NBA Player Analysis Project

Christopher Vu, Owen Ngabirano

December 7, 2024

Abstract

For this project, we look to study the statistical relationship that player performance metrics have with player's anthropometrics. The player performance metrics in question are points scored, rebounds, and assists, while the anthropometrics are height, weight, and wingspan. We used both linear models and non-linear models for the different response variables. Our non-linear model of choice for all the response variables was the Generalized Additive Model which uses splines to make a better fit for the non-linear nature of our data. Our analysis gives interesting results but certainly emphasizes that a lot more factors affect the metrics of performance that we're studying beyond just the anthropometric. While this is the case, having an analysis of these givens, allows us to make predictions that go a long way in sports analysis.

1 Introduction

The project analyzes the impact that a player's height, weight, and wingspan have on his performance in terms of points, rebounds, and assists in a National Basketball Association League game. This provides some tools for different beneficiaries of this growing aspect of the entertainment industry to make inferences and predictions on the player's performance. This can be used by coaches to game plan, scout players, and sports betting companies to make betting lines.

In preparation for this, we relied on our own knowledge of the NBA as well as articles on platforms such as ESPN and others. We also researched models that are typically used for the type of data we had and the type of variables we planned to use. We found that most prediction used in the industry at the moment was based on the player's recent averages in performance metrics and not much consideration of their anthropometrics or their opponents'. With our project we hope to find the relevant physical body build-related predictors for each of a player's points, rebounds, and assists in a game. Our goal is to predict a player's performance in any game.

We use generalized linear models and non-linear regression models particularly Generalized Additive models to model our relationships between the response variables and predictors. For the generalized linear model we use the Gaussian family with a identity link function. For the non-linear model, we selected GAM because it is an extension of the GLM but with a better study of the non-linear patterns of the data.

2 Methods

2.1 Dataset

We used two datasets: box scores from the 2023-2024 NBA season (13,188 rows, 18 columns) and anthropometric data for all NBA players since 2000 (1,631 rows, 7 columns). The box scores contained detailed game statistics, while the anthropometrics data included player measurements such as height, weight, and wingspan. To explore the impact of anthropometrics on performance, we calculated matchup differences and focused on players who averaged at least 12 minutes per game, ensuring meaningful data from those with sufficient court time. To select who the match ups are for each player for each game, we filtered the positions that qualified for each player. For point guards, only point guards and shooting guards were considered, for shooting guards, only point guards, shooting guards and small forwards were considered, for small forwards, only shooting guards, other small forwards and power forwards were considered, for power forwards, only other power forwards, small forwards and centers were considered, and finally for centers, only power forwards and other centers were considered. After this they were ranked in order of who had the most similar number of minutes played in the game and filtered for the top three in this list. These top three were averaged, in-terms of height, weight and wingspan.

We standardized the response variables—points, rebounds, and assists—into per-minute metrics. Standardization reduces the influence of extreme values caused by differences in playing time and improves the model's predictive accuracy. After cleaning and preprocessing, the final dataset consisted of 6,856 rows and 10 columns, with target variables including points per minute, rebounds per minute, and assists per minute. The predictors were position (categorical), player height, weight, wingspan, and the differences in these measurements compared to the opponent's averages.

2.2 Exploratory Data Analysis

Before building our models, we conducted exploratory data analysis to better understand the dataset's characteristics and relationships. Figure 1 illustrates the distributions of our response variables. Points per minute (PPM) follows a roughly normal distribution, while rebounds per minute (RPM) and assists per minute (APM) are right-skewed. This variation in distribution can be attributed to the nature of these metrics. Scoring points is a fundamental aspect of basketball, with most players contributing to it regularly. The consistency of scoring opportunities across games and positions leads to a more balanced distribution. In contrast, rebounding is heavily influenced by physical attributes such as height, strength, and positioning. Players in specific roles, like centers and power forwards, are typically tasked with rebounding, resulting in fewer individuals with high rebounding rates.

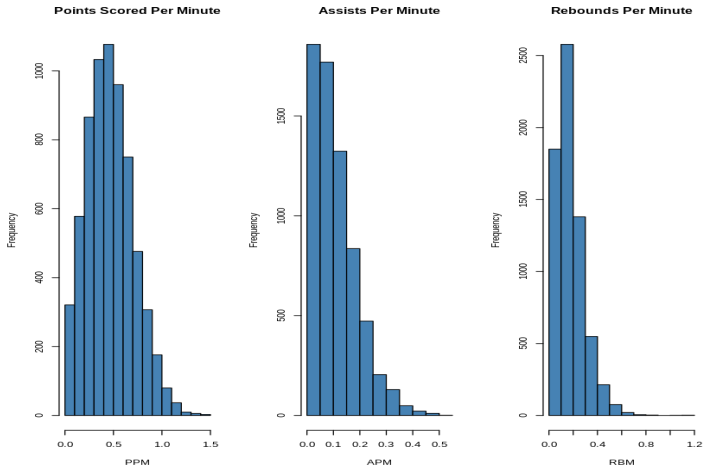


Figure 1: Distributions of Points Per Minute (PPM), Assists Per Minute (APM), and Rebounds Per Minute (RBM). Each histogram visualizes the frequency of observed values within the dataset.

