Ukrainian Catholic University

Faculty of Applied Sciences

IT and Business Analytics

Examining ceteris paribus relationship between an NBA player's performance and his salary in regular season 2019-2020

Econometrics course

Author:

Maxym KRYVAL

April 2021

1. Introduction

The only measurement of a basketball player's efficiency is indeed his/her on-court performance. Nevertheless, the mentioned statement may not apply to explaining his/her salary. This study aims to investigate whether statistical factors (points, assists, rebounds, personal fouls, shooting percentage) are sufficient to explain an NBA player's salary for a given season.

In this research, the main questions I tried to find detailed, analyzed, and proven answers to are the following:

- Do a player's performance measurements of the regular season 2019-2020 are sufficient to explain his financial compensation for the given year?
- What are the most essential game factors concerning annual salary?
- Can we build a statistically significant regression model based on game performance indicators to determine an NBA player's salary for a given season?

2. Data

2.1 Description

The data picked for analysis consists of two parts: personal statistics and salary. The information about the on-court performance was downloaded from *Kaggle.com* (Source). As for the salary information, due to the incompleteness and inaccuracies of a few datasets on Kaggle, it was web-scraped from resource *Hoopshype.com* (Source).

The performance dataset is divided into several *.csv* files out of which I picked two needed ones: *nba_2020_per_game.csv* and *nba_2020_advanced*. The first one contains basic information such as points, assists, rebounds, etc. As for the second one, the given information is about player's efficiency measures such as winning shares, player efficiency rate, box plus-minus, etc. A detailed explanation of the data columns is provided in this link.

The salary dataset contains the compensation in millions of US dollars of a specific NBA player.

Alongside the mentioned dataset, I used additional ones to retrieve information about an NBA player's pick number. The information towards NBA drafts for the previous years was obtained from Kaggle.com (Source) and the data about draft picks for season 2019-2020 was web-scraped from resource basketball-reference.com (Source).

2.2 Preparing

To efficiently perform the analysis of the obtained data about NBA statistics, cleaning and preparing the datasets is a crucial part.

The most important step was to delete statistical data about the players who did not perform in the NBA regular season 2019-2020. The reason for that is the following: such players as Kevin Durant, Klay Thompson, and John Wall did not participate in the games yet received their annual salary. Therefore, their salary data is huge, while all of the statistical factors are zeros. Including information about these players would cause the presence of numerous data outliers which would lead to the unreliability of the inferences made.

Missing values and outliers would make the modeling process difficult. Therefore, I used an imputation method to fill in the missing values. In this case, I choose the median number of the whole column to fill in the missing variables because a median value will not be influenced by outliers.

The mentioned replacement works for every column of the data except for the Pick column. Picked (or drafted) players possess a specific pick number which varies from 1 to 60: the smaller the number the better the player. However, there are undrafted players who did not make it to the NBA Draft event. There is an inference that these players are worse and this is why filling their value of the Pick column with a median is not correct. Therefore, the replacement value for them is number **61** which stands for the worst pick number. For a better explanation of the NBA Draft procedure, check the source.

Having pre-processed all of the datasets containing information about salaries, draft numbers, and personal statistics, I merged them into a single final file *NBA_stats_salary_2019-2020.csv* which contains 502 rows and 49 columns.

3. Methods and Results

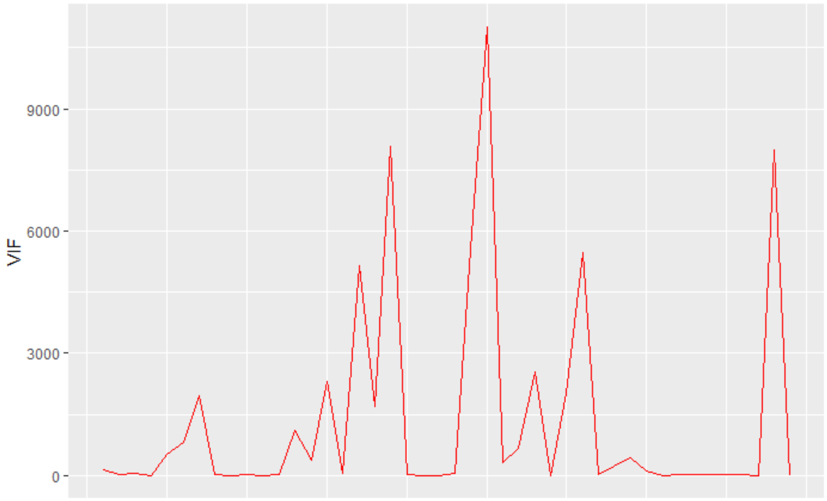
3.1 Significant variables detection

From now on, we should reformulate the phrase ‘rows and columns’. Our dataset contains 502 observations, 43 independent variables of on-court performance which explain the Salary dependent variable.

