NBA PLAYERS RANK RESEARCH REPORT

YING YU, JIACHENG GONG, KA YING MOK

December 15, 2017

Abstract

As basketball fans, we are curious how the Player Efficiency Rating (PER), a rating of a player's per minute productivity, is calculated since it is so popular and official to be use by fans or medias. However, the PER formula is not easy to understand, which involves lots of mathematics and statistical methods to implement out this formula, and it still has space to be advanced due to the change of how to measure overall score for basketball players in different positions. With this curious, we start the research and the purposes of this study are following. First of all, we want to understand the PER value based on our research and try to figure out whether Touches and Position affect the ranking system of PER. Then, we combine them with other variables to build an easier and simpler model to predict PER rankings and apply it to different seasons. This study involves the information data of all players in 2013-2014, 2014-2015 and 2015-2016 NBA seasons.

Introduction

John Hollinger uses Player Efficiency Rate (PER) to value a player's statistical performance in a detailed number during one season. This ranking system become the most commonly used method to rate a player or a team after it is accessed by public. However, this formula also has issue with too much favor on offensive performance so that it's not a completely reliable method to get a comprehensive rate for each player. There are only two defensive statistics incorporated into the formula: Blocks and Steals, which appears the unfairness on different positions. Meanwhile, with the hand-checking rule change in 2004, the number of three-point-shot attempts increased significantly instead of keeping plateaued for several years, and the formula of PER trends to be more unreasonable because of its overmuch weight on offensive statistics. Therefore, we are going to test positions and touches with other variables to see if PER rank can be predicted more precisely.

Method

Data resources

The data are searched and extracted from BasketballReference.com [1], and ESPN.com [2]. Information of all players among three seasons (2013-2014, 2014-2015 and 2015-2016) is our focus of this study. The information merged from two websites is used to create a new data frame to use.

Data Cleaning

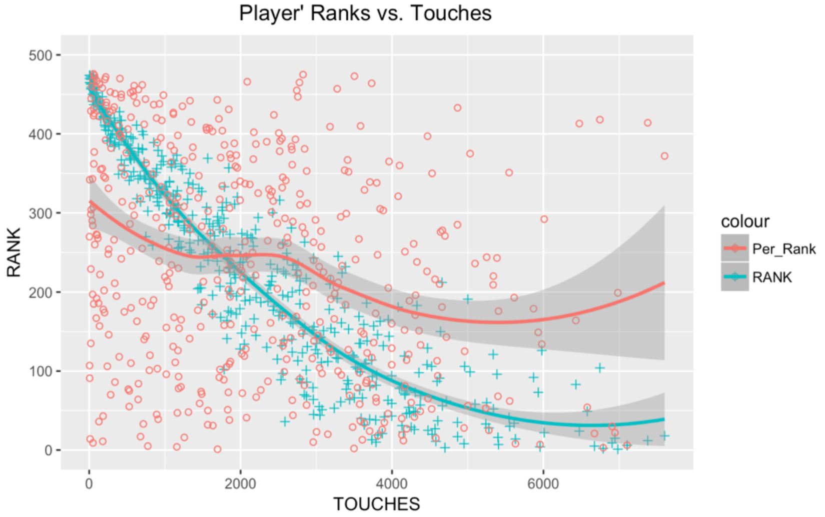
First, data cleaning is applied to three seasons. We keep rows whose PER, Touches, Three Points, Two Points, Free Throw(FT), Assist(AST), Block(BLK) and Steal(STL) are all positive and delete the meaningless rows. Then, we select the key features needed for the ranking system: Player, Touches, Point Per Touch, Field Goal(FG), Field Goal Percentage (FG%), Three Point, Three Point Attempt(ThreePA), Three Point Percentage(ThreeP%), Two Point, Two Point Attempt(TwoPA), Two Point Percentage(TwoPP), Free Throw(FT), Free Throw Attempt(FTA), Free Throw Percentage(FTP), Offense Rebound (ORB), Defense Rebound (DRB), Assist (AST), Block (BLK), Turn Over (TOV), Steal (STL). Then, we combine them with Team, Conference and Position to get a new data frame.

Measures

The methods used in this study include Simple Linear Regression, Multiple Linear Regression and stepwise regression to figure out the relationship between PER and two or more predict variables. Analysis of Variance (ANOVA) is applied to determine statistically significant differences between two or more means of groups. Marginal model Plotting is used to get useful diagnostic information for both linear and generalized linear models, while box-cox power transformation is used for normalizing data and improving the validity of models with normality regression diagnostics. And visualized graphs are used to access the relationship between variables and any notable trends of the dataset.

