Donald Vu

Player Performance Prediction

Research Question

We were interested in how we can predict the performance of NBA basketball players, specifically the number of points that a player would score per game. Because there are dozens of available statistics used to evaluate NBA players, we questioned which ones were most predictive of points per game. Therefore, we created a multivariable regression model that can predict the points per game of an NBA player (PS.G) using the variables Field Goal Attempts (FGA), Three Pointers Made (X3P), Free Throws Made (FT), and Minutes Played (MP). Understanding how these variables relate to scoring will allow coaches and executives to make decisions about playing time and player acquisition.

Data Preparation

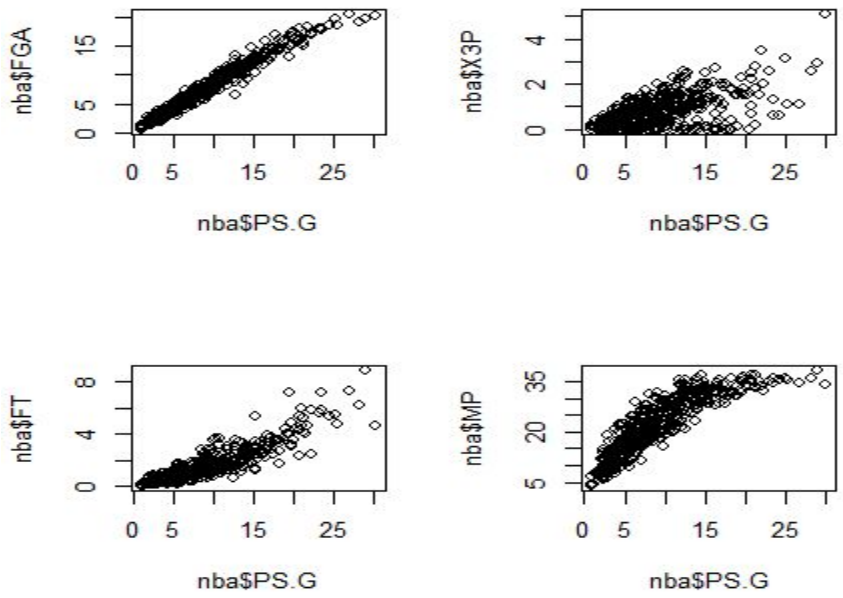
In order to collect our data, we used basketball-reference.com. They have numerous statistics for every active and former player in a per-game format, so we used this as our data set (http://www.basketball-reference.com/leagues/NBA_2016_per_game.html). The original data set included 476 players with 30 different variables over the course of the entire 2015 - 2016 NBA season. Except for the "Team" which is a categorical variable, all variables are numerical.

Data Cleaning

In order to get more accurate data, we created a subset that eliminated any player that appeared in fewer than 20 games. Because the NBA season is 82 games long, any player who failed to play in 20 would not offer much statistical significance in our model. After applying this filter, our sample size was 468 players. After reviewing our data set, the only "NA" value that appears occurs in the "X3P." category, which measures the percentage of 3-point field goals that are successful. Please note the period in "X3P." This denotes that it is three-point percentage, not three-pointers made. This only occurs if the player did not attempt a three-point field goal all season. We do not use this variable in our model, so it was not a factor for us.

Data Analysis

We used four variables as predictors: Field Goal Attempts, 3-point Field Goals Made, Free Throws Converted per Game, and Minutes Played to predict Points per Game. We ran both individual regressions and multiple variable regressions to try and see if the R output would lead to statistical and predictive significance. The individual results can be seen below.



The first variable was "FGA." This shows the total number of field goals that were attempted per game for each player. We chose this variable because it directly affects scoring by increasing the number of chances to score. The mean number of field goals attempted was 7.41. The median was 6.40. The standard deviation was 4.23. While this evidently is a very volatile variable, it is essential to our model. The R^2 value between PS.G and FGA was 0.9627. This is our strongest predictor, as increasing the number of shots will most likely increase the number of baskets scored.

