Boyu Chen 678 Midterm Project

Boyu Chen

December 11 2021

Abstract

This report explores the relationship between the team's total points for each season and their performance during the match. To support this research, a multilevel model is used to indicate performance effects during the match. Almost all the performance positively affects final points except the number of on-target shots. And total goals contribute the most to total points because the data are only from Premium League, so future studies can get more data from different leagues to do further research.

Introduction

The era of data is silently influencing and transforming human life in unimaginable ways, and in sports, the collection and analysis of information makes the sport harmonious and controllable. With the development of technology, soccer has also become surrounded by data and has many tools to interpret this data. While data analysis may not change the outcome of a game, it can have a huge impact on the outcome of a game. For soccer, one of the important ways to make team more competitive is to use data analytics; for data analytics to be used in soccer is a testament to its importance.

One example is the penalty that Pogba conceded in the match between Manchester United and Wolves. After collecting and analyzing data on Pogba's penalty-taking habits, Wolves' data analysis team found that Pogba would kick the penalty more to the left when his team was in a tie or trailing. Because they told Wolves goalkeeper Patricio in advance, Patricio was guarding the penalty was moving to the left in advance, and indeed pounced on the penalty. In combination, data analysis can touch players' habits and subconscious performance, something that is difficult for players to change and adjust. After all, in many cases, it is already set in stone.

The data shows that teams that change their starting lineups usually have a lower probability of winning. I think this is because when a team has a formation that they are used to for a long time, they show consistency in their match, and the players develop a good understanding. Here are some examples: Barcelona's long-standing 4-3-3 formation and Italy's 3-5-2 formation are often used.

My project is to study how the team's performance affects the team's end-of-season point total. I choose the following variables to do this study.

Data Processing

In my data, there is not a variable for total points, so I extracted the number of draws, wins, and losses from data and got the total points by the "number of draws * 1 + number of wins * 3 + number of losses * 0". In addition, because some variables such as total number of tackles and total number of passes are too big, I use log transformation for all the variables used in order to make model more accurate.

Here is the variable I used in this project.

Variables	Explanation
Goals	Number of goal for each team
Offside	Number of offsides in single season
Ontarget	Number of shots on target in single season
Pass	Total number of passing ball in single season
Tackle	Total tackle in single season
Total Points	Total points in single season
Season	Season

EDA Part

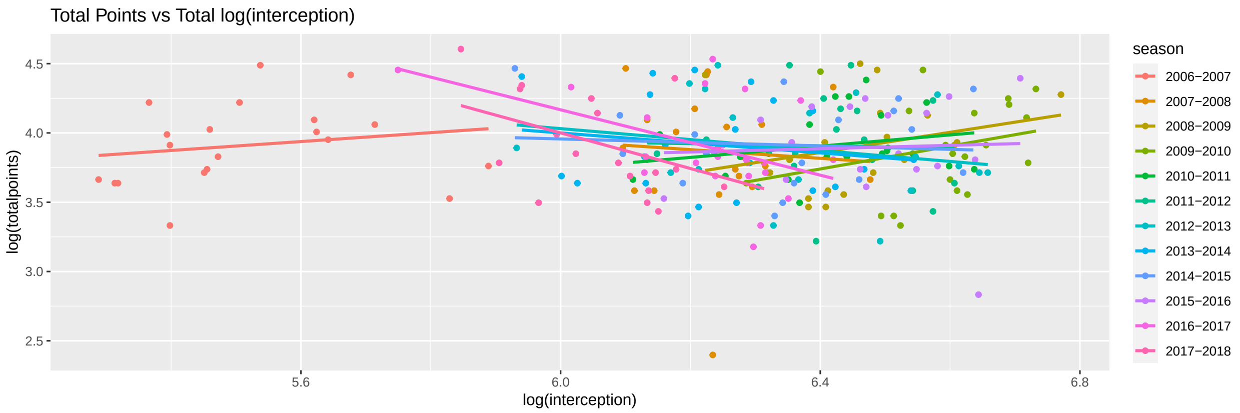


Figure 1 is the relationship between the log(rate) and log(total points) of each team for each season. According to the plot, we can see that at the earliest Premium League(season 2006 - 2007 and season 2008-2012) and most recently Premium League (season 2015-2018), the log(interception) and total points show a positive relationship. And others show a negative relationship.

