Investigating the Determinants of Total Points Scored in NBA Games

Brandon Montez

March 2023

1 Introduction

Basketball is a popular sport enjoyed by millions of fans around the world. The end points in the game can be attributed to various factors such as players' individual skills, team strategies, and the quality of the opposition. In order to enhance the performance of a team, it is essential to understand the factors that contribute to scoring points in a game. This analysis aims to investigate the influence of a team's game statistics on the total points scored, thereby providing insights into the important aspects of basketball that need to be focused on during training and practice sessions.

The research question of interest for this study is: What will be the total points in a game based upon a team's statistics for the game besides the points? By analyzing the relationship between these variables, I aim to build a strong linear model that can suggest the significance of various factors in contributing to the total points per game. This will help to understand which variables, such as rebounds or assists, can be ruled out as less important in determining the response variable. It is important to note that this study does not cover the defensive aspect of the game and does not consider the opposing team's statistics in determining the scores.

The data set used for this analysis has been sourced from Kaggle.com (https://www.kaggle.com/datasets/nathanlauga/nba-games?resource=download), where a user has compiled data from the NBA stats website (https://www.nba.com/stats). The choice of a linear model for this study is based on the assumption that there exists a linear relationship between the predictor variables (team statistics) and the response variable (total points scored in a game). Preliminary visual analysis of scatterplots for correlation between these variables and checking the summary of the linear model helped justify our choice of using a linear model.

This paper is organized as follows: The first section, Data Description, details the data set and its variables, followed by an explanation of the methods used to analyze the data, including the justification for choosing a linear model. The subsequent section presents the Results, discussing the results of our linear model. Then we cover the model fitting portion, to adjust our model and test

it against other models to find the most parsimonious model. Finally, the paper concludes with a summary of the study, its limitations, and suggestions for future research.

In the next section, we will delve into the Data Description, where we will describe the variables in the data set and provide the code for visualizing scatterplots and checking the summary of the linear model. This analysis will serve as a foundation for justifying our choice of a linear model and guide us in building an appropriate model to address our research question.

2 Data Description

To begin the Data Description, I will provide a summary of the main variables included in the data set. We will be using data from the 2020 NBA season:

Field Goals Made (FGM): The number of field goals made by the team during a game.
Field Goal Percentage (FG%): The percentage of field goals made by the team out of the total attempted.
Three-Point Field Goals Made (3PM): The number of three-point field goals made by the team during a game.
Free Throws Made (FTM): The number of free throws made by the team during a game.
Rebounds (REB): The total number of rebounds collected by the team during a game.
Assists (AST): The total number of assists made by the team during a game.
Turnovers (TOV): The total number of turnovers committed by the team during a game.

We can find the summary stats, correlation matrix and the density plots for each variable.

The summary stats:

```
GAME_DATE_EST          GAME_ID             SEASON     HOME_TEAM_NICKNAME    PTS_home      FG_PCT_home      FT_PCT_home      FG3_PCT_home       AST_home       REB_home
 Min.   :2020-12-11   Min.   :12000001   Min.   :2020   Length:1249        Min.   : 73.0   Min.   :0.2770   Min.   :0.3330   Min.   :0.0890   Min.   :10.0   Min.   :24.00
 1st Qu.:2021-01-24   1st Qu.:22000235   1st Qu.:2020   Class :character   1st Qu.:104.0   1st Qu.:0.4290   1st Qu.:0.7140   1st Qu.:0.3100   1st Qu.:21.0   1st Qu.:40.00
 Median :2021-03-12   Median :22000547   Median :2020   Mode  :character   Median :112.0   Median :0.4660   Median :0.7860   Median :0.3660   Median :24.0   Median :45.00
 Mean   :2021-03-08   Mean   :23113374   Mean   :2020                      Mean   :112.4   Mean   :0.4669   Mean   :0.7781   Mean   :0.3664   Mean   :24.8   Mean   :44.85
 3rd Qu.:2021-04-19   3rd Qu.:22000859   3rd Qu.:2020                      3rd Qu.:121.0   3rd Qu.:0.5050   3rd Qu.:0.8460   3rd Qu.:0.4240   3rd Qu.:28.0   3rd Qu.:49.00
 Max.   :2021-07-20   Max.   :52000211   Max.   :2020                      Max.   :154.0   Max.   :0.6540   Max.   :1.0000   Max.   :0.7200   Max.   :50.0   Max.   :70.00
```

