NBA Statistics Project 2

Giovanni Evans and Sebastian Furney

30 April 2023

Introduction

The NBA (National Basketball Association) is the highest level professional league for basketball in the world. The NBA was founded in 1946 and originally had 11 teams. Now, the NBA has 30 teams. The NBA has a regular season that is 82 games (in years not affected by the pandemic), and the playoffs where 16 teams make it, and one single team ends up winning the tournament. Our data set includes the 2019-2020, 2020-2021, and 2021-2022 NBA seasons with all 30 teams.

The following 27 parameters are in our NBA team statistics data set: Team Name, Games Played, Number of Wins, Number of Losses, Win Percentage, Average Minutes Played, Points Per Game, Field Goals Made Per Game, Field Goals Attempted Per Game, Field Goal Percentage, Three Pointers Made Per Game, Three Pointers Attempted Per game, Three Point Percentage, Free Throws Made Per Game, Free Throws Attempted Per game, Free Throws Percentage, Offensive Rebounds Per Game, Defensive Rebounds Per Game, Total Rebounds Per Game, Assists Per Game, Turnovers Per Game, Steals Per Game, Blocks Per Game, Blocks Against Per Game, Personal Fouls Taken Per Game, Personal Fouls Drawn Per Game, and Team Plus Minus.

We decided that the NBA playoffs involves too many random variables, deviations from expectations, and results in a binary outcome (won championship/did not win championship). For this reason, we decided to make a predictive model for the regular season. The main measure of a team's success in the regular season is win percentage, so this will be our response variable. We will throw out the variables: Team Name, Number of Wins, Number of Losses, and Average Minutes Played from our data set as they involve totals by season instead of averages. This is a problem because the 2019-2020, 2020-2021, and 2021-2022 seasons all had differing numbers of games played due to COVID. If we were to keep these variables in, than teams in 2021-2022 would have many more predicted wins since they played more games than teams in 2019-2020 and 2020-2021.

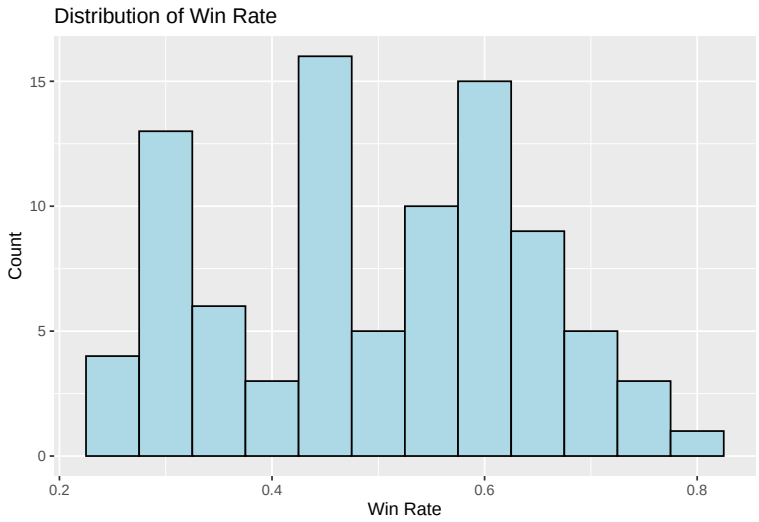
From this analysis of NBA team statistics, we hope to find out which team statistics are the most important in predicting team success. We also hope to determine which team statistics have little to no association with team success. For these reason, we will be testing every combination of our predictor variables to see which variables have the strongest association with team win percentage. Once we determine which variables individually have the strongest association with win percentage, we will test for covariance between the variables and then create a multiple regression model that best predicts our outcome variable.

Exploratory Analysis

Response Variable

As explained in the introduction, we will be using Team Win Percentage as the response variable for this model. Team Win Percentage is simply the number of games a team wins in a season divided by the total number of games that they play. To reiterate once more, due to complications regarding the pandemic, not

every team got to play the standard 82 games of a season in the 2019-2020 season and the 2020-2021. This is why it is important for us to use win rate instead of total wins as our response variable. Below is the distribution of win percentage among teams from the past three seasons.



