# NBA players' ability analysis

*678 midterm project by Shicheng Wang*

*11/1/2019*

## 1.Abstract

This is a statistical analysis report about NBA situation. The aim of this report is to develop series of methods to measure the comprehensive strengthes of players in NBA union and trying to figure out how each kind of player's abilities could influence his shooting results in the small ball era. Additionally, the project has also analyzed the strategies and preferences of players. Specifically, EDA(Exploratory Data Analysis) and modeling are involved in this report. Overall, readers will get the general idea about how various elements influence players' shooting result.

## 2.Introduction

### 2.1 Background

As NBA steps into a commercialize stage for almost 80 years, it have become a mature business alliance through decades' development, which has always been regarded as a focus and attracting bunches of capitals overseas. It is acknowledged that the players' performances is straightly related to the ratings and could determine the future of a team and related industries, making it is meaningful to conduct a depth analysis for these players' shooting ability.

More importantly, NBA has experienced a huge change since 2014-2015 season. It could be attributed to the rise of three-point shooters leading by Curry and Tompson. Along with Golden State Warriors win the NBA championship finally, it has initiated a new stage for NBA. That is to say, NBA has completed the transformation from old times (Tactical arrangement dominated by tall players inside the three-point line) to small ball era (Dominated by the three points shots outside the line). Hence, player's scoring choice would change since then.

### 2.2 Data sources

There are overall three datasets included in the report. For the EDA part, I will use NBA League Averages(season) which contains averages statistics of all players in NBA in each year to get the general idea of players' condition in the whole association throughout 70 years. This dataset is collected in Basketball Reference. Besides, I also use NBA Season Data(player) for EDA part to analyze players' performance throughout 70 years, and this dataset is collected in Pro Basketball Statistics. The main dataset(da) i used to further analyze is NBA Shot Logs during 2014-2015 season, this dataset detailly record each player's performance at each time period in each game, and this is collected in Kaggle, while they are originally scraped from NBA's REST API. The following are websites of each dataset.

- NBA League Averages - Totals: link
- NBA Season Data: link
- NBA Shot Logs during 2014-2015 season: link

### 2.3 Research question

For the former part of EDA in this project, I would analyze how players' characteristics and how average skill level could change throughout 70 years based on the relavant database. Main research question of this project is to analyze how each players' series of charactistics could influence the actural shooting ability in each single game.

## 3.Methodology

### 3.1 Data process

- Data access

I have imported three datasets. First dataset(named 'team') contains information about whole NBA average data in each year. Second dataset(named 'player') records each player's average data in their career. Third dataset includes every single shot's information took by every player during 2014-2015 season.

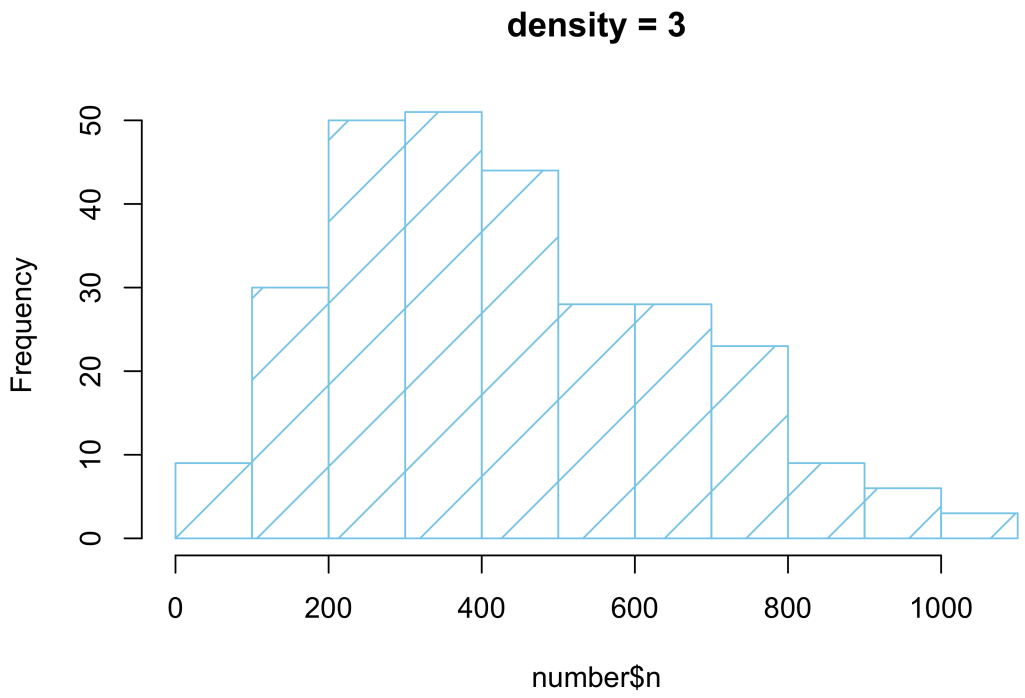
- Data organize

For the season average dataset, clear the NA data in the beginning. Because the row are arranged from big to small, so i reverse the data order by year, then select the columns which are going to be analyzed in EDA part. In order to be easily understood, I rename the variables' name.

For NBA players dataset, drop out players whose data are incompelete. Then i get all players' dataset from 1980 to 2017. This dataset is used for analyzing player's preference and ability of shooting.

For the main dataset, the data cleaning procession has a few steps:

- Step1: cleaning




*Firstly, clearing the NA value, then calculate how many rows does each player have, and draw a barplot based on each player's data amount. Notice that several players have less than 200 rows which could hinder modeling procession, so i drop these players.*

- Step2: adjustment

*After organizing the dataset well, I select specific information and rename the variable name. Additionally, i adjust variables' value which could be beneficial for later analysis.*

- Data overview
  The head of shot log.

Table 1: Overview of shot log

| name | location | winlose | margin | shotnumber | period | shotclock | dribble | touch | range | type | result | c |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| aaron brooks | 1 | 1 | 5 | 10 | 3 | 20.7 | 4 | 3.9 | 24.8 | 3 | 0 | |
| aaron brooks | 1 | 1 | 5 | 11 | 4 | 9.3 | 17 | 13.5 | 7.8 | 2 | 1 | |
| aaron brooks | 1 | 1 | 5 | 12 | 4 | 6.6 | 4 | 3.4 | 4.1 | 2 | 0 | |
| aaron brooks | 1 | 1 | 5 | 13 | 4 | 3.9 | 0 | 1.1 | 22.3 | 3 | 1 | |
| aaron brooks | 1 | 1 | 5 | 14 | 4 | 14.9 | 6 | 5.1 | 7.4 | 2 | 0 | |
| aaron brooks | 1 | 1 | 5 | 1 | 1 | 10.1 | 20 | 13.8 | 0.5 | 2 | 0 | |

Type variables's categories and overview.

