University of Tennessee – Knoxville

Can NBA Player's Salaries be Predicted from Statistics?

A Research Paper

Caleb Miller

Econ 381

Professor Benjamin Compton

Due 11/23/21

Abstract: This research paper looks at a dataset of statistics of 235 NBA players from the 2020-2021 season to attempt to predict their salaries. Previous studies have asked this same question and have found statistics such as games played, field-goal percentage, etc. to have the highest impact. The goal for this research paper is to build a model that can accurately predict NBA salaries from the chosen statistics and decide which ones have the highest impact. The analysis will include summary statistics, a correlation matrix, a VIF table, and a regression analysis. The findings from this study reveal that age and steals are the two most impactful on the salaries collected. These findings are similar, but not the same to other studies so further research is highly encouraged.

Introduction:

In this research paper, I will be walking you through my regression project. This project will be asking the question: Can NBA player's salaries be predicted from statistics? This question is very interesting because the NBA is a competitive market, and player empowerment picked up steam in the 2010s. This means that players now, more than ever, realize their true value on a market. It is now socially and professionally acceptable to leave the team that drafted you because another team is offering you more money. This being said, I think that a regression analysis will be able to predict variation in salaries because the NBA is an analytics driven league. I had to pick out the variables that will be used in order to create the perfect regression. I decided to use the following variables initially: points, personal fouls, turnovers, blocks, steals, assists, total rebounds, offensive rebounds, free-throw percentage, two-point percentage, three-point percentage, field goal percentage, minutes played, and age. These will be used to predict salary (in millions). The semi-log functional form is used for this analysis. The majority of the variables are fine as linear, but age should use the natural log form. This is because as an NBA player increases in age up to his prime, salary should increase. However, after a player's prime we should see diminishing returns on salary. To accomplish this, I generated a new variable called lnage which will replace the age variable in my regression.

Overview of Literature:

There have been previous studies looking at the same question. For example, "NBA Player Salaries Prediction with Linear Regression" attempts to predict salary from 9 variables including age, blocks, three-pointers made, rebounds, free-throw percentage, games played, games started, steals, and minutes. This regression had an adjusted r-squared of 0.674 and the F-statistic reveals that the model is statistically significant. This study concludes that free-throw percentage and games played had the strongest effects on NBA players' salaries (Annieshieh). The next study reviewed was "Predicting NBA Salaries from the 2019 Offseason using Various NBA Statistics". This research paper attempts to predict the square root of salary from 4 variables including points, rebounds, assists, and turnovers. This model had an r-squared of 0.9728 and was significant at the one percent level. This regression found that assists and turnovers had the most impact on salaries of NBA players (Baum). The last regression project evaluated was "Determinants of NBA Player Salaries". This study attempted to predict salary from nine variables including field-goal percentage, three-point percentage, free-throw percentage, rebounds, assists, steals, blocks, fouls, and points per game. This regression had an adjusted r-squared of 0.613 and found that field goal percentage and points per game had the biggest impact on salaries of NBA players (Lyons).

