IDS702 Final Project: NBA Salary Analysis

Junbo Guan

Summary:

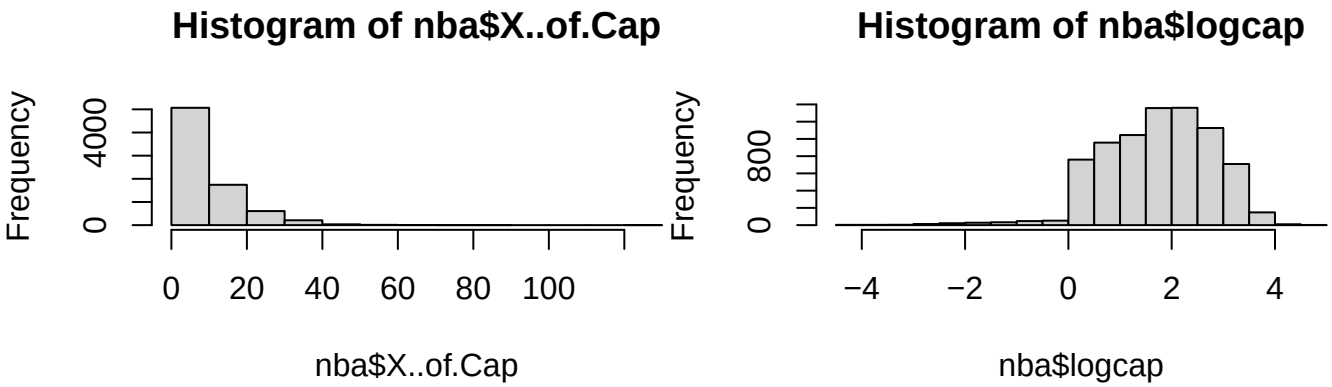
The project explores how to use various NBA statistics to predict the share of NBA players' salary cap (the total amount a team can spend on players in a given season) from 1995 to 2017. The project's goal was to find out which statistics are the best predictors of NBA players' salaries. 2. Interpret how three-point ability influence the salaries. In the project, I create two linear regression models, one for the regular statistics and one for the advanced statistics, and compare the two models. I also fit a multilevel linear regression model setting the seasons as the group variable. The conclusions are 1. the basic statistics predict NBA players' salaries better than the advanced statistics. 2. From 1995 to 2017, three-point shooters are getting a higher share of total salary caps.

Introduction

With the increasing influence of the NBA worldwide within the past decades, NBA players are getting incredibly high salaries. My project will reveal what kind of players tend to get high salaries from 1995 to 2017. The dataset used is from Kaggle, containing aggregate individual statistics for past NBA seasons. In this report, I want to quantify the effect of different statistics on NBA players' salaries and determine a likely range for the effect of them. With the sudden rise and success of Stephen Curry and his Golden Warriors, NBA players tend to shoot 3 points more than ever. I want to explore how a three-point shooter get paid during the past decades. I do exploratory data analysis to understand the relationship between different predictors and the response variable salary, fitted a multilevel linear regression model with appropriate features, and interpreted the results.