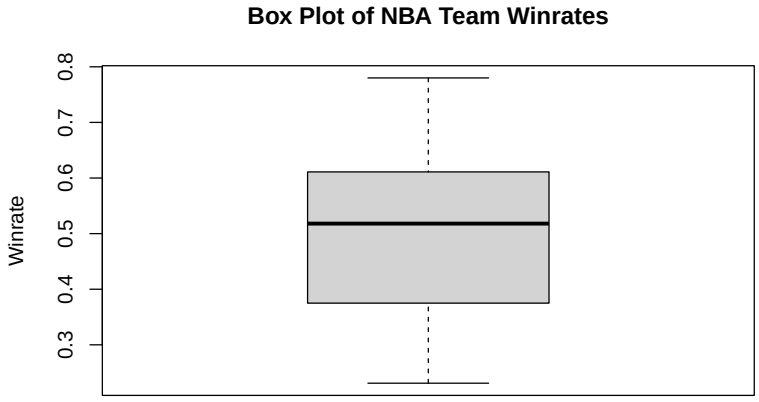
Looking at the histogram we can observe that there is no normal distribution in the win rates of nba teams. Additionally, the graph appears to be left skewed, this fact is reinforced by the median (0.518) being larger than the mean (0.4983). Mean:

```
## [1] 0.4983
```

Median:

```
## [1] 0.518
```

Below is a box plot visual of win percentage in order to provide a visual for the mean, median, quantiles, and potential outliers:



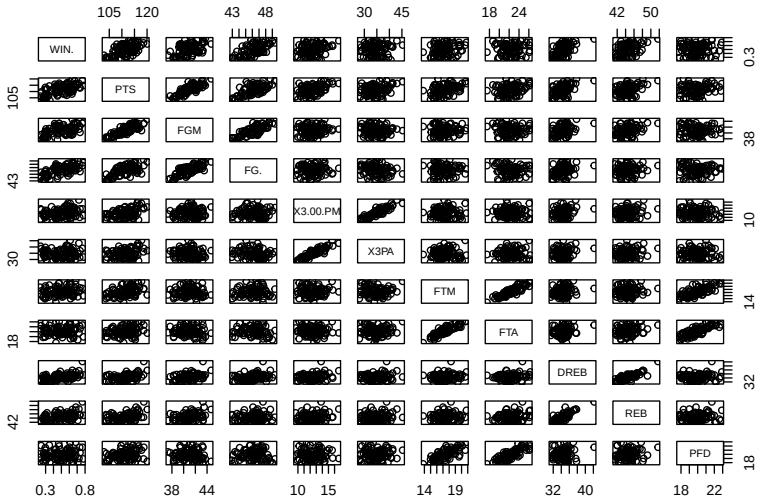
We can conclude that there are no outliers in our response variable because nothing falls outside of our lower and upper bounds. This can be visualized on the above box plot, where no data point falls outside of the 'whiskers' of the box plot. The actual quantiels are as follows:

Quantiles:

```
##      0%     25%     50%     75%    100%
## 0.23100 0.38175 0.51800 0.61100 0.78000
```

Predictor Variable

As detailed in the introduction, there are many possible predictor variables that we can include in our final regression model. The main aspect of the possible predictor variables that we want to check before we make our model is for collinearity. We looked at the collineairty between all predictor variables using the functions cor and pairs for both visual and numerical indicators of collineairty. Both these functions resulted in far too much output, so we will only explain and give visual output for the variables that have a strong association with each other (correlation coefficient of greater than 0.8 or less than -0.8).