| Var1 | Freq |
|---|---|
| A | 58405 |
| H | 58369 |

| Var1 | Freq |
|---|---|
| 1 | 31307 |
| 2 | 28390 |
| 3 | 30024 |
| 4 | 26068 |
| 5 | 797 |
| 6 | 149 |
| 7 | 39 |

| Var1 | Freq |
|---|---|
| 2 | 87071 |
| 3 | 29703 |

| Var1 | Freq |
|---|---|
| made | 53382 |
| missed | 63392 |

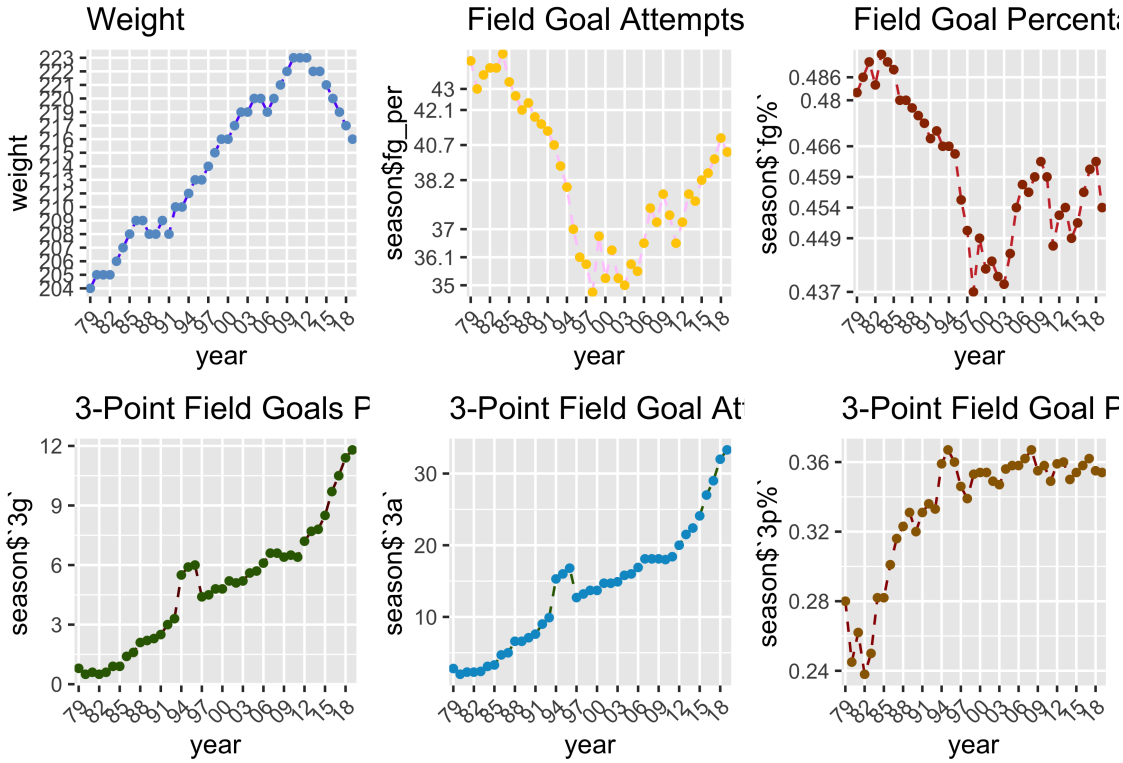
### 3.2 Variables interpretation

- The main data i use is named 'da', the following table shows variables interpretations that are included in the cleaned dataset.

| Variable Name | Interpretation |
|---|---|
| name | player's name |
| location | game home or away |
| winlose | game win or lose |
| margin | game final margin |
| shotnumber | play's shooting number |
| period | player's shoot in which game period |
| shotclock | the rest attacking time when shooting |
| dribble | player's dribble times before shooting |
| touch | player's touching times before shooting |
| range | shooting distance from the board |
| type | 2 points shoot or 3 points shoot |
| defence | the defence player's distance from shooting player |
| result | goal or miss |

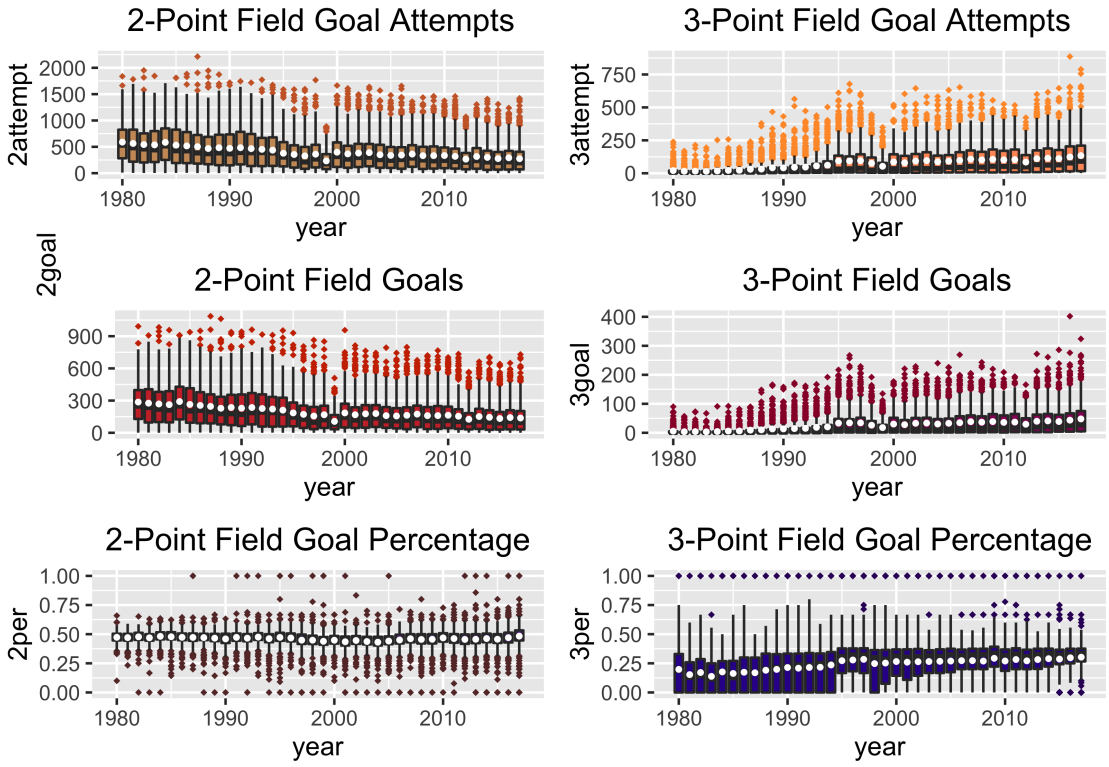
# 4.EDA visualization

## 4.1 Player's shooting average overyear




*For the weight aspect, we could see the weight requirement is incresing, while there is an obvious decrease in 2012, it can be ascribed to the fact that the center position has been dominated the whole union until 2012 and then comes the small ball era, which relies more on light and flexible players to shooting 3-point shots.*
*For the field goal condition, we could see it vibrate overyears and one characteristic is that the big drop 1990-2000, this is because players would like to concentrate more on defencing than offencing.*
*For 3-point field goals, there exists steady increase in recent 30 years and the growth rate increases since 2012 due to the small ball era.*

