An Exploration to Determine the NBA MVP

In what ways can the NBA's Most Valuable Player be predicted through the use of statistical analysis?

5277 Words

Mathematics

Table of Contents

I. Introduction

The NBA Most Valuable Player award is the most prestigious individual accolade in all of basketball. This honor is awarded to the player who has the biggest impact on the success of their team. However, since a player's impact cannot be measured numerically, the award winner is very difficult to predict throughout the season. At the end of every season, the basketball community, including myself, usually has some idea of who the award winner will be. Yet, when the statistics of the players are observed objectively, the outcome sometimes becomes unclear. This proposes the following question: In what ways can the NBA's Most Valuable Player be predicted through the use of statistical analysis?

First, it is important to know how the MVP is selected. The MVP is chosen through voting ballots by one hundred broadcasters and sports analysts (not affiliated with any NBA team) and one ballot that represents the fans. Each voter is required to rank their top five candidates in order, and the position in this order corresponds to a specific point total.

1st	10 points
2nd	7 points
3rd	5 points
4th	3 points
5th	1 points

The player with the highest cumulative point total wins the award. However, what do these analysts look at to determine how to vote? There are four main factors that determine who is the MVP:

1. Personal statistics of the player on offense and defense
2. His impact on the team's success through the **regular season** (playoffs are **not** included)

3. The athletes health: how many games they played in
4. The connotations around the player's name in the media and sports market

Therefore, the goal of this paper is to seek a mathematical model, in which these factors can be expressed numerically and predict who the MVP will be.

First, I will show how to determine the least squares regression line, which I will use to eventually create a mathematical model. I will also demonstrate how to find the least squares regression line in multivariate regression, to eventually create a different mathematical model.

Before I start creating my models, I will use the bivariate linear regression model to observe the correlation between player statistics and the share of MVP votes they won. Then, I will select the statistical categories that have the greatest effect on who wins the MVP, to be the factors in my experiment. Then, I will record the statistics from these categories for the top ten players in MVP voting over the past ten years.

Next, I will create my first mathematical model with the bivariate linear regression model. In order to simplify many factors into a single variable, I will compute the mean, standard deviation, and z scores of various statistics to ultimately generate a new statistical sum that represents each player. These new sums will then be plotted against the voting share for each corresponding player, and the regression line will be computed.

Then, I will create a mathematical model using the multivariate linear regression model. This model will be created through the program R, and will find the regression equation based on the statistical categories that I previously selected.

After both models are created, I will compare them against each other, to see which model is more accurate. This will be determined by examining the r^2 values for each model, and

by observing if they accurately predict the MVP in the 2010 season (data that was not used to create the model).

II. Determining how to create a prediction model

To conduct a regression analysis, the least squares regression line needs to be created. In order to determine the correct line that represents the data, the sum of the squared residuals needs to be minimized. This can be accomplished by taking the derivative of the sum of the squared residuals with respect to each coefficient and setting these equations to zero.

$$\sum (y-\hat{y})^2 = \sum (y-(a+bx))^2$$

In order to find the derivative of the equation, I must find the partial derivative with respect to both a and b.

Equation a:

$$\frac{\partial \hat{y}}{\partial a} = \sum \frac{\partial}{\partial a}(y-(a+bx))^2$$

$$= \sum 2(y-(a+bx))(-1)$$

$$= -2\sum (y-(a+bx))$$

Equation b:

$$\frac{\partial \hat{y}}{\partial b} = \sum \frac{\partial}{\partial b}(y-(a+bx))^2$$

$$= \sum 2(y-(a+bx))(-x)$$

$$= -2\sum x(y-(a+bx))$$

Set them equal to 0 to solve for the minimum

Equation a:

$$-2\sum (y-(a+bx)) = 0$$

Equation b:

$$-2\sum x(y-(a+bx)) = 0$$

-2 cancels, and we can solve for the unknowns

Equation a:

Express a in terms of b

$$\sum (y-(a+bx)) = 0$$

$$\sum y - \sum a - \sum bx = 0$$

$$\sum y - na - \sum bx = 0$$

$$na = \sum y - \sum bx$$

Divide by n

$$a = \frac{\sum y}{n} - \frac{\sum bx}{n}$$

$$a = \frac{\sum y}{n} - \frac{b\sum x}{n}$$

Since $\frac{\sum y}{n} = \bar{y}$, and $\frac{\sum x}{n} = \bar{x}$

$$a = \bar{y} - b\bar{x}$$

Now, we can find the slope os the LSRL, b, by subsituting $\bar{y} - b\bar{x}$ in for a.

Equation b:

$$\sum x(y-(a+bx)) = 0$$

$$\sum x(y-(\bar{y} - b\bar{x} + bx)) = 0$$

$$\sum x(y-\bar{y} - b(x-\bar{x})) = 0$$

$$\sum x(y-\bar{y}) - \sum bx(x-\bar{x}) = 0$$

$$\sum x(y-\bar{y}) - b\sum x(x-\bar{x}) = 0$$

$$b\sum x(x-\bar{x}) = \sum x(y-\bar{y})$$

$$b = \frac{\sum x(y-\bar{y})}{\sum x(x-\bar{x})}$$

Now that the equations of the unknowns are found for a bivariate LSRL, a similar process can be used to find the LSRL for multivariate a multivariate equation.

To solve for mininmum, consider $\frac{\partial \hat{y}}{\partial a} \sum_i \left(y_i - \left(a + \sum_j b_j x_{i_j} \right) \right)^2$

Set the derivative equal to zero.

$$0 = -2 \sum_i \left(y_i - \left(a + \sum_j b_j x_{i_j} \right) \right)$$

$$0 = \sum_i \left(y_i - a - \sum_j b_j x_{i_j} \right)$$

$$0 = \sum_i y_i - na - \sum_i \sum_j b_j x_{i_j}$$

$$a = \frac{\sum_i y_i - \sum_i \sum_j b_j x_{i_j}}{n}$$

$$a = \bar{y} - \sum_j b_j \bar{x}_j$$

This a can now be used to determine the equation for any b_m.

Consider $\frac{\partial \hat{y}}{\partial b_m} \sum_i \left(y_i - \left(a + \sum_j b_j x_{i_j} \right) \right)^2$

Set the derivative equal to 0.

$$0 = 2 \sum_i \left(y_i - \left(a + \sum_j b_j x_{i_j} \right) \right) \left(-x_{i_m} \right)$$

$$0 = \sum_i \left(y_i - a - \sum_j b_j x_{i_j} \right) \left(-x_{i_m} \right)$$

$$0 = \sum_i \left(y_i - a - b_m x_{i_m} - \sum_{j \mid j \neq m} b_j x_{i_j} \right) \left(-x_{i_m} \right)$$

$$0 = \sum_i \left(y_i - \bar{y} + \sum_j b_j \bar{x}_j - b_m x_{i_m} - \sum_{j \mid j \neq m} b_j x_{i_j} \right) \left(-x_{i_m} \right)$$

$$0 = \sum_i \left(y_i - \bar{y} + \sum_{j \mid j \neq m} b_j \bar{x}_j + b_m \bar{x}_m - b_m x_{i_m} - \sum_{j \mid j \neq m} b_j x_{i_j} \right) \left(-x_{i_m} \right)$$

$$0 = -\sum_i x_{i_m} y_i + \bar{y}\sum_i x_{i_m} - \sum_i x_{i_m} \sum_{j|j\neq m} b_j \bar{x}_j - b_m \bar{x}_m \sum_i x_{i_m} + b_m \sum_i \left(x_{i_m}\right)^2 + \sum_i x_{i_m} \sum_{j|j\neq m} b_j x_{i_j}$$

$$0 = -\sum_i x_{i_m} y_i + n\bar{y}\,\bar{x}_m - \sum_i x_{i_m} \sum_{j|j\neq m} b_j \bar{x}_j - n b_m \left(\bar{x}_m\right)^2 + b_m \sum_i \left(x_{i_m}\right)^2 + \sum_i x_{i_m} \sum_{j|j\neq m} b_j x_{i_j}$$

$$b_m = \frac{n\bar{y}\,\bar{x}_m + \sum_i x_{i_m} \sum_{j|j\neq m} b_j x_{i_j} - \sum_i x_{i_m}\left(y_i + \sum_{j|j\neq m} b_j \bar{x}_j\right)}{n\left(\bar{x}_m\right)^2 - \sum_i \left(x_{i_m}\right)^2}$$

$$b_m = \frac{n\bar{y}\,\bar{x}_m + \sum_i x_{i_m}\left(\sum_{j|j\neq m} b_j x_{i_j} - \left(y_i + \sum_{j|j\neq m} b_j \bar{x}_j\right)\right)}{n\left(\bar{x}_m\right)^2 - \sum_i \left(x_{i_m}\right)^2}$$

Now that the equation for both a and any b_m has been found, the second derivative should be calculated to verify that it is a minimum.

