Determining the Factors that Affect the Players' Salary In National Basketball Association

Student Name: Seung-Hyun Kim

Date: Dec.6th.2017

Instructor Name: Jiguo Cao

Course: STAT 350

Table of Content:

Abstract

In this report, we investigated the factors that affect the players' salary in National Basketball Association. In order to provide the dataset for the report, 523 entries of NBA players' statistics in the 2015 season were collected with total 21 variables: one categorical variable of Players' name and 20 numerical variables including the response variable Salary and 19 explanatory variables that we investigated. We first studied the scatterplots matrix to construct the full model and studied the multiple linear regression analysis on the full model. Then, we transformed the response variable and used model selection methods to get two best fitting models for AIC and BIC. We then used ANOVA and model summaries to determine the final model. The final model showed that the salary was affected by five explanatory variables: age, minutes played, height, assists, and points.

Keywords:

Salary, age, minutes played per game, height, assists made per game, points scored per game, linear regression.

Introduction

In the National Basketball Association (NBA), every team has an equal limit in total salary expenditure called salary cap, therefore the team that outputs the highest efficiency out of given salary cap has the highest chance of winning the championship. However, it is not easy to acquire superstars who outputs highest efficiency out of their salary, because of the competitions to acquire such player among 30 NBA teams and also because of the fact that there are only limited number of superstars. Therefore, during the free agent market season, many teams have a strategy not to give over-paying contract to their players. In order to execute such strategy in a professional way, a need of accurate salary prediction model is necessary. Therefore, we decided to use multiple linear regression model to determine what factors affect the NBA players' salary and how significant are those factors.

Dataset:

The first dataset was collected on 490 unique NBA players in 2015 season with 34 variables: four variables was categorical, which contained players' name, position, college and team, and the rest of 30 variables were numerical, which contained players' height, weight and basketball statistics. Moreover, the second dataset was the collected on all NBA players from 1990 to 2016 season with 6 variables including their salary. The first and second dataset were inner joined on players' name in 2015 season using SQL procedure in SAS, and then the salary

variable and 19 other variables of the most simple and common NBA statistics were selected to be an output-table to be analyzed in R. The sources of the first and second data set are shown in the reference 4 and 5 respectively.

Response variable:

1. SALARY: The Player's Salary in Million US dollars (continuous)

Explanatory variables:

1. FGM: Total Field Goals Made (integer)
2. FGA: Total Field Goals Attempted (integer)
3. FG_: Field Goal Percentage, which is a probability and ranges from 0 to 1 (continuous)
4. FTM: Free Throws Made (integer)
5. FTA: Free Throws Attempted (integer)
6. FT_: Free Throws Percentage, which is a probability and ranges from 0 to 1 (continuous)
7. AST: Total Assists Made (integer)
8. STL: Total Steals Made (integer)
9. BLK: Total Blocks Made (integer)
10. TOV: Total Turnovers Made (integer)
11. PF: Total Personal Fouls Made (integer)
12. AGE: Age of the Player (integer)
13. Height: Height of the Player in centi-meters (continuous)
14. GP: Total Games Played (integer)
15. MPG: Minutes Played Per Game (continuous)
16. POINTS: Total Points Scored (integer)
17. TRB: Total Rebounds Made (integer)
18. ORB: Total Offensive Rebounds Made (integer)
19. DRB: Total Defensive Rebounds Made (integer)

Analysis to do:

1. To construct scatterplots matrix to determine linear relationships among the explanatory variables and also between the response and explanatory variables.
2. To construct the full model and determine whether assumptions of multiple linear regression are satisfied.
3. To construct the transformed model if necessary.
4. To perform model selection method to find two best models for AIC and BIC.
5. To test the significance of the AIC and BIC model using ANOVA
6. To construct the final model and test the significance of the explanatory variables.
7. To interpret the effects of the explanatory variables on the response variable.

Methods

In this report, we used multiple linear regression models to analyze the NBA players' salary. Multiple linear regression models generally have response variable as a linear function of explanatory variables plus an error term ε. However, when using multiple linear regression models to analyze the data, the five assumptions must be checked before proceeding to data analysis. The five assumptions of multiple linear regressions are:

1. Assume that the response variable is a linear function of the explanatory variables.
2. Assume that the error term, ε has zero mean.
3. Assume that the error term, ε has constant variance.
4. Assume that the error terms are uncorrelated to each other.
5. Assume that the errors follow normally distribution.