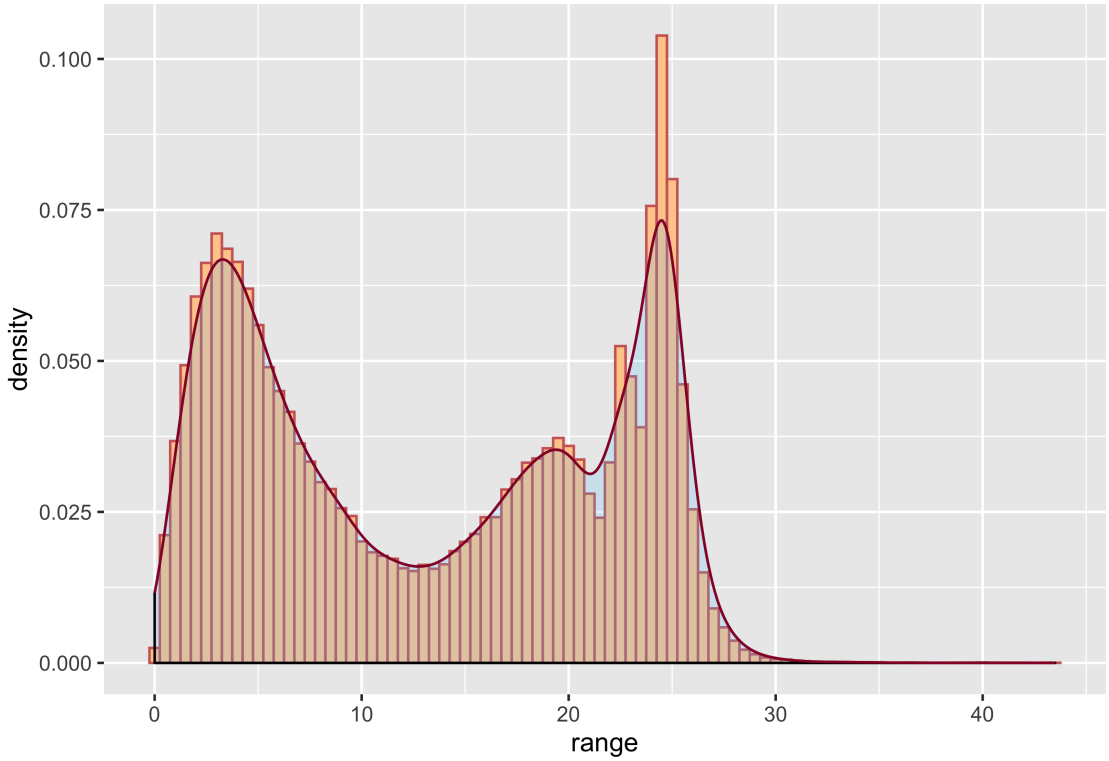
### 4.2 Distribution of players' 2-point and 3-point ability overyear



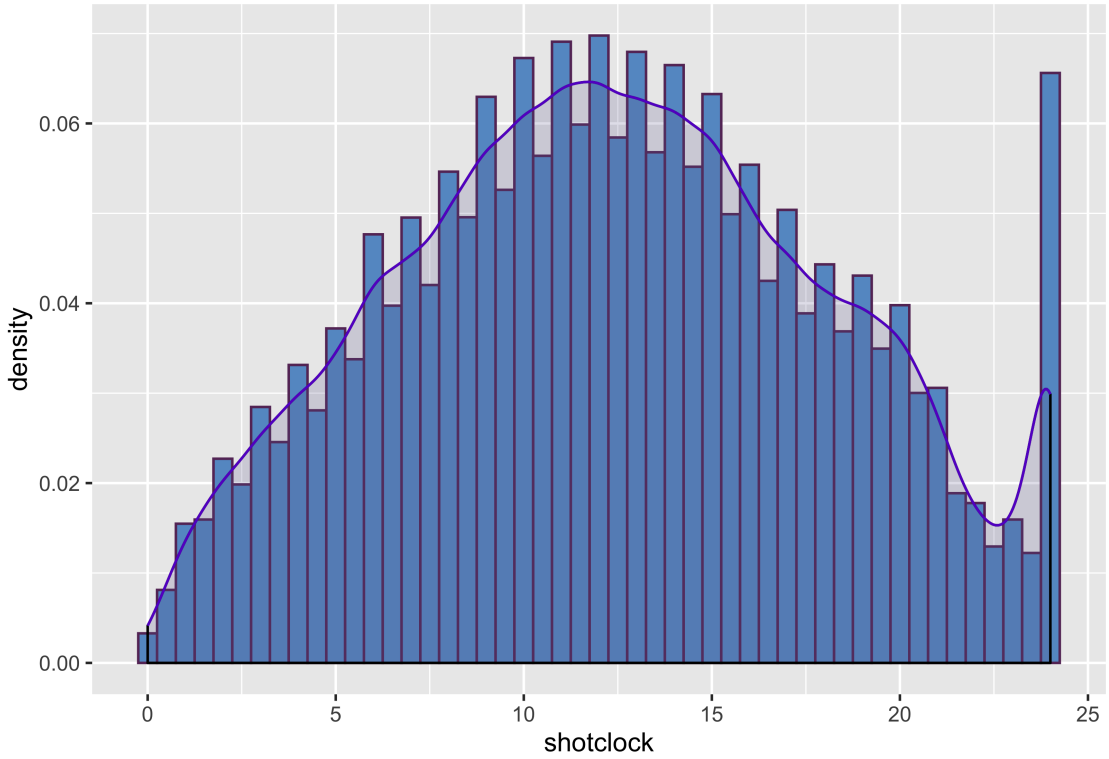

*Of all the players in the NBA from 1980 to 2017, their shooting trends are obvious. In the comparasion between 2-point field goals and 3-point field goal, 2-point field goal attempts/goals have been decreasing while 3-point field goal attempts/goals have been increasing. In the meantime, the 2-point field goal percentage keep developing steady, while the 3-point shooting quality increases to an upper level.*

### 4.3 Shooting preference 2014-2015 season.

- shoot type preference:
  It is well acknowledged that 2014-2015 season could be regarded as the beginning of the small ball era.




*From the first plot we could see player's main scoring distance distribute located in 0-5 meters and 22-25 meters area which indicates players nowadays prefer layup, dunk or throw 3-point balls during the game*

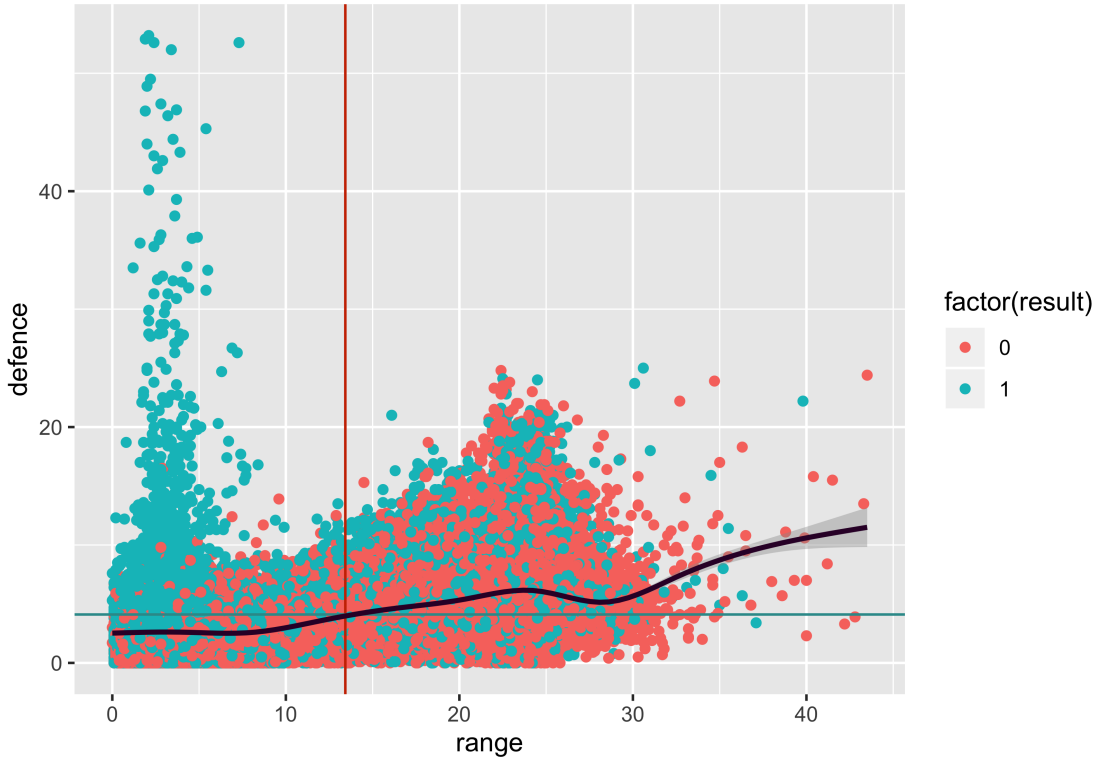



*From the second plot, we could see majority of players prefer more organized attacks, that is to say they are*

