Defending the Three: What traits make an NBA player an effective 3-point defender?

Ben Thorpe

Introduction

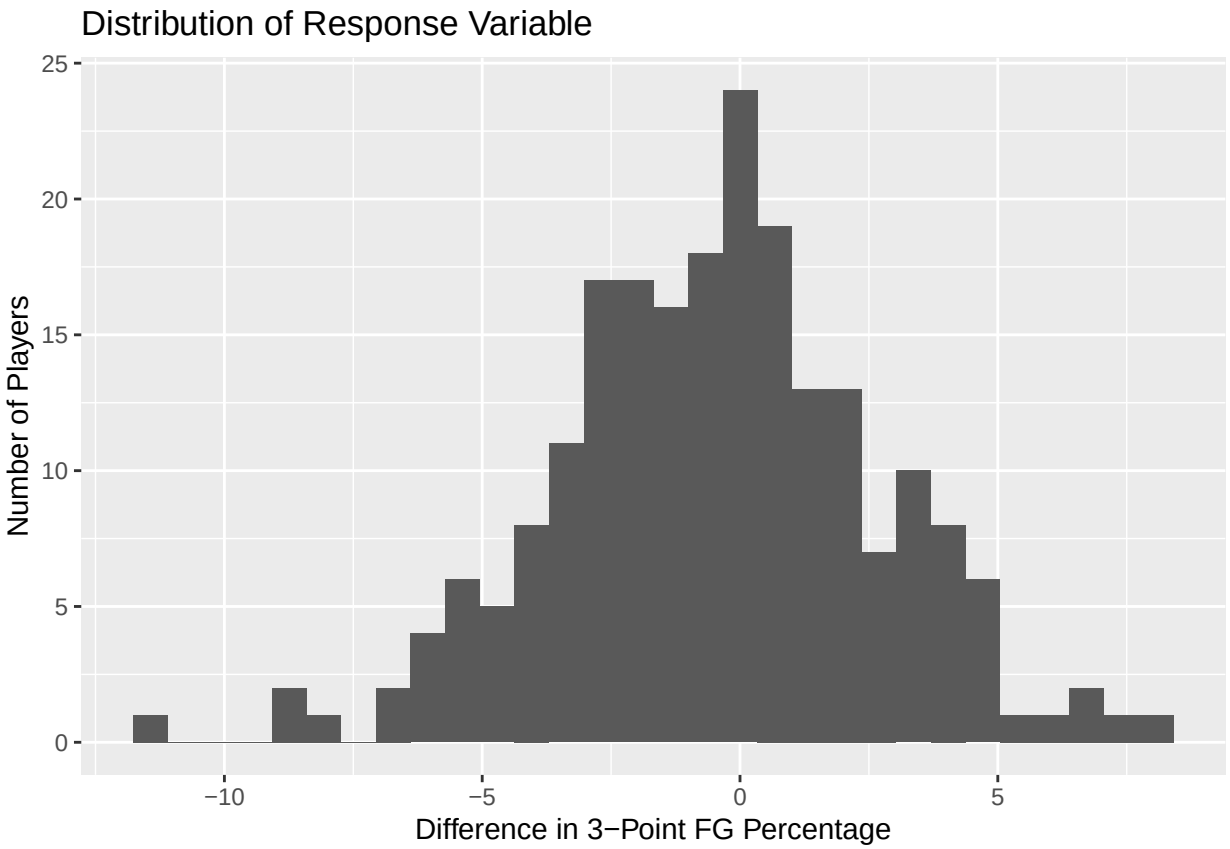
With the surge in importance of three-point shooting in the National Basketball Assoiciation (NBA) over the last decade or so, most teams have made creating open threes a focal point of their offensive system. Thus, the ability for teams to limit their opponents to a low three-point percentage has become paramount in today's NBA. But what kinds of players influence this statistic the most, and what traits do they have in common? This is the question I will be attempting to answer with my project.

For my data, I have gathered a few tables from the nba.com/stats website. I first copied the table containing advanced defensive statistics of individual players at the three point level from the 2021-2022 regular season into a csv, which includes the response variable for this study, the difference in guarded three point percentage (further description will be in the data section). I then repeated the same process with another similar dataset, except this one contained team-level three-point defense data as opposed to player level statistics. Lastly, I obtained measurables from the NBA combine (the pre-draft workout put on by the NBA) to act as level one predictors, as I am interested in whether athletic traits such as speed, quickness, and wingspan have a statistically significant effect on the response variable.

