Predicting Total Points Scored per Game by Each NBA Team Using Multiple Linear Regression

STATS 101A

Justin Phu

1 Introduction

Context

In today's modern NBA, sports analytics continues to make a revolutionary impact on the game, as it has become a key factor in optimizing team performance, evaluating player impact, and shaping strategic game planning. Through extensive analysis of previous performances, teams can uncover valuable trends and insights that were once difficult to quantify, such as understanding the value of individual player contributions and identifying the most effective offensive and defensive schemes. Although many factors contribute to winning a basketball game, one of the most important areas that teams aim to improve is their offensive prowess. Specifically, this means understanding the factors that will enable them to score more points. By focusing on these key factors, teams can evaluate the impact of each player, adjust rotations, and align strategies with their offensive goals. On the defensive end, predicting an opposing team's scoring potential helps inform defensive game planning, as knowing an opponent's likely offensive output allows for more targeted preparations that aims to limit their scoring. For this reason, a predictive machine learning model that uses in-game statistics as predictors to forecast total points scored per game can prove crucial to help teams make more informed decisions and gain a competitive edge.

Dataset Introduction

The dataset used in this analysis includes a quantitative stat sheet that details the statistics recorded by each team in all the games they have played since October 24, 2023, in chronological order. Each NBA game is represented by two rows, with each row corresponding to the total statistics of one of the two competing teams. The predictor and response variables of the dataset are outlined in the table below:

Table 1: Dataset Variables Table

Variable	Role	Description
Team	Not Used	The abbreviated name of 1 of 30 NBA teams playing in the game.
Match Up	Not Used	The opponent and home/away designation.
Game Date	Not Used	The date of when the game was played.
W/L	Not Used	Indicates whether the given team won or lost the game.
+/-	Not Used	Final score point differential.
FGM	Not Used	Field goals made (number of successful shots from the field).
FGA	Not Used	Field goals attempted.
3PM	Not Used	Three-point shots made.
3PA	Not Used	Three-point shots attempted.
FTM	Not Used	Free throws made.
FTA	Not Used	Free throws attempted.
REB	Not Used	Total combined rebounds.

FG%	Predictor	Field goal percentage.
3P%	Predictor	Three-point shooting percentage.
FT%	Predictor	Free throw percentage.
OREB	Predictor	Total number of offensive rebounds (rebounds made while on offense).
DREB	Predictor	Total number of defensive rebounds (rebounds made while on defense).
AST	Predictor	Total number of assists (passes directly leading to a made basket).
STL	Predictor	Total number of steals.
BLK	Predictor	Total number of blocks.
TOV	Predictor	Total number of turnovers (losing possession to the other team).
PF	Predictor	Total number of personal fouls committed by the team.
MIN	Predictor	Total amount of minutes played by the team in the given game.
PTS	Response	Total points scored.

Objective

By analyzing the predictor variables of the dataset, we are trying to find which predictor variables have the most profound impact on predicting the total amount of points scored in a given NBA basketball game. To determine this answer, we will create a multiple linear regression model and use stepwise regression to filter out insignificant predictor variables in order to find the model with the most accurate predictions.

2 Descriptive Statistics

Below is a table consisting of the descriptive statistics for each predictor variable in the dataset.

