# MA678_Midterm_Jiaqi

Jiaqi Sun

2022-11-24

## Abstract

The Premier League is the highest level of the men's English football league system. Contested by 20 clubs, it operates on a system of promotion and relegation with the English Football League. Seasons typically run from August to May with each team playing 38 matches. As 3 point for a win, 1 point for a draw and 0 point for a lose, the team with the highest points in the end of the season is the champion. For the past 20 seasons, six clubs have won the Premier League title: Manchester United, Chelsea, Manchester City, Arsenal, Leicester City and Liverpool. Thus, here comes the problem: with all kinds of match statistics being collected and analyzed, what factor determines whether a match win or lose?

To figure out this problem, I built a multilevel model with group level `Team` and `Referee`. The result indicates that different teams are influenced by different match statistics.

This report can be divided into 4 main parts: Introduction, Method, Result, and Discussion.

## Introduction

Usually, there are several Essential Premier League Stats including Number of Cards (yellow/red) Per Game,Goals Per Game, Shots on Target, Fouls Committed. These are important description of performance of the match and whether the home team can win the match is greatly decided by it.

Nevertheless, central to all of this are the referees, the individuals who uphold the laws of the game, who maintain authenticity and who make the decisions, which can be the difference between winning and losing. Despite their importance, referees are an often overlooked part of professional football. Referees are an integral part of the global game and are required for fixtures to take place and competitions to occur. Despite the help of VAR, referees continue to make errors on a weekly basis. Fans have also given humorous nicknames to Premier League referees over the years: "Fantastic Four Blind" – Refers to the four referees who enforce the Premier League, namely "Great Hunter" Clattenburg, "The Scourge" Atkinson, "Blind Man" Oliver and "North London Sniper" McDean, these four people often appear on the field Misjudgments and misjudgments, and turning a blind eye to many foul actions, are called the "Fantastic Four Blind" by fans. All these four "blind" have enforced countless matches compared to other referees so definitely, referees should be considered when we talked about the match result.

On the other hand, Unlike some of its counterparts across Europe, where one or two clubs are dominant, the Premier League features what has become popularly known as 'The Big Six', which are composed by Manchester United, Liverpool, Arsenal, Chelsea, Manchester City and Tottenham. They are the most consistently successful teams in the division. Not only that, but they boast the biggest stadiums, broadest fanbases and, as a result, the healthiest bank accounts. Meanwhile, "The Big Six" teams are willing to spend a lot on buying top players and hire dominant coached all over the world to form a better lineup and chase championship while some teams not. It's natural to come to the conclusion that teams in 'The Big Six' teams are more likely to win the matches. That to say, team is another factor we should take into account.

Therefore, I decide to introduce multilevel models to find out the influences of fixed effects (e.g. corners, goals, cards, shots and so on) and random effects (teams, Referees).

## Methods

### Data Preprocessing

I found the data set from Kaggle (https://www.kaggle.com/datasets/irkaal/english-premier-league-results/code?select=results.csv).

Firstly, I download 2000-2022 match results. The match statistics can be divided into Home Team statistics and away team statistics. For now, we are only focusing on both the Home Team statistics and the Away Team statistics.

Additionally, as the match statistics is by individual games, I calculate average game statistics for each team and transform `FTR` and `HTR` information to binary factor (0 or 1). I also need to filter out the referees that participated in over 50 matches so that they are really making an impact. I then combined the data to get the final data frame.

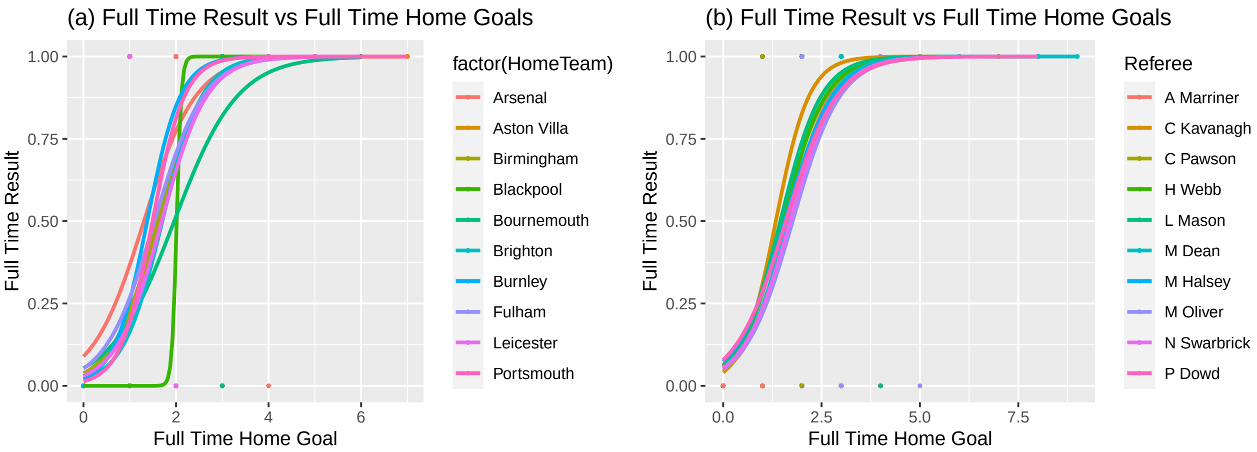
Here is the glossary of terms:

| column names | explanation |
|---|---|
| Season | Match Season |
| DateTime | Match Date and Time (yyyy-mm-dd hh:mm:ss) |
| HomeTeam | Home Team |
| AwayTeam | Away Team |
| FTHG | Full Time Home Team Goals |
| FTAG | Full Time Away Team Goals |
| FTR | Full Time Result(H=Home Win, D=Draw, A=Away Win) |
| HTHG | Half Time Home Team Goals |
| HTAG | Half Time Away Team Goals |
| HTR | Half Time Result(H=Home Win, D=Draw, A=Away Win) |
| Referee | Match Referee |
| HS | Home Team Shots |
| AS | Away Team Shots |
| HST | Home Team Shots on Target |
| AST | Away Team Shots on Target |
| HC | Home Team Corners |
| AC | Away Team Corners |
| HF | Home Team Fouls Committed |
| AF | Away Team Fouls Committed |
| HY | Home Team Yellow Cards |
| AY | Away Team Yellow Cards |
| HR | Home Team Red Cards |
| AR | Away Team Red Cards |

### Exploratory Data Analysis

By aforementioned part, I've got a `EPL_stats` with 7041 observations and 23 variables, among which there is 1 output `FTR` and 20 predictors. However, whether or not to use all of these 22 predictors is depended on following analysis.

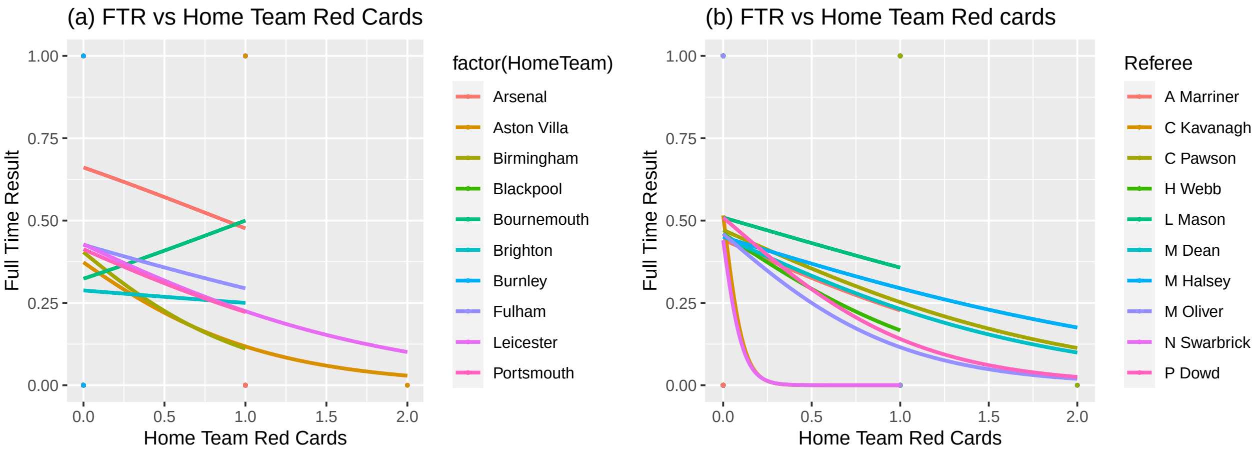
~\



~\

Figure 1 illustrates the relationship between FTR and Full Time Home Goals, while fig(a) is in team level and fig(b) is in Half Time Result level. However, whatever the level, Full Time Result show the winning trend as Full time Home Goals going up. And in different teams and Half Time Results, the intercepts and slopes show slights differences. After I draw the graph of Full Time Result versus appearance, rebounds, assists, steals and blocks, the figures are quite similar. Thus I put them in the appendix.