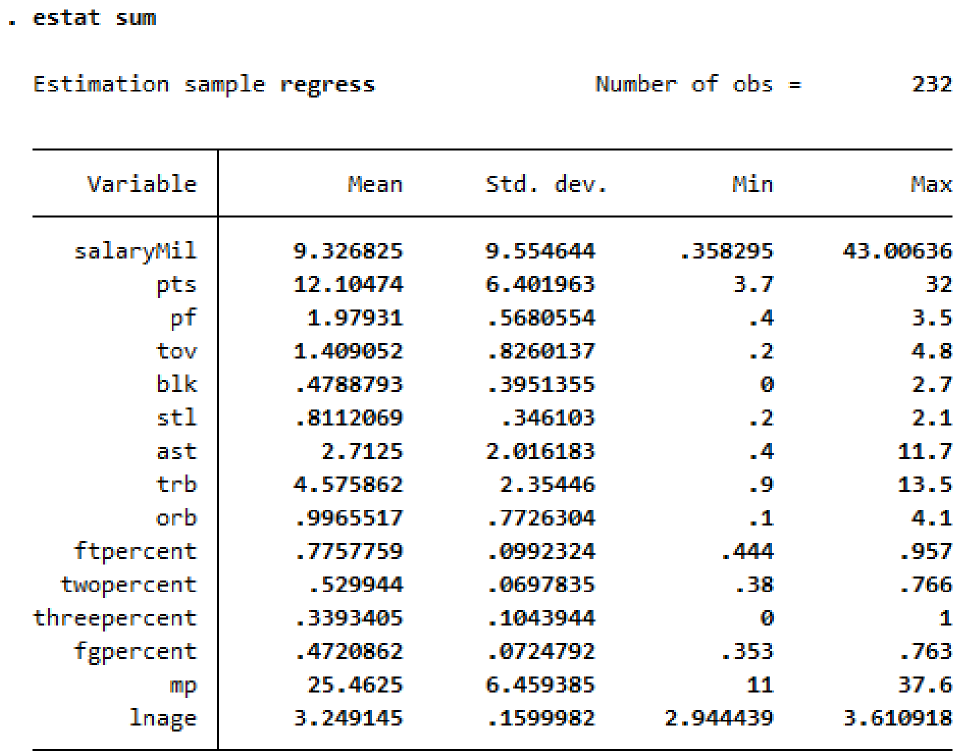
Description of my Model:

In this model, I am estimating the signs of the coefficients of points, blocks, steals, assists, total rebounds, offensive rebounds, free-throw percentage, two-point percentage, three-point percentage, field-goal percentage, minutes played, and age to be positive. On the other hand, I expect the coefficients of personal fouls and turnovers to be negative. I predict points and assists to have the highest impact on salary as I feel that these stats are highly coveted by NBA teams. All of the variables will be linear other than age which will be a natural log to account for diminishing returns on salary. The ideal set of variables for this problem will include statistics accounting for offense (points, field-goal percentage, etc.), defense (steals, blocks, etc.), and miscellaneous stats that account for usage/longevity (minutes played, age, etc.). Points, blocks, steals, assists, total rebounds, offensive rebounds, and minutes played will be the number per game. Free-throw percentage, two-point percentage, three-point percentage, field-goal percentage will be the percentage of the time that a player makes a certain type of shot. Lastly, age will be the natural log of number of years a player has been alive to compress the scale.

Description of my Data:

The data for this analysis was collected from basketballreference.com. This site is an extremely reliable source for up-to-data NBA statistics. All the variables listed above were collected for 235 NBA players who played more than ten minutes per game in the 2020-2021 season. This cutoff was chosen because this gives a broad range between superstar players making thirty-plus million and other players that are on minimum contracts. There was no dataset available with these specific NBA players’ salaries, so I manually copied and pasted them from their player page to my Stata dataset. This dataset will do a good job at helping estimate my model. There are a couple of features that could introduce simultaneity bias or multicollinearity. For example, total rebounds and offensive rebounds are connected to each other as if you get another offensive rebound, your total rebounds will increase. The same can be said for two-point percentage, three-point percentage, and field-goal percentage. If two-point or three-point percentage increase, field-goal percentage will do the same.

