Modeling NBA Salaries: A Multiple Linear Regression Approach to Analyzing Player Performance and Predicting Earnings

Yuxin Fan, Arnnav Aggarwal

December 6, 2024

Contribution

- Yuxin: Data processing, Data analysis (Manual Model), Report composing, Poster review, Editing demonstration
- Arnnav: Data processing, Data analysis (Automated Model), Report review, Poster composing, Editing demonstration

1 Introduction

The National Basketball Association (NBA) is the world's most renowned basketball league, known for its intense competition both on the court and in team management [8, 9]. NBA teams face strict limitations on total player salaries under a salary cap system, which makes it critical to ensure that every dollar is allocated efficiently to maximize on-court performance [1]. This brings forth the central research question of this report: What key factors influence a player's salary the most, and how can these insights guide optimizations for both players and team managers?

Sarlis and Tjortjis utilized unsupervised learning techniques to uncover valuable features such as age and health conditions, providing insights for salary optimization [9]. Sigler and Compton employed multiple regression with player salary as the response variable, refining their model using backward selection to find significant predictors [3]. Yang et al. analyzed wage and on-court efficiency using a DEA framework, highlighting the importance of wage efficiency in team operations and its subsequent impact on performance [2].

By building on these findings, this report investigates the relationship between player performance and salary using a multiple linear regression approach. The analysis highlights player age, minutes played, points per game, shot accuracies, and position on the field as the most significant factors impacting salary. While these findings align partially with existing research, a possible explanation for difference lies in evolving team management strategies, which may have shifted focus to different factors over time. The results offer practical insights for improving both individual performance and team efficiency.

2 Methods

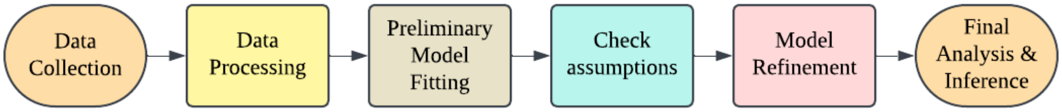

Figure 1: Linear Regression Research Steps Summary

This research used linear regression to model the relationship between various player-related factors and salary. The process began with data collection, followed by data cleaning to ensure relevance to the research scope. A preliminary model was fitted and all assumptions were checked. To address assumption violations, transformations were applied to the response variable. Subsequent model refinement included diagnosing and addressing problematic observations to improve model performance. Efforts were made to maximize the percentage of variation explained and minimize AIC by removing statistically insignificant variables, both manually and through automated selection methods. Finally, the best-fitting model was selected for comprehensive analysis and inference.

2.1 Data Collection

To address the research question of identifying key factors that influence a player's salary, we sought a dataset containing comprehensive information on NBA player salaries and relevant performance metrics. The original data was collected from Basketball Reference [6], a reputable open-data platform for official basketball statistics. The dataset used in this research was sourced from Kaggle where its original purpose was to predict All-Star Game scores for players [4]. It was preprocessed to focus on columns containing salary information and on-court performance indicators for each player. This dataset aligns with our research objective by providing sufficient variables for analysis.

2.2 Data Processing

Numerous studies have explored wage efficiency in the NBA and identified critical factors for evaluating player performance [2, 3, 9], but the conclusions were controversial. In particular, points per game, age, and experience have been consistently highlighted as major features influencing salary [3, 9]. However, factors such as field goals, assists, and blocks per game have shown varying levels of importance across different studies [5, 7]. During the data processing, we selected all features that have been discussed in previous research to ensure a comprehensive analysis and alignment with established findings.

2.2.1 Response Variable

The response variable is player salary (in dollars). The dataset contains 1,134 observations, with salaries right-skewed distributed from \$56,845 to \$37,457,154 (M = \$7,386,686, Median = \$4,152,481, SD = \$7,573,796).