$$\frac{\partial^2 \hat{y}}{\partial a^2} = -2\sum_i -1$$

$$= 2n$$

Since the second derivative is positive, this shows that the graph of the function is concave up, thus proving that we have found a minimum.

$$\frac{\partial^2 \hat{y}}{\partial b_m{}^2} = 2\sum_i \left(-x_{i_m}\right)^2$$

$$= 2\sum_i \left(x_{i_m}\right)^2$$

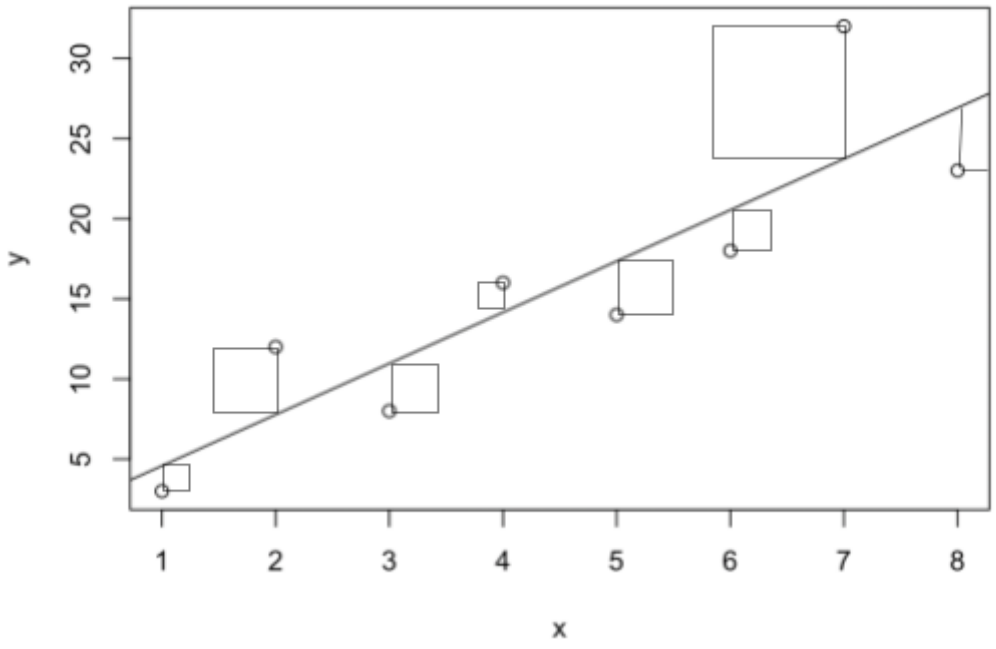
Once again, since the second derivative is positive, this shows that the graph of the function is concave up, thus proving that we have found a minimum.

Despite finding the formula for both a and b_m, all of the coefficients in the equation are in terms of one another. This means that in order to solve for a or b_m, one must create a system of equations to solve for the unknowns. Due to the amount of explanatory variables that I intend to include, I decided to create my regression equation using the program R, as opposed to by hand.

III. Determining how to statistically evaluate a prediction model

Once a regression equation has been constructed, how does one know if it accurately represents the data? There are many statistical methods to help determine whether or not a regression line is a good fit for the data. However, I will be focusing on the use of r^2, the coefficient of determination, to evaluate my models.

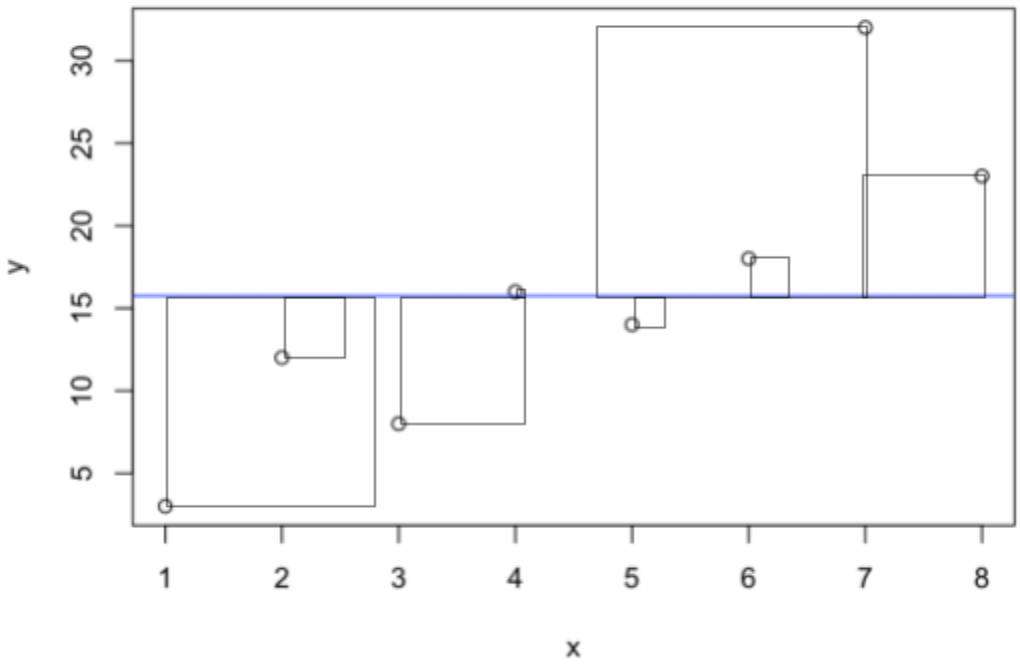
The coefficient of determination represents the amount of variability in the response variable that is represented by the least squares regression line. Finding the formula for r^2 is quite simple. To determine the amount of variability in y that the LSRL does capture, the amount of variability that is not captured must be examined, or in other words, the squared residuals. Here is an example of an approximation of the squared residuals, SSResid, with *arbitrary* data.



In this plot, it is clear to see the variability that the LSRL does not explain for each y value. Since the explanatory variables are ultimately trying to explain why certain data values

deviate from the mean, the squared deviations from the mean should be examined to capture the total variability in *y*.

Here is an example of an approximation of the squared deviations from the mean, SSTo, with the same *arbitrary* data.



So, if one knows the total variability in *y* (SSTo), and the amount of variability not captured by the LSRL (SSResid), then the amount of variability in *y* captured by the LSRL can be determined. First, one can simply calculate the percentage (in decimal form) of variability captured by the LSRL and then simply subtract that from 1. This gives us the following formula:

$$r^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}}$$

To show how this formula works, I will calculate the r^2 value for the *arbitrary* dataset in the previous graphs.

SSResid = 133.9762 SSTo = 561.5

$$r^2 = 1 - \frac{133.9762}{561.5}$$
$$= 1 - .2386$$
$$= .7614$$
$$= 76.14\%$$

So, 76.14% of the variability in *y* is captured by the LSRL. This measurement will become extremely valuable in my research, to make sure that I am accurately representing the factors that go into winning an MVP award.

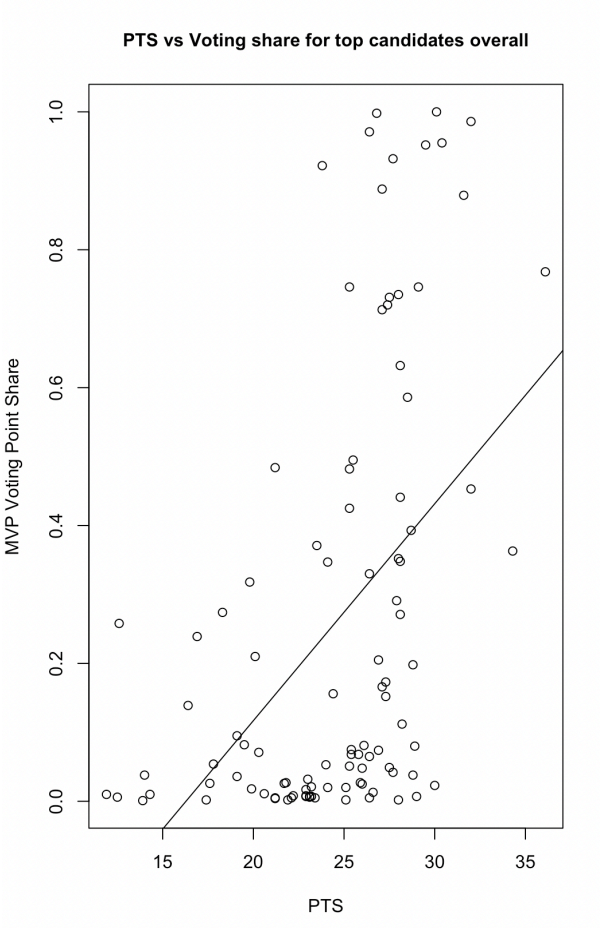
IV. Determining the different factors

In order to determine what statistics affect the MVP voting, a regression analysis can be conducted comparing the different statistical categories to the MVP voting share.

a. The impact of points per game

Points are essential in determining who deserves the MVP or not. For example, a player averaging 25 points per game (ppg) has a way better chance at winning the award than a player with 5 ppg. However, do points still matter when only the top players are compared? In other words, if one player averages 29 ppg, and another 27 ppg, is that enough of a difference to set them apart? We can try to answer this question through the examination of a regression model.