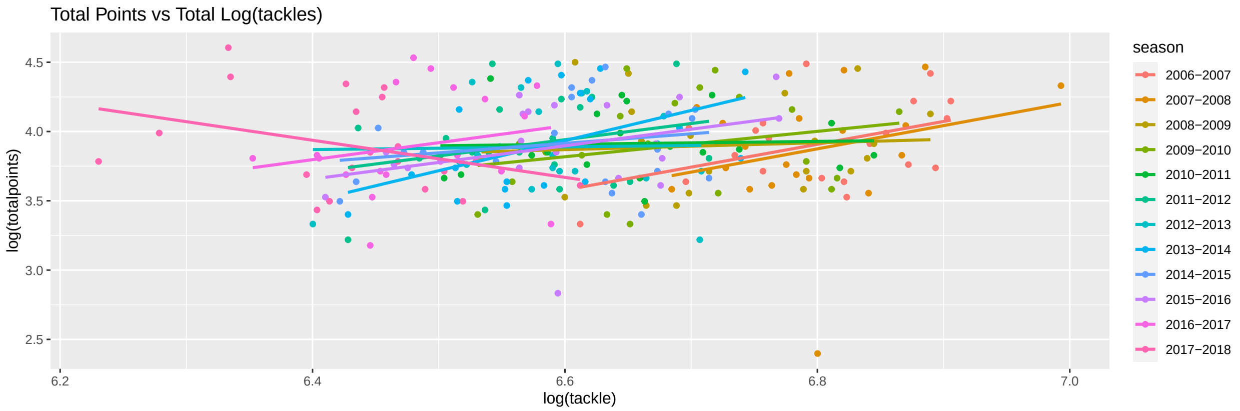


Figure 2 shows the relationship between log(tackles) and log(total points) for each season. We can see that for almost all the season, higher log(tackles) and the number of tackles mean the higher total grades, which means the team's good defence leads to the higher result. However, for the season 2017-2018, we can find a negative relationship between log(tackles) and total points. The possible reason log(tackles) and total points show the negative association in season 2017-2018 is that the matches may have style changes.

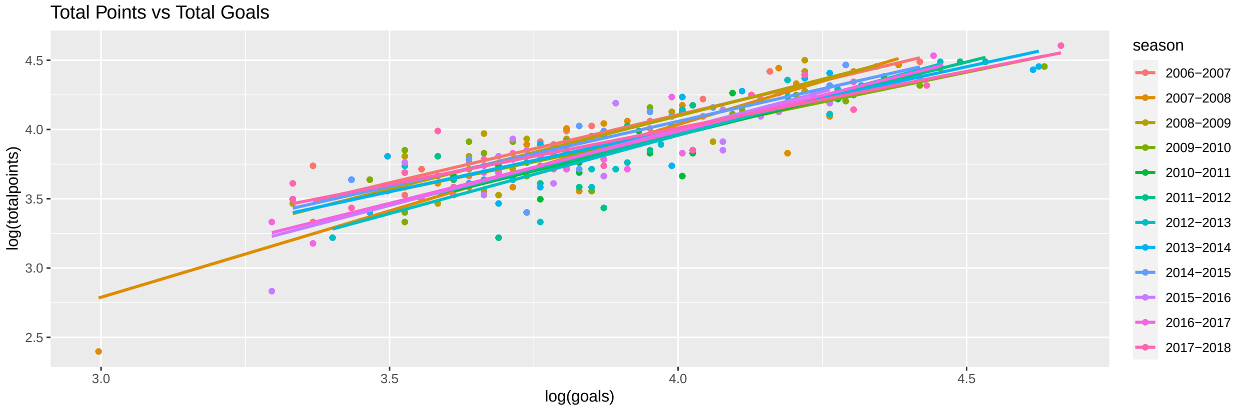


Figure 3 is the relationship between the number of goals and log(total points) for each season. From the plot, we can find that more goals for the team lead to higher total points for all the season. There is an obvious positive relationship between goals and total points. The largest slope for the line in season 2008-2009. Because there is an obvious relationship between goals and total points, so I add this indicator to my model.

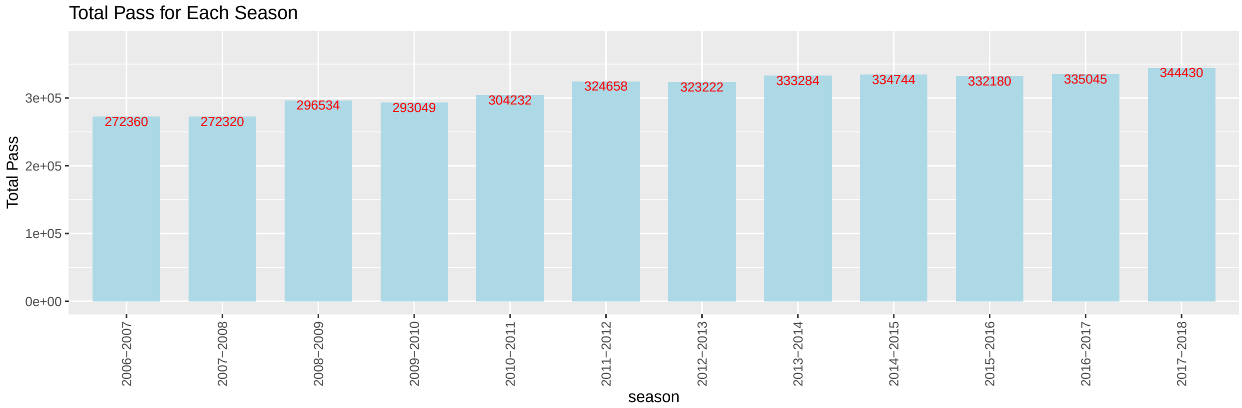


Figure 4 shows that the total pass increase every season which shows that the team try to focus more on controlling the match.