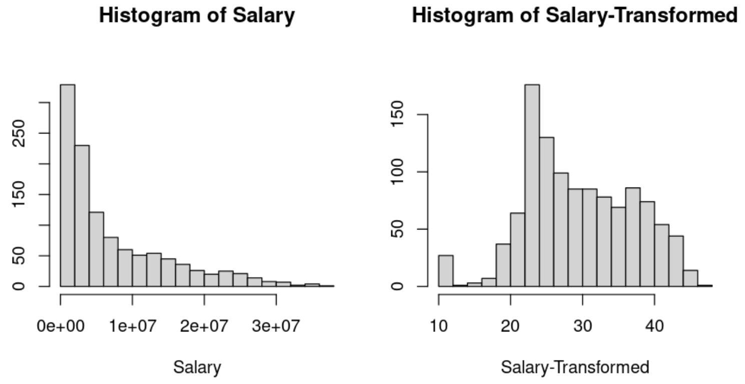

Figure 2: Histogram Summary of Response and Transformed Response (after assumption check)

2.2.2 Predictor Variables

There are 16 predictors in total, consisting of 13 numerical and 3 categorical variables. A summary of all variables is provided in Table 1-additional.

2.3 Preliminary Model Fitting and Checking Assumptions

All predictor variables are used to fit the preliminary model, as all are considered relevant. The assumptions of conditional mean response and predictor, linearity, constant variance, uncorrelated errors and normality are checked and addressed by applying transformations to the response variable, including logarithmic, square root, and Box-Cox transformations.

2.4 Model Refinement

Both manual and automated approaches are used for model refinement through variable reduction. From a manual perspective, we remove insignificant variables ($p > 0.05$) based on the t-test results of the fitted model. The model's goodness is then evaluated using an ANOVA test, along with comparisons of adjusted R^2 and VIF with the previous model. From an automated perspective, we apply an forward automated selection technique, allowing the computer to fit model with the least number of variable based on AIC optimization.

2.5 Model selection and analysis

The final model was chosen based on a balance between model goodness and simplicity, comparing the manually selected model and the automated model. All assumptions and additional conditions were rechecked to ensure no violations before the final analysis. For model goodness, higher adjusted R^2, lower AIC, and VIF values close to 1 were prioritized. Fewer variables were preferred to enhance model simplicity and interpretability. The research question was then addressed through inferences drawn from the selected final model.

3 Results

All the aforementioned methods were applied to the dataset, resulting in two refined models: one manually selected and the other generated through automated selection. These models were compared against each other and the preliminary model to evaluate improvements in performance and interpretability. All plots below are presented in the order of the preliminary model, the machine-refined model, and the manually refined model, from left to right.

3.1 Model fitting and Assumption Checking

Additional conditions were checked in advance to ensure that the patterns observed in the plots below accurately identified valid violations of assumptions. The Response-Fitted plot for the preliminary model in Figure 1-additional showed a significant shift from the diagonal line, indicating a violation of the conditional mean response due to a non-linear relationship between the predictors and the response. This was consistent with the right-skewed distribution of the response variable, suggesting that a transformation should be applied to the response. The plots for the refined models, which were built based on the transformed model, aligned with the diagonal line, indicating that the violation had been addressed. The pairwise scatter plots in Figure 2-additional displayed the relationships among numerical predictors selected in the refined models. Some scatter plots exhibited a filled curve shape due to the wide range of the response variable, but they did not violate the conditional mean predictor assumption, as there were no distinct curves. However, the plots for MP and PTS showed a curve-like shape, indicating a potential violation. This will be discussed further in the limitations section.

Figure 3 presents the residual plots for the preliminary model and the refined models. The leftmost plot exhibits a fan-out pattern with heavy clustering at the bottom, indicating violations of constant variance and uncorrelated errors in the preliminary model. To address these issues, logarithmic, square root, and Box-Cox transformations were explored. The Box-Cox transformation with a power of 0.22 on the response variable was ultimately chosen, as it resolved all violations while preserving model simplicity and interpretability. The refined models, shown in the two rightmost plots, are based on the transformed response and display relatively random scatters along $y = 0$, suggesting that the violations have been effectively addressed. However, a linear pattern persists in the lower-left corner, attributed to the presence of outliers with significantly lower salaries compared to the average.