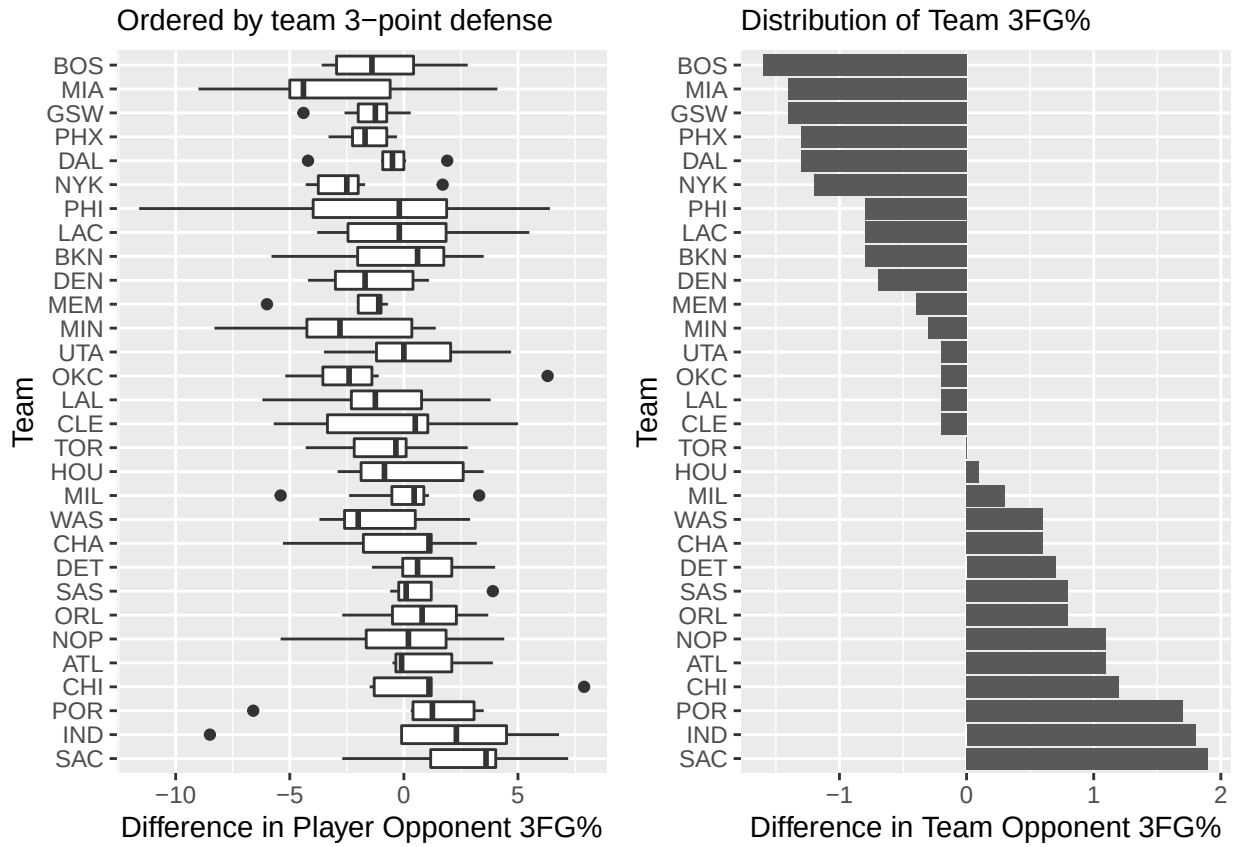
There has been some previous work done on this topic, most notably by Seth Partnow. Partnow, a data analyst who was the director of basketball research with the NBA franchise the Milwaukee Bucks and now works for sports analytics company StatsBomb, wrote an article just over 7 years ago on how player 3-point defense data "is so noisy that it appears to be meaningless." However, there are some differences between his article and my analysis. In his study Partnow was focusing on the variability of individual 3-point defense whereas I am more interested in the predictors which may influence this statistic. Even with a noisy response variable it may be possible to learn about the degree to which certain factors influence it. Also, due to the multilevel model structure I am using hopefully some of this variability Partnow described can be account for. There will be much greater detail on this process in the methodology section.

