics :

Despite the first major variable home-field advantage, there are many other variables which could have influenced the result of a game. There are six indicators listed below:



In the Figure above, it contains six variables, which are the total three points, rebounds, free throws, assists, blocks, and turnovers. Obviously, all of them do have a correlation with the final score differences. The first five are positively connected with score differences, while turnover is a negative factor. Among them, the free throw variable has the weakest correlation. In contrast, three points and assists have much higher coefficients, which means their additional change will have more impact on the score difference. This does make sense, because free throws could only get one point at a time; however, three points could gain three points per goal. Thus, the three points variable is more important. One more observation is that the assist variable has a steeper slope than the blocking and rebounding variables. This is because each single independent assist implies one score by someone in the team; however, rebounds and blocks could only give a sense of the team's defensive ability being not bad. It is only an attack opportunity, and does not necessarily mean that the team will gain more points from them. Thus unsurprisingly, the indicator assist is directly related to scores that gain a high correlation; while the rebound and block variables are indirect to a team's scores, which get lower correlation coefficients. Apart from those five variables, turnover is also important. The coefficient of turnover is negative, since when one team makes a turnover, they will give the opponent possession of the ball. The opportunity of the opponent's score implies the loss of score differences. They could make a good defense that does not let the opponent score, or the opponent takes this opportunity to score, and make the score difference larger.

Injury :

Injury is one of the key factors that affects the game results, especially the injury of a key player can have a huge impact on the game. For example, Kawhi Leonard from SPURS is injured and cannot play form 2017-05-16 to 2017-05-22 during the playoff games vs WARRIORS. After his injury, the SPURS loses the following games by a greater margin compared with Leonard is on the court.

```
##       Team        Player   GoneDate   BackDate
## 1302 SPURS Kawhi Leonard 2017-05-16 2017-05-22
```

```
##     Date_of_game                Team1             Team2 Pointspread Score1
## 793   2017-05-01       HOUSTON ROCKETS SAN ANTONIO SPURS         6.0    126
## 794   2017-05-03       HOUSTON ROCKETS SAN ANTONIO SPURS         5.5     96
```

```
## 795   2017-05-05       HOUSTON ROCKETS SAN ANTONIO SPURS        -5.0     92
## 796   2017-05-07       HOUSTON ROCKETS SAN ANTONIO SPURS        -5.5    125
## 797   2017-05-09       HOUSTON ROCKETS SAN ANTONIO SPURS         5.5    107
## 798   2017-05-11       HOUSTON ROCKETS SAN ANTONIO SPURS        -8.5     75
## 724   2017-05-14 GOLDEN STATE WARRIORS SAN ANTONIO SPURS       -10.0    113
## 725   2017-05-16 GOLDEN STATE WARRIORS SAN ANTONIO SPURS       -13.5    136
## 726   2017-05-20 GOLDEN STATE WARRIORS SAN ANTONIO SPURS        -9.0    120
## 727   2017-05-22 GOLDEN STATE WARRIORS SAN ANTONIO SPURS       -11.5    129
##     Score2 scorediff Location
## 793     99        27        V
## 794    121       -25        V
## 795    103       -11        H
## 796    104        21        H
## 797    110        -3        V
## 798    114       -39        H
## 724    111         2        H
## 725    100        36        H
## 726    108        12        V
## 727    115        14        V
```

For Zaza Pachulia, an average player from WARRIORS, his absent from 2017-05-20 to 2017-06-01 has no obvious impact on WARRIORS performance.

```
##          Team        Player   GoneDate   BackDate
## 1303 WARRIORS Zaza Pachulia 2017-05-20 2017-06-01
```

```
##     Date_of_game                Team1                Team2 Pointspread Score1
## 736   2017-05-08 GOLDEN STATE WARRIORS            UTAH JAZZ        -8.5    121
## 724   2017-05-14 GOLDEN STATE WARRIORS     SAN ANTONIO SPURS       -10.0    113
## 725   2017-05-16 GOLDEN STATE WARRIORS     SAN ANTONIO SPURS       -13.5    136
## 726   2017-05-20 GOLDEN STATE WARRIORS     SAN ANTONIO SPURS        -9.0    120
## 727   2017-05-22 GOLDEN STATE WARRIORS     SAN ANTONIO SPURS       -11.5    129
## 407   2017-06-01   CLEVELAND CAVALIERS GOLDEN STATE WARRIORS         7.5     91
## 408   2017-06-04   CLEVELAND CAVALIERS GOLDEN STATE WARRIORS         9.0    113
## 409   2017-06-07   CLEVELAND CAVALIERS GOLDEN STATE WARRIORS         3.5    113
## 410   2017-06-09   CLEVELAND CAVALIERS GOLDEN STATE WARRIORS         5.5    137
## 411   2017-06-11   CLEVELAND CAVALIERS GOLDEN STATE WARRIORS         8.5    120
##     Score2 scorediff Location
## 736     95        26        V
## 724    111         2        H
## 725    100        36        H
## 726    108        12        V
## 727    115        14        V
## 407    113       -22        V
## 408    132       -19        V
## 409    118        -5        H
## 410    116        21        H
## 411    129        -9        V
```

These examples show that injury may have an impact on game results, and who got injured matters a lot.

3. Model Building

Building a linear model to predict the future game point spread is complicated since there are so many factors. In this study, only several variables are chosen, including 30 teams as a whole, home-field advantages, six

indicators of each team per game mentioned above, and injury information. Four models have been tried and will be illustrated in detail below. All models are built to predict game results, i.e. score differences.

Model 1: Model with only 30 teams

In this model, only the competence of 30 basketball teams is considered. The model is built by making up a matrix with 1309 rows representing all games from the season 2016-2017, and 30 columns, which stand for 30 teams. Here, 29 dummy variables are needed. The formula is listed below.

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \ldots\ldots \alpha_{29} D_{29i} + U_i$$
$$where$$
$$Y_i = Socre\ difference\ based\ on\ each\ game$$
$$D_{2i} = 1, if\ the\ team\ is\ Team1\ in\ the\ game$$
$$D_{2i} = 0, if\ the\ team\ did\ not\ play\ the\ game$$
$$D_{2i} = -1, if\ the\ team\ is\ Team1\ in\ the\ game$$
$$\alpha_1 = 0, Intercept, which\ is\ not\ needed$$
$$\alpha_2, \alpha_3 \ldots\ldots = Each\ team's\ coefficient\ value,\ there\ are\ 29\ coefficients\ in\ total$$
$$D_{2i}, D_{3i} \ldots\ldots = Represents\ each\ team,\ there\ are\ 29\ teams\ in\ total$$
$$U_i = Error$$

The basic idea of making this dummy variable is to assign the team who is team 1 a value of 1, assign the team who is team 2 a value of -1, and assign all the other team's the value of 0. For instance, if one of the rows represents the game "LA Lakers" versus "Boston Celtics", the result in the dummy variable matrix will be shown the positive number 1 under the LA Lakers column corresponding to that date of game, and negative number 1 will be presented under the Boston Celtics column corresponding to the date of the game. Now, at the specific row, team 1 "LA Lakers" will have number one, team 2 "Boston Celtics" will have a number -1, and spontaneously the rest of other teams will all have number 0 for that row. This is because for this specific game, only two teams have participated, any other teams does not have any relation to that game at all. Thus, they all get number 0 for this row of game.

By running this model, as shown by the formula above , a positive intercept exists, which means the model will give extra points to team 1 but not team 2.