*not willing to hurry up shoots or shot until the last second.*

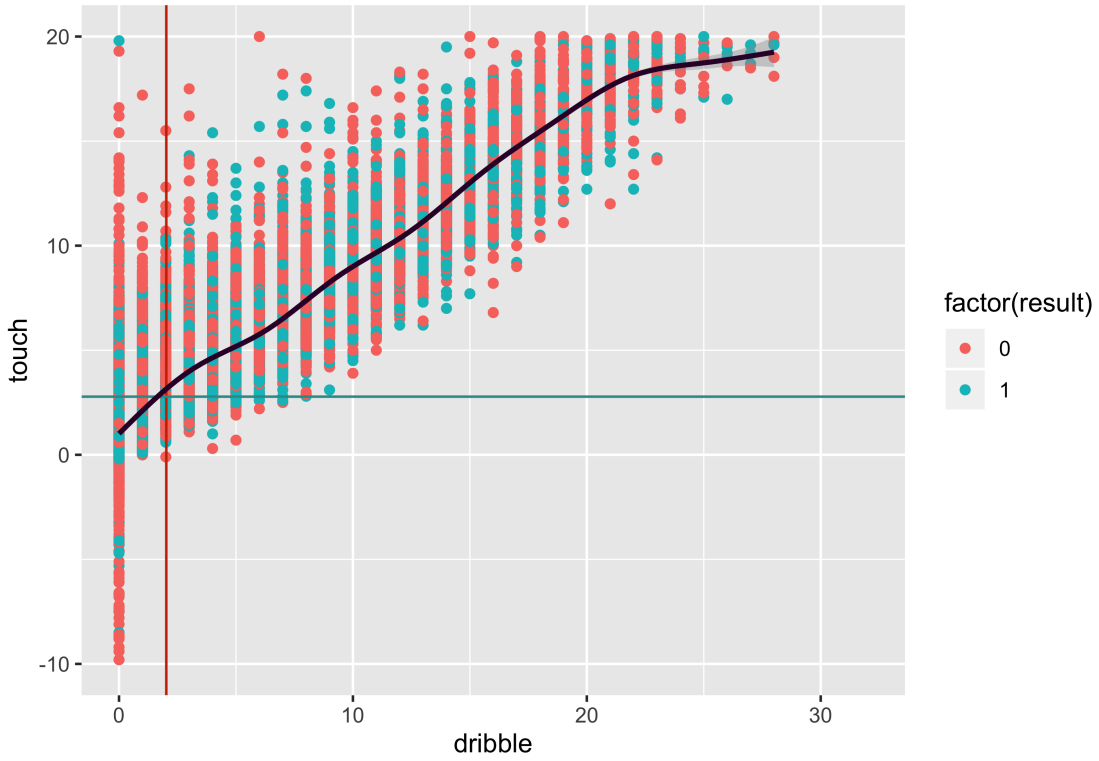
## 4.4 How variables influence player's shoot result.

**Range and defence classfied by result**



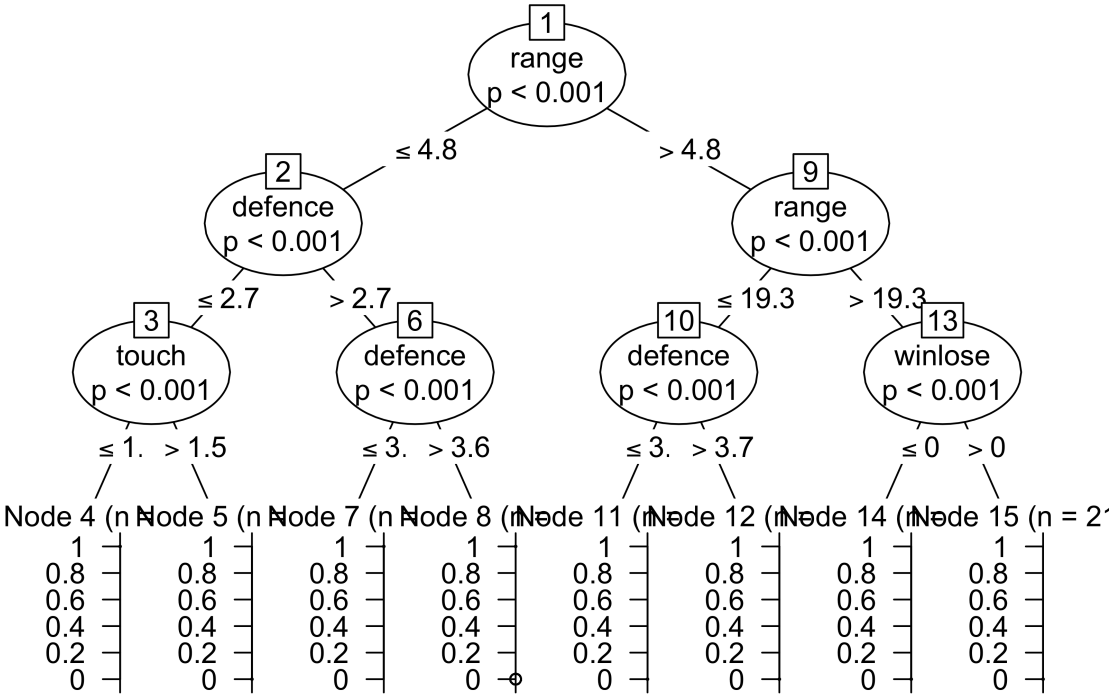

*In this plot we can discover the relationships between defence and range categorized by result. It turn out that players could make more shots in 0-9 meters and 25 meters around range. As the range rise, the defence range could correspondlingly rise because defender believes it is difficult to make a shot at far range.*

**Dribble and touch classfied by result**



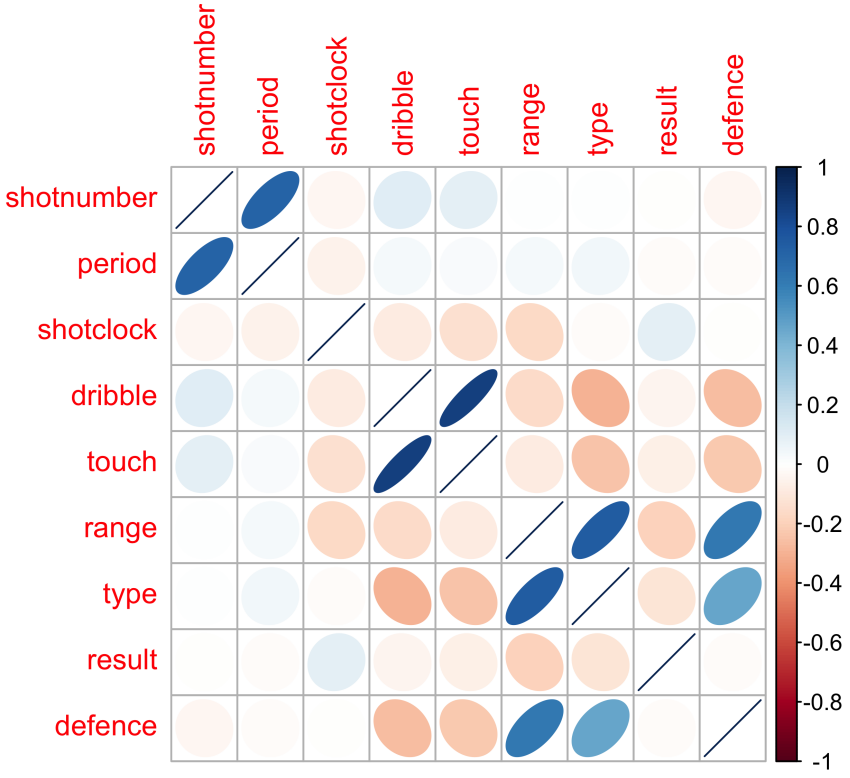

*As we can see, if a players dribble more, they would exert more touching time. While, in this plot we could not see an obvious difference that more dirbbles and touch time could bring to higher hit rate.*