Method

Considering different categories, I will use the multilevel model to fit the data. All of the variables are continuous variables. And I make log transformations for all the continuous variables and produce the log transformation outcomes. Because there is a positive relationship between on-target shots and total goals, more goals mean more on-target shots, so I added the interaction between log(on-target) and log(goals). Here is the function:

Result

We can get the function here

$$\begin{aligned} log(totalpoints) = -3.28 + 1.092 \cdot log(goals) + 0.09 \cdot log(offside) + 0.14 \cdot log(tackle) \\ 0.06 \cdot log(ontarget) + 0.2 \cdot log(pass) - 0.03 \cdot log(ontarget) \cdot log(goals) \end{aligned}$$

If a team doesn't have any performance during the match, it is obvious that the team will lose the point. We can see that almost all the parameter coefficients are bigger than 0, which means that nearly all the parameter coefficients have a positive effect on the outcome. For each 1% difference in goals, the team's predicted difference in total points will be 109%. For each 1% difference in offside, the team's predicted difference in total points will be 9%. For each 1% difference in the tackle, the team's expected difference in total points will be 14%. For each 1% difference in on-target shot, the team's predicted difference in total points will be negative 6%. For each 1% difference in the total pass, the team's expected difference in total points will be 20%, and the coefficient of interaction is -0.033.

Discussion

I can see that when a team scores more goals on the field, their season point total will go up more. This is one of the characteristics that a strong team must have - the ability to attack, and this reflects the trend in soccer where teams will be more on the attack rather than just cowering in defence.
The second most influential factor in total points is the number of passes, reflecting a team's control of the field. When a team has more passes, it will have a higher probability of scoring. In addition, Another variable is the number of passes, which reflects the team's control of the match. Passing the ball between players is always more save energy than stealing the ball from the opponent's feet. Having more stamina means that players can be more comfortable making technical moves and forcing the opposing team to make mistakes. When the opposing team makes more turnovers, the opposing forward players put less pressure on our defence, and our attacking players can put more pressure on the opposing defence and score goals.
The third most influential factor in total points is the number of steals. The number of tackles indicates the defensive ability of the individuals on the team. When the opponent breaks through with the ball, the defender needs to steal the ball and reduce the pressure of the opponent's attacking players on the defence, reducing the probability of the team conceding goals. Although other variables impact the final points, the offensive category is most influential.
These results also reflect the famous saying in the football world: attack is the best defence. With the improvement of individual players' ability, each team will try to improve the team's offensive ability, which is also in line with the current trend of soccer - extreme offence and strengthening the team's control on the field. And we can see that the number of the passing ball increased each season.
One of the interesting things I found is that the on-target shot has negative effects on total points. This shocked me. In my previous perception, the higher the number of shots on target is beneficial for the team. I need to do more research about this area.

Limitation

1. After fitting model, there are only three variables are significant. Below is the fix effects for my model. I can see that only log(goals) ,log(pass) and log(offside) are significant at 5% significant level.I need to do more research or find more data to deal with this problem.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.60	0.98	-2.65	0.008122254
log(goals)	0.92	0.07	13.22	0.000000000
log(offside)	0.10	0.05	2.10	0.035892632
log(tackle)	0.14	0.09	1.52	0.128995889
log(ontarget)	-0.01	0.10	-0.61	0.544973726
log(pass)	-0.01	0.09	2.22	0.026248164

2. Lack of data.I only have the data for Premium League. I still need other league data such as La Liga, Liga 1, Bundesliga, Series A to do more research.

