Statistical Analysis of Basketball Players’ Performance for the NBA season of 2018-19

Author: Himesh Buch

Submitted to: Lynn A. Agre, MPH, PhD

Statistics 467/567

July 27, 2020

Table of Contents

[Abstract: 3](#_Toc46763790)

[Introduction: 4](#_Toc46763791)

[Methods: 6](#_Toc46763792)

[Results: 10](#_Toc46763793)

[Discussion: 13](#_Toc46763794)

[Reference: 19](#_Toc46763795)

# Abstract:

In recent years, new player tracking technology provides new information about basketball game performance. The objective is to figure out who the best all round player is in the NBA season of 2018-19 by looking at some important predictors like points, blocks, steals, assist, and free throw percent per game, and also to see how all these predictors help or matter in concluding the results. One thing to also look at is whether the variable Player Efficiency Rating (PER), which is a reliable factor, is a proper indicator to deciding the best player. Coaching staffs may apply this data analysis to various players, while representing singular contrasts and utilitarian inconstancy, to streamline work on arranging and, thus, the game exhibitions of people and groups. Archival data are obtained from all 2018-2019 regular season games of over 500 players (n = 512). The dataset can also be used to find the best defender and best shooter of that season. To accomplish the above explained results, various statistical analysis methods are used for both univariate and bivariate analysis. Multiple regression methods can also be used to see the relation between different independent variables like points, blocks, steals, assists per game and the response variable which in this case is player’s efficiency rating. Depending on the response and independent variables, for all round best player, it would make sense to see a linear relationship between the independent variables like points, assist, and blocks per game. Graphing different plots and removing outliers can help to make more solid assertions**.** Not only the best player, questions like, what predictors help in making the best shooter and defender, can also be discussed. The results identified different playing profile of performers, particularly related to the game roles of scoring, passing, defensive and all-round game behavior. Comparing the results of who the best player is according to the analysis with 2018-19 season’s declared Most Valuable Player (MVP) by NBA will be the final step.

# Introduction:

Thesis: Checking whether or not Player Efficiency Rating (PER) is a reliable factor in deciding the Most Valuable Player (MVP)

Basketball performance relies fundamentally upon shooting 2-point field-goals and on making sure about cautious rebounds. In close games, in any case, fouls and free-throws show expanded significance for deciding game result than for lesser challenged games. Other remaining game statistics, for example, offensive bounce back, turnovers, steals and assists are not announced reliably as segregating performance factors for winning and losing, but that does not mean that it doesn’t play a part at all. While differentiating the best and worst teams, the best performance factors for long haul wins are identified with assists, steals and blocks. This is essentially the main goal of this research, which is to support these arguments and provide more certainty.

In major professional sports, the coach and team management are looking for ways to dominate more matches so as to pull in more fans to help their business. Sports statistical modeling analytics is turning into a basic way to deal with reveal the triumphant examples covered up with sports information gathered during each game played. The objective of this paper is to build a statistical model based on given limited number of variables in the dataset. There are several research talks presented in MIT Sloan Sports Analytics Conference. There are a few exploration talks introduced in MIT Sloan Sports Analytics Conference. These papers have utilized escalated Analytics to reveal players' playing pattern and assist coaches with building up every player so as to make and boost every player's qualities to their particular group.

The analysis can also help to determine if any of the predictors are co-related. Meaning, the analysis can prove the inter-relation of any of the variables (if any). By just looking at the dataset some variables could seem independent but might not be. After comparing multiple variables with the response variable player efficiency rating, the analysis should be able to find why an indicator is useful or needless.

Keeping that in mind, the hypothesis that needs to be answered are,

1. The best all round player (MVP) of the season of 2018-19
2. The best defensive player of the season of 2018-19
3. What variables come into effect while concluding above two results
4. Whether or not PER is a good standard

The dataset that has been chosen has a combination of both quantitative and qualitative variables. Not only that, most of the predictors that are considered are in quantitative or numeric form, which makes it easier to perform any statistical analysis as compared to categorical datapoints. The study can provide important results that can help teams in drafting new players as well. Things like how much each independent variable dominate in determining the result can help the team management help a lot as they will know what player statistics to look at while drafting a player.

