Player Performance, Social Power and Team Valuation in 2016-2017 NBA Season Project

Renxiong Liu, Lingfei Zhao

STAT 6950

1 Introduction

National Basketball Association(NBA) is no doubt one of the most successful sport league in the world. The successful commercialisation of NBA makes it possible for teams to offer high salaries to players, which encourages players to play harder for their future contracts. Besides, elite players who perform impressively usually have large numbers of followers on twitter, or are searched frequently on websites, which means they might have large social powers. In view of these, valuations of sports teams, whose main assets are players and especially the elite players, might be influenced by players' performance and their social powers. So for team managers, how can they determine suitable salary for players and what strategies can they take to improve team valuation?

Intuitively speaking, the salaries of players are mainly determined by their performance. But considering there are so many variables to evaluate players' performance, managers might wonder which groups of them are most important. In addition, lots of players have their own social media accounts, which is a good strategy to expose themself to public. We might have good reasons to cast doubts on whether the power of social media will have some impacts on their salary. All of these will be investigated and explained in the main body, which could be used to answer the question: what determines players' salaries in NBA league?

Another issue managers have to deal with is the team valuation. The valuation of a sport team is a bit hard to determine in corporation finance. Generally speaking, the income from tickets, the values of brands, the operation expenses are three main factors that make up for team valuations. These factors are associated with team performance, marketing strategy, salary structure, etc, many of which might be covered in the determimants of salaries as well. Given these information, the managers might still feel confused at what on earth determines the team valuation. Statistical analysis is provided to eliminate confusions and the managers can know what to do to increase the team valuation.

Our project aims to answer both questions raised above by using appropriate statistical models. Lending from these models, we hope to provide business insights for team managers to improve their teams in multi-dimensions.

2 Dataset

With the help of background information above, we find relevant variables from multiple sources. A main part of our dataset comes from Kaggle (https://www.kaggle.com/noahgift/social-power-nba). In this part, data about the team valuation, average attendence of fans per game, performace statistics, salaries and social power of elite players in a team are provided. Besides, we also find

regional GDP data from Bureau of Economic Analysis (https://www.bea.gov/) and 2016-17 NBA ticket prices from VIVIDSEATs (https://www.vividseats.com/blog/nba-ticket-prices). All of these make up for our dataset for analysis.

Some interesting issues occur in our exploratory data analysis. Firstly, players in different positions are good at different skills. For example, players in position center tend to have better performances in rebound statistics, which makes sense since they are usually higher then players in other position on average. More importantly, this informs us the importance of distinguish players in different position in the salary model since players have different strengths. Besides, we also depict the scatter plot matrix for relevant statistcis about player performances and social power, team performance and social power, etc. A good suggestion in these scatter plot matrix is we should do the logarithm transformation when investigate the team valuation. We also need to delete some outliers/influential points to regularize the dataset. All these plots are present in Appendix 6.1.

3 Player Salary: Method and Results

Since we are exploring the relationship between players' salary and their statistics, we will use players' salary as response variables, and their statistics as explantory variables, to build a multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_{p-1} X_{p-1,i} + \varepsilon_i, \quad i = 1, 2, \cdots, n. \tag{1}$$

where ε_i $iid \sim Normal(0, \sigma^2)$. Here, the Y_i represents the salary of the ith player, and $X_{i,j}$ represents the jth statistic of the ith player. Because we do not have any prior knowledge in basketball, we decided to consider all variables into our model at first in order to see whether these explantory variables are useful to explain our response varialbe. The R^2 of the full model is 0.7 (we omit the summary.lm here because it is too long), which is a relatively high number. That means our explantory variables can be helpful to explain about 70% variability of the response data around its mean. Although this full regression model is useful in predicting the player's salary, our goal is to provide an simple and interpretable model to help team manager make a decision. Therefore, to simplfy our model, we will do the variable selection at first and choose explantory variables as few as possible.

3.1 Potential Predictors

The dataset we collected contains many different type of potential predictors (explanatory variables). Some of these are continuous measurements, like the "Field Goals" or "Free Throws" of a player. Some of them are discrete but ordered, like a player's age. Others can be categorical, like the position of a player. All these types of potential predictors can be useful in our multiple linear regression (1). The potential predictors we consider are:

- **The intercept**: Suppose we define $\mathbf{1}$ be a predictors that is always equal to 1, then our mean function of (1) can be written as

$$E(Y_i|\mathbf{X}) = \beta_0 \mathbf{1} + \beta_1 X_{1,i} + \cdots + \beta_{p-1} X_{p-1,i}, \quad i = 1, 2, \cdots, n.$$

- **Continuous measurements and age**: All continuous measurement are included in our regression model without any transformation. We also include the "AGE" without any transformation because we believe it is approximately continuous.
- **Dummy variables and factors**: The "POSITION" with 4 levels are considered as *factor*, and they are included in 1 using *dummy vaiables*.
- **Interactions**: Based on our experience, we think that the "POSITION" of a player will interact with other potential predictors when they influence the salary of this player. For example, "POINTS" is more important to the point guard (PG), while "Defensive Rebounds" is more important to the center (C). Therefore, we include all interaction terms between "POSITION" and other predictors as our potential predictors.