One can be almost one hundred percent sure that some or a lot of the explanatory variables are not statistically significant, and, therefore, the goal is to define the factors which uniquely describe Salary. The mentioned can be easily proven by examining the **VIF** (Variance Inflation Factor) of the data.

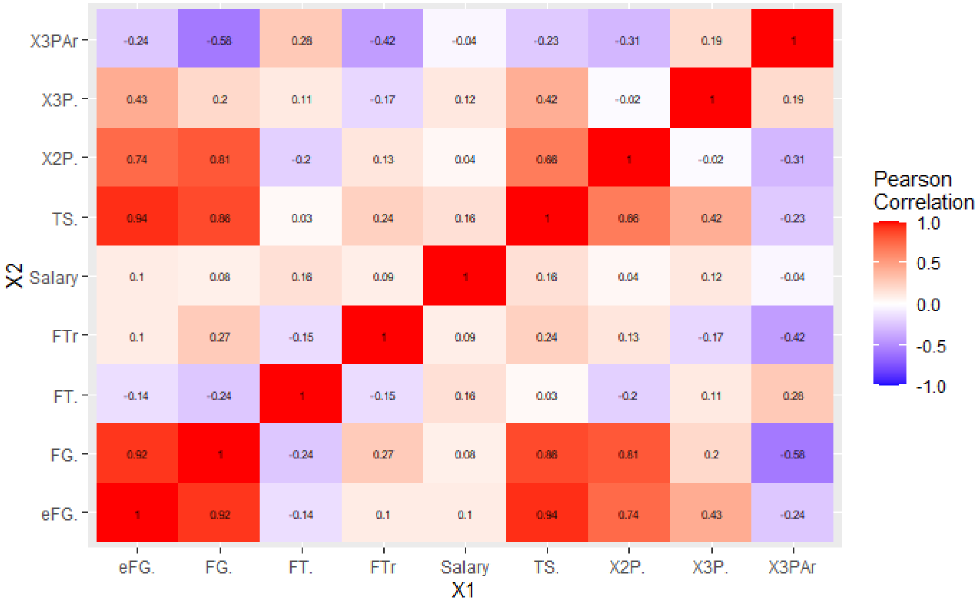
Variance Inflation Factor is used for determining the multicollinearity between independent variables. The method is about taking one variable and regressing it against all the other ones: the higher the R^2 the more one variable explains another one. The basic rule is to check if VIF is greater than 10: if so, then one may worry about multicollinearity issues.

Here is the VIF plot of the data:

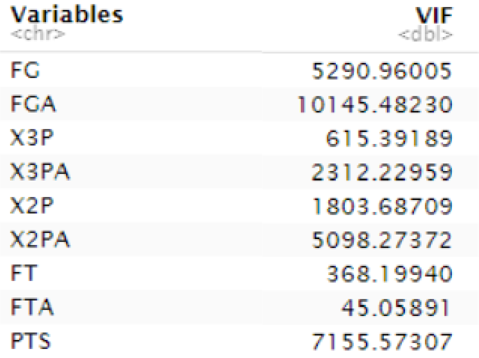


Almost all of 43 variables have VIF much greater than 10 which strictly states that numerous variables are highly correlated between each other and we need the ones that explain the Salary dependent variable uniquely and in the best way.
The first thing to do was to examine the correlation between salary and shooting rates/shooting percentages. The reason is the following: there may be some outsider players who may do 1-2 attempts to shoot during a game and get 50% realization; however, top players make 50-60 attempts to score and their realization could be 30%. This means that shooting percentages and rates are immensely biased indicators that give no significant explanation for a player's salary.