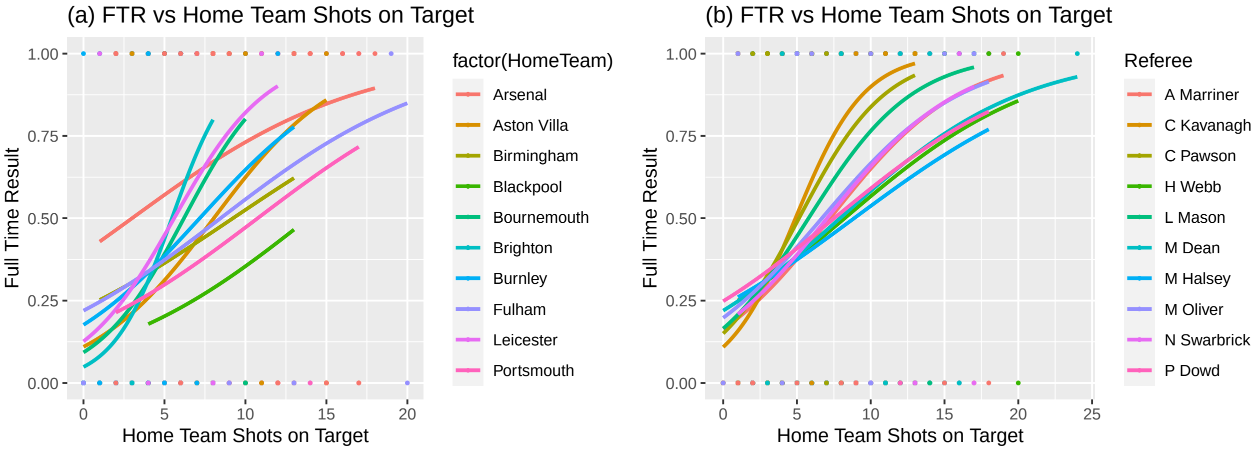
~\



~\

Figure above shows that whether from team or referee level, high Home Red Cards are bad and usually lead to worse match result. A red card means instant dismissal. In addition to having to fight with ten people, being fined and suspended, and sometimes suffering penalty kicks, this will break the balance of the field. Yet, the slopes and intercepts vary from team to team and that indicate random effects really matter in this case.

~\



~\

Figure above shows the correlation between the match's Full Time Result and Home Team Shots on Target. Similarly, figure(a) is in team level while figure(b) is in referee level. The results is very like the one shows the correlation between the match's Full Time Result and Home Team Shots. Because Home Team shots on Target is absolutely come from Home Team Shots. Thus, I decided to look into relationship between HST and HS.

~\

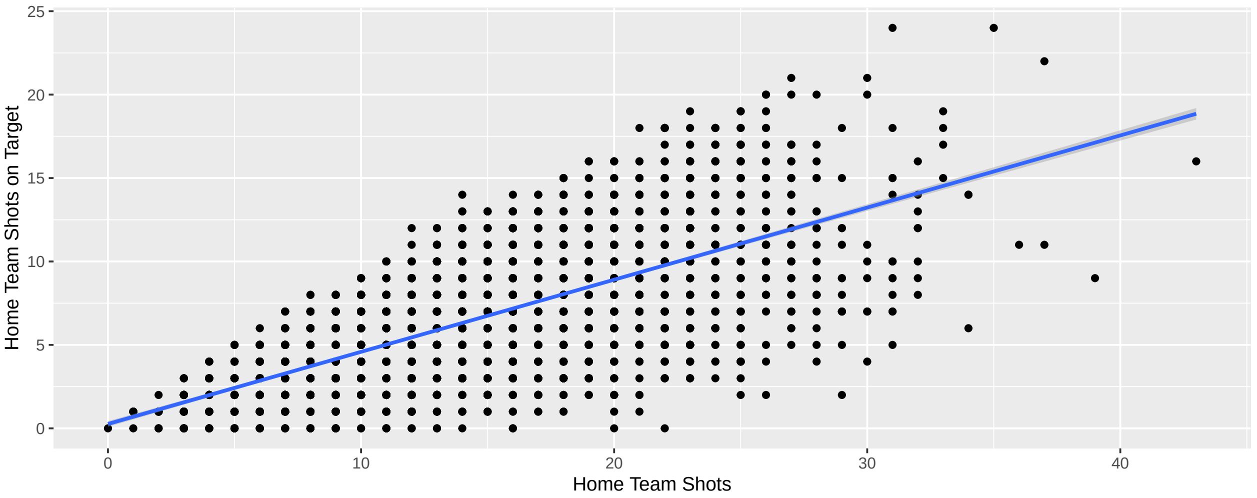
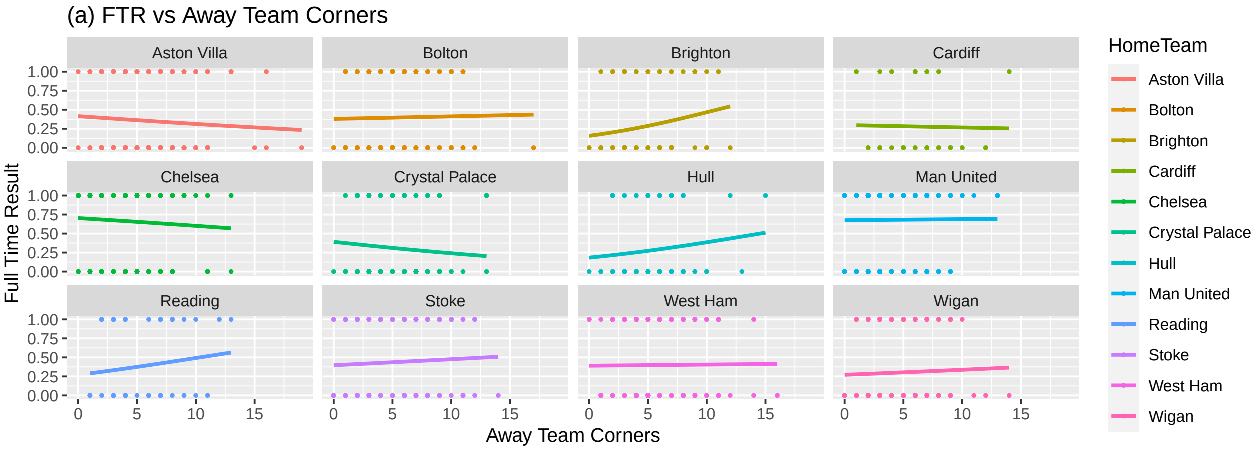