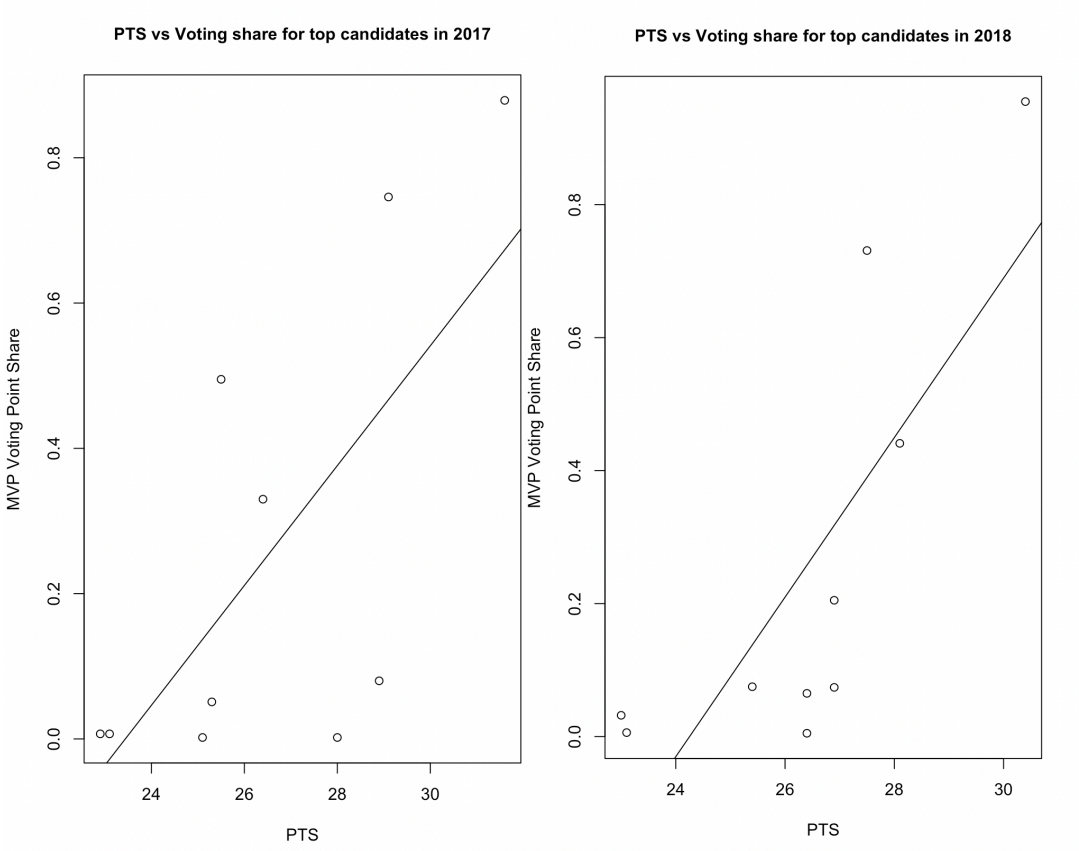
Figure 1.0



In Figure 1.0, the graph demonstrates the correlation between PPG and the MVP voting share. The correlation is positive between the two factors but is relatively weak. In fact, the r^2 value is only .2247, which shows the weak correlation in a numerical way. This is due to the fact that comparing all of the players together cannot accurately show the correlation between these two factors. For example, a player who had the most PPG in one year could have ended with 27 ppg, whereas a player who had the third most PPG in one year could have ended with 30

PPG. Therefore, the data should be studied in each individual year, when using bivariate regression.

Figure 1.1 Figure 1.2



Figures 1.1 and 1.2 show the correlation between points and the MVP voting share for the top ten candidates in 2017 and 2018. Now, each data point can only be compared to other points from that same year, which makes logical sense. Through this simple change, the positive correlation is now clear. The r^2 value for 2017 is .4658, and the value for 2018 is .6159, both of which are much greater than the r^2 value overall. While the correlation is still not particularly strong, the other factors in determining the MVP should be able to explain why. On top of that,

in the human sciences an r^2 value above .6 is strong enough to catch one's attention, due to the difficulty in predicting human behavior. Overall, due to the moderate positive linear correlation, I can conclude that points do factor into who wins MVP. .

b. The impact of assists per game

Figure 1.3

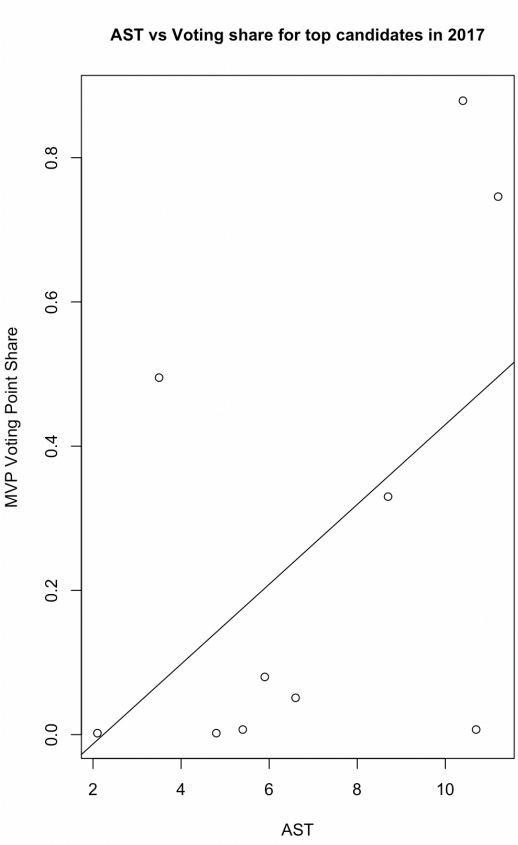
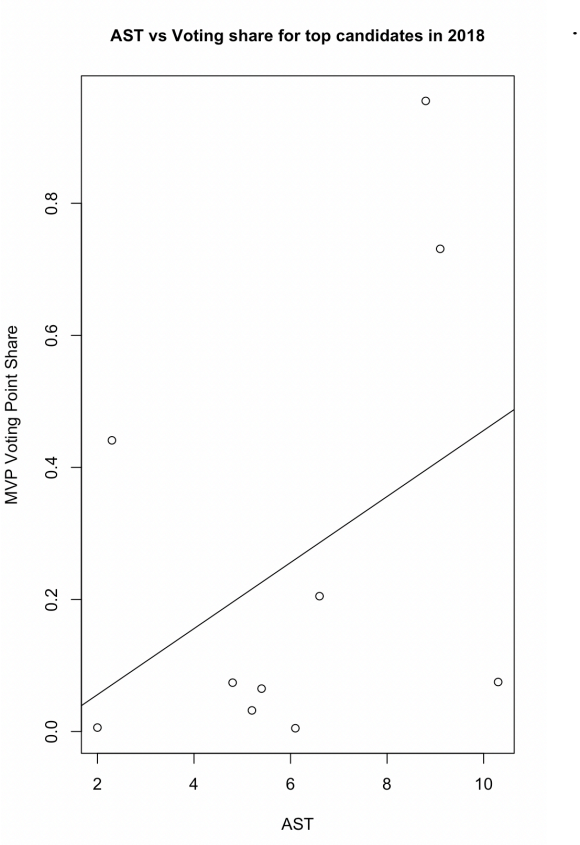


Figure 1.4



The relation between assists and the MVP voting share is weakly correlated, with the r^2 value for 2017 being just .2736 (Figure 1.3), and the value for 2018 just .1663 (Figure 1.4). However, these low values can be justified. Assists per game is a stat that is primarily significant to most guards. The guards on a basketball team are the ball handlers and play makers, so it makes sense that they have the most assists. While the center and forwards of a team still get

assists from time to time, they average a lot less assists because they are not the ones taking the ball up the court. So, many of the data values with a low amount of assists but a high voting share are simply players who are not guards. This is the reason for the low correlation in the models. Since the stat is very significant to guards, it must be a factor of any model that is created.

