PREDICTING NBA PLAYERS' SUCCESS USING ROOKIE YEAR STATISTICS

Statistics 101

Aryaman Babber

12/18/19

word count: 2,094

Table of Contents

1. INTRODUCTION

Professional sports has changed dramatically over the last 20 years. When teams were choosing players to draft or trade for, recruiters would watch and choose the players they want based on what they see and on their "gut feeling". Recently, Data Science has changed that. Instead of using scouts' beliefs to choose which players will be the most beneficial to their team, teams have started trusting statistics to determine every players' potential and likelihood of succeeding.

In this report, we will apply statistical methods to predict NBA players' success in the NBA. Specifically, we will create a multiple regression model using NBA players' rookie year statistics, and use this model to determine how successful those players will be in their whole NBA career.

2. LITERATURE REVIEW AND HYPOTHESIS

2.1 Researching Predicting Models and Traditional vs Advanced Statistics

In order to choose our predictor and response variables, I researched different existing NBA predicting models and professional analysts' articles and reports. Before explaining these models, it is important to understand the statistics of basketball.

In basketball, there are two different types of statistics: traditional (basic) statistics and advanced statistics. Traditional statistics are the statistics that we can directly measure from a basketball game, such as points per game, successful/made field goals, attempted field goals, etc. These statistics mainly come in three different forms: per game, per 48 minutes, and per 100 possessions. Data scientists and mathematicians have used these simple statistics, and created new equations, producing advanced statistics, such as player efficiency rating (PER), plus/minus (+/-), win shares (w/s), rebound/assist/steal percentage, etc. These advanced statistics usually reveal more about how effective a certain player is. For example, let us look at rebounds. While the number of rebounds per game (traditional) may tell us how many rebounds a player averages every game, the rebounding percentage (advanced statistic) tells us the percentage of available rebounds a player grabs. This statistic is much more effective, because it takes into play the number of times there was a potential rebound. Guards (players who usually stay further away from the basketball hoop) will average much fewer rebounds per game than forwards or centers (players generally closer to the hoop). Rebounding percentage levels this playing field and provides a better analysis of how effective and beneficial a player is.

Due to this advantage, almost all, if not all, predicting models opt to use advanced statistics. We will be choosing the predictor variables as a combination of traditional and advanced statistics, and the response variable will be a traditional statistics.

2.2 Choosing Predictor and Response Variables

Using news articles discussing different statistics in the NBA, as well as existing predicting models, I learnt that some highly valued statistics are true shooting percentage (TSP), value of replacement (VORP), and defensive box plus/minus (DBPM). While these statistics are generally part of a much more complex formula, we will use only these three variables as our predictor variables. We will use these variables to predict how many minutes are played (MP) in the players' entire NBA career. The higher the number of MP, the more valuable we consider that player. We are using MP because it takes into consideration both how successful the players are individually, as they have earned the right to stay in the game, and how successful they are at taking their teams to the playoffs and finals, as these games will lead to more MP.

** It is important to note that all predictor variables are measures of players' rookie season statistics, while the response variable is a measure of the players' entire career in the NBA*

2.3 Hypotheses

We hypothesize that as TSP, VORP, and DBPM increase, the MP will also increase. TSP is a measure for how effective a player is at shooting (how accurate they are, how many high–quality shots they take, etc.), so the greater this statistic, the greater MP we expect. As VORP increases, the value of the player increases, so we predict that as VORP increases, MP will increase.

Similarly, as DBPM increases, the better the player is on defense, leading to a higher MP over a player's career.

For these predictor variables, our null hypotheses are that the coefficients (β) are equal to 0, and our alternative hypotheses are that the coefficients are greater than 0. These are one–tail upper–tail hypotheses. We will use an α of 0.05.

