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 Importing three datasets.

```
team <- read.csv("/Applications/BU/BU/678 Applied Statistical Modeling/MA678 midterm project/678 midterm
player <- read.csv("/Applications/BU/BU/678 Applied Statistical Modeling/MA678 midterm project/678 midt
data <- read.csv("/Applications/BU/BU/678 Applied Statistical Modeling/MA678 midterm project/678 midterm
```

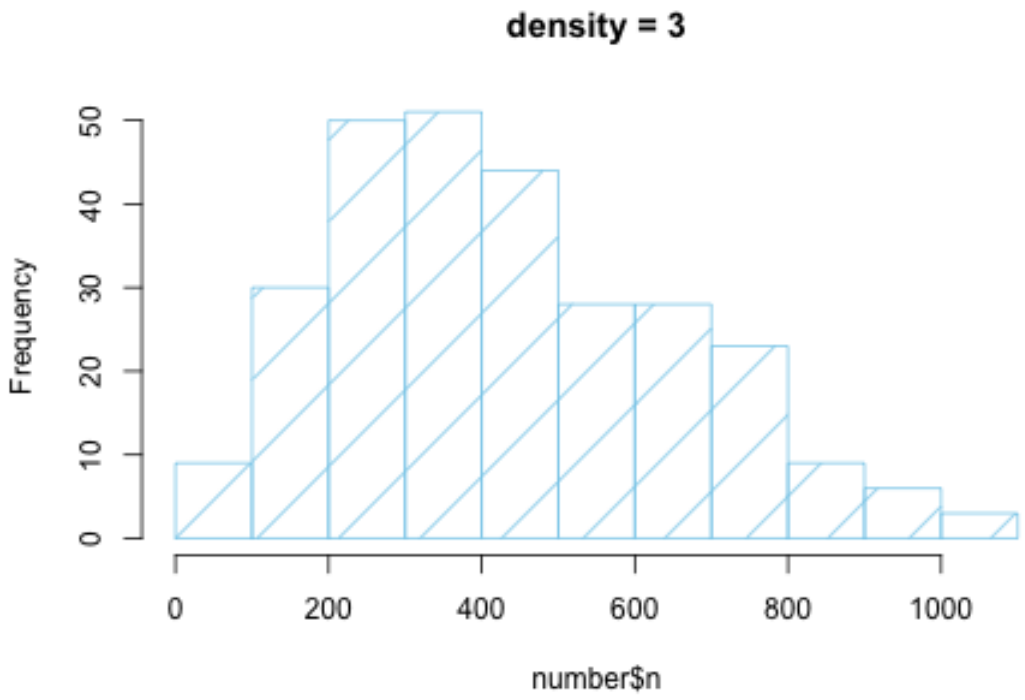
- Data organize

For the season average dataset, clear the NA data first and then reverse the data order by year, then select the columns which are going to be analyzed. 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 shows as below: * Step1: cleaning

```
### players data amount
hist(number$n, density = 3, col = "skyblue", main  = "density = 3")
```



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 the modeling procession, so i drop these players.

- Step2: adjustment

```
### select useful column
data <- dplyr::select(data,-n)
da <- data[,c(1,4,5,6,7,8,10:15,18)]

### rename variables' names
library(reshape)
```

```
##
## Attaching package: 'reshape'

## The following object is masked from 'package:dplyr':
##
##     rename

## The following objects are masked from 'package:tidyr':
##
##     expand, smiths

## The following object is masked from 'package:data.table':
##
##     melt

## The following object is masked from 'package:Matrix':
##
##     expand
```

```
da <- plyr::rename(da,c(player_name="name",LOCATION="location",W="winlose",FINAL_MARGIN="margin",SHOT_NU

### adjust variables
da$location=as.numeric(da$location)-1    # 1 for home, 0 for away
da$winlose=as.numeric(da$winlose)-1      # 1 for win, 0 for lose
da$result=2-as.numeric(da$result)        # 1 for made, 0 for miss
da$type=as.numeric(da$type)
```

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.

```
kable(table(data$LOCATION))
```

Var1	Freq
A	58405
H	58369

```
kable(table(data$PERIOD))
```

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

```
kable(table(data$PTS_TYPE))
```

Var1	Freq
2	87071
3	29703

```
kable(table(data$SHOT_RESULT))
```

Var1	Freq
made	53382
missed	63392

```
par(mfrow=c(2,2))
```

3.2 Variables interpretation

- The main data i use is called '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