Table 2: Descriptive Statistics Table

Variable	N	Mean	SD	Median	Minimum	Maximum	Skew	SE
+/-	2460	0.00000	15.785622	0.00	-62.0	62.0	0.0000000	0.3182688
FGM	2460	42.17033	5.343417	42.00	26.0	65.0	0.1693956	0.1077337
FGA	2460	88.90325	7.013117	89.00	67.0	119.0	0.3180949	0.1413981
FG%	2460	47.52122	5.498003	47.50	27.7	67.1	0.0609072	0.1108504
3PM	2460	12.83699	3.837044	13.00	2.0	27.0	0.3449438	0.0773623
3PA	2460	35.10366	6.541778	35.00	12.0	63.0	0.2608191	0.1318950
3P%	2460	36.49443	8.341066	36.55	6.9	64.5	0.0746111	0.1681721
FTM	2460	17.03374	5.890426	17.00	0.0	44.0	0.4163408	0.1187624
FTA	2460	21.71951	6.995393	21.00	0.0	52.0	0.4072017	0.1410407
FT%	2460	78.33379	10.154975	78.90	33.3	100.0	-0.4313135	0.2047857
OREB	2460	10.55366	3.817189	10.00	0.0	28.0	0.5384155	0.0769620
DREB	2460	32.98780	5.409375	33.00	16.0	55.0	0.1752726	0.1090635
REB	2460	43.54146	6.575082	43.00	25.0	74.0	0.2197716	0.1325665
AST	2460	26.67073	5.101064	27.00	11.0	50.0	0.2668577	0.1028474
STL	2460	7.47439	2.821905	7.00	0.0	20.0	0.3621052	0.0568951
BLK	2460	5.14187	2.598016	5.00	0.0	17.0	0.5723708	0.0523811
TOV	2460	13.60488	3.811539	14.00	3.0	29.0	0.2405962	0.0768481
PF	2460	18.72927	4.148803	19.00	4.0	34.0	0.2779681	0.0836479
MIN	2460	241.36179	6.351120	240.00	240.0	290.0	5.0526871	0.1280509
PTS	2460	114.21138	12.845883	114.00	73.0	157.0	0.1119432	0.2589980

3 Initial Multiple Linear Regression Model

In our initial multiple linear regression model, the predictor variables we used were FG%, 3P%, FT%, OREB, DREB, AST, STL, BLK, TOV, PF, and MIN. The REB predictor was taken out since it is a redundant

statistic that can be found by combining OREB and DREB, and the +/- predictor was taken out since it does not have a direct influence on the total points scored. The estimated regression equation is as follows:

PTS = -90.720 + 1.533(`FG%`) + 0.300(`3P%`) + 0.187(`FT%`) + 0.907(OREB) + 0.237(DREB) + 0.324(AST) + 0.327(STL) + 0.059(BLK) - 0.827(TOV) + 0.422(PF) + 0.336(MIN)

Coefficient Interpretation

Based on our model, we can interpret each predictor coefficient, while holding the other predictors constant:

1. The 1.533(`FG%`) term shows a one unit increase in FG% leads to a 1.580 point increase in the total scoring output in each NBA game.
2. The 0.300(`3P%`) term shows a one unit increase in 3P% leads to a 0.300 point increase in the total scoring output in each NBA game.
3. The 0.187(`FT%`) term shows a one unit increase in FT% leads to a 0.187 point increase in the total scoring output in each NBA game.
4. The 0.907(OREB) term shows a one unit increase in OREB leads to a 0.907 point increase in the total scoring output in each NBA game.
5. The 0.237(DREB) term shows a one unit increase in DREB leads to a 0.237 point increase in the total scoring output in each NBA game.
6. The 0.324(AST) term shows a one unit increase in AST leads to a 0.324 point increase in the total scoring output in each NBA game.
7. The 0.327(STL) term shows a one unit increase in STL leads to a 0.327 point increase in the total scoring output in each NBA game.
8. The 0.059(BLK) term shows a one unit increase in BLK leads to a 0.059 point increase in the total scoring output in each NBA game.
9. The -0.827(TOV) term shows a one unit increase in TOV leads to a 0.827 point decrease in the total scoring output in each NBA game.
10. The 0.422(PF) term shows a one unit increase in PF leads to a 0.422 point increase in the total scoring output in each NBA game.
11. The 0.336(MIN) term shows a one unit increase in MIN leads to a 0.336 point increase in the total scoring output in each NBA game.

Table 3: Regression Analysis Results Table

Variable	Beta	SE	tvalue	Pr...t...	F.value	Pr...F..	VIF
Intercept	-90.720	4.562	-19.884	<2e-16	N/A	N/A	N/A
FG%	1.533	0.029	52.200	<2e-16	7258.794	<2.2e-16	2.072
3P%	0.300	0.017	18.110	<2e-16	375.292	<2.2e-16	1.514
FT%	0.187	0.011	16.859	<2e-16	224.894	<2.2e-16	1.005
OREB	0.907	0.032	28.596	<2e-16	1040.275	<2.2e-16	1.166
DREB	0.237	0.022	10.652	<2e-16	105.362	<2.2e-16	1.153
AST	0.324	0.028	11.565	<2e-16	301.787	<2.2e-16	1.626
STL	0.327	0.040	8.035	1.44e-15	59.112	2.139e-14	1.045
BLK	0.059	0.044	1.324	0.186	0.219	0.64	1.056
TOV	-0.827	0.030	-27.328	<2e-16	621.017	<2.2e-16	1.058
PF	0.422	0.028	15.271	<2e-16	335.831	<2.2e-16	1.040
MIN	0.336	0.018	18.259	<2e-16	333.385	<2.2e-16	1.085