3. DATA COLLECTION AND ANALYSIS

3.1 Collecting the Data

I found data online which provides all NBA players' seasonal data from 1950 to 2017. I created a subset of this dataset, focusing on ten years of rookies, starting from the 1998–1999 season, going to the 2006–2007 season. This subset included all of these players' rookie season statistics. I created another subset containing these same players, but data of their entire NBA career. For the statistics in this dataset, I found the average of all of their seasons. For example, if someone played in the NBA for ten years, their TSP would be the average of their TSP from each season. I then created a third dataset that merged these two. I analyze this third dataset to see how the rookie season statistics will affect the NBA career statistics. This dataset includes 691 different observations (different players).

3.2 Explaining the Predictor and Response Variables

It is important to note the units of each predictor and response variable. The unit of TSP is the percentage of total shots made, treating free throws, two–pointers, and three–pointers equally. VORP is a representation of how valuable a player is over a replacement player. In other words, it estimates the player's overall contribution to the team. This is calculated by using the Box Plus/Minus (BPM), percent of minutes played, and number of team games. The average player in the NBA will have a VORP of 0, meaning any player with a lower VORP is below average and any player with a higher VORP is above average. DBPM is similar to BPM, but instead focuses on the defensive statistics to see how a player performed defensively. Similar to VORP, the average player is expected to get a DBPM of 0, worse than average below 0, and better than average above 0. Both VORP and DBPM are measured per 100 possessions. The response

variable MP is the total number of minutes played in players' NBA careers, and is measured in minutes.

3.3 Analyzing the Distributions and Plots of Variables

Looking at the density plot of TSP, we can see that the data is approximately symmetric and unimodal. The boxplot reveals many outliers of this variable. However, looking at the summary statistics, we see that the difference between the mean and median is only 0.16 (4.67 and 4.83 respectively), showing that the outliers are not significantly affecting the data. The standard deviation is only 0.12, and the IQR is 0.10.

Next, we will look at the density plot and box plot of VORP. We can see that this distribution is also roughly symmetric, unimodal, and contains many outliers. As we can see from the summary table, the mean and median (0.00 and -1.00 respectively) difference is slightly greater with value of 1. However, it is still relatively small, as is the standard deviation and IQR with a value of 0.71 and 0.30 respectively.

The density plot of BDPM is unimodal, but unlike the others, is slightly skewed to the left. The box plot again shows many outliers. The summary table shows a mean value of -1.23 and a median value of -1.10. Although the difference is tremendously smaller than the last variable (0.13), the standard deviation (2.69) and IQR (2.50,) show a greater variance in the data than the previous two variables.

The distribution and data of MP is very different from the previous variables. The density plot, although unimodal, is very right–skewed. This is likely due to many rookie players playing very few games in the NBA their first year, and playing the majority in the D–League, or not receiving any playing time over the veterans. This shows in the data, as the mean is over 150 minutes greater than the median. The standard deviation is also very large, with a value of 654.50, and an even larger IQR of 1,024.49. Interestingly, the box plot shows just one outlier in this data, a high outlier. This outlier represents Lebron James, someone who is considered one of the best players of all time, if not the best player of all time.

We will include all outliers in all the variables, because if we exclude any, then it is not an accurate representation of the data. Every single player should be considered equally when analyzing this data.

3.4 Analyzing the Correlation Matrix

Before we look at the matrix, let us make sure these variables pass the conditions to use a Pearson Correlation Matrix. All the variables are continuous quantitative variables, passing the Quantitative Variables Condition. The distribution of MP fails the No Outliers Condition, but the three predictor variables pass, as their outliers are not influential. We will analyze the scatterplots when creating the linear model.

Looking at the correlation matrix, we can see that MP has a moderately low correlation to VORP and DBPM, with a slightly stronger, but still weak, correlation to TSP. TSP has a low correlation to VORP and DBPM, while VORP and DBPM have a stronger correlation than the other

predictor variables, but still a moderately weak one. This correlation is because DBPM is part of what is used when calculating VORP, as mentioned earlier. I expect this relationship between DBPM and VORP to potentially cause a lower R^2 value.