3. It is hard for me to explain the interaction. It is too abstract to explain it.

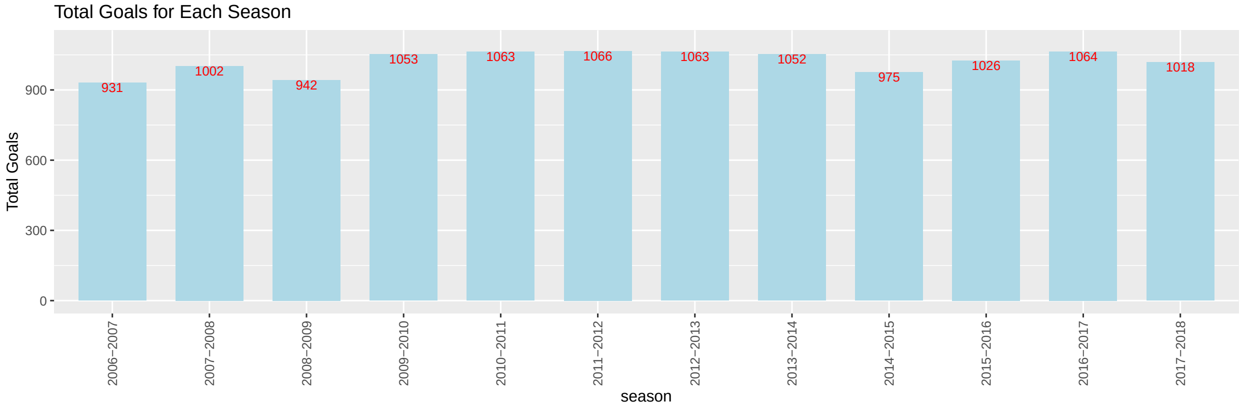
Citation

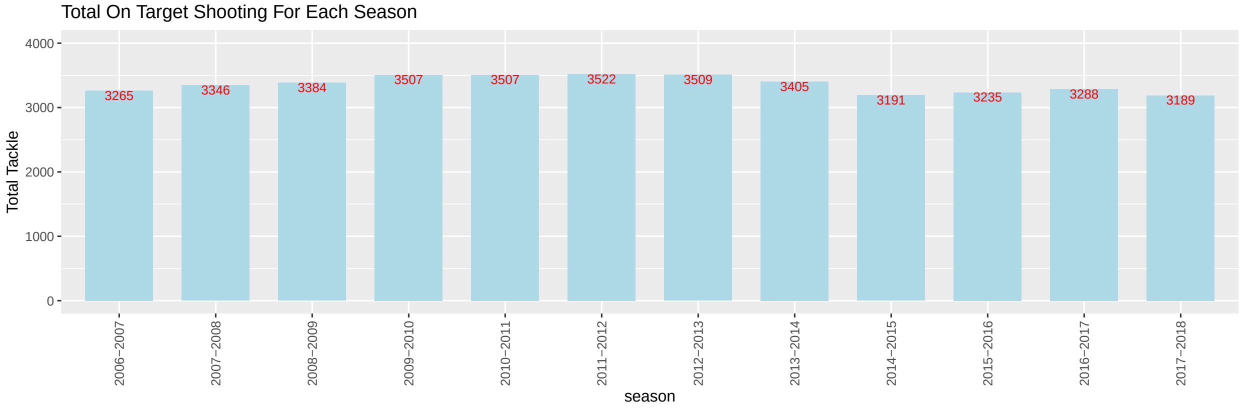
1. https://www.kaggle.com/zaeemnalla/premier-league?select=stats.csv
2. https://soccerblade.com/how-soccer-has-changed/#11-sports%0Ascience
3. https://www.quora.com/How-has-football-soccer-changed-in-the-last-10-years-in-terms-of-tactics-and-overall-game-play

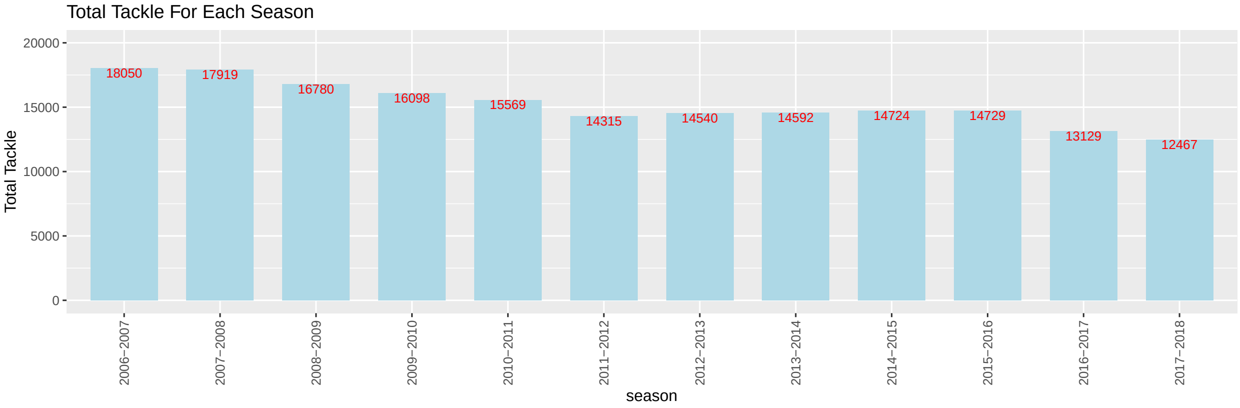
Appendix

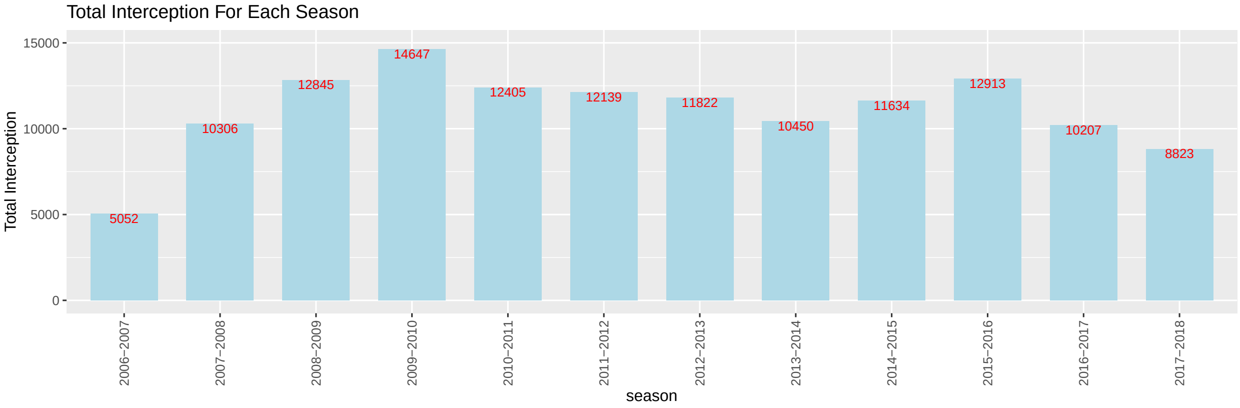
```
##                            Estimate Std..Error    t.value        p.z
## (Intercept)             -3.28561396 3.10483030 -1.0582266 0.28995216
## log(goals)               1.09232369 0.74193258  1.4722681 0.14094849
## log(offside)             0.09516510 0.04656328  2.0437797 0.04097531
## log(tackle)              0.14250553 0.09570192  1.4890561 0.13647261
## log(ontarget)            0.06644542 0.57862022  0.1148342 0.90857651
## log(pass)                0.19767316 0.08885722  2.2246156 0.02610705
## log(goals):log(ontarget) -0.03372843 0.14453232 -0.2333625 0.81547991
```