As you can see from the correlation coefficient table and the pairs grpahings, the following pairs of variables were strongly correlated with each other: Points Per Game(PTS) and Field Goals Made Per Game(FGM): 0.849, Three Pointers Made Per Game(3PM) and Three Pointers Attempted Per Game(3PA): 0.92, Free throws Made Per Game(FTM) and Free throws Attempted Per Game(FTA): 0.89, Free throws Attempted Per Game(FTA) and Personal Fouls Drawn per Game(PFD): 0.81, Defensive Rebounds Per Game(DREB) and Rebounds Per Game(REB): 0.83,

All of the variables that are collinear make intuitive sense. PTS and TPA are collinear because when a field goal is made, it results in points. 3PM and 3PA are collinear because more threes that are attempted per game results in more threes made per game. FTM and FTA are collinear because the more free throws that are taken, the more free throws that are made. FTA and PFD are collinear because when a team is fouled on a shot or in the bonus, that team get to attempt more free throws. Finally, DREB and REB are collinear because most rebounds that a team gets in game will be on its own defensive end, and when a defensive rebound is gotten, it also counts as a rebound in general. We will be careful when making our models to not include any of these strongly correlated predictor variables if we are trying to make inferences about our data.

Model Development

Model Selection

To begin the creation of the model, a linear model with nearly every variable from the original data set was created. The excluded variables were team, wins, losses, and average minutes played. Wins and losses were of course excluded because we are building a model to predict win percentage meaning both wins and losses essentially show the same information as win percentage. This project's scope is to inference on what statistics are the greatest determiners of win percentage. Additionally, since we want to look league wide,

the team column was excluded. Finally, average minutes played is roughly the same across all observations as a game will always be 48 minutes unless it goes in to overtime, so it was excluded from the original model.

The original model used PTS, FGM, FGA, FG., X3.00.PM, X3PA, X3P., FTM, FTA, FT., OREB, DREB, REB, AST, TOV, STL, BLK, BLKA, PF, PFD and plus.minus to predict win percentage. As you may expect using every variable does not lead to the strongest model. Additionally when doing this, very few of the variables proved to be statistically significant. In order to improve the model, backwards step wise regression was used on the model. A backwards step wise process removes variables from a model one at a time until the strongest version of the model is created. The variables removed are chosen in order to minimize the Akaike information criterion (AIC).

$$AIC = 2k - 2\ln(L)$$

where k is the number of model parameters and L is the maximized value of the likelihood function for the estimated model. The final model resulting from the backwards step wise process is represented by the equation below.

$$\begin{aligned} Win\ Percentage = {} & 0.4914 + 0.0028 \times FG + 0.0009 \times FT + 0.0026 \times DREB \\ & - 0.0171 \times OREB + 0.0270 \times plus.minus - 0.0034 \times TOV \end{aligned}$$

In this equation win percentage is represented as a decimal. So an output of 0.5 would represent a 50% win rate. As a reminder FG. represents field goal percentage, FT. represents free throw percentage, DREB represents defensive rebounds, OREB represents offensive rebounds, plus.minus represents plus-minus, and TOV represents turnovers.

An increase of FG.by 1 will result on average in an increase in win percentage by 0.28%. An increase of FT. by 1 will result on average in a increase in win percentage by 0.04%. An increase of DREB by 1 will result on average in an increase in win percentage by 0.26%. An increase of OREB by 1 will result on average in a decrease in win percentage by 1.72%. An increase of plus.minus by 1 will result on average in an increase in win percentage by 2.7%. An increase of TOV by 1 will result on average in a decrease in win percentage by 0.34%.