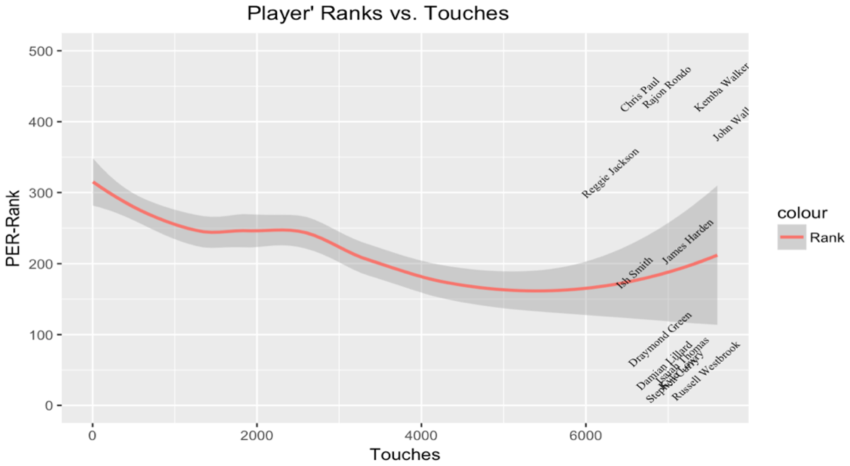
Major Findings

Figure 1: The graph below shows the relationship between ranks (PER Rank & Total Points Rank) and number of touches.



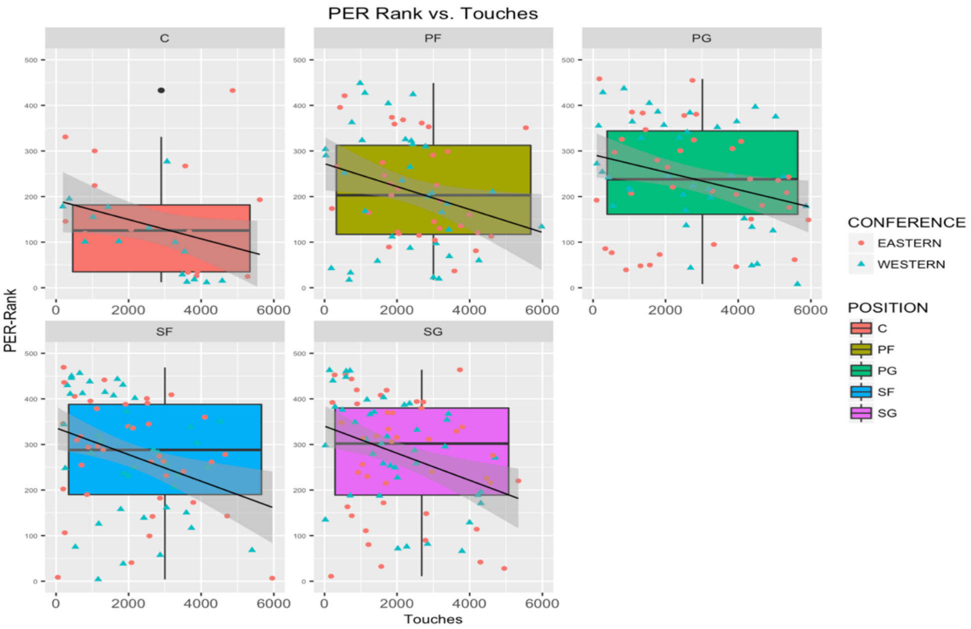
The ranking of player determined by both PER and Total Points have trend with touches. It suggests the higher number of touches will result in better ranking. But the trend of PER rank fluctuates more than points rank does, and it might be caused by some players with high touches and relatively low PER rank for some reason.

Figure 2: The graph below identifies all players with more than 6000 numbers of touches in 2015-2016 seasons