However, this is wrong here. In the case of not considering home-field advantages, if the status of team 1 and team 2 are reversed, the absolute value of the score difference will be different, which actually should not be. In other words, the formula of this model should pass the original point. Thus, the model needs to delete the intercept by -1 in model construction code. After that, there is one team variable that states NA value, and the team "Washington Wizards" in this model is actually being a role of a standard team. Each other 29 teams will be compared to the Washington Wizards and get a team value from it. Thus, the Washington Wizards will be dismissed from the model, and be used as a comparative variable to help make other teams' value more visualized. It is all the coefficients of other 29 teams that suggest the relative competence of this team against "Washington Wizards"'. For example, the coefficient of Golden State Warriors is 11.1106, which implies that this team plays much better than Washington Wizards. It is a good team and keeps doing great. The corresponding p-value is also quite small. The coefficient of Memphis Grizzlies is -0.4113, which shows that these two teams are really evenly matched, though Memphis is a little bit inferior. The multiple R-squared is 0.177, with residual standard error 13.04.

```
##
## Call:
## lm(formula = sdif1309 ~ -1 + teamMM1309[, 1:29])
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -45.311  -8.247   0.324   8.583  45.234
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## teamMM1309[, 1:29]ATLANTA.HAWKS           -2.3503     1.8346  -1.281 0.200400
## teamMM1309[, 1:29]BOSTON.CELTICS           0.4167     1.7766   0.235 0.814584
## teamMM1309[, 1:29]BROOKLYN.NETS           -7.9764     1.9265  -4.140 3.70e-05
## teamMM1309[, 1:29]CHARLOTTE.HORNETS       -1.2827     1.9266  -0.666 0.505690
## teamMM1309[, 1:29]CHICAGO.BULLS           -1.5194     1.8907  -0.804 0.421750
## teamMM1309[, 1:29]CLEVELAND.CAVALIERS      3.2926     1.8457   1.784 0.074665
## teamMM1309[, 1:29]DALLAS.MAVERICKS        -3.7588     1.9591  -1.919 0.055250
## teamMM1309[, 1:29]DENVER.NUGGETS          -0.5326     1.9591  -0.272 0.785789
## teamMM1309[, 1:29]DETROIT.PISTONS         -2.5126     1.9377  -1.297 0.194983
## teamMM1309[, 1:29]GOLDEN.STATE.WARRIORS   11.1106     1.8736   5.930 3.89e-09
## teamMM1309[, 1:29]HOUSTON.ROCKETS          4.5786     1.9039   2.405 0.016319
## teamMM1309[, 1:29]INDIANA.PACERS          -1.7840     1.9060  -0.936 0.349472
## teamMM1309[, 1:29]LOS.ANGELES.CLIPPERS     3.0669     1.9233   1.595 0.111035
## teamMM1309[, 1:29]LOS.ANGELES.LAKERS      -7.5250     1.9591  -3.841 0.000128
## teamMM1309[, 1:29]MEMPHIS.GRIZZLIES       -0.4113     1.9277  -0.213 0.831066
## teamMM1309[, 1:29]MIAMI.HEAT              -0.4431     1.9268  -0.230 0.818155
## teamMM1309[, 1:29]MILWAUKEE.BUCKS         -1.4880     1.8963  -0.785 0.432768
## teamMM1309[, 1:29]MINNESOTA.TIMBERWOLVES  -1.8588     1.9591  -0.949 0.342895
## teamMM1309[, 1:29]NEW.ORLEANS.PELICANS    -2.9082     1.9591  -1.484 0.137928
## teamMM1309[, 1:29]NEW.YORK.KNICKS         -5.0900     1.9267  -2.642 0.008348
## teamMM1309[, 1:29]OKLAHOMA.CITY.THUNDER   -0.2960     1.9324  -0.153 0.878289
## teamMM1309[, 1:29]ORLANDO.MAGIC           -7.8456     1.9273  -4.071 4.97e-05
## teamMM1309[, 1:29]PHILADELPHIA.76ERS      -7.0695     1.9369  -3.650 0.000273
## teamMM1309[, 1:29]PHOENIX.SUNS            -6.3599     1.9591  -3.246 0.001199
## teamMM1309[, 1:29]PORTLAND.TRAIL.BLAZERS  -1.6899     1.9372  -0.872 0.383176
## teamMM1309[, 1:29]SACRAMENTO.KINGS        -4.5135     1.9591  -2.304 0.021387
## teamMM1309[, 1:29]SAN.ANTONIO.SPURS        5.8109     1.8817   3.088 0.002058
## teamMM1309[, 1:29]TORONTO.RAPTORS          1.6294     1.8857   0.864 0.387718
## teamMM1309[, 1:29]UTAH.JAZZ                2.5946     1.9037   1.363 0.173151
##
## teamMM1309[, 1:29]ATLANTA.HAWKS
## teamMM1309[, 1:29]BOSTON.CELTICS
## teamMM1309[, 1:29]BROOKLYN.NETS           ***
## teamMM1309[, 1:29]CHARLOTTE.HORNETS
## teamMM1309[, 1:29]CHICAGO.BULLS
## teamMM1309[, 1:29]CLEVELAND.CAVALIERS     .
## teamMM1309[, 1:29]DALLAS.MAVERICKS        .
## teamMM1309[, 1:29]DENVER.NUGGETS
## teamMM1309[, 1:29]DETROIT.PISTONS
## teamMM1309[, 1:29]GOLDEN.STATE.WARRIORS   ***
## teamMM1309[, 1:29]HOUSTON.ROCKETS         *
## teamMM1309[, 1:29]INDIANA.PACERS
## teamMM1309[, 1:29]LOS.ANGELES.CLIPPERS
## teamMM1309[, 1:29]LOS.ANGELES.LAKERS      ***
## teamMM1309[, 1:29]MEMPHIS.GRIZZLIES
## teamMM1309[, 1:29]MIAMI.HEAT
## teamMM1309[, 1:29]MILWAUKEE.BUCKS
## teamMM1309[, 1:29]MINNESOTA.TIMBERWOLVES
## teamMM1309[, 1:29]NEW.ORLEANS.PELICANS
```

```
## teamMM1309[, 1:29]NEW.YORK.KNICKS        **
## teamMM1309[, 1:29]OKLAHOMA.CITY.THUNDER
## teamMM1309[, 1:29]ORLANDO.MAGIC          ***
## teamMM1309[, 1:29]PHILADELPHIA.76ERS     ***
## teamMM1309[, 1:29]PHOENIX.SUNS           **
## teamMM1309[, 1:29]PORTLAND.TRAIL.BLAZERS
## teamMM1309[, 1:29]SACRAMENTO.KINGS       *
## teamMM1309[, 1:29]SAN.ANTONIO.SPURS      **
## teamMM1309[, 1:29]TORONTO.RAPTORS
## teamMM1309[, 1:29]UTAH.JAZZ
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.04 on 1280 degrees of freedom
## Multiple R-squared:  0.177,  Adjusted R-squared:  0.1583
## F-statistic: 9.491 on 29 and 1280 DF,  p-value: < 2.2e-16
```

Model 2: Model with 30 teams and home-court-advantage

After analyzing variables of 30 teams, the linear model can be improved. The main purpose for this model is to combine teams with home-field advantages. Whether the game takes place at home or away will be considered. Similarly, the method of dummy variables is applied. Home games will be assigned 1, visit games will be assigned -1, and neutral games will be 0. Indeed, it is still a matrix with 1309 rows representing games, but only 1 column presenting the locations. Adding this variable directly after the formula above, the formula is shown below.