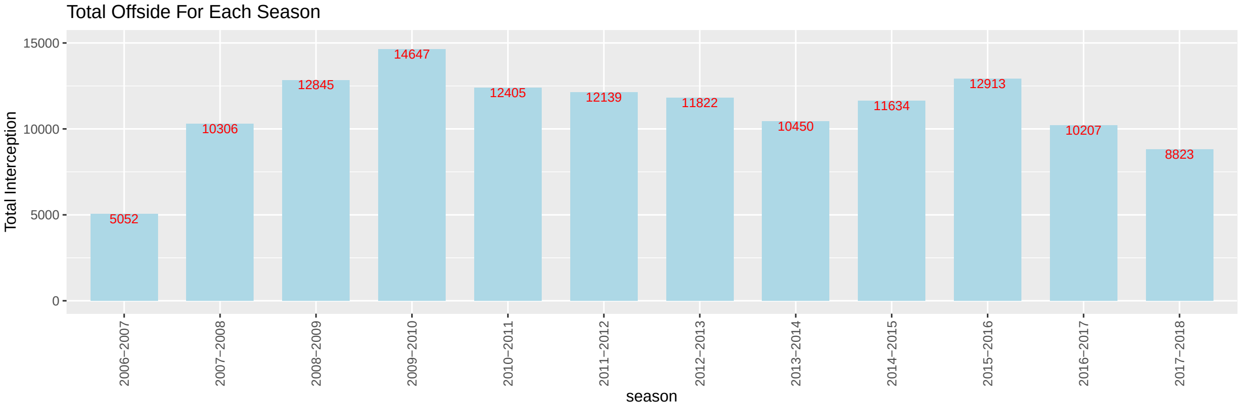
Other Eda

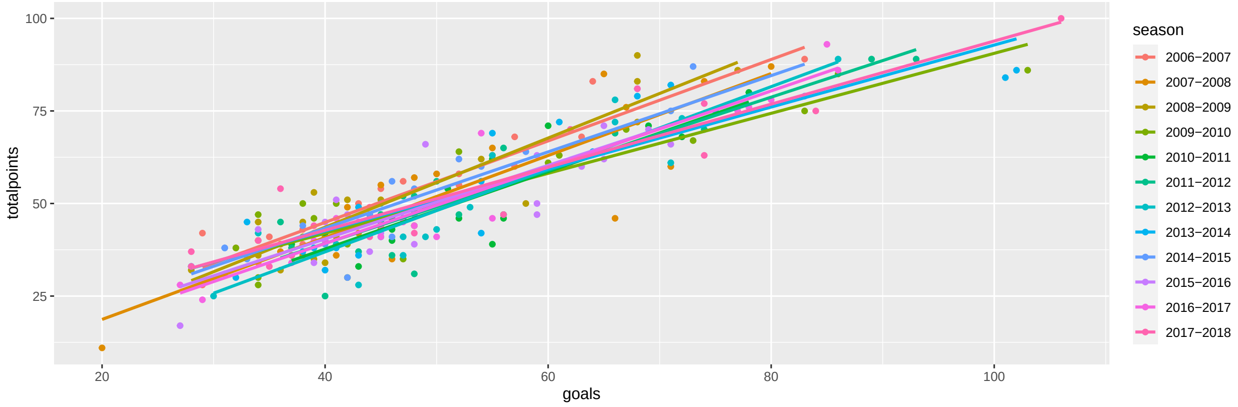


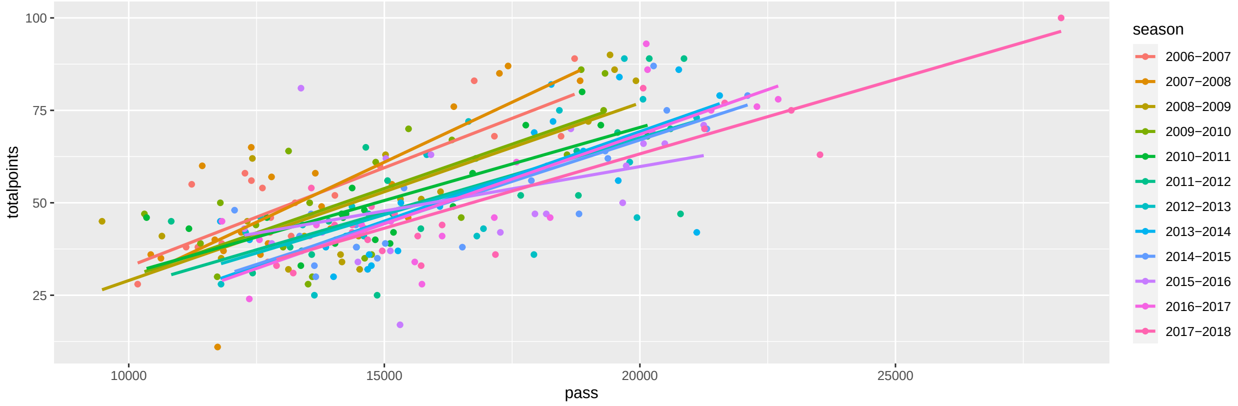


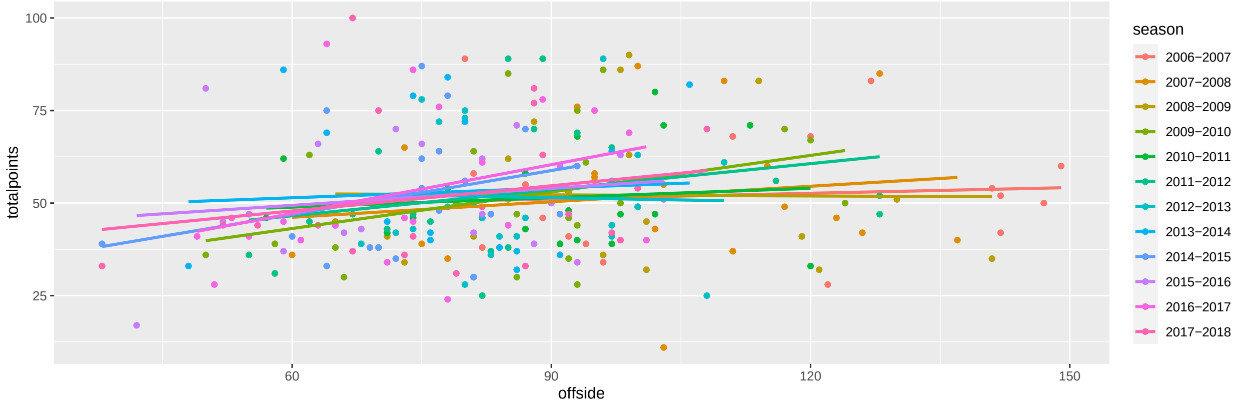

Total Tackle For Each Season

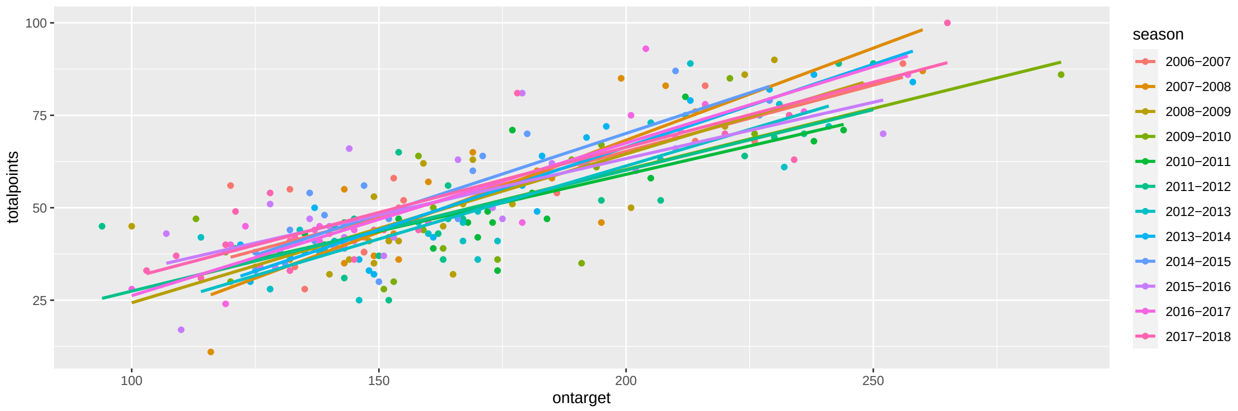