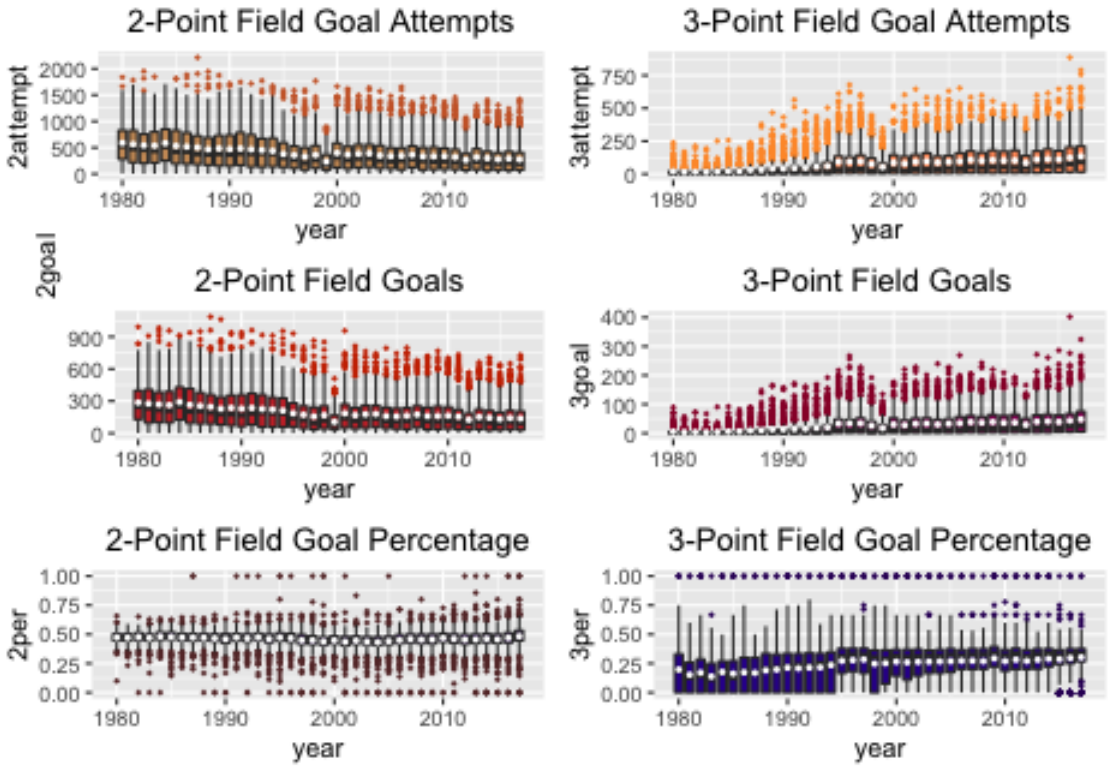
4.1 Player's shooting average overyear

For the weight aspect, we could see the weight requirement is incresing but there is an obvious decrease in 2012, it can be ascribed to the fact that the center dominate whole union until 2012 and then comes the small ball era, which rely 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

```
### 2-Point Field Goal Attempts
p7<-ggplot(players,aes(x=year,y=`2attempt`,group=year))+geom_boxplot(fill='#CC9966',outlier.colour="#CC
### 3-Point Field Goal Attempts
p8<-ggplot(players,aes(x=year,y=`3attempt`,group=year))+geom_boxplot(fill='#FF9966',outlier.colour="#FF
### 2-Point Field Goals
p9<-ggplot(players,aes(x=year,y=`
                           2goal`,group=year))+geom_boxplot(fill='#CC3333',outlier.colour="#CC3300",
### 3-Point Field Goals
p10<-ggplot(players,aes(x=year,y=`3goal`,group=year))+geom_boxplot(fill='#990066',outlier.colour="#9900
```

```
### 2-Point Field Goal Percentage
p11<-ggplot(players,aes(x=year,y=`2per`,group=year))+geom_boxplot(fill='#663399',outlier.colour="#66333
### 3-Point Field Goal Percentage
p12<-ggplot(players,aes(x=year,y=`3per`,group=year))+geom_boxplot(fill='#330099',outlier.colour="#33006

gridExtra::grid.arrange(p7,p8,p9,p10,p11,p12,ncol = 2)
```

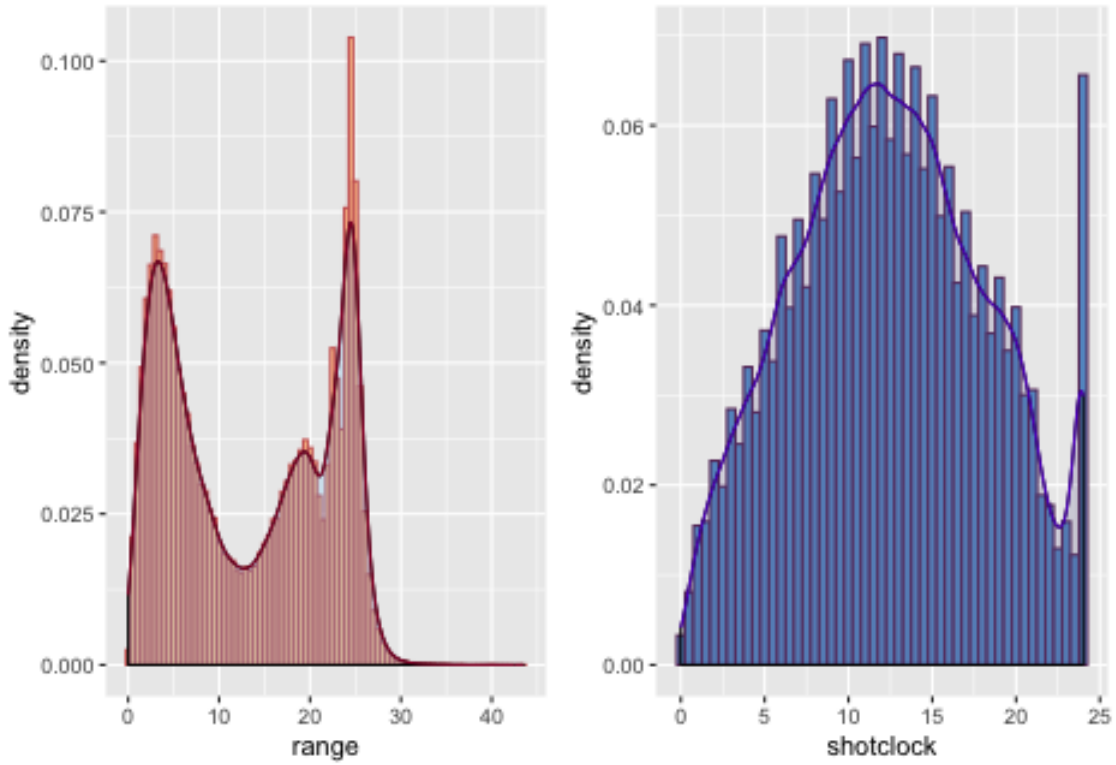


Of all the players in the NBA from 1980 to 2017, their shooting trends are obvious. In the comparasion between 2-point field goals and 3-point field goal, we could see there exists decreasing in 2-point field goal attempts/goals and increasing in 3-point field goal attempts/goals. In the meantime, the 2-point field goal percentage keep 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,

```
### range preference
p13<-ggplot(da, aes(x=range)) +
    geom_histogram(aes(y=..density..),binwidth=.5,colour="#CC6666", fill="#FFCC99") +
    geom_density(alpha=.2, fill="#66CCFF")+
    geom_line(stat="density",colour="#990033")
### shotclock preference
p14<-ggplot(da, aes(x=shotclock)) +
    geom_histogram(aes(y=..density..),binwidth=.5,colour="#663366", fill="#6699CC") +
    geom_density(alpha=.2, fill="#666699")+
    geom_line(stat="density",colour="#6600CC")
gridExtra::grid.arrange(p13,p14,ncol = 2)
```



From the first plot we could see the major scores come from 0-5 meters and 22-25 meters which indicates players nowadays prefer layup, dunk or throw 3-point balls. From the second plot, we could see majority of players prefer to play organized attack, they are not will to save time or shot until the last second.

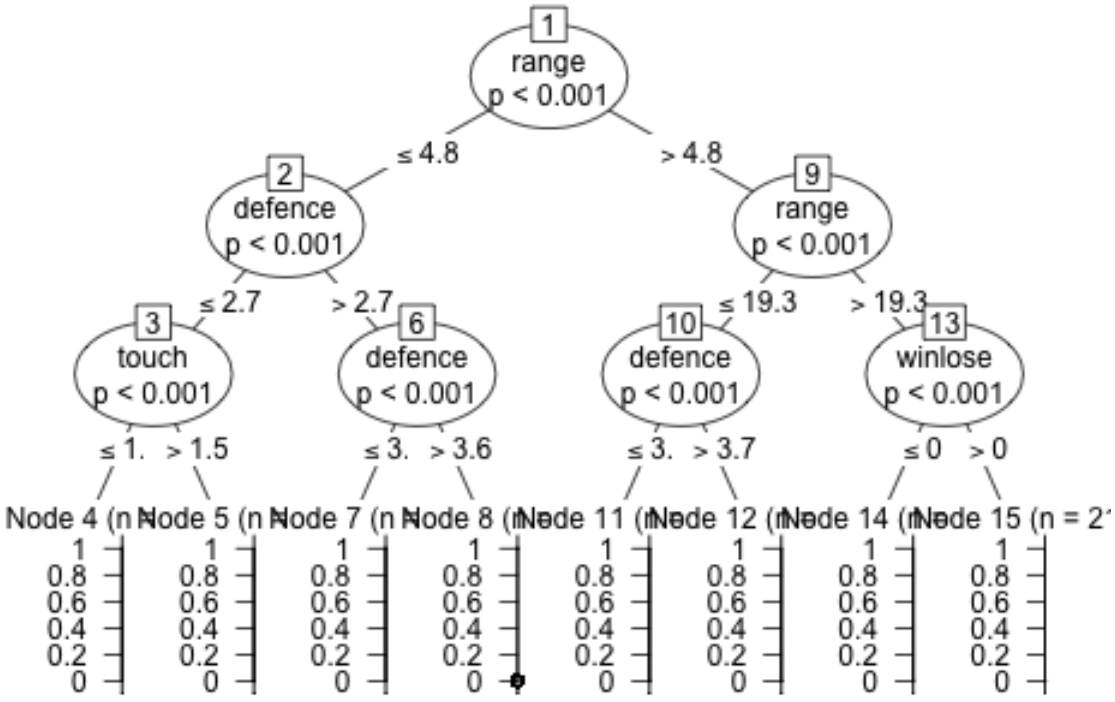
4.4 C-tree analysis

```
library(party)
```

```
## Loading required package: grid
## Loading required package: mvtnorm
## Loading required package: modeltools
## Loading required package: stats4
##
## Attaching package: 'modeltools'
## The following object is masked from 'package:car':
##
##     Predict
## The following object is masked from 'package:lme4':
##
##     refit
## Loading required package: strucchange
## Loading required package: zoo
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

##
## Attaching package: 'strucchange'

## The following object is masked from 'package:rstan':
##
##     monitor

## The following object is masked from 'package:stringr':
##
##     boundary
```