### 4.5 C-tree analysis




*From the C-tree, we could see the decisive relationship between variables. Defence distance could related to touch time and shooting range. Shooting range which largely influence the other variables and might directly related to game result.*

### 4.6 Corelation Test




*Corelation plot shows there exists obvious relationships ammong touch, dribbles, distance, range, shotclock, type, and shooting result.*

**After i have done EDA part, I believe the players shooting preference has changed so much in recent years that i decide to analyze how their shooting results are influenced by other factor. Based on NBA shot log dataset, I pick some of variables to build the model.**

# 5.Model analysis

## 5.1 General linear model

- Building model
  The first model i use is general linear model, cause output of the model is bineary distribution, I use the glm function to build model. Model shows below:

```
glm(formula = result ~ location + shotclock + shotnumber + period +
    dribble + touch + range + defence + type + as.factor(name),
    family = binomial, data = da)
```

| | coef.est | coef.se |
|---|---|---|
| (Intercept) | 0.01 | 0.11 |
| location | 0.04 | 0.01 |
| shotclock | 0.01 | 0.00 |
| shotnumber | 0.00 | 0.00 |
| period | 0.00 | 0.01 |
| dribble | 0.02 | 0.01 |
| touch | -0.06 | 0.01 |
| range | -0.07 | 0.00 |
| defence | 0.11 | 0.00 |
| type | 0.11 | 0.02 |
| as.factor(name)al farouq aminu | -0.35 | 0.16 |
| as.factor(name)al horford | 0.18 | 0.12 |
| as.factor(name)al jefferson | -0.01 | 0.12 |

*There are 242 players in my dataset, each player has their own constructed function.Picture above shows part*

*of the model summary.*

- Interpretation
  In this model, the majority of the coefficient are significant. Choose my favorite player called Damian Lillard whose coefficient estimate is 0.10, so his predicting function shows below:

$$logit(result) = 0.01 + 0.04location + 0.01shotclock + 0.02dribble - 0.06touch - 0.07range + 0.11defence + 0.11type + 0.1(player)$$

*In this model, the majority of the coefficient are significant. But the according to the binned residual plot, there exists an obvious pattern, and the outliers are pretty large, hence this model needs improvement.*

*Intercept: The log odds of shooting percetage of a player is 0.01 when he is playing away with no touch/dribble and no shoot intentation.*

*The coefficient of location: Gaming home would has 0.04 log odds of shooting percentage more than gaming away for players with all other variables the same.*

*The coefficient of type: Shooting 3-point field goal would has 0.11 log odds of shooting percentage more than shooting 2-point field goal with all other variables the same*

*The coefficient of shotclock: With every 1 level increase in shotclock level, the expected log odds of shooting percentage for specific player would increase by 0.04 unit with all other variables the same.*

*The coefficient of dribble: With every 1 level increase in dribble level, the expected log odds of shooting percentage for specific player would increase by 0.02 unit with all other variables the same.*

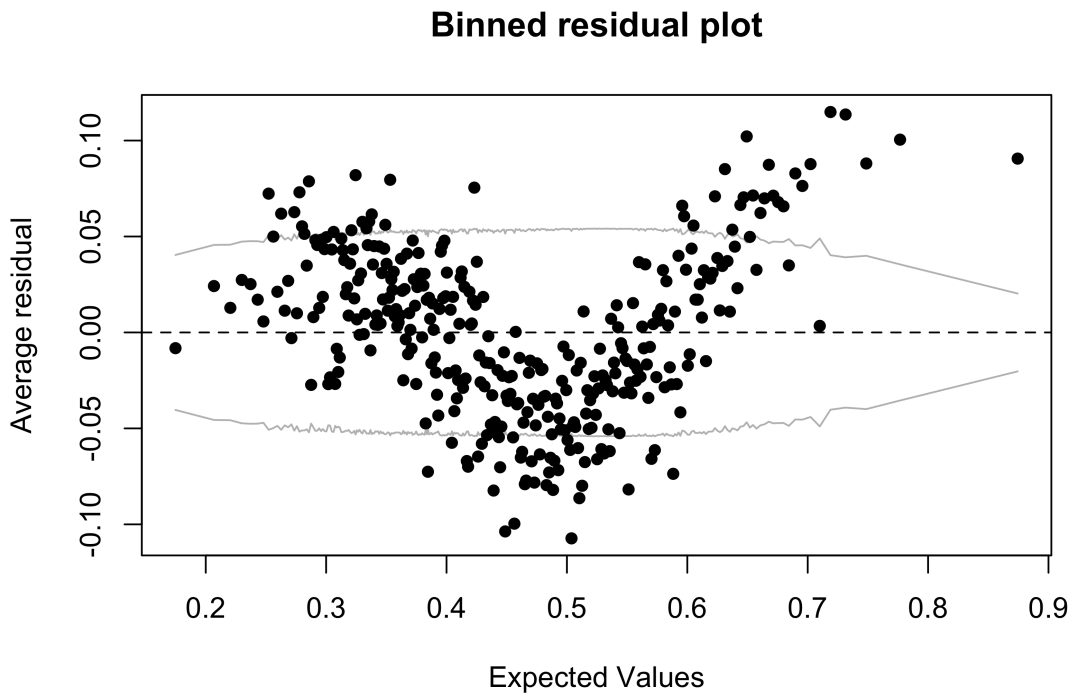
*The coefficient of touch: With every 1 level increase in touch level, the expected log odds of shooting percentage for specific player would decrease by 0.06 unit with all other variables the same.*

*The coefficient of range: With every 1 level increase in range level, the expected log odds of shooting percentage for specific player would decrease by 0.07 unit with all other variables the same.*

*The coefficient of defence: With every 1 level increase in defence level, the expected log odds of shooting percentage for specific player would increase by 0.11 unit with all other variables the same.*

*The coefficient of player: With each player, his log odds of shooting percentage would corespond with their own coefficient in the model summary.*

- Model check I draw the binned plots to check if the model fits well.




*In this model, the majority of the coefficient are significant. But the according to the binned residual plot, there exists an obvious pattern(like a character V). What's more the outliers are pretty large, showing that this model does not fit very well. Hence this model needs improvement.*