The following could be proven on the correlation heatmap:

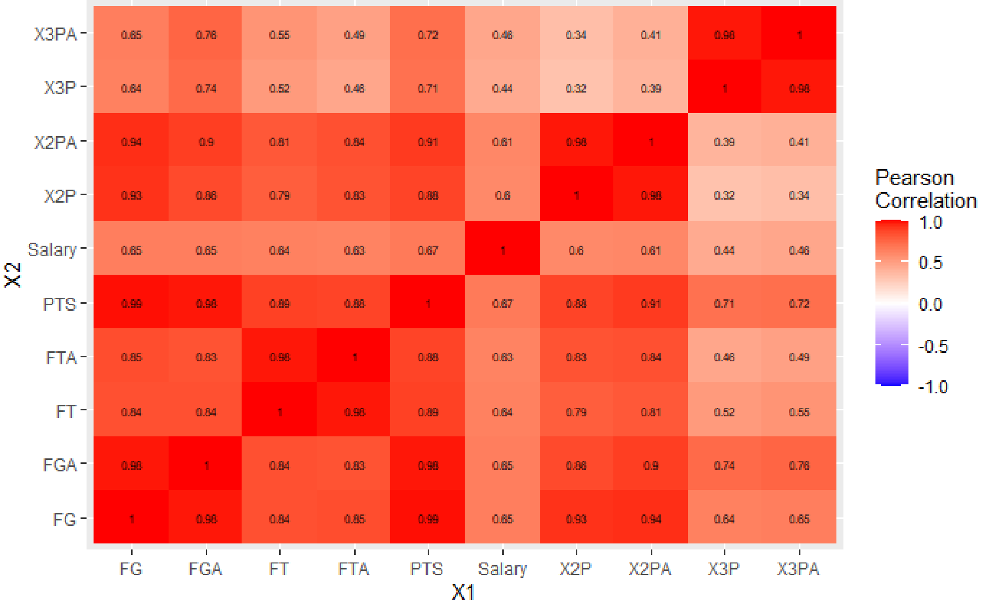


The next step is to define a group of shooting measures (PTS, FTA, X3PA, etc.). The reason is because of the hypothesis that all of these indicators answer to the same question “How well does the player shoot?” yet in a slightly different manner. Owing to VIF values, one can distinctly see below that all of the variables are highly correlated with each other:



Variables <chr>	VIF <dbl>
FG	5290.96005
FGA	10145.48230
X3P	615.39189
X3PA	2312.22959
X2P	1803.68709
X2PA	5098.27372
FT	368.19940
FTA	45.05891
PTS	7155.57307

This is why I picked the variable with the strongest correlation with a player’s salary. From the correlation heatmap below, it is seen that PTS has the greatest correlation with a salary which equals **0.67**:

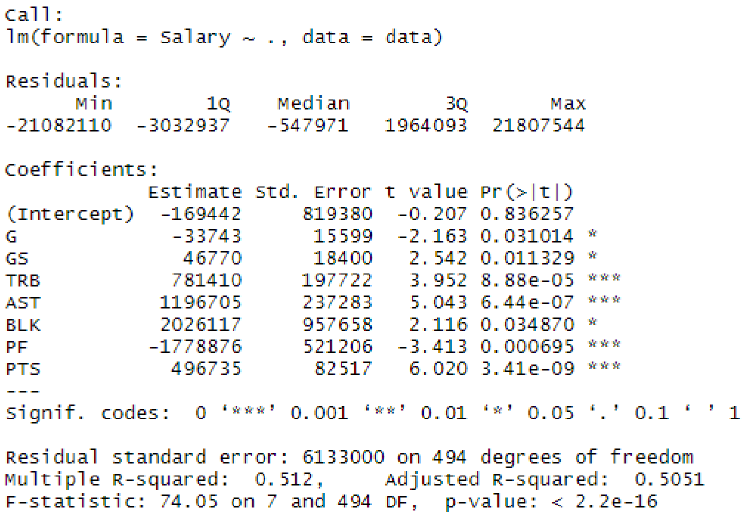


The final thing I did was the removal of all the efficiency rates (PER, BPM, WS, etc.). The cause for that is extremely simple: all of these rates are different formulas that are

based on the indicators such as PTS, AST, TRB, etc. This implies that all of these measures are perfectly explained by other variables.
As the result of the mentioned manipulations and variable analysis, we include only 7 explanatory variables: G, GS, PTS, AST, TRB, PF, BLK.