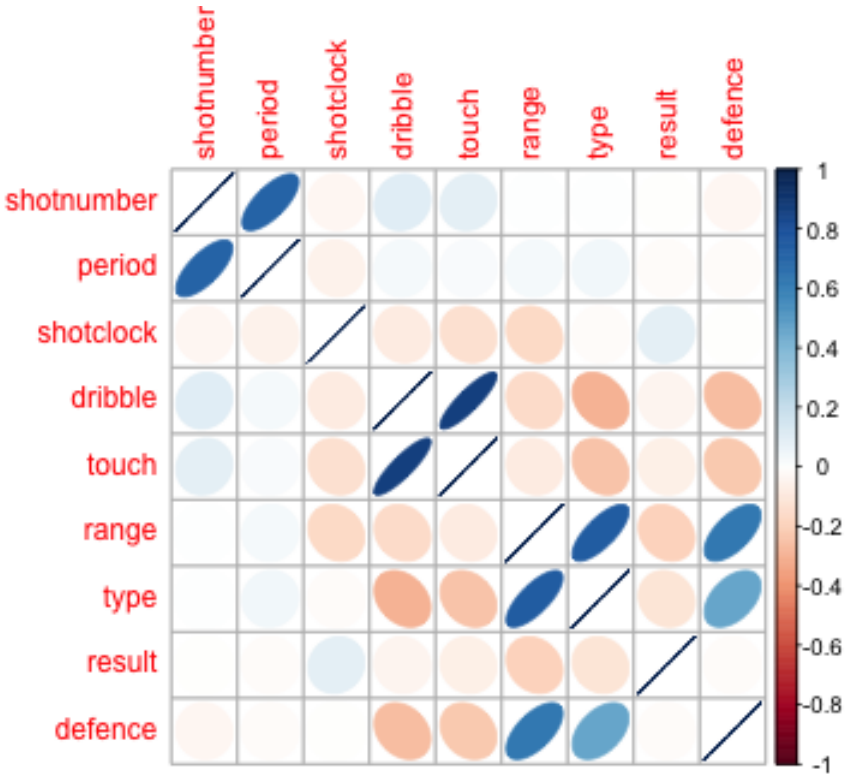
```
formula<-result~location+winlose+period+shotnumber+shotclock+dribble+touch+range+type+defence
dt<-ctree(formula,da,controls=ctree_control(minsplit=10,maxdepth=3))
plot(dt)
```



From the C-tree, we could see the decisive relationship between variables, the shooting range large influence the other variables.

4.5 Corelation Test

```
shot <- da[,c(5:13)]
M = cor(shot,method = "spearman")
corrplot::corrplot(M,method = "ellipse")
```



Corelation plot shows there exists obvious relationships between variables.

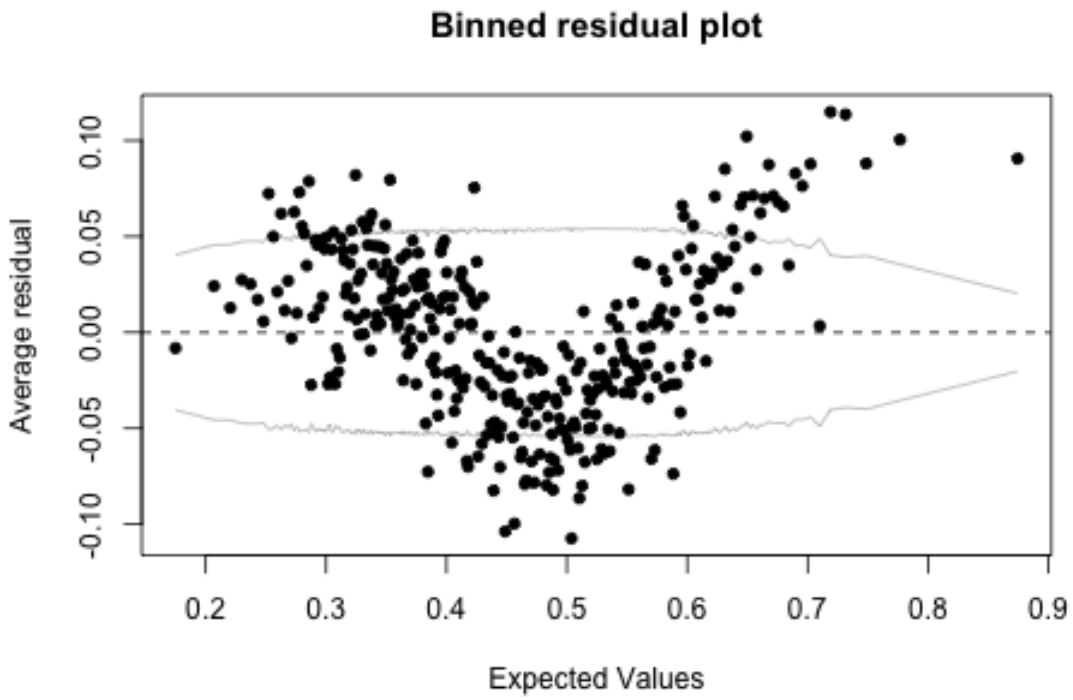
After i do these EDA plots, I believe the players shooting condition changes so much in recent years that i decide to analyze how their shooting results are influenced by other factor.
Based on NBA shot log dataset, I pick some of variables to build the model.