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \ldots\ldots \alpha_{29} D_{29i} + \alpha_h H + U_i$$

$$where$$

$$Y_i = Socre\ difference\ based\ on\ each\ game$$

$$D_{2i} = 1, if\ the\ team\ is\ Team1\ in\ the\ game$$

$$D_{2i} = 0, if\ the\ team\ did\ not\ play\ the\ game$$

$$D_{2i} = -1, if\ the\ team\ is\ Team1\ in\ the\ game$$

$$H = 1, if\ team1\ is\ the\ home\ team$$

$$H = 0, if\ the\ game\ is\ played\ at\ a\ neutral\ site$$

$$H = -1, if\ team1\ is\ the\ visitor\ team$$

$$\alpha_1 = 0, Intercept, which\ is\ not\ needed$$

$$\alpha_2, \alpha_3 \ldots\ldots = Each\ team's\ coefficient\ value,\ there\ are\ 29\ coefficients\ in\ total$$

$$\alpha_H = thehome\ court\ advantage\ per\ game\ in\ measured\ in\ scores$$

$$D_{2i}, D_{3i} \ldots\ldots = Represents\ each\ team,\ there\ are\ 29\ teams\ in\ total$$

$$U_i = Error$$

After running this model, the result is shown below.

```
##
## Call:
## lm(formula = sdif1309 ~ -1 + teamMM1309[, 1:29] + xxMA1309$hv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.068  -7.913   0.373   7.957  42.217
```

```
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## teamMM1309[, 1:29]ATLANTA.HAWKS          -2.3742     1.7833  -1.331 0.183314
## teamMM1309[, 1:29]BOSTON.CELTICS          0.3359     1.7269   0.194 0.845830
## teamMM1309[, 1:29]BROOKLYN.NETS          -8.0036     1.8727  -4.274 2.06e-05
## teamMM1309[, 1:29]CHARLOTTE.HORNETS      -1.3092     1.8728  -0.699 0.484634
## teamMM1309[, 1:29]CHICAGO.BULLS          -1.5496     1.8378  -0.843 0.399292
## teamMM1309[, 1:29]CLEVELAND.CAVALIERS     3.3195     1.7941   1.850 0.064511
## teamMM1309[, 1:29]DALLAS.MAVERICKS       -3.8206     1.9043  -2.006 0.045035
## teamMM1309[, 1:29]DENVER.NUGGETS         -0.5226     1.9043  -0.274 0.783794
## teamMM1309[, 1:29]DETROIT.PISTONS        -2.5395     1.8836  -1.348 0.177823
## teamMM1309[, 1:29]GOLDEN.STATE.WARRIORS  11.0571     1.8212   6.071 1.67e-09
## teamMM1309[, 1:29]HOUSTON.ROCKETS         4.5204     1.8507   2.443 0.014717
## teamMM1309[, 1:29]INDIANA.PACERS         -1.8416     1.8528  -0.994 0.320438
## teamMM1309[, 1:29]LOS.ANGELES.CLIPPERS    3.0103     1.8695   1.610 0.107601
## teamMM1309[, 1:29]LOS.ANGELES.LAKERS     -7.5517     1.9043  -3.966 7.73e-05
## teamMM1309[, 1:29]MEMPHIS.GRIZZLIES      -0.4410     1.8738  -0.235 0.813971
## teamMM1309[, 1:29]MIAMI.HEAT             -0.4693     1.8729  -0.251 0.802193
## teamMM1309[, 1:29]MILWAUKEE.BUCKS        -1.5139     1.8433  -0.821 0.411628
## teamMM1309[, 1:29]MINNESOTA.TIMBERWOLVES -1.8851     1.9043  -0.990 0.322393
## teamMM1309[, 1:29]NEW.ORLEANS.PELICANS   -2.9345     1.9043  -1.541 0.123574
## teamMM1309[, 1:29]NEW.YORK.KNICKS        -5.1166     1.8729  -2.732 0.006383
## teamMM1309[, 1:29]OKLAHOMA.CITY.THUNDER  -0.2897     1.8784  -0.154 0.877442
## teamMM1309[, 1:29]ORLANDO.MAGIC          -7.8722     1.8735  -4.202 2.83e-05
## teamMM1309[, 1:29]PHILADELPHIA.76ERS     -7.0970     1.8828  -3.769 0.000171
## teamMM1309[, 1:29]PHOENIX.SUNS           -6.3151     1.9043  -3.316 0.000938
## teamMM1309[, 1:29]PORTLAND.TRAIL.BLAZERS -1.7167     1.8830  -0.912 0.362117
## teamMM1309[, 1:29]SACRAMENTO.KINGS       -4.5398     1.9043  -2.384 0.017273
## teamMM1309[, 1:29]SAN.ANTONIO.SPURS       5.7504     1.8291   3.144 0.001706
## teamMM1309[, 1:29]TORONTO.RAPTORS         1.6051     1.8330   0.876 0.381378
## teamMM1309[, 1:29]UTAH.JAZZ               2.5963     1.8505   1.403 0.160850
## xxMA1309$hv                               3.0516     0.3508   8.699  < 2e-16
##
## teamMM1309[, 1:29]ATLANTA.HAWKS
## teamMM1309[, 1:29]BOSTON.CELTICS
## teamMM1309[, 1:29]BROOKLYN.NETS          ***
## teamMM1309[, 1:29]CHARLOTTE.HORNETS
## teamMM1309[, 1:29]CHICAGO.BULLS
## teamMM1309[, 1:29]CLEVELAND.CAVALIERS    .
## teamMM1309[, 1:29]DALLAS.MAVERICKS       *
## teamMM1309[, 1:29]DENVER.NUGGETS
## teamMM1309[, 1:29]DETROIT.PISTONS
## teamMM1309[, 1:29]GOLDEN.STATE.WARRIORS  ***
## teamMM1309[, 1:29]HOUSTON.ROCKETS        *
## teamMM1309[, 1:29]INDIANA.PACERS
## teamMM1309[, 1:29]LOS.ANGELES.CLIPPERS
## teamMM1309[, 1:29]LOS.ANGELES.LAKERS     ***
## teamMM1309[, 1:29]MEMPHIS.GRIZZLIES
## teamMM1309[, 1:29]MIAMI.HEAT
## teamMM1309[, 1:29]MILWAUKEE.BUCKS
## teamMM1309[, 1:29]MINNESOTA.TIMBERWOLVES
## teamMM1309[, 1:29]NEW.ORLEANS.PELICANS
## teamMM1309[, 1:29]NEW.YORK.KNICKS        **
```

```
## teamMM1309[, 1:29]OKLAHOMA.CITY.THUNDER
## teamMM1309[, 1:29]ORLANDO.MAGIC          ***
## teamMM1309[, 1:29]PHILADELPHIA.76ERS     ***
## teamMM1309[, 1:29]PHOENIX.SUNS           ***
## teamMM1309[, 1:29]PORTLAND.TRAIL.BLAZERS
## teamMM1309[, 1:29]SACRAMENTO.KINGS       *
## teamMM1309[, 1:29]SAN.ANTONIO.SPURS      **
## teamMM1309[, 1:29]TORONTO.RAPTORS
## teamMM1309[, 1:29]UTAH.JAZZ
## xxMA1309$hv                              ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.68 on 1279 degrees of freedom
## Multiple R-squared:  0.2229, Adjusted R-squared:  0.2047
## F-statistic: 12.23 on 30 and 1279 DF,  p-value: < 2.2e-16
```

The key point requiring attention here is that there is one more row added: coefficient of the home-field advantage variable 3.0516, which is presented above. This is consistent with what has been calculated before. If one team plays at home, it tends to have an advantage of about 3 points. It is a significant difference, and this is an important factor when predicting the game results.

