Breaking Down the Numbers: How NBA Player Stats Impact Salaries

Table of Contents

Breaking Down the Numbers: How NBA Player Stats Impact Salaries

Introduction

The purpose of this study is to investigate which NBA basketball player game stats affect the player's salary. Despite being a huge fan of basketball, one thing I never understood was how an NBA player's pay works, does it all depend on their skills? Several NBA players make millions of dollars in salary pay but others earn less than a million dollars. This study is meant to be informative to help others like myself who have this query to establish a link between an NBA players' on-court game stats and their salary to answer this question.

When searching online for related studies there were a handful of results in the form of a previously published study or a thesis paper. They looked at an NBA player's salary through different perspectives like determining if there was any pay discrimination based on a player's race or colour (Yang, & Lin, H.-Y., 2012), if NBA players were being overpaid than how much they are truly worth (Li, N., 2014), and another study that looked into the relevant variables that make up a player's pay (Lyons Jr, Jackson Jr, E. N., & Livingston, A., 2015). While the first two papers could be used to explain skews and potential biases in the data, they do not answer the question. However, the third article seems identical to this study but the difference in this study is that it looks at the pay structure from another angle using salary information from the current season and game stats from the prior season. Additionally, this study would be more relevant now than the other studies that were published more than five years ago as we all can agree there are many things that could change over a year in sports.

Methods

We use linear regression to answer the research question as this technique would allow us to see if there is any link between the two variables: a player's salary (response variable) and their game stats (predictor variables). Using linear regression, we would like to come up with a linear model that could best explain this link. In R, we would start by removing any observations that contains missing information as this could become problematic when performing calculations. An EDA should be performed next by creating histograms and boxplots on the data to see if there are any issues like skews or outliers in the variables that we may have trouble with later. We would then split our data into two equal parts, a training set, and a test set, through random sampling without replacement which would be used to validate our model in the end. We will investigate the two sets by looking at their respective mean and standard deviations to ensure they are not too different from each other so if we have issues with our conclusions this may potentially be a cause of it.

We proceed with using the training set to create a linear model this could be constructed using salary as our response variable and the player stats as the predictors in our model. Once this is created, its important to ensure two of our model conditions and four of our model assumptions hold. This is an essential step to take when building or modifying our model to guarantee the correct interpretation of the results from our statistical analysis, t-test. The first condition is valid if we create a response vs. fitted plot and see the presence of a simple function and/or scatters that follow a diagonal. If this does not hold, then linear regression would not be

appropriate for this study, this same conclusion applies when the second condition fails. To validate the second condition, we look at the pairwise plots between each predictor and look for a relatively linear relationship, any minor issue may be corrected later, but any obvious non-linear relationships would make this condition invalid. To satisfy all four model assumptions we create three plots: residuals vs. fitted plot, residuals vs predictors plot(s), and a normal q-q plot for the model and on the first two plots we want to see no systematic patterns, or obvious signs of separation of points between the residuals, and no fanning patterns either. The q-q plot is used to check for normality and the scatters should relatively follow a diagonal, the q-q line, for it to hold. If there is an issue with any of these then we need to apply a transformation on the response and/or predictor variable(s) so our assumptions are satisfied. Typical transformations involve taking the natural log or square root of the response and/or predictor(s), but to make an informed decision we use Box-Cox to narrow down the list of variables you would want to apply a transformation on and which ones. Note that the same transformations applied here would be applied to the data in the test set.

Once the model conditions and assumptions hold then can we proceed with using this model for analysis. We would ideally prefer to use a model that has limited to no multicollinearity present to ensure our variances are not too inflated, any VIF > 5 would be considered severe. If the variable(s) violating this condition are not essential to this study or were created by accident, then we would remove such variable(s) either individually or at once; Otherwise, we would acknowledge its existence and discuss it as it would be a limitation. If any adjustments were made to the linear model, it is essential to check the model conditions and assumptions before moving forward. Now, we want to identify any problematic observations in our model by calculating the number of leverage points, outliers, and influential points that may be present using appropriate cut-offs for each calculation. We would not delete any problematic observations but rather discuss these under limitations. Finally, we would like to use t-tests to compute p-values and generate a statistical summary of significant variables with their respective coefficients. Additionally, we would make note of the R^2 value, adjusted coefficient of determination (R^2_{adj}) value, and calculate the corrected Akaike's information criterion (AICc), and the Bayesian Information Criterion (BIC) for model selection purposes. This is just one possible model for this study, we could try to create better model(s) by reducing the number of predictors by removing any/or all non-significant predictors (if applicable) re-checking all model conditions and assumptions when doing so, and by comparing the $R^2_{adj,}$ AICc, and BIC values for each. We would like to conclude with a single model that has a relatively high R^2_{adj} value and relatively low AICc and BIC values with one another.