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:
- Model check

```
binnedplot(predict(model_1,type="response"), residuals(model_1,type="response"))
```



- Interpretation Choose my favorite player called Damian Lillard, his predicting function shows below:

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

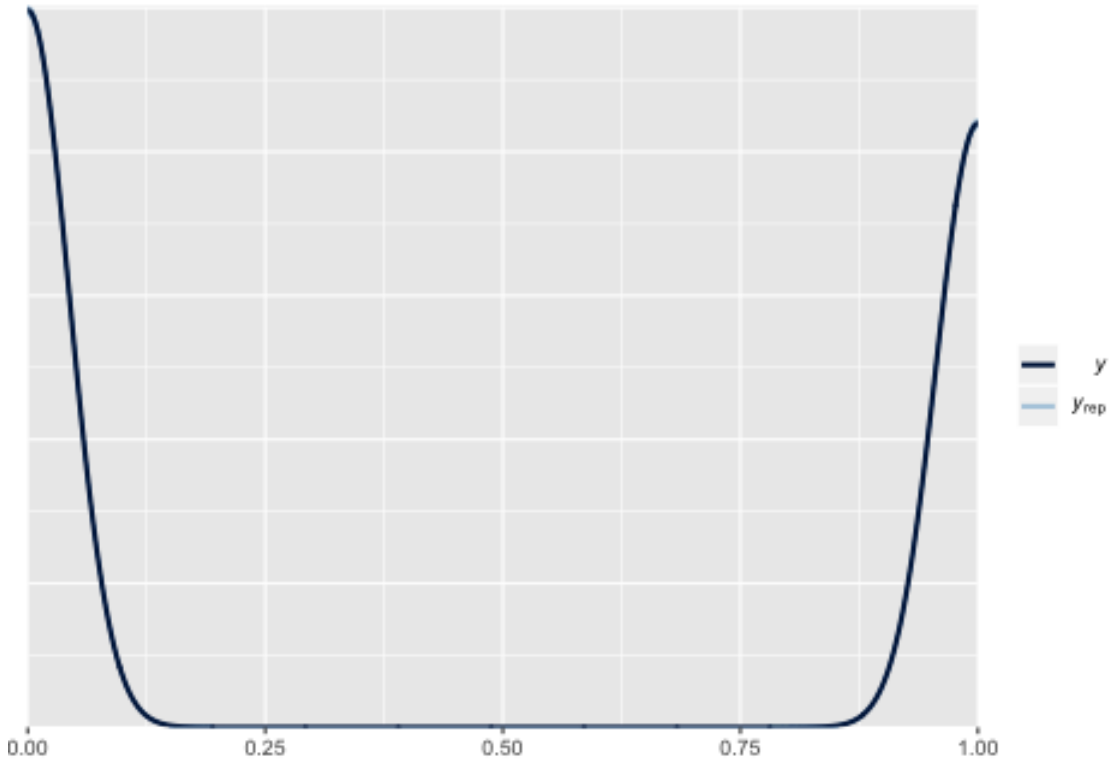
In this model, the majority of the coefficient are significant. But the according to the binned residual plot, there exists an obvious pattern, and the outliers are pretty large, hence this model needs improvement.

5.2 Bayesian generlized linear model

The second model i do is bayesian generlized linear model, shows below:

- Model check

```
rstanarm::pp_check(model_2)
```



The simulation result show the output of my model is quite corespond with the observation which is 0 or 1, indicating the model fits quite well.

5.3 Mutilevel model

random intercept

```
### m1
m1<-glmer(result~location+shotnumber+period+shotclock+dribble+touch+type+range+defence+
            dribble*touch+type*range+(1|name), data=da,family = binomial(),REML = T)
```

```
## Warning: extra argument(s) 'REML' disregarded
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.0249899
## (tol = 0.001, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?;Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```

```
### summary
summary(m1)
```

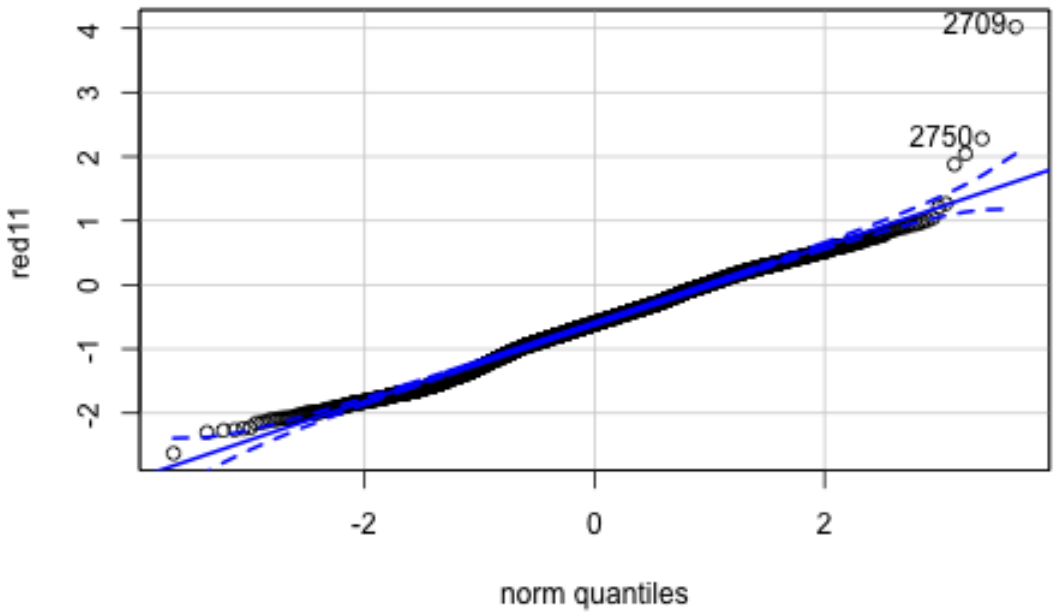
```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: result ~ location + shotnumber + period + shotclock + dribble +
##     touch + type + range + defence + dribble * touch + type *
##     range + (1 | name)
##    Data: da
```

```
##
##      AIC      BIC   logLik deviance df.resid
## 154521.8 154647.4 -77247.9 154495.8   116761
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -16.7085  -0.8768  -0.6401   0.9885   2.6931
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  name   (Intercept) 0.01638  0.128
## Number of obs: 116774, groups:  name, 242
##
## Fixed effects:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.7480432  0.4121106   1.815  0.06950 .
## location       0.0349269  0.0121303   2.879  0.00399 **
## shotnumber     0.0009547  0.0020350   0.469  0.63898
## period        -0.0093570  0.0078908  -1.186  0.23569
## shotclock      0.0144802  0.0011300  12.814  < 2e-16 ***
## dribble       -0.0089126  0.0061154  -1.457  0.14500
## touch         -0.0657122  0.0057935 -11.342  < 2e-16 ***
## type          -0.2668588  0.2048571  -1.303  0.19269
## range         -0.0938429  0.0170915  -5.491 4.01e-08 ***
## defence        0.1040335  0.0029685  35.045  < 2e-16 ***
## dribble:touch  0.0031906  0.0003346   9.536  < 2e-16 ***
## type:range     0.0140771  0.0084323   1.669  0.09503 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) locatn shtnmb period shtclc dribbl touch  type   range
## location    -0.012
## shotnumber  -0.012  0.006
## period      -0.017 -0.007 -0.732
## shotclock   -0.057 -0.006 -0.008  0.039
## dribble     -0.014  0.005  0.012 -0.023 -0.130
## touch       -0.055 -0.003 -0.016  0.025  0.173 -0.657
## type        -0.997 -0.002  0.016 -0.009  0.013  0.020  0.030
## range       -0.987 -0.002  0.019 -0.007  0.042  0.007  0.023  0.987
## defence      0.066  0.001 -0.003  0.015 -0.114  0.048  0.053 -0.078 -0.110
## dribble:tch  0.052  0.000 -0.035  0.013  0.045 -0.584 -0.131 -0.039 -0.022
## type:range   0.993  0.002 -0.017  0.007 -0.025 -0.008 -0.024 -0.994 -0.997
##             defenc drbbl:
## location
## shotnumber
## period
## shotclock
## dribble
## touch
## type
## range
## defence
## dribble:tch -0.090
```