## 5.2 Bayesian generlized linear model

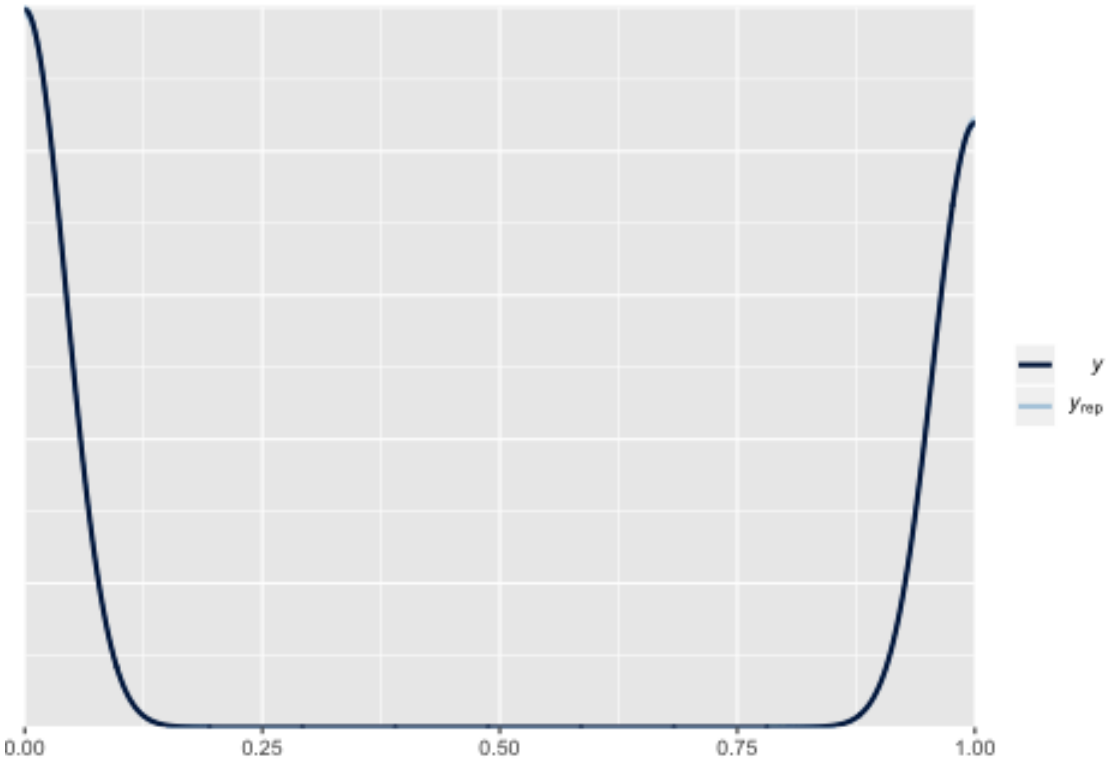
The second model i construct is bayesian generlized linear model with proper transformation and interaction, model shows below:

```
Model Info:
 function:     stan_glm
 family:       binomial [logit]
 formula:      result ~ shotnumber + I(shotclock^2) + period + dribble + touch +
    touch * dribble + type * range + I(range^2) + I(defence^2)
 algorithm:    sampling
 sample:       4000 (posterior sample size)
 priors:       see help('prior_summary')
 observations: 116774
 predictors:   12

Estimates:
                  mean   sd   10%   50%   90%
(Intercept)    -2.7    0.4 -3.3  -2.7  -2.1
shotnumber      0.0    0.0  0.0   0.0   0.0
I(shotclock^2)  0.0    0.0  0.0   0.0   0.0
period          0.0    0.0  0.0   0.0   0.0
dribble         0.0    0.0  0.0   0.0   0.0
touch          -0.1    0.0 -0.1  -0.1  -0.1
type            1.7    0.2  1.4   1.7   2.0
range           0.0    0.0  0.0   0.0   0.1
I(range^2)      0.0    0.0  0.0   0.0   0.0
I(defence^2)    0.0    0.0  0.0   0.0   0.0
dribble:touch   0.0    0.0  0.0   0.0   0.0
type:range     -0.1    0.0 -0.1  -0.1  -0.1
```

*Using the binomial family for bayesian general linear model and do the MCMC diagnostics. From the summary, we can see the sample's mean/standard devation and the standard error. The next step to examine this model is to see the simulate y-output of this model*

- Model check I use pp-check function to check the distribution of fitting output.




*The simulation result show the output of my model is quite corespond with the binomial observation which is 0 or 1, indicating the model fits quite well.*
*According to the EDA part and the formoer models, I found that there could exist interaction among variables. Moreover, different players could leads to different shooting range and defending strategies. So the next following model i will introduce is mutilevel model.*

## 5.3 Mutilevel model

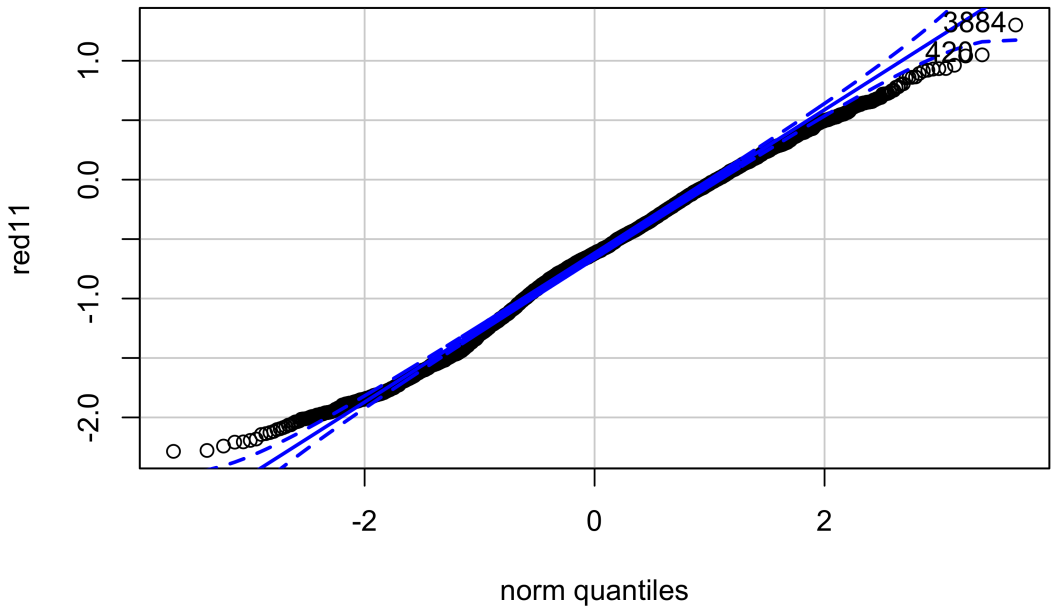
### Random intercept

First model i apply is mutilevel level with random intercept. In this model, different player would has different intercept.

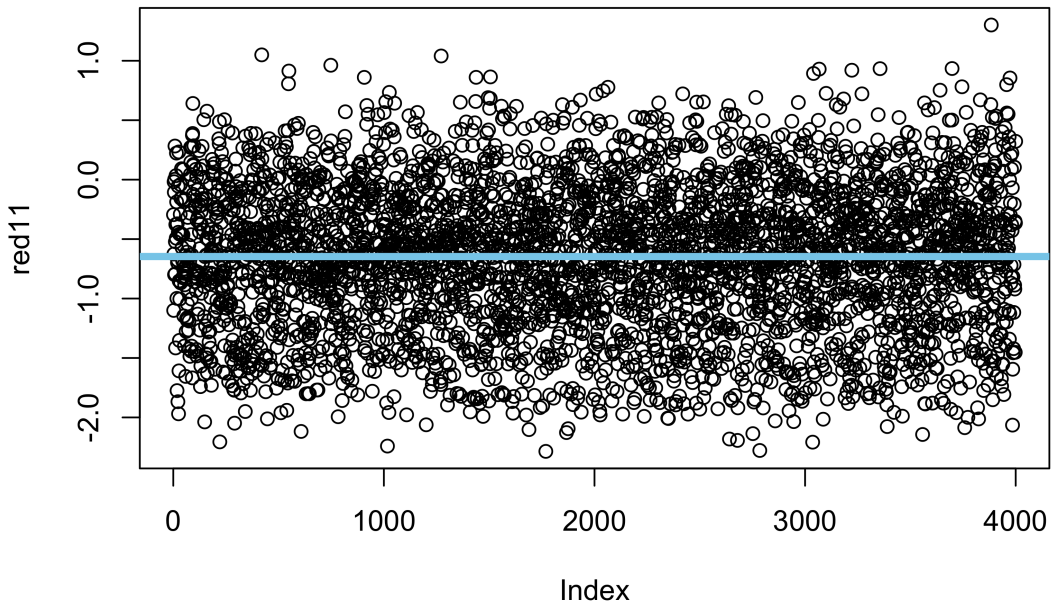
```
## glmer(formula = result ~ location + shotnumber + period + shotclock +
##     dribble + touch + type + range + defence + dribble * touch +
##     type * range + (1 | name), data = da, family = binomial(),
##     REML = T)
##               coef.est coef.se
## (Intercept)    0.75     0.41
## location       0.03     0.01
## shotnumber     0.00     0.00
## period        -0.01     0.01
## shotclock      0.01     0.00
## dribble       -0.01     0.01
## touch         -0.07     0.01
## type          -0.27     0.20
## range         -0.09     0.02
## defence        0.10     0.00
## dribble:touch  0.00     0.00
## type:range     0.01     0.01
```

```
##
## Error terms:
##  Groups   Name        Std.Dev.
##  name     (Intercept) 0.13
##  Residual             1.00
## ---
## number of obs: 116774, groups: name, 242
## AIC = 154522, DIC = 153704
## deviance = 154099.9
```