After selecting a model, we would validate it so our result could be made generalizable to the population rather than exclusively to the sample. To do this we want to see the coefficients from the summary of the t-test from the training set to lie within 2 standard deviations of the coefficients from the test set while ensuring the same number of predictors are significant in both sets. We would still validate the model if there were minor issues present only if they could be explained by issues from within the data like having completely different sets from splitting the data, leverage points, and/or influential observations. If the model could not be validated, then the results would not be generalizable outside of the sample population. We conclude with the

validated model as the appropriate model to answer the question in the study with the significant parameter(s) determined by the results of the t-test being the answer(s) to this study.

Results

Through an EDA of our data on 186 NBA player salary from the 2021-22 season, provided by ESPN, and player stats, provided by basketball-reference.com, from the 2020-21 season the histograms shown in Figure 1 reveals right-skews in salary, experience, rebounds, and assists, and left-skews for three-point % and free-throw %. The rest seem approximately normal. The skews may influence our results if they are deemed problematic when checking our model assumptions.

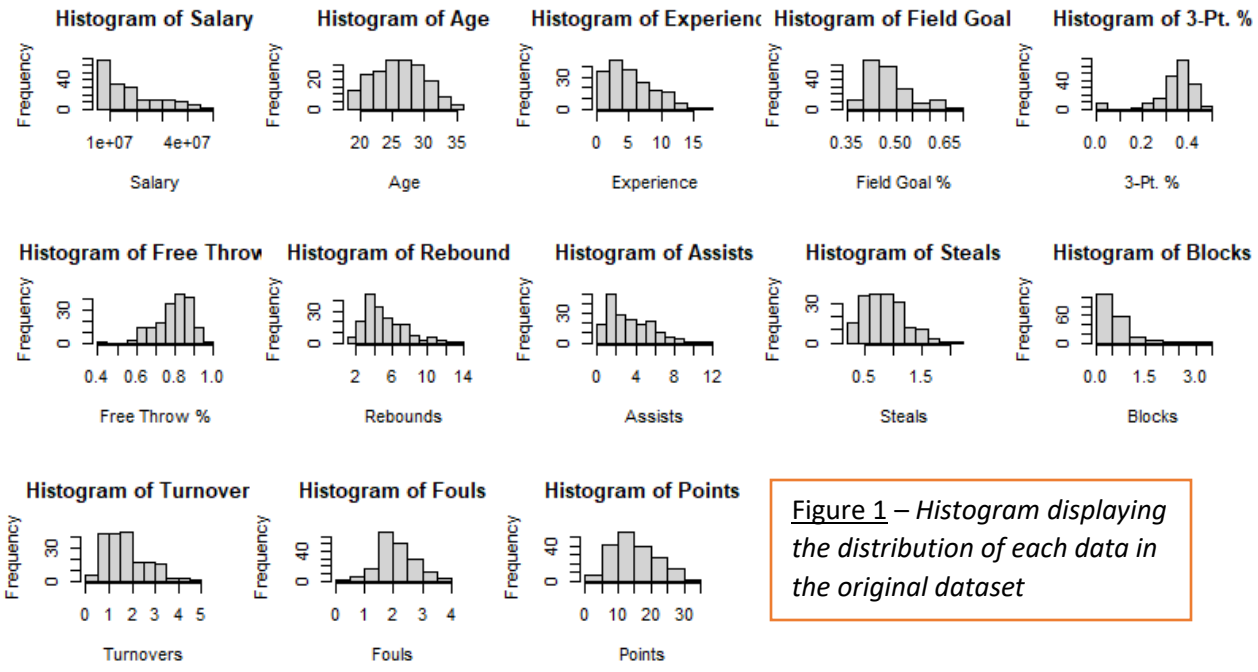

Figure 1 – *Histogram displaying the distribution of each data in the original dataset*

The boxplots in Figure 2 reveal several outliers in our data where field-goal % and average blocks per game contains the most outliers, these may explain discrepancies when validating our model if they are influential.