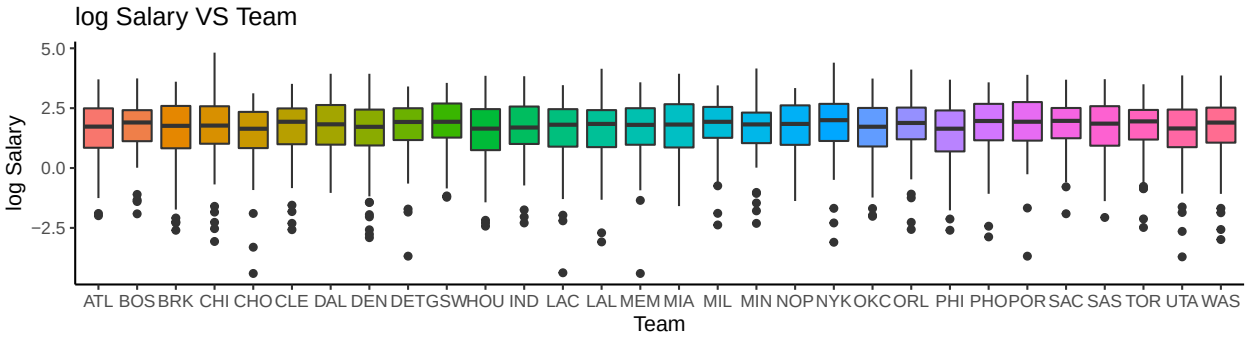
Data

Data Processing The original dataset contains aggregate individual statistics for 67 NBA seasons. Great changes have taken place in NBA during the 1980s, so I decided to use data after 1995, because in this year, two new teams were added, bringing the team total to 29 (the 30th team, the Charlotte Bobcats, was added in 2004). Since the salary cap differs each year, NBA players' salaries inflate a lot. So I include the salary caps for each year in my dataset. Instead of directly analyze salaries, I explore the percentage of the league's salary cap to normalize the data. Some of the teams moved or changed their name during the time period. For example, Seatle SuperSonics moved to Oklahoma City and changed its name to thunders in the 2008-2009 season. So I merged such kind of data. For players who were traded during the season, I only take the team's statistics he played more games in so that such players will be counted only once. I removed the players who only played less than 15 games for that season. I also transfer all the regular data on a per-game basis instead of totals to eliminate the effect of games played. Since the data set only has less than 2% observations with missing data, I discard such observations. I create a new factor variable 3 point shooter for those players who have three-point shooting percentages, and three-point shooting attempts greater than the league's median. I classifier him as a three-point shooter, and all the other players are not. I then divided the statistics into two subcategories: Regular and Advanced. Regular statistics include statistics that casual fans can understand: points per game, rebounds per game, percentage of field goals scored, and so on. Advanced statistics were created to assess player performance in more detail, PER (player efficiency level), victory share (share of the number of wins a player contributes to his team), usage rate (percentage of team games the player uses in the field)...



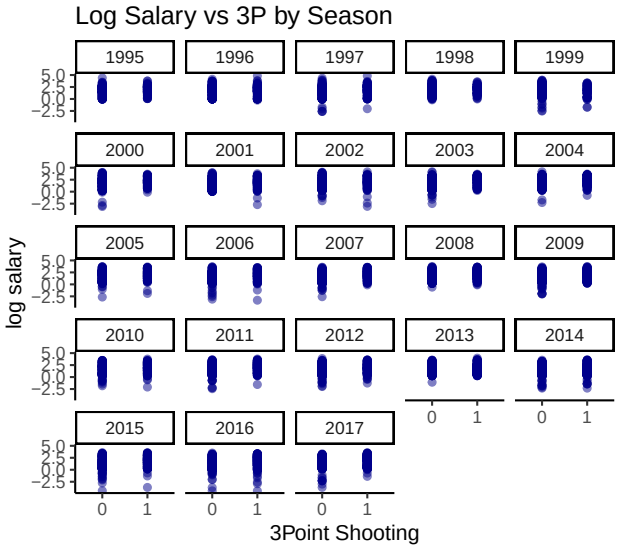
EDA Before starting the analysis, the predictors "three-point shooter", "position," and "team" are considered as factor variables. I plot the histogram for the response variable; the salary cap percentage does not follow the normal distribution. The log percentage of cap, however, follows the normal distribution. Thus, I use the log percentage of the cap(logcap) for the coming analysis.