Data

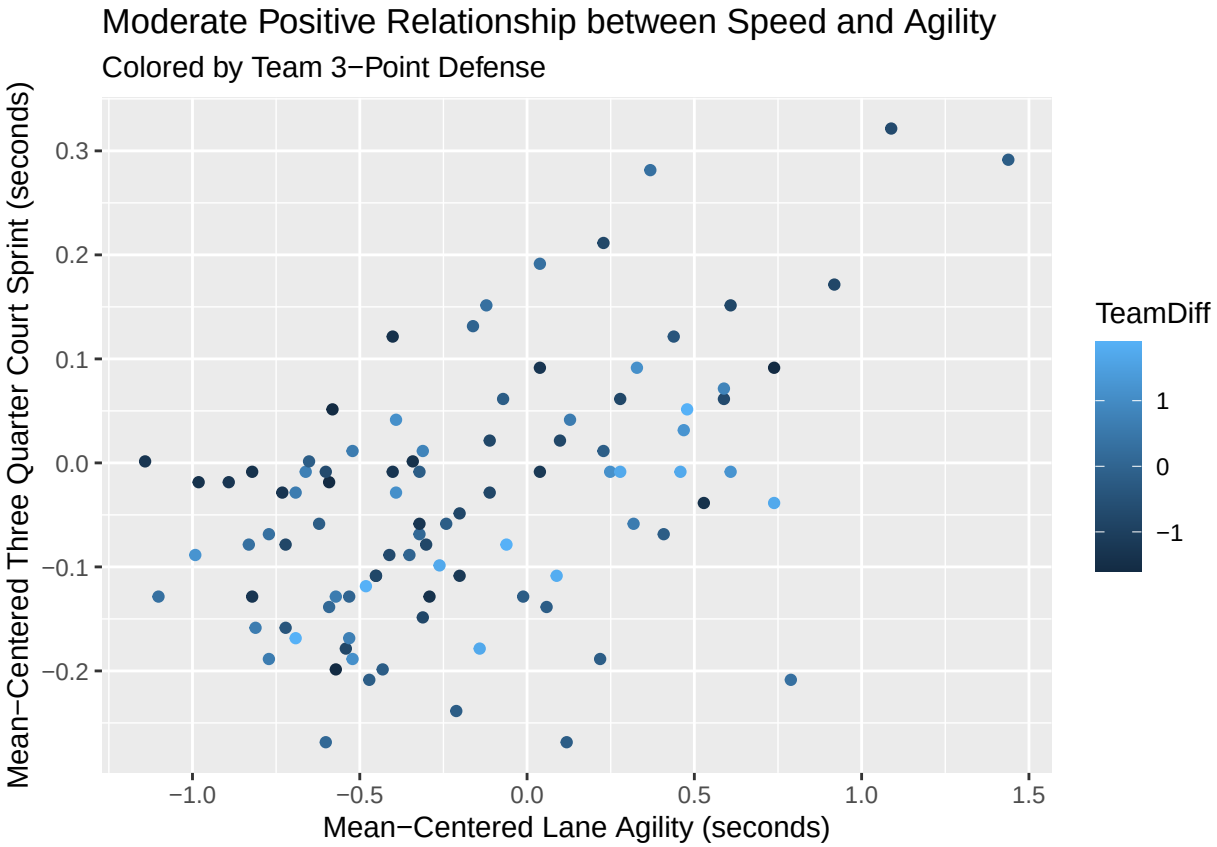
After combining the player three-point defense data, team three-point defense data, and player combine data (which contains metrics such as height, reach, and weight), the final dataset is ready for analysis. The players who were used as observations in the creation of the model played a minimum of 41 games (at least half of the regular season) and defended at least 3 three point attempts per game. The former condition was used to ensure the players selected were important contributors to their team and the latter was to make certain that the sample size of three point shots defended was sufficient for analysis. The qualifier of 3 was chosen arbitraily, though it did seem to remove most players who may have played in many games but did not play many minutes. For my EDA, I will go through some of the key variables looked at in the study and describe why they are important to the analysis of individual three point defense in the NBA. Any variables used in the final multilevel model not addressed here will be defined in the codebook in the readme of the data folder on the project's Github page.



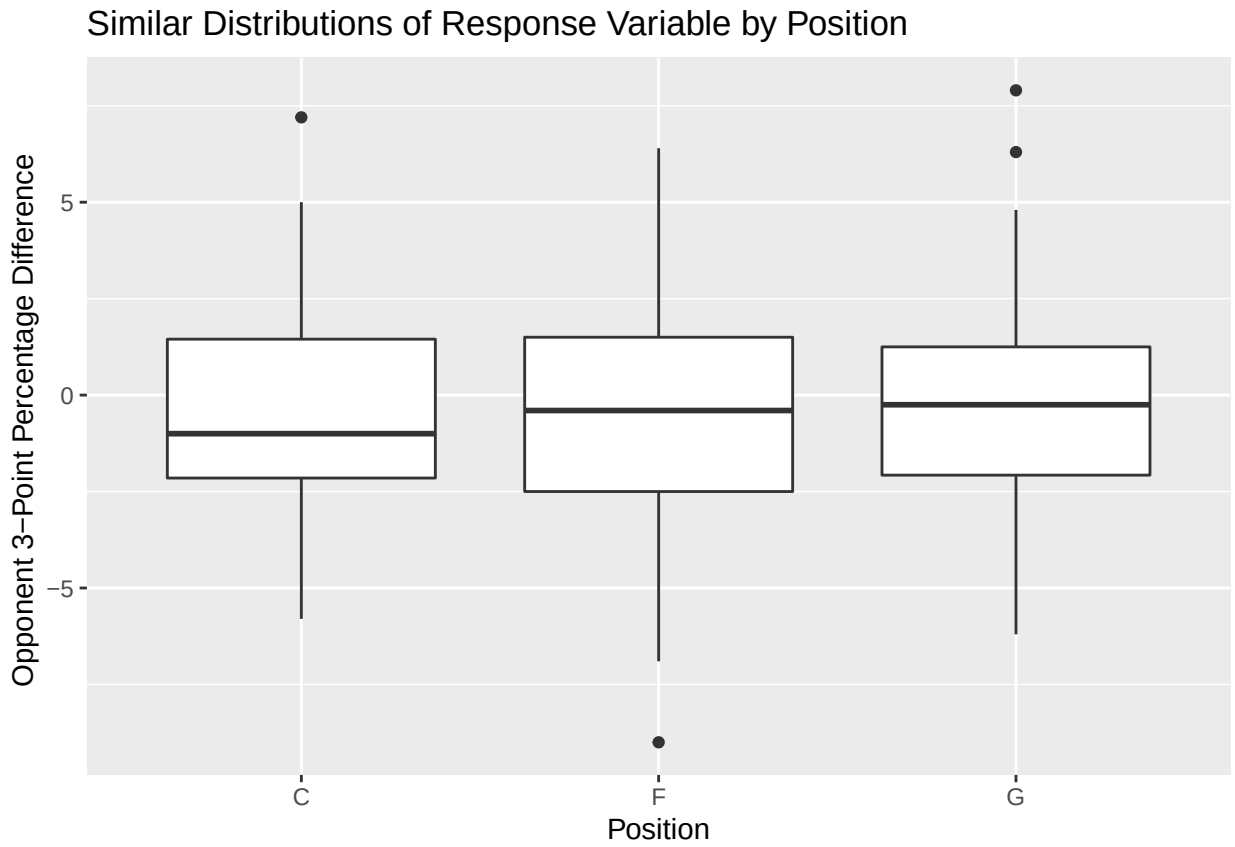
We can see from the above histogram that the response variable of "difference in 3-point field goal percentage" has an approximately normal distribution. This variable is calculated by taking the percentage shot on three point attempts when a certain player is defending and subtracting the opponent's season-long three point percentage from this value. In other words, it is the difference between the average 3-point percentage of all opposing players who have taken 3-pointers when a player is the primary defender and the percentage of threes made when a player is the primary defender. For example, if a player's opponents shot 33% from three when that player was guarding the shooter and the opponents average 3-point shooting percentage across the whole season was 35%, than the player's `diff` would be -2. Thus, defenders who limit their opponents to lower 3-point shooting percentages and are considered better at 3-point defense would have larger negative values for the response variable and vice versa for those who are poor at 3-point defense. Thus, if opponents shoot worse than expected from three against a certain player this will be negative, and if they shoot better than expected it will be positive.