Almost all, if not all, other predictor models use much more advanced statistical methods (including machine learning) and much more complicated predictor and response variables to create their models. This is to create stronger correlations to the response variables, resulting in more accurate predictor models.

4. MULTIPLE REGRESSION MODEL

4.1 Interpreting the Multiple Regression Coefficients

Using TSP, VORP, and DBPM as our predictor variables for MP, we result in the multiple regression model and the following multiple regression equation.

$$\widehat{MP} = 29.07 + 1{,}104.51(TSP) + 321.77(VORP) + 40.76(DBPM)$$

Interpreting the intercept, it means that when all other variables are zero, then the predicted minutes played is 29.07. The second coefficient tells us that, for every one extra percent true shooting increases, the predicted number of minutes played in an NBA career increases by 1,104.51. The third coefficient tells us that, for every one extra percent the value over replacement player increases, the predicted number of minutes played increases by 321.77. The fourth coefficient tells us that, for every one extra unit Defense Box Plus/Minus increases, the predicted number of minutes played increases by 40.76. This model has a multiplied R^2 of 0.2547. This R^2 means that 25.47% of the variance in MP can be predicted by TSP, VORP, and DBPM.

Looking at the p–values for the variables in the multiple regression model, we see that all p–values are less than our alpha of 0.05, indicating that predictor variables are statistically significant. Practically, we see that the coefficient for TSP makes a tremendous difference, as does the coefficient for VORP. The coefficient for DBPM, however, is a mere 40.76. Considering the range of DBPM in our data is roughly 28, it is very difficult to increase the DBPM by a full unit, and doing so will lead to a very small increase, especially compared to the

other predictor variables. Therefore, we may conclude that TSP and VORP are both statistically significant and practically significant, but DBPM is only statistically significant. Before we can officially state this, however, we need to check the conditions for a multiple regression model.

4.2 Checking the Conditions for the Model

Looking at the residual plot, we cannot see any clear pattern in the data. However, it clearly does not follow the Plot Thickens Condition, failing the Equal Variance Assumption. The data is also not well–distributed around the x–axis, but instead largely clumped together. The magnitude of the standard error of the residual plot is very large, with a value of 565.5, making this a bad model. Seeing the partial regression plots, all three plots do not pass the Straight Enough Condition as they are not close to the predicted values. This fails the Linearity Assumption.

Although all our p–values are less than our alpha, this model does not pass the required conditions, so we fail to reject all our null hypotheses.

5. Conclusion

We fail to reject our null hypotheses that TSP, VORP, and DBPM are statistically significant to MP. This means that we do not have enough evidence to reject our null hypotheses.

While I do believe these variables are significant in predicting a players' future success in the NBA, a more complex and advanced model is needed. Instead of using a multiple regression model, a gradient descent model (which uses machine learning) may have been more accurate. In addition, I do not believe it is enough to use a single response variable, such as MP, to summarize how successful a player's career in the NBA is. A more complicated mathematical equation needs to be formed that better summarizes this.

6. Works Cited

Goldstein, Omri. “NBA Players Stats since 1950.” Kaggle, 27 Apr. 2018, www.kaggle.com/drgilermo/nba-players-stats#Seasons_Stats.csv.

Khan, Ehran. “Advanced NBA Stats for Dummies: How to Understand the New Hoops Math.” *Bleacher Report*, Bleacher Report, 3 Oct. 2017, bleacherreport.com/articles/1813902-advanced-nba-stats-for-dummies-how-to-understand-the-new-hoops-math#slide10.