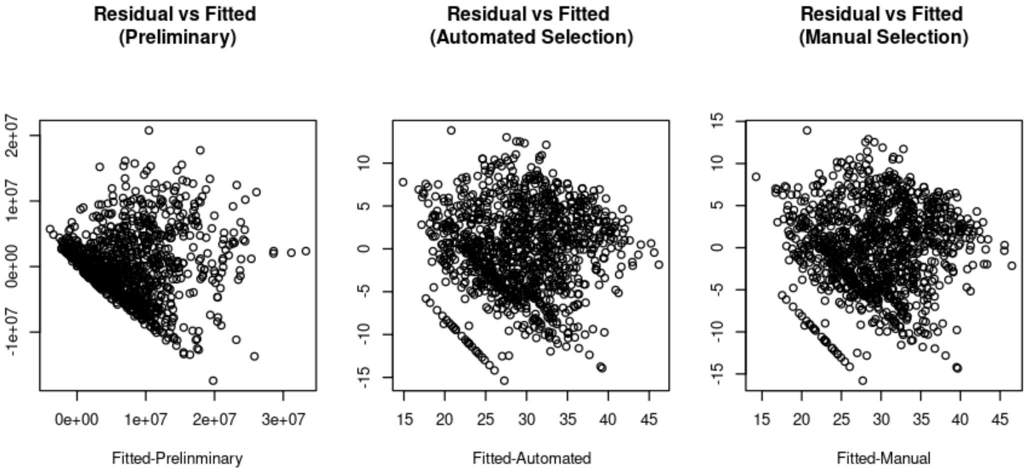

Figure 3: Residual Plots for Assumptions Checking

Figure 4 presents the normal Q-Q plots for the models. The preliminary plot shows a shifted tail compared to the red standard line ($y = x$), indicating a violation of normality. Box-Cox suggested power transformations were applied to the response, and the refined models were built based on the transformed model. They improved the tail issue, though slight deviations remain. Since almost all data points closely align with the red line, the violation of normality is considered to be addressed.

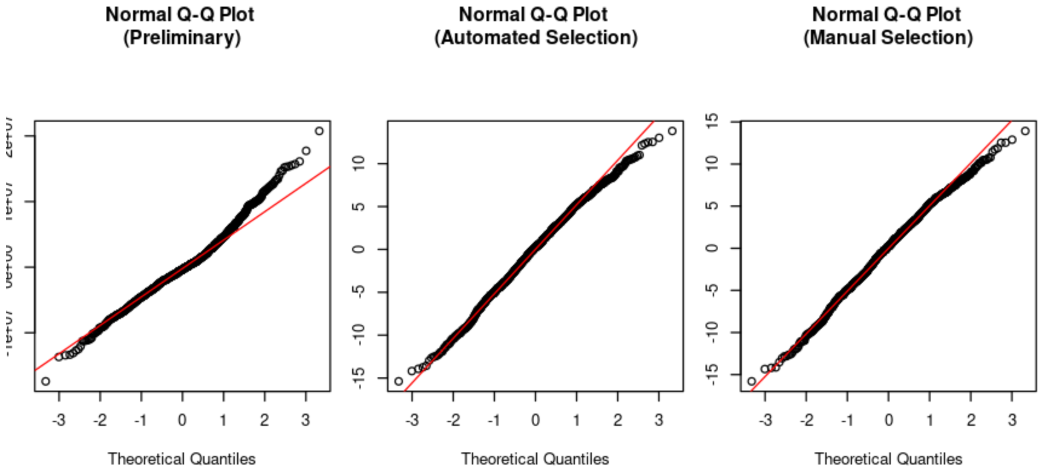

Figure 4: Normal Quantile-Quantile Plots for Assumptions Checking

3.2 Model Diagnostics and Model Goodness

Table 1 summarizes the results of the three fitted models, including significance levels. A partial F-test was performed to verify that the reduced models were significantly better than the preliminary model by failing to reject the null hypothesis. Both reduced models yielded $p < 0.05$, confirming that the variable reduction process was valid and improved model performance by filtering out irrelevant variables (manual: $p = 0.336$, automated: $p = 0.115$). Subsequently, only the manual and automated models were compared based on adjusted R^2, AIC, BIC, and VIF to finalize the model selection.