The two previous charts demonstrate why a multilevel model approach is applicable for studying 3-point defense. Here, difference in player opponent 3FG% is the response variable (meaning in previous paragraph) and the difference in team opponent 3FG% is the difference between the average 3-point percentage of all opposing players who have taken 3-pointers against a certain team and the percentage of threes made against a certain team; If opponents shoot worse than expected from three against a certain team this will be negative, and if they shoot better than expected it will be positive. Depending on their player personnel and defensive philosophy, teams defend the three point line differently and to varying degrees of success. The scale is smaller on the team level compared to the player level (difference only goes from about -2 to 2), however this is due to the much larger sample size of threes taken against a team as opposed to against a single player. Furthermore, we can see that players on a good 3-point defending team have good 3-point defense themselves, on average, but this is not always the case, as there are multiple outliers on certain teams as shown by the box plots. Thus, it makes sense to add a player's team's three point defense as a level two variable in the composite model, since team 3-point defense should be correlated to player 3-point defense.



Above is an example of a graph I used to determine which interaction terms to possibly use in the model. In the chart I looked at the relationship between results from a combine agility test and a three-quarter court sprint, since I thought quickness would be correlated with speed. The scatter plot demonstrates that there is a moderate positive correlation between these predictor variables so I will look at including an interaction term between them in the model selection process. I also colored the points by team 3-point defense to figure out if it had an effect on this relationship. Based on the plot, it does not seem to influence the pattern between agility and speed. This makes sense considering we would not expect the team a player is on (or how much they prioritize three point defense) to affect the athletic traits a player has, and this is why the level two variable of team 3-point defense will only be impacting the intercept in the composite model, rather than all of the level one athletic-based predictor variables as well. Using a very similar process I made two additional charts which helped me determine I should potentially include interaction terms between height with shoes and standing reach and standing vertical jump and maximum (or running) vertical jump, respectively, as well. These charts can be seen in the appendix.



Another component of the EDA looks at how a player's position might affect their 3-point defense. Here C is for centers, F is for forwards, and G is for guards. If this terminology is new, these positions generally describe the height of the player, with centers generally the tallest, followed by forwards, and lastly guards often being the smallest players on the court. The box plot above shows us that each distribution is fairly similar to the other two, so it does not seem necessary to include this categorical variable as a predictor in the final model.

	Mean
LaneAgility	11.232
ThreeQuarterSprint	3.243
StandingVert	30.312
MaxVert	36.255
HandLengthInches	8.745
HeightWithShoes	78.583
StandingReach	102.966

Lastly, here are the mean values of the predictor variables which have been mean-centered for analysis. Lane agility and three quarter court sprint time are measured in seconds, while standing vertical jump, maximum vertical jump, hand length, height with shoes, and standing reach are all in inches. The reason for the mean transform is described in the methodology section.

Methodology

I am fitting a two-level model since 3-point defense in a player level statistic but we would expect this to be affected by their teammates defensive ability and overall team defensive strategy as discussed above. Thus, the difference in 3-point percentage on the player level is the response variable (which has been shown to have a fairly normal distirbution in the EDA), individual players are the level-one observations, athletic measurables are the level one predictor variables, a player's team is the level two observation, and the level two predictor variable is the team's difference in 3-point percentage. The team level effect does not affect a player's athleticism or body makeup, so the level two variable will only affect the intercept of the composite model. Three interaction terms among level-one variables will be included as stated in the data section. The only decisions I had to make in my model selection process was regarding the interaction terms, as the other predictor variables essentially chose themselves based on the available dataset. Among the potential combine measurables available, only agility, speed, jumping ability, hand length, and reach could have an impact on 3-point defense from a logical standpoint. Agility and speed affect a defender's ability to quickly reach and close out on opposing shooters while hand length, height, and reach can affect the effectiveness of their closeout once near the shooter. Therefore, variables measuring these skills were included in the model. A player's position was considered as an additional categorical input to the model, however the distribution of the response variable was very similar when looking at each position separately (as shown in the EDA). Each of the level one predictor variables were mean-centered so the intercept can make sense in context and can be interpreted easily in the model output.

I used an ANOVA test to determine if I should include the interaction terms mentioned earlier in my model. The results are shown below, where `model0` contains the three interaction terms along with the level one predictors and the level two team 3-point defense predictor and `model1` contains the same minus the three interaction terms.

