How to analysis changes in NBA games' playing style by analyzing the changes of stats of players in NBA

Getong Zhong

2022-12-16

Introduction & literature review

In this project, I would like to analyze the change in NBA play style between 2011 to 2020 years by looking at the difference in stats that players produce on the court. Due to the big variation of players' stats between the playoff and regular season, especially for star players, I would only consider regular season stats in this project. The performance of NBA players across different positions can be compared to each other in the same standard with the same data set. Unlike the NFL, a player's performance across different positions is measured across different categories of stats.

I didn't find a paper that uses the same data as I do, but there are many of them which use similar data or write with a similar purpose as I do. In the paper "Predicting NBA Player Performance", the author used a more advanced machine learning algorithm to predict the performance of individual NBA players. The author mentioned the extension of predicting the outcomes of games by predicting the scoring performances of each player and summing them up. Therefore, the prediction of an individual's performance can also contribute to the win-loss classifier. In the results of this paper, the author found out that the standard deviation of a player's performance is usually big due to injury and coach strategies which are all unpredictable but always affect the player's performance a lot. The author also mentioned another important factor, the confidence level of a player. The confidence level of a player is always determined by not only how well he is but also by how well his teammates or opponents play in the game, and his confidence level could affect his performance. But the author has not yet found a way to present a player's confidence level in the regression model, since such a confidence level is not numeric and does not have a standard to measure. In another paper, "Trends in NBA and Euroleague basketball: Analysis and comparison of statistical data from 2000 to 2017", the author compares NBA and Euroleague basketball by the box-score statistics. They found out that the differences in win-loss determining factors are very small. From both leagues, game results are normally determined by four factors that could explain most of the variation in their game results. Considering the biggest difference between Euroleague and NBA, they mentioned the "game pace", whereas in NBA there are more numbers of possessions per game. In particular, the numbers of blocks per game, fouls per game, and the number of free throws per game (free throws are usually caused by opponent's fouls) in NBA games are higher than in Euroleague. This can be possibly explained by the fact that NBA players have better athleticism and the Euroleague puts more emphasis on plays and tactical aspects of basketball.

Such results from the previous study are helpful to my project. As known fact that players' performance always variates a lot, I will not only consider points they made in the game as the only standard to measure their performance, but I will also include other stats such as assists, rebounds, steals, blocks, turnovers into the measurements of their on-court performance. Besides, in the second article that contrasts the NBA and Euroleague, the author mentioned game pace difference and compare the difference in athleticism between NBA and Euro players which can explain most of the variation between the NBA and Euroleague. With that in mind, and knowing NBA has more stretch Four, Stretch Five in nowadays game than before, if I

Table 1: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
p/g	1880	13.971	5.304	2.927	10.118	17.054	36.128
3/g	1880	1.205	0.924	0	0.319	1.846	5.349
r/g	1880	5.459	2.549	0	3.524	6.799	15.987
a/g	1880	3.031	2.117	0.125	1.457	4.148	11.738
s/g	1880	0.975	0.399	0	0.685	1.188	2.517
b/g	1880	0.634	0.53	0	0.257	0.856	3.685
fgftto/g	1880	1.74	0.804	0	1.141	2.21	5.728

want to compare the difference in play style with the athleticism of players does not change a lot, I should take a look into other angles such as 3-pointers per game, assists per game to analysis the change of the game.

Methods and Sample

Table of Descriptive Statistics

In this paper, I'm using the box score from the top 188 NBA players (determined by YAHOO) from 2011 to 2020. In this data, there are 9 different variables.
-p/g, stands for numbers of points a player averagely makes per game.
-3/g, stands for numbers of 3-pointers a player averagely makes per game.
-r/g, numbers of stands for rebounds a player averagely grabs per game.
-a/g, numbers of stands for assists a player averagely makes per game.
-s/g, stands for numbers of steals a player averagely makes per game.
-b/g, stands for numbers of blocks a player averagely makes per game.
-fg%, stands for percentage of field goal a player averagely has per game (calculated by field goal make/ field goal attempts)
-ft% stands for percentage of free threw a player averagely has per game (calculated by free throw make / free throw attempts), to/g, stands for number of turnovers a player averagely makes per game.