Total Interception For Each Season


Total Offside For Each Season

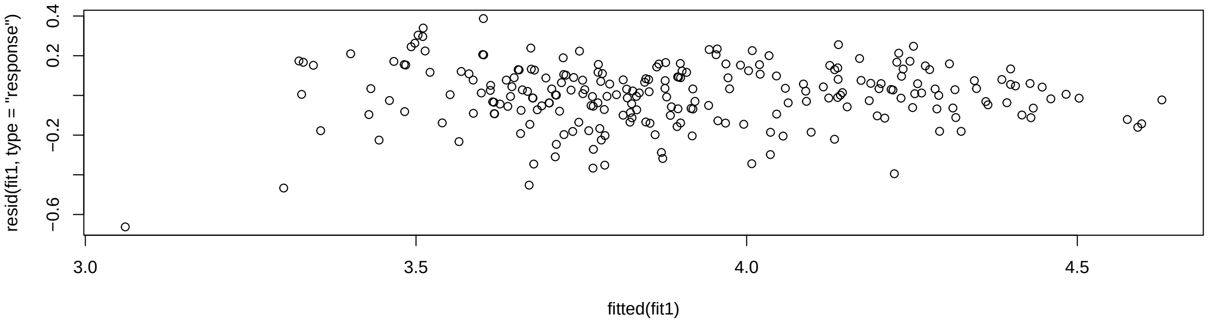

totalpoints
100
75
50
25
20
40
60
80
100
goals
season
2006−2007
2007−2008
2008−2009
2009−2010
2010−2011
2011−2012
2012−2013
2013−2014
2014−2015
2015−2016
2016−2017
2017−2018

