The link between statistics and Salary in sports

Branton Parsons

brantonparsons@oakland.edu

Oakland University

ECN 4050/5050: Econometrics

November 11, 2024

Abstract

The point of this paper is to analyze the relationship between player salary and the statistic player put up during the 2023-24 season. Using econometric tools such as linear regression modeling, t-statistics, F-statistics, p-values, Sum of Squared Errors (SSE), Sum of Squared Residuals (SSR), Total Sum of Squares (SST), as well as tests for collinearity and heteroskedasticity, this analysis identifies the key factors that should be used to determine a player’s salary. The insights from these results could provide valuable guidance for economists and NBA teams to optimize salary allocations in the current market.

# An increased emphasis on Analytics in sports

The NBA, one of the premier professional sports leagues, provides an opportunity to analyze how statistical output and salary are correlated. In a league that emphasizes statistical output that translates into salary, an understanding of how the two are related is essential. Variables such as points, rebounds, shooting percentage, assists, are central factors that determine a player’s value. This paper seeks to explore and analyze the relationship between player salary and the 2023-24 season statistics by employing different types of economic tools we have learned through this course.

Some of the primary questions that are answered along the way include: what statistics may hold the most weight in determining an NBA salary? How accurate can these models predict salary? And how can these models be refined to improve salary predictions? Linear regression serves as the foundation of this analysis creating a baseline for calculating key statistics like the t-statistic, F-statistic, p-statistic, Heteroskedasticity, and collinearity. These tools will be useful for refining and determining the reliability of the model.

The paper will begin with a review of past analysis on player salary, followed by discussion of the methods, models, and data that were used, along with the results of these findings and their takeaways.

# Key Terms review

Age = Age of a Player

Position = Position the play plays

GamesPlayed = How many games a player played in that year (out of 82)

MinutesPlayed = How many minutes that player averaged per game

FGM = Field goals made, how many shots a player made on average in a game

FGA = Field goals attempted, how many shots a player shot on average in a game

FGP = FGP is calculated as the ratio of Field Goals Made (FGM) to Field Goals Attempted (FGA), representing a player’s shooting efficiency.

REB = Average amount of times a player gets the ball after a missed shot.

AST = Average amount of times a player passes the ball to someone else who makes a shot per game

STL = Average amount of times a player steals the ball from an opponent per game.

BLK = Average amount of times a player Blocks a shot attempt from an opponent per game.

PTS = Average amount of points a player scores per game.

# The past, the present, and the future

An evolution of player salaries has been the status quo for the entire time the league has existed, with an average increase of 8% since the 1990-91 season (Curcic, 2024). This growth has dropped during different time periods shows a nonlinear relationship in which other economic and league specific factors could cause fluctuations. As an example, growth between 2002 and 2012 sat at 3% compared to the 7% increase between 2012 and 2022 (Curcic, 2024). These shifts underline that salary growth could be sensitive to the league and economic factors like a change in policy. In 2011, for example, owners proposed a higher luxury tax and salary cap while reducing the players' share of NBA-related revenue. These changes had the potential to impact on player salaries. (Lockhart, 2022)

In previous studies examining the relationship between NBA salaries and player performance, some key statistical indicators were identified. For instance, (Yang, 2024) gives six primary variables, Free Throw, 3 Points Attempts, 2 Points Attempts, Blocks Per Game, Assists and Win Shares for analysis. Similarly, a study by (William Wu, 2017), *Classification of NBA Salaries through Player Statistics,* highlights defensive rebounds, turnovers, points, and value over replacement player (VOIP) for salary. Both studies employed linear regression, collinearity and statistical tests like t-statistics and p-statistics to validate the strength of their findings.

However, both Yang’s and Wu’s research cautioned that while their models revealed correlation between statistics and salary, they could not be used as definitive tools for salary determination. This limitation exposes complexities when relating player performance to salary due to league complexities with marketability and team dynamics playing substantial roles.

The present study plans to build on the fundamentals of each report by analyzing the 2023-24 season, by incorporating economic tools used with a fresh data set to refine the linear regression model approach. This paper seeks to advance our understanding of the factors most reliable in predicting NBA salaries, offering an up-to-date perspective on the evolving dynamic of player salary.

# Data Gathering & The Model

To begin this project, data on both statistics and salary was needed and gathered from two main sources: Basketball Reference and HoopsHype. Basketball Reference is a comprehensive database tracking the history of basketball with official record dating back to the 1946-47 seasons. And before the ABA and NBA merged into one in 1976 (Britannica, 2024). HoopsHype is a site that has data about NBA salaries for each player which could be useful when trying to find the correlation between statistics and salary.

A close up of text

Description automatically generated The first model, referred to as Basis\_1, was developed using Ordinary Least Squares (OLS) regression. This model used eleven variables, Points (PTS), Assists (AST), Rebounds (REB), Steals (STL), Blocks (BLK), Age, Games Played, Minutes Played, Field Goals Made (FGM), Field Goals Attempted (FGA), and Field Goal Percentage (FGP). After running the regression, the resulting linear model is as below: A screenshot of a graph

Description automatically generated

The model shows certain variables, such as points, assists, steals, blocks, age, games played, and minutes played are statistically significant at a 5% level. The positive coefficients also suggest that an increase in each variable is associated with a higher predicted salary. Points (PTS) had a coefficient of .14 indicated that each point per game a player has there is a 14% increase in salary all else equal. This relationship also works with age and Games Played with .08 and .02 respectively. This suggests that experience and players that are older tend to have higher salaries.

