Aidan David
BU MET CS555
02/24/2022
Term Project

INTRODUCTION

Professional sports teams dedicate hundreds of thousands of dollars to scouts which spend their time travelling, spectating, and assessing the skill level of college and non-professional players to select for sports drafts[1,2]. Part of the reason for this is that it is difficult and time-consuming to determine how good a player is in team-sports with numerous player statistics. More than 25 unique statistics are recorded for each basketball player throughout the season which makes this process even more challenging. I hypothesize that there exists a regression model which accounts for all of these statistics together which may be used to quantify the impact a player has on their team's success. If this hypothesis is true, which statistic(s) have the largest impact on a player's impact?

The data I will use to conduct this analysis is from the NBA's 2021-2022 season player statistics (https://www.nba.com/stats/players/traditional/?sort=PLUS_MINUS&dir=-1&Season=2021-22&SeasonType=Regular%20Season&PerMode=Totals). There are 592 rows (each represents a current NBA player) and 29 columns (not including id). To quantify the overall impact a player has had for their team I will use the "+/-" (also known as the "diff") statistic. This represents the number of points a team has scored overall when a particular player of said team is on the court subtracted by the number of points this team scored when they were not playing.

Before I ran any tests or calculated models, I observed the distribution of each of the 26 numeric players statistics to identify potential outliers. Most outliers were in percentage-based statistics such as free throw percentage. The reason for this is 5% of the data is players with less than 10 minutes of play time. These players do not take many shots thus increasing the chances of a 100% or 0% in shot percentage. Seeing as these outliers are only in about 5% of the data, I chose to leave them in as to not bias the sample.

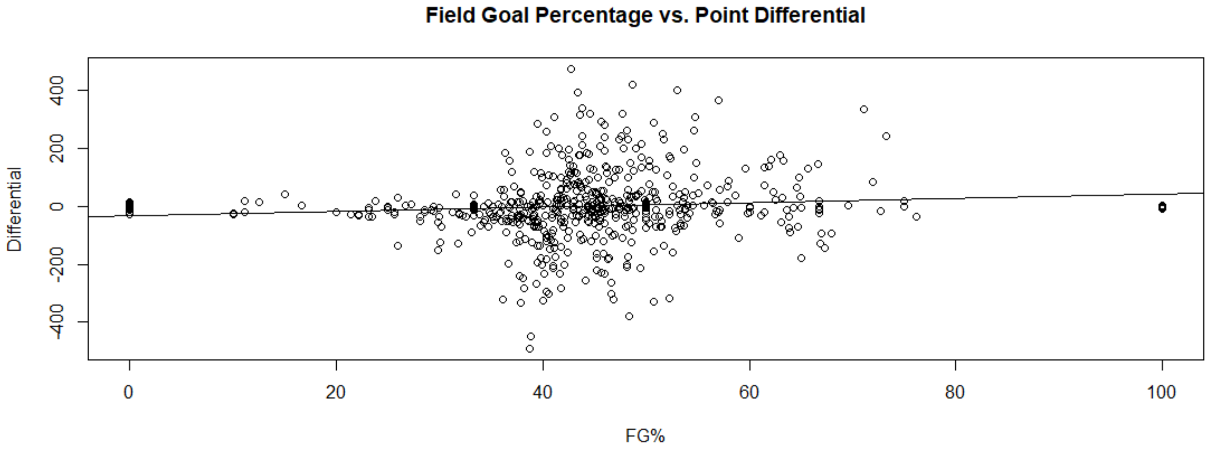


Figure 1: outliers lie at 0%, 100%, and 50% (r = 0.098)