Table 1: Multilinear Regression Results

Predictors	**Preliminary**	**Automated Selected**	**Manual Selected**
Constant	11.37***	11.21785***	10.57055***
Player Age (Age)	0.6258***	0.63609***	0.63812***
Minutes Played (MP)	0.2942***	0.30200***	0.30708***
3-Point Field Goal Accuracy (X3P.)	-4.062**	-4.12992**	-4.26809**
2-Point Field Goal Accuracy (X2P.)	-7.468***	-6.98728***	-7.44167***
Position (Pos1)	-0.5647***	-0.48980***	-0.49961***
Season (Season)	0.4655*	Not Included	0.023136*
Points Per Game (PTS)	0.3628***	0.42939***	0.42282***
Free Throw Accuracy (FT.)	-3.735*	-3.88209**	-3.87614**
in the All Star Game (Play)	1.053	Not Included	Not Included
Turnovers Per Game (TOV)	0.07753	Not Included	Not Included
Games Played (G)	0.005131	Not Included	Not Included
Games Started (GS)	0.013	Not Included	Not Included
Daily Views Online (mean_views)	0.00006325	Not Included	Not Included
Personal Fouls Per Game (PF)	-0.3726	Not Included	Not Included
Steals Per Game (STL)	0.5131	Not Included	Not Included
Blocks Per Game (BLK)	-0.1435	Not Included	Not Included
F-stats	91.7	203	178.9
DF	16	7	8

1134 observations; $*p < .05$, $**p < .01$, $***p < .001$.

Table 2 summarizes the key measures for comparing the goodness of fit between the manual and automated models. Before evaluating these measures, VIF analysis was conducted to assess multicollinearity. Most predictors in both models had VIF values near 1, indicating no severe multicollinearity. However, Minutes Played and Points Per Game showed VIF values close to 5 in both models, reflecting moderate but acceptable correlation with other predictors. This aligns with the pairwise scatter plots analyzed during additional condition checks and will be further discussed in the limitation.

Table 2: Model Goodness Comparisons

Model	**Automated Selected**	**Manual Selected**
R^2_{adj}	0.5551	0.5568
AIC	6873.372	6870.170
BIC	6918.673	6920.505
VIF-Player Age	1.042556	1.043235
VIF-Minutes Played	4.224687	4.240823
VIF-3-P FG Acc.	1.240468	1.242842
VIF-2-P FG Acc.	1.168291	1.179101
VIF-Position	1.252605	1.254336
VIF-Pts Per Game	4.337536	4.351502
VIF-FT Acc.	1.243565	1.243569
VIF-Season	Not Included	1.016210

Higher adjusted R^2 and lower AIC and BIC were prioritized in evaluating model goodness. As shown in Table 2, the manual model (adjusted $R^2 = 0.557$, AIC = 6870, BIC = 6921) achieved a slightly higher adjusted R^2 and lower AIC compared to the automated model (adjusted $R^2 = 0.555$, AIC = 6873, BIC = 6919). However, the automated model had a slightly lower BIC, favoring simplicity. While the differences between the models are minimal, they indicate a trade-off between fit and simplicity. Considering the context of this research, where robustness and interpretability are preferred, the automated model is selected for final analysis.

4 Conclusion

To address the research question of identifying key factors that influence an NBA player's salary and to guide optimizations for both players and managers, this study finds that age, minutes played, and points scored are the most influential factors. These factors are positively correlated with a player's salary, meaning that as a player's age, playing time, or points increase (while holding other factors constant), their salary also increases. Factors such as position, shooting accuracy for 3-point, 2-point, and free throw are also influential in determining a player's salary.

Therefore, some suggestions can be made to players and managers:

- **Players**: To maximize salary potential, they should focus on improving their on-court performance early in their careers. By enhancing their ability to play more minutes and score more points, players can build a long-term reputation and significantly increase their salary prospects as they age.
- **Teams**: managers may want to invest in older players who can consistently play more minutes and score more points, as these factors are positively correlated with higher salaries. Additionally, considering the player's position is important, as it can influence both scoring opportunities and overall performance.

Our findings of the positive correlations align with the conclusions of Sigler and Compton [3]. However, there is a discrepancy: Sigler and Compton also found that assists, blocks, and personal fouls were statistically significant, whereas our model did not find these factors to be significant. This mismatch may be due to differences in the datasets used in both studies, or the fact that athletes with different roles cannot be judged on the same scale. Sigler and Compton's research focused on 540 players from the 2017-18 season, while we studied 1134 players from the 2016-17, 2017-18, and 2018-19 seasons.