From the table, we can see that the standard deviation for every variable is high and the gap between the 75 percentile and the Max value is also big. Among all the variables Field goal percentage has the smallest standard deviation and half of the players can average their field goal percentage in the range between 43% to 50 %, which also shows field goal percentage is the variable that is the most predictable among others. "3/g" has the most variation with a mean value of it is 1.2 and half of the player average of around 0.319 to 1.846 but the maximum is 5.349 means that 3 pointers ability among players variates a lot, there are some of the players are much better at 3-point-shooting than others.

Graphical and Statistical methods

In this project, I will perform an exploratory analysis of trends by plotting the variables over time. In each scatterplot, I will use the mean value of different variables over time to show the trend. For example, if I get a plot of average turnovers per game and observe a line with a positive slope (in an increasing trend), that means that NBA players tend to make more and more turnovers in today's game than in the past. This outcome is not only caused by one factor but a net outcome that is affected by multiple factors, I will analyze each trend and give my interpretation based on each result. In addition, I want to create a model to predict a player's scoring ability based on other variables. To do that I will first show the correlation table between all the variables and the scatter plots between each variable and "p/g" (average points per game) so that I can decide what kind of regression model I would like to use, and what variables I would

like to add to the model, and check if transformation is necessary to the model.

Before all of that, I will do some modifications to the data, so that I can make my later process easier. The initial data I got had 16 variables in total, I first deleted some of them and remain the variables that I think can present the player's performance more directly. And then I put a 0 value to the missing data to make it in good integrity. I also make 9 data frames from the initial data to get the mean value of each variable with the corresponding Year, so that it is convenient for me to create plots about those values later on.

Findings

Visulization and Graphs

Before presenting the graphs, I would like to present the correlations between each variable and interpret some of them that might support my later interpretation of the graphs. In the correlation chart we can see that points (p/g) and turnovers (to/g), points (p/g) and threes (3/g) have a high positive relationship (0.72 and 0.47 respectively) which means a player tends to score more points in a game usually might also turn the ball over more times, and a player who can make more threes can also likely to score more points. These positive relationships between them are reasonable, since the player who can score more usually touches the ball more, and if they have the ball in their hand longer than other players, the chance to turn the ball over for them is also higher. The positive relationship between threes and points is also straightforward, if you can make 3 points per shot rather than 2 points per shot, you can easily score more points. However, it is more difficult to make a 3-point shot than 2 point shot, so the relationship between points per game and 3-pointer per game is not that high (only 0.44). Rebound (r/g) and blocks (b/g) also has a high positive relationship (0.62). This also makes a lot of sense, since taller players are more likely to protect the rebounds and they are also easier to block other players as they are taller.

