Bradley Chen, Lingrui Fan, Sean Ream

29 November 2019

Predicting an NBA Player's Salary From Their Stats

Introduction

Basketball players in the NBA are some of the world's top paid athletes, with players like Lebron James, Stephen Curry, and Kevin Durant making up three of the top ten highest paid athletes in the world. There are a multitude of factors that contribute to a player's salary. We will investigate key player stats to see which contribute to a player's salary and how much they contribute.

About The Data

The dataset we used was a collection of player metrics from a basketball reference website and focuses on the 2017-2018 season and includes basic statistics such as position, team, and free throws as well as advanced statistics such as true shooting percentage which is a measure of shooting efficiency that takes into account field goals. The dataset has 651 observations and the original dataset contains 31 different variables. Through discussion, we came up with these predictors we would like to focus on in our analysis. Variable descriptions in more detail such as the formula used to calculate can be found here:

USG%: Usage percentage is an estimate of the percentage of team plays used by a player while he was on the floor.

G: Games played

Age: Age of player in years old

Pos: What position the player plays

TS%: True shooting percentage is a measure of shooting efficiency that takes into account field goals, 3-point field goals, and free throws

AST%: Assist percentage is an estimate of percentage of teammate field goals a player assisted with while he was on the floor

STL%: Steal percentage is an estimate of percentage of opponent possessions that end with a steal by the player while he was on the floor

ORB%: Offensive rebound percentage is an estimate of offensive rebounds grabbed by the player while on the floor

DRB%: Defensive rebound percentage is an estimate of defensive rebounds grabbed by the player while on the floor

PER: Player efficiency rating is a per-minute rating of a player's performance.

MP: Minutes played

VORP: Value over replaced player is the estimated increase or decrease in contribution that a player provides over who he replaced

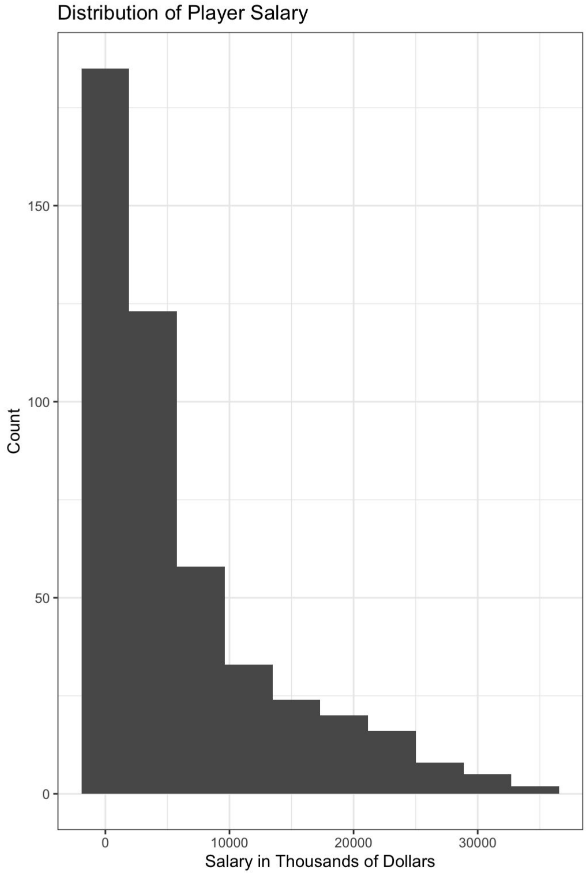
WS: Win shares is an estimated number of wins contributed by the player

