Report of MA678 Midterm Project

Franky Zhang

11/29/2021

Abstract

The National Basketball Association (NBA) is a professional basketball league in North America, which is absolutely the premier men's professional basketball league in the world. However, in 2020-2021 season, the highest paid player is Golden States top star, Stephen Curry, who earned a salary over $43,000,000 while the lowest paid player's salary is around $150,000. Thus, here comes the problem: which kind of players is favored most in NBA currently or which kind of players can sign a big NBA contract.

Introduction

Usually, boxed score is important description of performance of players(e.g. points, rebounds, assists and so on) and whether a player can sign a big contract is greatly decided by it. Nevertheless, different stages of NBA development seem to have different "most important" positions. For example, during a relatively long period, a truly dominant center could simply dominate the game, like Kareem Abdul-Jabbar and Wilt Chamberlain who could score almost ever time he touched the ball. Then, as Michael Jordan dominated 1990s and won 6 NBA championships, he undoubtedly led a wave of shooting guard. Recently, Warrior's Death lineup was too dominant that nearly the whole league was reorienting around to compete with Hampton Five. Thus, the combo guards with great offensive ability (e.g. Stephen Curry, James Harden, Russell Westbrook) or small forwards(e.g. Lebron James, Kevin Durant and Kawhi Leonard) are absolutely most favored in NBA and have the biggest chance to earn a fairly high salary. On the other hand, the luxury tax threshold is $136,606 million and meanwhile, some teams are willing to take risk of paying high tax luxury to form a better lineup and chase championship while some teams not. It's natural to come to the conclusion that players in big franchise teams are more likely to receive higher salaries. 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. points, rebounds, assists and so on) and random effects (teams, positions).

Methods

Data Preprocessing

I found the data set from a public github repository(https://github.com/MattC137/Open_Data/tree/master/Data/Sports/NBA). Firstly, I download 2017-2020 box score and 2021 players information because the players' 2021 salaries are greatly depended on their performances in recent years. Hence, I need to combine them and create the appropriate data frame. Additionally, as the box data is by individual games, after combining the data, I calculate average game statistics(non-Playoff games) for each player and transform `played` information to binary factor (0 or 1) then calculate average `appearances` for each player in non-Playoff games. Additionally, to avoid the bias that some players are labelled with f (forward) and g (guard) while some are labelled more specific with SF, PF, SG, PG, I combined the levels and used foward and guard identically. For the next step, I wiped out 'NA' data in players' information and merge it with aforementioned data frame. Up till now, I get full information of 314 players,which means 11 players for each team on average. Here is the glossary of terms:

column names	explanation
Player_Id_Str	Structured id name of player
Team	Team of player
Position	Position of player(SF, PF, C, SG, PG)
Position1	Position of player(forward, center, guard)
Salary	Annual salary of player (dollars)
Minutes	Average playing time of player (minutes)
Appearance	Average appearance of player (times)
Points	Average points of player
FG_Made	Average field goal made by player
FG_per	Average field goal percentage of player
Threes_Made	Average 3-point field goal made by player
Threes_pre	Average 3-point field goal percentage of player
FT_Made	Average free throws made by player
FT_pre	Average free throws percentage of player
Rebounds	Average rebounds of player
Assists	Average assists of player
Steals	Average steals of player
blocks	Average blocks of player
Turnovers	Average turnovers of player
Fouls	Average fouls of player
Height	Height of player (foot)
weight	Weight of player (lb)
Draft_Pick	Draft pick of player

Exploratory Data Analysis

By aforementioned part, I've got a `NBA_data` with 314 observations and 21 variables, among which there is 1 output `salary` and 20 predictors. However, whether or not to use all of these 20 predictors is depended on following analysis.