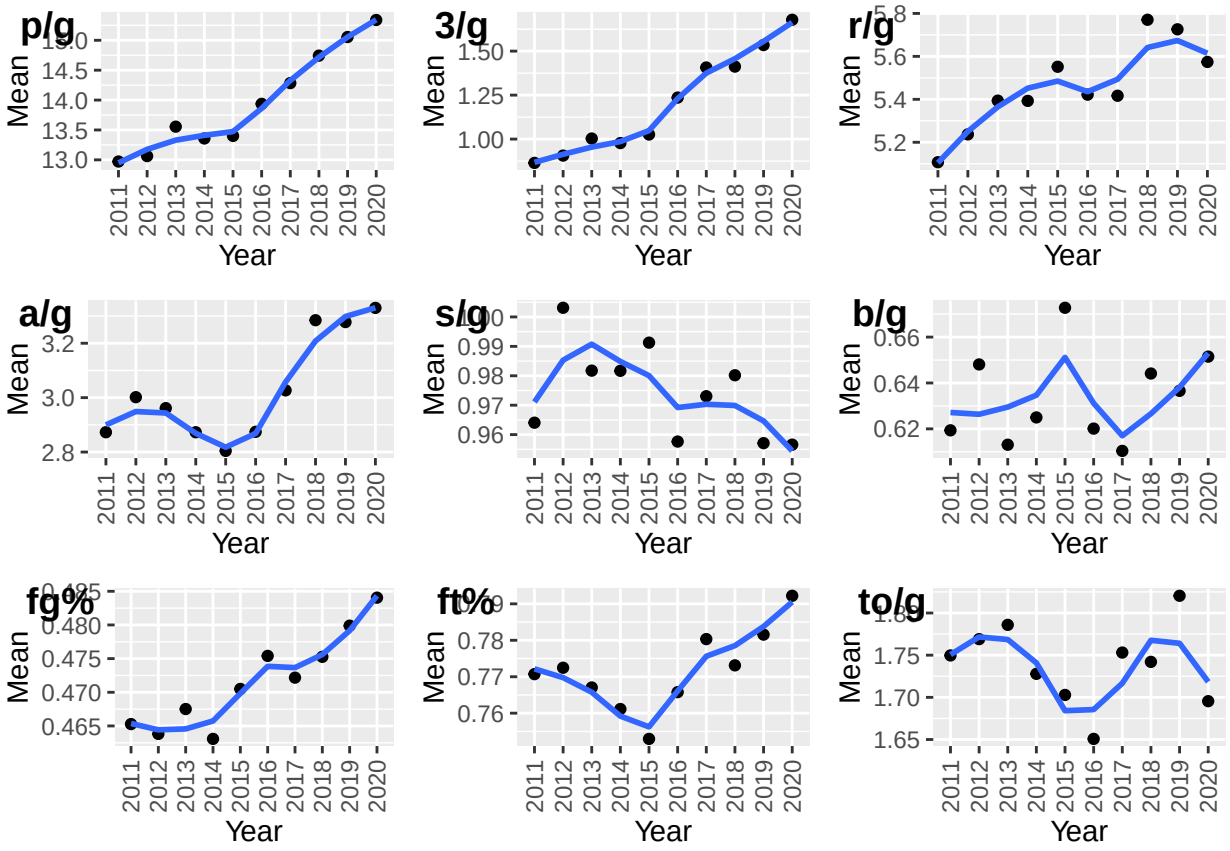
```
##              p/g        3/g        r/g        a/g        s/g         b/g
## p/g   1.00000000  0.4461201  0.2147169  0.4916469  0.3002498 -0.03188566
## 3/g   0.44612009  1.0000000 -0.4168977  0.3002136  0.1950160 -0.44922537
## r/g   0.21471692 -0.4168977  1.0000000 -0.1464473 -0.0683118  0.61691210
## a/g   0.49164693  0.3002136 -0.1464473  1.0000000  0.5422372 -0.31426057
## s/g   0.30024976  0.1950160 -0.0683118  0.5422372  1.0000000 -0.18528892
## b/g  -0.03188566 -0.4492254  0.6169121 -0.3142606 -0.1852889  1.00000000
## fg%  -0.07249609 -0.5398602  0.5145510 -0.2812634 -0.2740825  0.53144308
## ft%   0.32615079  0.5129999 -0.3921333  0.2276559  0.0578033 -0.42799175
## to/g  0.72438743  0.2039304  0.1921424  0.7806362  0.4687238 -0.05642223
##              fg%        ft%        to/g
## p/g  -0.07249609  0.3261508  0.72438743
## 3/g  -0.53986022  0.5129999  0.20393043
## r/g   0.51455096 -0.3921333  0.19214244
## a/g  -0.28126337  0.2276559  0.78063617
## s/g  -0.27408249  0.0578033  0.46872381
## b/g   0.53144308 -0.4279918 -0.05642223
## fg%   1.00000000 -0.4486309 -0.12150298
## ft%  -0.44863087  1.0000000  0.14382947
## to/g -0.12150298  0.1438295  1.00000000
```

Now we can take a look at the graphs. In the graph of "p/g", points scored per game by an NBA player, we can see a clear positive line that shows NBA players tend to score more points from 2011 to 2020. Meanwhile, a similar positive line shows in the graph of "3/g", this can easily relate to the positive relationship between "p/g' and"3/g", as more three-pointers are made in the game, average points a player scores and the total scores are all going to increase. However, as another variable that has a strong positive relationship with