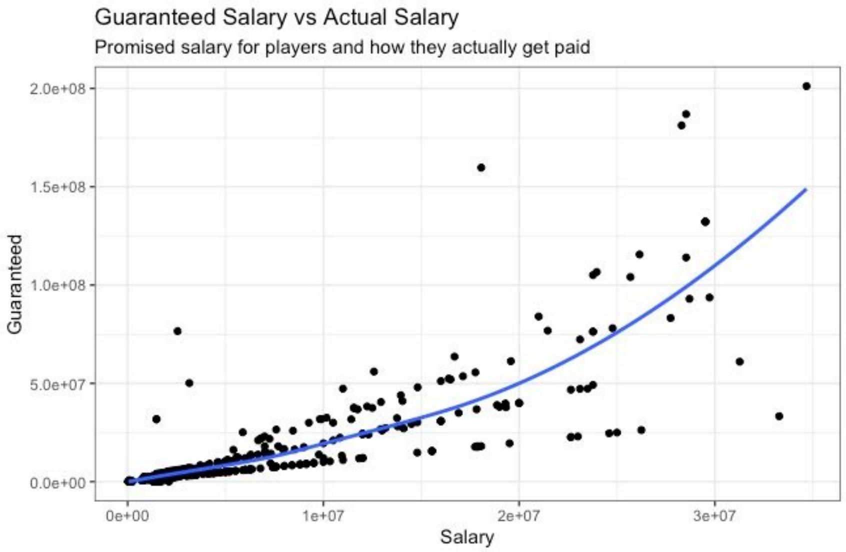
Draft Number: The player's draft number

We picked these variables because we felt that these were key statistics pertaining to the performance of a player and could ultimately determine how much a player is worth. These variables quantify key activities of a player during a game, such as catching rebounds, stealing balls, and shooting baskets as well as important personal statistics such as age and how experienced the player is.

Variable Cleanup

We decided to remove all players who are not from the United States because the sample size for players not from the US was not large enough, the largest being Canada which was around 15 players. The variables Salary and Guaranteed were converted from character type into an integer type. Looking at the distribution of the response variable salary below, we can see that the distribution is unimodal and skewed right. The minimum salary for a player in the United States is $46,080 and the maximum is $34,682,550. Some of the observations in the data set contained salaries of $0, so they were removed as well.

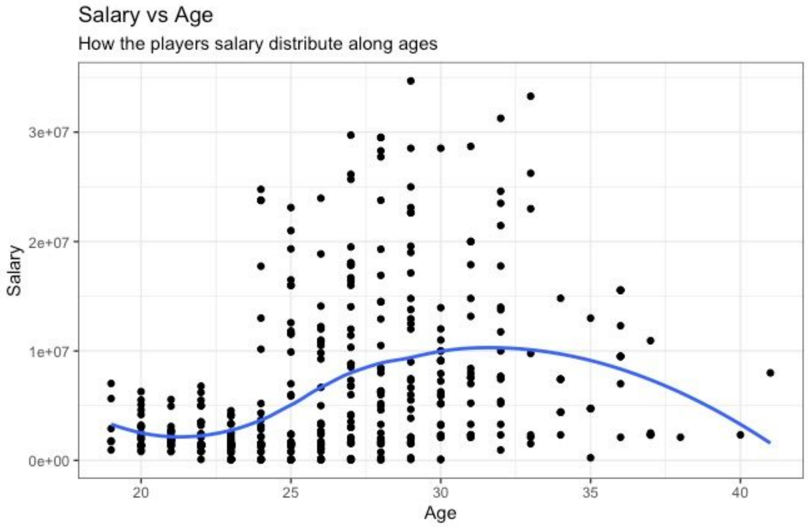




We also looked at the relationship between guaranteed salary of players and the actual salary they received. Guaranteed salary does not always match the actual salary, in fact it seems that predicted salary tends to be lower than the guaranteed salary. This could be because the salary of some players is guaranteed with multiple seasons and we are currently just looking at one season.

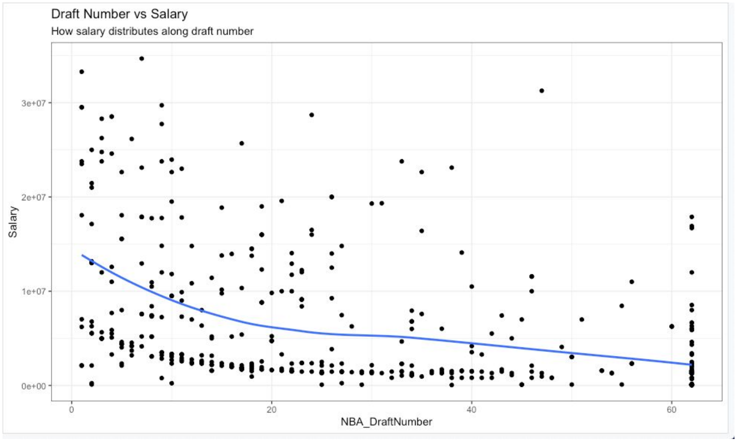
Initial Model Analysis

We estimated the regression models of natural log of salary with every single variable we have selected to get a basic sense of how these variables affect salary individually. One of the most important and common variables that can affect an athlete's salary is age, which is the very first variable we used for estimation. The regression estimation for Salary against Age indicates that with every additional increase on a player's age, his salary is expected to increase by 0.512% on average. The predicted salary of an average aged player (26 and a half years old) is $2,703,343. The age factor is statistically significant on affecting the Salary and the residuals are normally distributed. We also plotted the distribution of Salary against Age since it is so important. The plot shows a normally like distribution of Salary over ages from 19 to 41 with a peak at around 32. It is seen that most of the players started to have their salary increase around age 23 and received a relatively high salary from 26 to 34. Their salary starts to drop as they get older.