Figure 1: relationship between Home Team Shots on Target and Home Team Shots

~\

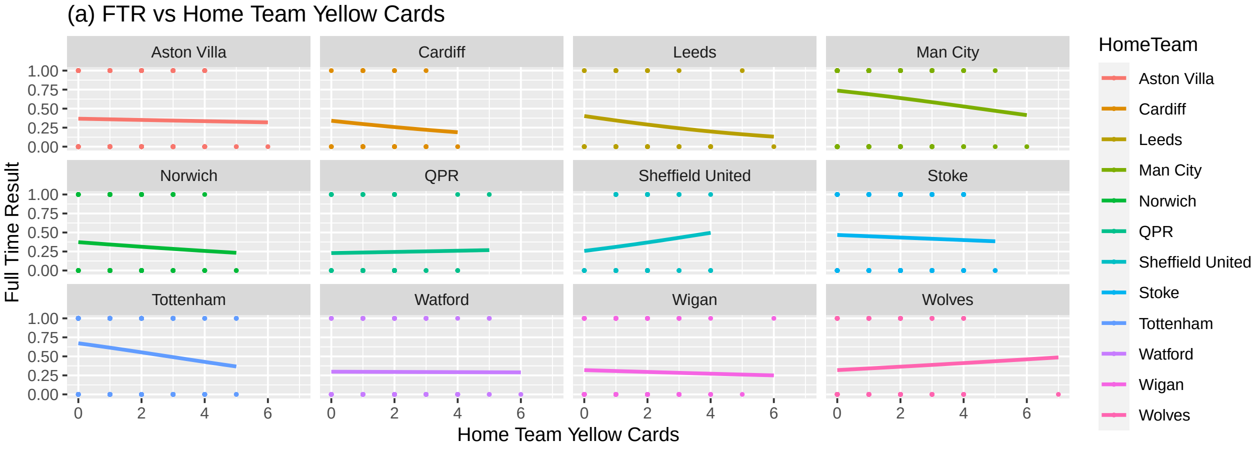
Figure 3 verifies that home team players' shots on target are closely related with their shots with no surprise. Thus, I decide to exclude variable `Home Team shots`.



~\

Figure above shows that whatever from team or referee level, high away team Corners are good and usually lead to better match result. Yet, the slopes and intercepts vary from team to team and that indicate random effects really matter in this case.

~\



~\

Figure above shows that whatever from team or referee level, high home team yellow cards are bad and usually lead to worse match result. Yet, the slopes and intercepts vary from team to team and that indicate random effects really matter in this case.

### Model fitting

Since different teams and positions have quite large impacts on the model, I decide to use multilevel model to fit `NBA_data`. As to selection of variables, I also include `three made per game` because under the crazy trendency of small ball, the ability to shoot 3-points is really important. Meanwhile, since all variables are more or less skewed and have heavy tails, I took `log(variable + 1)` to create new ones. All original distribution plots of variables can be found in Appendix of this report. For the next step, I draw the Pearson correlation matrix to do the predictor selection.