the “p/g”, “to/g” doesn’t have the same graph, it goes up and down every year since 2011, and we can conclude a kind of negative trend is there more the turnovers per game. Similar to “to/g”, “s/g” (steals per game) has almost the same graph as “to/g”, except for the year 2019. Recall the correlations chart, we can observe a positive relationship (0.468) between these two variables as well. In real life, a correlation between turnovers and steals is also reasonable. Players get one turnover when they lose control of the ball, sometimes the ball goes out of the bound and switches the possessions of the ball or it falls to an opponents player’s hand, and that count as a steal for that opponent player, that why the positive relationship makes sense. Although not all turnovers will result in a steal many of them did. In the graph of “a/g” (assist per game) and “ft%”, the trend is surprisingly similar to each other, worth mentioning that the year 2015 has the lowest in assists, free-throw percentage, and highest in blocks. These strange results can only be explained by the data of 2015, it included more C and PF players (who are usually good and blocking, but bad in assists and ft%). Besides, the negative relationship between blocks and free-throw percentage (-0.42) can also prove this interpretation. Considering the changes in the playing style of NBA players in general, the biggest difference is that they likely to make more three-pointers. And the NBA game is more pleasant to watch nowadays as an audience than in the past: you can see they make more shots with a higher field goal percentage, and they are not likely to turn the ball over so the game has a better fluency than before, they play more like a team, and try harder to share the ball with their teammate on the court. Those are all the improvements in NBA, improvements in modern Basketball.

```
## Warning: Continuous limits supplied to discrete scale.
## Did you mean `limits = factor(...)` or `scale_*_continuous()`?
## Continuous limits supplied to discrete scale.
## Did you mean `limits = factor(...)` or `scale_*_continuous()`?
## Continuous limits supplied to discrete scale.
## Did you mean `limits = factor(...)` or `scale_*_continuous()`?
## Continuous limits supplied to discrete scale.
## Did you mean `limits = factor(...)` or `scale_*_continuous()`?
## Continuous limits supplied to discrete scale.
## Did you mean `limits = factor(...)` or `scale_*_continuous()`?
## Continuous limits supplied to discrete scale.
## Did you mean `limits = factor(...)` or `scale_*_continuous()`?
## Continuous limits supplied to discrete scale.
## Did you mean `limits = factor(...)` or `scale_*_continuous()`?
## Continuous limits supplied to discrete scale.
## Did you mean `limits = factor(...)` or `scale_*_continuous()`?
## Continuous limits supplied to discrete scale.
## Did you mean `limits = factor(...)` or `scale_*_continuous()`?
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Regression model

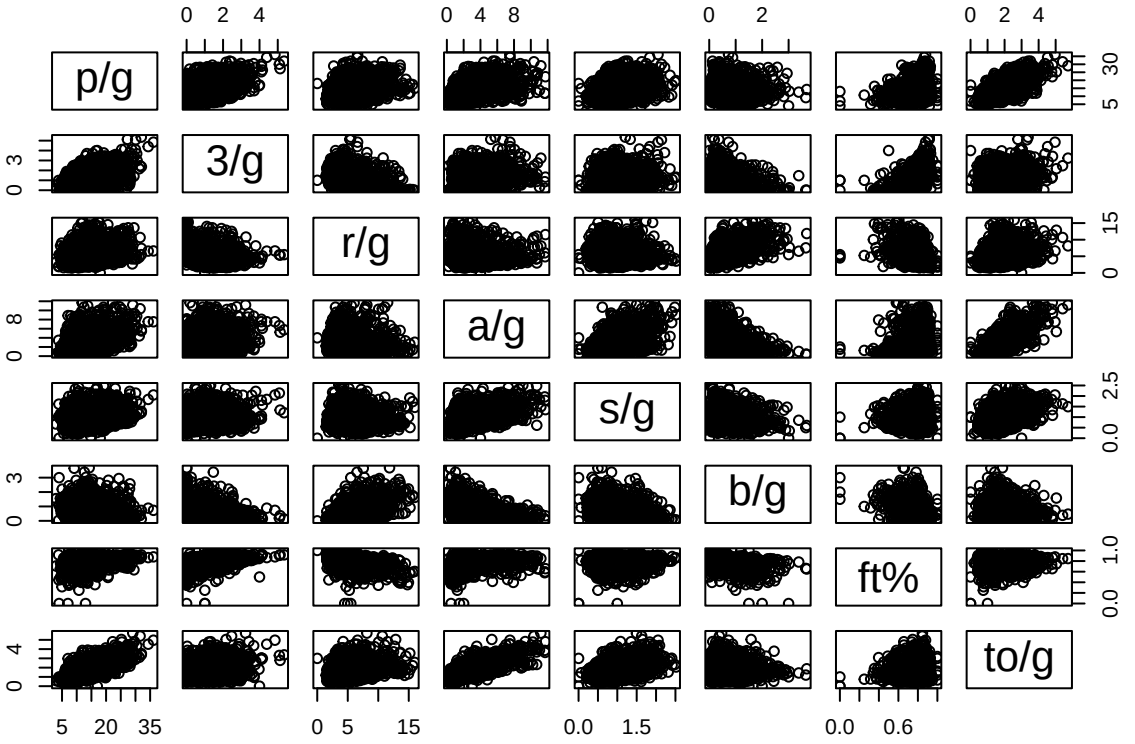
In the second part of the Findings, a regression model will be used to predict the points per game a player could averagely predict based on all the data from 2011 to 2020. Recall the correlation table block per game and fg% have a very weak relationship with "p/g". Therefore, the initial model will only include the other 6 variables.