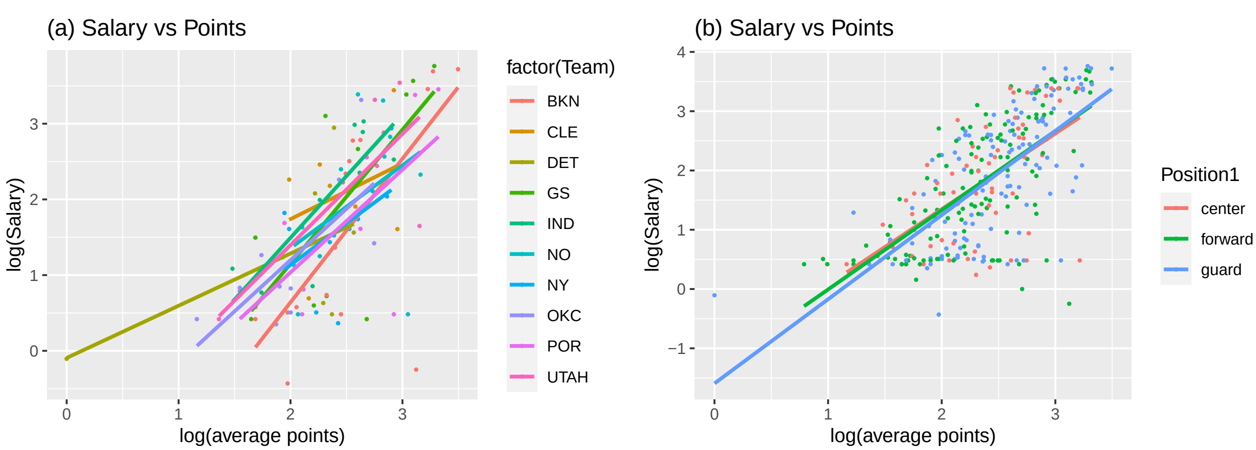

Figure 1: relationship between salaries and points of players

Figure 1 illustrates the relationship between salaries and average points, while fig(a) is in team level and fig(b) is in position level. However, whatever the level, salaries show the increasing trend as points going up. And in different teams and positions, the intercepts and slopes show slights differences. After I draw the graph of salaries versus appearance, rebounds, assists, steals and blocks, the figures are quite similar. Thus I put them in the appendix.

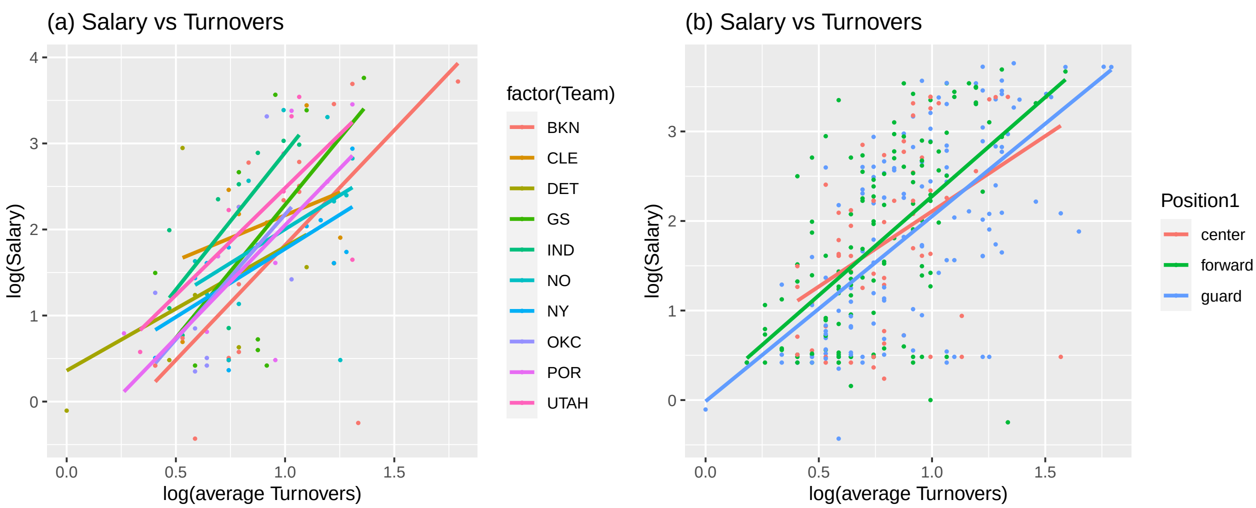

Figure 2: relationship between salaries and turnovers of players.

Figure 2 shows the correlation between players' salaries and turnovers. Similarly, figure(a) is in team level while figure(b) is in postion level. The results is weird because turnover is absolutely a negative statistic on basketball court and no one would sign huge contract with players making over 10 turnovers per game! However, as only those most high-maintenance guards have the chance to make high turnover, merely look at number of turnovers is misleading. Thus, I decided to look into relationship between turnovers and assists.