# Methods:

The dataset is adapted and modified accordingly for this data analysis from Kaggle. It has six different variables that can help to determine who the best player is in the NBA season of 2018-19. It is very important to understand and analyze why and how each variable helps to get to the results.

In order to understand the spread of all these variables, mean, median, mode along with interquartile range (IQR), range, standard deviation, variance, are computed using R and is presented along with the analysis.

Here is a chart to understand the types of the variables:

**Unit of analysis of each variable:**

* **Not considered in the analysis:**

Name: Factor (string)

Team: Factor (string)

Age: num (integer)

Height: num (double)

* **Considered in the analysis:**

Weight: num (double)

Points: num (double)

Blocks: num (double)

Assists: num (double)

Steals: num (double)

Rebounds: num (double)

Free throw %: num (double)

Field Goal %: num (double)

Player efficiency rating: num (double)

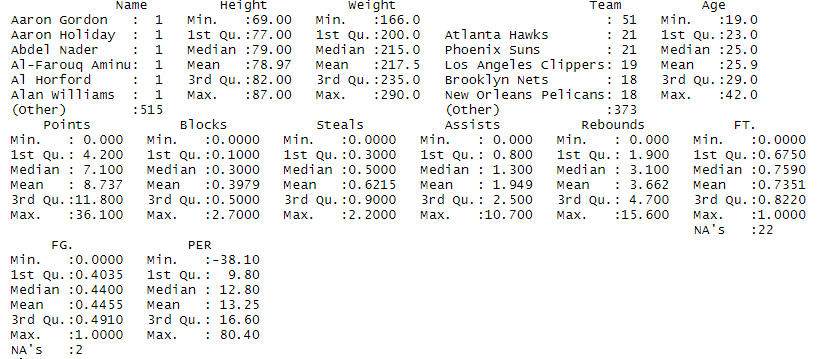
The variables analyzed are:

* Points: as the name suggests, it describes points scored per game by each player. It plays a very important role in determining the MVP.
* Assists: basketball is a team sport and assisting the ball properly to players is very important in successfully executing different strategies and eventually winning the games with it. The variable “assists” describes number of times the ball is passed to a player and is successfully turned into a point, per game. These numbers are generated by taking an average of all assists throughout the season. This attribute helps to understand whether a player is a team player or not, which again, helps in deciding the MVP
* Blocks: this predictor describes average number of times the shot is blocked by the player per game, throughout the season. These numbers are generated by taking an average of all blocks throughout the season. This independent variable helps in choosing the best defender.
* Steals: this attribute represents the number of times a ball is stolen by a player in transition. This an essential step in breaking the oppositions offensive strategy. This predictor can also help in determining the best defender of the season.
* Rebounds: “rebounds” is the number of times when a player acquired the ball after an unsuccessful attempt at shooting the ball. The ball usually rebounds off the hoop and the player who attains it after, gains point for rebound.
* Free throw percentage (FT%): usually, two or three free throws are given to a player upon a foul play. If the player successfully shoots the ball from a given fixed distance, they get a point. This predictor will help in checking whether a player is a good defender or not, as this type of play usually happens because of the mistake of a defender. It also helps in deciding whether a player is a good shooter from a closer range.
* Field goal percentage: a field goal point is given to a player when the shot that is not a free throw, is successful. The final points consist of the sum of both the field goals and free throws scored in a game. This again plays an important role in determining the MVP.
* Player efficiency rating: The player efficiency rating (PER) is John Hollinger's all-in-one basketball rating, which attempts to boil down all of a player's contributions into one number. Using a detailed formula, Hollinger developed a system that rates every player's statistical performance. This variable is used in plotting the normal probability plot and linear regression. It is used as a response variable and all the other variables are regressed upon PER.

