To what extent is height a reliable metric to be used to predict an NBA player's stat line?

William See

December 2023

1 Introduction

Every year, the National Basketball Association (NBA) takes in 60 new players through 'the draft'. The draft is used to bring in new young talent, with the first fourteen picks go to the worst-performing teams across the season in hopes these young stars can help turn around the franchise that selects them. Each team will send scouts to college (University) games, private workouts, tournaments, or the NBA's 'combine' [1] to help determine which player will be the best fit for their team. However, if NBA teams could have access to only the player's height, would this one metric alone offer any insight into how well the players will perform?

2 Dataset and data preparation

One dataset was used as part of my investigation, found as [2]. The former of the two is a data set contains twenty-six seasons of data for every NBA player contracted to a roster across that period. The dataset has 23 rows, which can be categorised into personal and professional data, such as name, date of birth, and team represented. Then, performance-based data such as points, assists and rebounds are measured per game averages. When producing the regression models, I have chosen only to use players who took part in 41 games (50%) or more in a single season. Any player who takes part in less than 50% of games would not be playing enough minutes per game and, therefore, have stats that may negatively affect the correlation.

3 Applied Methods

3.1 Polynomial regression

The method applied to the dataset was polynomial regression, a form of regression analysis used to model the relationship between a dependent variable and one or more independent variables. In terms of the research, I wanted to see how much impact height had on a player's performance-based statistics and if any correlations existed. The R-squared value in polynomial regression measures how well the chosen polynomial curve fits the data. It shows the proportion of variance in the dependent variable explained by the model's independent variable(s). A higher R-squared (closer to 1) indicates a better fit, meaning the model captures more of the variability in the data. Polynomial regression aims to find the best-fitting curve by adjusting coefficients in a polynomial equation to minimise the difference between predicted and observed values.

3.2 Justification

In an attempt to answer the research question, the R-squared value will indicate how much height impacts each type of performance metric in basketball. Polynomial regression was chosen over linear regression because it provides more accurate answers to the research question. While linear regression fits a linear equation to the data, polynomial regression extends this idea by introducing polynomial equations to better fit complex patterns in the data. Classification and clustering were contemplated but ultimately disregarded. Classification was dismissed due to their inability to offer a comparative analysis of how height impacts each specific statistical category. While it could potentially group players based on height ranges, it wouldn't allow an exploration of the

relationships between height and various performance metrics. Similarly, clustering methods, while proficient in identifying patterns within data, wouldn't provide a direct means to assess the impact of height on individual statistical aspects.

3.3 Limitations

Data on either end of the height spectrum was scarce, so the regression line can easily be influenced, or any correlations could be affected by a single player performing over or under the expected level for a season. Further data cleaning was considered but not applied as a way to prevent oversimplifying.

The model makes predictions for the most 'average' player at each height and does not account for any level of 'talent' a player will have when being drafted by a team.

The model views the history of the NBA as a single entity, so it does not account for the evolution of the game, which may make specific statistics more or less critical in the modern game.

4 Results

Figure 2 found in Appendix "A" can also be used to show the correlation between height and the different statistics.

4.1 Points by height

Figure 1a shows the predicted points per game by a player's height; a visual description would suggest players in the 180-190cm range and above 225cm are expected to average the most points. However, the range of the data is not that large, suggesting height does not really affect point scoring. This is supported by an R-squared value of 0.0066, meaning only about 0.66% of the variability in the number of points scored by NBA players can be explained by their height using the polynomial regression model. In other words, points scored by NBA players can be explained by their height.