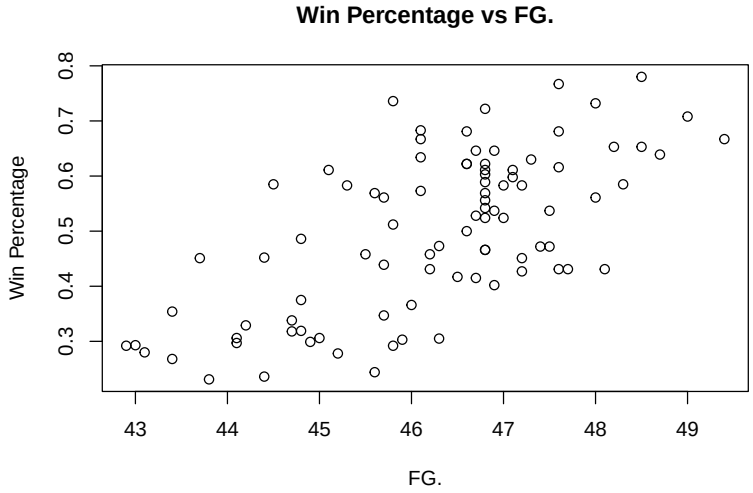
The most surprising insight is that an increase in offensive rebounds actually results in a decrease in win rate. Most people would expect that getting an offensive rebound results in an increase of win rate because having possession of the ball tends to be a good thing. One possible explanation for this negative relationship is that getting an offensive rebound can only occur when your team shoots and misses. So even though getting the ball is a positive outcome the offensive rebound can only occur in a negative situation.

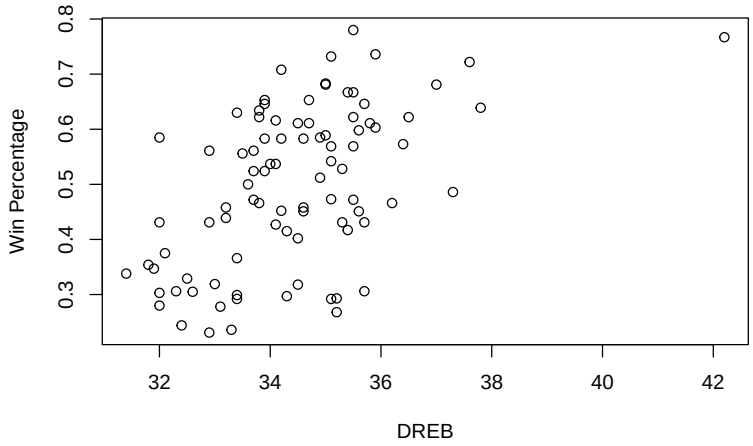
Assumptions

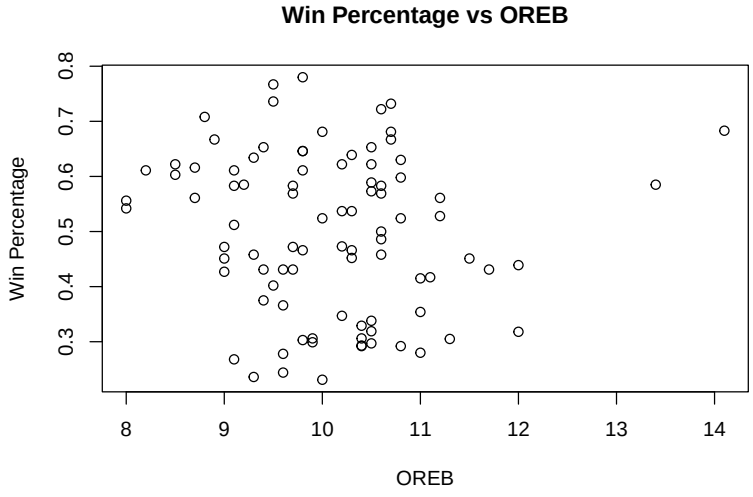
There are five assumptions that must be met for the model to be valid. One, there must be a linear relationship between each predictor variable and the response variable (win percentage). Two, each observation should be independent of one another. Three, there should be no multicollinearity among predictor variables, in other words predictors must not be highly correlated. Four, there needs to be homoscedasticity among each predictor, meaning each predictor variable should have relatively similar variance. Five, the residuals received from the model should be normally distributed.