~\

Additionally, as different teams have quite different on-court strategies and seasonal goals, random effect of teams is quite important for variables: `FTHG, HST and HC`. On the other hand, A foul is an unfair act by a player, deemed by the referee to contravene the game's laws, that interferes with the active play of the game. Fouls could be pretty bias as each referee has his own standard. For example: Mike Dean is officially the strictest referee in the Premier League, according to results of a new study, `HF` per game: 10.80. `HY` per game: 1.93. Overall `HF` per booking: 5.59. Fewest `HF` per game: Burnley (9.06). Thus, I prefer changing slopes and intercepts of 'referee" for different ones. Here is the function I built:

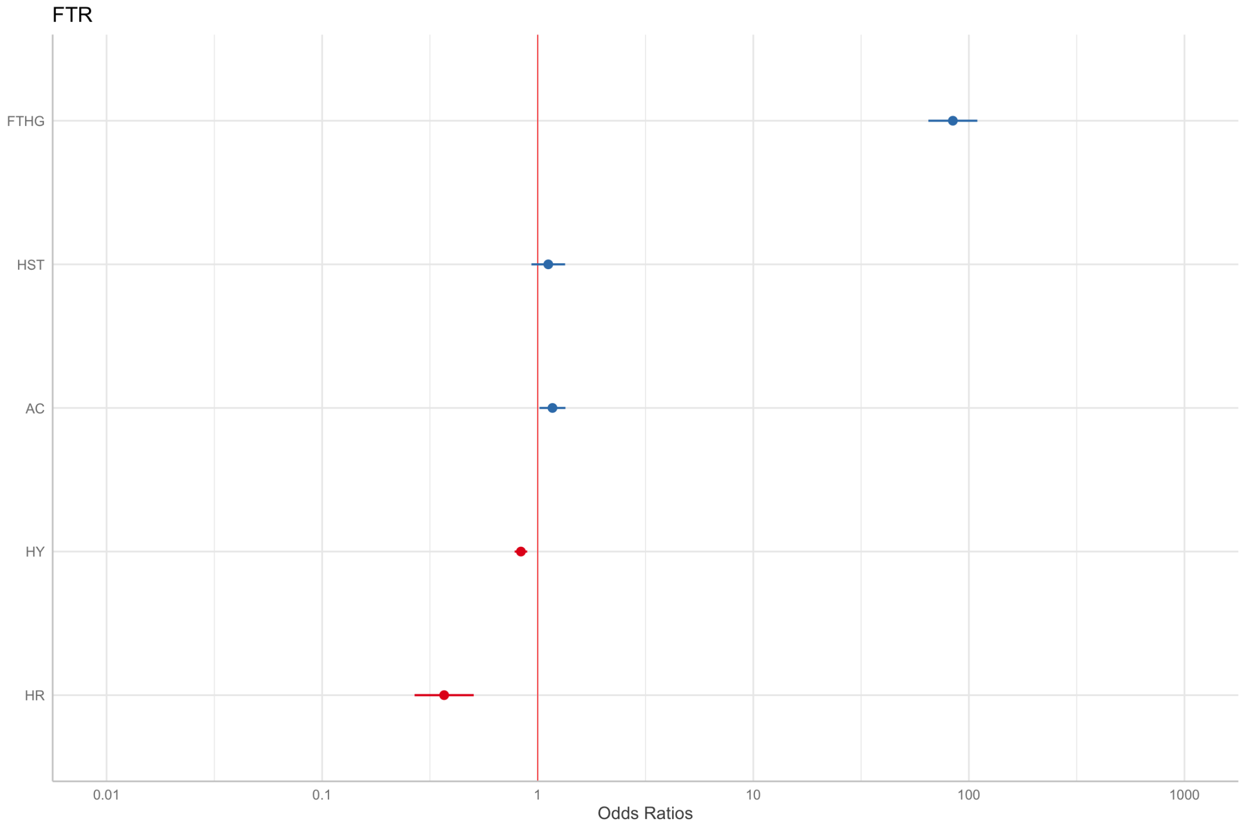
```r
model <- glmer(formula = FTR ~ FTHG + HST + AC + HY + HR
              + (1 + FTHG + HST + HC| HomeTeam)
              + (1 + HF + HY| Referee),
              data = log_EPL_stats,
              family = binomial(link = "logit"))

summary(model)
```

~\

Here is the summary of model(fixed effect) and all variables here are considered as statistically significant at $\alpha = 0.5$ level. To be more clear, a fixed effect parameters are also include in figure 6

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -4.21 | 0.23 | -18.3 | 0.00 *** |
| log_FTHG | 4.43 | 0.13 | 33.4 | 0.00 *** |
| log_HST | 0.11 | 0.09 | 1.16 | 0.25 * |
| log_AC | 0.16 | 0.07 | 2.23 | 0.03 * |
| log_HY | -0.17 | 0.03 | -5.67 | 0.00 *** |
| log_HR | -0.99 | 0.16 | -6.21 | 0.00 ** |



| (Intercept) | FTHG | HST | AC | HY | HR |
|---|---|---|---|---|---|
| -4.2114098 | 4.4340482 | 0.1127628 | 0.1574599 | -0.1794364 | -0.9994892 |