In the summary result table, the coefficients of teams have exactly the same meaning as the original model. Compared with coefficients obtained from the original model, they all decrease somewhat. One of the reasons is that a new variable has been introduced and helped to explain the score difference. The residual standard error here decreases a little bit to 12.68, and the Multiple R-square rises to 0.2229. Both of these changes imply that the second model can fit better, and have higher accuracy.

Model 3: Model with 30 teams, home-court-advantages and 6 game statistics

Based on the second model, the advanced purpose is to check the model result when there are more variables. In the exploratory data analysis, there is strong evidence that these six indicators: three points, rebounds, free throws, assists, blocks and turnovers, all have an impact on score differences. The expected result is that each variable would be significant with high correlation and contribute to the new model. However, the outcome is not satisfying.

Firstly, start with constructing the model. To do this, only historical data and information could be used. In other words, the data for the match to be predicted cannot be taken into account. The method used here is to find the data of team 1 and team 2 in the last three matches and average them separately to predict the next match. Just take the rebound variable as an example. If team 1 had 6, 10, 8 rebounds and team 2 had 4, 8, 6 rebounds in their last three games respectively, then the average rebound for team 1 and team 2 would be 8 and 6. Their difference is 2, which is the statistic that will be built into the model. Such a rolling method requires a re-selection of games, since the first several games lack historical data to use. After that, there are 1260 games left in total. Therefore, the matrix used here to build the model has 1260 rows.

After running the model, the result is surprising.

```
##
## Call:
## lm(formula = sdif ~ -1 + teamMM[, 1:29] + xxMA$hv + statMM6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.576  -8.066   0.353   8.077  42.384
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
```

```
## teamMM[, 1:29]ATLANTA.HAWKS          -2.70255    1.88151  -1.436 0.151152
## teamMM[, 1:29]BOSTON.CELTICS          1.59778    2.02347   0.790 0.429902
## teamMM[, 1:29]BROOKLYN.NETS          -7.62127    2.01251  -3.787 0.000160 ***
## teamMM[, 1:29]CHARLOTTE.HORNETS      -0.78989    2.17977  -0.362 0.717136
## teamMM[, 1:29]CHICAGO.BULLS          -2.13170    1.98457  -1.074 0.282975
## teamMM[, 1:29]CLEVELAND.CAVALIERS     4.48537    2.21273   2.027 0.042871 *
## teamMM[, 1:29]DALLAS.MAVERICKS       -3.07254    2.33117  -1.318 0.187743
## teamMM[, 1:29]DENVER.NUGGETS          0.24351    2.07051   0.118 0.906395
## teamMM[, 1:29]DETROIT.PISTONS        -2.25904    2.17821  -1.037 0.299890
## teamMM[, 1:29]GOLDEN.STATE.WARRIORS  12.92145    2.15577   5.994 2.69e-09 ***
## teamMM[, 1:29]HOUSTON.ROCKETS         6.00746    2.20112   2.729 0.006438 **
## teamMM[, 1:29]INDIANA.PACERS         -1.51093    2.01253  -0.751 0.452939
## teamMM[, 1:29]LOS.ANGELES.CLIPPERS    3.53514    2.11142   1.674 0.094328 .
## teamMM[, 1:29]LOS.ANGELES.LAKERS     -7.37836    2.06118  -3.580 0.000357 ***
## teamMM[, 1:29]MEMPHIS.GRIZZLIES       0.47493    2.12124   0.224 0.822879
## teamMM[, 1:29]MIAMI.HEAT             -0.34867    2.04523  -0.170 0.864662
## teamMM[, 1:29]MILWAUKEE.BUCKS        -0.68278    1.97954  -0.345 0.730213
## teamMM[, 1:29]MINNESOTA.TIMBERWOLVES -2.07675    1.99435  -1.041 0.297932
## teamMM[, 1:29]NEW.ORLEANS.PELICANS   -2.45120    2.10669  -1.164 0.244842
## teamMM[, 1:29]NEW.YORK.KNICKS        -4.76486    2.00223  -2.380 0.017475 *
## teamMM[, 1:29]OKLAHOMA.CITY.THUNDER  -0.55917    1.97892  -0.283 0.777559
## teamMM[, 1:29]ORLANDO.MAGIC          -7.11057    2.09955  -3.387 0.000730 ***
## teamMM[, 1:29]PHILADELPHIA.76ERS     -6.46361    2.03533  -3.176 0.001532 **
## teamMM[, 1:29]PHOENIX.SUNS           -6.26377    2.07025  -3.026 0.002533 **
## teamMM[, 1:29]PORTLAND.TRAIL.BLAZERS -1.10479    2.12586  -0.520 0.603372
## teamMM[, 1:29]SACRAMENTO.KINGS       -4.27823    2.03155  -2.106 0.035417 *
## teamMM[, 1:29]SAN.ANTONIO.SPURS       6.25313    2.03047   3.080 0.002119 **
## teamMM[, 1:29]TORONTO.RAPTORS         1.71045    2.15666   0.793 0.427873
## teamMM[, 1:29]UTAH.JAZZ               3.35337    2.06644   1.623 0.104895
## xxMA$hv                               3.08445    0.35735   8.631  < 2e-16 ***
## statMM63Pdif                         -0.20639    0.17393  -1.187 0.235610
## statMM6RBdif                         -0.01307    0.07607  -0.172 0.863622
## statMM6FTdif                          0.03788    0.08531   0.444 0.657097
## statMM6ASTdif                        -0.08705    0.11479  -0.758 0.448393
## statMM6BLKdif                         0.08423    0.21613   0.390 0.696801
## statMM6TOVdif                         0.05328    0.13580   0.392 0.694851
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.67 on 1224 degrees of freedom
## Multiple R-squared:  0.2286, Adjusted R-squared:  0.2059
## F-statistic: 10.08 on 36 and 1224 DF,  p-value: < 2.2e-16
```

On the one hand, the figure shows that none of these 6 variables reveals significance to the model prediction. The coefficients of these 6 variables are all small with high p-values, which could lead to the assumption that the three points, rebounds, free throws, assists, blocks, and turnovers do not affect one game's result a lot. This statement can be supported by using analysis of variance. [The figure]. The sum of squares of the six indicators is 7660, which is much smaller than other variables. This suggests that the new added six indicators does not well explain the score differences. On the other hand, the home advantage coefficient in this model stays nearly the same around 3 with high significance level. The coefficients of 30 teams also change upwards a little bit, with p-values decreasing. It is also due to the reason that new indicators have helped to explain the score difference, leading to less significance of original variables. Currently with this new model, residual standard error is 12.67, and multiple R-squared is 0.2286. Overall, this model has been improved a little bit with these new statistics, since residual standard error falls, and multiple R-squared

rises.

Model 4: Model with 30 teams, home-court-advantages and player injury

As shown in the exploratory data analysis, injuries of players may hugely impact the game result, especially the injury of star players. Since model 3 does not provide significantly more information than model 2, we propose adding player injury information, based on model 2, to predict the score difference.