```
##
## Call:
## lm(formula = `p/g` ~ `3/g` + `r/g` + `to/g` + `fg%` + `ft%` +
##     `s/g` + `a/g`)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5579 -1.8293 -0.1868  1.5229 16.5796
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.11809    1.01271 -14.928  < 2e-16 ***
## `3/g`         2.49083    0.08994  27.695  < 2e-16 ***
## `r/g`         0.47233    0.03481  13.569  < 2e-16 ***
## `to/g`        4.58618    0.14936  30.706  < 2e-16 ***
## `fg%`        16.49739    1.40949  11.705  < 2e-16 ***
## `ft%`        11.09076    0.77498  14.311  < 2e-16 ***
## `s/g`         0.42016    0.19553   2.149   0.0318 *
```

```
## `a/g`        -0.40221    0.05691  -7.068 2.21e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.752 on 1872 degrees of freedom
## Multiple R-squared:  0.7318, Adjusted R-squared:  0.7308
## F-statistic: 729.8 on 7 and 1872 DF,  p-value: < 2.2e-16
```

An R-squared value of 0.73 refers to a good fit and all the predictors, are highly statistically significant. But we should check out the scatter plots and diagnostic plots.

From the scatterplots, we can see that there exists some evidence of a non-linear relationship between the response and the predictors, though not a lot. However, inter-predictor scatters are quite curvy, which will lead to problems with std. residuals analysis and unknown model misspecification. Also, some predictors seem very highly correlated among themselves. Let us recall the correlation matrix again:



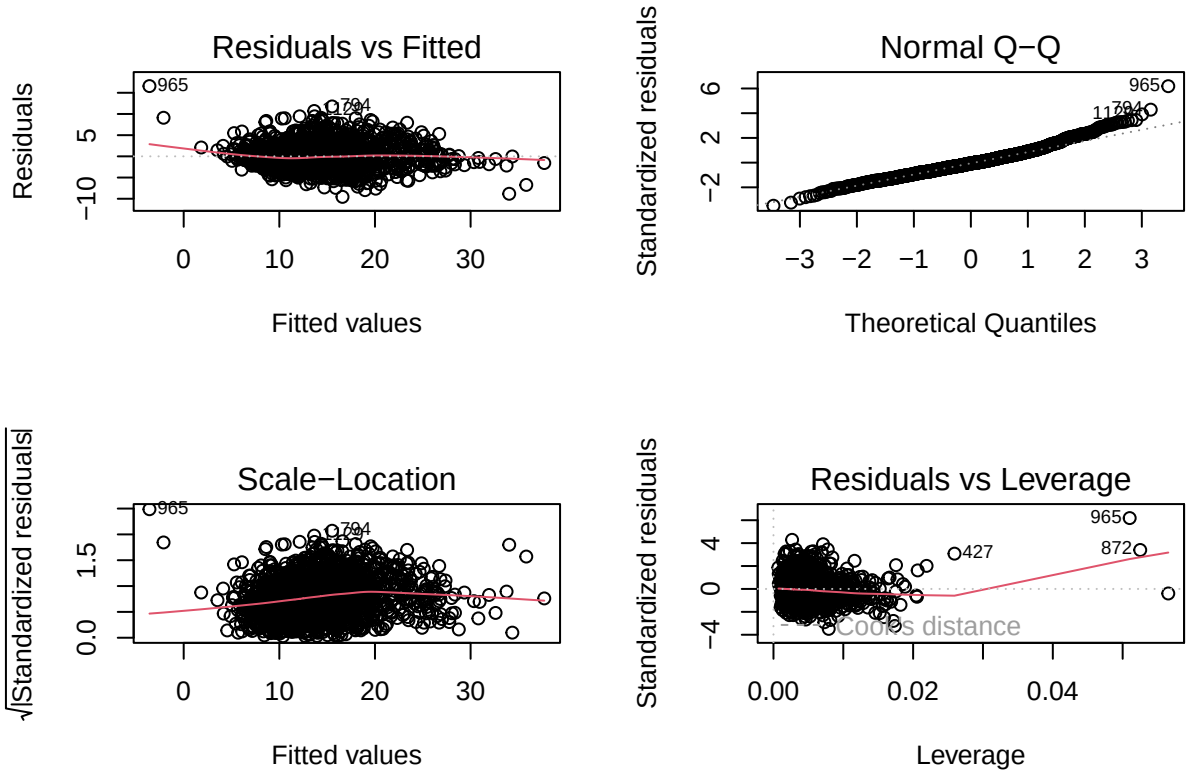