Table 1: Summary Statistics



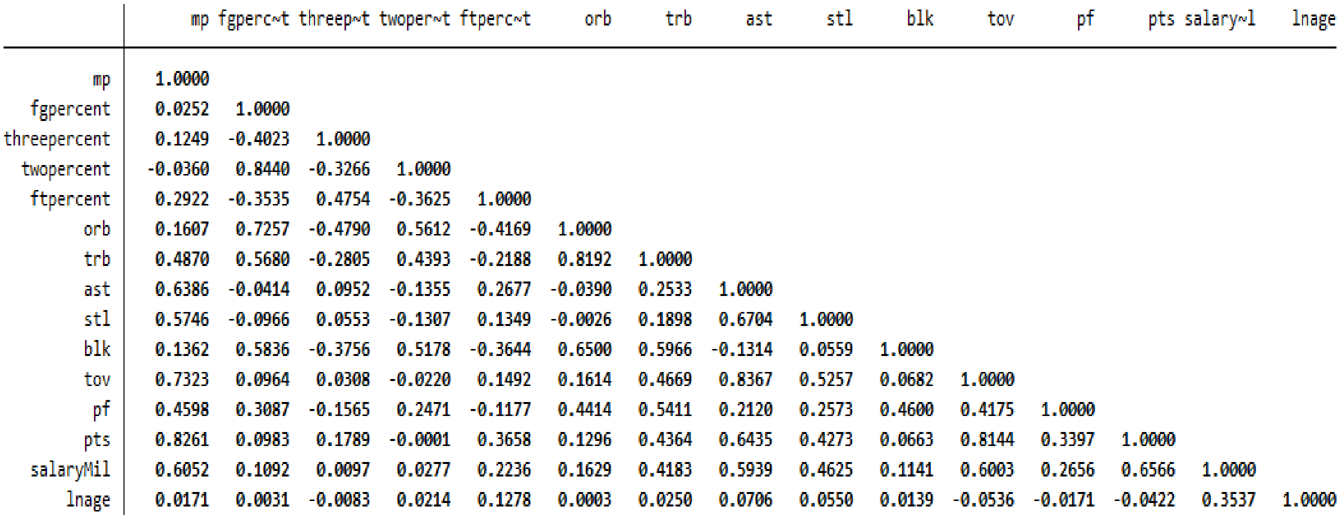
```
. estat sum

  Estimation sample regress              Number of obs =        232

  ---------------------------------------------------------------------
      Variable |         Mean    Std. dev.          Min          Max
  -------------+-------------------------------------------------------
     salaryMil |     9.326825    9.554644       .358295     43.00636
           pts |     12.10474    6.401963           3.7           32
            pf |      1.97931    .5680554            .4          3.5
           tov |     1.409052    .8260137            .2          4.8
           blk |     .4788793    .3951355             0          2.7
           stl |     .8112069     .346103            .2          2.1
           ast |       2.7125    2.016183            .4         11.7
           trb |     4.575862     2.35446            .9         13.5
           orb |     .9965517    .7726304            .1          4.1
     ftpercent |     .7757759    .0992324          .444         .957
    twopercent |      .529944    .0697835           .38         .766
  threepercent |     .3393405    .1043944             0            1
     fgpercent |     .4720862    .0724792          .353         .763
            mp |      25.4625    6.459385            11         37.6
         lnage |     3.249145    .1599982      2.944439     3.610918
  ---------------------------------------------------------------------
```

Some interesting features of this dataset include that the average salary (in millions) is 9.33 with a standard deviation of 9.55. The maximum salary was 43.01 and the minimum was 0.358. This illustrates that this variable is highly skewed right. This means that there are many more salaries that are respectively low than there are high salaries (such as 43.01). Many of our variable have the same skewedness such as points, turnovers, blocks, assists, and offensive rebounds. On the other hand, free-throw percentage is skewed left with most players being fairly good free throw shooters. Lastly, the data shows that somehow, a player in our sample shot 100 percent from three-point distance. This can be seen as an outlier as this is near impossible to do on high volume. This player is Drew Eubanks, and the second highest three-point percentage was 47.5 percent from Joe Harris. Drew Eubanks only shot two three-pointers all season and happened to hit both illustrating a small sample size (“2020-21 NBA Player Stats”).

Results and Analysis:

Before I run the regression on my model, I need to deal with the simultaneity bias or multicollinearity. I accomplished this by running a correlation matrix.