And the following tables are the summary of random effects. The first one is random effect of Team (only display first ten teams alphabetically) and the second one is Positions.

| | (intercept) | FTHG | HST | HC |
|---|---|---|---|---|
| Leicester | -0.10 | -0.04 | -0.05 | 0.06 |
| liverpool | 0.28 | 0.11 | 0.15 | -0.17 |
| Man City | 0.55 | 0.21 | 0.29 | -0.34 |
| Man United | 0.62 | 0.24 | 0.33 | -0.38 |

| | (intercept) | HF | HY |
|---|---|---|---|
| M Atkinson | 0.54 | -0.19 | -0.03 |
| M Clattenburg | -1.04 | 0.36 | 0.07 |
| M Dean | 0.08 | -0.03 | -0.01 |
| M Halsey | 0.01 | 0.00 | 0.00 |
| M Jones | -0.50 | 0.17 | 0.03 |
| M Oliver | -0.54 | 0.19 | 0.03 |

Additionally, a random effect plot for `Team` level are included. From upper left plot of Picture 7, we can come to the conclusion that baseline of salary for each team are quite different. This exactly verify that championship team are willing to pay more money even over the luxury cap. Another parameter that differs most is `Points`, which means scorers suits tastes of some team while not for others.

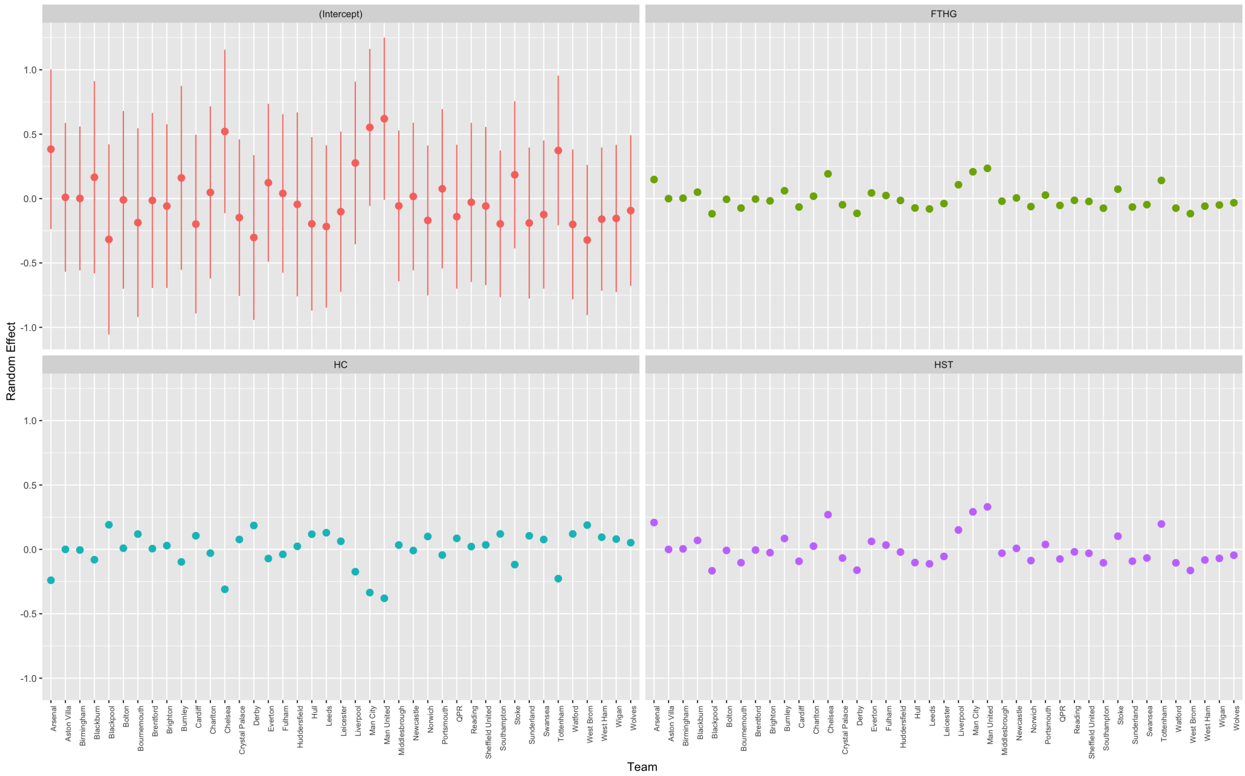


Figure 2: Random Effect of EPL Model

## Result

### Interpretation

Let's take Manchester United Team for example. I fitted multilevel model using predictors FTHG (Full Time Home Team Goals), HST (Home Team Shots on Target), AC (Away Team Corners), HY (Home Team Yellow Cards) and HR (Home Team Red Cards), for the random effect is (1 + FTHG + HST + HC| HomeTeam) and (1 + HF + HY| Referee).