However, since we did not have any time variable or spatial variable in our dataset, it was impossible to check whether the error terms were uncorrelated to each other, therefore we assumed that the assumption 4 holds for our data and we only checked assumption 1, 2, 3 and 5 in this report.

Firstly, we used boxplot on the variable SALARY to check whether there are any outliers. In the boxplot, outliers would be detected and we would remove these outliers, as well as the observations with missing data.

We then set the initial model, which contained all 19 numerical explanatory variables:

Initial Model: $SALARY = \beta0 + \beta1 * FGM + \beta2 * FGA + \beta3 * FG\_ + \beta4 * FTM + \beta5 * FTA + B6 * FT\_ + \beta7 * AST + \beta8 * STL + \beta9 * BLK + \beta10 * TOV + \beta11 * PF + \beta12 * AGE + \beta13 * Height + \beta14 * GP + \beta15 * MPG + \beta16 * POINTS + \beta17 * TRB + \beta18 * ORB + \beta19 * DRB$

After, we constructed scatterplots matrix to check the linear relationships between the explanatory variables. One of each pairs of the explanatory variables with exceptionally strong linear relationships were removed from the initial model. Then interaction terms of our best interest were added to construct the full model:

Full Model: $SALARY = \beta0 + \beta3 * FG\_ + \beta6 * FT\_ + \beta7 * AST + \beta8 * STL + \beta9 * BLK + \beta10 * TOV + \beta11 * PF + \beta12 * AGE + \beta13 * Height + \beta14 * GP + \beta15 * MPG + \beta16 * POINTS + \beta18 * ORB + \beta19 * DRB + \beta20 * TOV:AST + \beta21 * TOV:POINTS + \beta22 * DRB:ORB + \beta23 * FG\_:POINTS$

The diagnostic plots were constructed to check whether the five assumptions of multiple linear regression were satisfied in the full model. Then transformation of variables would be used if it is necessary to satisfy the five assumptions, and diagnostic plots would be constructed to check whether the transformed model satisfy the five assumptions.

We then used forward, backward and stepwise model selection using Akaike's Information Criterion (AIC) and Bayesian's Information Criterion (BIC) method to construct the two most significant models. In the model selection, we used two base models: one was the full model shown above, and the other was the simple model, which only contains intercept:

Simple Model: $SALARY = \beta0$

We then obtained total six models using two information criterion AIC and BIC, in forward, backward and stepwise model selection. In each criterion, the model with smallest criterion would be selected as the best AIC model and the best BIC model (Weisberg, 2005):

Best AIC Model: $sqrt(SALARY) = \beta0 + \beta1' * AST + \beta2' * AGE + \beta3' * Height + B4' * MPG + \beta5' * POINTS + \beta6' * STL$

Best BIC Model: $sqrt(SALARY) = \beta0 + \beta1' * AST + \beta2' * AGE + \beta3' * Height + \beta4' * MPG + \beta5' * POINTS$

Then, in order to decide which model is suitable as a final model, we used ANOVA test using the best AIC model and the best BIC model. In the end, the best AIC model was chosen as the final model. We then constructed diagnostic plots of the final model to make sure whether the five assumptions are satisfied.

Results

Outliers:

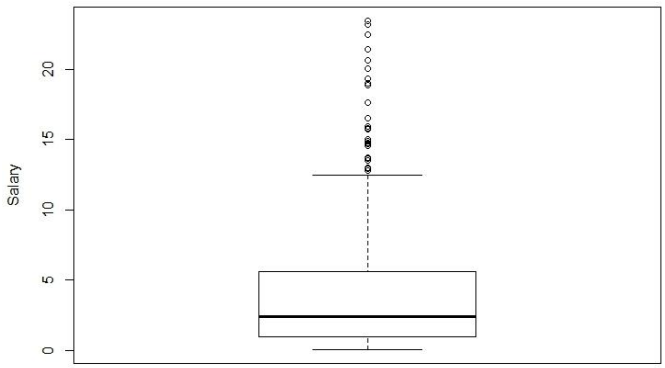

Figure 1: Boxplot of the Response variable SALARY

The boxplot in Figure 1 clearly showed that salaries larger than 13 are considered as outliers. In order to build an accurate model, all the outliers and the observations containing missing data were removed.

