Assignment #5: Regularized Regression

Introduction

A team's ability to efficiently shoot the ball is paramount to success in the NBA. Even if a team shoots the ball substantially more than the opposition, if that team consistently shoots poorly, whether through poor shot selection or general poor shooting ability, it becomes much more difficult to find success offensively.

On the other side of the ball, the ability to force teams to shoot inefficiently, through either strong shot contests or proper area denial which forces offensive teams to shoot shots at the shot-clock, is just as important for teams to find success in the NBA. It does not matter how well a team does on the offense if they allow the opposition to score just as much, if not more, on the other side of the court.

While shooting metrics like True Shooting Percentage(TS%) and Effective Field Goal Percentage(eFG%) attempt to capture a player's shooting efficiency and defensive metrics like Defensive Rating attempt to quantify a player's defensive ability, those measurements do not account for the other team's shooting or defensive ability. A good shooter may have a low True Shooting Percentage because he is simply facing a very good defender, and a good defender may have a very low Defensive Rating because he is facing a very strong shooter.

Methodology

To account for the skill of the opposition, this report looks to estimate the shooting ability and shot defense ability of all players during the 2020 - 2021 NBA season using a Regularized Rasch Model, where the offensive adversary is the shooter and the defensive adversary is the five defenders on the other team, with an outside variable describing the distance of the shot, where each shot is placed in bins used by the NBA Statistics Website(e.g. a shot from 11 feet away would be placed in the 10 - 14 ft bin). I can then predict the odds of a shot going in as:

$$P(score) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

where I calculate η_i as:

$$\eta_i = \beta_0 + \beta_{Shooter} + \sum_{p \in D} \beta_p + \beta_{Distance}$$

with β being the assigned effect on the shot from each variable(i.e. a higher β for the defender will increase the odds of the shot going in, so negative is best) found from ridge regression, where β_0 is the intercept of the regression and D is the set of all the defenders on the court during the possession. Following ridge regression, I got an intercept of $\beta_0 = \mathbf{0.4117}$.

Data

In running the Rasch Model, I took only the event data from the 2020 - 2021 NBA season from a Kaggle Dataset containing play by play data and roster from the 2015 - 2016 to the 2020 - 2021 NBA season. I then found all shot events, filtering out free throws, as well as each shot's shooter, the five defensive players, the distance of the shot, and whether the shot went in or not.

I then used a sparse matrix to one-hot encode player data such that there is are two columns for each player, one describing whether the player is shooting and the other describing whether the player is defending as well as columns which describe whether the shot is within a certain distance threshold. That sparse matrix was then fed into a ridge regression algorithm to generate defensive and offensive βs for each player.

In addition, since the event dataset only had Basketball Reference IDs, I used an NBA Player ID Dataset to get the actual names of each shooter and defender.

Results

Top 5 Shooters and Top 5 Defenders

Using ridge regression, I found the Rasch Model coefficients for every single player who took at least one shot and defended one possession in the NBA, finding the top 5 best shooters and 5 best defenders:

Best Shooters

Player	# Shots	Offensive β
Nikola Jokic	1346	0.0215
Deandre Ayton	861	0.0203
Mitchell Robinson	334	0.0201
Seth Curry	764	0.0194
Jarrett Allen	529	0.0189

Best Defensive Players

Player	# Possessions	Defensive β
Garrett Temple	1991	-0.0139
Jerami Grant	2753	-0.0118
Paul Millsap	700	-0.0114
Rajon Rondo	1251	-0.0105
Semi Ojeleye	747	-0.0102

While notable names like Nikola Jokic, the 2021 NBA MVP, and Deandre Ayton, the starting center for the Suns who made the NBA Finals that year, appear on the best shooters list, we

see some surprises like Mitchell Robinson and Seth Curry. However, it becomes clear that 4 of the 5 players on the best shooters list are centers, which follows, as centers tend to take closer and more efficient shots in the paint, leading the Rasch model to inherently become more biased towards centers, as opposed to point guards, who may more take higher value, but less efficient three-point shots.

The only non-center on the list, Seth Curry, is most likely highly rated by the model because he is able to act as a pure shooter, being able to benefit from his superstar teammate Joel Embiid soaking up defensive position, leading him wide open to make much easier shots, as opposed to a superstar shooter like his brother, Steph Curry, who is usually the target of defensive pressure and is forced to create his own shot.

On the other hand, every single player on the best defensive players list was a role-player that season, meaning that they are non-superstars who might come off the bench. This implies that strong defensive players tend to be bench players who specialize in defense, as opposed to superstars who might specialize only in shooting or other offensive aspects of the game.