To simplify our model and make it interpretable, we do not consider the transformations or the polynomials of our predictors. We would like to consider them as our future jobs.

3.2 Variable Selection

In our last subsection, we have identified the potential predictors in our model, but the question for us is how to select the useful variables and why we need to do the selection. Firstly, we need to interpret our model to a team manager, so we have to make the model simple. Too many predictors will definitely increase the complexity of our model. Also, the collinearity between variables can make our model unreasonable. For example, in our full model, which contains all potential predictors, the coefficient of "POINTS" is negative, which means the player will get fewer salary if he get more points in the game. That is obviously wrong, so we must do the variable selection to let our model seem reasonable. What's more, we have fewer than 250 data in our dataset, but the potential predictors are more than 70. Therefore, variable selection can enhance generalization by reducing overfitting, so it can make more precise prediction.

The approach to finding useful predictors we use here is considering all potential predictors, and then select one subset that optimizaes the criterion (we consider AIC and BIC in this section). We have many different methods to deal with it. For example, the "All Subsets" method and Stepwise Search Methods are both very useful. However, because we have more than 70 potential explantory variables in our model, the running time of "All Subsets" method will be extremely long, so it is more sensible for us to use the Stepwise Search Methods here.

At first, we employ the "Backward Elimination", "Forward Selection" and "Stepwise Regression" to minimize the AIC criterion. We run these three method, and our results are as follows:

1. **Backward Elimination** It chooses almost all potential predictors, and it gets AIC=641.89.
2. **Forward Selection** It chooses "POINTS + AGE + DRB + PF + AST + eFG. + W + FG. + TOV + ORPM + FGA + PAGEVIEWS + TWITTER FAVORITE COUNT + TWITTER RETWEET COUNT + X2P" as predictor, and it gets AIC=671.14.
3. **Stepwise Regression** It chooses "POINTS + AGE + DRB + PF + AST + eFG. + W + TOV + ORPM + FGA + PAGEVIEWS + TWITTER FAVORITE COUNT + TWITTER RETWEET COUNT" as predictor and it gets AIC=670.52

Similarly, we employ the BIC criterion and get the following result:

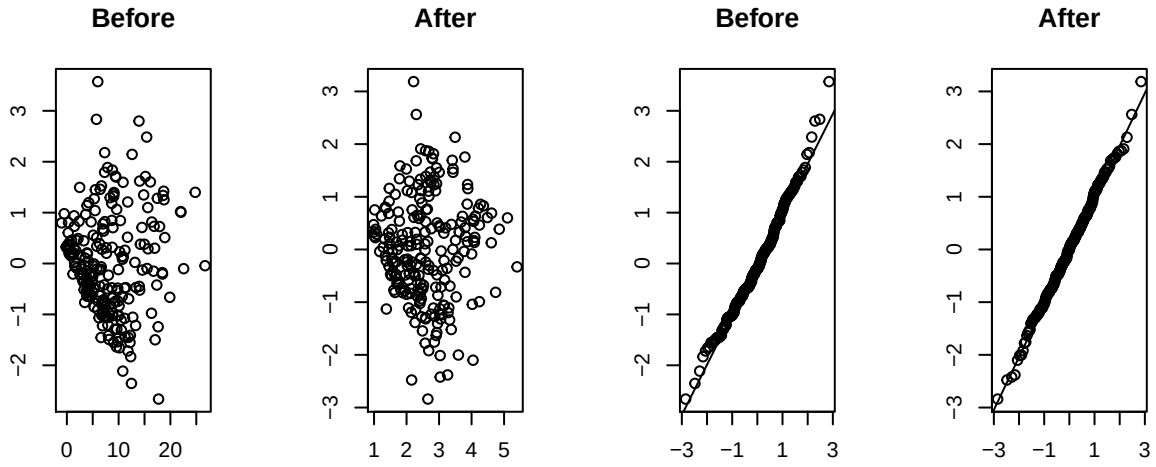


Figure 1: (1) The first two scatterplots are standardized residuals against the fitted value before and after the transformation. (2) The third and forth plots are QQ-plots of standardized residuals before and after the transformation.

1. **Backward Elimination** It chooses "AGE + MP + FT + MPG + PAGEVIEWS + TWITTER FAVORITE COUNT" as predictors, and it gets BIC=708.97.
2. **Forward Selection** It chooses "POINTS + AGE + DRB + PF" as predictor, and it gets BIC=711.48.
3. **Stepwise Regression** It gets the same result as Forward Selection.