3.2 Checking variable credibility

To check whether the factors derived in the previous section are credible and sufficiently explain behaviour of an NBA player's salary, regression analysis will be performed.
The first model to run is the naïve model which includes all 43 explanatory variables. Having run the linear regression, almost none of the 43 variables are statistically significant which implies the violation of the multicollinearity assumption. The **adjusted R^2** equals **0.519**.
To check the validity of the theoretically significant variables obtained in the previous section, the proper step to do is running the model including only them, checking for statistical significance, and comparing its adjusted R^2 with the naïve one to see if some explanation of the data was lost. The summary of the theoretically appropriate model is:



```
Call:
lm(formula = Salary ~ ., data = data)

Residuals:
      Min        1Q    Median        3Q       Max 
-21082110  -3032937   -547971   1964093  21807544 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -169442     819380  -0.207 0.836257    
G             -33743      15599  -2.163 0.031014 *  
GS             46770      18400   2.542 0.011329 *  
TRB           781410     197722   3.952 8.88e-05 ***
AST          1196705     237283   5.043 6.44e-07 ***
BLK          2026117     957658   2.116 0.034870 *  
PF          -1778876     521206  -3.413 0.000695 ***
PTS           496735      82517   6.020 3.41e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6133000 on 494 degrees of freedom
Multiple R-squared:  0.512,	Adjusted R-squared:  0.5051 
F-statistic: 74.05 on 7 and 494 DF,  p-value: < 2.2e-16
```

One can see that the **p-values** for all of the variables are statistically significant. The mentioned implies that all of them uniquely describe the dependent variable and the model meets the assumption of multicollinearity absence. The **adjusted R^2** is almost the same as the naïve model's one which means that almost no explanation of the data was lost after removing 35 explanatory variables.

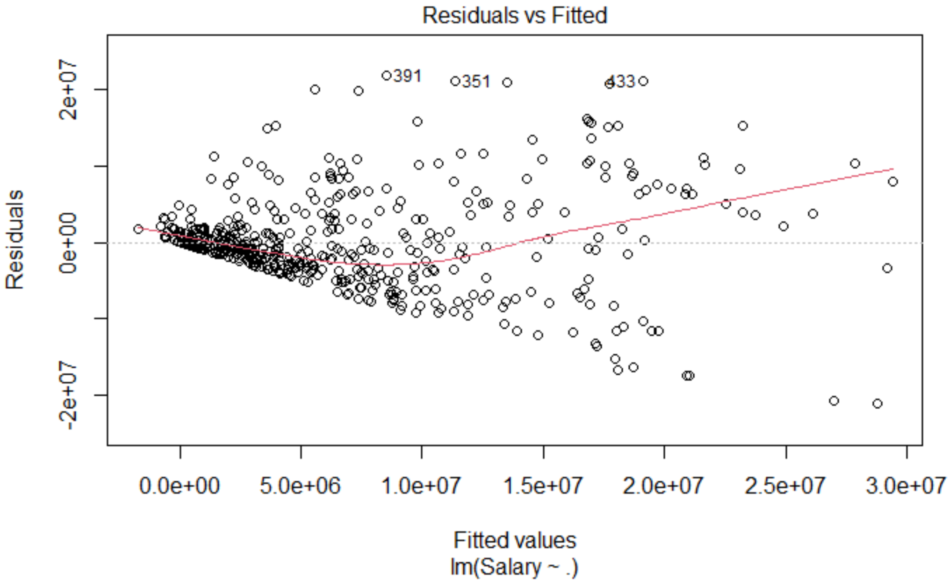
3.3 Validating and transforming model

In the previous section, we obtained a model with theoretically and statistically significant variables which explain an NBA player's salary. Generally, the R^2 value of approximately **0.51** does not tend to explain sufficient variability. However, to make significant inferences about the competency of the explanation by the obtained variables, regression analysis needs to be performed.

To validate the linear model from the previous section, the check for linear model assumptions should be made:

The assumption I: Linear relationship between the Y and X variables.