The Distance Parameter

As stated in the Methodology section, the Rasch Model also accounts for distance by assigning weights to shots from distances falling within certain buckets.

Distance(ft)	β
$(0, 5]$	0.1965
$(5, 10]$	-0.0014
$(10, 15]$	0.0066
$(15, 25]$	-0.0393
$(25, 30]$	-0.0692
$(30, \infty]$	-0.0983

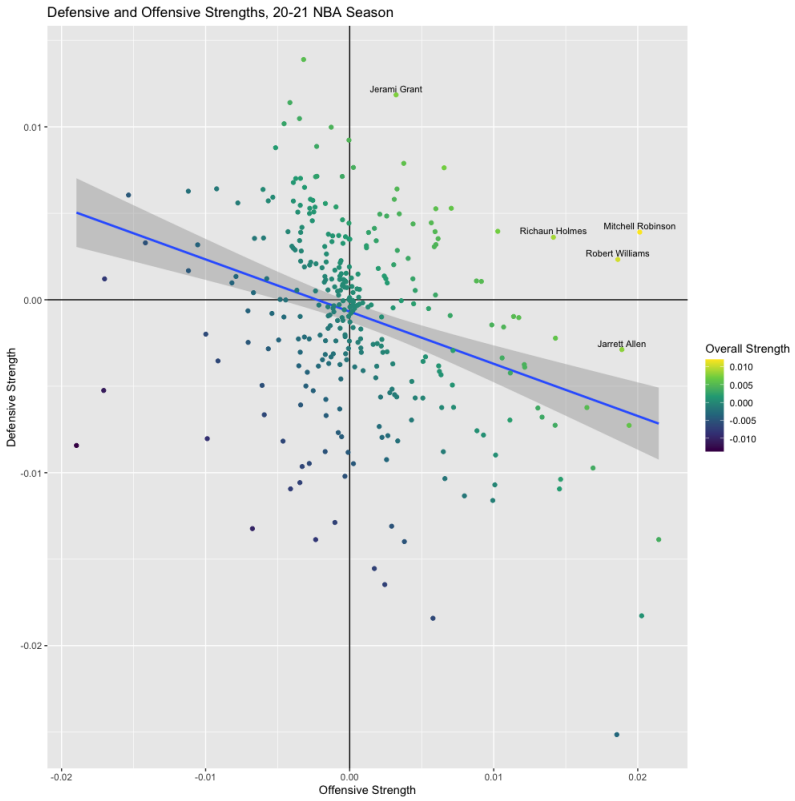
Overall, the coefficients follow the expected trend where farther away shots tend to make the shot more difficult and decrease the odds of a shot scoring. However, there is a notable exception of shots between 10 - 15 feet from the basket. With the slow decline of the midrange[1], this sudden efficiency in midrange shooting can likely be explained by the fact that only players who are excellent midrange shooters are the only players shooting midrange shots at high volumes, as poorer midrange shooters will either stick to three point shots or layups closer to the basket.

Overall Ability

Because players in basketball must be on both defense and offense, I am able to find the defensive and shooting Rasch Model abilities of all players in the NBA and then amalgamate

[1]Kilcoyne, Shawn. *The Decline of the Mid-Range Jump Shot in Basketball: A Study of the Impact of Data Analytics on Shooting Habits in the NBA*. Nov. 2020.

these two statistics by taking the average, yielding the following graph[2], with the overall ability being correlated with a brighter color:



We once again see the top 5 players in overall ability, labeled in the graph, being centers, with the minor exception of Richaun Holmes, who also played as Power Forward. This should be expected as centers tend to shoot closer and more efficient shots, increasing their offensive ability, while their larger size can contest shots much better, increasing their defensive ability.

Furthermore, based on the somewhat negative correlation of $\mathbf{r = -0.328}$ between offensive and defensive ability(with defensive ability multiplied by -1), players tend to either specialize in shooting or defending, focusing less on the other factor, becoming more likely to be very good at either shooting or defending, but not both.

Limitations

However, this model fails to account for differences between positions. Because of their tendency to shoot efficient layups, centers are inherently favored more by the Rasch Model than point guards, with their tendency to shoot longer range shots.

Additionally, the defensive side of the model, as I do not have the on-ball defender for each shot, acts more as a plus-minus model, with defenders being rewarded for having a shot missed on that possession for just being on the court at the same time, not for any action that they take during the possession.

[2]The Defensive Strengths were multiplied by -1 to make the graph more readable.