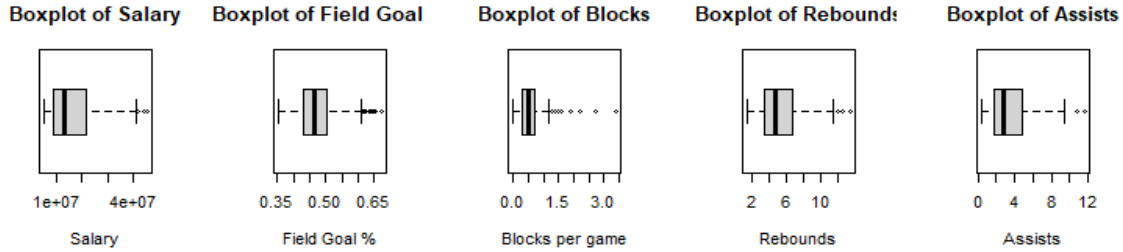

Figure 2 – *Boxplot displaying variability and outliers for select variables in the original dataset*

We observe the following summary statistics in Table 1 after splitting the original dataset equally into a training and test dataset and conclude both sets are not too different from each other as the means in both sets are within 2-standard deviations of each other.

Table 1 - *Summary statistics in training and test dataset, each of size 93*

Variable	mean (s.d.) in training	mean (s.d.) in test
Salary $	16124825 (10117625)	17384681 (10703429)
Age	25.903 (3.663)	27.151 (4.102)
Experience	4.978 (3.131)	6.247 (3.847)
Field Goal %	0.479 (0.064)	0.471 (0.064)
3-Pt. %	0.346 (0.091)	0.34 (0.096)
Free-throw %	0.784 (0.091)	0.809 (0.086)
Rebounds	5.246 (2.211)	5.213 (2.556)
Assists	3.028 (2.044)	3.778 (2.379)
Steals	0.886 (0.346)	0.901 (0.368)
Blocks	0.605 (0.412)	0.534 (0.51)
Turnovers	1.665 (0.857)	1.804 (0.919)
Fouls	2.209 (0.571)	2.052 (0.626)
Points	14.786 (6.725)	15.167 (6.174)

A full linear model was built with the model conditions verified but there were issues with the assumptions that would require transformations to fix them. In Figure 3 the Residual vs. Fitted plot, an obvious fanning pattern was present which meant that we may have a constant variance assumption violation with the response variable, salary. The Residual vs. Predictor plots showed an obvious non-linear pattern for the points variable and a minor upward curve for the assists variable which both could indicate a linearity violation. Normal QQ-Plot showed no issues with normality.

Using the suggestions from Box-Cox (see Appendix 1) we applied a natural log transformation to salary, rebounds, assists, turnovers, and fouls and took the square root of experience and points per game then verified that the model conditions and assumptions hold. Then we tested for multicollinearity (see Appendix 2) and found the following variables that had severe multicollinearity issues present in order of severity: turnovers, age, experience, and assists.