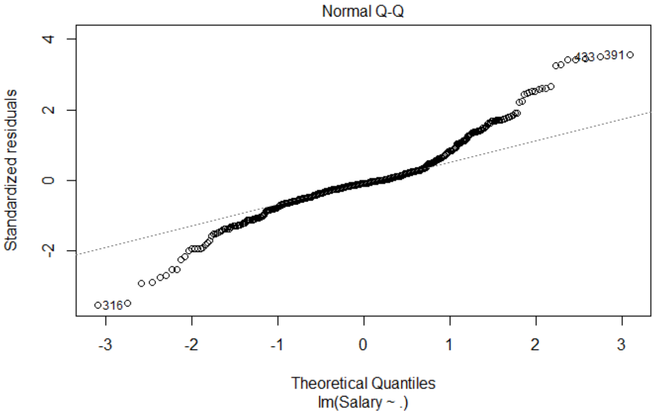
Here is the Residuals vs Fitted plot:



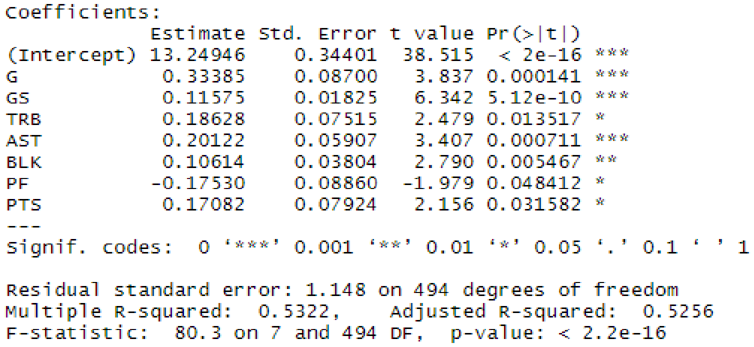
The curviness of the red line indicates the absence of linear relationship in the model which violates the first assumption.

The assumption II: Normal distribution of error term.

Here is the Q-Q plot of the residuals:



One can see that the observed residuals differ a lot from the theoretical ones. This implies that the assumption is violated as well.
The assumption III: no perfect multicollinearity. The assumption was already checked in the previous section and it is not violated. All of the variables uniquely describe the dependent variable.
The assumption IV: no heteroskedasticity of the data. To check the following assumption, I ran Breusch-Pagan test. The null hypothesis of the test states that there is constant variance while the alternative hypothesis states the opposite. It resulted in an extremely small p-value which means that heteroskedasticity is present and the assumption IV is violated as well.
As 3 out of 4 assumptions are violated, the model was transformed into Log-Log type. After the executed transformation we met an approximately linear relationship between the dependent variable and independent ones and the error term became normally distributed. All of the variables still remained significant and the R^2 became a bit better.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.24946    0.34401  38.515  < 2e-16 ***
G            0.33385    0.08700   3.837 0.000141 ***
GS           0.11575    0.01825   6.342 5.12e-10 ***
TRB          0.18628    0.07515   2.479 0.013517 *  
AST          0.20122    0.05907   3.407 0.000711 ***
BLK          0.10614    0.03804   2.790 0.005467 ** 
PF          -0.17530    0.08860  -1.979 0.048412 *  
PTS          0.17082    0.07924   2.156 0.031582 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.148 on 494 degrees of freedom
Multiple R-squared:  0.5322,	Adjusted R-squared:  0.5256 
F-statistic:  80.3 on 7 and 494 DF,  p-value: < 2.2e-16
```

The problem with non-constant variance is still unsolvable. Hence, it is another indicator that supports the theory about the insufficiency of on-court performance factors to explain an NBA player's salary.

3.4 Executing and evaluating models

In the previous sections, I obtained significant variables that uniquely affect the dependent variable. I transformed the data into the logarithmic form so that more of the linear model assumptions were met. To make final inferences to answer the question of whether on-court performance sufficiently explains an NBA player's salary, different types of models will be run and analyzed: linear, Ridge, and Lasso regressions. The assumption of Ridge and Lasso regressions are identical to the ones of basic linear regression which were checked in the previous chapters.
Ridge regression is quite similar to simple linear regression. However, instead of just fitting a line with the regular **OLS** (Ordinary Least Squares) algorithm to the training data, it adds a **degree of bias** to the regular line so that predicted testing values would be less biased. In other words, ridge regression executes bias-variance trade-off: worse initial fit for better prediction of future values. The degree of bias is determined by the sum of

squared coefficients near the independent variables multiplied by a parameter **ƛ** (optimal ƛ is defined using **cross-validation** technique).

Lasso regression in its turn is quite similar to Ridge regression. The difference is in the degree of bias which is determined by sum of absolute values of slopes multiplied by a parameter **ƛ**. The benefit of such a degree of bias performs well when the model has some insignificant variables: their slope will be shrunk to 0.

To finally evaluate regression models and check the validity and error term of the predictions, the dataset was divided into train and test parts with the relation 2:1 respectively.

As predicted values of logarithmic form are not suitable for interpretation, I retransformed them back into regular numeric form by exponentiating the value.

Having run the model, for an average salary of **$7** mln, the average prediction **RMSE** (Root Mean Square Error) for Ridge, Lasso, and LM predicted values are **$4.46**, **$4.83**, and **$4.4** mln respectively. Although multiple linear regression predicts most efficiently, none of the models are reliable to make inferences about an NBA player's salary as the error terms are too high for the aforementioned average of salary. Eventually, on-court performance data is not sufficient to explain an NBA player's salary.

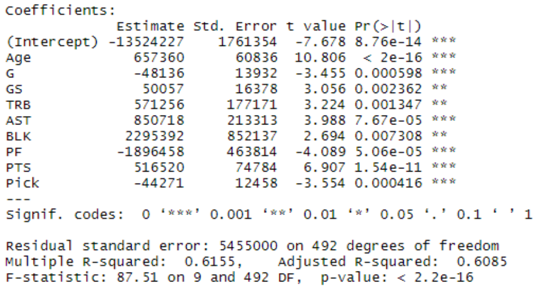
4. Conclusions, further steps

4.1 Conclusions

The inferences of the work are the answers to the questions set in the introduction of the research.

Question 1: Do a player's performance measurements of the regular season 2019-2020 are sufficient to explain his financial compensation for the given year?

The answer is indeed negative. The initial explanation of the data by different models was relatively small. However, to prove that with more details I tried to examine the data and build several models. None of the models gave an adequate error term of salary prediction. The problem was in the heteroskedasticity of the data which is surely caused by the lack of significant explanatory variables. To prove the conclusion with an example, I added a few more explanatory variables which are not related to an NBA player's on-court performance: age and draft number. Here is the summary of a linear model which involves the added variables:



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -13524227    1761354  -7.678 8.76e-14 ***
Age            657360      60836  10.806  < 2e-16 ***
G              -48136      13932  -3.455 0.000598 ***
GS              50057      16378   3.056 0.002362 ** 
TRB            571256     177171   3.224 0.001347 ** 
AST            850718     213313   3.988 7.67e-05 ***
BLK           2295392     852137   2.694 0.007308 ** 
PF           -1896458     463814  -4.089 5.06e-05 ***
PTS            516520      74784   6.907 1.54e-11 ***
Pick           -44271      12458  -3.554 0.000416 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5455000 on 492 degrees of freedom
Multiple R-squared:  0.6155,	Adjusted R-squared:  0.6085 
F-statistic: 87.51 on 9 and 492 DF,  p-value: < 2.2e-16
```