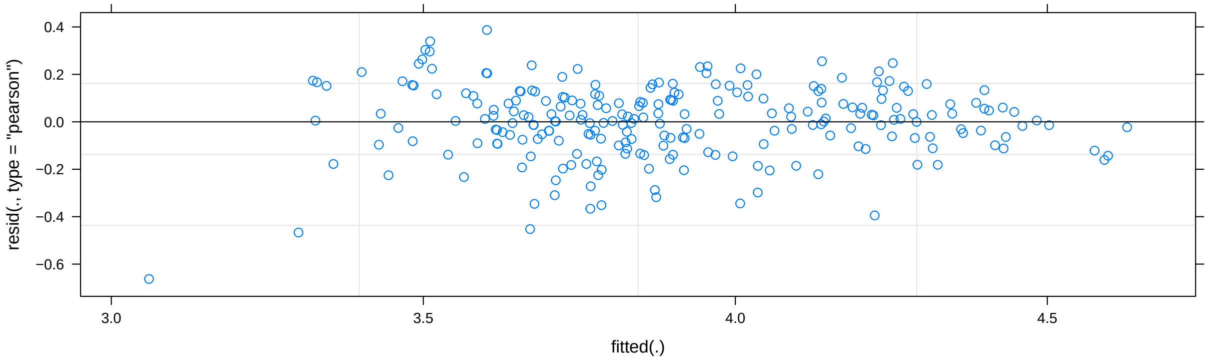

totalpoints
100
75
50
25
10000
15000
20000
25000
pass
season
2006−2007
2007−2008
2008−2009
2009−2010
2010−2011
2011−2012
2012−2013
2013−2014
2014−2015
2015−2016
2016−2017
2017−2018

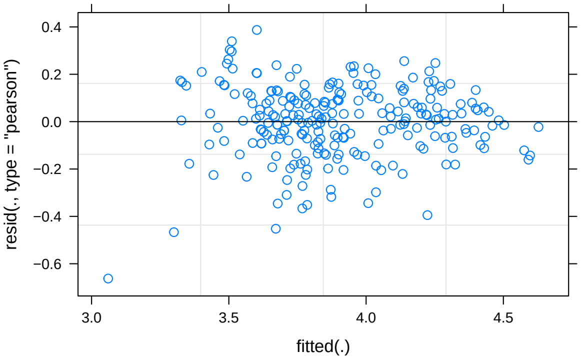

totalpoints
100
75
50
25
60
90
120
150
offside
season
2006−2007
2007−2008
2008−2009
2009−2010
2010−2011
2011−2012
2012−2013
2013−2014
2014−2015
2015−2016
2016−2017
2017−2018