To measure the impact of injured players, the player stats are used to determine if the player is important. As convention mentioned above to avoid the use of future information, the estimation of player statistics per game in one game is the average statistics per game of the player in the last 3 games. For example, Kawhi Leonard is absent from 2017-05-16 to 2017-05-22.

```
##       Team        Player   GoneDate   BackDate
## 1300 SPURS Kawhi Leonard 2017-05-16 2017-05-22
```

So he will not be able to play the 2017-05-16 game.

```
##             dt                    t1                t2  s1  s2 hv  Team1FG
## 697 2017-05-16 GOLDEN STATE WARRIORS SAN ANTONIO SPURS 136 100  1 39.66667
##     Team1FGA  Team1FG.  Team13P Team13PA  Team13P.  Team1FT Team1FTA Team1FT.
## 697       83 0.4776667 10.33333 28.66667 0.3633333 22.33333 27.66667    0.814
##     Team1ORB Team1TRB Team1AST Team1STL Team1BLK Team1TOV  Team1PF  Team2FG
## 697 10.33333       48 22.33333        8        5 12.33333 23.33333 39.66667
##     Team2FGA  Team2FG. Team23P Team23PA  Team23P. Team2FT Team2FTA  Team2FT.
## 697       84 0.4723333      12 28.33333 0.4206667      18 21.33333 0.8363333
##     Team2ORB Team2TRB Team2AST Team2STL Team2BLK Team2TOV  Team2PF  t1short
## 697 9.666667 45.33333 27.33333        5 7.666667 16.33333 18.33333 WARRIORS
##     t2short
## 697   SPURS
```

The estimation the impact of Leonard's absence from the game is his estimation of statistic if he plays the game. We use the avg statistic of the last theree games as the estimation of his statistic if he plays the 2017-05-16 game.

```
##               player       Date  Tm MP FG FGA   FG. X3P X3PA X3P. FT FTA   FT.
## 15534 Kawhi Leonard 2017-05-07 SAS 30  7  14 0.500   1    4 0.25  1   3 0.333
## 15535 Kawhi Leonard 2017-05-09 SAS 39  8  21 0.381   1    4 0.25  5   6 0.833
## 15536 Kawhi Leonard 2017-05-14 SAS 24  7  13 0.538   1    4 0.25 11  11 1.000
##       ORB DRB TRB AST STL BLK TOV PF PTS GmSc X...
## 15534   2   4   6   4   1   1   3  2  16 11.5  -15
## 15535   3  12  15   4   2   2   2  1  22 19.6   -2
## 15536   4   4   8   3   1   0   1  2  26 25.0   21
```

```
##        MP        FG       FGA       FG.       X3P      X3PA      X3P.        FT
## 31.000000  7.333333 16.000000  0.473000  1.000000  4.000000  0.250000  5.666667
##       FTA       FT.       ORB       DRB       TRB       AST       STL       BLK
##  6.666667  0.722000  3.000000  6.666667  9.666667  3.666667  1.333333  1.000000
##       TOV        PF       PTS      GmSc      X...
##  2.000000  1.666667 21.333333 18.700000  1.333333
```

We sum up the estimation of statistics of all missing players in one game to use as the aggregate impact of many injured players. If no player is injured in one game, the value is 0. We only select the sum of the average minute played per game for the team's injured player as the indicator for the injured players impact on games, and use it in regression. The intuition behind this indicator is that major players tend to stay on the court for longer in every game, whereas for ordinary players, their minutes played are much less. When fitting the model, we take the difference between Team2's injured players' impact indicator minus the Team1's, with the same assumption that the difference of certain characters of the two teams in a game results in the score difference of the game.