All of these three methods choose too many predictors when we use the AIC criterion. However, when we use BIC criterion, they tend to select fewer predictors. We think it is because the BIC criterion is largely influenced by the sample size, so since our dataset is large, it tends to choose fewer predictors comparing with the AIC. Because we want to simplify our model and make it interpretable, the predictors we got in BIC are preferable, and since the "Forward Selection" and "Stepwise Regression" get the same result, we will adopt the predictors selected from them and do the analysis.

3.3 Diagnostics

In our multivariate regression model (1), we have assumed $\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\beta, \mathrm{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2\mathbf{I}$, and our error terms ε_i are $iid \sim Normal(0, \sigma^2)$. In order to check the correctness of our model's assumptions, we analyze the residual $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}}$. The residuals should satisfy $\mathrm{E}(\hat{\mathbf{e}}) = \mathbf{0}$ and $\mathrm{Var}(\hat{\mathbf{e}}) = \sigma^2(\mathbf{I} - \mathbf{H})$, where $\mathbf{H} = \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$. We draw the scatterplot of standardized residual $\mathbf{r} = \frac{\hat{\mathbf{e}}}{\hat{\sigma}\sqrt{\mathbf{I}-\mathbf{H}}}$ against the fitted value. Therefore, if our assumptions are correct, the scatterplot 1 should not have obvious patterns.

It is obvious that our standardized plot (Figure 1) has right-opening megaphone shape, which means our constant variance assumption has not been satisfied. We can see that $\mathrm{Var}(\mathbf{Y}|\mathbf{X}) \propto \mathrm{E}(\mathbf{Y}|\mathbf{X})$, so we do the square root transformation for our response variable (players' salary) at first, and then repeat our variable selection by "Stepwise Regression" and BIC criterion. This time we choose `POINTS + AGE + DRB` as our response variables and the BIC=-93.8. We draw a new scatterplot of standardized residual, and it does not show any pattern this time, which means our constant

variance has been satsified. Also, in the QQ-plot (Figure 1), we can see that most points fit the straight line perfectly after the transformation. We also perform the Shapiro-Wilk test of normality. The corresponding p-value is close to 1, which is a perfect support of the normality assumption of the residuals. Besides, the maximum of Cook's distance in this model is 0.1, which means no obvious influential point in our data.

After examining the residuals, checking for influential observations there is no compelling evidence that any of the usual MLR assumptions are violated.

3.4 Cross Validation

In this subsection, we will adopt the K-fold Cross-Validation method to assess if our fitted model will be useful for prediction. The idea is to divide our available data into k equal sized groups. Each time a single group is retained as the validation data for testing, and the remaining k-1 groups are used to fit the model. The process will repeat k times, and k groups used exactly once as the validation data. The so-call K-fold CV estimator is

$$\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2,$$

where $\hat{Y}_i$ is the predicted value of Y_i when the model was fit without the group that case i belongs to. We run the K-fold CV 1 time, and the K-fold CV estimator is 0.6, which is close to the MSE of our regression model. The result is shown in the Appendix 6.5.

3.5 Conclusion

After the cross validation, we firmly believe that our model can be useful in predicting the salary of a basketball player. This multivariate regression model only contains three predictors, `POINTS + AGE + DRB` and the intercept, which is easy to interprete and understand. Based on our result (Appendix 6.2), the final model is:

$$\sqrt{\text{Salary}} = -1.3582 + 0.0986POINTS + 0.0975AGE + 0.1012DRB \tag{2}$$

The interpretation for the coefficient of `DRB` is, for example, when Defensive Rebounds increases one unit and other predictors are fixed, the square root of salary will approximately increase 0.1012 units. From this model, we can see that the coefficients of these predictors are all positive, which means they have positive contribution to the salary. We only choose these three predictors does not means other predictors are not important, instead it means there may exist strong correlations between these three predictor and other potential predictors. However, since the $R^2 \approx 0.56$, which means these three predictors can explain 56% of variability in salary, which are sufficient enough for us to do the prediction and provide a useful suggestion to the team manager.

4 Team Valuation: Method and Results

4.1 Potential Predictors

Intuitively, reasonable predictor variables include regional GDP, tickets prices, and performace statistics, salaries, social power of elite players. Notice not all performance statitics are reasonable for our analysis. With the help of information in 1 part, we can know only the basic performance statistics seem reasonable since it can stimulate the purchase of tickets directly. So we only include `PTS`, `rebounds`, `assists` and some other 5 performance statitiscs variables to reduce the number of variables we will consider. After that, we use the sum of personal performance statistics within a team to represent the corresponding team performance statistics, for example, use the sum of `PTS` within a team to represent `team_PTS`. Besides, we add up the salaries and social power variables of players within a team to represent the corresponding team variables. Lastly, other accessory informaton like `TICKET_PRICE` is also included. Considering the different scale of the variables, we take the logarithm for some of them. We plot a summary scatter plot matrix of typical variables below to describe the relationship between these variables in Appendix 6.6.