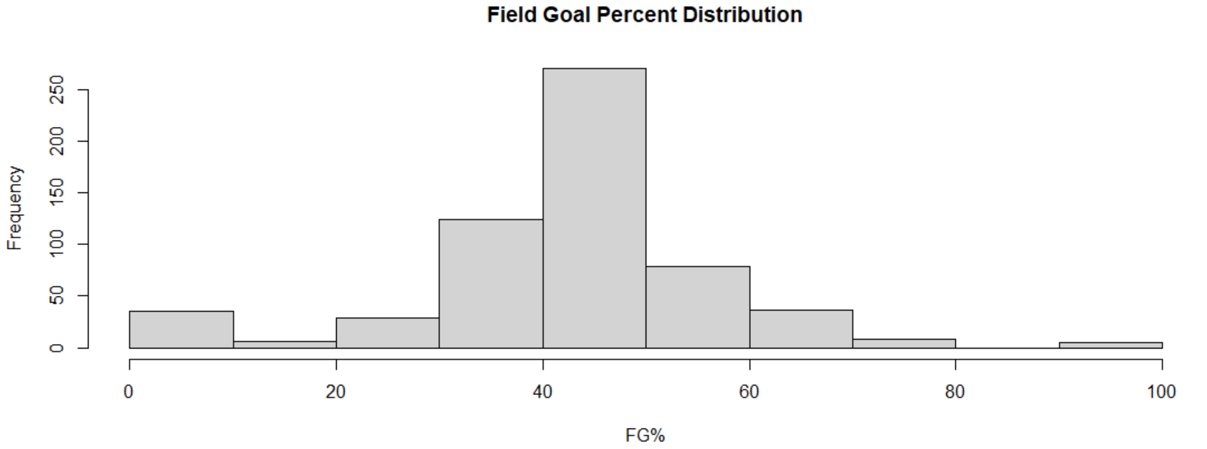


Figure 2: Normally distributed FG%

Lastly, I selected the explanatory variables to use for the model. I decided to leave out the following variables: wins, losses, field goals made, field goals attempted, three pointers made, three pointers attempted, free throws made, free throws attempted, rebounds, and fantasy points. Wins and losses are already represented by the games played statistic, the attempted and made shots are summarized by the shot percentages and the rebounds statistic is just offensive and defensive rebounds summed (the rebound variables are independent from one another, so I include them). I left out fantasy points as these are not true player statistics and reflect how spectators predict a player will perform. This leaves the following variables which I included in the model: age, games played, minutes played, points scored, overall field goal percentage, overall free throw percentage, overall three-pointer percentage, offensive rebounds, defensive rebounds, assists, turnovers, steals, blocks, personal fouls, double doubles, and triple doubles.

I use multiple variables to predict the response value of the point differential variable, therefore I will calculate a multiple linear regression model. The model's explanatory variables multiplied by a corresponding coefficient estimate (calculated using R) represents the predicted point difference of a player. To test my hypothesis, whether at least one of the model's variables has an impact on a player's total point differential, I will use an F-test for multiple linear regression at the alpha level $\alpha = 0.05$. If the p-value for this test is less than this alpha, I will conduct individual t-tests to determine the association each explanatory variable has with the response variable and then provide confidence intervals for the coefficient estimates.

RESULTS

Hypothesis:

H_0 : $\beta_{age} = \beta_{games\,played} = \beta_{minutes}$ ... $= 0$ (none of the explanatory variables are significant predictors of point differential)

H_1 : $\beta_{age} \neq 0$ and/or $\beta_{games\,played} \neq 0$ and/or $\beta_{minutes} \neq 0$ ... (at least one of the slope coefficients is different than 0; at least one of the explanatory variables is a significant predictor of point differential)

Test Statistic:

Since there are multiple explanatory variables contributing to the response variable, an F-test is the appropriate statistic to use for this procedure.

Decision Rule:

I will use the p value to determine if there is a linear association between variables.
The alpha value provided is 0.05 therefore if the p value is less than or equal to 0.05, we reject the null hypothesis.

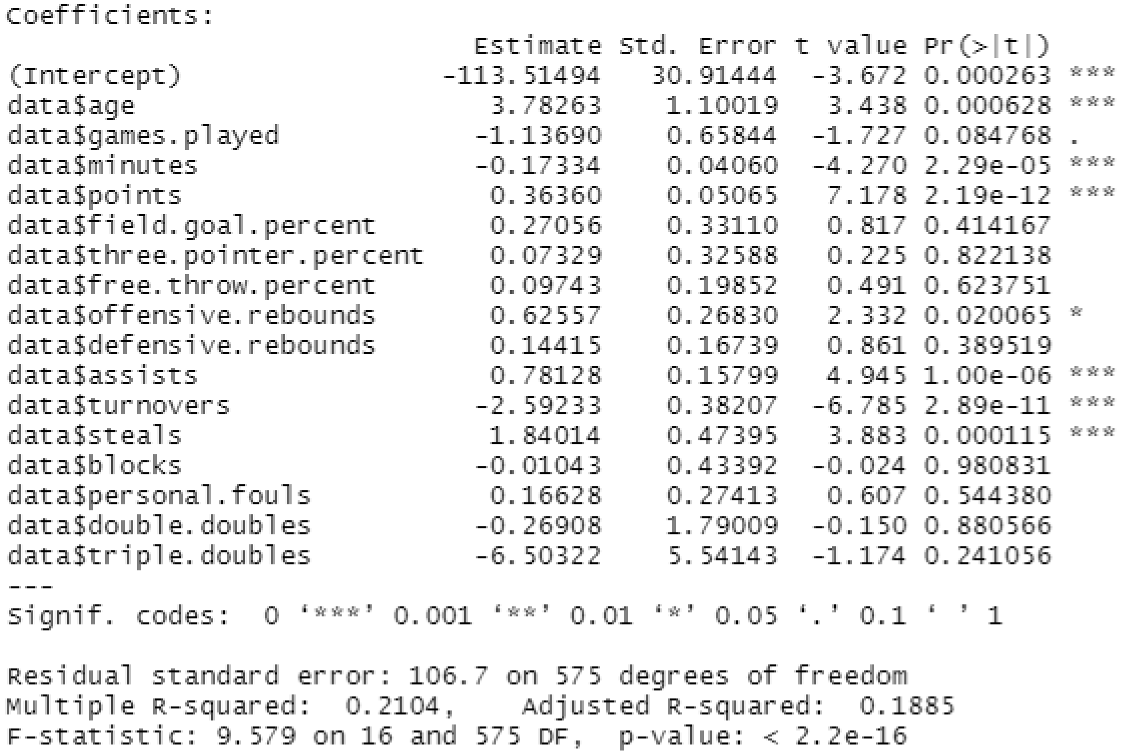
Compute Test Statistic:

F-statistic: 9.579 on 16 and 575 DF, p-value: < 2.2e-16,
Residual standard error: 106.7 on 575 degrees of freedom, Multiple R-squared: 0.2104,
Adjusted R-squared: 0.1885

Conclusion:

Reject the null hypothesis since 2.2e-16 <= 0.05. We have significant evidence at the α=0.05 level that age, games played, minutes, points, field goal percentage, three pointer percentage, free throw percentage, offensive rebounds, defensive rebounds, assists, turnovers, steals, blocks, personal fouls, double doubles, and triple doubles are significant predictors of point differential. That is, there is evidence of a linear association between the previously stated explanatory variables and point differential.

Because there is an association with the variables together and the point differential, we must now conduct a t-test for each explanatory variable to determine if it is a predictors of point differential when controlling for the other explanatory variables.



```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 -113.51494   30.91444  -3.672 0.000263 ***
data$age                       3.78263    1.10019   3.438 0.000628 ***
data$games.played             -1.13690    0.65844  -1.727 0.084768 .  
data$minutes                  -0.17334    0.04060  -4.270 2.29e-05 ***
data$points                    0.36360    0.05065   7.178 2.19e-12 ***
data$field.goal.percent        0.27056    0.33110   0.817 0.414167    
data$three.pointer.percent     0.07329    0.32588   0.225 0.822138    
data$free.throw.percent        0.09743    0.19852   0.491 0.623751    
data$offensive.rebounds        0.62557    0.26830   2.332 0.020065 *  
data$defensive.rebounds        0.14415    0.16739   0.861 0.389519    
data$assists                   0.78128    0.15799   4.945 1.00e-06 ***
data$turnovers                -2.59233    0.38207  -6.785 2.89e-11 ***
data$steals                    1.84014    0.47395   3.883 0.000115 ***
data$blocks                   -0.01043    0.43392  -0.024 0.980831    
data$personal.fouls            0.16628    0.27413   0.607 0.544380    
data$double.doubles           -0.26908    1.79009  -0.150 0.880566    
data$triple.doubles           -6.50322    5.54143  -1.174 0.241056    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 106.7 on 575 degrees of freedom
Multiple R-squared:  0.2104,    Adjusted R-squared:  0.1885 
F-statistic: 9.579 on 16 and 575 DF,  p-value: < 2.2e-16
```

As shown in the above image, the variables with p values less than 0.05 are significant at the 0.05 alpha. Age, minutes, points, offensive rebounds, assists, turnovers, and steals are all significant predictors of point differential. According to the estimates, for every increase in 3.7 years of age, 0.36 points, 0.62 offensive rebounds, 0.78 assists or 1.84 steals the model predicts that a player improves their point differential by 1. With every 0.17 minutes a player has played or 2.59 turnovers a player has, the model predicts that a player's point differential decreases by 1.

We are 95% confident that the true value of the slopes of the following variables fall within their corresponding interval:

age: (1.62, 5.94)
minutes: (-0.25, -0.09)
points: (0.26, 0.46)
offensive rebounds (0.09, 1.15)
assists: (0.47, 1.09)
turnovers: (-3.34, -1.84)
steals: (0.90, 2.77)

CONCLUSION AND LIMITATIONS

According to the results, there is an association with general player statistics and the impact they have on their team's success. Additionally, age, minutes played, points, offensive rebounds, assists, turnovers, and steals are the statistics which are most important for a professional basketball player. Furthermore, age, points scored, offensive rebounds, assists, and steals all positively contribute to a player's impact whereas the number of minutes played, and turnovers negatively contribute to their team impact.