```
## type:range   0.078  0.024
## convergence code: 0
## Model failed to converge with max|grad| = 0.0249899 (tol = 0.001, component 1)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
## Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```

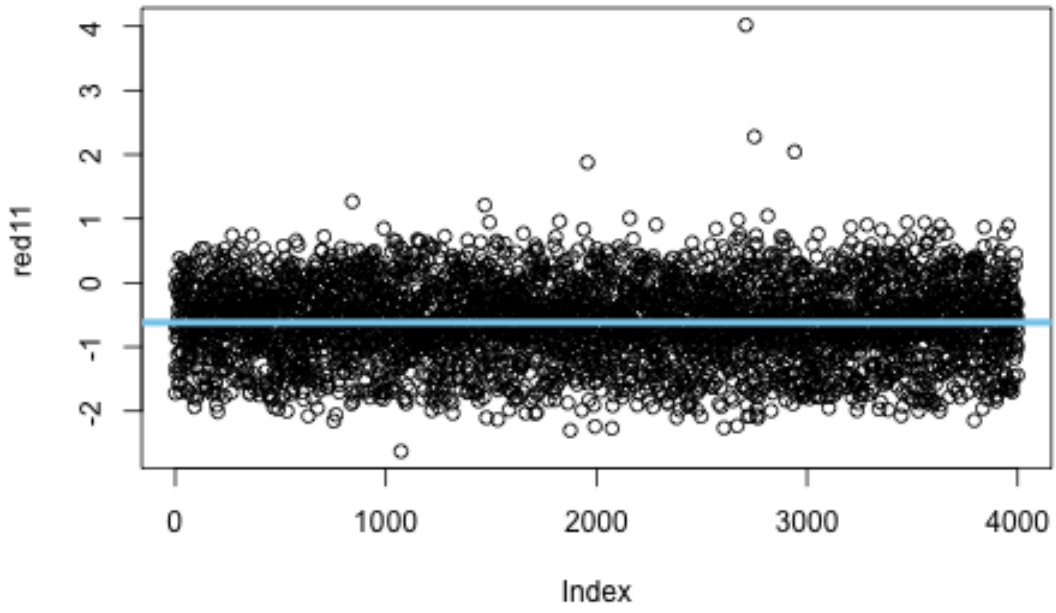
From model1 we can see the most coefficients's p-value are quite small and they are significant. I set players name as pooling principle, so i get nearly 250 groups. From the resultm different players would have different shooting ability and different intercept. Additionally, there exists corelationships between these variables.

```
### check
pred1 = predict(m1,type = "link")
red1 = as.numeric(as.character(pred1))-da$result
com1 = data.frame(obs = da$result,pre = pred1) %>% pivot_longer(cols = 1:2,names_to = "type",values_to =
red11 = sample(red1,4000)
### check visualization
car::qqPlot(red11)
```



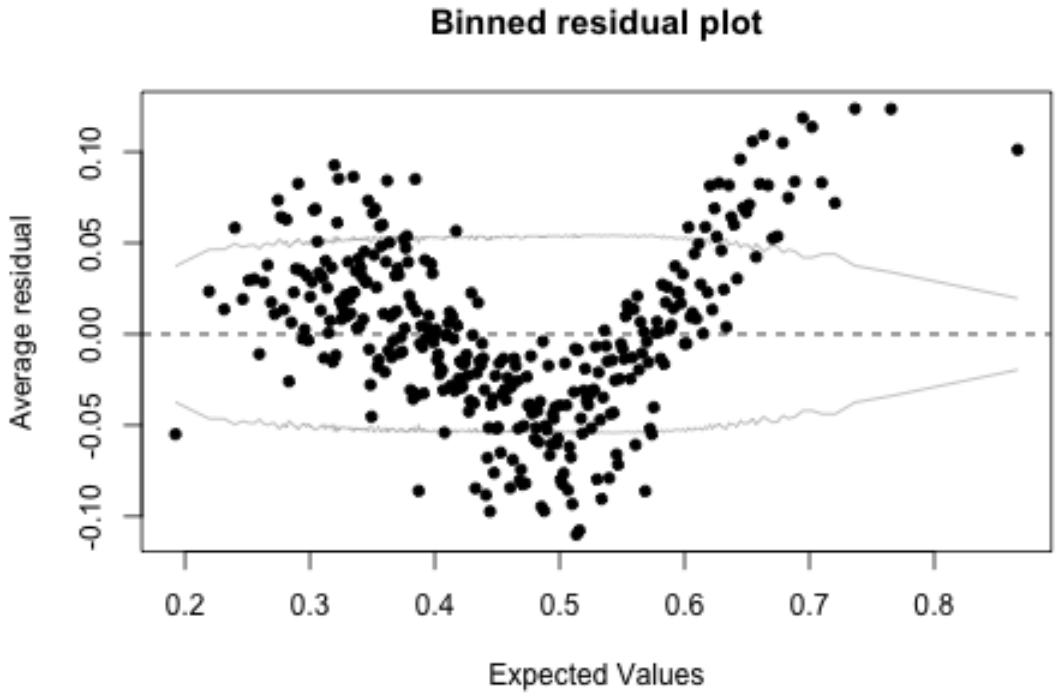

```
## [1] 2709 2750
```

```
plot(red11)+abline(h=mean(red11),lwd=4,col="skyblue")
```




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

```
binnedplot(fitted(m1),residuals(m1,type="response"))
```



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. The model fits well in this way.

random slope

```
### m2
m2<-glmer(result~location+shotnumber+shotclock+dribble+touch+type+range+defence+
          dribble*touch+type*range+(touch-1|name),
          data=da,family = binomial(),REML = T)
```

```
## Warning: extra argument(s) 'REML' disregarded
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.254732
## (tol = 0.001, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?;Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```

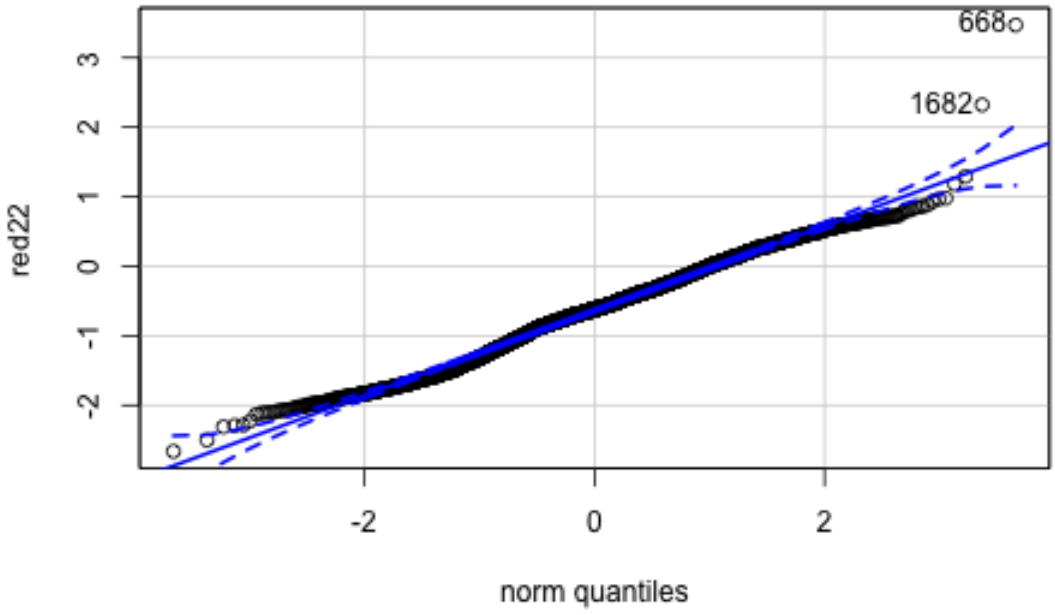
From model2 we can see the most coefficients's p-value are quite small and they are significant. I set players name as pooling principle,and the slope of range varys form different players. From the result different players would have different slope of range. We could see the range preference could be different because some players(like Lilard) prefer to shoot long distance ball. Additionally, there exists corelationships between these variables.