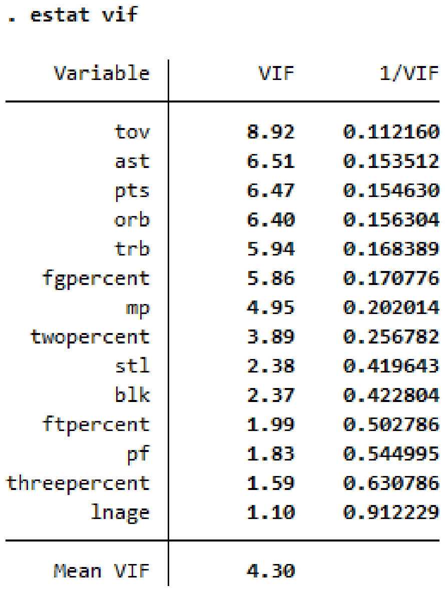
Table 2: Correlation Matrix



```
             |       mp fgperc~t threep~t twoper~t ftperc~t      orb      trb      ast      stl      blk      tov       pf      pts salary~l    lnage
-------------+--------------------------------------------------------------------------------------------------------------------------------------
          mp |   1.0000
  fgpercent |   0.0252   1.0000
threepercent |   0.1249  -0.4023   1.0000
  twopercent |  -0.0360   0.8440  -0.3266   1.0000
   ftpercent |   0.2922  -0.3535   0.4754  -0.3625   1.0000
         orb |   0.1607   0.7257  -0.4790   0.5612  -0.4169   1.0000
         trb |   0.4870   0.5680  -0.2805   0.4393  -0.2188   0.8192   1.0000
         ast |   0.6386  -0.0414   0.0952  -0.1355   0.2677  -0.0390   0.2533   1.0000
         stl |   0.5746  -0.0966   0.0553  -0.1307   0.1349  -0.0026   0.1898   0.6704   1.0000
         blk |   0.1362   0.5836  -0.3756   0.5178  -0.3644   0.6500   0.5966  -0.1314   0.0559   1.0000
         tov |   0.7323   0.0964   0.0308  -0.0220   0.1492   0.1614   0.4669   0.8367   0.5257   0.0682   1.0000
          pf |   0.4598   0.3087  -0.1565   0.2471  -0.1177   0.4414   0.5411   0.2120   0.2573   0.4600   0.4175   1.0000
         pts |   0.8261   0.0983   0.1789  -0.0001   0.3658   0.1296   0.4364   0.6435   0.4273   0.0663   0.8144   0.3397   1.0000
   salaryMil |   0.6052   0.1092   0.0097   0.0277   0.2236   0.1629   0.4183   0.5939   0.4625   0.1141   0.6003   0.2656   0.6566   1.0000
       lnage |   0.0171   0.0031  -0.0083   0.0214   0.1278   0.0003   0.0250   0.0706   0.0550   0.0139  -0.0536  -0.0171  -0.0422   0.3537   1.0000
```

We know that a high correlation typically leads to problems so we will be using a threshold of 0.8 as too high. Immediately when looking at this matrix, I see that offensive rebounds' correlation to total rebounds is too high at 0.8192. When looking deeper, I see that two-point percentage is highly correlated with field-goal percentage with a measure of 0.8440. Interestingly, assists and turnovers have a correlation of 0.8367 and points and minutes played have one of 0.8261. In order to further investigate this issue, I will examine the variance inflation factor (VIF). Typically, the closer to 10+ you get, the more problematic a variable can be seen. I have also read that some people consider a VIF of 5 to be moderately problematic (Choueiry).