Despite this result, there might be unaccounted factors which contribute to how relevant this analysis is. This dataset is a collection of players of the 2021-2022 NBA season exclusively, thus, this may not be representative for any NBA season. If I wanted to be able to more confidently generalize this to any NBA player, I would need to collect player statistics for all players across all NBA seasons and then take a random sample from the dataset. Besides this, the assumptions of multiple linear regression are all met. As shown below, the variation of the response variable around the regression line is constant (Figure 3) and the residuals are normally distributed (Figure 4).

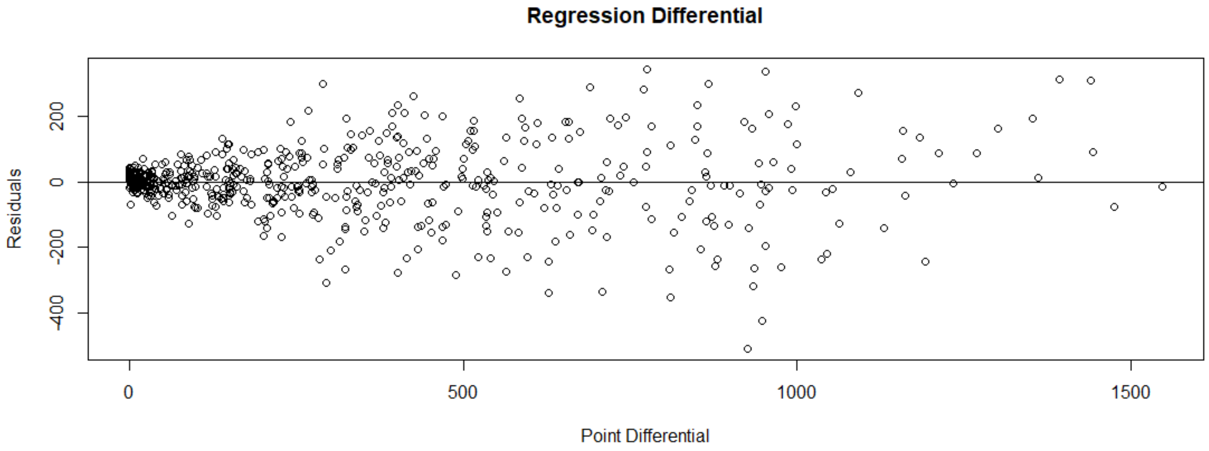


Figure 3: Regression residuals

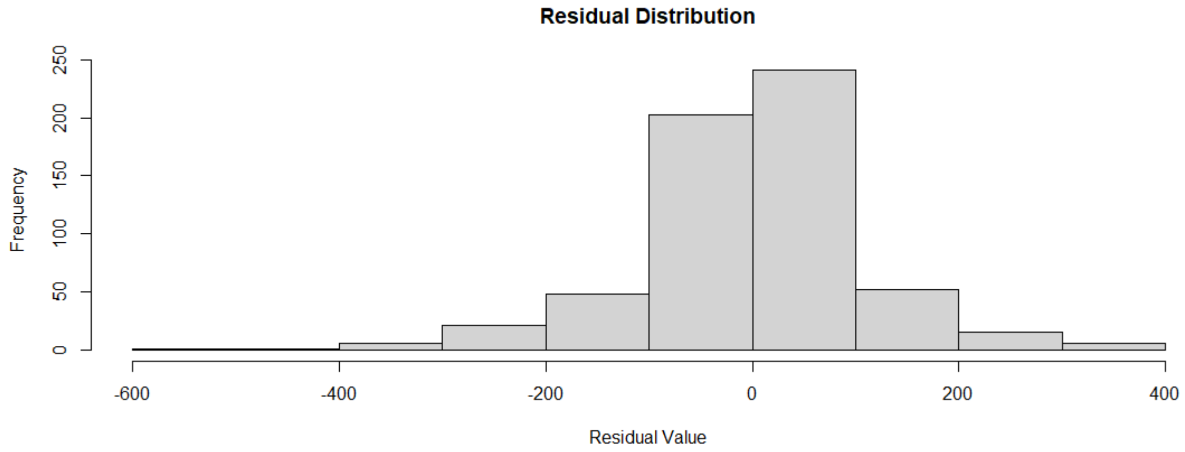


Figure 4: Residual Distribution

Sources

1

https://www.comparably.com/salaries/salaries-for-national-basketball-association-scout-nba-scout

2

https://www.boydsbets.com/highest-paid-coaches-nba/#:~:text=The%20median%20stands%20at%20around,That's%20about%20%2425%2C000%20per%20game.