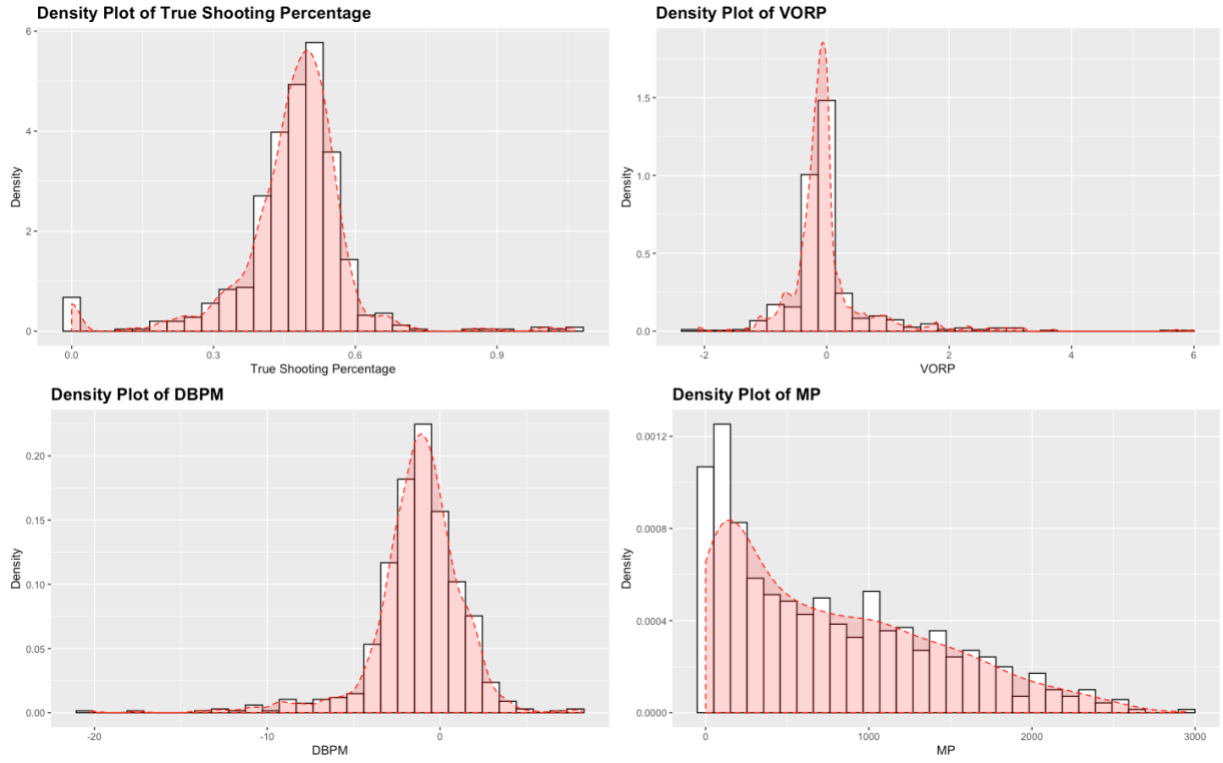
Fromal, Adam. “Understanding the NBA: Explaining Advanced Defensive Stats and Metrics.” *Bleacher Report*, Bleacher Report, 3 Oct. 2017, bleacherreport.com/articles/1040309-understanding-the-nba-explaining-advanced-defensive-stats-and-metrics#slide1.

“About Box Plus/Minus (BPM).” Basketball Reference, Basketball Reference, www.basketball-reference.com/about/bpm.html.

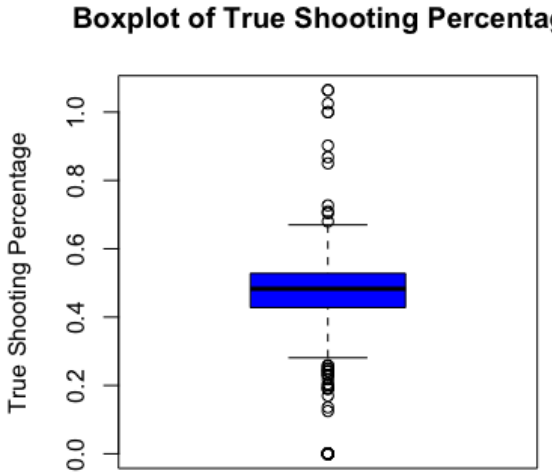
Fromal, Adam. “Understanding the NBA: Explaining Advanced Comprehensive Stats and Metrics.” Bleacher Report, Bleacher Report, 3 Oct. 2017, bleacherreport.com/articles/1040320-understanding-the-nba-explaining-advanced-comprehensive-stats-and-metrics#slide0.