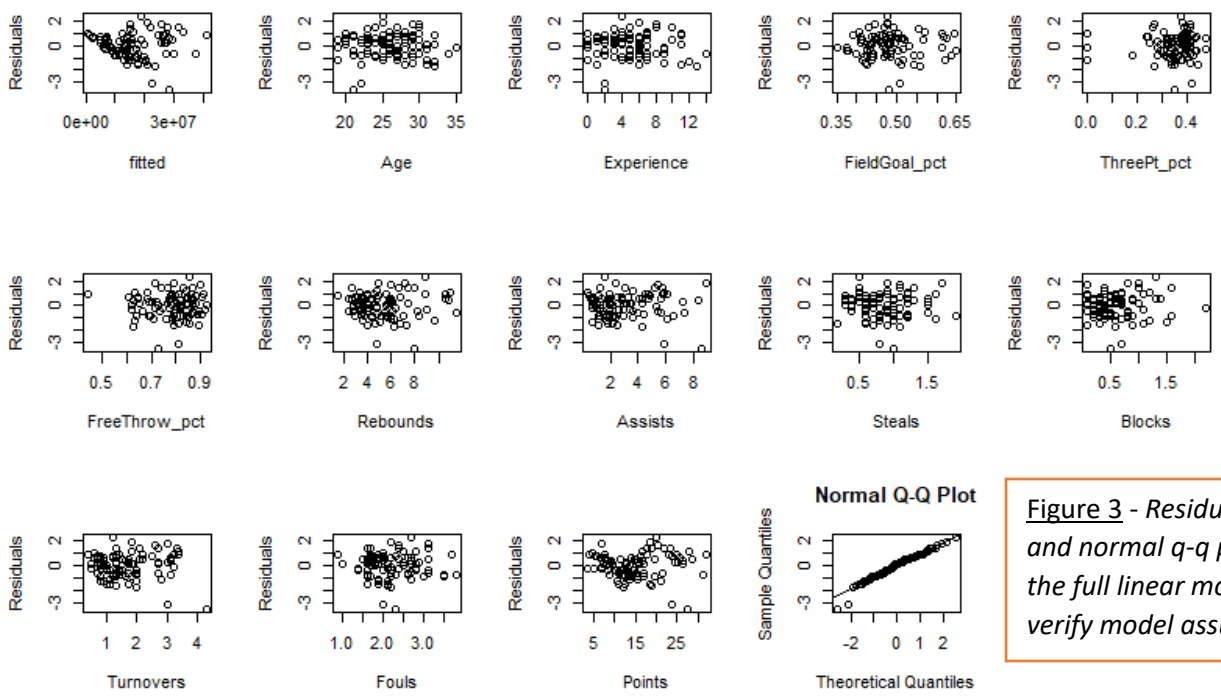

Figure 3 - *Residual plots and normal q-q plot for the full linear model to verify model assumptions*

Age and experience cannot co-exist, same with turnovers and assists. To resolve this, age was removed from the model followed by assists and turnover but removed separately as we managed to get models with VIF<5 when removing one of them, so we managed to create two test models model 3 and model 2 respectively. A third model was formed by removing all four of the multicollinear variables to test to see which one of these three models would be best. The conditions and assumptions of the three models hold. A model that best explains this study is to be determined after comparing results from a t-test, R^2_{adj}, AICc, and BIC values between models, and by determining if the model is valid. A summary of these calculations is listed in Table 2.

After comparing the values between the three models in the test and training sets from Table 2, model 3 is selected as the most appropriate linear model to answer the question. Model 2 fails to be valid as there is a lot of discrepancy between the coefficients of the training and test set as highlighted in Table 2. Despite model 3 having minor discrepancies it is still a valid model because the issues could be explained by the leverage points and/or placement of the influential points in its dataset. The same explanation holds for why there is an extra significant predictor in all three of the test models than in their respective training sets. Of the two valid models, models 1 and 3, model 3 has a high R^2_{adj} value, low AICc and BIC value hence why it is the best model to answer the question in study. Model 3 is the linear model that contains the following data: salary, field goal %, 3-pt %, free-throw %, steals, blocks, experience, rebounds, assists, fouls, points per game, where rebounds, experience and points per game are the three significant predictors in this model.

Table 2 -

Summary of characteristics of three candidate models in the training and test datasets. Model 1 uses Field Goal %, Three-point %, Free throw %, Steals, Blocks, Rebounds, Fouls, and Points per game as predictors, while Model 2 uses all predictors from Model 1 but adds Experience and Assists into its model, Model 3 add Experience and Turnovers to Model 1. Response is log(Salary) in all three models. Coefficients are presented as estimate ± 2SE (= significant t-test at $\alpha = 0.05$). Issues are highlighted.*

Characteristic	Model 1 (Train)	Model 1 (Test)	Model 2 (Train)	Model 2 (Test)	Model 3 (Train)	Model 3 (Test)
Largest VIF	2.4080479	2.3623475	3.106858	3.2977183	3.6624552	4.11643
# Cook's D	0	0	0	0	0	0
# DFFITS	7	7	7	5	8	6
# DFBETA	53	69	77	72	75	70
# Leverage	6	8	5	5	6	5
# Outliers	5	5	5	5	5	5
Violations	none	none	none	none	none	none
R^2	0.562	0.48	0.696	0.584	0.701	0.583
Adjusted R^2	0.52	0.431	0.659	0.533	0.664	0.532
AICc	115.9963098	128.0305317	87.2266728	112.5437496	85.746597	112.7278899
BIC	138.7032571	150.737479	113.8129886	139.1300655	112.3329129	139.3142058
Intercept	14.265 ± 1.314 (*)	14.626 ± 1.63 (*)	14.747 ± 1.142 (*)	14.43 ± 1.492 (*)	14.827 ± 1.13(*)	14.405 ± 1.488(*)
Field Goal %	-0.334 ± 1.726	-1.414 ± 2.11	-0.59 ± 1.464	-0.969 ± 1.928	-0.472 ± 1.468	-0.967 ± 1.956
3-Pt %	-0.672 ± 1.274	-0.678 ± 1.224	-0.805 ± 1.076	-0.323 ± 1.122	-0.704 ± 1.076	-0.323 ± 1.132
Free-throw %	0.635 ± 1.226	0.803 ± 1.382	-0.238 ± 1.074	0.238 ± 1.276	-0.157 ± 1.07	0.28 ± 1.3
Steals	0.204 ± 0.28	0.284 ± 0.294	0.064 ± 0.314	0.184 ± 0.308	0.025 ± 0.278	0.213 ± 0.274
Blocks	-0.088 ± 0.292	0.191 ± 0.22	-0.021 ± 0.26	0.173 ± 0.214	-0.009 ± 0.25	0.157 ± 0.202
Experience	-	-	0.303 ± 0.102 (*)	0.226 ± 0.106 (*)	0.313 ± 0.102 (*)	0.232 ± 0.104 (*)
Rebounds	0.145 ± 0.33	0.396 ± 0.3 (*)	-0.11 ± 0.292	0.338 ± 0.274 (*)	-0.137 ± 0.294	0.338 ± 0.276 (*)
Assists	-	-	0.048 ± 0.178	0.054 ± 0.224	-	-
Turnovers	-	-	-	-	0.169 ± 0.268	0.042 ± 0.318
Fouls	-0.031 ± 0.402	-0.313 ± 0.322	0.191 ± 0.348	-0.229 ± 0.294	0.137 ± 0.358	-0.238 ± 0.294
Points per Game	0.456 ± 0.14 (*)	0.357 ± 0.146 (*)	0.46 ± 0.126 (*)	0.321 ± 0.166 (*)	0.416 ± 0.148 (*)	0.323 ± 0.212 (*)