Initial Model:

After the data cleaning, we constructed the initial model, which contained only 19 explanatory variables without any interaction terms:

Initial Model:	**SALARY = β0 + β1 * FGM + β2 * FGA + β3 * FG_ + β4 * FTM + β5 * FTA + B6 * FT_ + β7 * AST + β8 * STL + β9 * BLK + β10 * TOV + β11 * PF + β12 * AGE + β13 * Height + β14 * GP + β15 * MPG +β16 * POINTS + β17 * TRB + β18 * ORB + β19 * DRB**

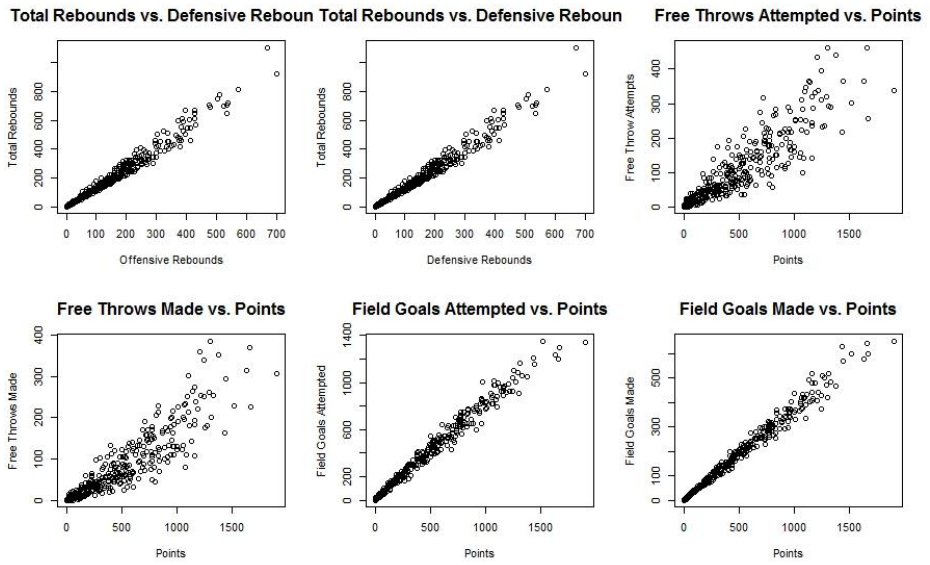

Figure 2: Scatterplots of Predictors with strongest linear relationships

Then we used scatterplots matrix to check which explanatory variables have the linear relationships. Among these scatterplots, six plots in Figure 2 showed the strongest linear relationships. If we kept the explanatory variables having exceptionally strong relationship with each other in the model and proceeded to model selection, then there would be a possibility of model selection process being messed up. Therefore, one of each pairs of explanatory variables in each scatterplot was removed from the Initial model: FGA, FGM, FTM, FTA and TRB were removed, therefore there were only 14 explanatory variables left in the second initial model:

Second Initial Model: SALARY = β0 + β3 * FG_ + β6 * FT_ + β7 * AST + β8 * STL + β9 * BLK + β10 * TOV + β11 * PF + β12 * AGE + β13 * Height + β14 * GP + β15 * MPG + β16 * POINTS + β18 * ORB + β19 * DRB

Full Model:

In order to check whether the response variable has linear relationships with the explanatory variables, which is the assumption 1), we again used scatterplots matrix of the dataset.

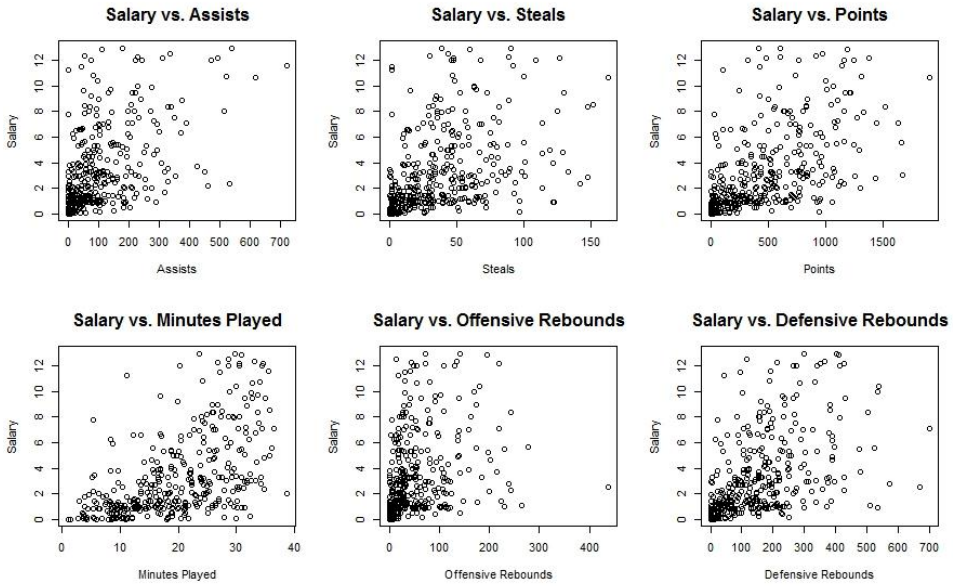

Figure 3: Scatterplots of the Response variable vs. Explanatory variables (1/2)