In the scatter plot matrix, we have already delete the strange points in our dataset, which has been explained in our EDA. There seems to be no strange behavior after the deletion, which means we can carry on to the main part of our analysis.

4.2 Variable Selection

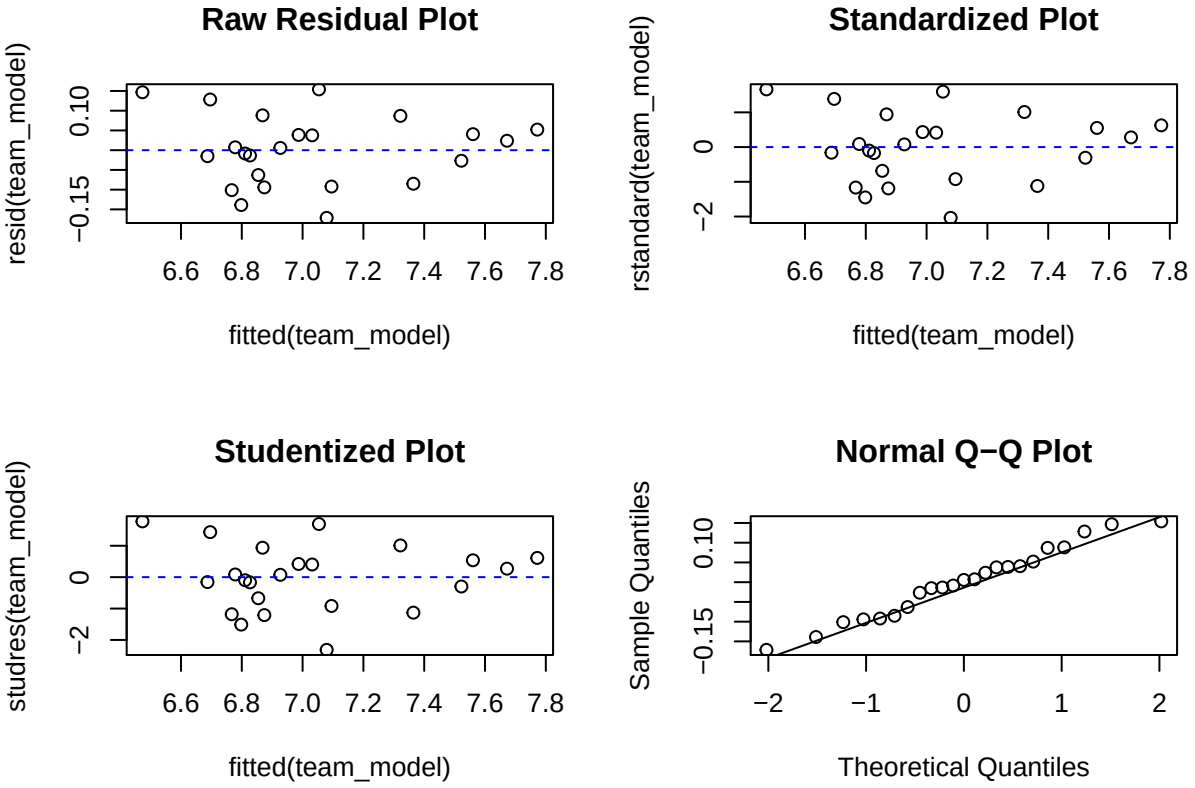
Given the reasonable variables we analyze in part 4.1, we employ backward selection method with BIC criterion to choose the "best" linear regression model for analysis. Before implementing the stepwise search methods, let's first look at the plot of all subsets 6.7 to have a basic evaluation of important variables we should pay more attention. A quick conclusion from this plot is `lg_GDP` and `lg_TICKET_PRICE` are important for BIC selection criterion. Besides, the inclusion of a social power variables, `lg_TWITTER_RETWEET_COUNT` is somewhat surprising since it might inform us the importance of social power on team valuation. As for the basic performance statistics, notice only variable `POINTS` is excluded, this might be due to the fact that the highest points in each team will not be so different compared with the other 4 basic performance statistics.

Now let's use the backward selection method with BIC criterion to implement model selection, the result is present in 6.4. The final result is the same as what the plot above indicates. So we use the chosen model as the standard model for further analysis, which selects `AST`, `TRB`, `STL`, `BLK`, `lg_TICKET_PRICE`, `lg_TWITTER_RETWEET_COUNT` and `lg_GDP` as predictors.