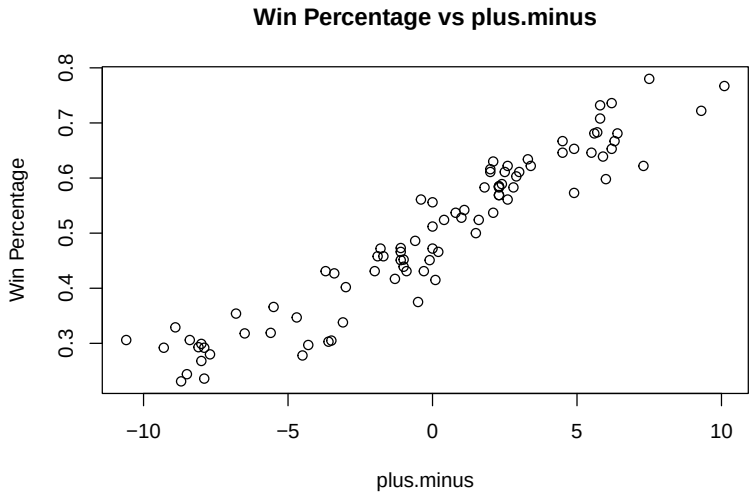
Linearity

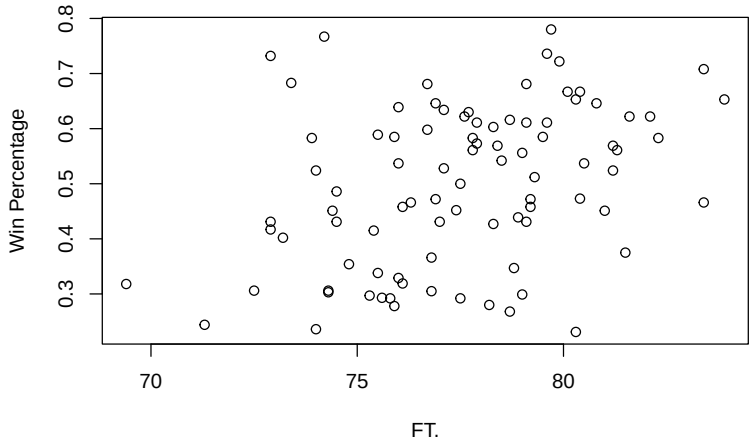
To check that each of our variables roughly has a linear relationship with Win Percentage we can view a set of scatter plots with predictors on the x-axis and Win Percentage on the y-axis.

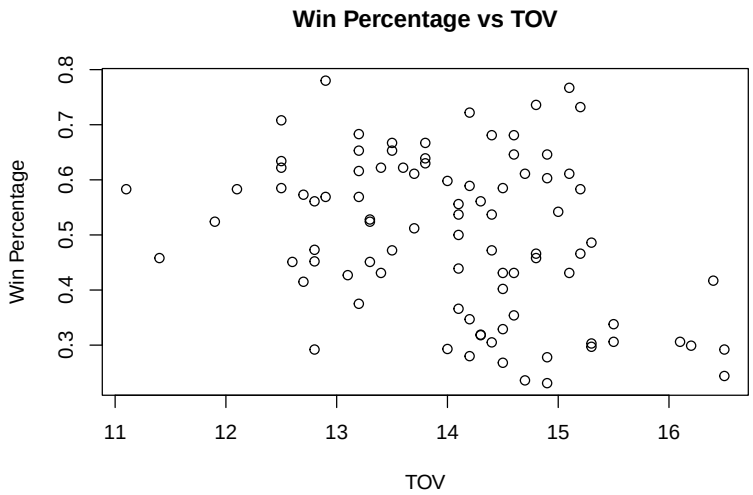

Win Percentage vs FG.
Win Percentage
0.3
0.4
0.5
0.6
0.7
0.8
43
44
45
46
47
48
49
FG.