Other data columns like point\_logistic is generated by having all the points greater than 15 as 1 and others as 0, and blocks\_logistic and per\_logistic are made with the similar approach. All these columns will be used in logistic regression and likelihood ratio test. Not only that, for eigen vectors and eigen values, the first 8 rows and 8 columns (all variables) are selected to build a square matrix to be computed on

# Results:

In order to understand each variable, it is important to understand how each variable is distributed throughout. To do that, very basic but highly effective statistical methods like computing mean, median, and mod, and to look at the spread, computing variance and standard deviation are used. In R, it is very easy to do this by using the summary() command. Here is the output of from R:



Different methods other than what has been described can also be used to understand a bit more about the data. Statistical analysis can be done most accurately if the spread and distribution of the data is understood properly. Other methods like graphing different plots can help a lot. For this analysis, these plots have been graphed for both univariate and bivariate analysis:

1. Dot plot (for univariate analysis)
2. Scatter plot (for bivariate analysis)
3. Box plot (for univariate analysis)
4. Histogram (for univariate analysis)
5. Density plot (to understand skewness)
6. Chi-square plot (for univariate analysis)
7. Normal Probability plot (for univariate analysis)

All these plots, in one way or another, help in visualizing the spread of the data to better understand it and derive conclusions from it, eventually. Dot plots are used for continuous, quantitative, univariate data. Dot plots are one of the simplest statistical plots and are suitable for small to moderate sized data sets. In this study, they are useful for highlighting gaps and outliers. Scatter plots are used to plot data points on a horizontal and a vertical axis in the attempt to show how much one variable is affected by another. As there are more than one variable are involved, it is used mainly in bivariate and multivariate analysis. In this study, scatter plots are drawn by comparing each variable with the response variable Player Efficiency Rating (PER). A box and whisker plot are a way of summarizing a set of data measured on an interval scale. It is often used in explanatory data analysis. This type of graph is used to show the shape of the distribution, its central value, and its variability. A histogram is only used to plot the frequency of score occurrences in a continuous data set. In this case, most of the variables are continuous and thus, histogram plays an important role in univariate analysis of the variables. The normal probability plot is a graphical technique for assessing whether or not a data set is approximately normally distributed. It is important for the research to see if the data is normally distributed, which means to check, if the variables are skewed to either sides. This can be also achieved by drawing density plots and checking if the shape of the curve is bell, in that case the data is symmetric or normally distributed, in all the other cases it might be skewed to either left or right side.

The skewness of each variable is different and that means that there are some outliers that exist in each data column. It is important to remove as many outliers as possible to get the most accurate results. There are basically three types of skewness:

1. Left skewed
2. Right skewed
3. Symmetric

The graph of a left skewed data column is skewed to the left, meaning, most of the data points are smaller and are to the left of the distribution and some bigger values are to the right. Unlike this, right skewed distribution is skewed to the right, meaning most of the data points are bigger than the rest, to the right, and very few small data points are to the left. On the other hand, symmetric distribution has almost, if not completely, equal distribution of the data points. If graphed, it creates a bell shape. In any data analysis, symmetric distribution is preferred as there will be very few occurrences of outliers in that case. Density plots can be plotted to visually see the skewness or in R, commands like skewness() can be used to get the numeric result. If the data is skewed a lot to either of the sides, then that indicates that it needs to be cleaned, meaning, might have to remove some outliers from it to get the most perfect result. Methods like transformation of variables is used to reduce the skewness in this case.

For bivariate analysis, methods like Hotelling’s T2 methods, correlation matrix (tells which pairs have the highest correlation), variance and covariance matrix (Table 4, 5, 6, 7), Likelihood ratio test (two variables points and blocks are converted into binary form along with the response variable PER), and eigenvalues and eigen vectors are used, and all the results are presented with the analysis. R programming language is used to do all the analysis and packages like e1071, mvdalab, mvoutlier are extremely useful in doing that.