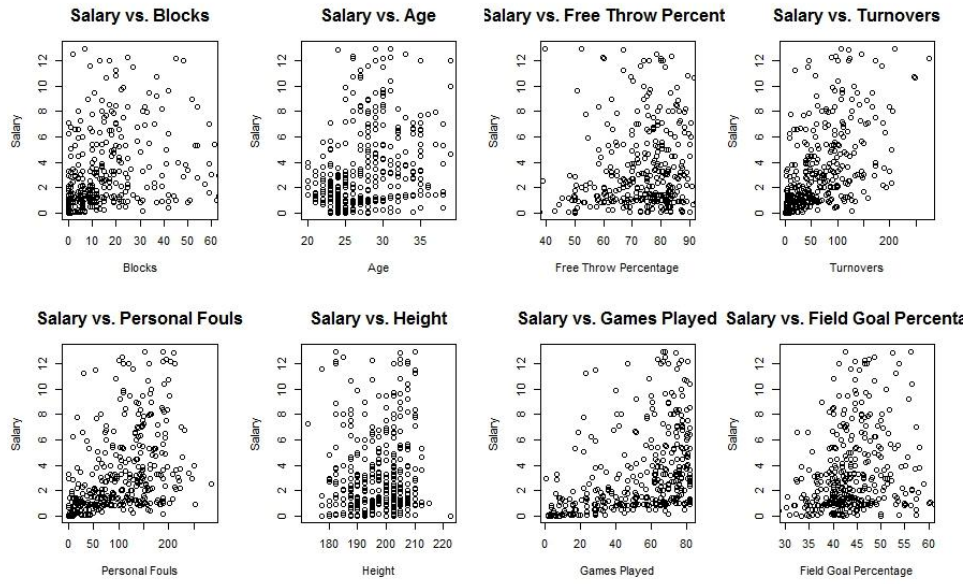

Figure 4: Scatterplots of the Response variable vs. Explanatory variables (2/2)

Figure 3 and 4 clearly showed that SALARY is in linear relationships with all 14 explanatory variables. We were assured that SALARY was a linear function of 14 explanatory variables, therefore the assumption 1 was satisfied. Then, four interaction terms, which fitted best interest of ours, were added to the initial model: TOV:AST, TOV:POINTS, DRB:ORB, and FG_:POINTS. Therefore, our full model contained 18 variables including four interaction terms:

Full Model: **SALARY = β0 + β3 * FG_ + β6 * FT_ + β7 * AST + β8 * STL + β9 * BLK + β10 * TOV + β11 * PF + β12 * AGE + β13 * Height + β14 * GP + β15 * MPG + β16 * POINTS + β18 * ORB + β19 * DRB + β20 * TOV:AST + β21 * TOV:POINTS + β22 * DRB:ORB + β23 * FG_:POINTS**

Analysis on the Full Model:

After constructing the full model, we constructed the diagnostic plots to analyze the model.

