Does height affect performance in basketball?

Term Project for Data Analysis 2

Grigoryan Anna, Monika Molnar
2018/12/17

Abstract

We address the question, whether taller athletes have better performance in basketball. Our results based on data from NBA 2014-2015 season, suggest that the relationship between height and efficiency ratings is significant and can be approximated by linear regression with quadratic polynomial. Additionally, the average height among across the positions is significantly different; moreover, central players are on average 13.35 cm taller than others, and the shooting guards are the ones to demonstrate significant increase in performance just because of height. This can be important for sport managers, as they can make sure their athletes are in the best positions to succeed and achieving higher efficiency ratings for their athletes thus aiming for higher NBA salaries.

NBA basketball is the highest level of basketball one can play in the world. Watching that league (or just know basketball in general), it is evident that usually very tall people play it, because with longer arms it is easier to throw and dunk better. However, there are certain aspects of basketball that shorter players do better, because of their lower center of gravity they can be faster and more agile. With this in mind, in our research we wanted to examine the relationship between the player's performance, and their height.

We are interested in the relationship of average height of athletes and their efficiency scores for the NBA 2014-2015 season. We argue, that the height has a strong association on the efficiency rating. We believe that if an athlete is taller than potentially he will have higher efficiency rating on average. This also give a comparison for sports managers, while given their athlete's potential performance from their body statistics (and other variables) they can have an idea how to benchmark the NBA contracts salaries. In order to address this question, we select a specific data set from https://www.basketball-reference.com and propose a level-level linear model, which gives some insight for this question.

1. Data

We restrict our attention to NBA 2014-2015 season players. The data was available on Kaggle, scraped from https://www.basketball-reference.com website. The data is official, and the data is continuously reviewed, crosschecked and used by numerous sport organizations, thus quality of the data is good, therefore we do not need to worry about systematic measurement errors in our variables.

The final set of variables is the following with definitions:

- **Height:** height of athletes.
- **Age:** age of athlete as of season 2014-2015
- **BMI:** BMI (body mass index)
- **Points (PTS):** Points made during the season
- **Efficiency Rating (EFF):** Efficiency ratings calculated by NBA
- **Position (Pos):** Position the athlete played.

The following table shows the descriptive statistics of these variable.

Table 1: Descriptive statistics of the variables

	Height	Age	BMI	Points	Efficiency Rating
Min	172.5	20	20.02	0	0
1st quartile	190	24	24.26	166.5	175.5
Median	197.5	27	25.33	433	517
Mean	197	27.48	25.22	514.5	576
3rd quantile	205	30	26.18	780.5	836
Max	215	39	27.98	2217	2202