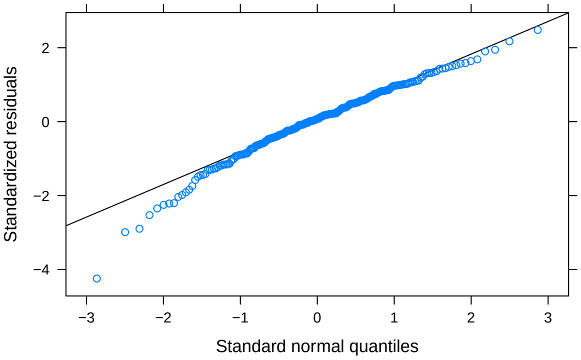


Model Validation









AIC(fit1) -166.001

Supplement

```
# ##data
# library(tidyverse)
# library(lme4)
# library(arm)
# library(gridExtra)
# df<-read.csv("stats.csv",header = TRUE)
# goals<-df$goals
# offside<-df$total_offside
# ontarget<-df$ontarget_scoring_att
# interception<-df$interception
# pass<-df$total_pass
# season<-df$season
# tackle<-df$total_tackle
# win<-df$wins
# losse<-df$losses
# tie<-38-win-losse
# totalpoints<-win*3+tie*1+losse*0
# colnames<-c("Goals","Offside","Ontarget","Pass","Tackle","Total Points","Season")
# explaination<-c("Number of goal for each team","Number of offsides in single season","Number of shots
# table <- cbind(colnames, explaination)
# colnames(table) <- c("Variables", "Explanation")
# knitr::kable(table, "pipe")
#
# ##eda
# ggplot(data=df, mapping=aes(x=log(interception), y=log(totalpoints), group=season)) +
#   geom_point(aes(color=season))+
#   geom_smooth(se = F,aes(color = season), method = "lm")+
#   labs(title = "Total Points vs Total log(interception)")
#
# ggplot(data=df, mapping=aes(x=log(tackle), y=log(totalpoints), group=season)) +
#   geom_point(aes(color=season))+
#   geom_smooth(se = F,aes(color = season), method = "lm")+
#   labs(title = "Total Points vs Total Log(tackles)")
#
```

```
# ggplot(data=df, mapping=aes(x=log(goals), y=log(totalpoints), group=season)) +
#   geom_point(aes(color=season))+
#   geom_smooth(se = F,aes(color = season), method = "lm")+
#   labs(title = "Total Points vs Total Goals")
#
# totalpass<-aggregate(pass, by=list(season=df$season), FUN=sum)
# ggplot(data = totalpass, mapping = aes(x = season, y = x,label = season))+
#   geom_bar(stat = ’identity’,fill = ’lightblue’,width = 0.7)+
#   geom_text(mapping = aes(label = x), size = 3, colour = ’red’, vjust = 1, hjust = .5,
#             position = position_dodge(0.9))+
#   theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
#   labs(x = "season",y = "Total Pass",title = "Total Pass for Each Season")+ylim(0,380000)
#
#
# ##Model
#
# fit1<-lmer(log(totalpoints)~log(goals)+log(offside)+log(tackle)+log(ontarget)+log(pass)+log(goals)*log
#              (1+log(pass)|season),data = df)
#
#
# ##Appendix
#
# extract coefficients
# coefs <- data.frame(coef(summary(fit1)))
# use normal distribution to approximate p-value
# coefs$p.z <- 2 * (1 - pnorm(abs(coefs$t.value)))
#
#
#
#
#
# totalgoals<-aggregate(goals, by=list(season=df$season), FUN=sum)
# ggplot(data = totalgoals, mapping = aes(x = season, y = x,label = season))+
#   geom_bar(stat = ’identity’,fill = ’lightblue’,width = 0.7)+
#   geom_text(mapping = aes(label = x), size = 3, colour = ’red’, vjust = 1, hjust = .5,
#             position = position_dodge(0.9))+
#   theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
#   labs(x = "season",y = "Total Goals",title = "Total Goals for Each Season")+ylim(0,1100)
#
# totalontarget<-aggregate(ontarget, by=list(season=df$season), FUN=sum)
# ggplot(data = totalontarget, mapping = aes(x = season, y = x,label = season))+
#   geom_bar(stat = ’identity’,fill = ’lightblue’,width = 0.7)+
#   geom_text(mapping = aes(label = x), size = 3, colour = ’red’, vjust = 1, hjust = .5,
#             position = position_dodge(0.9))+
#   theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
#   labs(x = "season",y = "Total Tackle",title = "Total On Target Shooting For Each Season")+ylim(0,4000
#
# totaltackle<-aggregate(tackle, by=list(season=df$season), FUN=sum)
# ggplot(data = totaltackle, mapping = aes(x = season, y = x,label = season))+
#   geom_bar(stat = ’identity’,fill = ’lightblue’,width = 0.7)+
#   geom_text(mapping = aes(label = x), size = 3, colour = ’red’, vjust = 1, hjust = .5,
#             position = position_dodge(0.9))+
#   theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
#   labs(x = "season",y = "Total Tackle",title = "Total Tackle For Each Season")+ylim(0,20000)
```

```
#
# totalinterception<-aggregate(interception, by=list(season=df$season), FUN=sum)
# ggplot(data = totalinterception, mapping = aes(x = season, y = x,label = season))+
#   geom_bar(stat = 'identity',fill = 'lightblue',width = 0.7)+
#   geom_text(mapping = aes(label = x), size = 3, colour = 'red', vjust = 1, hjust = .5,
#             position = position_dodge(0.9))+
#   theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
#   labs(x = "season",y = "Total Interception",title = "Total Interception For Each Season")+ylim(0,1500
#
# totaloffside<-aggregate(offside, by=list(season=df$season), FUN=sum)
# ggplot(data = totalinterception, mapping = aes(x = season, y = x,label = season))+
#   geom_bar(stat = 'identity',fill = 'lightblue',width = 0.7)+
#   geom_text(mapping = aes(label = x), size = 3, colour = 'red', vjust = 1, hjust = .5,
#             position = position_dodge(0.9))+
#   theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
#   labs(x = "season",y = "Total Interception",title = "Total Offside For Each Season")+ylim(0,15000)
#
# ggplot(data=df, mapping=aes(x=goals, y=totalpoints, group=season)) +
#  # geom_line(aes(linetype=season,color=season))+
#   geom_point(aes(color=season))+
#   geom_smooth(se = F,aes(color = season), method = "lm")
#
# ggplot(data=df, mapping=aes(x=pass, y=totalpoints, group=season)) +
#  # geom_line(aes(linetype=season,color=season))+
#   geom_point(aes(color=season))+
#   geom_smooth(se = F,aes(color = season), method = "lm")
#
# ggplot(data=df, mapping=aes(x=offside, y=totalpoints, group=season)) +
#  # geom_line(aes(linetype=season,color=season))+
#   geom_point(aes(color=season))+
#   geom_smooth(se = F,aes(color = season), method = "lm")
#
# ggplot(data=df, mapping=aes(x=ontarget, y=totalpoints, group=season)) +
#  # geom_line(aes(linetype=season,color=season))+
#   geom_point(aes(color=season))+
#   geom_smooth(se = F,aes(color = season), method = "lm")
# re <- plot(fit1)
# qq <- lattice::qqmath(fit1)
# plot(fitted(fit1), resid(fit1,type = "response"))
# plot(fit1)
# grid.arrange(re,qq,nrow=1)
# aic(fit1)
```