Salary vs Age

How the players salary distribute along ages



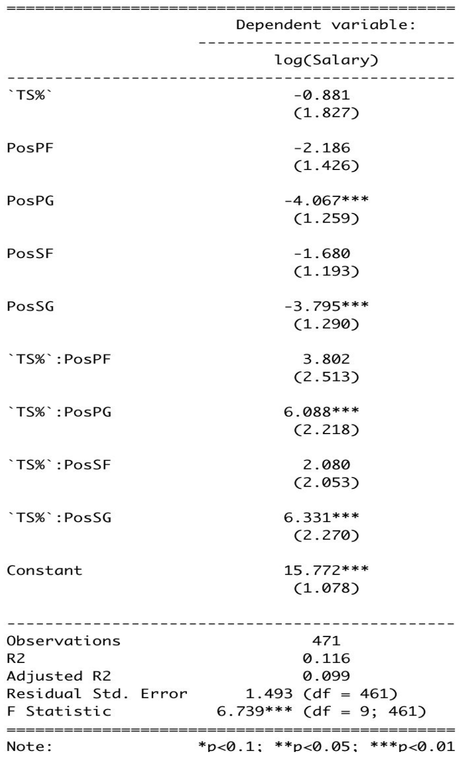
Draft Number vs Salary

How salary distributes along draft number

The NBA draft number is also a strong indicator of salary. The draft format suggests that the lower the number, the better the player. The regression shows a clear negative correlation between the salary and draft number and also a statistical significance on the effect. The estimations by player's usage rate, minutes played, and games played indicate that a player's salary can also increase if he spends more time playing. Though the regressions for these three variables suggest their significant influence on salary, the residuals are heteroskedastic so the separate regressions do not fit the data very well. The regression models on other variables such as true shot rate and steal percentage all show a positive correlation with salary and all suggest statistical significance for their influences. But the residuals plots for those estimations all indicate that they do not fit the data very well.

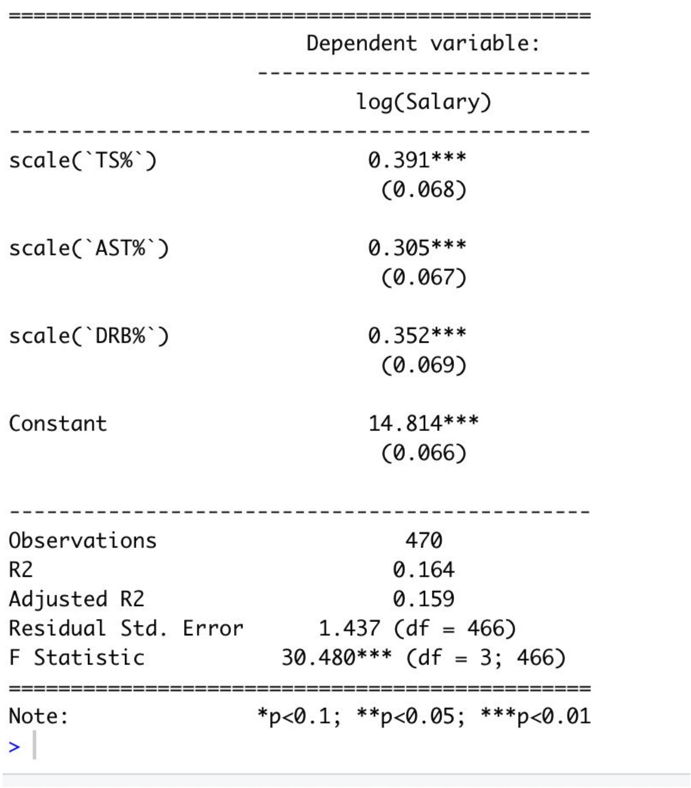
Effect Position Has On Salary Using True Shooting Percentage

Since the variables evaluating players' performance may have an indirect effect on salary with different positions, we chose one of the most representative variables, true shooting percentage, to estimate the additional influence position has on salary measured by it. We regressed a model predicting salary using true shooting percentage (TS%), position (Pos), and an interaction between them. From the summary of the model, it is seen that the increase rate of salary over true shooting percentage is particularly large for two position: Point Guard (PG) and Shooting Guard (SG), and the increase are 6.09% and 6.33% respectively with 1% increase in true shooting percentage. The effect of true shooting percentage for those two positions all show statistical significance.