Similarly, assists are position-specific, as point guards are primarily responsible for playmaking. This specialization results in a concentration of low assist rates among players in other roles. The skewness observed in RPM and APM highlights how these metrics are more role-dependent compared to PPM, which is influenced by broader contributions across the team.

We also examined the position-wise variation in PPM, RPM, and APM through boxplots, as shown in Figure 2. For PPM, there is minimal variation across positions, confirming the uniform contribution of most players to scoring.

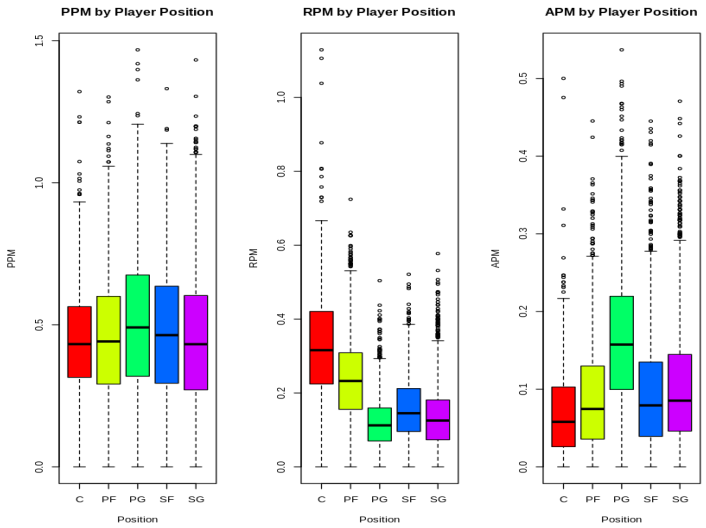


Figure 2: Boxplots of PPM, RPM, and APM by player position. The plots show distribution variations, with PPM exhibiting limited spread, RPM having more variability, especially for centers, and APM concentrated at lower values for non-point guards.

This aligns with the roughly normal distribution seen earlier. In contrast, RPM shows a wider spread, particularly among centers, reflecting the role-specific nature of rebounding. Centers, tasked with securing rebounds, exhibit a broader range compared to point guards,

who contribute less to this stat. Lastly, APM reveals a similar trend, with point guards showing higher values due to their playmaking role, while other positions exhibit lower values. These boxplots reinforce the role-dependent nature of RPM and APM, with PPM remaining more consistent across positions.

In Figure 3, the correlation matrix reveals the relationships between all variables, with a particular focus on points, rebounds, and assists. Notably, points scored show little correlation with the predictors. In contrast, rebounds are positively correlated with the predictors, likely due to the physical attributes (e.g., height and size) that benefit players in roles focused on rebounding.

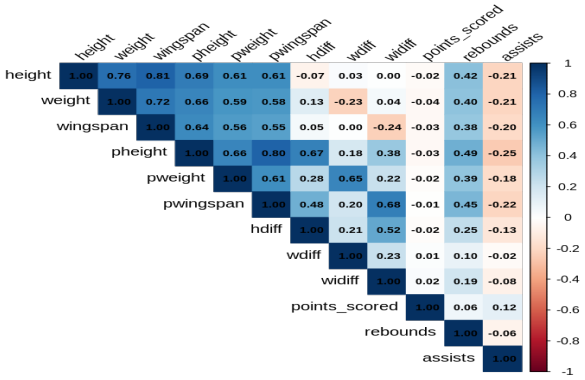


Figure 3: Correlation matrix showing the relationships between points, rebounds, assists, and predictor variables, with rebounds positively correlated and assists negatively correlated with the predictors.

Conversely, assists have a negative correlation with the predictors, reflecting the specialized nature of the point guard role, where agility and playmaking ability are more relevant than physical stature. These patterns align with positional roles, where players in positions like centers tend to show stronger correlations with rebounding, while point guards are more aligned with assist metrics.