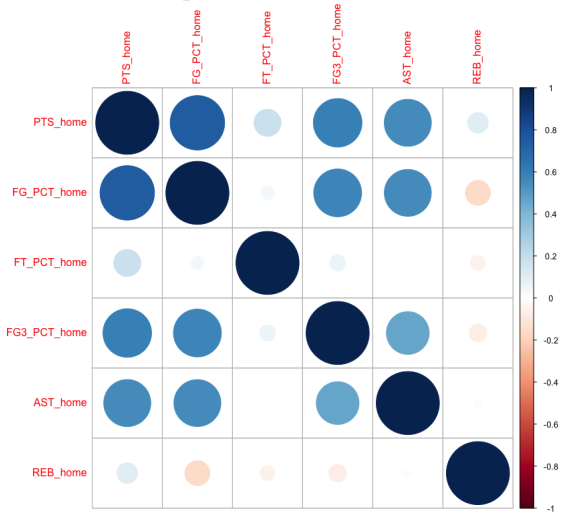
The standard deviations:

```
   PTS_home  FG_PCT_home  FT_PCT_home FG3_PCT_home     AST_home     REB_home     PTS_away  FG_PCT_away  FT_PCT_away FG3_PCT_away     AST_away     REB_away
12.53261419   0.05524905   0.10272652   0.08612009   4.98355920   6.44866844 12.56554534   0.05353707   0.10533304   0.08662366   4.96426342   6.37751265
```

The correlation matrix:

```
print(correlation_matrix)
              PTS_home FG_PCT_home  FT_PCT_home FG3_PCT_home     AST_home    REB_home
PTS_home     1.0000000  0.74570949  0.184120613   0.59383047  0.557506756  0.10744510
FG_PCT_home  0.7457095  1.00000000  0.042042910   0.57615516  0.554405266 -0.15689682
FT_PCT_home  0.1841206  0.04204291  1.000000000   0.06358466  0.009015923 -0.05450234
FG3_PCT_home 0.5938305  0.57615516  0.063584665   1.00000000  0.460659132 -0.07677188
AST_home     0.5575068  0.55440527  0.009015923   0.46065913  1.000000000 -0.01542704
REB_home     0.1074451 -0.15689682 -0.054502342  -0.07677188 -0.015427042  1.00000000
```

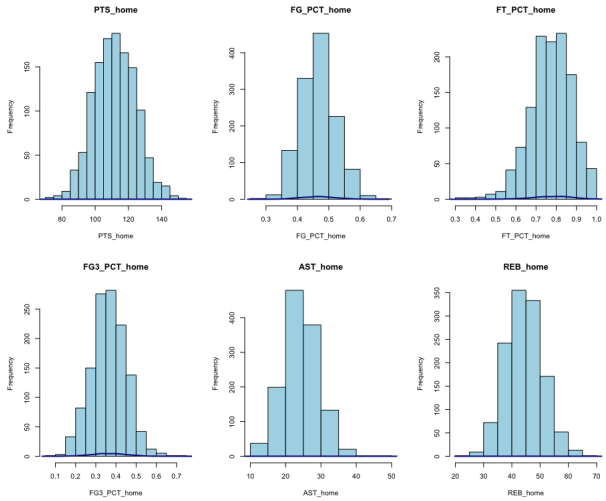
The correlation plot:



Looking at the correlation matrix and the correlation plot, it does not appear that there is an issue with multicollinearity as none of the variables have a very strong correlation with each other.

The most influential variables on the total points scored by the home team appear to be the field goal percentage, 3-point field goal percentage, and assists. Free throw percentage and rebounds seem to have a weaker influence on the points scored.

We can then create density plots for each variable:



These show a normal distribution among each of the variables.

3 Results

We create a linear model using all of the variables as shown below:

```
# Build linear regression model
model <- lm(PTS_home ~ FG_PCT_home + FT_PCT_home + FG3_PCT_home + AST_home + REB_home, data

# Summarize model
summary(model)
```

```
> summary(model)

Call:
lm(formula = PTS_home ~ FG_PCT_home + FT_PCT_home + FG3_PCT_home +
    AST_home + REB_home, data = selected_columns)

Residuals:
    Min      1Q  Median      3Q     Max
-21.792  -4.682  -0.402   4.690  34.301

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.16229    2.84180  -1.113    0.266
FG_PCT_home  130.88244    4.88244  26.807  < 2e-16 ***
FT_PCT_home   19.27506    1.94990   9.885  < 2e-16 ***
FG3_PCT_home  29.21757    2.90284  10.065  < 2e-16 ***
AST_home       0.37010    0.04939   7.493 1.27e-13 ***
REB_home       0.43585    0.03149  13.839  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.051 on 1243 degrees of freedom
Multiple R-squared:  0.6847,Adjusted R-squared:  0.6835
F-statistic: 539.9 on 5 and 1243 DF,  p-value: < 2.2e-16
```

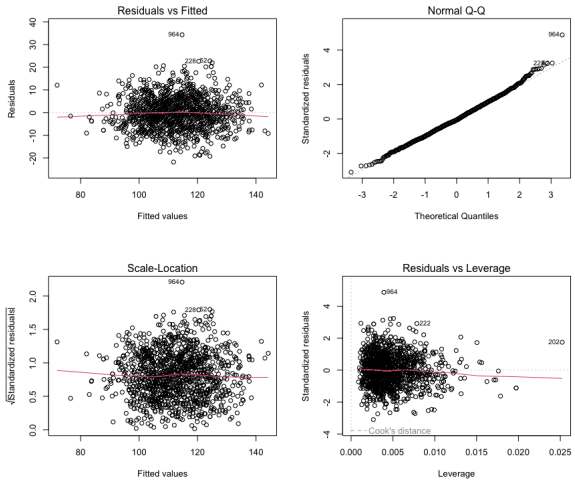
From these results, we can see that the full linear model is statistically significant, with a p-value very close to 0. This suggests that at least one of the independent variables has a significant relationship with home team points (PTS_home).