c. The impact of rebounds per game

Figure 1.5

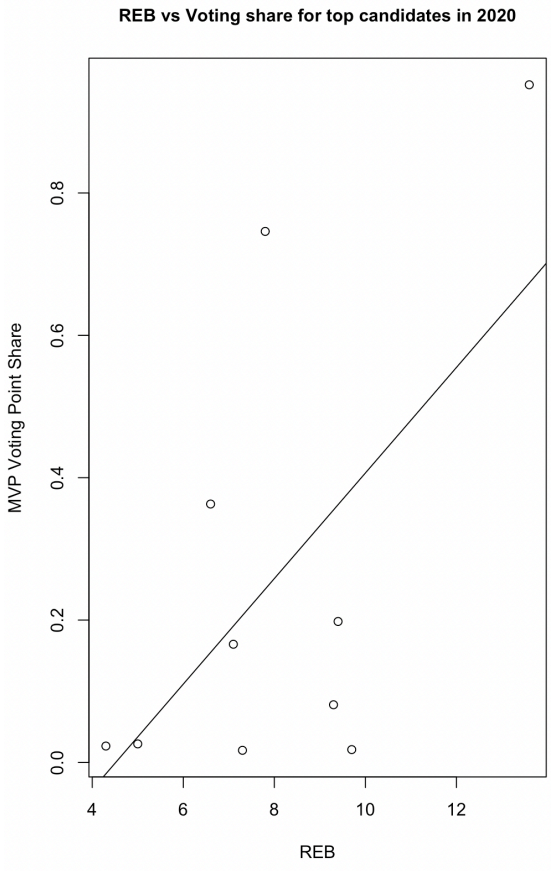
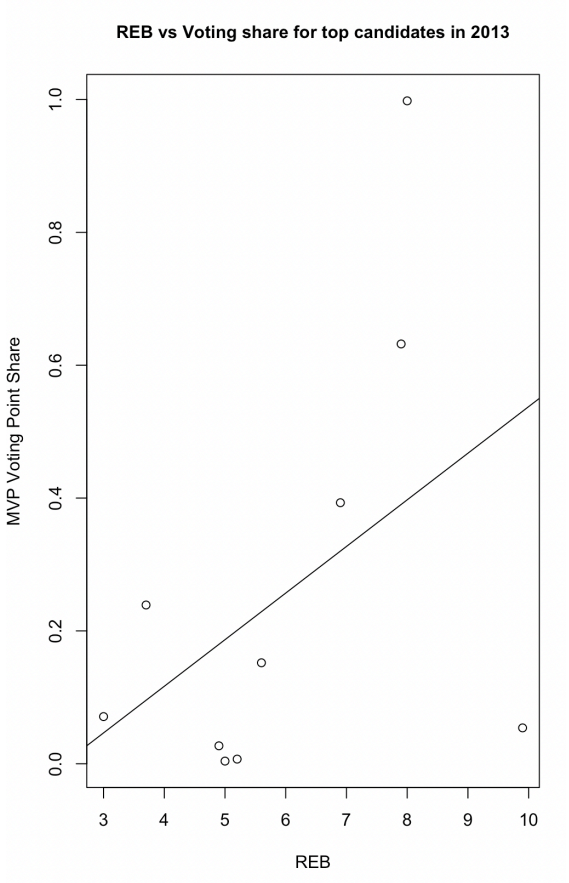


Figure 1.6



Similar to assists, rebounds and MVP voting share are weakly correlated, with the r^2 value for 2020 being .3513 (Figure 1.5), and for 2013 just .2068 (Figure 1.6). However, the weak correlation can be justified as well. While assists is a statistic significant to guards, rebounds is a

statistic significant to forwards and centers. Being these bigger players on the court, players at these positions get most of the rebounds for their team in a game. Therefore, many data points with a low amount of rebounds but a high voting share are typically guards. So, since the stat is very significant to forwards and centers, it must be included in any model.

d. The impact of steals per game

Figure 1.7

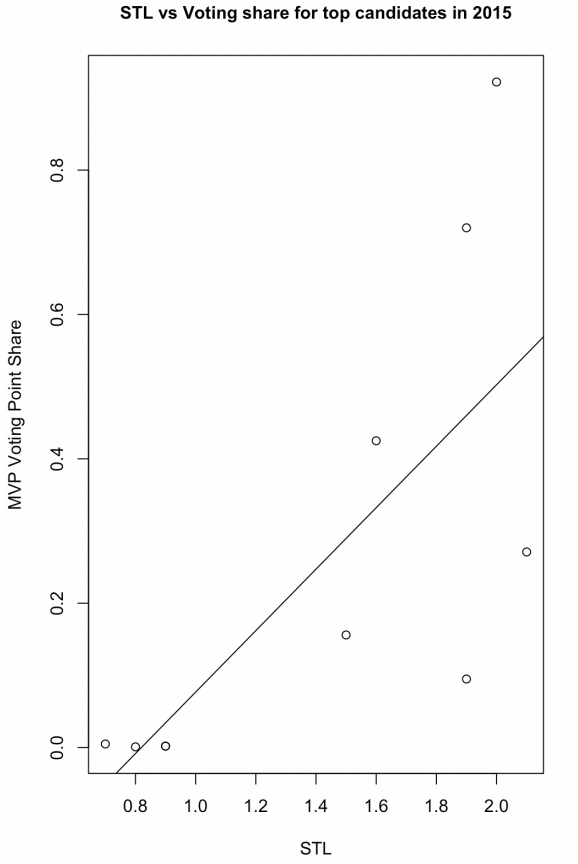
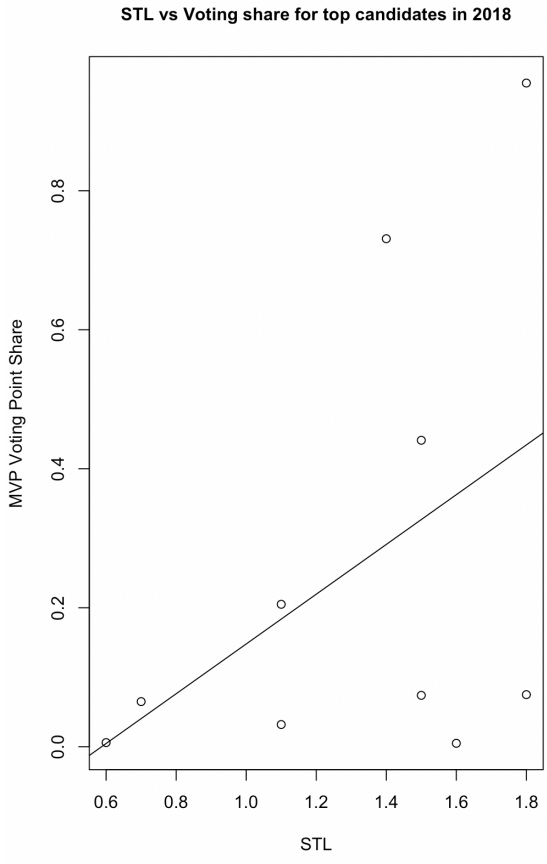


Figure 1.8



Steals per game is another statistic significant to guards, thus resulting in a weaker correlation. The r^2 value for 2015 was .5067 (Figure 1.7), and just .2005 for 2018 (Figure 1.8). Steals are generally obtained by guards because they are faster, tend to intercept passes, and are

usually guarding the ball handler. Despite the weak correlation, steals are key in determining the defensive impact of a player, which is very hard to do numerically. Therefore, steals should be a factor in any equation moving forward.

e. The impact of blocks per game

Figure 1.9

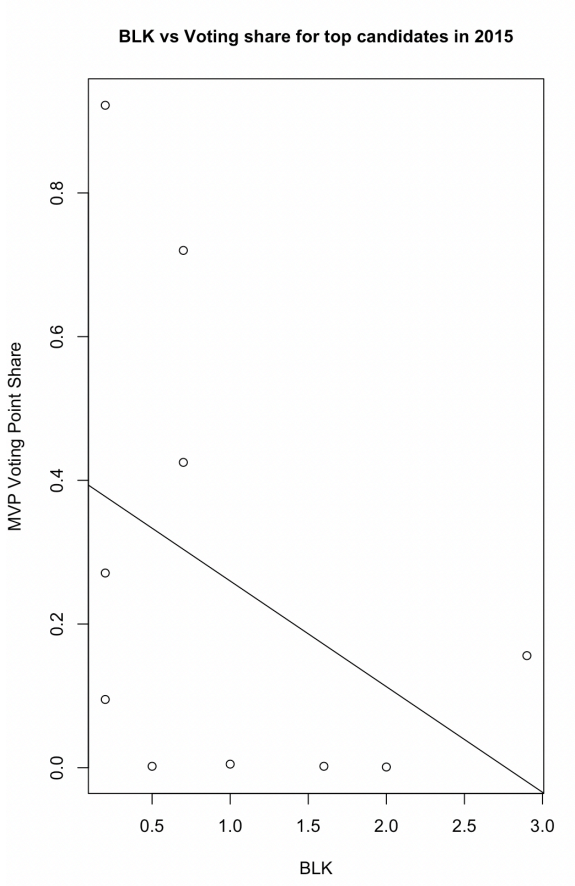
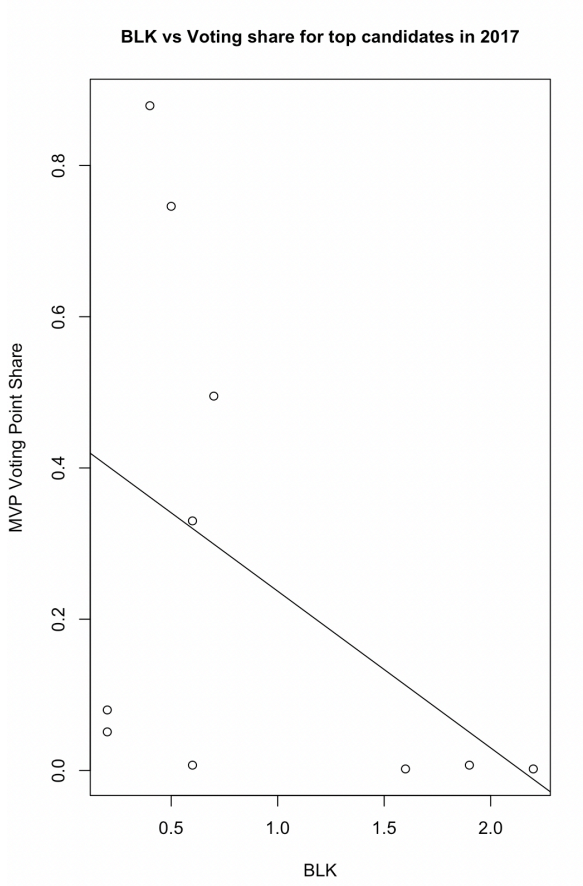