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
model1	11	404.175	429.957	-191.088	382.175	NA	NA	NA
model0	14	406.018	438.831	-189.009	378.018	4.158	3	0.245

The p-value from the chi-squared test (0.245) is much greater than the significance level of 0.05, so we cannot conclude that the interaction terms are meaningful in the context of the model. Thus, based on these results, the final model will not include the interaction terms. Below is the mathematical notation of the final model is shown.

Final model equation

Level One

$$\begin{gathered} Diff_{ij} = a_i + b_i(LaneAgility)_{ij} + c_i(ThreeQuarterSprint)_{ij} + d_i(StandingVert)_{ij} \\ +e_i(MaxVert)_{ij} + f_i(HandLengthInches)_{ij} + g_i(HeightWithShoes)_{ij} \\ +h_i(StandingReach)_{ij} + \epsilon_{ij}, \\ \epsilon_{ij} \sim N(0, \sigma^2) \end{gathered}$$

Level Two

$$a_i = \alpha_0 + \alpha_1 TeamDiff_i + u_i, u_i \sim N(0, \sigma_u^2)$$

Composite Model

$$Diff_{ij} = \alpha_0 + \alpha_1 TeamDiff_i + u_i + b_i(LaneAgility)_{ij}$$
$$+c_i(ThreeQuarterSprint)_{ij} + d_i(StandingVert)_{ij}$$
$$+e_i(MaxVert)_{ij} + f_i(HandLengthInches)_{ij} + g_i(HeightWithShoes)_{ij}$$
$$+h_i(StandingReach)_{ij} + \epsilon_{ij},$$
$$\epsilon_{ij} \sim N(0, \sigma^2), u_i \sim N(0, \sigma_u^2)$$

Results

effect	group	term	estimate	std.error	statistic
fixed	NA	(Intercept)	0.034	0.480	0.070
fixed	NA	LaneAgility	0.461	0.981	0.470
fixed	NA	ThreeQuarterSprint	9.856	4.366	2.258
fixed	NA	StandingVert	0.300	0.223	1.343
fixed	NA	MaxVert	-0.132	0.225	-0.585
fixed	NA	HandLengthInches	-2.139	1.163	-1.840
fixed	NA	HeightWithShoes	-0.092	0.316	-0.291
fixed	NA	StandingReach	-0.020	0.232	-0.085
fixed	NA	TeamDiff	1.603	0.401	4.001
ran_pars	Team	sd__(Intercept)	0.894	NA	NA
ran_pars	Residual	sd__Observation	2.959	NA	NA

The results from the final model has some surprising elements, however there are a few aspects of it that were expected. Mainly, the effect of team 3-point defense is a statistically significant predictor of an individual's 3-point defense, with a relatively high t-value of about 4. This was hypothesized earlier in the paper due to the common style of defense teammates must adhere to when on the court together. The positive coefficient estimate here tells us that when a player is on a team, that plays good 3-point defense it increases the likelihood that the player themselves will be a good 3-point defender when using our response variable to measure this concept. We see that for every one percent increase in a player's team's opponent 3-point field goal difference, we would expect the player's opponent 3-point defense percentage to increase by about 1.603, holding all else constant. Also, sprint speed has a relatively large t-value of around 2.26, meaning that this is a fairly significant predictor of individual 3-point defense as well. The model exhibits that for every 0.1 second increase in time it takes a player to sprint three quarters of a basketball court, we would expect the player's opponent's 3-point percentage to increase by 0.986, holding the other predictor variables constant. Lastly, the intercept of 0.034 makes sense in context. It means that for a player who with an average lane agility time, three quarter court sprint time, standing vertical jump, maximum vertical jump, hand length, height with shoes, standing reach, and a team 3-point field goal differential of zero percent, we would expect their individual 3-point field goal differential to be around 0.034% (all predictor variable averages are not listed here as they are shown in the data section). This number is very close to zero, which aligns with the idea that an average NBA athlete would be about average at 3-point defense, which lends credence to the argument that athletic traits would influence this variable. There are no other predictor variables that have a t-value greater than two. This was unexpected, as I originally believed that the other athletic measurables, specifically standing reach and maximum vertical, would be significant predictors of the response variable. However, both had t-values relatively close to zero.