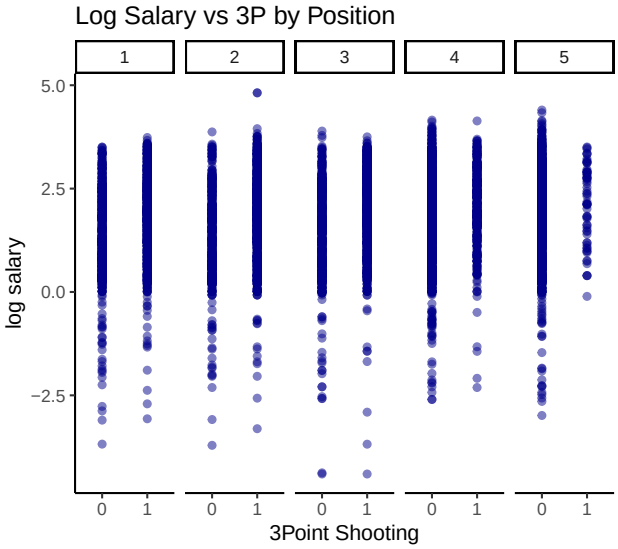
Relationship between predictors and the response variable Next, explore the relationship between salary and each predictor. I use boxplots for categorical predictors. I use scatter plots for continuous/numeric predictors. We can find out the log cap is almost the same in each season. For the predictor, I add a quadratic term. Nba players' salaries increase before the age of 29 and decrease after age 30. I will use log(WS) for win share since the log(WS) has a linear relationship with the log cap. Generally speaking, a three-point shooter gets paid more than a non-three three point shooters. The salary difference between teams is not that much, but several teams would offer higher salaries to players than other teams: Detroit Pistons, Cleveland Cavaliers, Golden State Warriors, and New York Knicks. Generally, Eastern conference teams are more generous. It's interesting because western conference teams won 14 championships from 1995 to 2017 while eastern conference teams only won 9 championships. During the period, the most successful dynasty teams: the Los Angeles Lakers and San Antonio Spurs do not pay players higher than other teams.



Interactions between predictors For the interaction terms between a factor variable and a continuous variable, I draw a set of boxplots to explore potential interaction. The interaction between 3 point shooter and position is interesting. The salary only goes high slightly for the three-point shooter for guards and forwards compared to the non-three-point shooter. However, for the centers, three-point shooters' salary is significantly higher than the non-three point shooters. It is important, especially for big men, to shoot and create space. As for the interaction between three-point shooting and season, the plot barely shows any difference before 2006. After 2006, it is clear that three-point shooters are getting paid relatively higher each year. The salaries of players in different positions vary as the NBA evolves. In the late 1990s and early 2000s, the power forwards and the center gets relatively higher salaries. That is the golden years for the big man. Then in the mid-2000s, The NBA league decided to modify the rules to encourage offense for the sake of ratings. The guards get paid higher than other positions. From the 2010s, forwards get paid more than other positions.

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```





Model

Model Selection, Assessment snd Comparision (Regular Statistic and Advanced Statistics) Before fitting the model, I center all numerical predictors to help interpret the final model. For regular statistics and advanced statistics, I use BIC stepwise model selection to choose predictors. For both BIC selection, I only put position and age(including quadratic term) in the null model and put all other data in the full model. The model for regular statistics is :

$$logcap = \beta_0 + \alpha Pos + \beta Agec + \gamma Agec2 + \delta PPGc + \epsilon RPGc + \zeta APGc$$
$$+ \eta BPGc + \theta Seasonc + \iota GPc + \lambda FGAPG + \mu X2PPGc + \varepsilon$$

The model for advanced statistics is:

$$logcap = \beta_0 + \alpha Pos + \beta Agec + \gamma Agec2 + \delta WSc + \eta USGc + \theta BPMc + \lambda WS48c$$
$$+ \mu VORPc + \nu Seasonc + \iota STLc + \sigma PERc + \tau TSc + \omega TOVc + \sigma TRBc + \varepsilon$$

Then I checked the four assumptions: linearity, independence of errors, equal variance, and normality. For both residual plots, the residuals seem to spread randomly along the x-axis with no pattern. For both residuals vs. fitted plots, residuals seem to spread randomly along the x-axis with no clear pattern. For both Q-Q plots, some of the observations lie away from the 45-degree line. My models slightly validate the normality assumption. I also check the multicollinearity for both models. Most of the VIF values are below 5 and only several between 10 to 15, which is acceptable. Then I implement k-fold cross-validation to calculate the RMSE for these two models. The regular statistics have slightly lower RMSE values, which indicates the regular statistics fit the data better than the advanced statistics.