We can see that adjusted R^2 got instantly greater by approximately 0.1. This distinctly states that other factors of an NBA player are relevant to the explanation of his salary. The **RMSE** for Ridge, Lasso, and LM predicted values are **$4.01**, **$4.12**, and **$4.06** mln which is smaller than the errors obtained by the statistics only model by more than **$0.5** mln.

Hence, we can see that more than just on-court performance data is needed to make inferences about an NBA player's salary.

Question 2: What are the most essential game factors concerning annual salary?

By performing both theoretical and statistical analysis, the factors which uniquely describe a player's annual salary are PTS, AST, TRB, G, GS, PF, and BLK. The reason is in the fact that these factors are the basis for performance explanation; all of the other measures are simply derivatives of these main factors.

Question 3: Can we build a statistically significant regression model based on game performance indicators to determine an NBA player's salary for a given season?

The answer to this question intercepts with the answer to the first one and is negative as well. The reasoning is the following: not all of the explanatory variables are present in the model which causes heteroskedasticity and inadequate error term in predictions made.

4.2 Limitations and next steps

The next steps would involve gathering many more factors besides on-court performance that influence an NBA player's salary. As it was seen in the answer to Question 1 in the previous section, salary explanation and significant model building are more than possible. I reduced the error by **$0.5** mln by just adding two more variables. Knowing there are numerous of them (nationality, race, anthropometric parameters, team rating, player's position, average career performance, recognizability of a player, etc.) one can minimize error term of predictions and as a result, build a statistically significant model.

As for the limitations, the NBA data collection process will be complicated because the information about other out-court factors is dispersed and hard-to-reach on the web.

4.3 Other

Github repository of the project

Data sources:

- On-court statistics
- NBA Draft
- NBA Draft 2019-2020
- Salaries