Firstly, we are able to get the following formula of fixed effect, the multilevel model can be written as below:

$$FTR = -4.21+4.43\times log(FTHG+1)+0.11\times log(HST+1)+0.16\times log(AC+1)-0.18\times log(HY+1)-1.00\times log(HR+1)$$

Then add the random effect of Manchester United Team's random effect to the intercepts and slopes and get the estimated formula:

$$FTR = -3.59+4.67\times log(FTHG+1)+0.44\times log(HST+1)+0.16\times log(AC+1)-0.18\times log(HY+1)-1.00\times log(HR+1)$$

Therefore, we can interpret the model fitted by using, for instance, Full Time Home Team Goals. For everyone extra change in Full Time Home Team Goals for all the teams in premier league, when other variables are constant, the Full Time Result(probability of winning the game scale from 0 to 1) will increase by exp(4.43).

Same interpretation with Manchester United. For everyone extra change in Full Time Home Team Goals for Manchester united, when other variables are constant, the Full Time Result(probability of winning the game scale from 0 to 1) will increase by exp(4.67).

## Model Checking

I used two individual plots to check the model, “Residuals vs. Fitted” and “QQ Plot”.

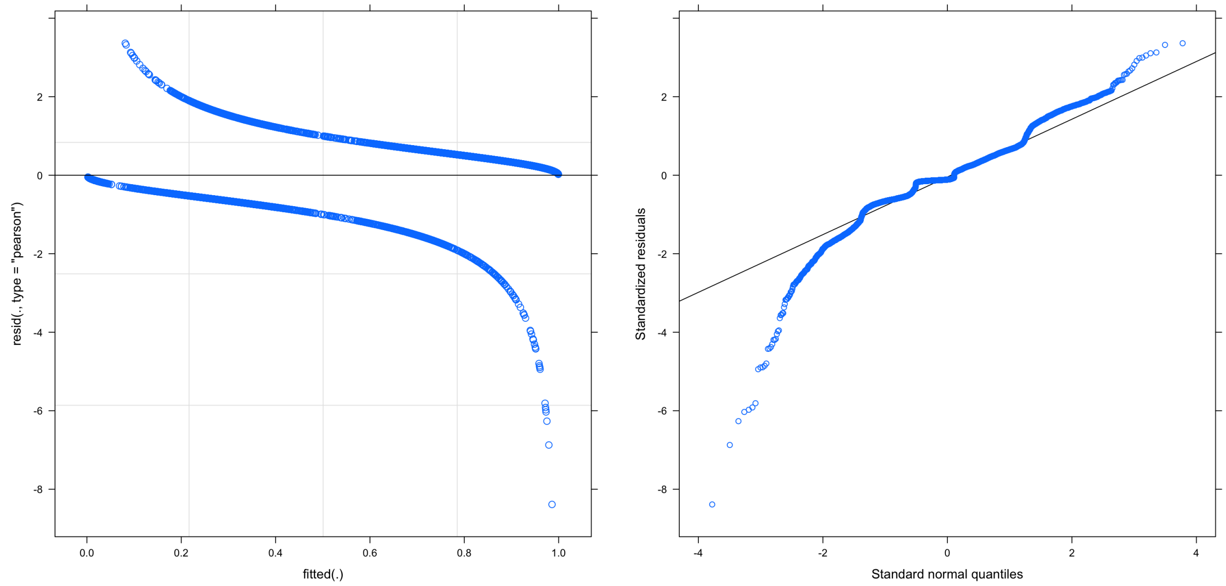


Figure 3: Residual plot and Q-Q plot.

~\

According to the residual analysis, I used several different predictors to fit the multilevel model, the Residual plot indicated that the residual may be ok to fit the model. According to it, the mean value of residuals is approximately 0.

On “QQ Plot”, there are plenty of residual points not on the lines with long tails, it has the issue of low values too low and high values too high for normal, so it might not follow the normal distribution. Thus the normality check fails.

## Discussion

In this report, multilevel model is used to figure out the relationship between matchs' full time result and their several basic on-court stats. Also, this model take two kinds of group level into consideration: match's' home teams and matchs' referees. Generally, from the perspective of fixed effects, predictors like full time home team goals and home team shots on target have positive impacts on winning the game while Fouls, yellow and red cards are always bad. In addition, both in team and position level, the random effects sound reasonable, which means the results can be explained by the characteristics of teams or referees. Finally, several model checks are not that good to support the validity of the model.