As shown in the model summary below, The residual standard error here decreases to 12.59, and the Multiple R-square rises to 0.235. The p-value of the coefficient for the MinutePlayedOfInjuredPlayer is 0.0001, which means this variable has a very convincing relationship with the final score difference. Even though the coefficient is only about 0.05, the MinutePlayedOfInjuredPlayer variable has maximum value of 140, which mean the impact of this variable can be as high as 7 points.

```
##
## Call:
## lm(formula = sdif ~ -1 + teamMM[, 1:29] + xxMA$hv + MinutePlayedOfInjuredPlayer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.508  -7.780   0.591   8.086  43.781
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## teamMM[, 1:29]ATLANTA.HAWKS          -2.911363   1.809305  -1.609 0.107850
## teamMM[, 1:29]BOSTON.CELTICS          0.815999   1.750066   0.466 0.641107
## teamMM[, 1:29]BROOKLYN.NETS          -7.404623   1.912435  -3.872 0.000114 ***
## teamMM[, 1:29]CHARLOTTE.HORNETS      -0.785847   1.911342  -0.411 0.681035
## teamMM[, 1:29]CHICAGO.BULLS          -1.761952   1.861116  -0.947 0.343969
## teamMM[, 1:29]CLEVELAND.CAVALIERS     4.315444   1.846360   2.337 0.019585 *
## teamMM[, 1:29]DALLAS.MAVERICKS       -3.685453   1.927456  -1.912 0.056098 .
## teamMM[, 1:29]DENVER.NUGGETS          0.005319   1.937325   0.003 0.997810
## teamMM[, 1:29]DETROIT.PISTONS        -2.789934   1.911237  -1.460 0.144613
## teamMM[, 1:29]GOLDEN.STATE.WARRIORS  11.715632   1.840490   6.365 2.74e-10 ***
## teamMM[, 1:29]HOUSTON.ROCKETS         5.044875   1.874144   2.692 0.007203 **
## teamMM[, 1:29]INDIANA.PACERS         -1.448532   1.873325  -0.773 0.439528
## teamMM[, 1:29]LOS.ANGELES.CLIPPERS    3.005825   1.890772   1.590 0.112152
## teamMM[, 1:29]LOS.ANGELES.LAKERS     -7.692393   1.925064  -3.996 6.83e-05 ***
## teamMM[, 1:29]MEMPHIS.GRIZZLIES       0.523411   1.921633   0.272 0.785377
## teamMM[, 1:29]MIAMI.HEAT              1.435078   1.971127   0.728 0.466722
## teamMM[, 1:29]MILWAUKEE.BUCKS        -0.589158   1.874226  -0.314 0.753311
## teamMM[, 1:29]MINNESOTA.TIMBERWOLVES -2.773131   1.923294  -1.442 0.149595
## teamMM[, 1:29]NEW.ORLEANS.PELICANS   -2.236048   1.936471  -1.155 0.248437
## teamMM[, 1:29]NEW.YORK.KNICKS        -4.796606   1.894269  -2.532 0.011460 *
## teamMM[, 1:29]OKLAHOMA.CITY.THUNDER  -0.358409   1.897717  -0.189 0.850231
## teamMM[, 1:29]ORLANDO.MAGIC          -7.372143   1.906109  -3.868 0.000116 ***
## teamMM[, 1:29]PHILADELPHIA.76ERS     -6.849000   1.902430  -3.600 0.000331 ***
## teamMM[, 1:29]PHOENIX.SUNS           -6.391973   1.929317  -3.313 0.000950 ***
## teamMM[, 1:29]PORTLAND.TRAIL.BLAZERS -1.725743   1.903385  -0.907 0.364759
## teamMM[, 1:29]SACRAMENTO.KINGS       -4.729525   1.931910  -2.448 0.014500 *
## teamMM[, 1:29]SAN.ANTONIO.SPURS       5.677263   1.858161   3.055 0.002297 **
## teamMM[, 1:29]TORONTO.RAPTORS         1.975382   1.876359   1.053 0.292651
## teamMM[, 1:29]UTAH.JAZZ               3.551257   1.883581   1.885 0.059615 .
## xxMA$hv                               3.007760   0.355339   8.464  < 2e-16 ***
## MinutePlayedOfInjuredPlayer           0.056717   0.014774   3.839 0.000130 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.59 on 1229 degrees of freedom
## Multiple R-squared:  0.235,  Adjusted R-squared:  0.2157
## F-statistic: 12.18 on 31 and 1229 DF,  p-value: < 2.2e-16
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

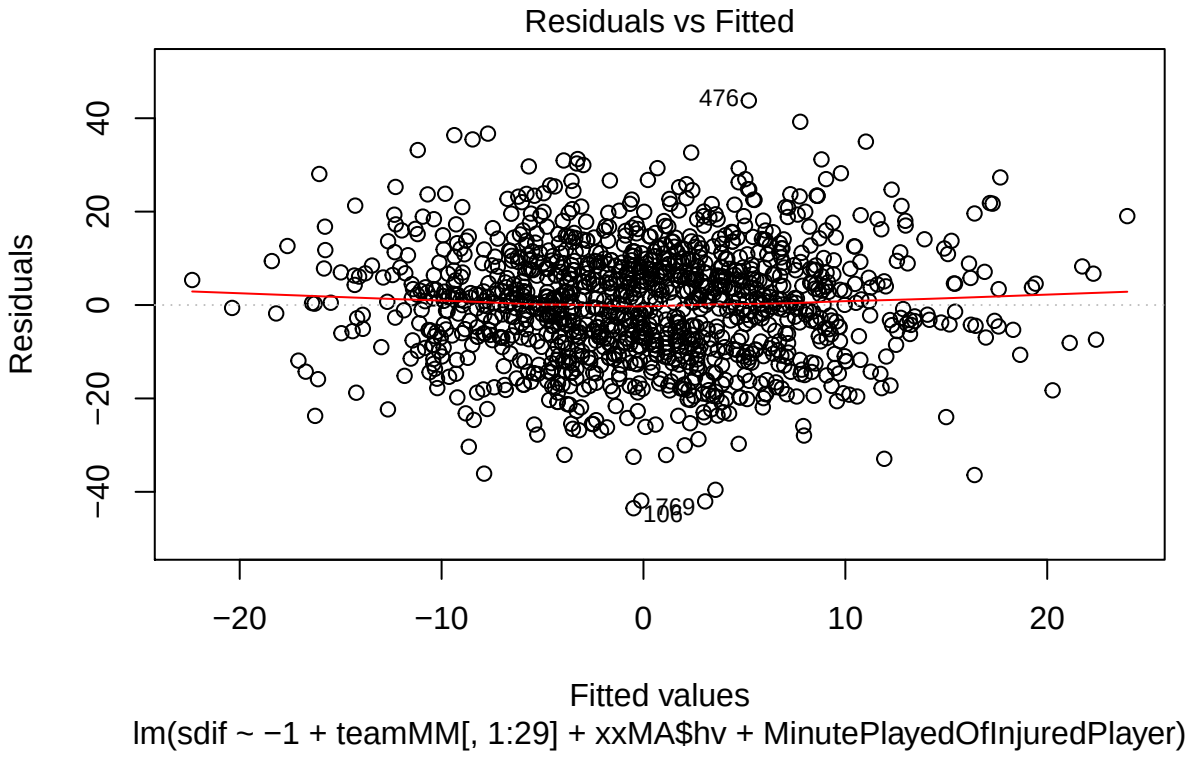
```
## -127.67  -18.00    0.00   -2.79   10.33  140.33
```

Model Checking

1. Regression Diagnostics

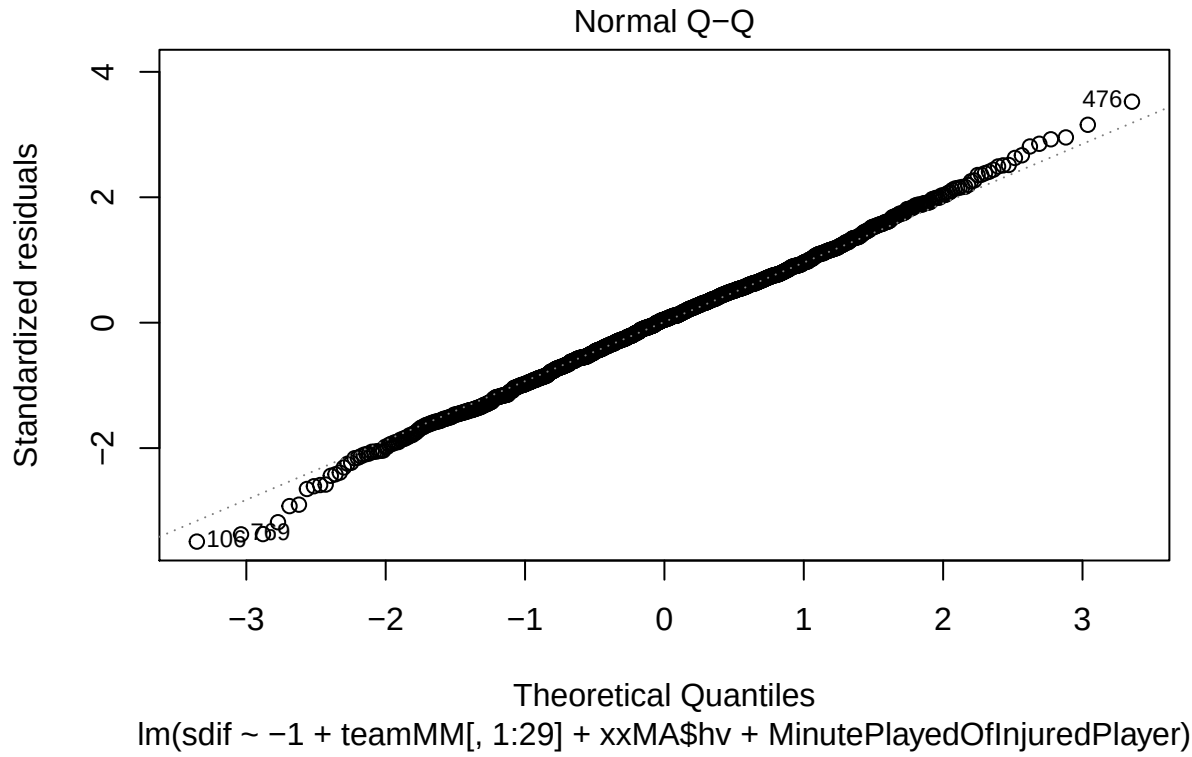
It is necessary to make sure the linear regression model could hold Gauss-Markov assumptions as we try to find the relationship between the variables and the score difference.

a.Assumption of linearity



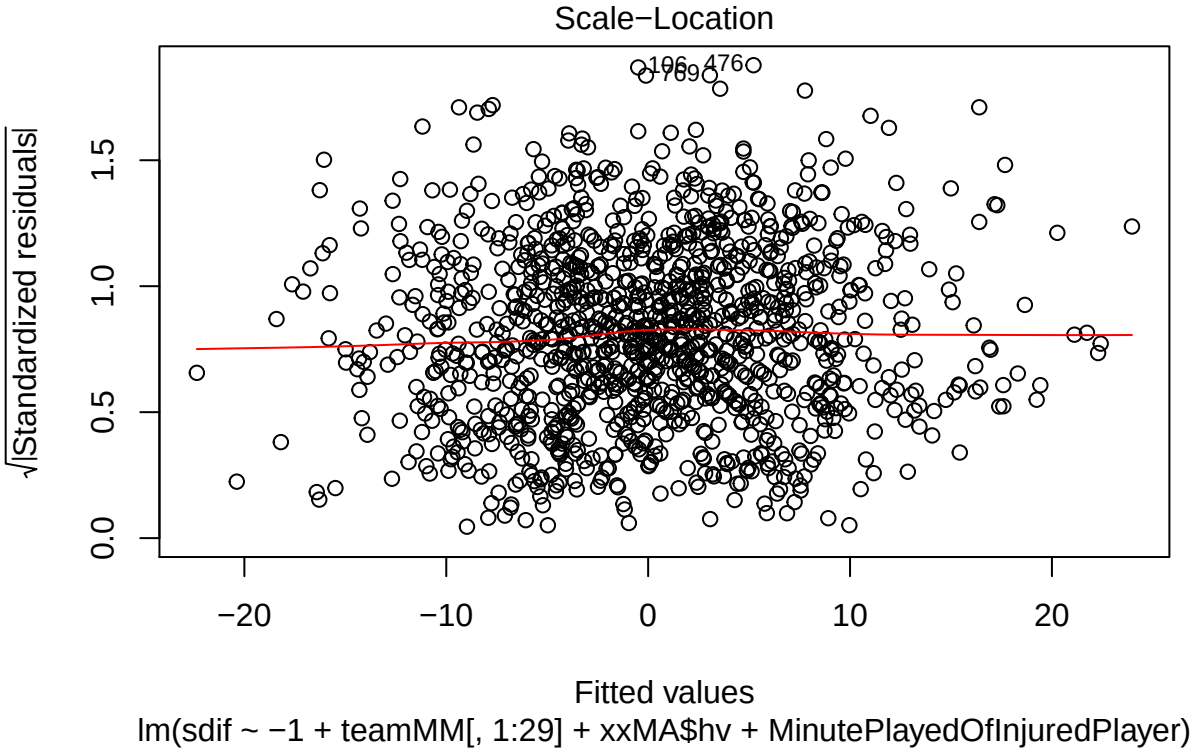
As the scatterplot of the standardized residuals versus numeric covariates shows, the points are scattered randomly around the average line, and obviously there are no significant systematic patterns between residuals and predicted values. Thus, the assumption of linearity is basically reasonable.

b. Assumption of error normality



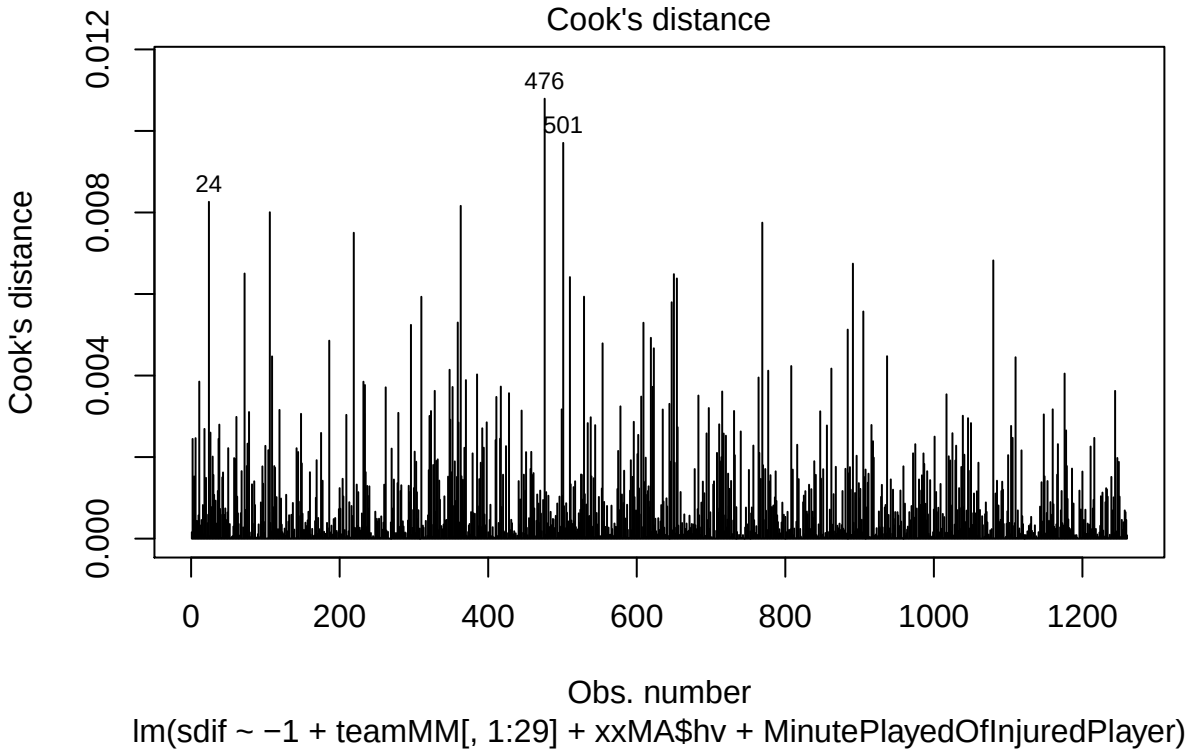
As the Q-Q plot shows, most of the points scattering closely to the reference line, it shows that the distributions of the standardized residual quantiles and theoretical residual quantiles are very similar. Thus, the assumption of error normality is basically reasonable.

c. Assumption of constant error variance



As the Scale-Location graph shows, the points are still randomly distributed around the horizontal line, which indicates the rationality of homoscedasticity.

d.Assumption of independence of errors



The assumption of independence of errors is reasonable because each basketball game happens individually. And as the figure shows, the distribution of residuals has no obvious pattern. Thus, the assumption could be basically approved.

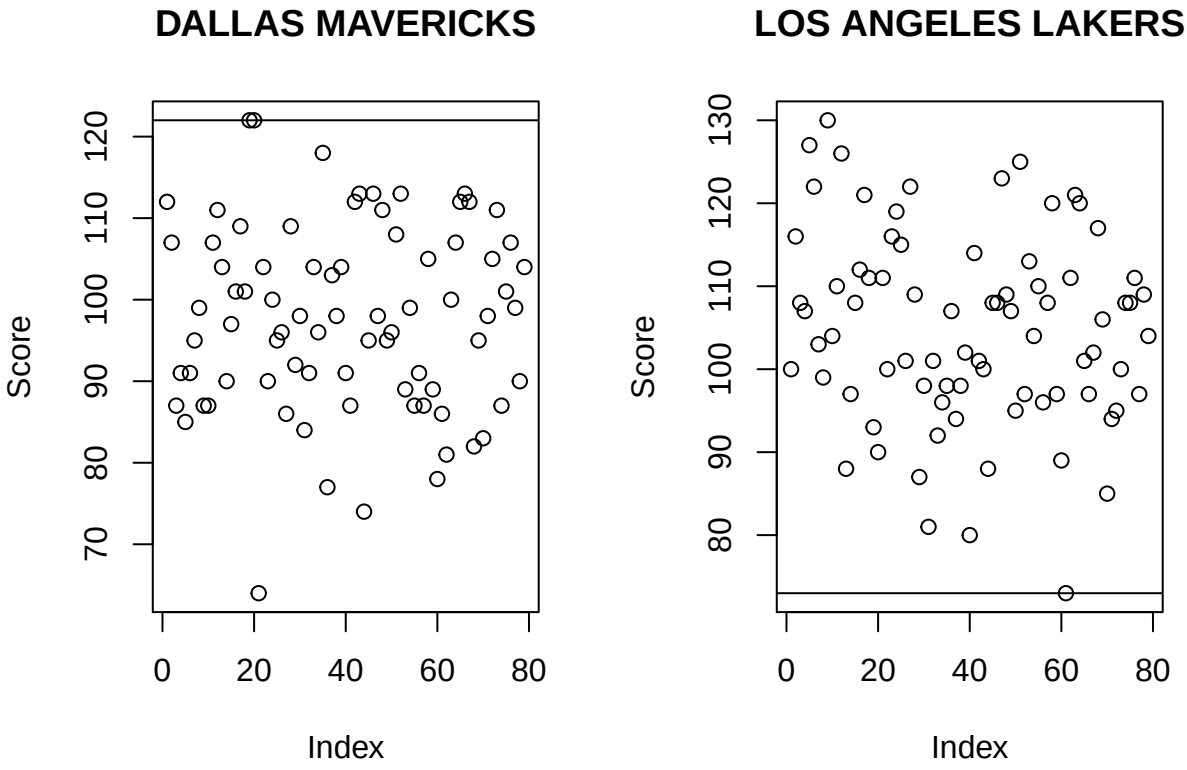
2.Unusual Observation

The plots above and Cook's distance plot shows that #476 has the largest positive residual and the largest cook's distance, which indicates the underestimation of score difference in this game.