4.2 Assists by height

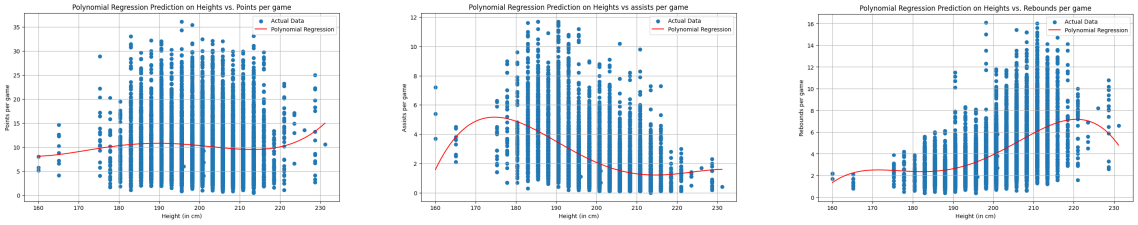
From figure 1b, the regression line suggests players considered to be on the smaller side of the average will average more assists, with the average dropping off as height increases before a small rise at the end. The R-squared value of 0.2747 suggests there is a much higher relationship between assists and height. While 27.47% is not a large value, it can suggest height does play a factor in the number of assists a player gets.

4.3 Rebounds

The prediction made by the regression line in figure 1c shows as height increases, the number of rebounds a player get significantly increases. Although, at the end of the graph, we do see a fall-off. Theoretically, this trend largely makes sense; a taller player will often be the one positioned nearer the basket to claim the rebounds of missed shots. The hypothesis I stated is supported by a R-squared value of 0.299 which tells us a correlation does exist, although again it cannot be considered to be strong.

5 Conclusion

To answer the research question, we can conclude height is not a good enough metric alone to predict a player's stat line. Despite the regression lines being a visual aid acting as a rough guide to what we can expect a player of any given height to average, the R-squared values prove that there is not enough of an impact of height on the statistics to be a reliable predictor. However, height seems to contribute to the number of assists and rebounds a player gets, and further analysis may provide insight into what other factors are.



(a) Points (b) Assists (c) Rebounds

Figure 1: Polynomial regression line shown on graphs predicting statistics against Height

References

[1] National Basketball association. *Draft Combine*. 2023. URL: `https://www.nba.com/stats/draft/combine`.

[2] Justinas Cirtautas. *NBA Players*. 2023. URL: `https://www.kaggle.com/datasets/justinas/nba-players-data`.

A Appendix

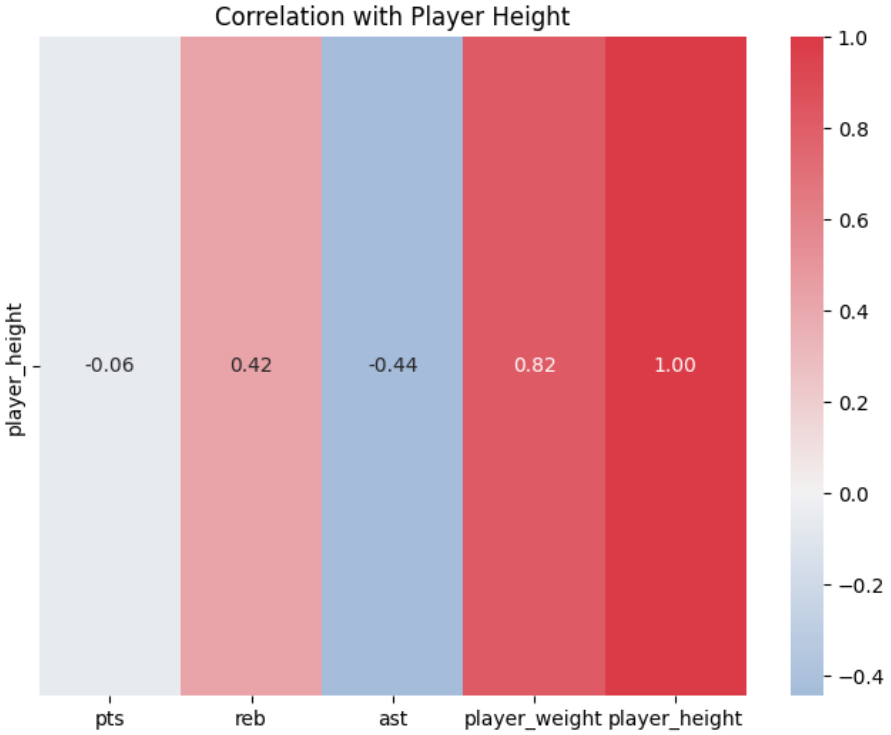


Figure 2: Figure showing correlation between statistics and height