Win Percentage vs DREB
Win Percentage
0.3
0.4
0.5
0.6
0.7
0.8
32
34
36
38
40
42
DREB


Win Percentage vs OREB
Win Percentage
0.3
0.4
0.5
0.6
0.7
0.8
8
9
10
11
12
13
14
OREB

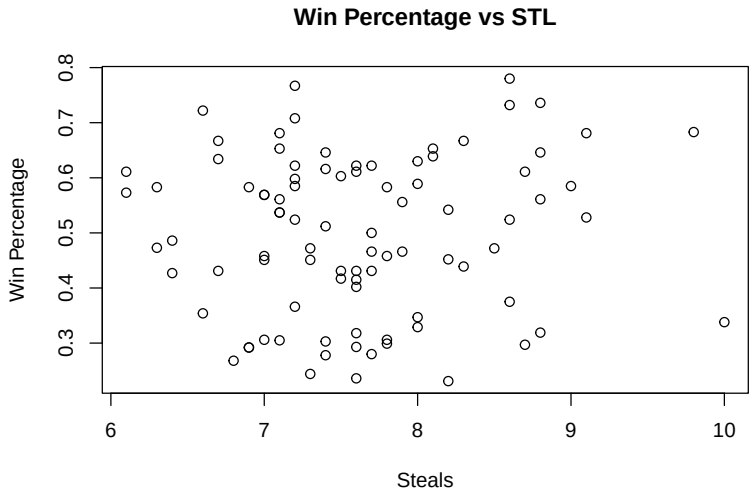






The level of present linearity varies among each predictor. Win rate and plus/minus have the strongest linear relationship. While no other variable has as strong of a relationship as plus/minus, there is still a present linear relationship between each of them and win rate. As an added point, the relationships between win percentage and offensive rebounds and win percentage and turn overs are both negative and linear.

Originally steals was also included in the model however after seeing the scatter plot relating steals and win percentage it was removed. That scatter plot can be seen below

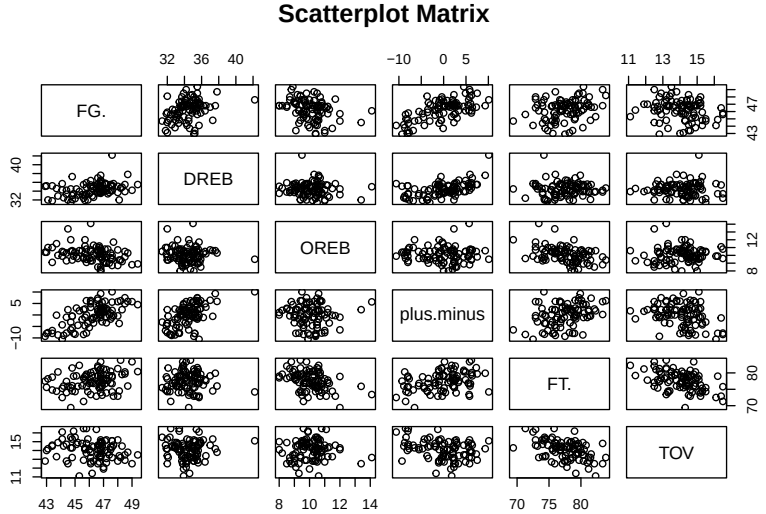


Independence

To insure independence among observations a few steps were taken in the collection of the data and creation of the model. To start data was collected from three different NBA seasons. This combats the fact that win rates in a given season or to some level dependent on other teams win rates that season. As each game requires one team to win and one team to lose an increase in one teams win rate will result in a decrease in another teams win rate. However by using data from three seasons we ensure that there are not necessarily shared games between two observations records.

In addition to using data from multiple seasons, at the beginning of the model making process, the data was randomly split into test and training data. This random selection also increases the probability that individual observations are independent of one another.