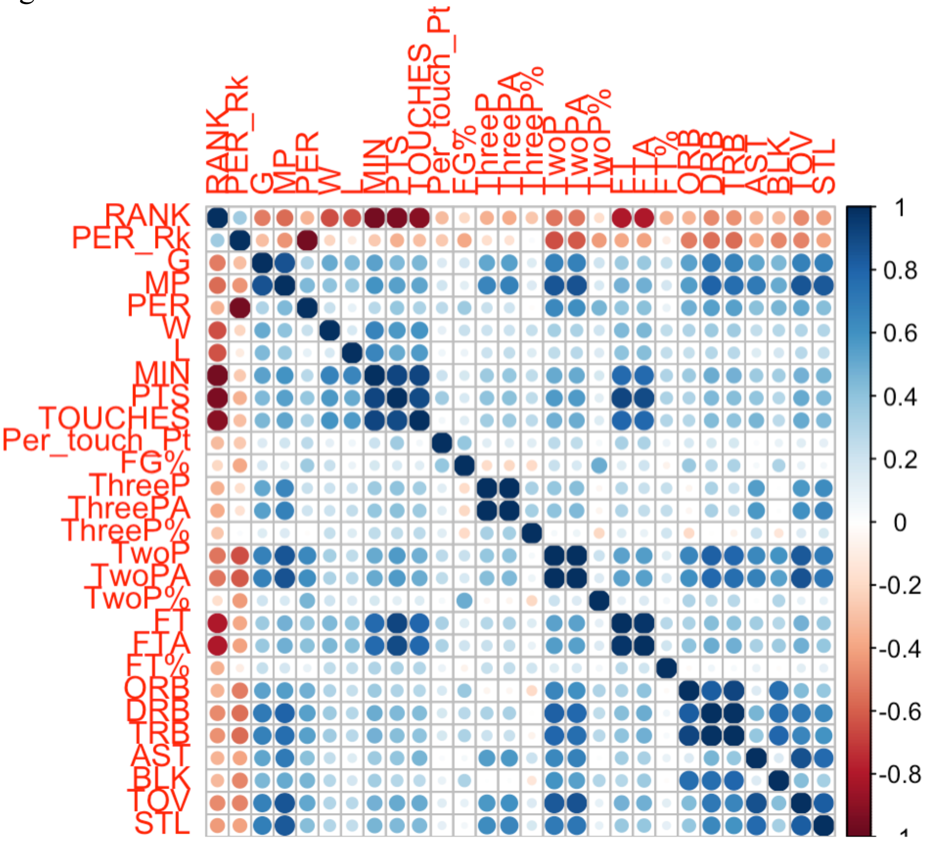
The explanation of high touches with low PER ranks of a player. Take Stephen Curry and Rajon Rondo as an example. Stephen Curry is an all-star player and ranked in top level by PER score. There is only a small difference of touches between these two players in 2015-2016 season, their PER ratings are significantly different. The points per touch of Rajon Rondo is 0.14 while Stephen Curry gets 0.35. Although the touches statistics of Rajon Rondo is as high as Stephen Curry's, the comparatively few points per touches demonstrate that Rajon Rondo has the problem of lacking efficiency when controlling the ball. This finding can be applied to explain the reason of some players who have low rank but with high touches.

Figure 3: The graph below shows the relationship of players' PER rank and number of touches with boxplot, and is separated by five different positions (Center (C), Power Forward (PF), Point Guard (PG), Small Forward (SF) and Shooting Guard (SG)). Blue points represent players from Western Conference and red points represents players from Eastern Conference.



The proposal of this assessment is to figure out if there is any significant relationship between PER Rank, Touches and Positions. The graph above shows that PER rank vs. touches, separating into 5 positions and 2 conferences. It appears that Center position has a better rank compare to other four positions under same touches score level. And the western players have higher PER rank than eastern players do, which consists that the western conference always has better performance on win ratio.

Figure 4: The plot below shows a correlation hear map with variables that were used in building a linear Regression model



Correlations map among variables is used to enhance model. First, we use touches and positions to predict the PER rank and get the R square of 0.21, which means the model with these two variables can only explain 21% of the response variable. So we consider to add more variables to enhance the model by trying whom are highly correlated with PER rank and less correlated with other independent variables for avoiding multicollinearity issue through the correlation map. In the process of determination, we decide to use points instead of points per touches for the reason that the latter variable is contrary to our original intention of balancing the weight of defense and offense. Then, we test the multicollinearity existing in this regression model by Variance Inflation Factors (VIF) to figure out the variables whose VIF are larger than 5 and delete the highest VIF one by one. And we get the model lm (PER Rank ~ Position + Touches + Three Points + Two Points + FT + AST) with all VIF smaller than 5 and R square of 0.5865, which is better than the former model. After that, we utilize the inverse transformation method to transform uniform samples into samples from different distribution, so we decide to use sqrt(PER Rank) instead of raw PER Rank by figuring out the estimated transformation parameter of 0.63, and we get a new R-square increasing to 0.6345. Since the rank must be a positive number, we enforce the model cross the zero by adding 0 into the equation. The output gives a R-square of 0.9642. At the same time, we also try to use powerTransform function to transfer independent variables and get the R-square of 0.5669, and some of variables after transforming are not still significant in the new model. So we determine to use the model which only transform response variable as our final model: lm (sqrt (PER Rank) ~ Position + Touches + Three Points + Two Points + FT + AST+0).