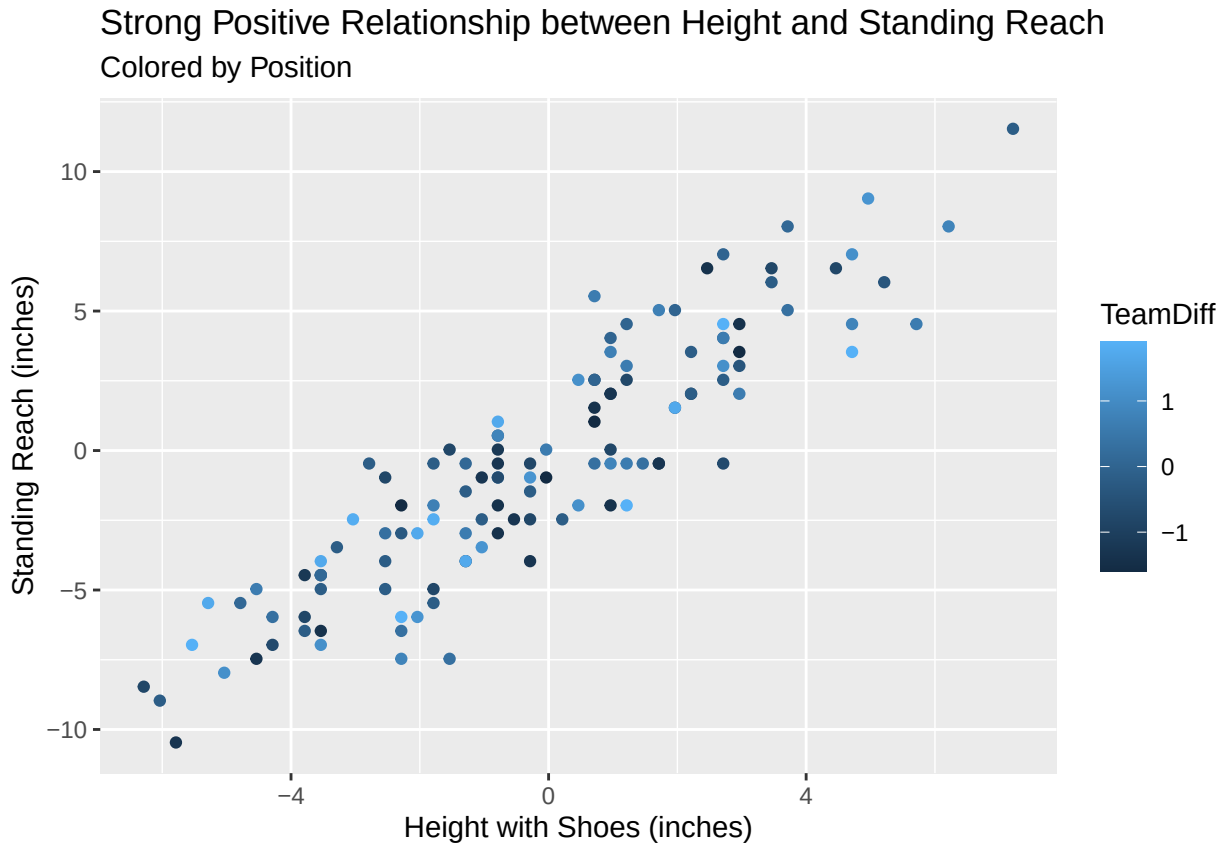
Discussion & Conclusion

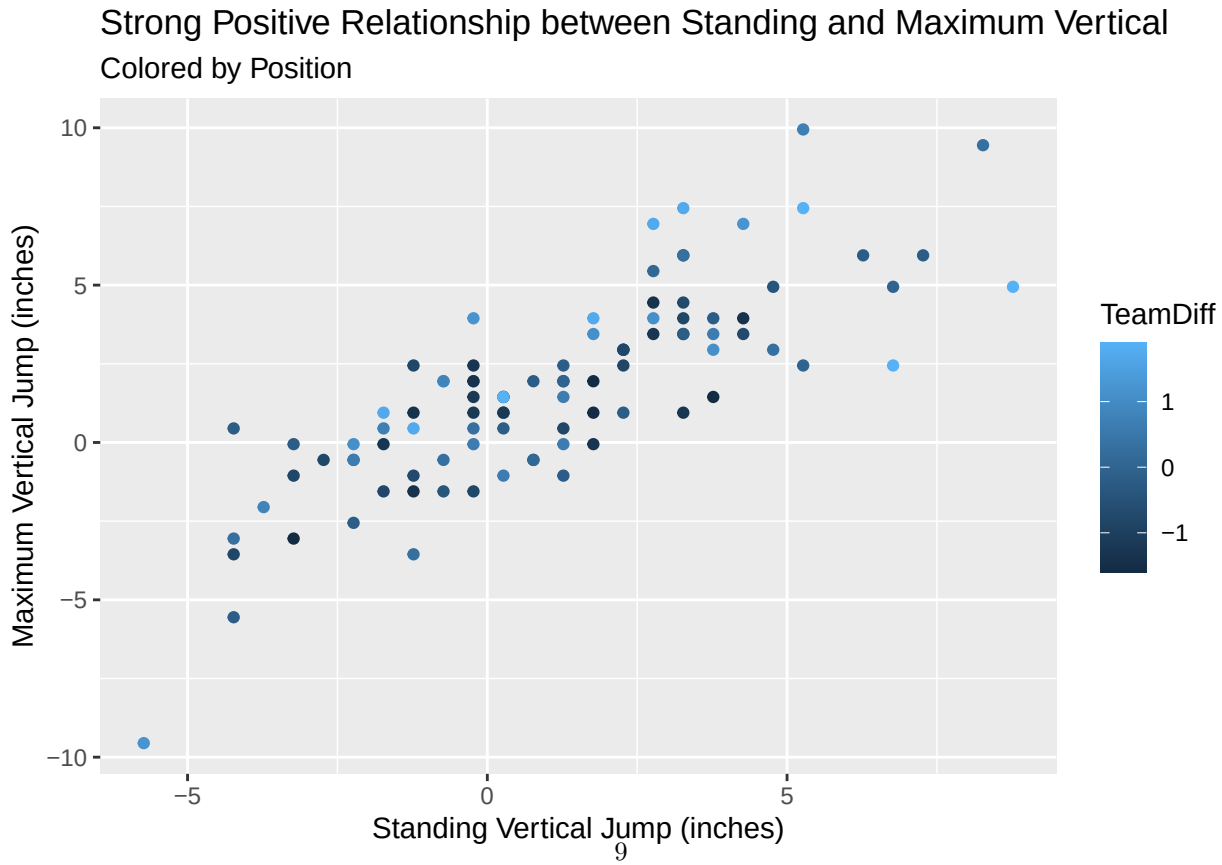
There are two main takeaways from the results of the multilevel model. One is that team has a significant effect on the effectiveness of a player's 3-point defense. This finding applies particularly to when NBA teams are looking to sign free agents or trade for players, as it demonstrates that it is important to look at an individual's defense against the three in context of their team's 3-point defense. In other words, when looking for a 3-point defender and comparing two players with equal opponent 3-point percentage differences, the player from the team with the worse 3-point defense will most likely be better at defending the three on a new team. The second intriguing takeaway is that the model indicates that speed is more important than length in predicting a player's 3-point defense. Intuitively we would expect players with a longer reach would be able to more effectively contest shots, but the model output tells us that the two main predictor variables involving length, height and standing reach, have t-values with a magnitude of below 0.3 and thus do not have a significant effect on 3-point defense. As discussed in the results section, we do see that sprint speed has a relatively large t-value. So based on the model having the ability to quickly closeout on a 3-point shooter makes a larger difference in whether they make the shot compared to how well a player is able to contest the shot with their length.