```
##             dt               t1                 t2  s1 s2 hv  Team1FG Team1FGA
## 476 2017-01-22 DALLAS MAVERICKS LOS ANGELES LAKERS 122 73  1 37.66667 82.33333
##      Team1FG. Team13P Team13PA  Team13P. Team1FT Team1FTA  Team1FT. Team1ORB
## 476 0.4576667      11       29 0.3766667      14       19 0.7213333 6.666667
##     Team1TRB Team1AST Team1STL Team1BLK Team1TOV  Team1PF  Team2FG Team2FGA
## 476 37.33333 21.66667 4.333333 2.333333       11 22.33333 43.66667 92.33333
##      Team2FG.  Team23P Team23PA  Team23P. Team2FT Team2FTA Team2FT. Team2ORB
## 476 0.4736667 8.666667 23.33333 0.3766667      23 27.66667    0.847 13.33333
##     Team2TRB Team2AST Team2STL Team2BLK Team2TOV Team2PF   t1short t2short
## 476       50 22.66667        9 4.333333 11.66667      19 MAVERICKS  LAKERS
##     t1pmp t2pmp     t1tp t2tp t1aa t2aa ATLANTA.HAWKS BOSTON.CELTICS
## 476     0     0 31.66667    0    0    0             0              0
##     BROOKLYN.NETS CHARLOTTE.HORNETS CHICAGO.BULLS CLEVELAND.CAVALIERS
## 476             0                 0             0                  0
##     DALLAS.MAVERICKS DENVER.NUGGETS DETROIT.PISTONS GOLDEN.STATE.WARRIORS
## 476                1              0               0                     0
##     HOUSTON.ROCKETS INDIANA.PACERS LOS.ANGELES.CLIPPERS LOS.ANGELES.LAKERS
## 476               0              0                    0                 -1
```