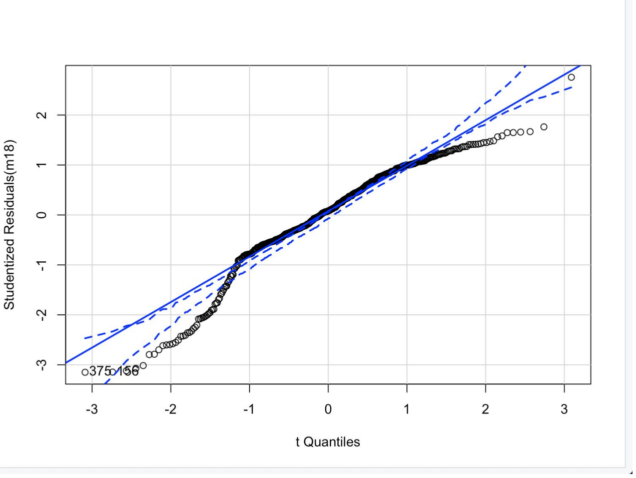
```
=============================================
                      Dependent variable:
                  ---------------------------
                          log(Salary)
---------------------------------------------
`TS%`                       -0.881
                            (1.827)

PosPF                       -2.186
                            (1.426)

PosPG                      -4.067***
                            (1.259)

PosSF                       -1.680
                            (1.193)

PosSG                      -3.795***
                            (1.290)

`TS%`:PosPF                  3.802
                            (2.513)

`TS%`:PosPG                6.088***
                            (2.218)

`TS%`:PosSF                  2.080
                            (2.053)

`TS%`:PosSG                6.331***
                            (2.270)

Constant                  15.772***
                            (1.078)

---------------------------------------------
Observations                  471
R2                           0.116
Adjusted R2                  0.099
Residual Std. Error     1.493 (df = 461)
F Statistic          6.739*** (df = 9; 461)
=============================================
Note:              *p<0.1; **p<0.05; ***p<0.01
```