This report has limitations. For example, the data set I use covers the past 20 seasons (that is, 20 years). In such a long period of time, whether the rules of the Premier League or the way the game is judged (such as the introduction of VAR), must have been through earth-shaking changes. Therefore, it is unreasonable to ignore the factor of time. For possible future improvements, maybe I'll consider introducing time series analysis. Additionally, the model I fit does not fit as well as expected, which may be related to other unconsidered predictors.

## Reference

[1] Ben Bolker and others. *GLMM FAQ*. http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html

# Appendix

## Variable distributions

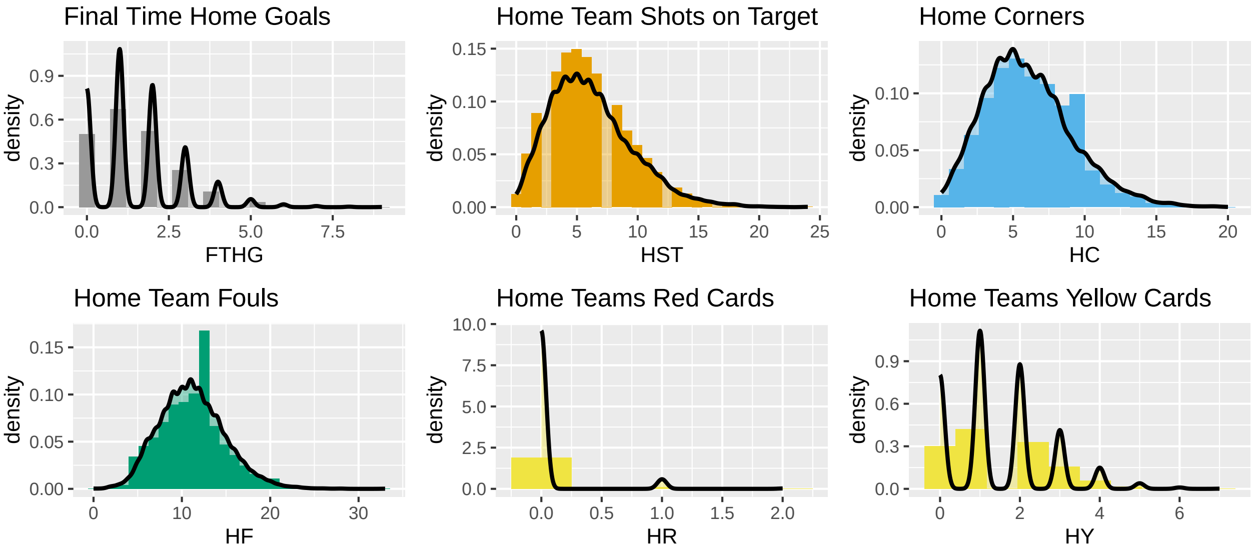


Figure 4: EDA: distribution plots (1)

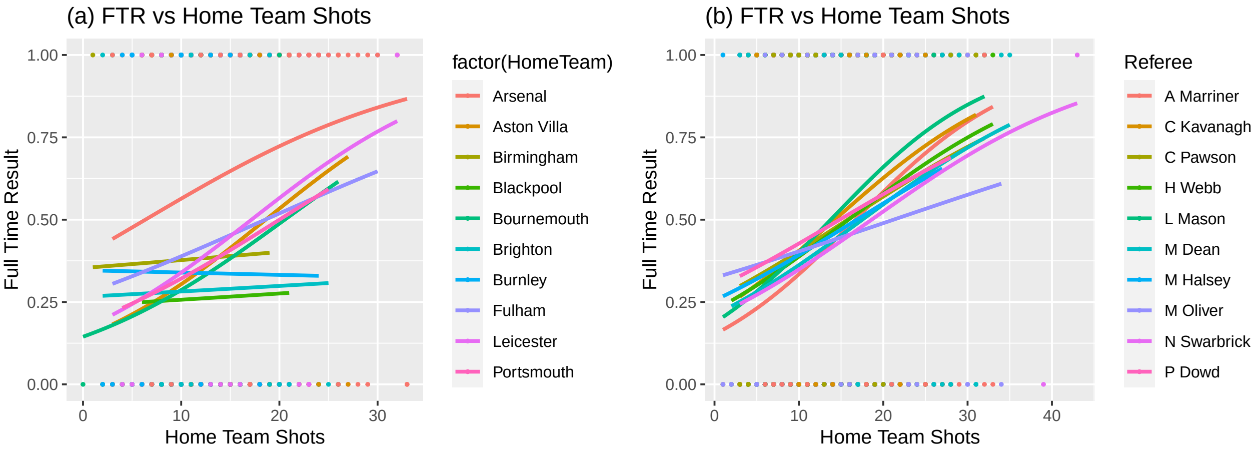


Figure 5: relationship between Full Time Result and Home Team Shots