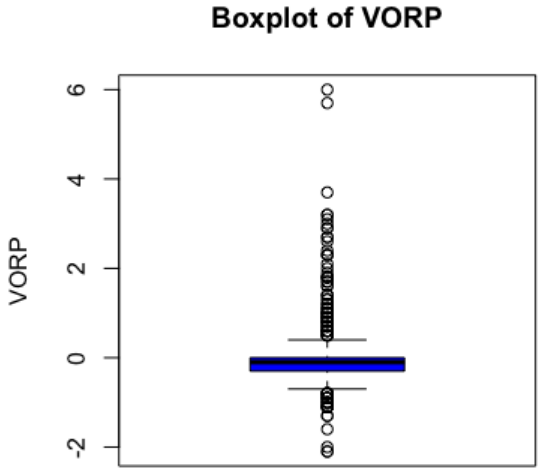
7. APPENDIX

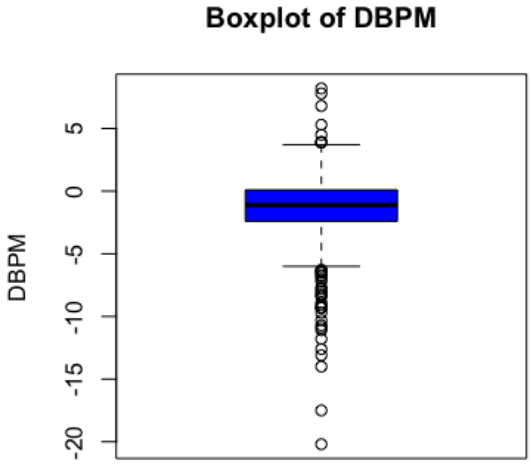
7.1 Density Plots of Predictor and Response Variables