```
### summary
summary(m2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: result ~ location + shotnumber + shotclock + dribble + touch +
##     type + range + defence + dribble * touch + type * range +
##     (touch - 1 | name)
##    Data: da
##
##      AIC      BIC   logLik deviance df.resid
## 154590.2 154706.2 -77283.1 154566.2   116762
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6535 -0.8747 -0.6445  0.9870  2.6971
##
## Random effects:
##  Groups Name  Variance Std.Dev.
##  name   touch 0.00105  0.03241
## Number of obs: 116774, groups:  name, 242
##
## Fixed effects:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.6854744  0.3993713   1.716  0.08609 .
## location       0.0343534  0.0121129   2.836  0.00457 **
## shotnumber     0.0004964  0.0013514   0.367  0.71339
## shotclock      0.0145781  0.0011270  12.935  < 2e-16 ***
## dribble       -0.0077733  0.0061910  -1.256  0.20927
## touch         -0.0695342  0.0063356 -10.975  < 2e-16 ***
```

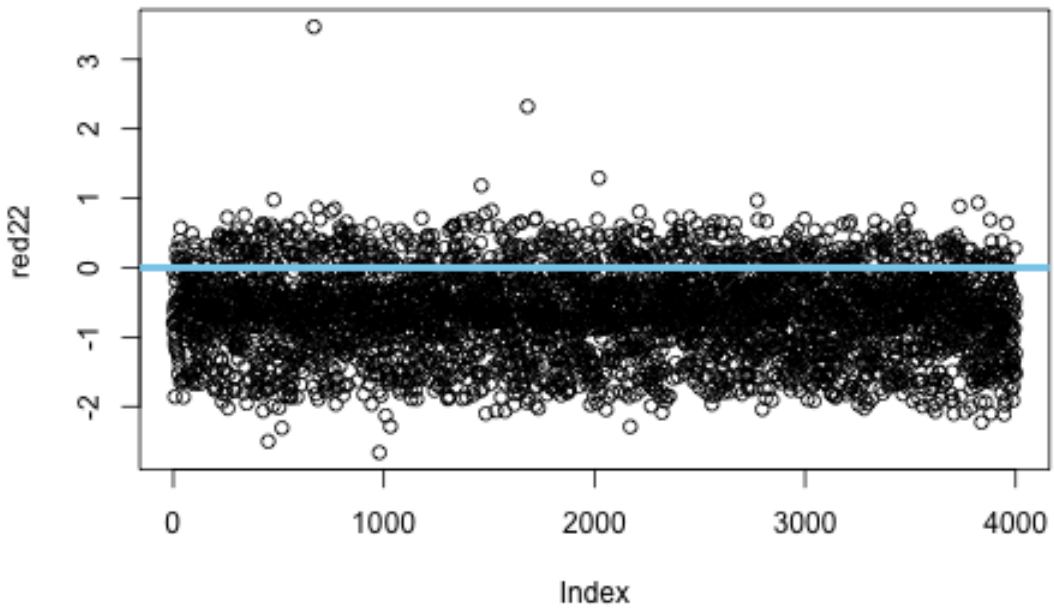
```
## type          -0.2440369  0.1986120  -1.229  0.21918
## range         -0.0909796  0.0165782  -5.488 4.07e-08 ***
## defence        0.1024842  0.0029366  34.899  < 2e-16 ***
## dribble:touch  0.0025804  0.0003419   7.548 4.43e-14 ***
## type:range     0.0131100  0.0081801   1.603  0.10901
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) locatn shtnmb shtclc dribbl touch  type   range  defenc
## location    -0.013
## shotnumber  -0.045  0.001
## shotclock   -0.055 -0.006  0.026
## dribble     -0.010  0.005 -0.011 -0.113
## touch       -0.055 -0.002  0.027  0.154 -0.614
## type        -0.998 -0.002  0.022  0.012  0.015  0.031
## range       -0.988 -0.001  0.025  0.041  0.005  0.025  0.986
## defence      0.064  0.000  0.022 -0.108  0.060  0.045 -0.076 -0.110
## dribble:tch  0.044  0.001  0.001  0.074 -0.506 -0.105 -0.036 -0.019 -0.094
## type:range   0.993  0.001 -0.025 -0.024 -0.006 -0.026 -0.994 -0.997  0.078
##             drbbl:
## location
## shotnumber
## shotclock
## dribble
## touch
## type
## range
## defence
## dribble:tch
## type:range   0.024
## convergence code: 0
## Model failed to converge with max|grad| = 0.254732 (tol = 0.001, component 1)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
## Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```

```
### check
pred2 = predict(m2,type = "link")
red2 = as.numeric(as.character(pred2))-da$result
com2 = data.frame(obs = da$result,pre = pred1) %>% pivot_longer(cols = 1:2,names_to = "type",values_to =
red22 = sample(red2,4000)
### check visualization
car::qqPlot(red22)
```




```
## [1]  668 1682
```

```
plot(red22)+abline(h=0,lwd=4,col="skyblue")
```