3 Models

3.1 Linear Regression Models

We began by using linear regression models to analyze the relationship between player performance (points per minute, rebounds per minute, and assists per minute) and several predictors. The models were as follows:

$$\text{PPM} \sim \text{Pos} + \text{Hdiff} + \text{Wdiff} + \text{Pheight} + \text{Wingspan} \tag{1}$$

$$\sqrt{\text{RPM}} \sim \text{Pos} + \text{Wdiff} + \text{Pheight} + \text{Pwingspan} + \text{Weight} \tag{2}$$

$$\sqrt{\text{APM}} \sim \text{Pos} + \text{Wdiff} \tag{3}$$

The inclusion of position (Pos), player height difference (Hdiff), weight difference (Wdiff), player height (Pheight), wingspan (Wingspan), and weight (Weight) as predictors allows us to explore how these factors influence performance. The square-root transformation of RPM and APM was applied due to skewness in their distributions, as indicated by exploratory data analysis.

We performed *stepwise regression* for variable selection, iterating through models by adding and removing predictors to minimize the *Akaike Information Criterion (AIC)*. This process ensures only the most relevant predictors are included, balancing model fit and complexity.

3.2 Generalized Additive Models (GAM)

Next, to account for potential non-linear relationships between continuous predictors and the response variables, we applied Generalized Additive Models (GAMs). GAMs are flexible models that allow for non-linear effects through smoothing functions. The GAMs used for each response variable were as follows:

$$\text{PPM} \sim \text{Pos} + s(\text{Hdiff}) + s(\text{Wdiff}) + s(\text{Pheight}) + s(\text{Wingspan}) \tag{4}$$

$$\text{RPM} \sim \text{Pos} + s(\text{Wdiff}) + s(\text{Pheight}) + s(\text{Pwingspan}) + s(\text{Weight}) \tag{5}$$

$$\text{APM} \sim \text{Pos} + s(\text{Wdiff}) \tag{6}$$

The inclusion of the smoothing function $s(\cdot)$ allows for modeling non-linear relationships between continuous predictors and the response variables, providing more flexibility than the linear regression models.

3.3 Model Justification

The linear regression models provide a simple and interpretable approach to quantifying the influence of various predictors on performance. The use of stepwise regression and AIC ensures the models remain efficient and relevant. However, given the potential for non-linear relationships, GAMs were employed to capture more complex trends. These methods are appropriate given the nature of the data, where performance metrics are influenced by a combination of linear and non-linear factors.

3.4 Assumptions and Limitations

The linear regression models assume that residuals are independent, normally distributed, and exhibit homoscedasticity. We assessed these assumptions through residual diagnostics, including normality and variance checks. The square-root transformation of RPM and APM was applied to address issues with skewness. For the GAMs, smooth functions for continuous variables were chosen to model non-linear trends; however, careful tuning of smoothing parameters was necessary to prevent overfitting. Both model types are sensitive to the choice of predictors and may be influenced by multicollinearity, which we addressed by selecting variables through stepwise regression.

4 Results

4.1 Generalized Linear Models

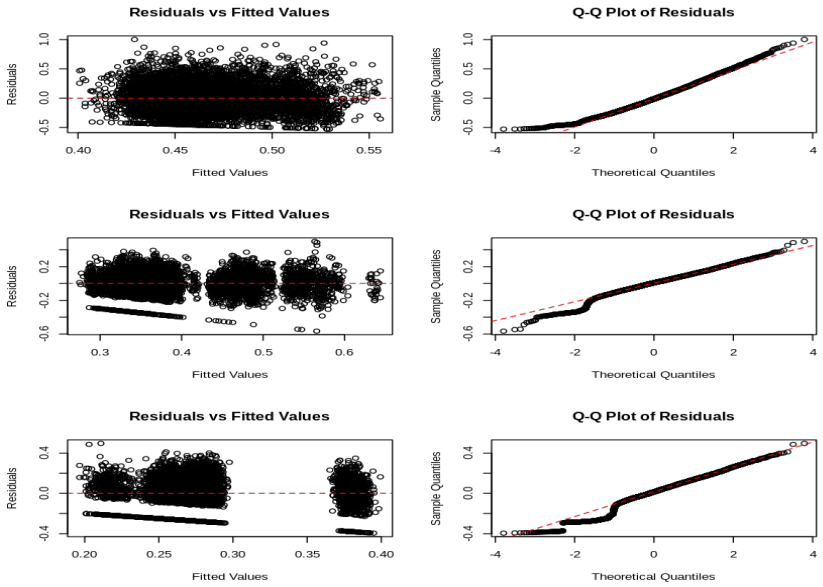
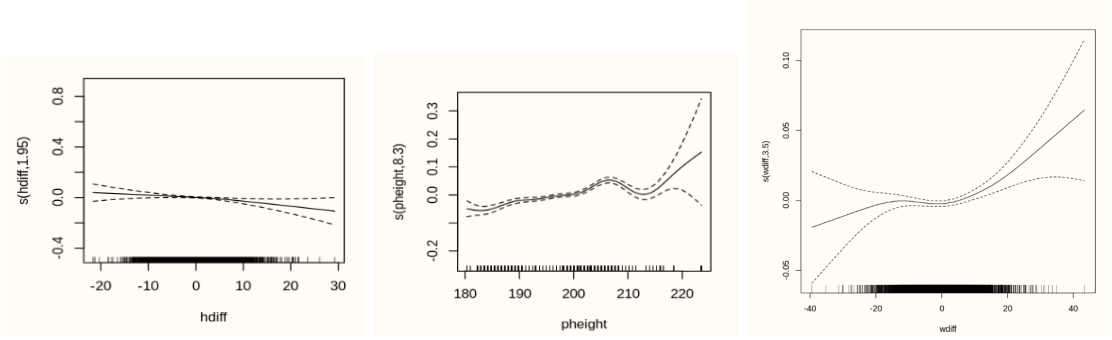