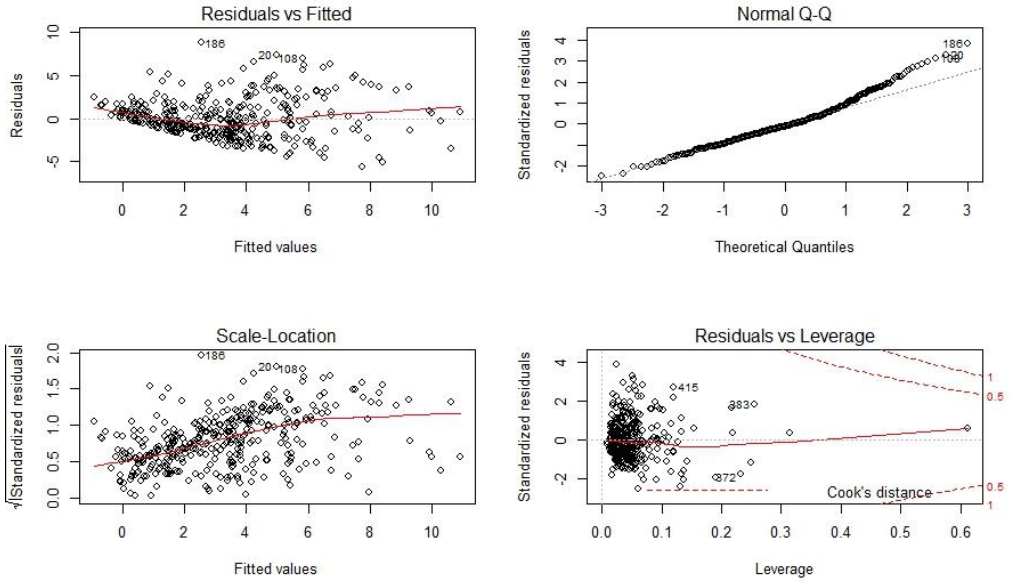

Figure 5: The Diagnostic Plots of the Full Model

In Figure 5, the Residual vs. Fitted plot showed a funnel shape with narrow left end and wide right end, which means that the residual variance is not constant (Pagano and Gauvreau, 2000), therefore the assumption 3 was not satisfied. Also, the Normal Q-Q plot showed the large departure from the straight line at the top right tail, which mean the residuals did not follow normal distribution (Weisberg, 2005), and the assumption 5 was not satisfied. Moreover, the Scale-Location plot showed non-linear trends, which also showed that residual variance is not constant. In Residuals vs. Leverage plot, any observations outside the Cook's distance are considered as influential cases (Kim, B., 2015). Since our residual vs. leverage plot did not show any influential cases, therefore we did not worry about any possible outliers influencing the full model.

Transformation:

Since the full model did not satisfy the assumption 3 and 5, it was necessary to make some changes in order to improve the full model, therefore we transformed the model by performing square root transformation on the response variable SALARY.

Transformed Model: sqrt(SALARY) = β0 + β3 * FG_ + β6 * FT_ + β7 * AST + β8 * STL + β9 * BLK + β10 * TOV + β11 * PF + β12 * AGE + β13 * Height + β14 * GP + β15 * MPG + β16 * POINTS + β18 * ORB + β19 * DRB + β20 * TOV:AST + β21 * TOV:POINTS + β22 * DRB:ORB + β23 * FG_:POINTS

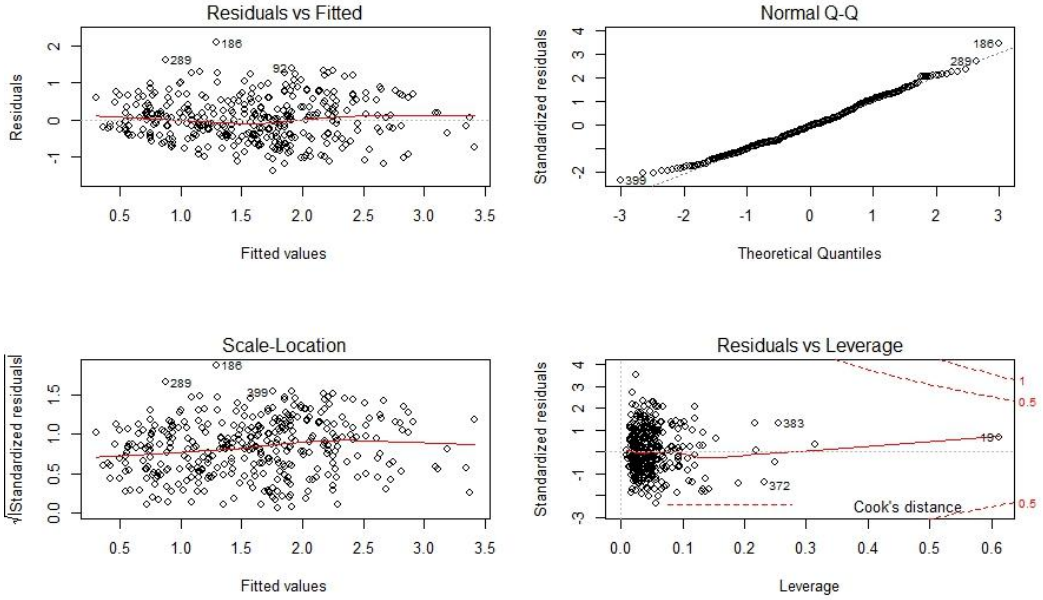

Figure 6: The Diagnostic Plots of The Transformed Model

In Figure 6, the Residuals vs. Fitted plot showed that residuals are randomly spreaded out along the horizontal center line, which means the residual variance is constant. Therefore, the assumption 2 and 3 were satisfied. Also, the Normal Q-Q plot showed fairly well fitted straight line along the straight line, which means that residuals follow normal distribution; therefore the assumption 5 was satisfied. Moreover, the Scale-Location plot showed linear trend and showed randomly spreaded out residuals, therefore we were reassured that the assumption 3 was satisfied. The Residual vs. Leverage plot did not show any outlier as before, therefore we did not worry about any possible outliers influencing the transformed model.