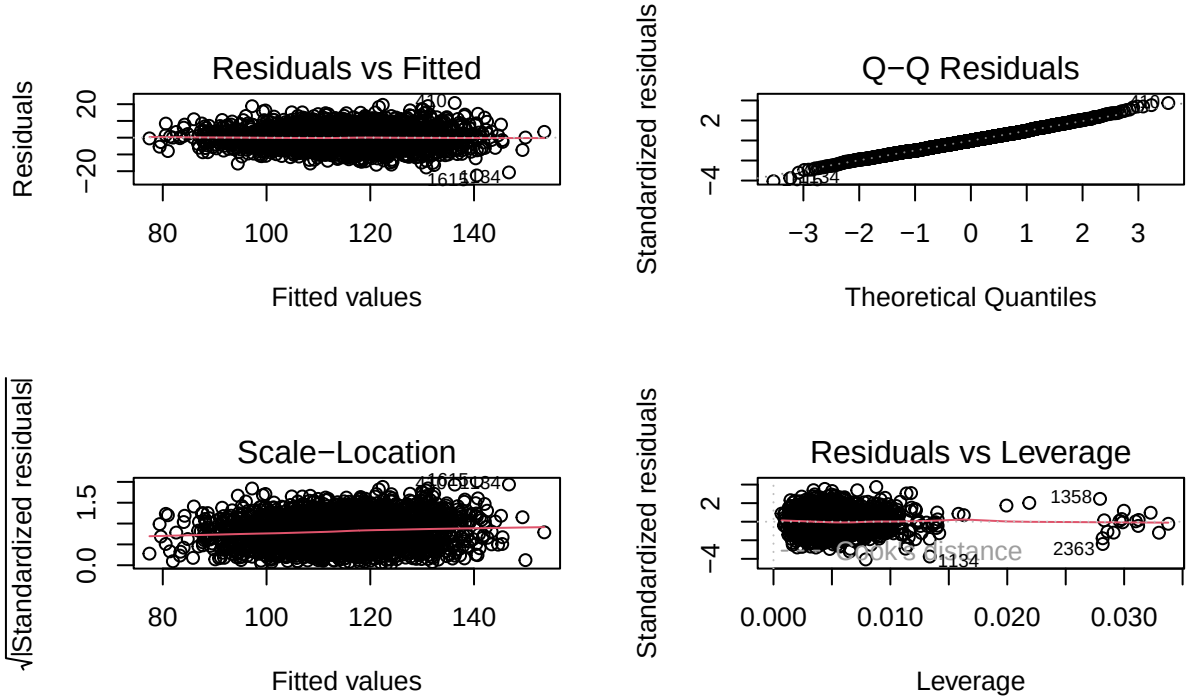
Result Explanation

Looking at the model results, we have an R^2 value of 0.8132 and an adjusted R^2 value of 0.8124 indicates that approximately 81% of the variance in the PTS response variable can be explained by the model. This shows that the predictors that we currently have in our model are performing well when it comes to predicting the total points scored in an NBA game. To further analyze the predictors in the initial model, the t-test and F-test results were computed, as shown in the table above. Based on the values obtained, the t-statistic values showed that all predictors were statistically significant except the BLK statistic. The statistically significant predictors all had corresponding p-values below the threshold of 0.05, but for the BLK statistic, its p-value was above the threshold at 0.186. As for the F-statistic, all predictor variables have relatively large F-statistic values and are statistically significant except for the BLK statistic, which has a p-value that is greater than the significance threshold of 0.05. This shows that the BLK statistic does not significantly contribute to explaining the total score while controlling for other predictors, and that if we were to remove the predictor variable, it would make little difference in the prediction. Lastly, in order to check for multicollinearity in the dataset, the VIF score is computed for each predictor variable. As the results show, all predictors have a VIF score that is less than 5, so multicollinearity does not exist in any of the predictor variables.