Discussions

From our results, model 3 proved to be the appropriate linear model to answer the study question. We note from Table 2 that rebounds, experience and points per game are all significant predictors in this model. We can generalize this conclusion to the population of every player in the NBA outside of the sample of 186 players because the model was validated. All three player stats: average rebounds, experience, and average points per game contain positive coefficients (0.338, 0.232, 0.323 respectively) so a unit increase to any of these stats would have a positive effect on a player's pay while holding the other player stats constant.

Since we had transformed salary by its natural log an example of how we can interpret the numbers would be as follows: for a one-year increase in experience, it is estimated your salary would increase by approximately 100 x ($e^{(0.232)}$-1) = 26.11 percent. Similar conclusions could be made for the other two by substituting their coefficients. A lot of things change over the years in sports, from a player not being able to play because of an injury or a player not giving their best performance during a basketball season, these unexpected variabilities could be classified as influential points as we found during our analysis, and these could skew the results and cause issues during model validation and interpretation. The limitation to this study is it could only prove to be valid for some time given the unpredictable nature of the data. In summary, the results highlight an NBA player's salary is partially determined by the rebounds a player makes, their average points per game, and their experience.

References

Li, N. (2014). *The determinants of the salary in NBA and the overpayment in the year of signing a new contract* (Order No. 1582938). Available from ProQuest Dissertations & Theses Global. (1654415245). Retrieved from http://myaccess.library.utoronto.ca/login?qurl=https%3A%2F%2Fwww.proquest.com%2Fdissertations-theses%2Fdeterminants-salary-nba-overpayment-year-signing%2Fdocview%2F1654415245%2Fse-2%3Faccountid%3D14771

Lyons Jr, Jackson Jr, E. N., & Livingston, A. (2015). Determinants of NBA Player Salaries. The Sport Journal. https://doi.org/10.17682/sportjournal/2015.019

Yang, & Lin, H.-Y. (2012). Is There Salary Discrimination by Nationality in the NBA?: Foreign Talent or Foreign Market. Journal of Sports Economics, 13(1), 53–75. https://doi.org/10.1177/1527002510391617

2021-22 NBA Player Contracts. (n.d.). Basketball-Reference.Com., from https://www.basketball-reference.com/contracts/players.html

NBA Player Salaries - National Basketball Association - ESPN. (n.d.). ESPN.Com. from http://www.espn.com/nba/salaries

Appendix

Appendix 1 –

Table with Box-Cox results on the full linear model of the original dataset displaying the variables that could benefit from a transformation

	Rounded Power	Suggested Transformation
Salary	0.00	Natural Log
Age	1.00	
Experience	0.50	Square Root
Field Goal %	-1.00	
3-Pt. %	5.53	
Free-throw %	2.00	
Rebounds	0.00	Natural Log
Assists	0.00	Natural Log
Steals	1.00	
Blocks	0.33	
Turnovers	0.00	Natural Log
Fouls	0.00	Natural Log
Points	0.50	Square Root

Appendix 2 –

Table showing multicollinearity between variables after transformations have been applied

	VIF	VIF > 5
Age	6.500580	Y
Experience	6.264752	Y
Field Goal %	1.637160	
3-Pt. %	1.874649	
Free-throw %	1.965901	
Rebounds	2.813682	
Assists	6.254482	Y
Steals	2.164830	
Blocks	2.340856	
Turnovers	7.414420	Y
Fouls	1.730853	
Points	3.547424	