4.3 Diagnostics and Analysis

Considering the small sample we have, We cannot do the lack of fit test in our problem setting. However, from the diagnostic graph, we can see there are no severe violations to the regression model, which means the model we fit is good. So it's safe to use our model selected above. The fitted result is given below.

Now let's look at the model we use to explore our dataset. One interesting question is whether team's social power has impacts on team's valuation. The answer of the question is important since if it is true, a direct strategy for managers to increase the team valuation is: sign free players who used to be the all-star. In our model, the only variable associated with social power is `lg_TWITTER_RETWEET_COUNT`, which is siginifcant according to t test. The result tells us the social power does have impacts on the team valuation. Also notice the sign of `lg_TWITTER_RETWEET_COUNT` is positive, so the managers can sign the palyers with high `TWITTER_RETWEET_COUNT` to increase the team valuation. Another interesting result is that variable `AST` is not significant in the summary table. A possibile interpretation for this is: The small ball trend in NBA speeds up the games, so a team good at blocks, steals will stand out in that trend. This argument is also supported by the sign of variable `TRB`, which is negative since the team with more rebounds tends to play more slowly.




```
##
## Call:
## lm(formula = lg_TEAM_EVALUATION ~ AST + TRB + STL + BLK + lg_TWITTER_RETWEET_COUNT +
##     lg_GDP + lg_TICKET_PRICE, data = teammat)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.171450 -0.074029  0.005667  0.046353  0.154039
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.18765    0.49924   2.379  0.03108 *
## AST                       -0.02764    0.01622  -1.704  0.10898
## TRB                       -0.04159    0.01315  -3.163  0.00643 **
## STL                        0.22708    0.10530   2.157  0.04767 *
```

```
## BLK                        0.12958    0.06060   2.138  0.04934 *
## lg_TWITTER_RETWEET_COUNT   0.23663    0.04443   5.326 8.47e-05 ***
## lg_GDP                     0.29129    0.03562   8.178 6.56e-07 ***
## lg_TICKET_PRICE            0.24447    0.09782   2.499  0.02454 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1076 on 15 degrees of freedom
## Multiple R-squared:  0.9373, Adjusted R-squared:  0.9081
## F-statistic: 32.06 on 7 and 15 DF,  p-value: 6.405e-08
```

4.4 Conclusion

Our fitted model suggests a positive association between the counts of Twitter retweet and Team valuation. A direct conclusion for this is: if the manager wants to increase the team valuation, one strategy he/she could take is to sign players in free player market with relatively large social power. For example, some players who used to be all-stars but are now in the free player market could be the potential targets. Reasonable as our result sounds, the value of the coefficient of social power might be a little confusing, since it indicates 1% increase in team's counts of Twitter retweet could increase the team valuation up to 23.6%. So if we can find more available data and arrange them into a panel form, we could get a more accurate relationship between the social power and team valuation.

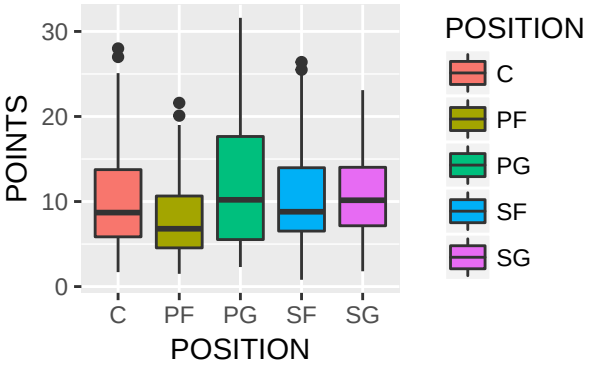
5 Discussion

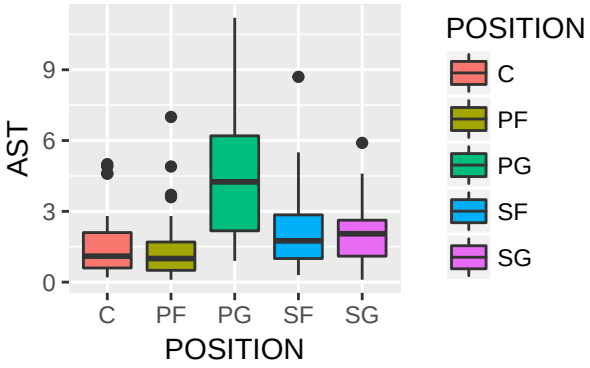
In this project, we investigate the important factors to determine the players' salary and team valuation, and we provide our specific answer and interpretation in each section. Our results show that the models fit the data pretty well. However, there are still some limitations in our project. Firstly, we only consider the linear relationship between our response variables and potential predictors, but we do not consider the transformations or the polynomials of our potential predictors. So if there exist any nonlinear relationships between them, our model will miss these information. Second, in the model selection part, we choose stepwise regression and BIC criteria to get our final predictors since it can simplify our model. However, there is no evidence which can supports our conclusion. To deal with these problems, we can consider more complex model which includes the transformation of predictors as potential predictors. Also, asking for some expertise or getting some prior knowledge can help us choose the correct predictors.