# Discussion:

As explained in previous section, in order to analyze the data properly, it is important to understand how it is distributed. In order to do so, looking at the mean, median, and mod of each variable, it looks like, each player approximately score about ~9 points per game, while the most points scored in any game is 36. There are very few blocks and steals are made in each game on average, while there are about ~3 assists made by players in each game. This result gives an insight of the strategies that the teams might have used, and by looking at the results, the teams or players have not done a great job in defending the players. The highest points scored are about 36 in a game which says a lot about players playing offensive and breaking the defense. This is also supported by a bigger number of about 10 assists per game. Hence, by just looking at a very high level data distribution, it can be said that teams and/or players have been using some offensive strategies to win, and as explained in the introduction, this analysis has thus helped on a minor scale to understand the team and player strategies and mindsets. There are more rebounds made than both steals and blocks combined, which again supports the previously made assertions.

Discussing more about the distribution, looking at the density plots of each variables, there are very few variables that have (almost) symmetric, bell shaped, distribution. These variables are points and steals (Table 3). All the other variables are either skewed to the left or right, meaning most of the datapoints, for left skewed, are to the left, and for right skewed, are to the right. The variables that are skewed to the left are blocks, assists, and rebounds. In order to make the data more symmetric, for univariate analysis, the method of transforming the variable is used and variable “rebound” is transformed.

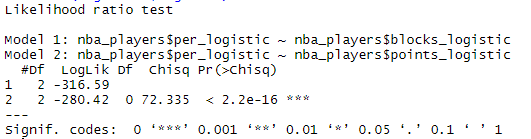
**Transformation of one variable**: For this method, the variable “rebounds” has been chosen to be transformed. The reason to choose this variable is because of its skewness. In order to transform the variable, the square root of the datapoints is taken. Skewness before the transformation is 1.42 which is not as close as 0 as needed. After the transformation, the skewness is reduced to 0.2 which is far closer to 0 than before and makes the distribution a lot more symmetric than before.

Looking at the dot plots of each variables, almost all variables have outliers. These can affect in a negative way to the analysis, as outliers are the values that are either way too large or way too small. These outliers can be removed, and the analysis can be re-done to get even more accurate results. Not only that, if many outliers are removed, it can raise a problem of data being too symmetric which can also lead to false results.

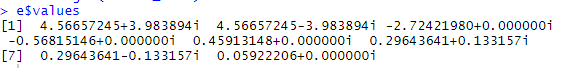
The Normal Probability Plots can also give information about the data being symmetric. In R, the command qqnorm() is used to draw the plots. These results (the graphs) also comply and support the results of symmetry, gotten from drawing the density plots and computing skewness.

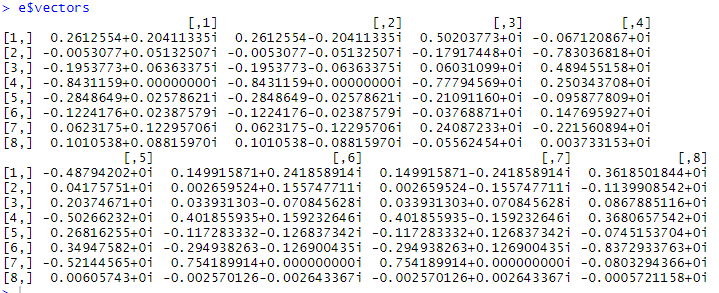
Looking at the covariance matrix (Table 4), if both variables tend to increase or decrease together, the coefficient is positive, and if one variable tends to increase as the other decreases, the coefficient is negative. It looks like all the variables have positive coefficient. Since all the coefficients are positive, it can be said that all the variables move together, as increasing one variable will increase the other too and vice-versa. This is also backed by the correlation matrix, as the relationship between one or more variables move together and are of a linear fashion.