Figure 1.10



The correlation between blocks and voting share is the most surprising by far. It is counterintuitive to think that the more blocks you have, the less votes you will receive. However, since this trend is prominent over the past ten years, there must be an explanation. The r^2 values for each year are still relatively low, with a value of .1614 in Figure 1.9 and a value of .2024 in Figure 1.10. So, what could possibly be the reason for this weak negative correlation? First, the

players with the most blocks are almost always centers. Over the past ten years, only one center has won the award. Therefore, this negative trend continues to suggest that centers tend to not win the MVP award. On top of that, the players with the most blocks are usually candidates for the Defensive Player of the Year award. These candidates tend to be strong on defense, but not as strong on offense. So, these defensive players receive some MVP votes for their defensive prowess, but their offensive impact does not push them over the top. Overall, the negative correlation between blocks and voting share shows that defensive players and centers rarely win the award.

f. The impact of field goal %

Figure 1.11

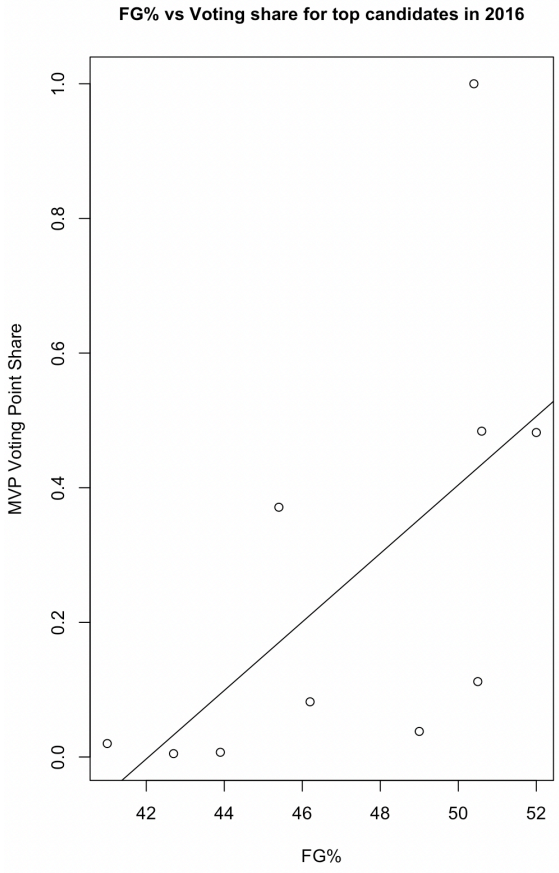
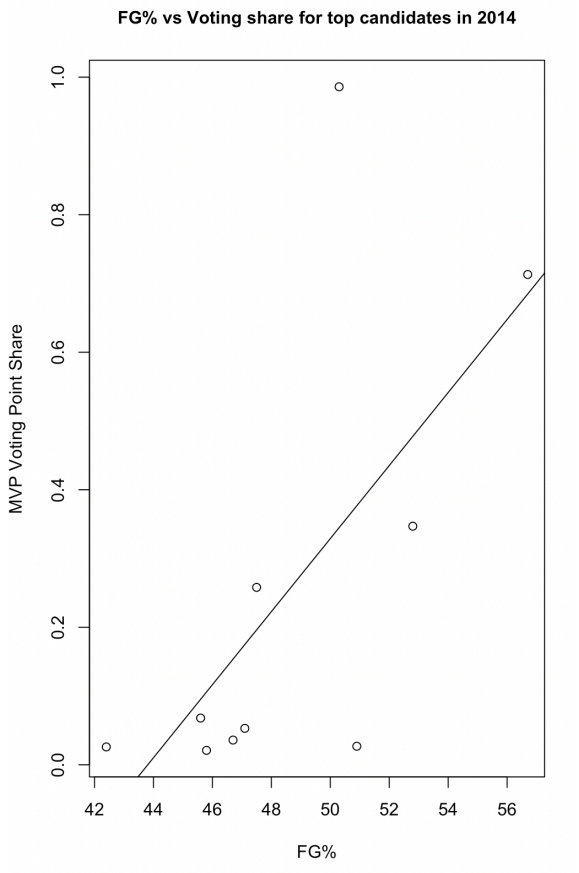


Figure 1.12



Field goal percentage is a key statistic in observing a player, because it measures that player's efficiency. A low field goal percentage means more missed shots and empty possessions, whereas a high field goal percentage means more made shots and scoring possessions. This ultimately affects the game for the entire team. As shown in Figures 1.11 and 1.12, there is a positive linear correlation between field goal percentage and voting share. While the r^2 values are moderately weak (.4188 for 2014 and .3622 for 2016), they are still greater than the r^2 values for many other statistics. This shows that a player's efficiency is essential to determining the impact they have on their team and if they will become the MVP.

g. The impact of win/loss differential

Figure 1.13

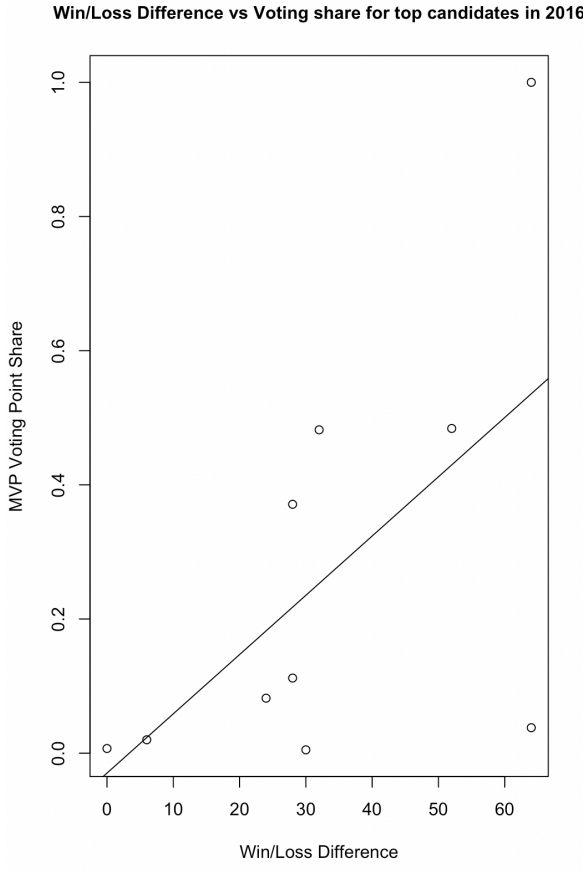
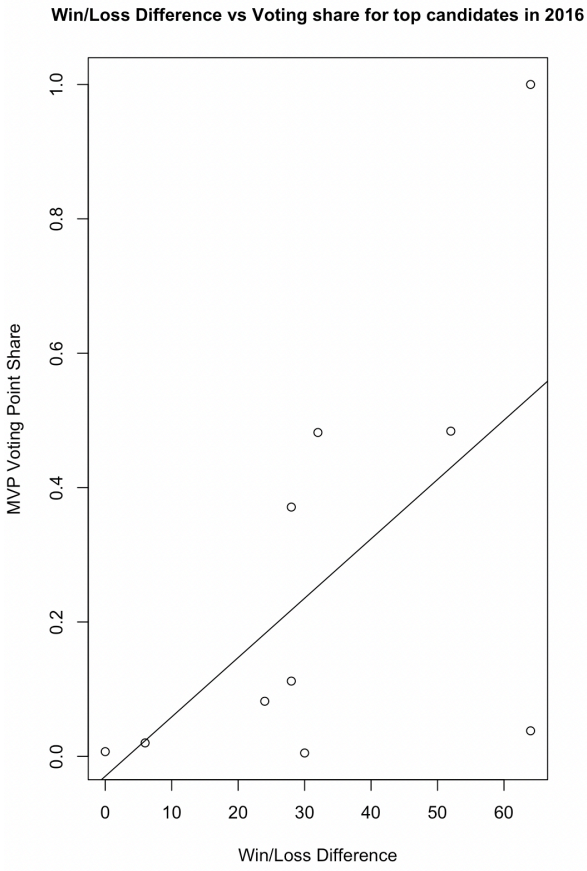


Figure 1.14



The win/loss differential represents the candidate's team's overall success, based on their record in the regular season. Figures 1.13 and 1.14 support the notion that the better a player's team does, the more votes they are likely to receive. With the r^2 values equal to .3489 for 2016 and .3778 for 2015, this positive linear correlation is stronger than most other statistics, and is very similar to that of field goal percentage. Therefore, the overall team's success of a player affects their chances of winning the award.