Limitations

One limitation of the model is the potential multicollinearity between MP and PTS, which is indicated by the pairwise scatter plot and VIF values. The correlation between these two variables suggests that players who play more minutes tend to score more points, which is reasonable given the nature of the sport. However, MP and PTS cannot be removed from the model since the coefficients are significant, thus may affect the precision of the estimates due to multicollinearity. Future improvements are applying regularization techniques or conduct PCA to address this limitation.

5 Ethics Discussion

This research explored both manual and automated selection methods, comparing them in the analysis of NBA player salaries within the context of ethical considerations. Both approaches are valuable in maintaining ethical integrity, and combining them can help avoid blameworthy practices. From the perspective of motivation, researchers bring conscious intent to select variables that are directly relevant to basketball performance, enabling meaningful conclusions. In contrast, computers operate without motivation, executing statistical tasks under human guidance. While this ensures objectivity, it also means that ethical responsibility lies entirely with the researchers who design and interpret the models. From the perspective of foreseeability, researchers may inadvertently introduce bias based on prior knowledge, such as assumptions about race, gender, or misinterpretation of the data, which can lead to misleading conclusions. Automated methods, on the other hand, identify patterns autonomously and can uncover relationships that researchers might overlook, offering new insights. However, these methods can still perpetuate bias if the data used to train them is skewed or unrepresentative. Finally, the NBA is inherently a team sport, where collaboration and diverse roles define success. Research should reflect this by considering the collective performance and interactions among players, rather than focusing solely on individual contributions. Balancing manual and automated methods can help ensure a comprehensive and fair analysis while mitigating the risks of bias and ethical lapses.

A Variable Summary

Table 3: Table 1-additional: Variable Summary

Variable	Mean	Std. Dev	Min.	Max.	Type
Salary	7386686	7573796	56845	37457154	Response
Player Age (Age)	26.21	4.318613	19.00	42.00	Numerical
Minutes Played (MP)	21.79	8.414531	3.10	37.80	Numerical
3-Point Field Goal Accuracy (X3P.)	0.3135	0.1186585	0.0000	1.0000	Numerical
2-Point Field Goal Accuracy (X2P.)	0.4988	0.07698729	0.0000	0.8330	Numerical
Position (Pos1)	3.086	1.427477	1.000	5.000	Categorical
Season (Season)	2.021	0.8133333	1.000	3.000	Categorical
Points Per Game (PTS)	9.706	6.056916	0.400	36.100	Numerical
Free Throw Accuracy (FT.)	0.7513	0.1149212	0.0000	1.0000	Numerical
in the All Star Game (Play)	0.06349	0.2439535	0.00000	1.00000	Categorical
Turnovers Per Game (TOV)	1.235	0.8127374	0.000	5.700	Numerical
Games Played (G)	58.61	20.89106	1.00	82.00	Numerical
Games Started (GS)	29.48	29.18633	0.00	82.00	Numerical
Daily Views Online (mean_views)	1042.52	2225.554	1.14	34147.96	Numerical
Personal Fouls Per Game (PF)	1.831	0.6864644	0.000	3.900	Numerical
Steals Per Game (STL)	0.6981	0.4116097	0.0000	2.4000	Numerical
Blocks Per Game (BLK)	0.4267	0.4018936	0.0000	2.7000	Numerical