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

random slope and random intercept

```
### m3
m3<-glmer(result~location+shotnumber+shotclock+dribble+touch+type+range+defence+
          dribble*touch+type*range+(1+range|name),
          data=da,family = binomial(),REML = T)
```

```
## Warning: extra argument(s) 'REML' disregarded
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.0725334
## (tol = 0.001, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?;Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```

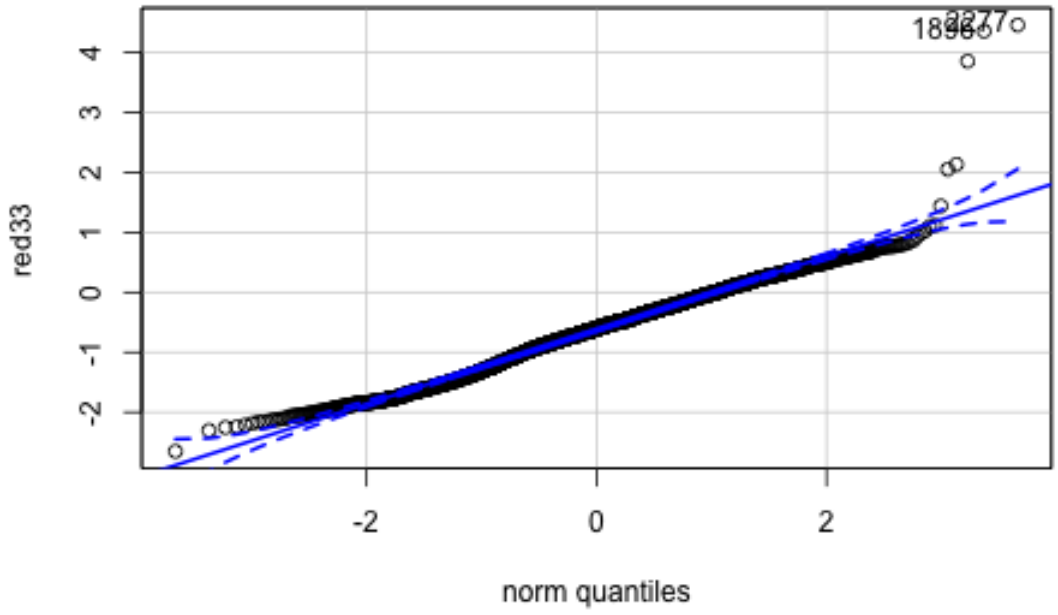
```
### summary
summary(m3)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: result ~ location + shotnumber + shotclock + dribble + touch +
##     type + range + defence + dribble * touch + type * range +
##     (1 + range | name)
##    Data: da
##
##      AIC      BIC   logLik deviance df.resid
## 154428.4 154563.7 -77200.2 154400.4   116760
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -14.5112  -0.8753  -0.6369   0.9905   2.7188
##
## Random effects:
##  Groups Name        Variance  Std.Dev. Corr
##  name   (Intercept) 0.0344049 0.18549
##         range       0.0001873 0.01369  -0.78
## Number of obs: 116774, groups:  name, 242
##
## Fixed effects:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.4016785  0.4090876   0.982  0.32615
## location       0.0344213  0.0121496   2.833  0.00461 **
## shotnumber    -0.0008310  0.0013920  -0.597  0.55050
## shotclock      0.0139196  0.0011371  12.241  < 2e-16 ***
## dribble       -0.0068525  0.0061468  -1.115  0.26493
## touch         -0.0650106  0.0058092 -11.191  < 2e-16 ***
## type          -0.1081019  0.2033960  -0.531  0.59508
## range         -0.0773900  0.0170390  -4.542 5.57e-06 ***
## defence        0.1091225  0.0030583  35.681  < 2e-16 ***
## dribble:touch  0.0030564  0.0003357   9.103  < 2e-16 ***
## type:range     0.0054606  0.0084045   0.650  0.51587
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Correlation of Fixed Effects:
##             (Intr) locatn shtnmb shtclc dribbl touch  type   range  defenc
## location    -0.011
## shotnumber  -0.033  0.001
## shotclock   -0.048 -0.006  0.033
## dribble     -0.017  0.006 -0.001 -0.125
## touch       -0.052 -0.003  0.004  0.169 -0.654
## type        -0.997 -0.003  0.010  0.005  0.022  0.027
## range       -0.986 -0.003  0.018  0.032  0.010  0.020  0.984
## defence      0.046  0.000  0.008 -0.126  0.044  0.054 -0.057 -0.084
## dribble:tch  0.052  0.000 -0.042  0.044 -0.586 -0.132 -0.039 -0.021 -0.089
## type:range   0.992  0.003 -0.014 -0.014 -0.010 -0.022 -0.993 -0.996  0.050
##             drbbl:
## location
## shotnumber
## shotclock
## dribble
## touch
## type
## range
## defence
## dribble:tch
## type:range   0.023
## convergence code: 0
## Model failed to converge with max|grad| = 0.0725334 (tol = 0.001, component 1)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
## Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```

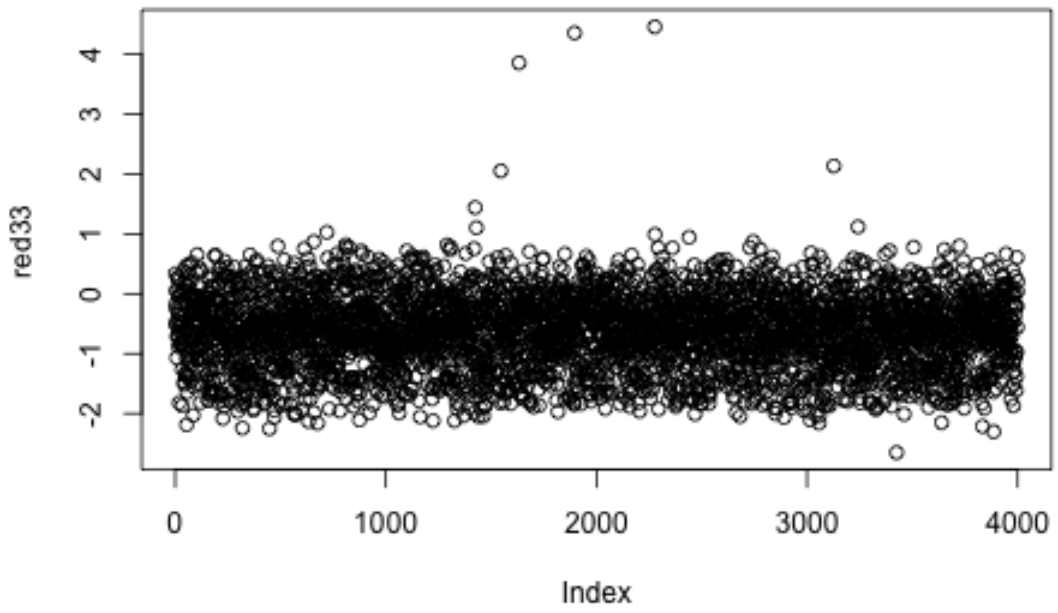
From model3 we can see the most coefficients's p-value are quite small and they are significant. I set players name as pooling principle,and the slope of range varys form different players, the intercept could vary for different players. We could see the range preference could be different because some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

```
### check
pred3 = predict(m3,type = "link")
red3 = as.numeric(as.character(pred3))-da$result
com3 = data.frame(obs = da$result,pre = pred1) %>% pivot_longer(cols = 1:2,names_to = "type",values_to =
red33 = sample(red3,4000)
### check visualization
car::qqPlot(red33)
```




```
## [1] 2277 1896
```

```
plot(red33)
```



The qq-plot and residual plot shows the residuals are randomly normal distributed, while the mean of residual

turn to be -1 which is abnormal.