Overall, I have confirmed that the seven statistics presented above are in influential in determining the MVP. These seven categories (points, assists, rebounds, steals, blocks, fg%, win/loss differential) will be the factors in the equations I create.

V. Approach 1: Bivariate linear regression

The first approach I took to creating a mathematical model to predict who wins the award was based on creating a linear regression model between a number representing all of a player's stats and their voting share. But how would I create one number that could represent seven categories?

First, I took the mean and standard deviation of all of the statistics for each year. From these calculations, I realized that I could measure a player's impact by comparing him to the other candidates in that same year. So, I calculated the z-scores for each player in every statistic for every year. See Figure 2.0 for an example of the dataset.

Figure 2.0. - Z-Scores for each player in each stat, 2014

Name	Z-Score PTS	Z-Score AST	Z-Score REB	Z-Score STL	Z-Score BLK	Z-Score FG%	Z-Score Win/Loss
Kevin Durant	1.752110	-0.1103481	0.0150513	-0.3501868	0.1761660	0.4161073	1.336781
LeBron James	0.7874653	-0.2883290	0.3160784	0.2677899	-0.7046642	1.964413	0.2587318
Blake Griffin	0.1968663	0.6371716	-0.5870028	-0.5561791	-0.04404151	1.020914	0.90556
Joakim Noah	-2.067096	1.277902	-0.02257703	-0.5561791	1.937826	-0.2612767	-1.034927
James Harden	0.4527925	-1.071445	0.2408217	0.2677899	-0.4844566	-0.7209302	0.2587318

Stephen Curry	0.1771796	-1.213829	1.143903	0.2677899	-0.9248718	-0.3580458	-0.3880977
Chris Paul	-0.7874653	-1.213829	1.971727	2.121720	-1.145079	-0.4548150	-0.1724878
Al Jefferson	-0.2559262	1.099921	-1.264313	-1.174155	1.056996	0.5612611	-2.112976
Paul George	-0.2756128	-0.3239252	-0.7375164	0.8857667	-0.7046642	-1.495083	0.6899514
LaMarcus Aldridge	0.0196866	1.2067105	-1.076171	-1.174155	0.8367888	-0.6725456	0.2587318

Then, I added up all of the z-scores for each player to generate a new sum. This sum demonstrates where a player lies compared to other players (above average = + sum, below average = - sum). The sums for the players above are listed in Figure 2.1.

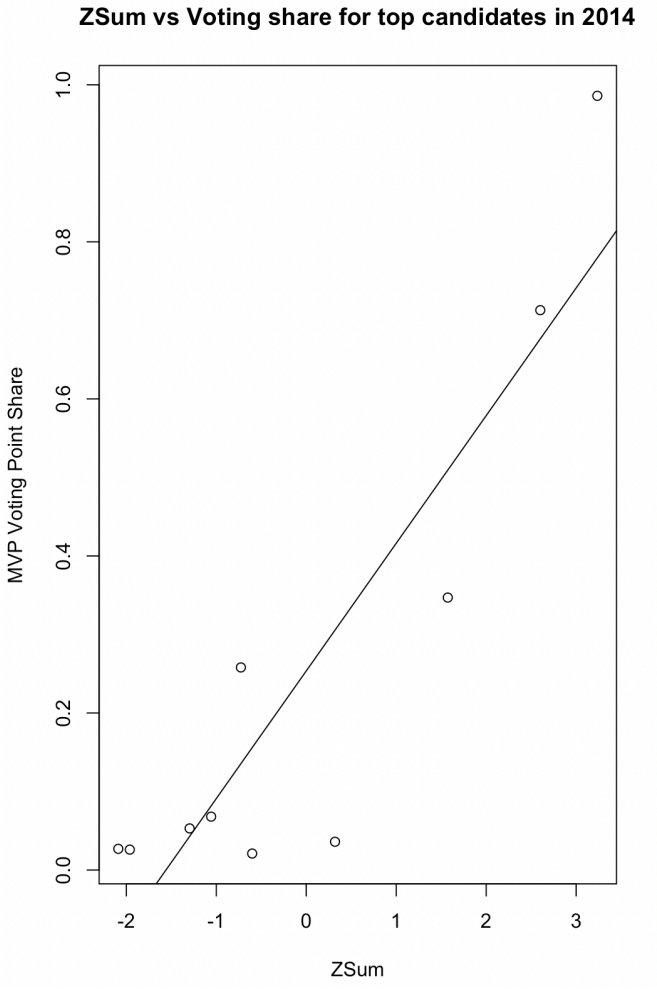
Figure 2.1

Name	Z Sum	Voting Rank
Kevin Durant	3.235681133	1
LeBron James	2.601486155	2
Blake Griffin	1.573290385	3
Joakim Noah	-0.7263269669	4
James Harden	-1.056695942	5
Stephen Curry	-1.295972489	6
Chris Paul	0.3197706089	7
Al Jefferson	-2.08919302	8

Paul George	-1.961084089	9
LaMarcus Aldridge	-0.6009557753	10

From the data, it is clear that the players with the higher Z sum tend to have more votes. The top three candidates all have high positive values, which shows that they are above average in most of the statistical categories. Now that a relationship has seemingly arisen, let's create a linear regression model to see the correlation for the data in 2014.

Figure 2.2



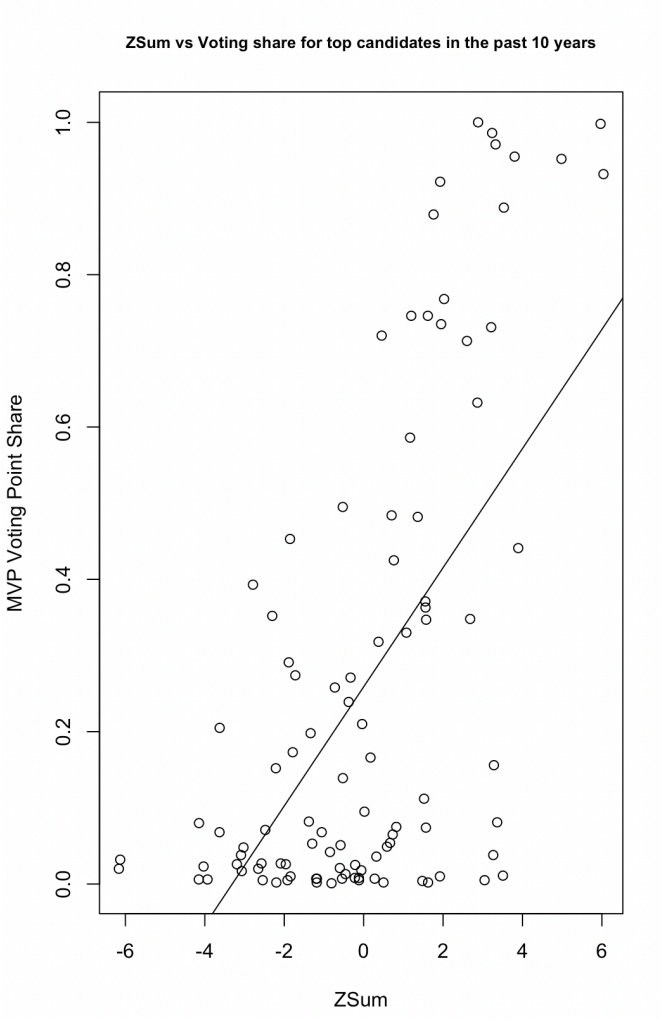
The graph in Figure 2.2 clearly shows a strong linear correlation between the Z Sum and the voting share. In fact, the r^2 value was .8092, the highest value across all of the regression models I have created so far. This demonstrates that this line can accurately represent about 80% of the data, which is very high in the human sciences. This line can be expressed by the following equation:

$$y = .25350 + .16256x$$

Although the correlation between the Z Sum and voting share is very strong, this

equation only tells us how many votes a player would receive in 2014. Therefore, we need to perform a regression analysis using the entire dataset. The Z sum can be calculated for the entire dataset, however it is important to calculate each z-score with respect to only the players in each year. This is because when analysts vote, they examine players in comparison to the other players in that same season. Therefore, the model should follow that same approach.