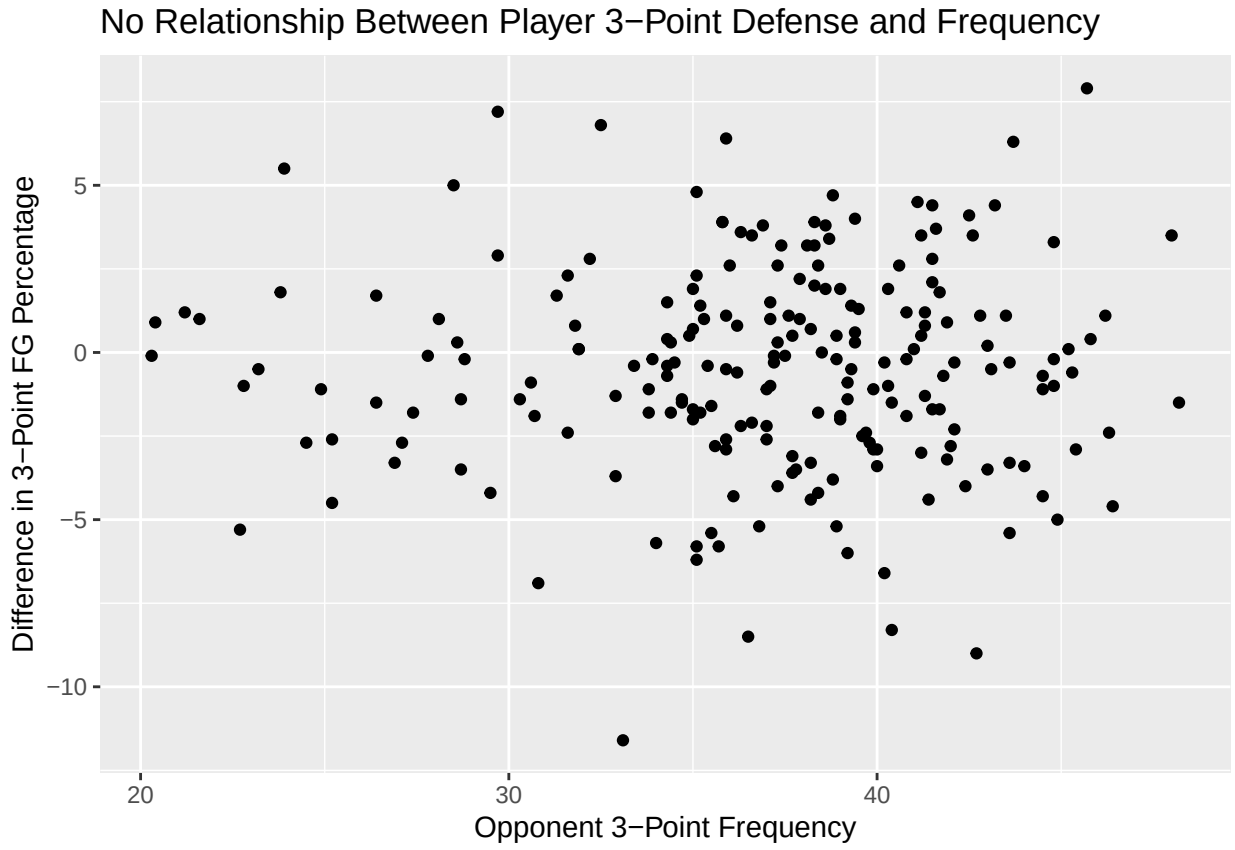
Although there have been previous studies concluding that the difference in opponent 3-point percentage is fairly random from season to season due to random variance in shot making, the takeaways from this project are still practical and contain meaning. The multilevel model approach is more meant to find statistically significant predictor variables and address the impact of team defense on individual defense rather than be used to predict the response variable in this study. The findings provide a good idea of what athletic traits NBA general managers can focus on when scouting players who can help defend the three, either in free agency or the draft.