Multicollinearity

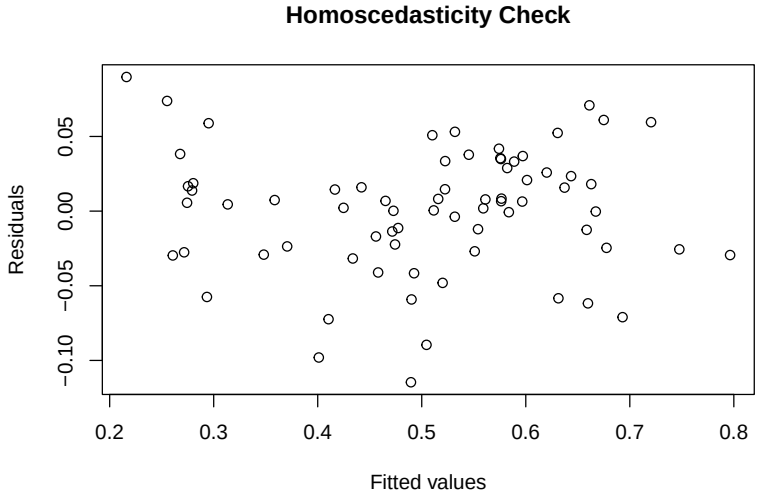


The above graphs and table display the strength of the relationships between each pair of predictor variables. While there is certainly some moderate correlation in some cases such as field goal% and plus/minus, there

is no case where this correlation coefficient exceeds 0.8 meaning there is no significant enough correlation to remove any of the predictors.

Homoscedasticity

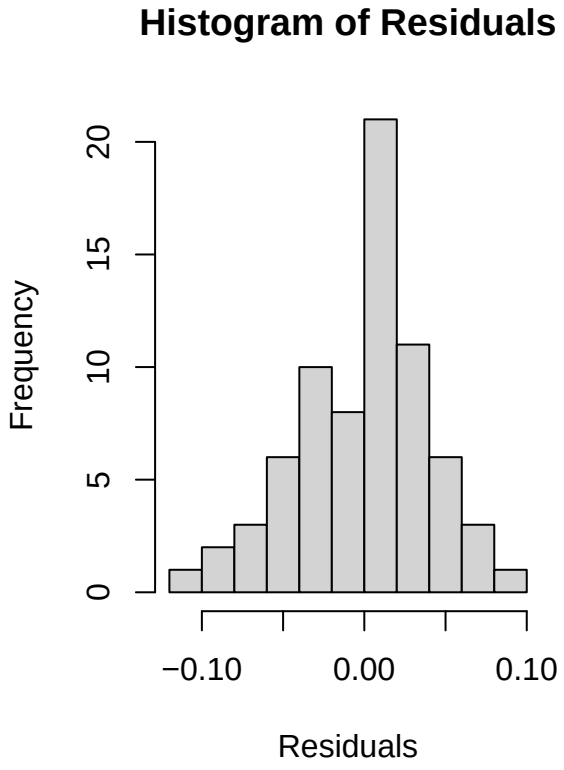
Next we will observe whether or not each predictor has relatively equal variance. We can do this by plotting residuals relative to fitted values.



In this scatter plot we are looking for any clear pattern, which would suggest that homoscedasticity is not met. There is no clear pattern in the plot meaning each predictor variable has similar variance.

Normality

Finally, we will create a histogram of residuals to check if they are normally distributed.



The histogram seems to have a slight left skew but overall is normally distributed meaning all are conditions are met.

Model Analysis

What Is the Best Predictor?

Our primary goal at the start of this project was to determine which statistic or statistics are the most effective at predicting a NBA team's win rate. Based on our model and the model making process, one statistic clearly proved to be the strongest when it comes to win rate predictions. This was a team's plus minus score. When looking at the plus minus for a team you are looking at the difference in score for that team in a game. So for this data set plus minus represents the average difference of score a team had throughout a given season.

While looking at p-values for each predictor in the model we noticed that defensive rebounds and turnovers had very high p-values 0.62 and 0.53 respectively. These high p-values indicate that both variables are not particularly significant despite the fact that both of their inclusions led to the strongest model. This indicates that neither variable is a particularity good indicator of win rate. In contrast however plus minus had an extremly low p-value indicating its relavence when predicting win rate. In addition a unit change in plus minus has the greatest effect on win rate of all predictor variables.