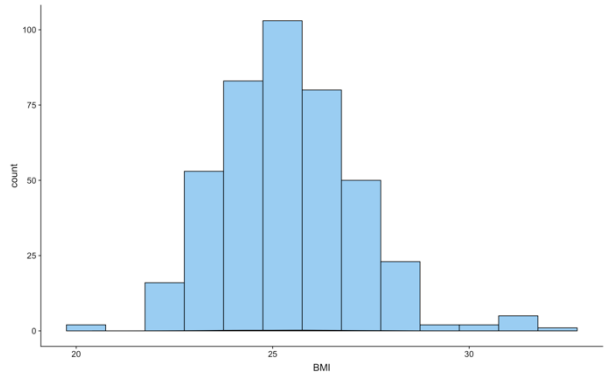

Figure 1: BMI density plot

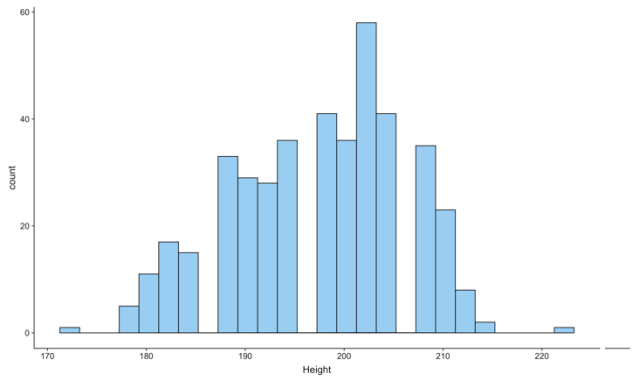

Figure 2: Height density plot

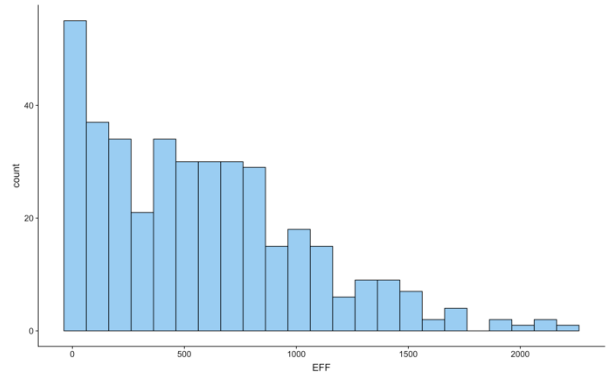

Figure 3: Efficiency density plot

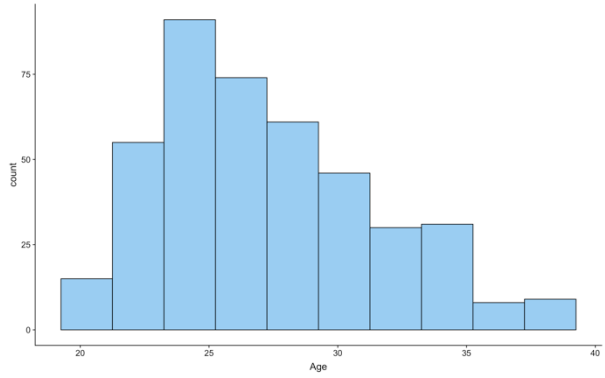

Figure 4: Age density plot

We want to model average performance during the season on average heights. Our aim, is to compare athletes that are similar in many ways but differ in their heights. To this purpose, we also include controls in the analysis, such as the BMI. We filtered our data, dropping observations with BMI below 21 and above 28. In our final dataset there are 391 observations.

2. Model

We want to regress efficiency ratings- ***EFF*** on ***Height***. First, we check a non-parametric estimator – lowess smoother – to have a general idea about the functional form between these two variables. Here we see a clear convex curve.

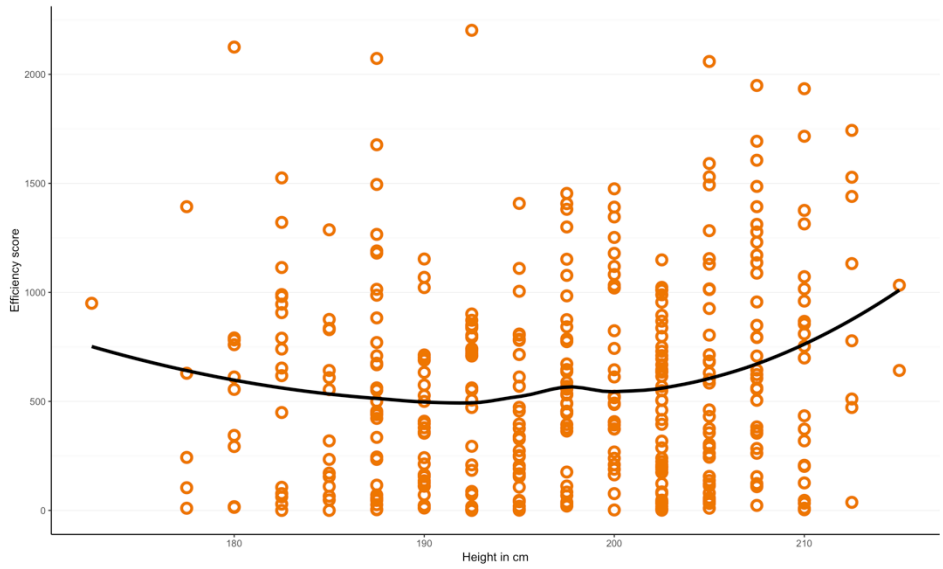

Figure 5: Loess Regression of EFF on Heights

Hence, the linear regression with splines will not be a good approximation. Instead, we will capture this connection with a linear model using a quadratic polynomial.