*I set players name as pooling principle, so i get nearly 250 groups. From summary of m1 we can see the most coefficients's p-value are quite small and they are significant, indicting the varibales could influence player's shooting result, there might exist mix effect in this model. From the result, we can see the variance is 0.01638, the random effect exists among different players. The AIC and BIC are 154521.8 and 154647.4. Additionally, there exists corelationships between these variables.*




```
## [1] 3884  420
```




```
## integer(0)
```

*The residual plot shows the residuals are randomly distributed but mean value also below 0 which could be regarded as a flaw.*

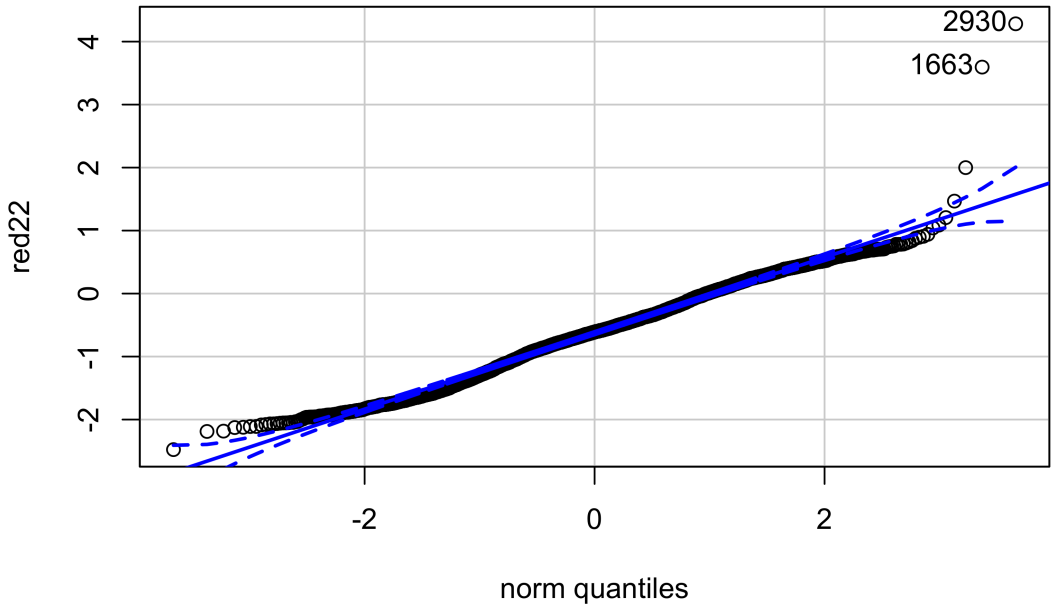
### Random slope

Second model i apply is mutilevel level with random slope. In this model, different player would has different range slope.

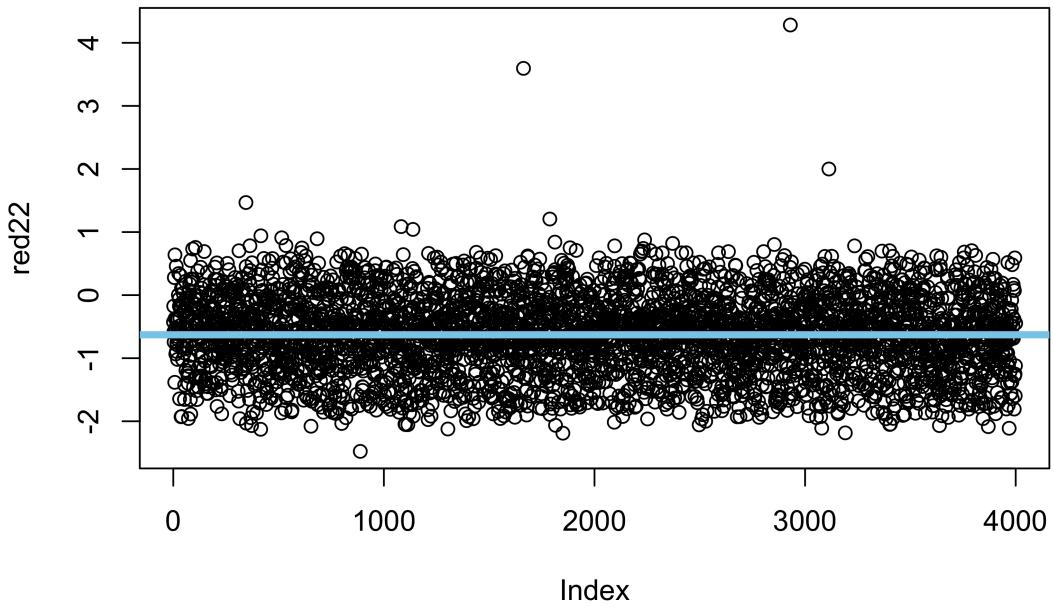
```
## glmer(formula = result ~ location + shotnumber + shotclock +
##     dribble + touch + type + range + defence + dribble * touch +
##     type * range + (touch - 1 | name), data = da, family = binomial(),
##     REML = T)
##                coef.est coef.se
## (Intercept)     0.69     0.40
## location        0.03     0.01
## shotnumber      0.00     0.00
## shotclock       0.01     0.00
## dribble        -0.01     0.01
## touch          -0.07     0.01
## type           -0.24     0.20
## range          -0.09     0.02
## defence         0.10     0.00
## dribble:touch   0.00     0.00
## type:range      0.01     0.01
##
## Error terms:
##  Groups   Name  Std.Dev.
##  name     touch 0.03
```

```
##  Residual       1.00
## ---
## number of obs: 116774, groups: name, 242
## AIC = 154590, DIC = 153917.2
## deviance = 154241.7
```

*This time, two interaction variables are introduced in the model and i define a random slope mutilevel model. The slope of the range could vary from players.From model2 we can see the most coefficients's p-value are quite small and they are significant, the AIC and BIC are 154590.2 and 154706.2. We could see the range preference could be different because some players(like Lilard) prefer to shoot long distance ball. Additionally, random effect exist between players and shooting range.*




```
## [1] 2930 1663
```




```
## integer(0)
```

*The residual plot shows the residuals are randomly distributed but mean value below 0. The qq-plot indicate the residual do not has more information but the mean still below 0. The model fits not so well in this way.*