Figure 4: Residual analysis plots, residuals vs fitted values on the left and qq-plots on the right for Generalized Linear Models. Top is Points Per Minute, middle is Rebounds Per Minute and bottom is Assits Per Minute.

For points per minute, our GLM returned that all our coefficients besides the Power Forward position and the intercept were statistically significant with a 95% CI meaning they had most likely no impact on points scored per minute. The strongest predictor was the Point Guard position with a 0.1444377 positive effect on the points scored in a game. This model achieved an AIC of -148.71. The analysis of the residual plots implied a somewhat linear relationship between the response and it's predictors. For rebounds made in a game, the linear model described earlier returned that all but weightDifference and player weight were significant, with the highest absolute effect being the small forward position which has an effect of -0.1229826 meaning that being a small forward and not

a center lowered the rebounds made by that much. This model had an AIC of -8599.6. As shown in figure 4, the plot for residuals against fitted values has clusters and other patterns and the q-q plot has a divergence on the lower end which is evidence against a linear relationship of rebounds and the selected predictors. Finally for assists, our linear model returned that all the coefficients were significant with 95% certainty. The largest coefficient estimate being the point guard position which has a positive effect of 0.1647955. Our residual analysis in figure 4, like with the rebounds linear model shows clustering and other proof against linear relationship of the data being analyzed.

4.2 Non-linear Generative Additive Models.



(a) HeightDifference in PPM **(b)** Player Height in RPM **(c)** Weight Difference in APM

Figure 5: This figures show the different examples of smooth terms. The more curves in the plot, the more significant the terms is.

As shown by the residual plots, we did no have sufficient evidence to claim a linear relationship between any of the response variables with their predictors. This prompted analyzing using a non-linear model. We selected GAM because it allowed for the linear intervals of our data to be analyzed that way while the non-linear intervals are treated appropriately as well. For the first response variable, PPM, a smoothing term was made from the continuous predictors from the prior linear model. An example is shown in figure 5 where the height difference shows some very subtle non-linearity on the ends of its range. The effect barely deviated from zero meaning low significance of the term and low effect on the points scored per minute. For this model, the player's wingspan and his height were the most significant and also had the highest estimated effect among the continuous variables. Among the positions, the point guard still had the highest effect, though now, all the position effects were significant. The AIC of this model was -505.1898, better than the linear model of -148.71.

Secondly, the rebounds per minute response variable, when analyzed using GAM, showed significance among all the position and continuous predictors besides, the smooth term of weight difference. The example provided of the smoothing terms in figure 4b for the player's height and the smoothing term for player wingspan had the most significance with strong influence on the maximum end of their respective ranges i.e. predicted

rebounds better for taller groups of players and for players with longer wingspans respectively. The positions which all were significant, had the same trend of effect as in the linear model. The AIC of this model was -12350.08 better than the linear model.

For Assists, the singular continuous predictor, was given a smooth term and which had the most influence on the extreme ends of the range i.e. the tallest group of players and the shortest. All the positions were still significant. This model had an AIC of -14854.79 which was an improvement from the linear model of AIC -6672.877.

5 Discussion

This study examined the effects of a player's height, weight, and wingspan on basketball performance, specifically focusing on points per minute (PPM), rebounds per minute (RPM), and assists per minute (APM). We found that GAM performed better than the linear regression models. This indicates that the relationship between the predictors and performance is non-linear. This highlights the complexity of how player attributes interact, particularly across different positions, which a simple linear model would fail to capture.

Our findings align with previous studies showing the importance of physical attributes in basketball performance, but extend this by demonstrating the need for non-linear modeling. However, the analysis has limitations, primarily due to the omission of key factors like player health, injuries, fatigue, and the quality of opposing teams, all of which significantly affect performance. Game-specific variables like home-court advantage were also not included, which are known to influence player performance.

To improve future models, we could add more factors such as player efficiency ratings, shot selection, and team strategy is crucial. Also, experimenting with advanced machine learning techniques such as decision trees or random forests to see if they can improve predictive accuracy. Our goal one day would be to predict game-specific player performance.