However, the model identifies some unexpected relationships. Rebounds (REB), Field goals made (FGM), and Field goals attempted (FGA) contain negative coefficients, suggesting that as these statistics increase salary would be associated with a decrease in salary. This is cause for some concern as there may be some unexpected relationships which could imply collinearity or other model limitations.

The Models overall fit, R-squared and adjusted R-square, are .6798 and 6792 respectively, suggesting around 67% of the variability in salary is explained by the Basis\_1 model. While this is a high R-squared value, the presence of both negative and positive coefficients for some variables leads to further scrutiny.

Two diagnostics were used to address issues in the Basis\_1 model:

**Collinearity**

With a high correlation between PTS and FGM at .99, FGM and FGA at .98, MiniutesPlayed and GamesPlayed at .69, and MiniutesPlayed and FGM at .89. These tests indicated potential multicollinearity and suggested some variables should be removed or combined to create a more complete model.

**Heteroskedasticity**A diagram of a graph

Description automatically generated with medium confidence

Initial tests for heteroskedasticity, indicated from a > shaped pattern in the residual plot, show the variability of errors changes across different levels of the independent variables. The Breusch-Pagan test backs the graph up with a p-value of 1.825e-14 suggesting the model may be unreliable due to non-constant variance in the residuals. This heteroskedasticity impacts the reliability of our coefficient estimation and significant tests, warranting model refinement.

While the Basis\_1 model provides valuable insight into the variables of NBA player salaries, these findings should be pursued with caution. The positive relationships for key variables support the idea that players who contribute statistically more tend to earn higher salaries. However, the model has its limitations. Collinearity, heteroskedasticity, and unexpected negative coefficients cause concern in accuracy of the model and if it can reliably predict salary.

# Interpreting results

Results from the regression analysis indicate there are variables that correlate with salary, points, age, games played, and minutes played being just a couple that showed positive coefficients. Other variables like rebounds, field goals made, and field goals attempted have a negative coefficient. There could be a variety of reasons each has its value, but we can try and explain why they might be. Points, for example, are a direct indicator of a player's ability to contribute to the team's success. This gives the reason that a player who scores more points per game has more value and should be compensated for due to that. Age, another key variable, can be correlated to experience and years in the NBA.

The positive coefficient could be explained by if you have more experience, you are less likely to have inconsistent performance in games or that you provide value in different ways to your team. As for minutes and games played, we can explain it by the more minutes you play in an NBA game and the more games you play, the team or organization believes that you provide value on the court. Another reason could be the amount of time you are on court correlated to how many statistics you put up. For instance, a player who plays 10 minutes per game may score only 8 points and make 2 assists, while a player who plays 30 minutes might achieve 30 points, 8 assists, and 6 rebounds. This difference could be one explanation as to why they have a positive coefficient.

However, the negative coefficients for Rebounds, field Goals Made, and Field Goals Attempted are more challenging to explain. One explanation could be multicollinearity, where these variables are highly correlated with each other, distorting their individual effects. The detection of collinearity and heteroskedasticity during diagnostic tests highlights the need for further refinement of the model. By combining games played and minutes played along with points, FGM, and FGA to get shooting efficiency, we could better predict salary.

# Conclusion

In this analysis, we developed a model that explored the relationship between statistics and player salary. Our results suggest that statistical values like points, age, games played, and minutes played have a positive correlation with salary. This positive correlation makes sense with players who contribute on the court, through experience and scoring, provide more value and are typically compensated more. On the other hand, the negative correlation of variables like rebounds, Field Goals Made, and Field Goals Attempted remain less clear and cause deep dive into the model itself.

Despite the model's R-squared value of 0.68, tests performed on the model revealed other issues such as multicollinearity and heteroskedasticity, which could distort results. Given these concerns, it is suggested that the model could be refined by combining certain variables, such as games played, and minutes played or by incorporating alternative statistical measurements like shooting efficiency to better capture the value of a player.

While this model provides important insight into the factors that influence NBA salary, it is not fully conclusive. Future models could build on this regression, addressing the limitations identified to improve the accuracy and validity of the result.

# References

*Basketball Stats and History*. (2024, November 01). Retrieved from Basketball Reference: https://www.basketball-reference.com/

*Britannica*. (2024, November 07). Retrieved from American Basketball Association: https://www.britannica.com/topic/National-Basketball-Association

Curcic, D. (2024, March 25). *NBA salaries analysis (1991-2022)*. Retrieved from RunRepeat: https://runrepeat.com/salary-analysis-in-the-nba-1991-2019

Lockhart, M. (2022, November 08). *NBA lockouts: Why did they happen, what came out of it, and more*. Retrieved from Sportskeeda: https://www.sportskeeda.com/basketball/nba-lockouts-why-happen-came-it

*NBA Player Salaries*. (2024, November 01). Retrieved from HoopsHype: https://hoopshype.com/salaries/players/

William Wu, K. F. (2017). *Classification of NBA Salaries through Player.* Berkeley: Sports Analytics Group at Berkeley.

Yang, Z. (2024). *Analysis of the Relationship between NBA Player Salary and.* London: Creative Commons Attribution License 4.0.