B Additional Condition Check

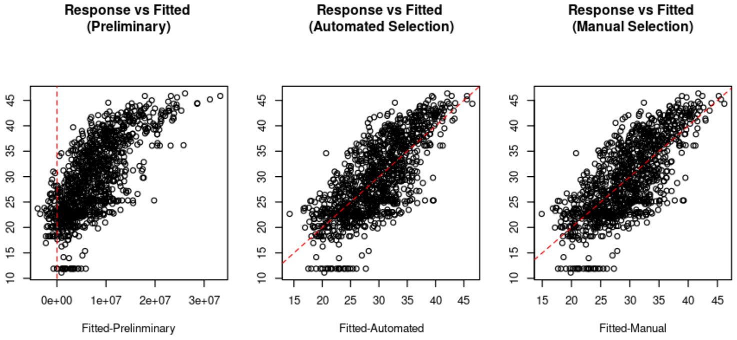

Figure 1-additional: Scatter Plot for Conditional Mean Response Check

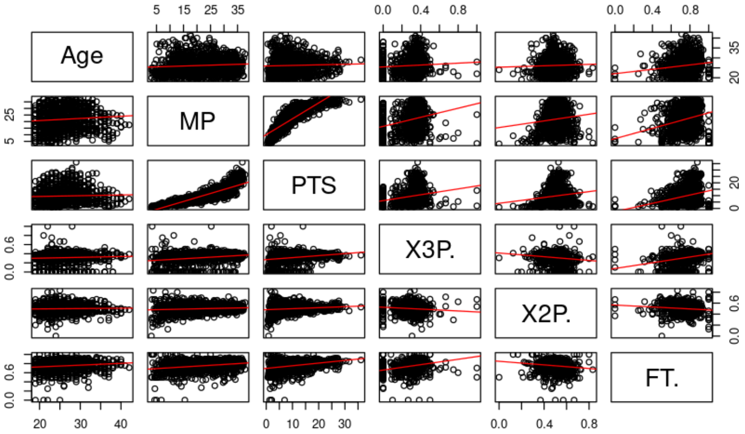

Figure 2-additional: Pairwise Scatter Plot for Conditional Mean Predictor Check

References

[1] Nikos Chatzistamoulou, Kounetas Kostas, and Antonakis Theodor. “Salary Cap, Organizational Gap, and Catch-up in the Performance of NBA Teams: A Two-Stage DEA Model Under Heterogeneity”. In: *Journal of Sports Economics* 23.2 (2022), pp. 123–155. DOI: 10.1177/15270025211022253. URL: https://doi.org/10.1177/15270025211022253.

[2] Chiang-Ping Chen Chih-Hai Yang Hsuan-Yu Lin. “Measuring the efficiency of NBA Teams: Additive Efficiency Decomposition in two-stage DEA”. In: *Annals of Operations Research* 217.1 (2014), pp. 565–589. DOI: 10.1007/s10479-014-1536-33.

[3] Kevin Sigler William Compton. “NBA Players’ Pay and Performance: What Counts?” In: *The Sport Journal* 21 (2018). URL: https://thesportjournal.org/article/nba-players-pay-and-performance-what-counts/.

[4] Riccardo Rubini Davide Ratto Alfredo Galli. *NBA PLAYERS 2016-2019*. Accessed: 2024-12-05. 2019. URL: https://www.kaggle.com/datasets/davra98/nba-players-20162019.

[5] Justin Ehrlich, Shane Sanders, and Christopher J. Boudreaux. “The relative wages of offense and defense in the NBA: a setting for win-maximization arbitrage?” In: *Journal of Quantitative Analysis in Sports* 15.3 (2019), pp. 213–224. DOI: doi:10.1515/jqas-2018-0095. URL: https://doi.org/10.1515/jqas-2018-0095.

[6] Sports Reference LLC. *Basketball Reference*. Accessed: 2024-12-05. URL: https://www.basketball-reference.com/.

[7] Rodney Paul, Andrew Weinbach, and Jeremy Losak. “Market Efficiency and the Setting of Salaries for NBA Daily Fantasy on FanDuel and Draft Kings”. In: *The Journal of Prediction Markets* 14.2 (Dec. 2020), pp. 45–60.

[8] John Robst et al. “Skin Tone and Wages: Evidence From NBA Free Agents”. In: *Journal of Sports Economics* 12.2 (2011), pp. 143–156. DOI: 10.1177/1527002510378825. URL: https://doi.org/10.1177/1527002510378825.

[9] Vangelis Sarlis and Christos Tjortjis. “Sports Analytics: Data Mining to Uncover NBA Player Position, Age, and Injury Impact on Performance and Economics”. In: *Information* 15.4 (2024). ISSN: 2078-2489. DOI: 10.3390/info15040242. URL: https://www.mdpi.com/2078-2489/15/4/242.