All five independent variables (FG_PCT_home, FT_PCT_home, FG3_PCT_home, AST_home, and REB_home) are statistically significant predictors of PTS_home, as their p-values are all less than 0.001. This indicates that each of these variables contributes to the explanation of the variation in home team points.

The model explains approximately 68.35% of the variation in home team points, as indicated by the adjusted R-squared value of 0.6835.

Finally, we can create summary plots of our linear model, where we see evidence of a linear trend. Mainly, the QQ-plot looks linear, the distribution of

residuals looks normal, and the Residuals vs Leverage plot also appears to be normal.



4 Model Fitting

I tried three different candidate models to find the best predictive model:

Full Model: Including all five independent variables

Backward elimination using AIC

Backward elimination using BIC

The full model and both backward elimination models resulted in the same model with the same adjusted R-squared value of 0.6835. The AIC and BIC values for the full model were 8431.561 and 8467.471, respectively. Neither AIC nor BIC removed any variables. Since all three models are identical, we can conclude that the full model is the best predictive model.

```
    summary(backward_aic)

Call:
lm(formula = PTS_home ~ FG_PCT_home + FT_PCT_home + FG3_PCT_home +
    AST_home + REB_home, data = selected_columns)

Residuals:
    Min      1Q  Median      3Q     Max
-21.792  -4.682  -0.402   4.690  34.301

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.16229    2.84180  -1.113    0.266
FG_PCT_home  130.88244    4.88244  26.807  < 2e-16 ***
```

```
FT_PCT_home   19.27506    1.94990   9.885  < 2e-16 ***
FG3_PCT_home  29.21757    2.90284  10.065  < 2e-16 ***
AST_home       0.37010    0.04939   7.493 1.27e-13 ***
REB_home       0.43585    0.03149  13.839  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.051 on 1243 degrees of freedom
Multiple R-squared:  0.6847,Adjusted R-squared:  0.6835
F-statistic: 539.9 on 5 and 1243 DF,  p-value: < 2.2e-16

> summary(backward_bic)

Call:
lm(formula = PTS_home ~ FG_PCT_home + FT_PCT_home + FG3_PCT_home +
    AST_home + REB_home, data = selected_columns)

Residuals:
    Min      1Q  Median      3Q     Max
-21.792  -4.682  -0.402   4.690  34.301

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.16229    2.84180  -1.113    0.266
FG_PCT_home  130.88244    4.88244  26.807  < 2e-16 ***
FT_PCT_home   19.27506    1.94990   9.885  < 2e-16 ***
FG3_PCT_home  29.21757    2.90284  10.065  < 2e-16 ***
AST_home       0.37010    0.04939   7.493 1.27e-13 ***
REB_home       0.43585    0.03149  13.839  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.051 on 1243 degrees of freedom
Multiple R-squared:  0.6847,Adjusted R-squared:  0.6835
F-statistic: 539.9 on 5 and 1243 DF,  p-value: < 2.2e-16
```

5 Discussion

Our final linear regression model was statistically significant and explained approximately 68.35% of the variation in home team points.

Multiple R-squared: 0.6847
Adjusted R-squared: 0.6835

All of the independent variables were found to be statistically significant with p-values less than 0.001. This indicates that the model is able to explain approximately 68.35% of the variation in home team points.

In conclusion, the full model with all five independent variables is the best predictive model for home team points in the 2020 NBA season when considering more parsimonious models utilizing these variables. The model has an adjusted R-squared of 0.6835, and the diagnostic plots support the validity of the model assumptions. Still, we may need to include more variables in order to provide stronger predictive power.

Our final model makes some sense in a real-world situation, as the chosen variables are known to have an impact on a team's scoring. However, the model does not utilize many other variables available in NBA statistics, such as injured starters, that could potentially improve its predictive power. Moreover, the model relies on in-game statistics like assists and rebounds, which are not known before or during the game, whereas shooting percentages can be estimated as a running average throughout the game. To make the model more applicable in real-world situations, we could develop formulas to predict the total rebounds and assists during a game based on their running total.

Our results show that the selected factors are individually important in determining the overall score of a team. However, we did not include several other potentially relevant factors in our model, which could have led to a higher Adjusted R-squared value if included. We also only focused on the 2020 NBA season, whereas our model would become more robust if it were built on multiple seasons. In addition, the model's reliance on in-game statistics may limit its applicability for real-time predictions.

To improve this analysis in the future, we could consider incorporating additional variables that are relevant to a team's performance, such as injured starters, opponent statistics, or historical performance data. We could also group the data by team, to get a better understanding of how well our model works for each team, rather than as a general predictor. We could also explore other types of models, such as time series models, to better capture the dynamics of the game and improve the predictive accuracy of our model.