Our second variable was "X3P", which denotes the number of three-point field goals made per game. We chose this variable because converting three-point instead of two-point field goals increases the player's points per game at a higher rate. The mean number of 3-point field goals made was 0.75, the median was 0.60, and the standard deviation was 0.70. The R^2 value between PS.G and X3P was the weakest of our four variables at 0.3455. This is because three-point field goals are converted at a lower rate than standard field goals. In our sample, the NBA field-goal percentage was 44.8% while the NBA three-point percentage was 35.2%. Therefore, the importance of three-point field goals varies drastically by player.

Our third variable was "FT," which denotes the number of free throws converted per game. We chose this variable because physical players rely on free throws to score when they are fouled. The mean number of free throws was 1.52, the median was 1.15, and the standard deviation was 1.29. The R^2 value between PS.G and FT was good, at 0.7793. This reveals the implicit relationship that a player with a strong free throw percentage is also a skilled shooter that will convert shots at a higher rate.

Our final variable was "MP," which denotes the minutes played per game. We chose this variable because a player must be in the game to score. The mean number of minutes played was 21.43, the

median was 20.7, and the standard deviation was 8.32 minutes. The R^2 value between PS.G and MP was also good, sitting at 0.7690. This is natural, for a player who spends more time on the court will have more opportunities to score.

Combining these four variables by using the lm() function of R allowed us to form an equation to model our findings. We can predict points scored per game through the equation:

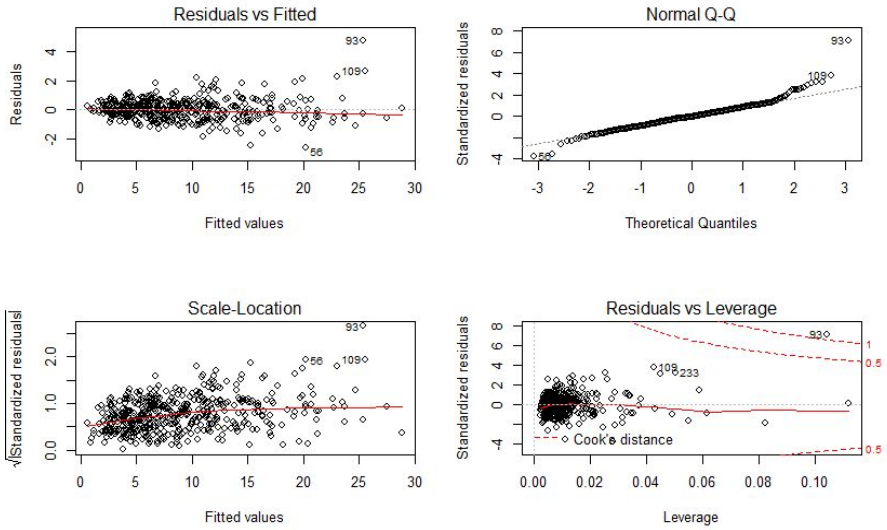
$$\text{Expected } PS.G = -0.3119 + 0.9153FGA + 0.2460X3P + 1.0576FT + 0.0306MP$$

This is our multiple variable regression model. Using this equation, we would expect the "average" NBA player to score 8.92 points per game. We found this by using the mean of each variable (FGA = 7.41, X3P = 0.75, FT = 1.52, MP = 21.43).

It gives meaningful context to discuss the coefficients of our model. To preserve the integrity of the model, we will hold all other variables constant when discussing change. For every field goal attempted per game, we expect a player to add 0.9153 points per game. For every three-point field goal made per game, we expect a player to add 0.246 points per game. For every free throw made per game, we expect a player to add 1.0576 points per game. Finally, for every minute played per game, we expect a player to add 0.0306 points per game. Interpreting the intercept in this case is largely irrelevant, because it is impossible to score less than 0 points per game.