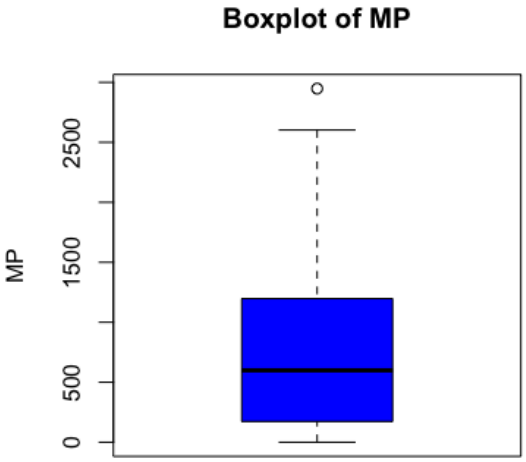


7.2 Box Plots of Predictor and Response Variables









7.3 Summary Statistics Tables

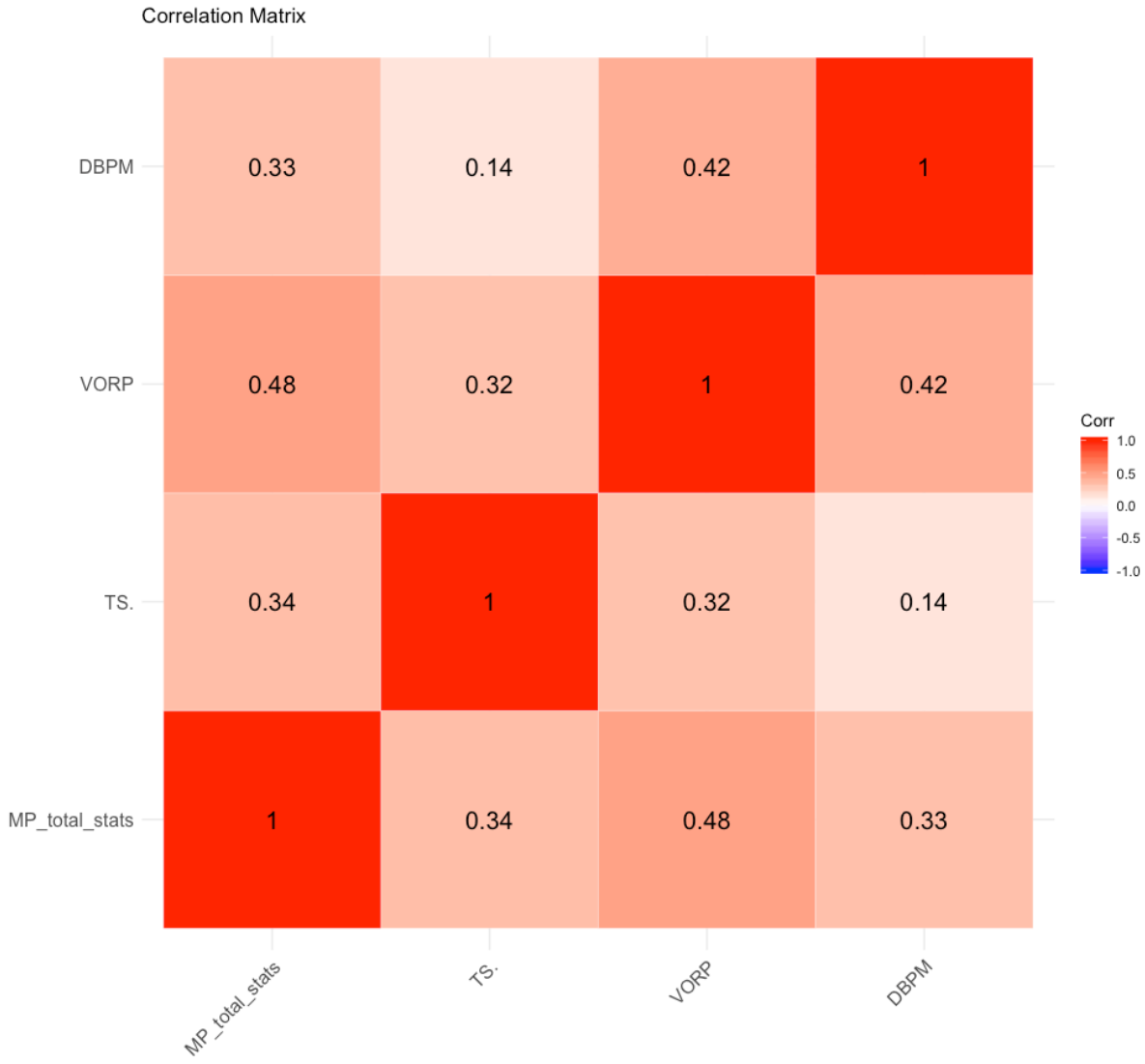
Summary Statistics for TSP	
	TSP
Mean	4.67
Median	4.83
Standard Deviation	0.12
Q1	0.43
Q3	0.53
IQR	0.10
Min	0.00
Max	1.06

Summary Statistics for VORP	
	VORP
Mean	0.00
Median	-1.00
Standard Deviation	0.71
Q1	-0.30
Q3	0.00
IQR	0.30
Min	-2.10
Max	6.00

Summary Statistics for DBPM	
	DBPM
Mean	-1.23
Median	-1.10
Standard Deviation	2.69
Q1	-2.40
Q3	0.10
IQR	2.50
Min	-20.20
Max	8.20