Figure 2.3



In Figure 2.3 there seems to be a distinct positive correlation between a player's Z Sum and their voting share. However, the data still seems to have high variability. The r^2 value for the data is .3737, which can be defined as moderately-weak correlation in the human sciences. The major issue with using the Z Sum to predict voting share is that players at different positions naturally have high z scores in some areas and low z scores in others. For example, many guards will have a high z score for assists, since it is a guard dominant statistic. However, guards will also have very low z scores for rebounds, since it is a center dominated stat. This is an issue

because stats out of the ordinary are being canceled out. For example, let's say the mean for both assists and rebounds in a dataset is 5, and the standard deviations are both 2. Guard #1 averages 5 assists and 5 rebounds a game, giving his Z Sum for these two statistics as 0. However, Guard #2 averages 8 assists a game but only 2 rebounds. His Z Sum would still be 0, despite having a very high total in assists. Since rebounds are a statistic not as important to guards, Guard #2 should receive a lot more recognition, due to his high assists per game. However, the Z Sum method does not do certain players justice for exceeding the norm. Therefore, another approach to modeling this relationship should be determined.

VI. Approach 2: Multivariate linear regression

Multivariate linear regression models allow for all factors to have an effect on the predicted value. While the Z Sum approach canceled out important data values, the regression model will have a coefficient for each factor to determine the impact on the voting share. Let's examine a multivariate regression model for a single year first.

The following regression model was calculated through the use of the program R, using the data from 2012.

Equation of the regression line:

$$y = -3.8085 + .08530x_1 + .17441x_2 + .07778x_3 - .13880x_4 - .02320x_5 + .01291x_6 + .01303x_7$$

Where each x represents to corresponding factors:

Figure 3.0

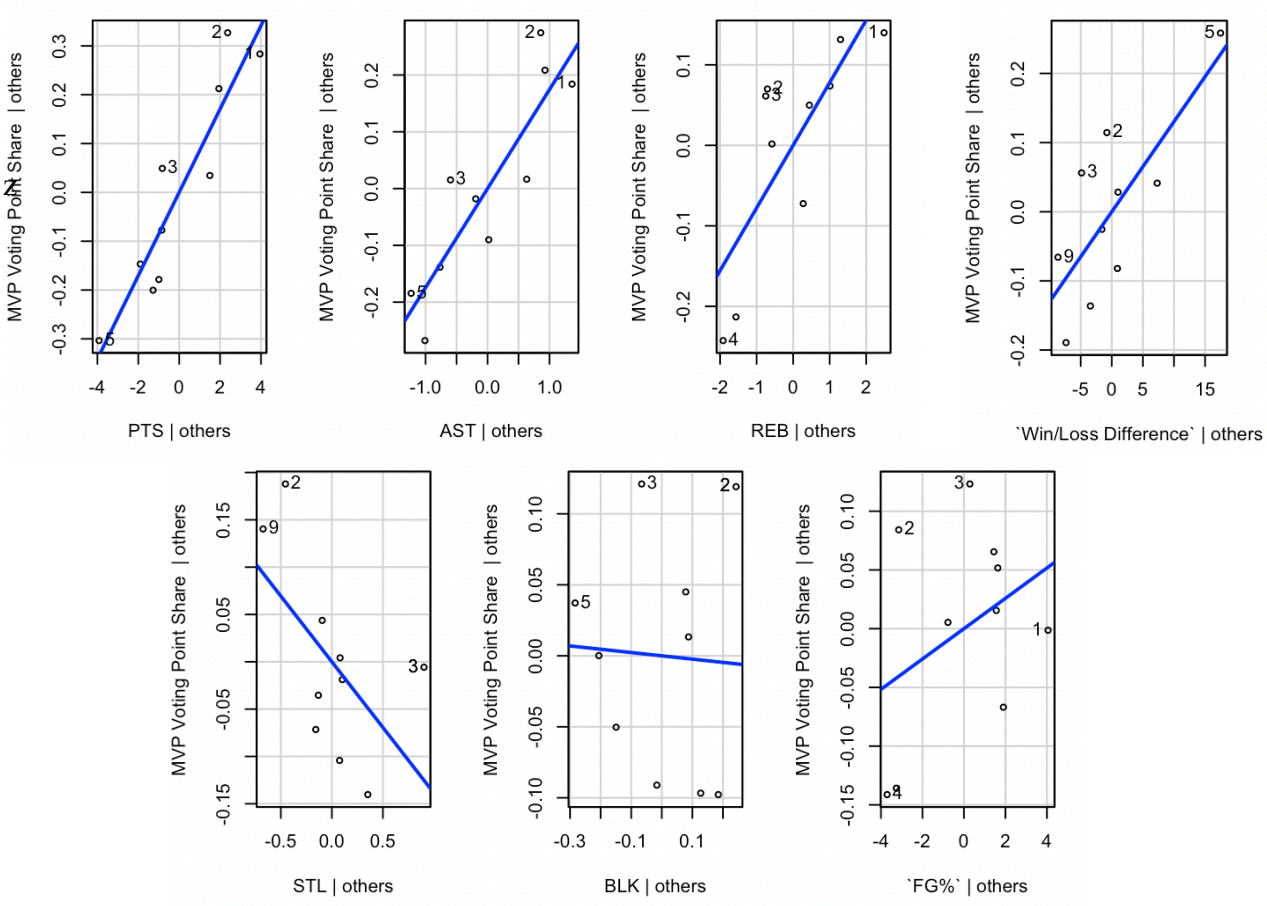
x_1	Points
x_2	Assists
x_3	Rebounds
x_4	Steals
x_5	Blocks
x_6	FG%
x_7	Win/Loss Diff.

The regression line is useful in a variety of ways. First, it can provide an accurate prediction for other players' voting share in this year. In fact, the r^2 value is .9328, which is extremely strong in any area of knowledge, not just the human sciences. On top of that, the coefficients of each x variable show how much a certain factor affects the dependent variable. In this case, the highest coefficient is .17441 for the x variable corresponding to assists, showing that assists were a very important statistic in this year's voting. The lowest coefficient is .01291 for the x variable corresponding to field goal percentage, demonstrating that this statistic was not a strong factor in the MVP decision. However , the coefficient alone cannot illustrate the effect a predictor has on the response variable. This is

because each predictor has a different range of values. A player's voting share must be from 0 to 1, which means that the predictor variables must be scaled down to fit this range. For instance, just because the coefficient for assists is greater than the coefficient for points does not mean that assists have a bigger impact on voting share. Points tend to be on a scale from around 20 to 30, whereas assists are on a scale from around 2 to 10. Therefore the coefficient for points needs to be small, in order to scale the numbers down to fit the range for voting share. But, the effect each factor has on the dependent variable can be visualized using added variable plots. These plots show the relationship between one factor and the response variable, while controlling the other factors.

Figure 3.1

Added Variable Plots for 2012 Regression Model



The added variable plots clearly show the relationship between each factor and the voting share. Each correlation between each variable and the response variable has the same sign as the x coefficient. For example, the regression line on the added variable plot for points has a positive slope, therefore the x_1 coefficient is positive. Also, the strength of the correlation for each individual plot can be related to the effect that a factor has on the voting share. For instance, the added variable plot for assists has a strong correlation, therefore the coefficient for x_2 will be high. Overall, multivariate regression is able to create an accurate prediction model, which portrays the effect that each predictor has on the response variable. However, the model that was created only used data from 2012, which means that the regression line can only accurately predict the voting share a player got in 2012. So, a multivariate regression model must be created for the entire dataset.

The following regression model was calculated through the use of the program R, using the data from 2012-2021. The x values correspond to the same factors as in Figure 3.0.