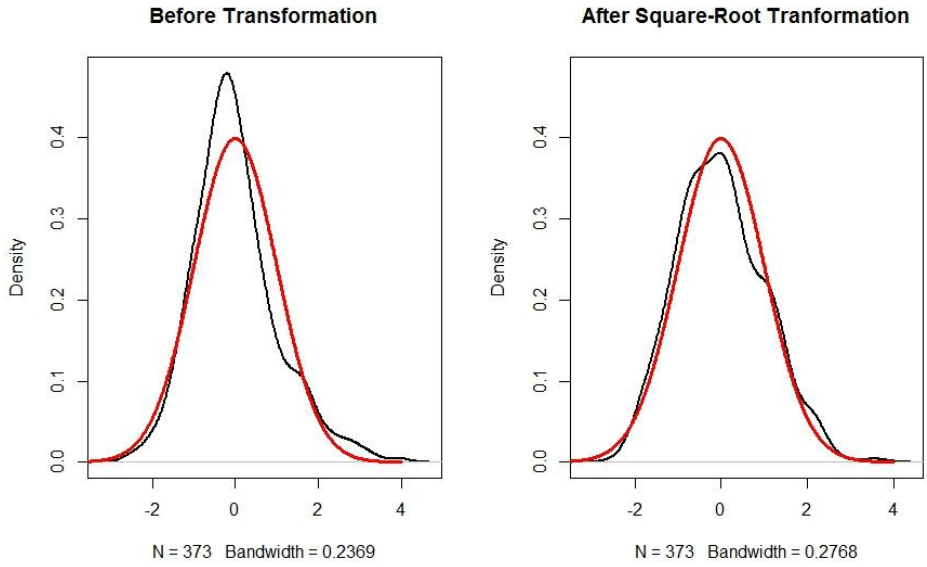

Figure 7: The studentized-residuals distribution plot

Figure 7 clearly showed that the square-root transformation improved the model; the residuals follow approximately normal distribution and have zero mean, therefore we were assured that the assumption 5 was satisfied. Therefore the transformed model satisfied the five assumptions of the linear regression; hence we were assured that the model was ready for the model selection procedure.

Model Selection:

We used the forward, backward and stepwise model selection methods to construct the best models in AIC and BIC with most significant variables. The best AIC Model was:

Best AIC Model: sqrt(SALARY) = β0 + β1' * AST + β2' * AGE + β3' * Height + B4' * MPG + β5' * POINTS + β6' * STL

The summary of the Best AIC Model is:

Table 1: Summary of the Best AIC model		
Coefficients	Estimates	P-value
(Intercept)	-5.0222	3.29e-08 ***
AST	0.00257	0.000997 ***
AGE	0.0564	4.68e-14 ***
Height	0.0211	3.88e-06 ***
MPG	0.0261	0.000607 ***
POINTS	0.000588	0.0125 *
STL	-0.00432	0.0132 *

On the other hands, the best BIC model was:

Best BIC Model: sqrt(SALARY) = β0 + β1' * AST + β2' * AGE + β3' * Height + β4' * MPG + β5' * POINTS

The summary of the Best BIC Model is:

Table 2: Summary of the Best BIC model		
Coefficients	Estimates	P-value
(Intercept)	-5.050	4.38e-09 ***
AST	0.00218	2.68e-06 ***
AGE	0.0581	1.27e-14 ***
Height	0.0211	6.40e-07 ***
MPG	0.0221	0.00269 **
POINTS	0.000452	0.00975 **

According to Table 1 and 2, both AIC and BIC model had estimated coefficients with p-value smaller than 0.05, therefore all the explanatory variables in either models were significant. On the other hands, the main difference of the best AIC model and BIC model was that the best AIC model had one more explanatory variable, which was STL. In order to test which model would be suitable as the final model, we used ANOVA test on the two models as shown in the following table:

Table 3: ANOVA table for BIC Model vs. AIC Model						
Model	Residual DF	RSS	DF	Sum of Sq	F	Pr (> F)
BIC Model	367	132.23				
AIC Model	366	130.15	1	2.0736	5.831	0.0162 *

The null hypothesis of F-test in the ANOVA table was {H0: The best AIC model, which contained one more explanatory variable, was not significantly better than the best BIC model}. Since the p-value of F-statistic was 0.0162, which was smaller than 0.05, we rejected the null hypothesis at the 0.05 level of significance. Therefore, there was strong evidence that the best AIC model is significantly better than the best BIC model and we decided to select the best AIC model as our final model:

Final Model: sqrt(SALARY) = β0 + β1' * AST + β2' * AGE + β3' * Height + B4' * MPG + β5' * POINTS + β6' * STL

Analysis on the Final Model:

Since we constructed the final model, we again used diagnostic plots to check whether the final model still satisfied the five assumptions:

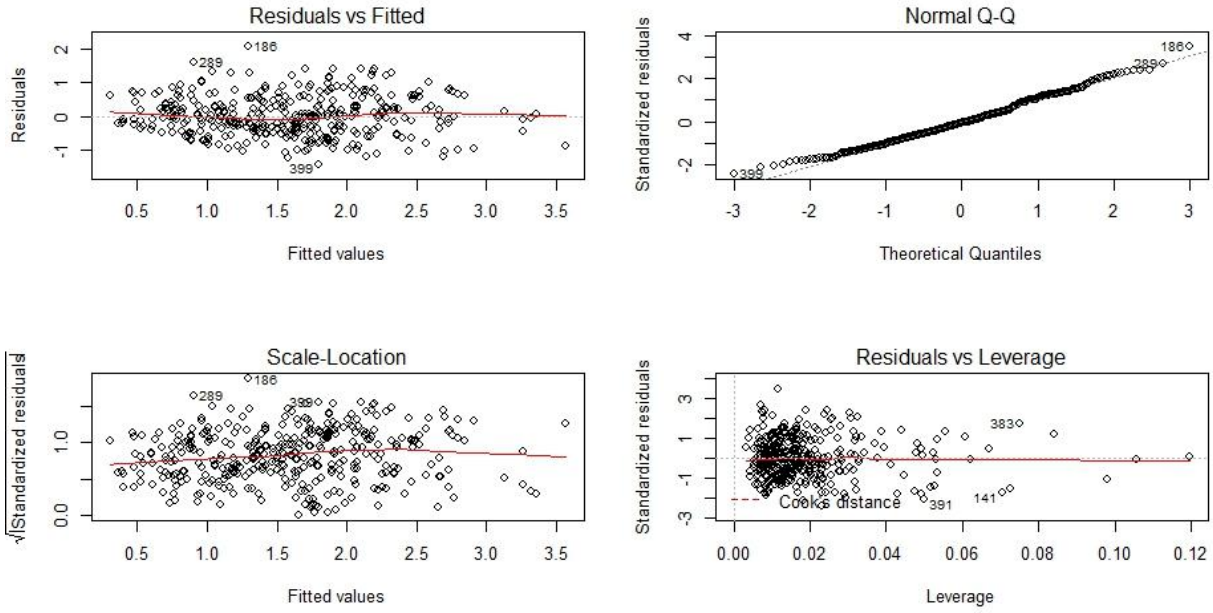

Figure 8: The Diagnostic Plots of the Final Model

In Figure 8, the residuals vs. fitted plot showed a linear trend, which means that the response variable is a linear function of the explanatory variables; therefore the full model satisfied the assumption 1. Also, the residuals were randomly spreaded out along the horizontal center line at zero, which means the residual variance was constant and the residuals have zero mean; therefore the full model satisfied the assumption 2 and 3. Moreover, the Normal Q-Q plot showed well fitted graph along the straight line, which means the residuals followed

approximately normal distribution; therefore the full model satisfied the assumption 5. The scale-location plot also showed linear trend with randomly spreaded out residuals, which reassured us that the assumption 3) was indeed satisfied. In addition, the residual vs. leverage plot showed no influential cases, therefore we did not need to worry about any possible outliers those could influence the final model.