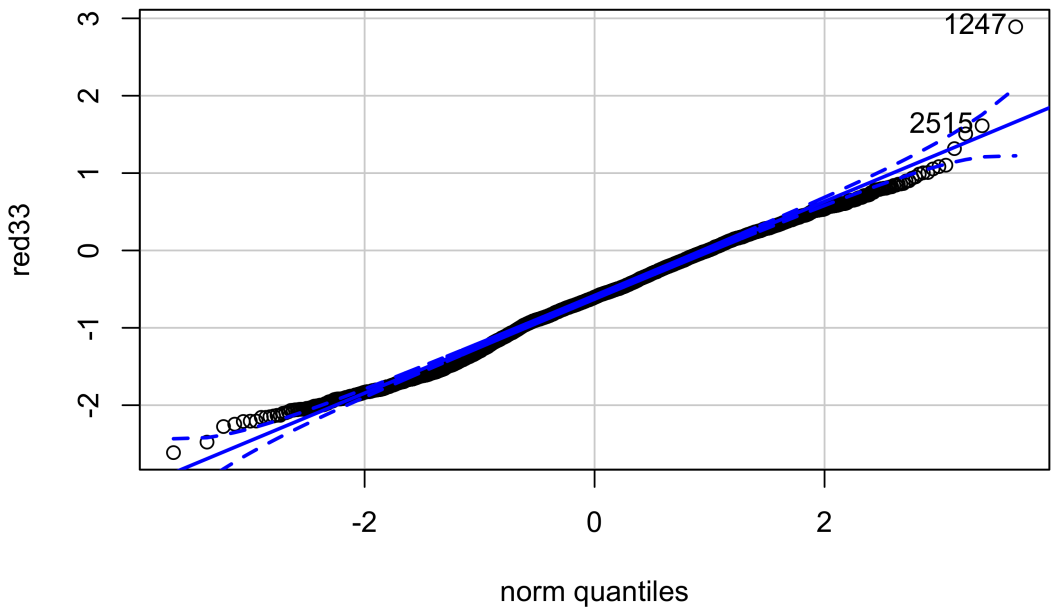
### Random slope and random intercept

Last model i apply is mutilevel level with random intercept and slope. In this model, different player would has different intercept and different range slope.

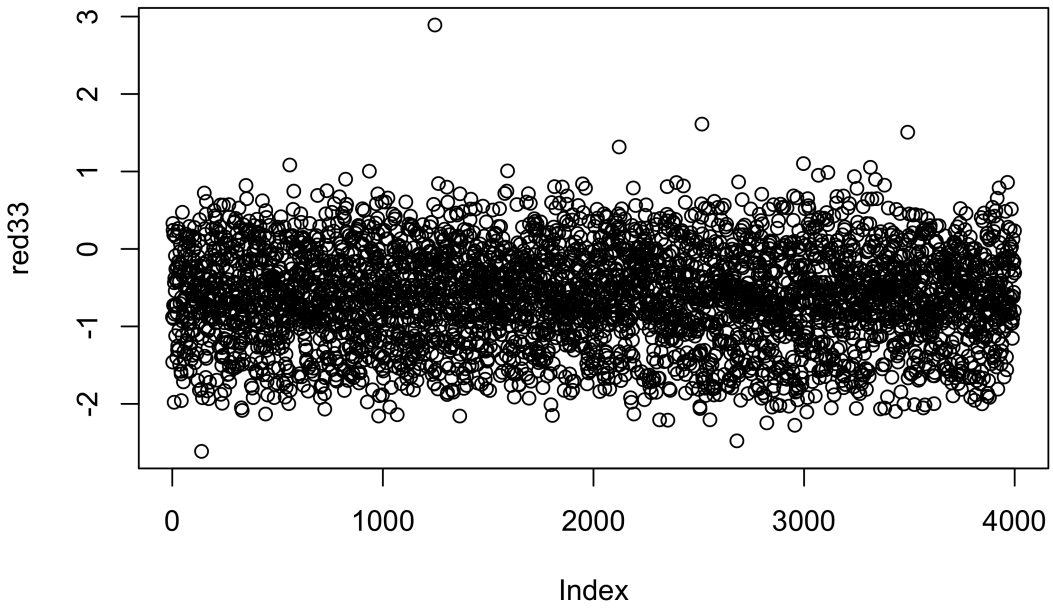
```
## glmer(formula = result ~ location + shotnumber + shotclock +
##     dribble + touch + type + range + defence + dribble * touch +
##     type * range + (1 + range | name), data = da, family = binomial(),
##     REML = T)
##                coef.est coef.se
## (Intercept)     0.40     0.41
## location        0.03     0.01
## shotnumber      0.00     0.00
## shotclock       0.01     0.00
## dribble        -0.01     0.01
## touch          -0.07     0.01
## type           -0.11     0.20
## range          -0.08     0.02
## defence         0.11     0.00
## dribble:touch   0.00     0.00
## type:range      0.01     0.01
##
## Error terms:
##  Groups   Name        Std.Dev. Corr
##  name     (Intercept) 0.19
```

```
##           range       0.01     -0.78
##  Residual             1.00
## ---
## number of obs: 116774, groups: name, 242
## AIC = 154428, DIC = 153023.5
## deviance = 153711.9
```

*From model3 we can see the most coefficients's p-value are quite small and they are significant. I set players name as pooling principle,and the slope of range varys form different players, the intercept could vary for different players. We could see the range preference could be different because some players(like Lilard) prefer to shoot long distance ball. Additionally, there exists corelationships between these variables.*




```
## [1] 1247 2515
```




*The qq-plot and residual plot shows the residuals are randomly normal distributed, while the mean of residual turn to be -1 which henders model's effectiveness.*

## 5.4 Model evaluation and selection

In order to measure which kind of mutilevel model is best fitted my dataset, the anova test is conducted.

```
## Data: da
## Models:
## m2: result ~ location + shotnumber + shotclock + dribble + touch +
## m2:     type + range + defence + dribble * touch + type * range +
## m2:     (touch - 1 | name)
## m1: result ~ location + shotnumber + period + shotclock + dribble +
## m1:     touch + type + range + defence + dribble * touch + type *
## m1:     range + (1 | name)
## m3: result ~ location + shotnumber + shotclock + dribble + touch +
## m3:     type + range + defence + dribble * touch + type * range +
## m3:     (1 + range | name)
##    Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## m2 12 154590 154706 -77283   154566
## m1 13 154522 154647 -77248   154496 70.422      1  < 2.2e-16 ***
## m3 14 154428 154564 -77200   154400 95.385      1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*From the anova test we can see the difference between three model are slight. Nonetheless, according to AIC and BIC, the model3 is quite small so we can regard this as a better model. Hence, model 3 has more perfect performance in predicting shooting result.*

# 6.Result and discussion

## 6.1 Result

- The analysis i present in this report does good to predict a player's shooting performance during the game. From the glm model we could see a player could be influenced by various elements, such as shooting distance, defending distance, touch time, dribble number, and the like. After we pooling the players issue, we could see there exists slightly relationship between players and shooting distance/touch time/dribbles number.

## 6.2 Limitation

- We all know the internal relations, mood, and other potential factors might impact the shooting result, these factors are differ from places and difficult to track.

- Regular season is quite different with the off season. Some team probably play less agressive in regular season because the risk of losing a single game is low. But when it comes to the play offs, all teams would spare no effort to chasing the championship, so some strategies might change.

## 6.3 Improvement

- The defensive strength could vary from teams, some teams has many tall players(LA Lakers), so teams are consist of low players(Portland Trail Blazers). Some information about opponents could largely improve the analysis quality.

- Players are easily to get injured during games and might absent several games if injured. They need time to go back to former condition after recovery. If more information could be collected, the analysis would be more accurate.

- NBA has the trading market machanism before a new season start, so each team's player could vary through years, how to organize each player's data needs to think again.

### 6.4 Future direction

- Some other method could be added into this analysis such as statistical learning to make this analysis more strengthful. Additionally, this analysis could be used into sports field(not limited to NBA), NFL, FIFA, hockey, and the like.

# 7.Reference website


- https://www.kaggle.com/

- http://www.stat-nba.com/

- https://www.basketball-reference.com/

- https://www.hupu.com/

- https://blog.csdn.net/

# 8.Appendix

## shooting period preference