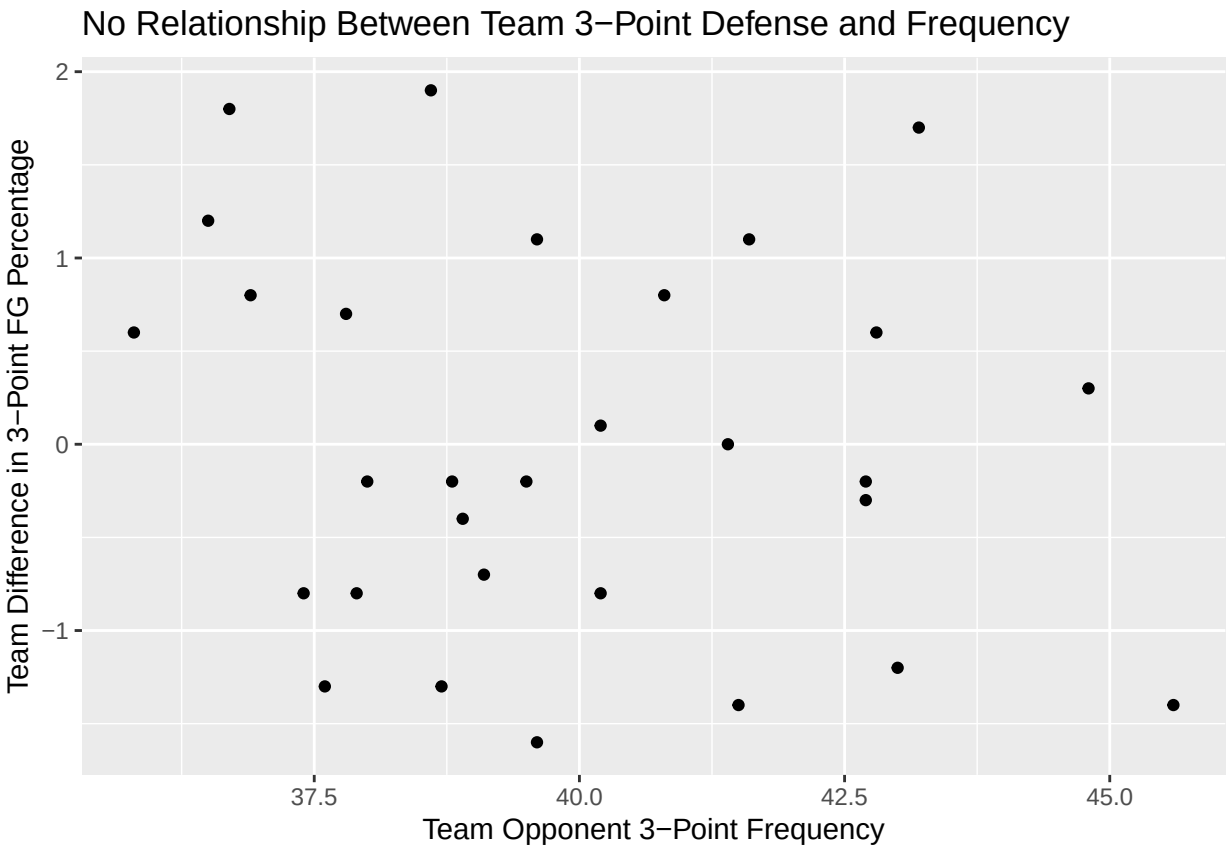
One limitation from this study is the relatively small sample size. This sample of players was taken from the most recent regular season and only those whose combine statistics were recorded were used for model creation to allow for easier analysis. Perhaps using the data available some of the missing data could be imputed using additional models in the future to enable more observations to be used in analyzing the predictor variables from this study if further work was to be done on the subject. Another limitation is that it is expected that some of the level one predictors such as agility and speed are expected to decline with age and/or injury, and this is not taken into account in the model since these effects are vary drastically on a player-by-player basis. However, NBA teams have the ability to bring in potential signings for workouts so these traits can easily be measured at any time, so this more poses an issue to future research on the subject by the public rather than in practical use NBA general managers and scouts. Further study might include expanding the scope of the project to identifying whether these same athletic abilities that were statistically significant in this study are also important in determining a team's three point defense. This research question might be: What characteristics are common among players who play for the best 3-point shot defending teams and which players around the league have the greatest impact on opponent three point percentage difference? This would involve taking the analysis from this project and turning it from a player-centric to a team-centric approach to measuring 3-point defense.

Appendix


Strong Positive Relationship between Height and Standing Reach
Colored by Position
Standing Reach (inches)
10
5
0
−5
−10
−4
0
4
Height with Shoes (inches)
TeamDiff
1
0
−1


Strong Positive Relationship between Standing and Maximum Vertical
Colored by Position
Maximum Vertical Jump (inches)
10
5
0
−5
−10
−5
0
5
Standing Vertical Jump (inches)
TeamDiff
1
0
−1


No Relationship Between Player 3−Point Defense and Frequency
Difference in 3−Point FG Percentage
5
0
−5
−10
20
30
40
Opponent 3−Point Frequency


No Relationship Between Team 3−Point Defense and Frequency
Team Difference in 3−Point FG Percentage
2
1
0
−1
37.5
40.0
42.5
45.0
Team Opponent 3−Point Frequency