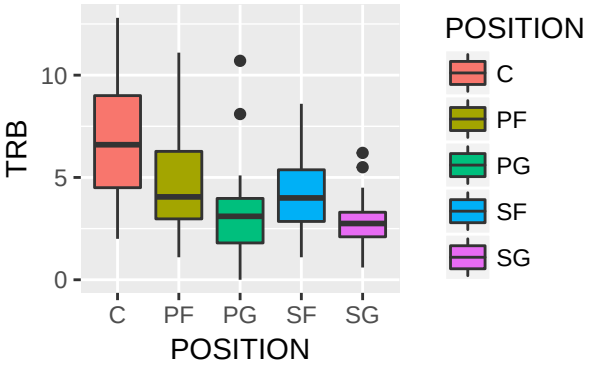
What's more, we also notice that 100 players are considered as elite players based on their performance in the 2016-2017 season. An interesting direction of future work is to predict the whether a player would be a elite player in the following season. Because the simple variable selection methods are not useful here, we choose the same predictors as in our multivariate regression model. The preliminary result is shown in Appendix 6.3, and we would like to investigate it deeper in the future.

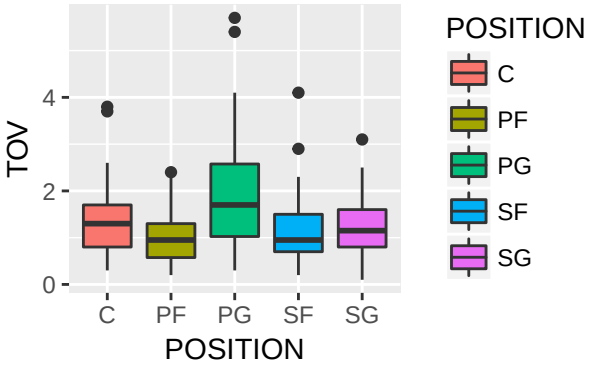
6 Appendix

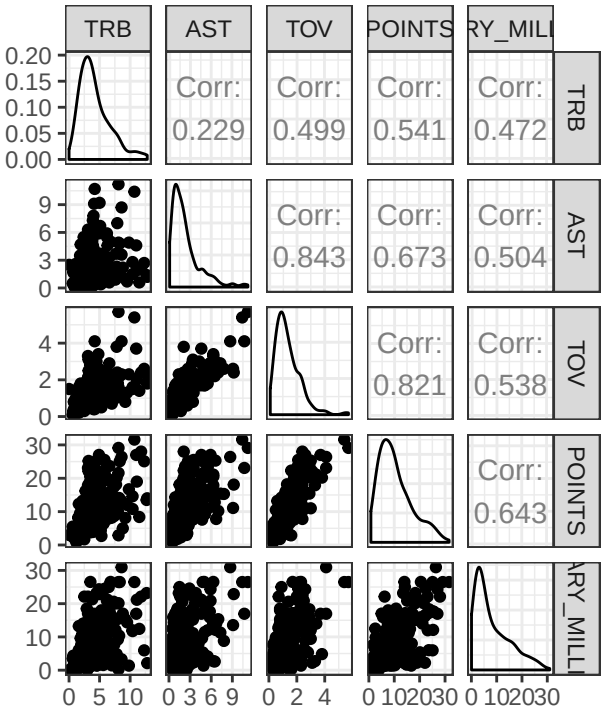
6.1 Exploratory data analysis

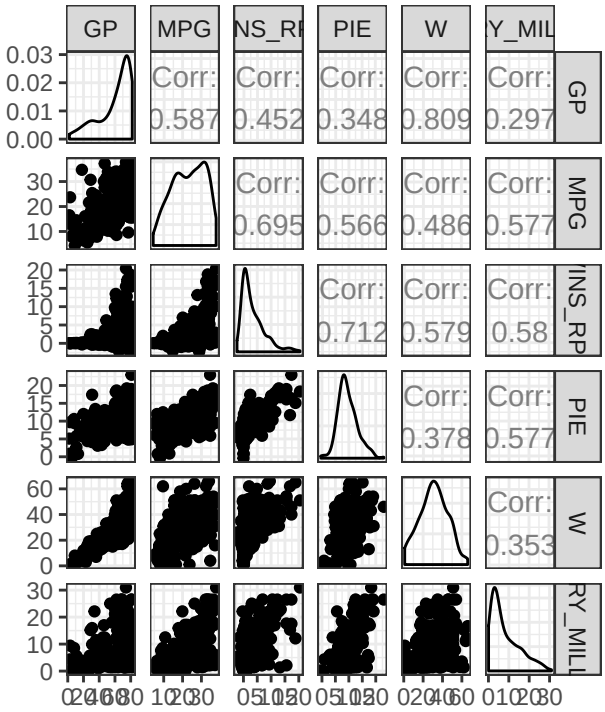












6.2 Salary model

```
##
## Call:
## lm(formula = sqrt(SALARY_MILLIONS) ~ POINTS + AGE + DRB, data = NBA_stat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16618 -0.55411  0.00095  0.49505  2.48650
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.35822    0.32635  -4.162 4.50e-05 ***