```
##             3/g        r/g       to/g        fg%        ft%        s/g
## 3/g   1.0000000 -0.4168977  0.2039304 -0.5398602  0.5129999  0.1950160
## r/g  -0.4168977  1.0000000  0.1921424  0.5145510 -0.3921333 -0.0683118
## to/g  0.2039304  0.1921424  1.0000000 -0.1215030  0.1438295  0.4687238
## fg%  -0.5398602  0.5145510 -0.1215030  1.0000000 -0.4486309 -0.2740825
## ft%   0.5129999 -0.3921333  0.1438295 -0.4486309  1.0000000  0.0578033
## s/g   0.1950160 -0.0683118  0.4687238 -0.2740825  0.0578033  1.0000000
## a/g   0.3002136 -0.1464473  0.7806362 -0.2812634  0.2276559  0.5422372
##             a/g
## 3/g   0.3002136
```

```
## r/g  -0.1464473
## to/g  0.7806362
## fg%  -0.2812634
## ft%   0.2276559
## s/g   0.5422372
## a/g   1.0000000
```

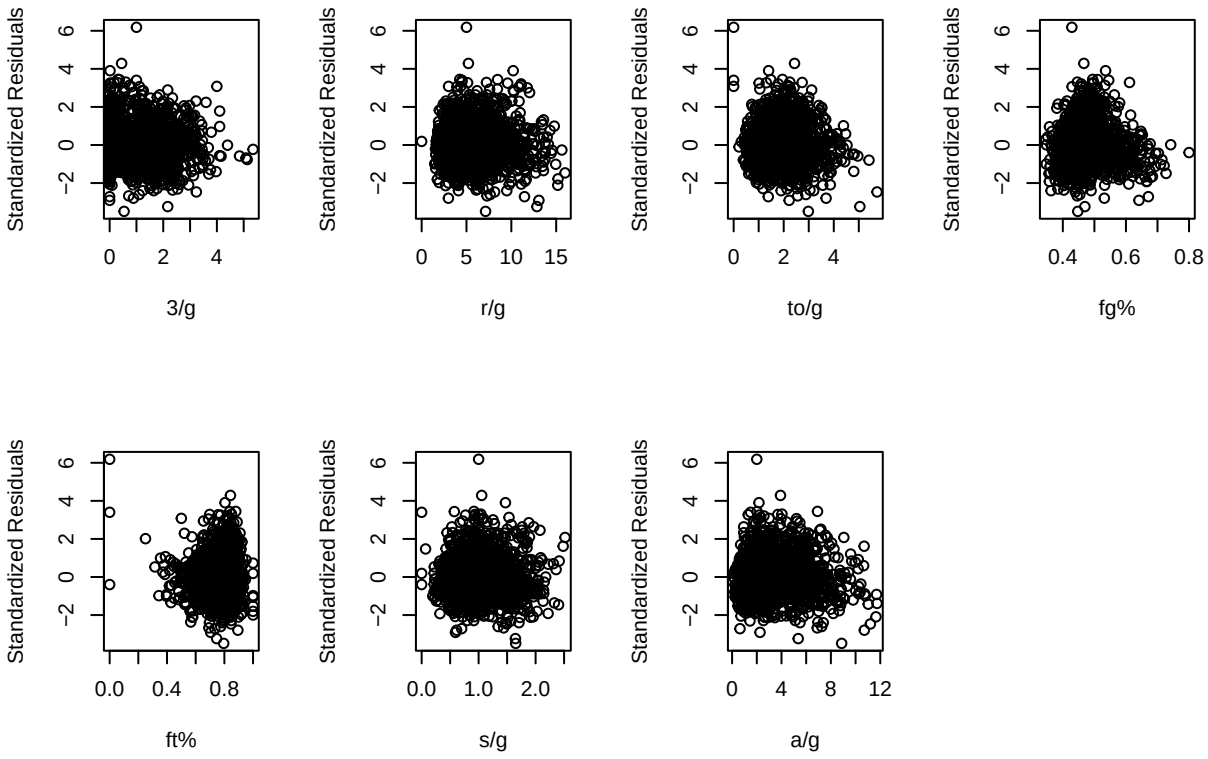
As suspected, we have there are some predictors with high correlation, such as "a/g" and "t/g", and that might result in a sort of multicollinearity. Let us look at VIF:

```
##    `3/g`    `r/g`   `to/g`    `fg%`    `ft%`    `s/g`    `a/g`
## 1.713345 1.953627 3.578168 1.797872 1.525862 1.508020 3.600205
```

This shows that the predictors "to/g" and "a/g" are relatively poorly estimated compared to others, this also proves our analysis of the potential multicollinearity of the model. Lets then take a look at the standard diagnostic plots:



Normality of std. residuals is an issue and the sq-root of std residuals do have a slight pattern, indicating this model might not be good enough. We should take a look at the std. residuals vs predictor plots:



Patterns seem not apparent for most predictors, but we have to say that ft% and 3/g present some evidence of curves. However, due to the big variation in players' stats, outliers seem to be really apparent. Overall, the model has a good performance in R-square value, and good significance on each variable, but poorly in outliers and leverage points with some high-correlation predictors that might lead to a sort of multicollinearity.

Discussion

In the first half of my findings, the trend of Player performance on the court from 2011 to 2020 in the NBA catches a lot of similarities with the second paper I mentioned in the literature review. Especially for the 3-pointers made per player, the positive curve line in the graph in early of my report indicates a substantial shift in how the game is played in the NBA. The author of my second literature review also stated a similar trend of more three-point is observable in the Euroleague as well, but on a smaller scale of increment. What He also mentioned is the decreasing number of free-throw attempts is highly associated with the increasing number of 3-pointer attempts, since I didn't include the number of shot attempts, it is hard to see the relationship between the attempts. Besides, more assists fewer steals, and fewer blocks from year to year also give us much insight into the game of Basketball in the NBA. Combined my work and the results from that paper, we can confirm the statement that the basketball game has been through a sort of "revolution" that more and more players tend to have shorter possessions, less emphasis on defense and tactical play, and, as a consequence, more turnovers, and more attractive basketball for the audience to view. That's also why NBA has become a successful league not only professionally, but also commercially.

In the second part, I created a regression model to predict the numbers of points a play can score per game by using all the data from 2011 to 2020. It is arguable that since the stats of players changed a lot over time, this model might not able to precisely predict players nowadays, but this model still has a sort of authority due to the good performance of several statistical tests. Compare to the model used in the first

article of my literature review, my model seems to be too simple to predict a player's score. The author in his article used the Naive Bayes model and maximized the likelihood expression, and that's some direction I can consider to improve my model. Despite more advanced modeling techniques, my model also has a sort of non-constant variant and potential multicollinearity that might need to improve in the future.

Overall, the research process gives me a lot of motivation for my research question. I love basketball and have been watching a lot of NBA games as well. Using the techniques I learned to analyze the stats of NBA players not only give me a different angle to view the game, but also know the process that the NBA has been through these years, and have a more insightful understanding of the NBA games. In the future, I would like to connect the changes in NBA games with its commercial development, so that I can probably dig more into the reason why players in NBA play so much differently than they did years ago.

Reference

Mandić, R., Jakovljević, S., Erčulj, F., & Štrumbelj, E. (2019, October 7). Trends in NBA and Euroleague Basketball: Analysis and comparison of statistical data from 2000 to 2017. PLOS ONE. Retrieved December 18, 2022, from https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0223524

Wheeler, K. (2012). predicting nba player performance. Santa Clara; Standford University

Lloyd, J., McKeown , K., & Smith , M. (n.d.). Player rankings 11-20. Player Rankings - Basketball Monster. Retrieved December 18, 2022, from https://basketballmonster.com/playerrankings.aspx