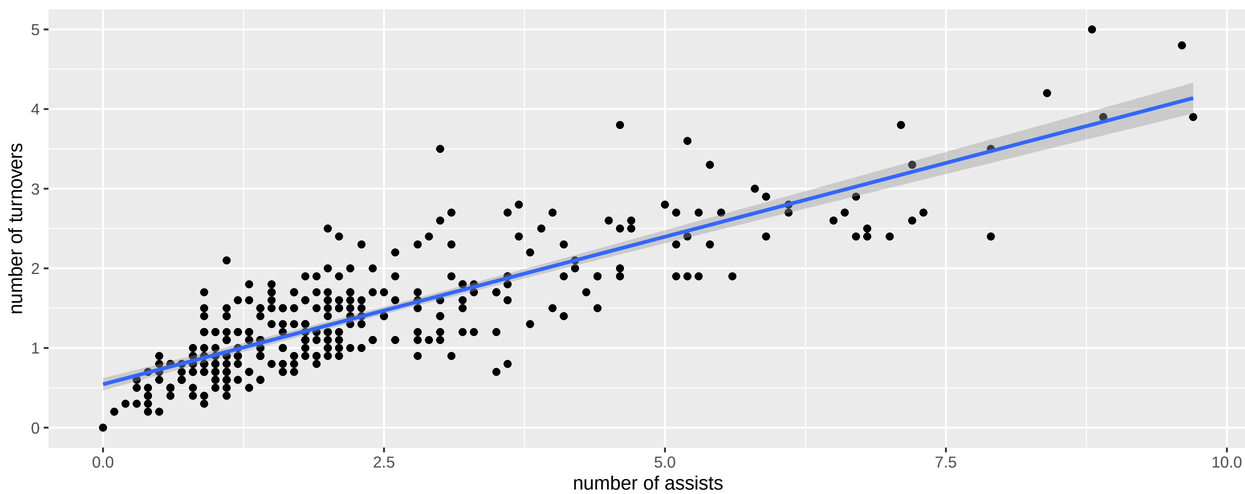

Figure 3: relationship between turnovers and assists of players

Figure 3 verifies that players' turnovers are closely related with their assists while the later is obviously positive stats. What's more, the Pearson covariance between turnovers and assists is over .85! Thus, I decide to exclude variable `Turnovers`. Variable `Fouls` is similar situation that it is highly correlated with `Minutes`, which shows Pearson covariance stats over .75. And `Heights` and `Weights` is another pair of high correlated variables. Hence, I only kept variable `Heights`.

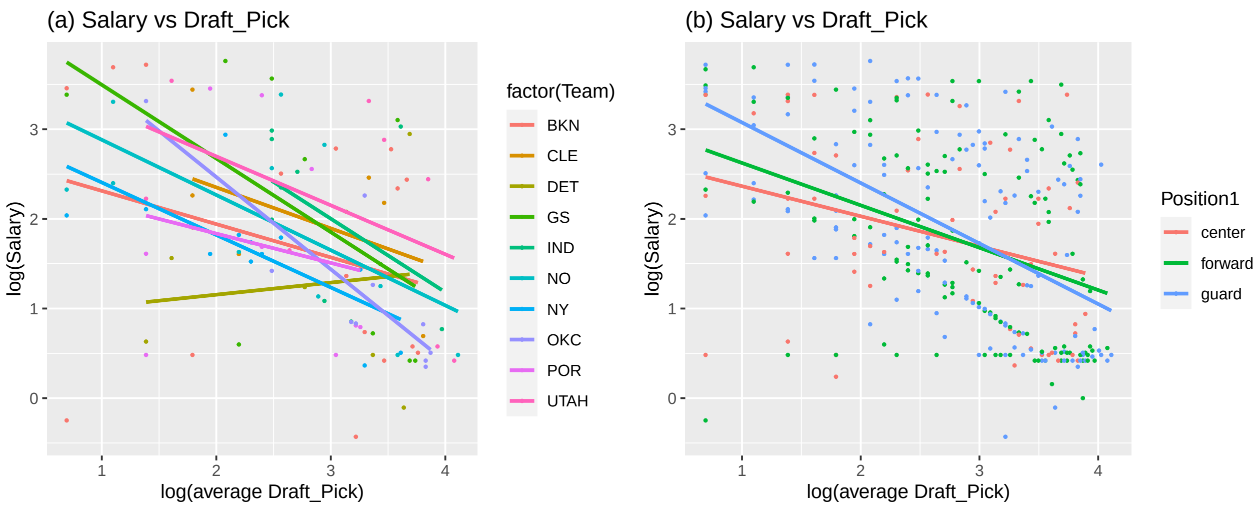

Figure 4: relationship between salaries and draft picks of players

Appendix

https://simpleblitz.com/nba/important-position-basketball/