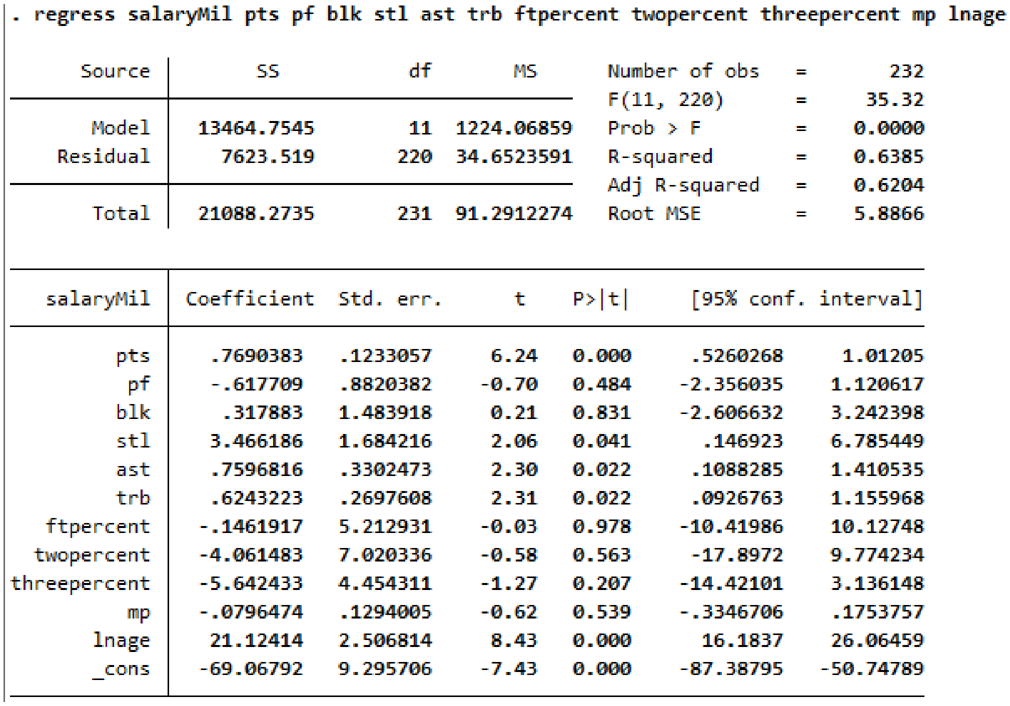
Table 3: VIF Table



```
. estat vif

    Variable |       VIF       1/VIF  
-------------+----------------------
         tov |      8.92    0.112160
         ast |      6.51    0.153512
         pts |      6.47    0.154630
         orb |      6.40    0.156304
         trb |      5.94    0.168389
   fgpercent |      5.86    0.170776
          mp |      4.95    0.202014
  twopercent |      3.89    0.256782
         stl |      2.38    0.419643
         blk |      2.37    0.422804
   ftpercent |      1.99    0.502786
          pf |      1.83    0.544995
threepercent |      1.59    0.630786
       lnage |      1.10    0.912229
-------------+----------------------
    Mean VIF |      4.30
```

All of our variables have a VIF of less than 10 so I will be observing all the ones that had a problem with correlation: offensive rebounds, total rebounds, two-point percentage, field-goal percentage, assists, and turnovers. When looking at these, one can see respective VIFs of 6.4, 5.94, 3.89, 5.86, 6.51, 8.92. To find a remedy, redundant variables are further evaluated. Two-point percentage, three-point percentage, and field-goal percentage are redundant as two- and three-point percentage helps calculate field-goal percentage. To solve this, field-goal percentage will be omitted from my regression. Assists and turnovers are redundant as well because if you have high assists (passes to a made shot), you typically have the ball in your hands a lot. This will lead to more turnovers. The same can be said for offensive and total rebounds because if you get an offensive rebound, your total rebounds will increase. Turnovers and offensive rebounds will be excluded from my regression to combat these issues. Now that I have finalized my list of variables, it is time to run the actual regression.