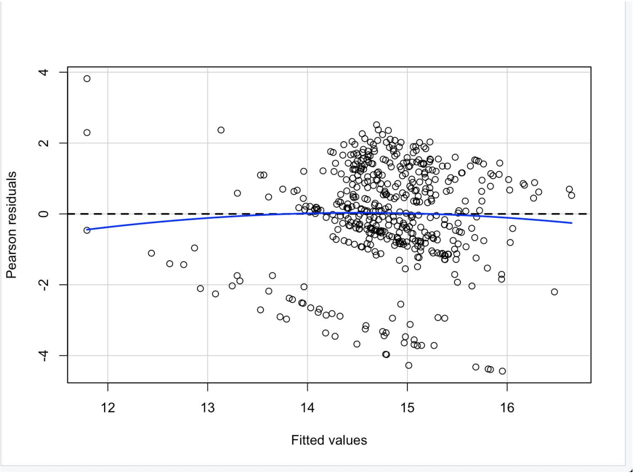
Final Models

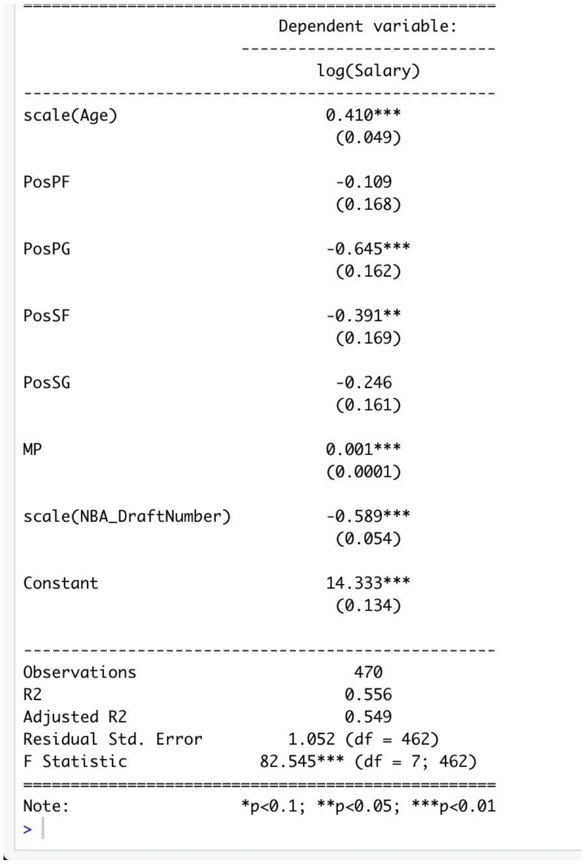


```
===============================================
                        Dependent variable:
                    ---------------------------
                            log(Salary)
-----------------------------------------------
scale(`TS%`)                 0.391***
                              (0.068)

scale(`AST%`)                0.305***
                              (0.067)

scale(`DRB%`)                0.352***
                              (0.069)

Constant                     14.814***
                              (0.066)

-----------------------------------------------
Observations                    470
R2                             0.164
Adjusted R2                    0.159
Residual Std. Error      1.437 (df = 466)
F Statistic           30.480*** (df = 3; 466)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
>
```

One of the final models that we came up with predicted salary from key player metrics. Of the original predictors that we picked that had to do with player performance on court, we found that true shooting percentage, assist percentage, and defensive rebound percentage were significant in predicting a players salary. We mean centered all of these variables, since players with a 0 in any of these stats would not be in the NBA. We found that a player with average stats in all these predictors would have a salary of $2,714,178. To interpret the coefficients individually, a 1% increase in True shooting percentage is associated with a 0.391% increase in salary for a player with average assist and defensive rebound percentages, a 1% increase in assist percentage corresponds with a 0.305% increase increase in salary for a player with average true shooting percentage and average defensive rebounds, and a 1% increase increase in defensive rebound percentage corresponds with a 0.352% increase in salary for a player with average true shooting percentage and average assist percentage. It seems that an increase in true shooting percentage is associated with the largest increase to a player's salary for player statistics and that if a player wants to increase their salary, they should consider improving their shooting. The VIFs for this model were all around 1 suggesting there is no danger of multicollinearity. Looking at diagnostic plots below of the residual plot and QQPlot we found that the residuals seem to be mostly homoskedastic and although some of the points in the QQPlot fall outside of the bounds, it is reasonable to assume the data is relatively normal.