Discussion

This model proved to be valid and very accurate, except for a handful of extremely high-scoring players. This is shown through the residual plots below. It is clear that the vast majority of the predictions are within two points of the observed values, and the Residuals vs Fitted plot has no patterns in its distribution. Because a standard field goal in basketball is worth two points, our model is even more accurate than it seems, especially for better players, who take more than 20 shots per game. An error of one shot per game is not very significant in the context of the NBA's 82-game season.



The model is useful to predict a hypothetical player's points per game. For example, say player X had FGA = 12.4, X3P = 2.2, FT = 3.3, MP = 34.2. Based on our model, we would predict player X to score 16.11 points per game. In addition to predicting a hypothetical player's points per game, we can use our model to calculate an actual player's residual score. Take Bradley Beal as an example, the starting Shooting Guard for the Washington Wizards. Beal attempted 14.5 field goals per game (FGA), 1.9 three-point field goals per game (X3P), converted 2.5 free throws per game (FT), and played for 31.1 minutes per game (MP). Based on these parameters, we predicted that Beal would score $-.3119 + 0.9153(14.5) + 0.2460(1.9) + .0306(31.1) = 17.2$ points per game. In reality, he scored 17.4 points per game. Beal's residual score of 0.2 shows that our model is very accurate.

One thing that was evident through the residual plots is that player 93 does not fit our model. This player was reigning NBA Most Valuable Player Stephen Curry. He has already set numerous NBA records, including the number of three-point field goals made in a single season. Because of his historically unmatched ability to make three-pointers, he scored significantly more points than expected through our model. Additionally, players 56 and 109 are specifically marked in the residual plots. These were Kobe Bryant and Kevin Durant, respectively, who are also two of the best scorers in NBA history. One thing that all three of these players have in common is their FGA values, which were abnormally high. Keep in mind that the mean number of field goals attempted was 7.41, with a standard deviation of 4.23. Curry's, Bryant's, and Durant's FGA values were 20.2, 16.9, and 19.2, respectively.

For an example, we will explore Steph Curry. In the 2015-2016 NBA season, he attempted 20.2 field goals per game, made 5.1 three-point field goals per game, made 4.6 free throws per game, and played 34.2 minutes per game. Our model predicted Curry would score 25.34 points per game. In reality, he scored 30.1 points per game. This discrepancy derives primarily from Curry's historic ability to convert three-point field goals. In this season, he made 402 "threes," shattering the all-time NBA record of 286, which was set by Curry two seasons ago. Because Curry is statistically the best three-point shooter in NBA history, he scores at a much higher rate than the rest of the league. Because our model is based on the league as a whole, his unmatched marksmanship exceeds any reasonable projections.

The antithesis of Curry is a defensive specialist. Andre Roberson, a Guard for the Oklahoma City Thunder is an example. In the 2015-2016 season, he attempted 3.9 field goals per game, made 0.5 three-point field goals per game, made 0.5 free throws per game, but still played 22.2 minutes per game. This means that while he spent nearly half the game on the floor, he had very limited involvement in the offense. Our model predicted that he would score 4.6 points per game. In reality, he scored 4.8. Roberson was one of four players in our sample that played more than 20.0 minutes per game and scored fewer than 5.0 points per game. He proves the strength of our model as an unusual type of player that still fits.

Our model taught us that it is much more difficult to predict an extraordinary player's points per game with statistics far beyond the average. Because there are certain athletes in any given sport that defy quantification, there will always be some players that cannot be modeled. In our case, three of these players include Curry, Bryant, and Durant. Outside of these outliers, nearly all of our players lie within one basket of our predicted values. We thought it was important to still include these outliers in our model to show just how extraordinary these players are.

Conclusion

Our model effectively predicts the points per game for the vast majority of NBA players. This can influence and aid the decision making of both coaches and executives. By making reasonable and accurate projections, this provides a method of evaluating free agency, opposing players, and even a team's own players. If this study were to continue, we would begin to introduce defensive and efficiency metrics into a more advanced model. Through this, we may even be able to create our own statistic to be used for development and evaluation. Through linear regression in R, we were able to build this model and create the opportunity for even more to come.