## POINTS       0.09857    0.01026   9.607  < 2e-16 ***
## AGE          0.09753    0.01178   8.279 1.12e-14 ***
## DRB          0.10122    0.03578   2.829  0.00509 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7832 on 224 degrees of freedom
## Multiple R-squared:  0.5669, Adjusted R-squared:  0.5611
## F-statistic: 97.72 on 3 and 224 DF,  p-value: < 2.2e-16
```

6.3 Logistic model

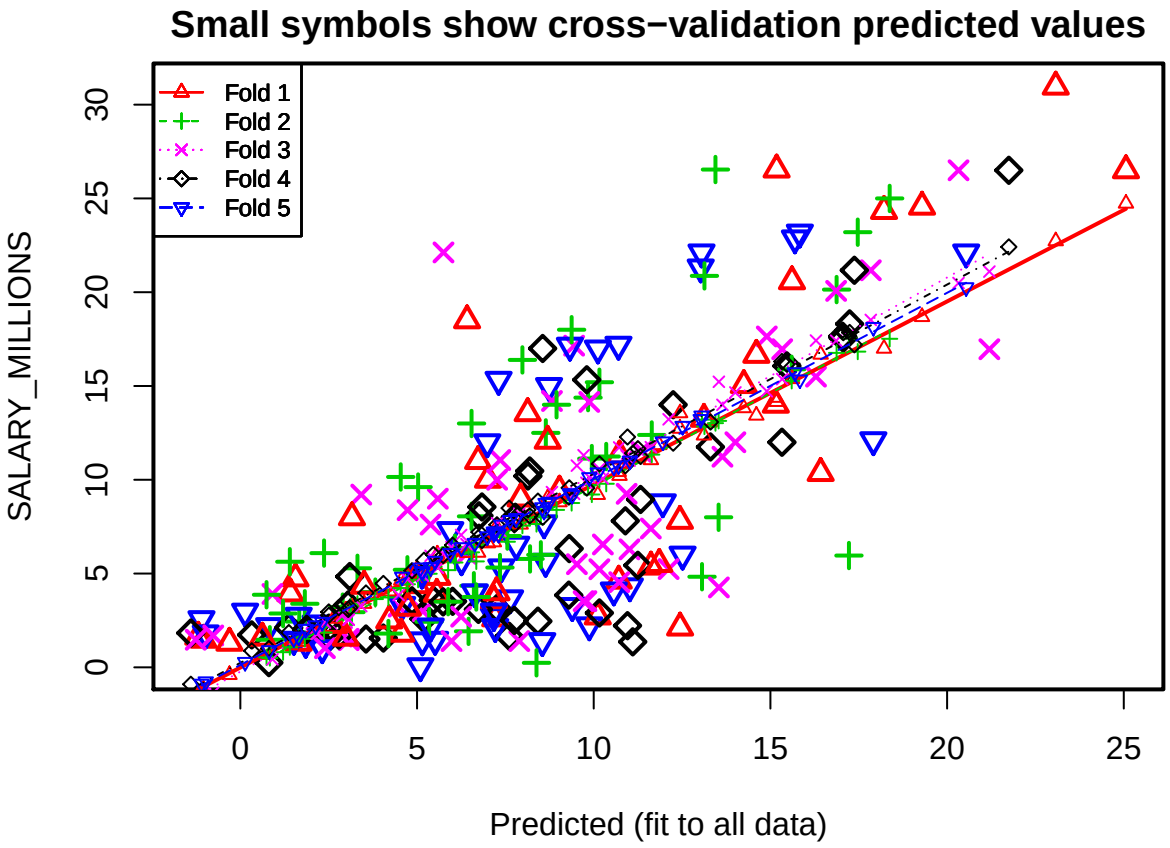
```
##
## Call:
## glm(formula = eli ~ POINTS + AGE + DRB, family = binomial(link = "logit"),
##     data = NBA_eli)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0365  -0.5404  -0.2321   0.2954   2.7786
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.38966    1.71713  -5.468 4.55e-08 ***
## POINTS       0.23350    0.04254   5.488 4.05e-08 ***
## AGE          0.12336    0.05024   2.455   0.0141 *
## DRB          0.64795    0.15119   4.286 1.82e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 272.30  on 224  degrees of freedom
```

```
## Residual deviance: 148.69  on 221  degrees of freedom
## AIC: 156.69
##
## Number of Fisher Scoring iterations: 6
```