Summary Statistics for MP	
	MP
Mean	755.20
Median	599.40
Standard Deviation	654.50
Q1	172.90
Q3	1,197.40
IQR	1,024.49
Min	0.00
Max	2,948.00

7.4 Correlation Matrix



7.5 Multiple Regression Data

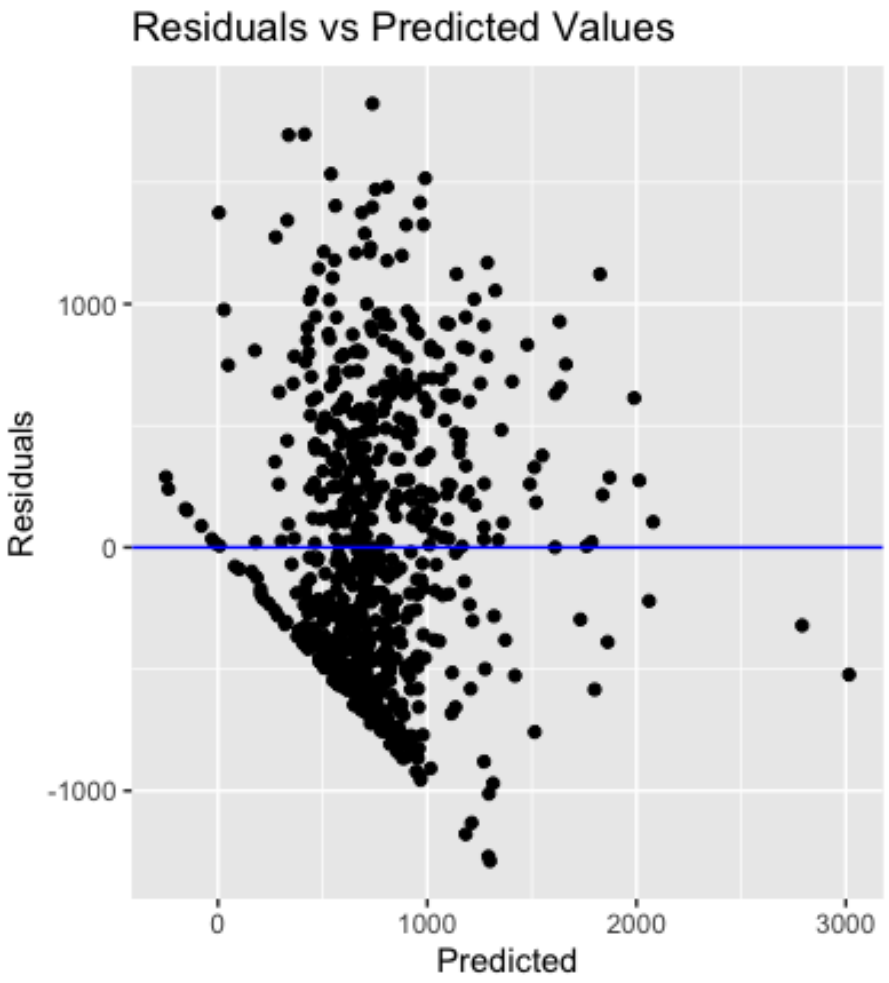
Variable	**Coefficient**	**T–Value**	**P–Value**
(Intercept)	297.074	3.466	0.000562
TSP	1,104.506	6.248	$7.320 * 10^{-10}$
VORP	321.769	10.006	$2.000 * 10^{-16}$
DBPM	40.756	4.887	$1.280 * 10^{-6}$

$R^2 = 0.2547$

$p-value = 2.200 * 10^{-16}$

$Residual\ Standard\ Error = 565.5\ on\ 681\ degrees\ of\ freedom$

7.6 Residual Plot



$Residual\ Standard\ Error = 565.5\ on\ 681\ degrees\ of\ freedom$

7.7 Partial Regression Plots