Final Model Selection and Assessment: Then I put all the statistics: regular and advanced in the full model, and use the same BIC stepwise method to fit the final model. I also set season as a factor variable and use it as the group variable to quantify how three-point shooters get paid in different seasons. The final model:

$$logcap_{ij} = \beta_{0j} + \beta_{ij} X_{ij} + \varepsilon_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}(X3PLevel|Seasonf)_j + \upsilon_{0j}$$
$$\beta_{1j} = \gamma_{10} + \upsilon_{1j}$$

The assessment of the assumption is the same process as mentioned above. The final model slightly violates the normality assumption. Then I use the same cross-validation to calculate the RMSE and find out the RMSE value is almost the same as the model with regular statistics. It means pure regular statistics capture the features pretty well. It makes sense to me since most advanced statistics come from the regular data we have in the model.

```
Scaled residuals:
    Min      1Q  Median      3Q     Max
-7.1275 -0.5293  0.0945  0.6402  3.6374

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Seasonf  (Intercept) 0.008356 0.09141
          X3PLevel1   0.002620 0.05118  -0.40
 Residual             0.536586 0.73252
Number of obs: 7682, groups:  Seasonf, 23

Fixed effects:
             Estimate Std. Error t value
(Intercept)  1.696958  0.039246  43.239
Pos2         0.079097  0.036162   2.187
Pos3         0.104771  0.041690   2.513
Pos4         0.219823  0.049758   4.418
Pos5         0.348634  0.056346   6.187
Agec         0.056901  0.002309  24.643
Agec2       -0.005649  0.000399 -14.150
PPGc         0.041069  0.006075   6.760
RPGc         0.124537  0.010826  11.503
APGc         0.223350  0.017136  13.034
BPGc         0.116119  0.027168   4.274
Seasonc     -0.004919  0.002986  -1.648
ASTc        -0.041506  0.003788 -10.957
TRBc        -0.062301  0.006556  -9.503
GPc          0.002310  0.000494   4.672
TOVc         0.030724  0.003210   9.570
USGc         0.029177  0.004424   6.596
WS48c        3.287660  0.461482   7.124
TSc         -2.768797  0.311165  -8.898
VORPc       -0.094636  0.012642  -7.486
BPMc         0.046727  0.009574   4.881
STLc        -0.058014  0.017910  -3.239
```

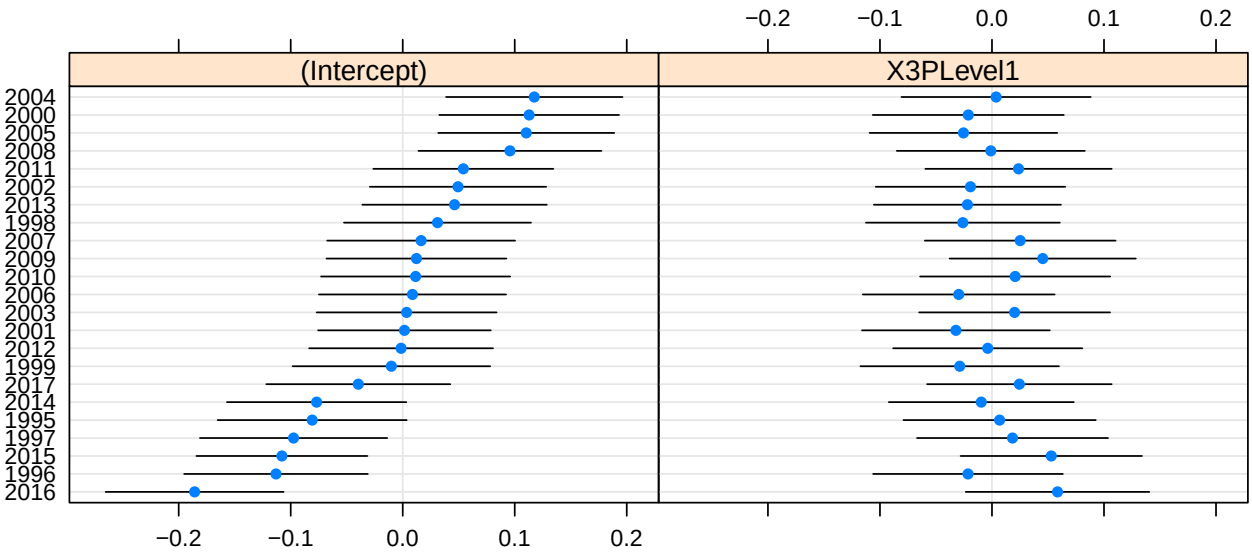
	2.5 %	97.5 %
(Intercept)	1.6244275	1.7623071
Pos2	0.0062438	0.1485824
Pos3	0.0207928	0.1847191
Pos4	0.1231230	0.3186482
Pos5	0.2450791	0.4662856
Agec	0.0518247	0.0609108
Agec2	-0.0063560	-0.0047835
PPGc	0.0286458	0.0524947
RPGc	0.1062673	0.1485779
APGc	0.1905676	0.2579720
BPGc	0.0551534	0.1619874
Seasonc	-0.0087372	-0.0034645
ASTc	-0.0499260	-0.0350933
TRBc	-0.0782511	-0.0528319
GPc	0.0014242	0.0033350
TOVc	0.0254160	0.0378819
USGc	0.0207880	0.0381289
WS48c	2.6189224	4.3988582
TSc	-3.6569643	-2.4776340
VORPc	-0.1197839	-0.0702331
BPMc	0.0293831	0.0667853
STLc	-0.1007446	-0.0306154