Results

Outcome 1

As the Figure 3 shows, the order of average ranking under the same level of Touches for each position is: Center, Power Forward, Power Guard, Shooting Forward, Shooting Guard, and the higher order means better PER performance with the same level of Touches. And it also suggests the slope of Power Forward and Shooting Forward are smaller than other three positions.

Outcome 2

For the linear regression after transformation and enforcing the model cross the zero, all of independent variables, including Position, Touches, Three Points, Two Points, Free Throw, Assist, appear to be statistically significant. And the test of multicollinearity using VIF function shows the VIF of all predictors are smaller than 5, indicating that they are not highly correlated to each other. Moreover, even the ranking system of PER does not conclude position as an independent variable, every positions is significant. While the coefficients of Three Points, Two Points, Free Throw, Assist are negative, it indicates that the player will get higher ranking if his performance is better on these statistics, and the coefficients of Two Points are almost two times of others, which shows getting points has a relatively large proportion in PER ranking. However, the variable Touches gets the coefficient of 0.0006736, which is unexpectedly positive and almost close to zero, and it demonstrates the calculation of PER is less relative to Touches and a high number of Touches is not equal to good efficiency. Finally, the output of summary shows that the model can explain 96.42% of 2015-2016 season’s data and it meets out expectation.

Outcome 3

After building the model based on the data of 2015-2016 season, we apply the same model on the data of 2013-2014 season and 2014-2015 season. Except the p-value of Free throw in 2014-2015 season is 0.051850, which is a little bit larger than 0.05, every other variable is statistically significant. The R-square of each season is larger than 0.96, and it meanwhile proves the model is applicable to the PER ranking in past years. The model can also be used to predict the future rank of players, which can give a guide of player trading among different teams and adjust players' positions to perform better during the game.

Discussion

We start our study by assuming the position and touches will affect the PER rank and create a simpler model by combining these two variables and the original predictors used in the PER formula. We apply the model based on 2015-2016 season's data to the other two NBA seasons to validate our findings. And we conclude that PER rank can be predicted in a much easier way. Moreover, it is recommended that the separation of assisted and unassisted field goal be taken into higher consideration because of its significant effect on valuing a player's productivity. In our study, we do not distinguish between an assisted and unassisted field goal, but it might make a significant difference since player who have ability to create their own shots are entirely different than players who can catch and shoot. The players of the former type change defensive schemes; they draw more attention and break down defenses. We can use Derrick Rose as an example, who had more unassisted field goals than any players in the 2015-2016 season. When he moves towards the rim, defensive implode to stop him, leaving his teammates with wide-open jumpers so that Rose passes the ball to his teammates, and thy are 6%

more likely to make the show than when he does not, which proves the value of a player being able to create his own shot. In order to value a player comprehensively, we need to consider the distinguish between assisted and unassisted goal in our future study because we predict that players with high unassisted shots would have the ability to penetrate and draw more fouls in the games.

Limitations

Due to limited to NBA data, there are only three palpable defensive measurements (Block, Steals, Defensive Rebounding Percentage) are open to public. Even though ESPN's DRPM (Defensive real Plus-Minus) goes deeper, attempting to ascribe a basic plus-minus number to a player, while also taking into account the teammates and opponents who share the court with him. However, publicly available DRPM data only the 2017-2018 season. Moreover, obscure defensive performances are hard to count. For example, set screen is a block that player does to help team members, but those data are lack of tracking. The lack defensive measurements limited the scope of our analysis by failing to quantify a player's relative value and productivity, which is a significant obstacle in scoring the efficiency of players. Meanwhile, due to the reason that the PER formula doesn't have very detailed implement online for public to learn and do research, there is some limitation when we use PER Rank as our response variable.

References

[1] Hollinger, J. “What is PER?” *ESPN.com*, ESPN.COM, 08 August, 2011, Web. 20 October 2017

[2] “NBA 2013-2014 Season.” *Basketball-Reference.com*, N.p., n.d. Web. 20 October 2017.

[3] “NBA 2013-2014 Season.” *EPSN.com*, N.p., n.d. Web. 20 October 2017.

[4] “NBA 2014-2015 Season.” *Basketball-Reference.com*, N.p., n.d. Web. 20 October 2017.

[5] “NBA 2014-2015 Season.” *EPSN.com*, N.p., n.d. Web. 20 October 2017.

[6] “NBA 2015-2016 Season.” *Basketball-Reference.com*, N.p., n.d. Web. 20 October 2017.

[7] “NBA 2015-2016 Season.” *EPSN.com*, N.p., n.d. Web. 20 October 2017.

[8] Sheather, S. J. *A Modern approach to regression with R*. New York: Springer. 2010. Web. 30 October 2017