Table 4: Regression Analysis



```
. regress salaryMil pts pf blk stl ast trb ftpercent twopercent threepercent mp lnage

      Source |       SS           df       MS      Number of obs   =       232
-------------+----------------------------------   F(11, 220)      =     35.32
       Model |  13464.7545        11  1224.06859   Prob > F        =    0.0000
    Residual |    7623.519       220  34.6523591   R-squared       =    0.6385
-------------+----------------------------------   Adj R-squared   =    0.6204
       Total |  21088.2735       231  91.2912274   Root MSE        =    5.8866

------------------------------------------------------------------------------
   salaryMil | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         pts |   .7690383   .1233057     6.24   0.000     .5260268     1.01205
          pf |   -.617709   .8820382    -0.70   0.484    -2.356035    1.120617
         blk |    .317883   1.483918     0.21   0.831    -2.606632    3.242398
         stl |   3.466186   1.684216     2.06   0.041      .146923    6.785449
         ast |   .7596816   .3302473     2.30   0.022     .1088285    1.410535
         trb |   .6243223   .2697608     2.31   0.022     .0926763    1.155968
   ftpercent |  -.1461917   5.212931    -0.03   0.978    -10.41986    10.12748
  twopercent |  -4.061483   7.020336    -0.58   0.563     -17.8972    9.774234
threepercent |  -5.642433   4.454311    -1.27   0.207    -14.42101    3.136148
          mp |  -.0796474   .1294005    -0.62   0.539    -.3346706    .1753757
       lnage |   21.12414   2.506814     8.43   0.000      16.1837    26.06459
       _cons |  -69.06792   9.295706    -7.43   0.000    -87.38795   -50.74789
------------------------------------------------------------------------------
```

$$
\begin{aligned}
salaryMil = \ & -69.068 + 0.769(pts) - 0.618(pf) + 0.318(blk) + 3.466(stl) + 0.760(ast) \\
& + 0.624(trb) - 0.146(ftpercent) - 4.061(twopercent) \\
& - 5.642(threepercent) - 0.080(mp) + 21.124(lnage)
\end{aligned}
$$

Looking at the F-statistic, one can see that this model passes the hypothesis test and is statistically significant. Five of the variables pass the t-test and are statistically significant. Unfortunately, personal fouls, blocks, free-throw percentage, two-point percentage, three-point percentage, and minutes played failed the t-test and are problematic. The adjusted r-squared of this model is 0.6204 meaning that it can account for 62.04 percent of variation in salary (in millions) from these variables. The coefficients reveal that the five most impactful statistics on salary (listed most impactful to least) include age, steals, two-point percentage, three-point percentage, and points. The coefficients of two- and three-point percentage are misleading however, since the results are listed as a decimal.

Summary and Conclusion:

To summarize, the question this research paper is attempting to answer is can NBA players' salaries be predicted from various statistics. This question is very interesting because the

NBA is often referred to as an analytics driven league and I feel that there must be some sort of data analysis when offering players contracts. To answer this question, I used a semi-log regression of eleven variables that are considered important to predict salary (in millions). The results of this regression show that as a whole, the model was statistically significant at the 0.05 level. Five of the eleven variables were statistically significant according to a t-test at the same significance level. The model concluded that the five most impactful statistics on salary include age, steals, two-point percentage, three-point percentage, and points. While there are some commonalities between my findings and those of other literature, they are not all the same or even similar in some cases. The key takeaway from this is that the statistics that are most impactful on salary varies from dataset to dataset. Further research is needed to conclude what the most impactful statistics on NBA player's salaries are. However, the answer to the title of this project is yes – to an extent. Data can be extremely helpful at estimating salaries of NBA players, but this estimate is not completely reliable.

Work Cited

"2020-21 NBA Player Stats: Per Game." *Basketball*, www.basketball-reference.com/leagues/NBA_2021_per_game.html.

Annieshieh. "NBA Player Salaries Prediction with Linear Regression." *Medium*, Analytics Vidhya, 28 Feb. 2020, medium.com/analytics-vidhya/nba-player-salaries-prediction-with-linear-regression-2b90280ff4e8.

Baum, Joe. *Predicting NBA Salaries from the 2019 Offseason Using Various NBA Satatistics*. Brockport, 6 Dec. 2019, brockport.edu/academics/mathematics/directory/docs/baum_joe.pdf.

Choueiry, George. "What Is an Acceptable Value for VIF? (with References)." *Quantifying Health*, 2021, quantifyinghealth.com/vif-threshold/.

Lyons, Robert. "Determinants of NBA Player Salaries." *The Sport Journal*, 13 July 2018, thesportjournal.org/article/determinants-of-nba-player-salaries/.