Final model Interpretation

Fixed Effects The intercept is 1.69, which means for a 27-year-old average point guard in season 2007-2008, who played 58 games, scored 9.4 points, 4 rebounds, 2 assist per game will earn e^1.69 = 5.4 percentage of the total salary cap. Holding all variables constant, a point guard gets the same statistics; he will get paid 8.2% higher than a point guard; a small forward will get paid 11.1% higher than a point guard; a power forward will get 24.6% higher, and a center will get 41.8% higher salary than a point a guard. As position goes from 1 to 5, players' salaries get higher. Holding all variables constant, for every one point a player score per game, his salary will increase by 4.1%; for every one more assist a player gets per game, he will get paid 24.5% higher; for every one more rebound a player gets per game, he will get paid 12.7% higher; for every one more block a player gets per game, he will get paid 12.3% higher. For every game a player plays, the 95% confidence interval is between 0.14% and 0.33%. For every one-unit increase in usage percentage, the player's salary will increase at best 3.9%, at worst 2.1%. For every one-unit increase in win share per 48 minutes, the player's salary will increase at best 1373.6%, at worst 8136.9%. Win share per 48 minutes is a powerful predictor for several reasons: first, the scale of WS48 is relatively small. The mean of WS48 is only 0.09, and the maximum is only 0.35. Thus a one-unit increase will influence the salary significantly. For every one-unit increase in turnover percentage, the player's salary will increase at best 3.9%, at worst 2.6%. It makes sense because superstars who handle the ball on the court tend to have a high turnover rate.

Random Effects: For any player in season 1996-1997, the baseline salary is actually lower than the overall season's average; the effect of three-point shooting is also lower than the overall season's average. For any player in season 2016-2017, the baseline salary is actually higher than the overall season's average; the effect of three point shooting is also higher than the overall season's average. The estimated standard error is 0.73, which describes the unexplained within-season variation. The estimated standard deviation of intercept is 0.09, which describes the within-season variation attributed to the random intercept. The estimated standard deviation of three-point shooting is 0.05, which describes the within-season variation attributed to the random slope(the predictor). The estimated random correlation is -0.40, which is low.

Then I draw the dot-plot of random effects to find the potential outliers. As shown in the plot, season 2000-2001, 2004-2005, 1996-1997, 2016-2017 are the potential outliers since they are far away from other seasons. We can also find out for seasons after 2005; the dots tend to lie on the right side of 0, while the dots representing seasons before 2005 tend to lie on the left of 0, indicating three-point shooting is valued more after 2005.



Conclusion

As obvious as it seems, players who score more points and make more field goals will be paid a higher salary, which is why the regular statistics produce a better model than the advanced model. Even the regular statistics and advanced statistics fit the data set well; the regular data is slightly better at predicting NBA players' salaries. In this analysis, the association between three-point shooting and salary was explored. The final model confirmed that the salary of a three-point shooter varies by season. Before 2005, three-point shooters did not get paid more than the non-three-point shooters; such circumstances changed after 2005. As the position goes from 1 to 5, the position does matter; players' salaries get higher significantly.

There are several potential limitations to this analysis. First, as mentioned in the data processing part, around 2% of the rows had null values. I do not have any information about these values to make further speculation or to conduct further processing. Hence, these observations were dropped before modeling. The missing observations could cause bias in my analysis. Moreover, my analysis is based on over 20 years period. In recent years, with the introduction of data science to sports, managers have based their salary offers to players much more on advanced statistics than in years past. Lastly, the NBA salary rules changed significantly during my analysis period, and my analysis did not reflect that.