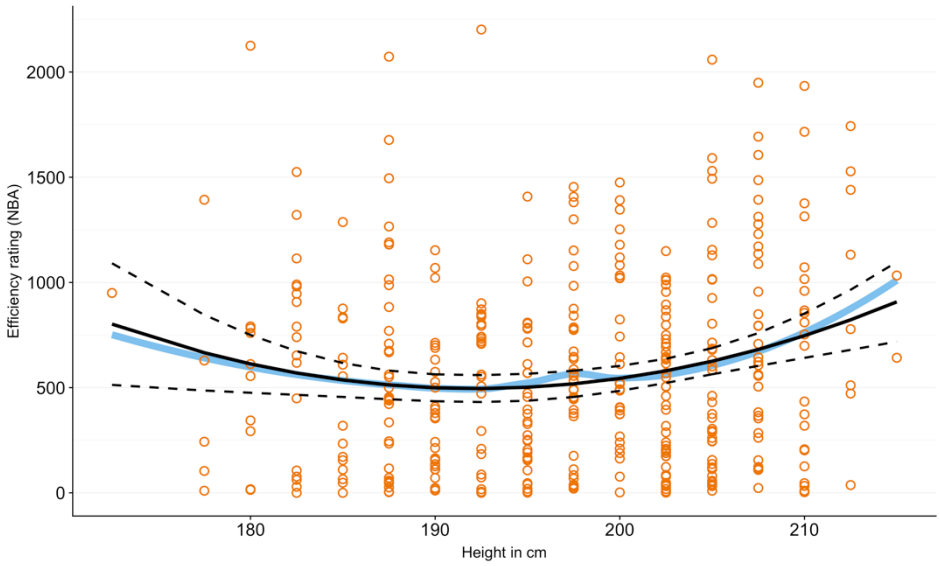

Figure 6: Linear regression of Efficiency on Height using quadratic polynomial.

Table 2: Regression results

	Dependent variable:		
	EFF		
	(1)	(2)	(3)
young			-148.819
			(46.630)***
Height	5.793	-304.207	-281.191
	(2.920)**	(111.485)***	(108.395)***
Heightsq		0.791	0.732
		(0.285)***	(0.277)***
Constant	-565.419	29,726.440	27,554.080
	(574.574)	(10,900.110)***	(10,586.350)***
Observations	391	391	391
R^2	0.012	0.033	0.058
Note:			*p<0.1; **p<0.05; ***p<0.01
			Robust standard errors in parentheses

The results suggest that without controlling for any other variable, taller athletes are associated with 5.793 higher efficiency. In other words, athletes with 1 cm taller height, on average, perform better by 5.793 in average efficiency ratings. To capture the pattern of association between the efficiency and height, the quadratic polynomial regression has been used. The results suggest that the relationship is convex, as the coefficient of the quadratic variable β_{height^2}=0.791 is positive. To learn more about the role of height, let us try comparing athletes that are similar to each other but differ in age. To this end, let us add a dummy variable for age (young, if less than 25 years; not young, if older).

$$(1)\ EFF = -565.419 + 5.793 height$$

$$(2)\ EFF = 29726 - 304.207 height + 0.791 height^2$$

$$(3)\ EFF = 27554 - 281.191 height + 0.732 height^2 - 148.819 young$$

We find that just being young is associated with 148.819 points less efficiency compared to players who are older than 25 years. Based on our extended model we can state with 95% confidence the association between the efficiency and selected variables in significant. However, age and height can explain only 5.8% of variation efficiency ratings.

3. Robustness analysis

In order to assess the robustness of our model, we decided to look for interaction with the player positions, as the latter have a strategic role during the game. We run three models, one including age, $height$, $heightsq$ and dummies for the position; another model only including the interaction between efficiency rating and the dummy variables of position; and a model describing the association between height and position.

Table 3: Regression results.