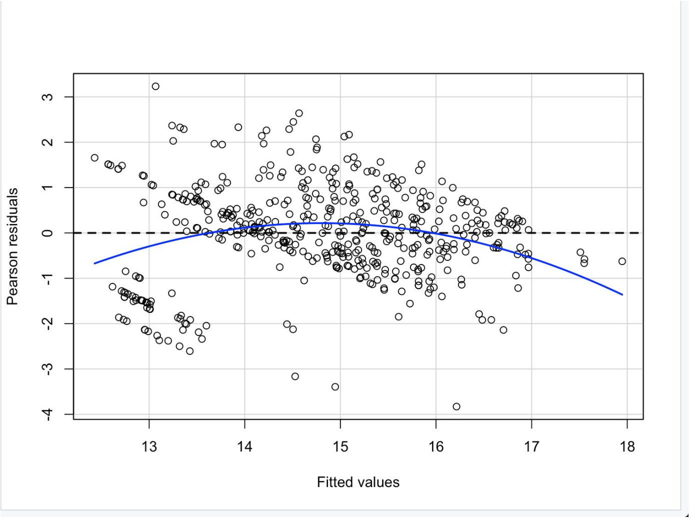


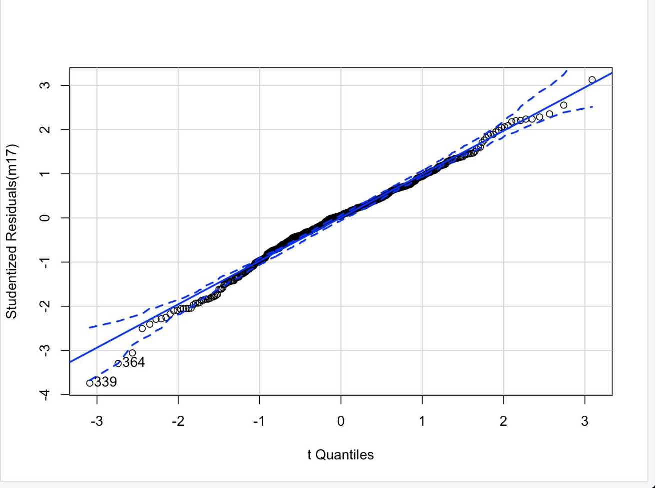


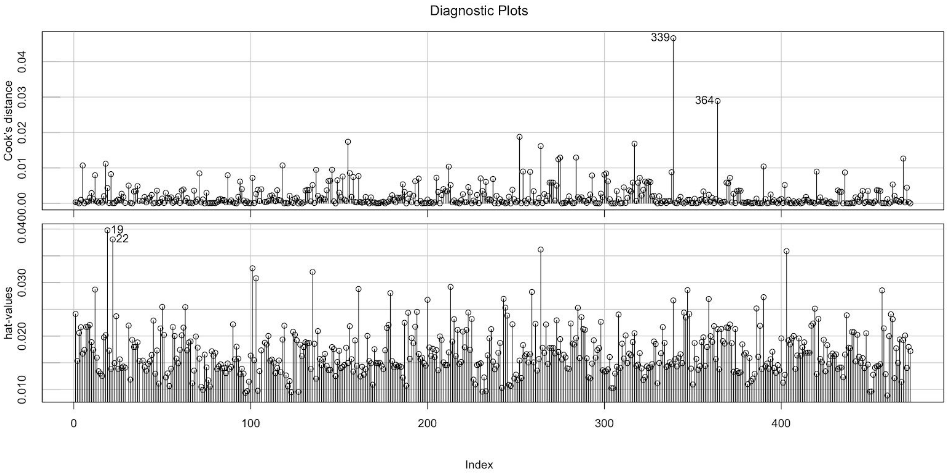


```
==================================================
                         Dependent variable:
                     ---------------------------
                             log(Salary)
--------------------------------------------------
scale(Age)                    0.410***
                              (0.049)

PosPF                         -0.109
                              (0.168)

PosPG                        -0.645***
                              (0.162)

PosSF                        -0.391**
                              (0.169)

PosSG                         -0.246
                              (0.161)

MP                            0.001***
                             (0.0001)

scale(NBA_DraftNumber)       -0.589***
                              (0.054)

Constant                     14.333***
                              (0.134)

--------------------------------------------------
Observations                    470
R2                             0.556
Adjusted R2                    0.549
Residual Std. Error       1.052 (df = 462)
F Statistic           82.545*** (df = 7; 462)
==================================================
Note:                  *p<0.1; **p<0.05; ***p<0.01
>
```

The last model we created predicted a player's salary through player information such as their age, the position they play, total minutes played, and their draft number. Again, mean centering needed to be done, in this case to age and draft number since it does not make sense to have zeroes in either of these values. We found that a player who is a center with average age, 0 minutes played and an average draft number will earn $1,677,810. A small forward of average age with 0 minutes played and an average draft number will have a salary that is 0.391% lower than a center and a point guard of average age with 0 minutes played and an average draft number will have a salary that is 0.645% lower than a center. For each additional year of age, salary goes up by 0.41% for a player who is a center that had an average draft number and 0 minutes played. Each additional minute played is associated with an increase in salary of 0.001% for a player who is a center of average age with an average draft number. For each unit increase in a player's draft number, his salary decreases by 0.589% for a player who is a center with 0 minutes played and of average age.







Looking at the diagnostics, we can see that there is some heteroskedasticity, but we believe that this is not too big of an issue since this is from raw, real world data. Looking at the QQPlot, we see that all residuals are well within bounds and that the model looks normally distributed. Moving onto the Cook's Distance and hat-values, we see that the Cook's distance is well under 0.5 save for a couple points, and the hat-values are also within range save for a couple of points.

Takeaways, Potential Weaknesses, and Conclusion

Through our analysis, we found that players aged 30-35 earn comparatively more than players in other age groups, an increase in true shooting percentage is associated with the largest increase to a player's salary for player stats and finally a player who is a center is associated with the largest salary compared to the other four positions. One potential weakness is that there was not more player data in other countries, which would have been helpful in determining if the factors that we found that influenced salary applies in other countries. Another weakness in our model is that we were not able to analyze player data from their college careers, which could influence their draft number and salary. In summary, using our models can help predict an NBA player's salary using game performance information and personal factors. This type of analysis could be useful for teams in the NBA to determine how much a player is worth and how much they should increase or decrease current players' salaries when renewing contracts and how much to offer incoming players.