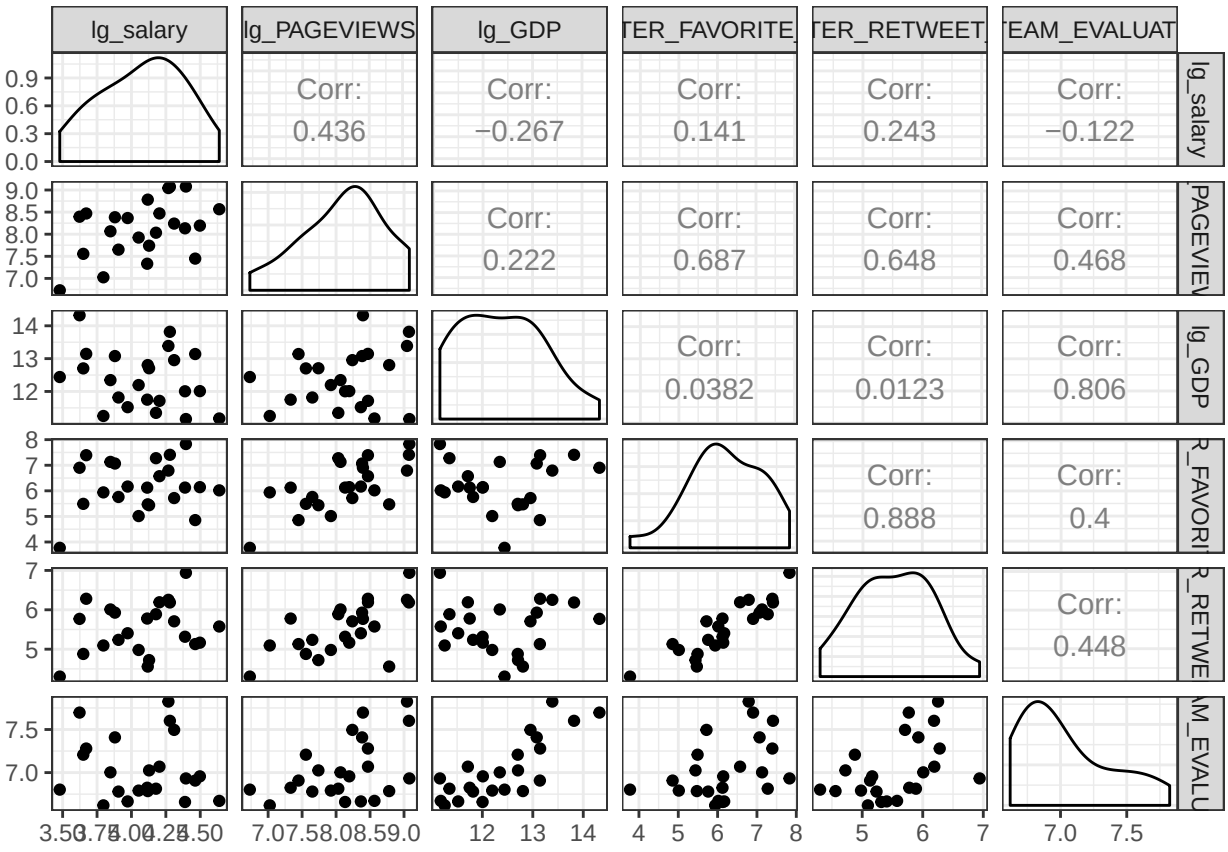
6.4 Team Valuation: variable selection

```
##
## Call:
## lm(formula = lg_TEAM_EVALUATION ~ AST + TRB + STL + BLK + lg_GDP +
##     lg_TWITTER_RETWEET_COUNT + lg_TICKET_PRICE, data = subset(teammat,
##     select = c(-TEAM, -GDP, -GDP_POS, -TEAM_EVALUATION, -PAGEVIEWS,
##         -TWITTER_FAVORITE_COUNT, -TWITTER_RETWEET_COUNT, -SALARY_MILLIONS,
##         -TICKET_PRICE)))
##
## Coefficients:
##              (Intercept)                       AST
##                  1.18765                  -0.02764
##                      TRB                       STL
##                 -0.04159                   0.22708
##                      BLK                    lg_GDP
##                  0.12958                   0.29129
## lg_TWITTER_RETWEET_COUNT           lg_TICKET_PRICE
##                  0.23663                   0.24447
```

6.5 K-fold CV plot



6.6 Scatter plot matrix of typical variables



6.7 Variable selection plot with BIC criterion