	Dependent variable:		
	EFF		Height
	(1)	(2)	(3)
young	-155.420		
	(47.045)***		
Height	-208.399		
	(143.453)		
Heightsq	0.571		
	(0.371)		
c	-118.272	221.376	15.769
	(146.662)	(78.891)***	(0.553)***
pf	-114.127	101.670	11.429
	(102.477)	(70.952)	(0.517)***
pg	132.006	93.415	-7.164
	(79.608)*	(67.018)	(0.651)***
sf	-110.613	13.534	7.345
	(77.672)	(62.590)	(0.522)***
Constant	19,521.540	499.133	192.551
	(13,894.170)	(42.767)***	(0.396)***
Observations	391	391	391
R^2	0.068	0.027	0.832
Note:			*p<0.1; **p<0.05; ***p<0.01
			Robust standard errors in parentheses

Although the first regression model, produces only two significant coefficients, the height is not one of them. In contrast, the results from the second regression with efficiency over player positions, suggests that the interaction between efficiency and being a central player is significant at a 95% level. Moreover, on average, central players are scoring 221.376 point more than players in the other positions. This incremental significance of position variable after removing the height variables, brings us to the final model, where we look for the association between the height and position of the player, with the hypothesis that taller players are initially selected as central players. The result suggests, that indeed height is significantly different (95% level) across the positions, i.e. central players (Cs) are on average expected to score 15.769 points more in efficiency ratings, than SGs (shooting guards). Power forwards (PFs) are on average expected to score 11.429 points more, point guards (PGs) -7.164 less and small forwards (SFs) 7.345 point more in average efficiency ratings compared to SGs (shooting guards).

$$(4)\ height = 192.551 - 15.7c + 11.43pf - 7.16pg + 7.35sf$$

4. Causality and external validity

Do taller athletes perform better in basketball? Is there a causal relationship?

Indeed, for athletes that are of almost same body composition (22<BMI<28) we found a positive correlation between height and their efficiency ratings.

According to basketball rules, some positions have better chance of shooting and scoring point than others. We have also proved this through our regression. On the other hand, players are positioned according to their height. Thus, the association of efficiency and height is sensitive to the positions.

Hence, we believe we found a useful pattern, but we would need much more control variables to believe we are close to causality, which is why after controlling for the position, we got the following results.

Table 5: Regression results.

	Dependent variable:				
	EFF				
	(1)	(2)	(3)	(4)	(5)
Height	28.987	21.866	-1.725	1.164	21.871
	(18.886)	(22.096)	(9.910)	(15.048)	(11.163)*
Constant	-5,318.138	-3,859.398	912.429	280.061	-3,712.061
	(3,924.831)	(4,498.163)	(1,849.219)	(3,005.508)	(2,150.649)*
Observations	61	76	84	72	98
R^2	0.028	0.016	0.0003	0.0001	0.041
Note:					*p<0.1; **p<0.05; ***p<0.01
					Robust standard errors in parentheses

$$(6)\ EFF = -3712 + 21.871 height\ (for\ shooting\ guards)$$

Although, on average the tallest players are positioned as centrals in our sample, after adding the control variable for position, appears that, the height is significant (90%) only in the performance of shooting guards (SGs). Specifically, on average 1 cm taller athlete playing as a shooting guard, is expected to score 21.871 more compared to the other players in the same position.

Another factor is the sample selection and to what extent NBA season 2014-2015 is representative to all the basketball athletes. Even though results suggest strong internal validity, the external validity may be weak, due to selection bias, as NBA players are the best of the best among the players worldwide, which assumes less variation in efficiency scores and body statistics compared to the variation in the whole population of basketball players. Further investigation is needed, if we would like to extend our suggestions.

5. Summary

We analyzed the relationship between efficiency rate during the NBA season 2014-2015 and height of the players. We built linear model between efficiency rating and height using quadratic polynomial and used age as a significant control variable. Intending to introduce further we came across with cofounding biases due to coaches initially positioning players according to their height.

Although our final extended model has a (5.8%) explanatory power, after controlling for position, thus looking for association between height and performance when players are in the same position, we figured out that the only position where height is associated to higher efficiency is the shooting guard. We believe that this a great example which demonstrates why correlation does not always imply causation.