Applying The Model

To observe the model on new data we will be applying the equation to statistics from NBA teams in the recently concluded 2022-2023 NBA regular season.

As a reminder this equation is:

$$WinPercentage = 0.4914 + 0.0028 \times FG + 0.0009 \times FT + 0.0026 \times DREB - 0.0171 \times OREB + 0.0270 \times plus.minus - 0.0034 \times TOV$$

For this test we will be using the statistics of six teams the Memphis Grizzlies, Boston Celtics, Toronto Raptors, Oklahoma City Thunder, Detroit Pistons, and Houston Rockets. These six teams were selcted to represent teams from the top, middle, and bottom of each conference.

Memphis Grizzlies

The Grizzlies had: FG% = 47.5, FT% = 73.3, DREB = 34.6, OREB = 12, plus/minus = 3.9, and TOV = 13.6. Plugging these into the equation we get a predicted win rate of 62.88%. The teams actual win rate was 62.2% in this case the model was extremely close in its prediction.

Boston Celtics

The Celtics had: FG% = 47.5, FT% = 81.2, DREB = 35.6, OREB = 9.7, plus/minus = 6.5, and TOV = 13.4. Plugging these into the equation we get a predicted win rate of 69.49%. The team's actual win rate was 69.5%.

Toronto Raptors

The Raptors had: FG% = 45.9%, FT% = 78.4%, DREB = 30.3, OREB = 12.7, plus/minus = 1.5, and TOV = 11.7. Plugging these into the equation we get a predicted win rate of 49.91%. The team's actual win rate was 50%, once again these predictions are very close indicating the strength of the model.

Oklahoma City Thunder

The Thunder had: FG% = 46.5, FT% = 80.9, DREB = 32.3, OREB = 11.4, plus/minus = 1.1, and TOV = 13.0. Plugging these into the equation we get a predicted win rate of 46.34%. The team's actual win rate was 48.8%.

Detroit Pistons

The Pistons had: FG% = 45.4, FT% = 77.1, DREB = 31.3, OREB = 11.2, plus/minus = -8.2, and TOV = 15.1. Plugging these into the equation we get a predicted win rate of 23.67%. The team's actual win rate was 20.7%.

Houston Rockets

The Rockets had: FG% = 45.7, FT% = 75.4, DREB = 32.9, OREB = 13.4, plus/minus = -7.9, and TOV = 16.2. Plugging these into the equation we get a predicted win rate of 28.04%. The team's actual win rate was 26.8%.

Overall the model preformed very well, however it appears that the lower a teams win rate is the worse the model becomes. Of the six tests two were within 0.1% accuracy and one more fell within 0.5%. The three others however were all off by over 1% and in one case by nearly 3%.

Conclusion

As a result of our multiple regression model, we were able to very accurately predict a NBA team's win percentage based on their Field Goals Made Per Game, Free Throw Percentage, Defensive Rebounds Per Game, Offensive Rebounds Per Game, Plus-Minus, and Turnovers Per Game. As shown above, this model was extremely accurate in predictions for teams that had actual moderate to high win percentages, and somewhat accurate for teams with actual lower win percentages.

Also as a result of this project, we came to the conclusion that Team Plus-Minus was the best solo predictor of regular season success. This makes intuitive sense as a team that has the highest Plus-minus scores more points and allows less points on average than other teams which leads to more wins.

One oddity that we discovered as a result of the model analysis was that Offensive Rebounds Per Game actually had a negative correlation with win percentage. That is, when a team gets one more offensive rebound per game, the win percentage drops, on average. We deduced that this is because offensive rebounds are only attainable when a team misses a shot. Thus, an offensive rebound can only be had when a team does not finish its chances on the offensive end and fails to score points.

Finally, we wanted to explore advanced team statistics such as Effective Field Goal Percentage or Net Offensive Efficiency, but it was difficult to find large enough data sets containing these types of statistics. However, if we wanted to go further with our model and make a more robust predictive model of win percentage, we could add advanced statistics to our model and go through the same process of backwards step wise modeling with all variables.