5.4 Model evaluation and selection

```
anova(m2,m3,m1)
```

```
## 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, while according to he AIC and BIC the model3's value is quite small. In addition, model 3 has more perfect performance in predicting shooting result.

6.Result and discussion

- 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, based on this acknowledgement, the
- Limitation

1. 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.

2. 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.

- Improvement

1. 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.

2. 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.

3. 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.

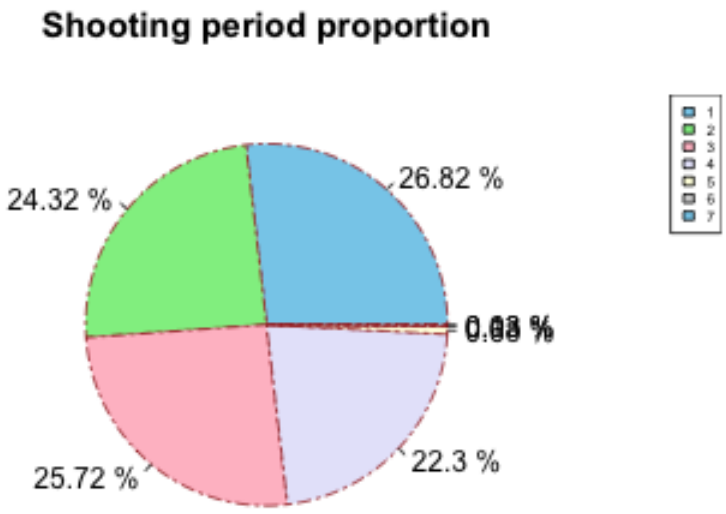
- Future direction Some other method could be added into this analysis such as statistical learning to make this analysis more strengthful. Additionally, this analysis could be used into sports field(not limited to NBA), NFL, FIFA, hockey, and the like.

7.Reference

- https://www.kaggle.com/
- http://www.stat-nba.com/
- https://www.basketball-reference.com/
- https://www.hupu.com/
- https://blog.csdn.net/

8.Appendix

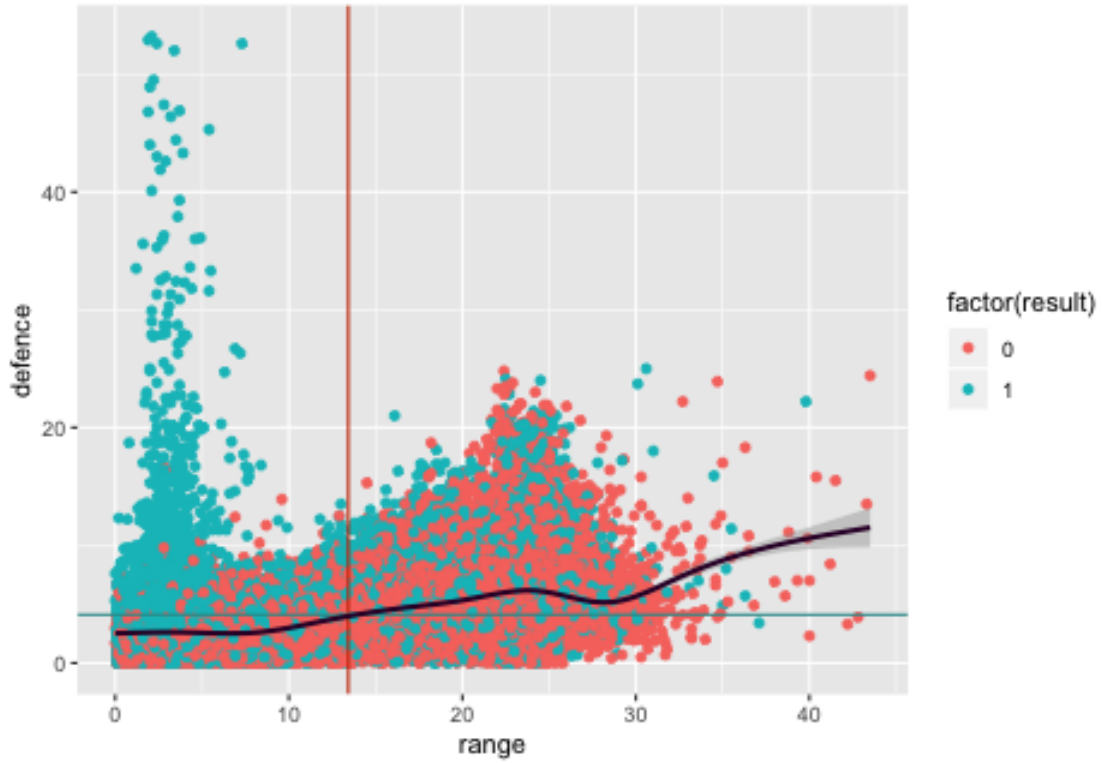
shooting period preference



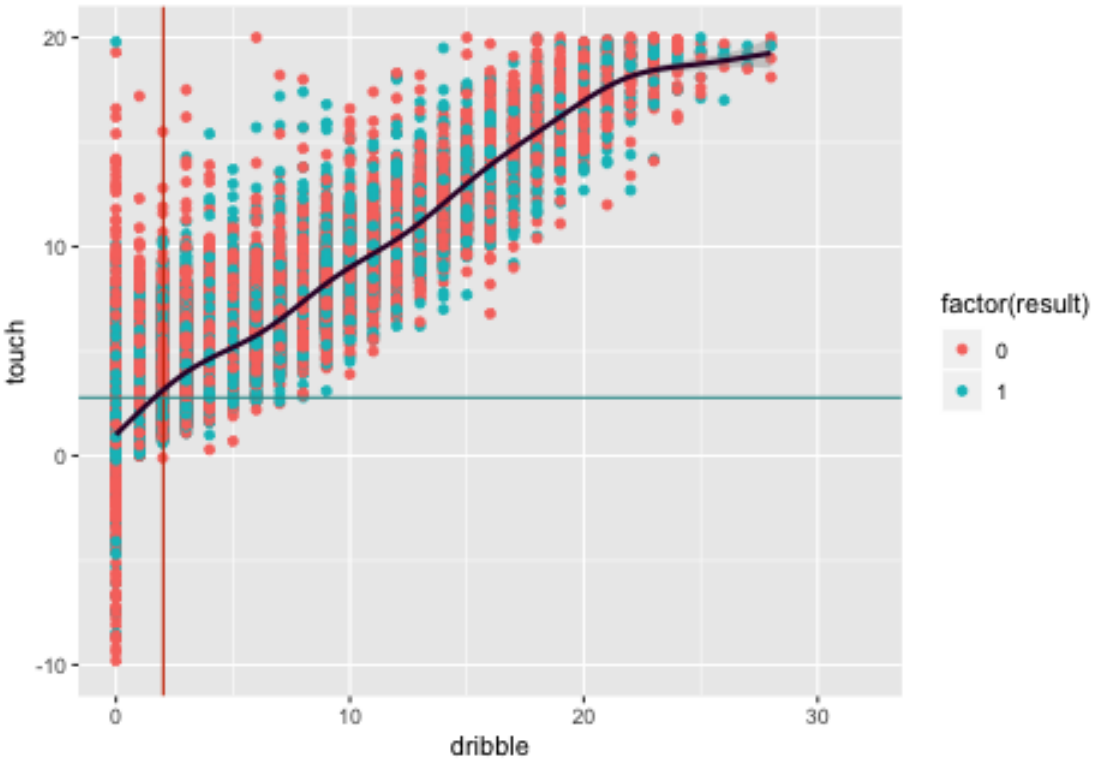
How variables influence player's shoot result.

```
## Warning: Ignoring unknown parameters: type
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
## Warning: Ignoring unknown parameters: type
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## Warning: Removed 194 rows containing non-finite values (stat_smooth).
## Warning: Removed 194 rows containing missing values (geom_point).
```


20
10
0
-10
touch
0
10
20
30
dribble
factor(result)
0
1