```
##     MEMPHIS.GRIZZLIES MIAMI.HEAT MILWAUKEE.BUCKS MINNESOTA.TIMBERWOLVES
## 476                 0          0               0                      0
##     NEW.ORLEANS.PELICANS NEW.YORK.KNICKS OKLAHOMA.CITY.THUNDER ORLANDO.MAGIC
## 476                    0               0                     0             0
##     PHILADELPHIA.76ERS PHOENIX.SUNS PORTLAND.TRAIL.BLAZERS SACRAMENTO.KINGS
## 476                  0           0                      0                 0
##     SAN.ANTONIO.SPURS TORONTO.RAPTORS UTAH.JAZZ WASHINGTON.WIZARDS
## 476                 0               0         0                  0
```

As we can see in this game, 2 teams got 122 and 73 points respectively. When looking back to all games played in this season, it is not hard to figure out that the Mavericks have rarely done so well, and the Lakers have rarely performed so poorly. In goldsheet website, this game was predicted to have a 6 point spread, while the real score difference actually reached 49.



This game surprised many people, but it seems hard to find the reason since there are no abnormal game statistics that influenced the result directly. The sports media speculated that it was the Mavericks' consecutive losses in the first few games that inspired the players to fight, while the Lakers happened to be out of shape. It looks like a pure accident. However, deleting the data of this game did not contribute much to the improvement of the model fit. So the data was kept, and maybe the abnormality will be explained in future work.

4. Discussion & Limitation

By comparing these four models, it can be easily seen that the second model tends to be the core one, and it can be improved by adding technical indicators and injury data. Besides, there are some other key points requiring attention and exploration.

First of all, the major estimation that was used for team statistic indicators is the rolling average of the 3 previous games; there are some other paths of estimation that might be more useful. For instance, using the rolling average of the previous 5 games to run the model would be an interesting idea. However, in that case, games need to be selected again and data volume size will decrease, which might lead to more errors.

Besides, it is preferably not to use data too long ago, since there are many changes between years among teams. For example, there may be some important player transactions or changes of coaches.

In addition, the data used in the model is the direct statistical data of each game, such as the number of three point shots and the number of field goals; however, if the model uses the data that is more comprehensive, like the percentage of this team's field goal rate, three points rate, average scoring points, and average losing points instead of the simple number of goals, surprising things might come out and would make the model result more interesting. One method could be to try some combinations of variables or interactive variables, since nonlinear relationships might work better.

Thirdly, the data that rolls in the model is just the season 2016-2017, this model is computing a single year instead of generalizing multiple times of years' data, its uses could be limited. Moreover, the summary result of the model has an R square around 0.235 and anova residual around 194785. The small R square and big residual value indicates that there are still some places for improvement, and the model is not perfect for now. It could be better by adding other variables.

Lastly, the injury data that was used to build the model only contains the minutes that each player played. Analyzing the importance of a player by the minutes that the player played compared to the team's total players' time could be biased, since the time of one player played in a game could not well represent the competence of him. If the data could include each player's scores, defensive rate, field goals rate, and fouls number will be an advanced approach.

5. Reference

(1)https://www.thelines.com/betting/point-spread/

(2)https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

Project Code: https://github.com/hupidan98/NBA16pred

Conclusion

The main purpose of this study is to predict the final score difference. In summary, many variables have been explored and analyzed, including competence of 30 teams, home-field advantages, three points, rebounds, free throws, assists, blocks, turnovers and injury. In addition, four models have been tried in total, which demonstrate that the outcome of a game has the strongest correlation with the home-field advantage, competence of 30 teams as well as the injury information; on the other hand, team's basic technical statistics show small correlation coefficient. In the future, data could be more completed. For instance, the data of coaches, the trading progress, the draft each team picked, and the current financial situation could all contribute some effects in this model. Moreover, the player injury part could be explored more. They could be shown more visually by adding each player's personal technical statistical data and their ability.