The output of log likelihood test also helps to in understanding how PER is only higher for players who score more than defend. Looking at the output,



the log likelihood ratio test rejects the null hypothesis of PER being a good indicator as the value of blocks is not as low as value of points. As explained in previous sections, the matrix for computing the eigen values is of dimensions 8x8 and here are the eigen values and eigen vectors for that,





As seen in the output, the values that all of the predictors are stretched by these values after transformation (eigen vectors). The negative sign represents the reversed direction.

In the analysis, PER is the response variable and all the other predictors are independent variables that help in predicting the results. After building a linear regression model (table 9), it looks like, according to the formula of PER, the variables “Points” (p-value: 6.11e-09) and “field goal%” (p-value: < 2e-16) are the only ones that are prioritized in getting those results, and hence, to answer the third hypothesis, what helps the most in becoming the MVP, are points scored and a better field goal percentage. Given these results, by looking at the variable PER, the MVP of the season should be, the player Zhau Qi, which is way off for a lot of reasons. Keep in mind that PER is not the variable that NBA committee keeps in mind while deciding the MVP but is an important factor to look at. Zhau Qi, by looking at the data, does not have a good record in any of the fields like shooting, scoring, defending, stealing, and assisting. Hence, it explains that the formula of PER can be a little buggy or might not be complying with the given dataset and a new formula, specific to this study should be made.

By doing some research, given the scope of the study and limited number of variables, here is the formula that is used in building the MVP index that will help to choose the MVP of the season, and answers the first hypothesis:

*Points scored + 0.4(Field goal%) + 0.4(Free throw%) + 0.7(Rebounds) + steals + 0.7(Blocks) + 0.7(assists)*

This formula, as explained above, is the result of understanding how one variable is more important than other, by going through different formulas used by other researchers, and by understanding the formula of the player efficiency rating. A new column “mvp\_index” has been added to the dataset which is the result of this formula. Now, each variable is given equal importance which can be proven by looking at the p-values of the of all the variables, which are very close to each other. Given the result and formula, now, the MVP according to our analysis is James Harden, who is a very skilled and renowned basketball player because of his abilities, and he is also one of the candidates NBA announced for the race of MVP of that season. However, the declared MVP of the season by NBA is Giannis Antetokounmpo, which according to our study, is on number two, with a very close score of mvp\_index to James Harden (Table 10). This is a lot better than the results found from the official PER rating.

Till now, two hypotheses, the MVP of the season, and what factors come in play in deciding that, has been answered. To answer the hypothesis of who the best defender is for the season, a new formula needs to be made. As defenders do not score as much as other players, variables like points and field goal percentages will not affect as much. And hence, after some research, the new formula is,

*0.4(Free throw%) + 0.7(Rebounds) + steals + Blocks + 0.7(assists)*

The formula given above, is a subset of the main formula, with some variables removed. In the dataset, the defender\_index has the results of this formula. And according to our analysis (Table 11), the best defender of the season of 2018-19 should be, Russell Westbrook (defender\_index of 17.92), but according to NBA, it is Rudy Golbert (defender\_index of 13.78). The results are a bit off, but that is because of the very few variables being analyzed. Variables like different fouls, and blocks and rebounds missed are very important, but the data of those variables is hard to create and not available at the moment.

For further analysis, a player’s body weight, height can also be compared to the analysis and see if there is a right Body Mass Index (BMI) that players who are at the top of the list have. This can help the coaching staff create proper meal plans and workout routine for players. This analysis of univariate and bivariate analysis asks a lot of questions like why the results are off, if they are (in finding the best defender), are there any other factors that play an important role other than the ones that are analyzed (like foul plays). The best way to answer those questions is to use multiple regression analysis and using various methods like Ridge and LASSO (can help a lot for the analysis of data which has no defined or well-accepted formulas) regression, Principal Components Regression (mainly used when the dataset has many independent variables or multicollinearity exist in the data) etc. The main impact of the study is that the PER is not as efficient as it looks. In basketball community, results of PER are respected and used for a lot of studies. Not only the results are way off, it also did not help in identifying why one variable is better or more important than the other. Our analysis overcame those limitations and answered the questions that could not be answered by PER and more.