Conclusion and Discussion

Even though the summary table of the final model is equivalent to that of the best AIC model, we presented the table again for convenience. The following table is the summary of the Final Model:

Table 4: Summary of the Final Model

Coefficients	Estimates	P-value
(Intercept)	-5.0222	3.29e-08 ***
AST	0.00257	0.000997 ***
AGE	0.0564	4.68e-14 ***
Height	0.0211	3.88e-06 ***
MPG	0.0261	0.000607 ***
POINTS	0.000588	0.0125 *
STL	-0.00432	0.0132 *

In Table 4, all six explanatory variables have estimated coefficients with p-value smaller than 0.05, therefore all six explanatory variables are significant.

Based on the presented statistics in Table 4, we drew six conclusion statements:

1. One year increase in the NBA player's age is associated with $56,400 increase in the player's salary, given that all the other explanatory variables are fixed.
2. One minute increase in the NBA player's minutes played per game is associated with $26,100 increase in the player's salary, given that all the other explanatory variables are fixed.
3. One centi-meter increase in the NBA player's height is associated with $21,100 increase in the player's salary, given that all the other explanatory variables are fixed.
4. One assist increase in the NBA player's total assists made is associated with $2,570 increase in the player's salary, given that all the other explanatory variables are fixed.
5. One point increase in the NBA player's total points scored is associated with $588 increase in the player's salary, given that all the other explanatory variables are fixed.
6. One steal increase in the NBA player's total steals made is associated with $4,320 decrease in the player's salary, given that all the other explanatory variables are fixed.

However, in order to benefit from using the conclusion statements in real cases, careful interpretation would be required, because there exist complicated reasons behind the linear relationships between the salary and the six explanatory variables.

For statement 1, even though players tend to get higher salary as they age due to their improvement in basketball skills, at certain age around 32, players' athletic ability start to decrease due to the aging, therefore such factors must be taken into account and we should not expect players' salary would increase indefinitely. Moreover, the increase in player's salary is also highly related to the maximum possible contract; maximum possible salary is non-linearly related with the player's number of years played in NBA, which is linearly related with players' age.

For statement 2, even though players with high minutes played per game tend to have high salary, it is may not solely because of their amount of minutes played. Generally, only well-performing players get long minutes to play the games, therefore we should not hesitantly make conclusion about causality between minutes played per game and the salary.

For statement 3, because the facts those taller players are much harder to find and tall players are required for the role of center, tall players tend to get higher salary then the shorter players if their performance is equivalent. However, it is not rare to find short players with higher salary then the tall players.

For statement 4 and 5, it was a surprising result that the total assists made had higher influence in players' salary than the total points scored. The reasons behind these results could be that since basketball is a team game and there is only one ball available during each game, not only it is important to score with the ball, but it is much important to assist teammates to score. From the results of the statements 4 and 5, we realized that among NBA teams, the ability of making high number of assists was viewed much valuable than the ability of scoring high number of points without assisting teammates.

For statement 6, it was also surprising result that the total steals made had negative influence in the players' salary, because it is generally known that making steals in the game has positive impact in winning the game. The reasons behind the statement 6 could be as following: Man-to-man defense is widely used defense tactic for NBA teams these days and making steals requires a player to accelerate towards the opponent's basket; away from the player's own basket, therefore such movement leaves his/her mark man wide open and potentially increases the chance of giving out easy points to the opponents. Since such behavior is considered risky in basketball, high number of total steals made could indicate that the player is not a solid defender, but a risky defender.

Moreover, there are some limitations in our model: there exist uncertain factors that could highly influence NBA players' salary, such as situation of free agent market, injuries, and work ethic, but our model did not consider these uncertain factors. Therefore our model can only be used in limited occasions. For the future analysis project, we would like to research about a method to quantify these uncertainties to include such factors into our model.

In conclusion, we have constructed a multiple linear regression model to determine the factors that affect the NBA players' salary; we identified the six most significant factors that influenced the NBA players' salary: player's age, minutes played per game, player's height, total assists made, total points scored, and total steals made, and these six factors were linearly related with NBA players' salary. Therefore, based on our study, as a professional consultant to help NBA players to capture highest possible salary contract, we would recommend the players to focus on increasing the number of years played in NBA, increasing minutes played per game, increasing total assists made, increasing total points scored, and decreasing total steals made.

Reference

1. Weisberg, Sanford (2005) Applied Linear Regression:Third Edition.

2. Pagano, M. , Gauvreau, K. (2000) Principles of Biostatistics: Second Edition.

3. Kim, B, (2015, September 21) Understanding Diagnostic Plots for Linear Regression Analysis from http://data.library.virginia.edu/diagnostic-plots/

4. https://www.kaggle.com/drgilermo/nba-players-stats-20142015/data

5. https://github.com/datadavis2/nbasalaries