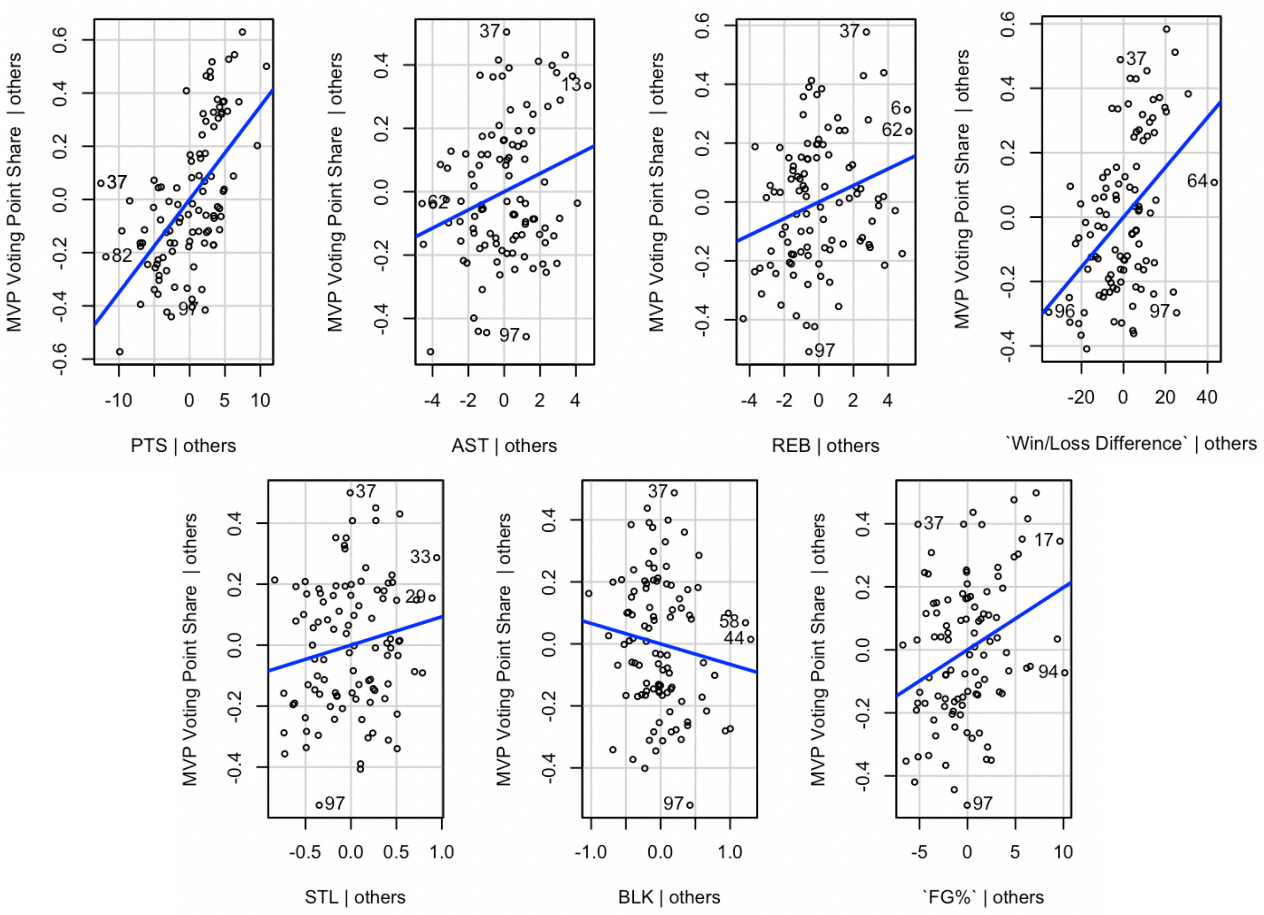
Equation of the regression line:

$$y = -2.21726 + .035019x_1 + .028741x_2 + .028119x_3 + .093009x_4 + .066039x_5 + .019789x_6 + .007734x_7$$

The initial observations of the regression line coincide with conclusions previously made. For example, the coefficients of the x values for both assists and rebounds are almost exactly the same. Given that the statistic is represented in the same range (typically 2-10), this shows that they have an equal effect on who wins MVP. The added variable plots can be studied to view other relationships similar to this one.

Figure 3.2

Added Variable Plots for Regression Model Overall



The majority of the variable plots seem to have a moderately weak to weak correlation. While this makes sense in which human behavior will vary from year to year, the weak correlation demonstrates that it is difficult to predict a winner using regression. However, the r^2 value of the data is .5365, which can be deemed moderate in the human sciences. On top of that, this r^2 value is higher than the r^2 for the bivariate regression model with Z Sums (.3737). Therefore, the multivariate linear regression model is the most accurate predictor I have found so far. In order to test this new model, data from the 2010 MVP race can be used to see if this model will produce correct predictions.

VII. Assessing and applying my model

In order to assess the success of my model, data from the three top candidates in 2010 will be used to determine if accurate predictions are measured. Data from the 2010 season was chosen to test my model because of the inability to use data that exists in the dataset used to create the least squares regression line.

Figure 4.1

Name	PTS	AST	REB	STL	BLK	FG%	Win/ Loss	Voting Share
LeBron James	29.7	8.6	7.3	1.6	1.0	50.3	40	.980
Kevin Durant	30.1	2.8	7.6	1.4	1.0	47.6	18	.490
Kobe Bryant	27.0	5.0	5.4	1.5	0.3	45.6	32	.487

This data can be entered into the regression equation, to get a prediction of voting share. This prediction will then be compared to the actual value.

LeBron James:

y = -2.21726 + .035019(29.7) + .028741(8.6) + .028119(7.3) + .093009(1.6) + .066039(1.0) + .019789(50.3) + .007734(40)

= .795

Kevin Durant:

y = -2.21726 + .035019(30.1) + .028741(2.8) + .028119(7.6) + .093009(1.4) + .066039(1.0) + .019789(47.6) + .007734(18)

= .408

Kobe Bryant

y = -2.21726 + .035019(27.0) + .028741(5.0) + .028119(5.4) + .093009(1.5) + .066039(0.3) + .019789(45.6) + .007734(32)

= .333

The values predicted by the regression model accurately correlate to the player's predicted finish in the MVP race. LeBron James had the highest predicted voting share by far, which makes sense considering he won the award. A voting share of .795 sits right in between the common voting shares for first and the high end of second place. This shows that LeBron has a high chance of placing either first or second. For Kevin Durant, a voting share of .408 lies right in between second and third place, thus giving an accurate prediction of how he finished. Finally, Kobe Bryant's projection of .333 voting shares sits in between third and fourth place, coinciding with his finish as well. On top of that, the reason why all of the values are lower than their actual values is due to the regression effect. So, even though the values are being pulled back towards the mean, they still hold a firm prediction.

Now that the model has been confirmed to be successful, I can apply it to the 2021-22 NBA season to create my first prediction.

Data for the top 5 candidates in the MVP race via nba.com as of January 13th, 2022:

Figure 4.2

Name	PTS	AST	REB	STL	BLK	FG%	Prj. Win/ Loss
Giannis Antetokounmpo	28.4	6.0	11.4	1.1	1.4	54.1	17.2
Kevin Durant	29.7	5.9	7.5	0.8	0.9	52.0	24.6
Nikola Jokic	25.7	7.0	14.1	1.4	0.8	56.3	3.0
Joel Embiid	27.1	4.3	10.5	1.1	1.4	48.2	12.3

Stephen Curry	26.8	6.1	5.5	1.4	0.5	42.0	41.0

$$y = -2.21726 + .035019x_1 + .028741x_2 + .028119x_3 + .093009x_4 + .066039x_5 + .019789x_6 + .007734x_7$$

Here are the candidates ranked in order of highest projected voting share:

Figure 4.3

Rank	Name	Voting Share
1	Giannis Antetokounmpo	.669
2	Nikola Jokic	.601
3	Kevin Durant	.556
4	Joel Embiid	.394
5	Stephen Curry	.363

Based on the data in figure 4.3, I can conclude that as of January 13th, 2022, Giannis Antetokounmpo, Nikola Jokic, and Kevin Durant have the best chances at winning the MVP award. However, this does not mean that other candidates should be counted out. If Joel Embiid, Stephen Curry, or another player in the NBA have a phenomenal streak of form, the odds could easily shift. So, the race to the MVP award should be closely monitored.

VIII. Conclusion

The goal of this paper was to find a mathematical model that could accurately predict the NBA Most Valuable Player through the use of statistical analysis. Through the progress of different approaches, I can conclude that this goal was achieved. The final model created through multivariate linear regression produces accurate predictions that can be used to determine a player's finish in the MVP ladder. This was justified through the use of data in the 2010 season, to see how close the predicted values were to the actuals. I was able to then calculate a prediction for the 2021-22 season.

However, the process that led to this point was the driving force behind the success of this model. By proving the equation for the least squares regression line, I was able to start conducting my research by determining the most important factors that go into the MVP decision. Through this bivariate statistical analysis, seven key predictor variables were determined, and the modeling process could begin. The first approach that was taken, using the Z Sum, seemed to be effective when using data from one season. However, when the dataset was expanded to the past ten years, problems began to arise. This led to the modeling of multivariate linear regression, which ultimately proved fruitful to the question at hand. Due to the complexity of the model I derived, I decided to use the program R to conduct my regression analysis. This allowed me to create a complex model easily and efficiently. As a result, the model created through a multivariate regression equation was more accurate than the bivariate model. So, the multivariate model was ultimately used for my final predictions.

Through this process, I learned how to propose and answer a research question in mathematics. I expanded my knowledge in the field by deriving equations I had not seen before, and I hope to apply this determination to other research questions in the future. Overall, I was not only able to create a successful model that predicted the NBA MVP, but was also able to discover how I can succeed as a researcher.

Bibliography

Sean Lahman. “Basketball Statistics and History.” *Basketball Reference*, www.basketball-reference.com/.

Venables, W. N., and D. M. Smith. "An Introduction to R." Cran.r-project.org. August 10, 2021. Accessed October 29, 2021. https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf.

YouTube. June 06, 2019. Accessed October 29, 2021. https://www.youtube.com/watch?v=_V8eKsto3Ug.