# Reference:

1. Data retrieved from Kaggle: <https://www.kaggle.com/schmadam97/nba-regular-season-stats-20182019/metadata>
2. Karipidis A, Fotinakis P, Taxildaris K, Fatouros J (2001). Factors characterizing a successful performance in basketball. J Hum Movement Stud. Retrieved from <https://scholar.google.com/scholar_lookup?journal=J+Hum+Movement+Stud&title=Factors+characterizing+a+successful+performance+in+basketball&author=A+Karipidis&author=P+Fotinakis&author=K+Taxildaris&author=J+Fatouros&volume=41&issue=5&publication_year=2001&pages=385-97&>
3. Malarranha J, Figueira B, Leite N, Sampaio J (2013). Dynamic Modeling of Performance in Basketball. International Journal of Performance Analysis in Sport. Retrieved from <https://scholar.google.com/scholar_lookup?journal=International+Journal+of+Performance+Analysis+in+Sport&title=Dynamic+Modeling+of+Performance+in+Basketball&author=J+Malarranha&author=B+Figueira&author=N+Leite&author=J+Sampaio&volume=13&publication_year=2013&pages=377-86&>
4. Sampaio J, Janeira M (2003). Statistical analyses of basketball team performance: understanding teams’ wins and losses according to a different index of ball possessions. International Journal of Performance Analysis in Sport. Retrieved from <https://scholar.google.com/scholar_lookup?journal=International+Journal+of+Performance+Analysis+in+Sport&title=Statistical+analyses+of+basketball+team+performance:+understanding+teams%E2%80%99+wins+and+losses+according+to+a+different+index+of+ball+possessions&author=J+Sampaio&author=M+Janeira&volume=3&issue=1&publication_year=2003&pages=40-9&>
5. Ibanez SJ, Sampaio J, Feu S, Lorenzo A, Gomez MA, Ortega E (2008). Basketball game-related statistics that discriminate between teams' season-long success. European journal of sport science. Retrieved from <https://scholar.google.com/scholar_lookup?journal=European+journal+of+sport+science&title=Basketball+game-related+statistics+that+discriminate+between+teams%27+season-long+success&author=SJ+Ibanez&author=J+Sampaio&author=S+Feu&author=A+Lorenzo&author=MA+Gomez&volume=8&issue=6&publication_year=2008&pages=369-72&>
6. Niiler, T. (2020). Comparing groups of time dependent data using locally weighted scatterplot smoothing alpha-adjusted serial T-tests. Gait & posture, 76, 58-63.
7. Sarlis, V., & Tjortjis, C. (2020). Sports analytics–Evaluation of basketball players and team performance. *Information Systems*, 101562.
8. Berri, D. J. (1999). Who is ‘most valuable’? Measuring the player's production of wins in the National Basketball Association. *Managerial and decision economics*, *20*(8), 411-427.
9. A. M. El, M. S. Kamel, F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 2011, 44(3):572-587.
10. M. Bejani, D. Gharavian, M. N. Charkari. Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks. Neural Computing and Applications, 2014, 24(2):399-412.
11. L. Wang, K. Wang, R. Li. Unsupervised feature selection based on spectral regression from manifold learning for facial expression recognition. JET Computer Vision, 2015, 9(5):655-662.
12. W. F. Gu, C. Xiang, Y. V. Venkatesh, et al. Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. Pattern Recognition, 2012, 45(1):80-9I.
13. Y. Lv, Z. Feng and C. Xu. Facial expression recognition via deep learning/ International Conference on Smart Computing. IEEE, 2015:303 - 308.
14. Chen, Y., Dai, J., & Zhang, C. (2019, August). A neural network model of the NBA most valued player selection prediction. In Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence (pp. 16-20).
15. Thriveni, K., & Mahanthesh, B. Nonlinear Boussinesq buoyancy driven flow and radiative heat transport of magnetohybrid nanoliquid in an annulus: A statistical framework. Heat Transfer.