4 Initial Model Assumption Checks

The model assumptions for multiple linear regression are:

1. Linearity : There exists a linear relationship between the independent and dependent variables.
2. Normality : The error terms are normally distributed.
3. Homoscedascity : The variance of the error terms is constant.
4. Check for any outliers, leverage points, and influential points.
5. Independence : The error terms are independent of each other.
6. No multicollinearity : The independent variables should not be highlly correlated with each other.



Residuals vs Fitted Plot

In this plot, we are looking to check the linearity assumption of the model. Since the line is horizontally straight, with the average of the residuals being close to 0, the model fulfills the linearity assumption.

Normal Q-Q Plot

In this plot, we are looking to check the normality assumption. Since the data points follow a straight line with an approximate slope of 1, that implies that the error terms are normally distributed.

Scale-Location Plot

In this plot, we check for homoscedascity, where we want the error terms to have constant variance. The plot does not show any apparent patterns, but after running the Breusch-Pagan test, we get a p-value of 3.7956e-14, which shows that the model's residuals exhibit non-constant variance. Therefore, it is still inconclusive as to whether or not our model fulfills the constant variance assumption.

Residuals vs Leverage Plot

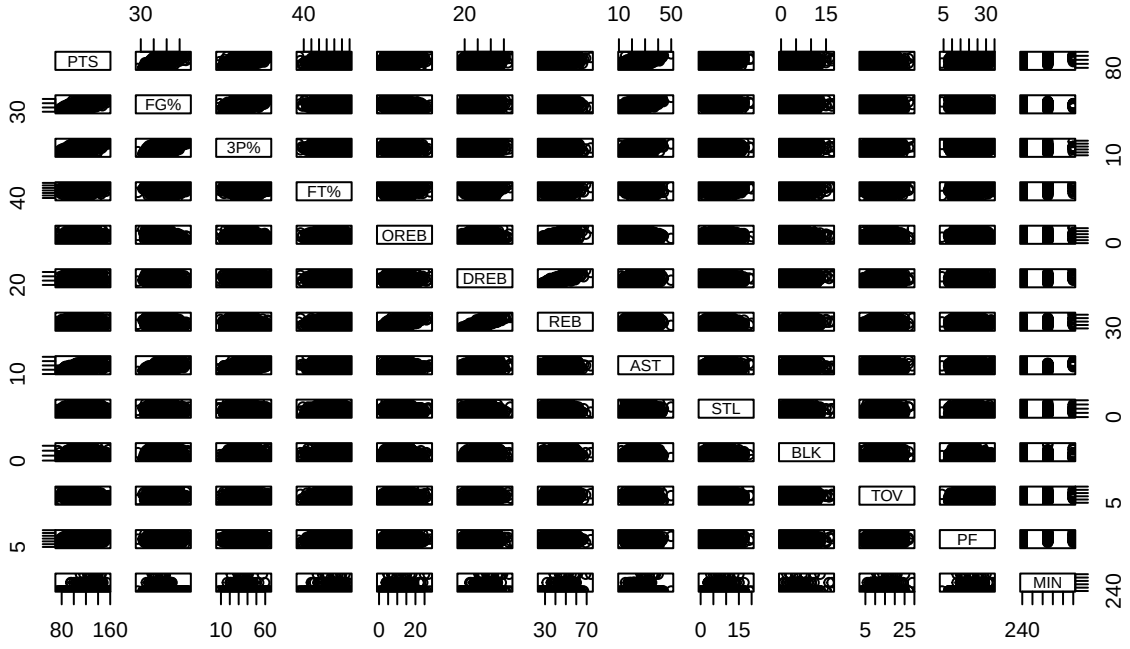
In this plot, we check for any outliers, leverage points, and influential points. There are a few points that have standardized residuals greater than 3 and -3, so some outliers are evident in the model. As for leverage points, there are a few points with leverage above 0.015, but the rest of the data points have a smaller leverage. Lastly, the plot has very few influential points, with most data points falling within Cook's distance.

Durbin-Watson Test for Independence

To check the independence assumption, we use the Durbin Watson test to see if performances in one game correlate with following games. The results show that the model has a Durbin-Watson statistic of 1.937683 and a p-value of 0.108. With a decently high p-value, it suggests little to no autocorrelation in the residuals, so the model fulfills the independence assumption.

Pairwise Correlation Plot

Furthermore, by analyzing the pairwise correlation plot, we see that the predictor variables showcase a linearly correlation with the response variables.



5 Model Selection

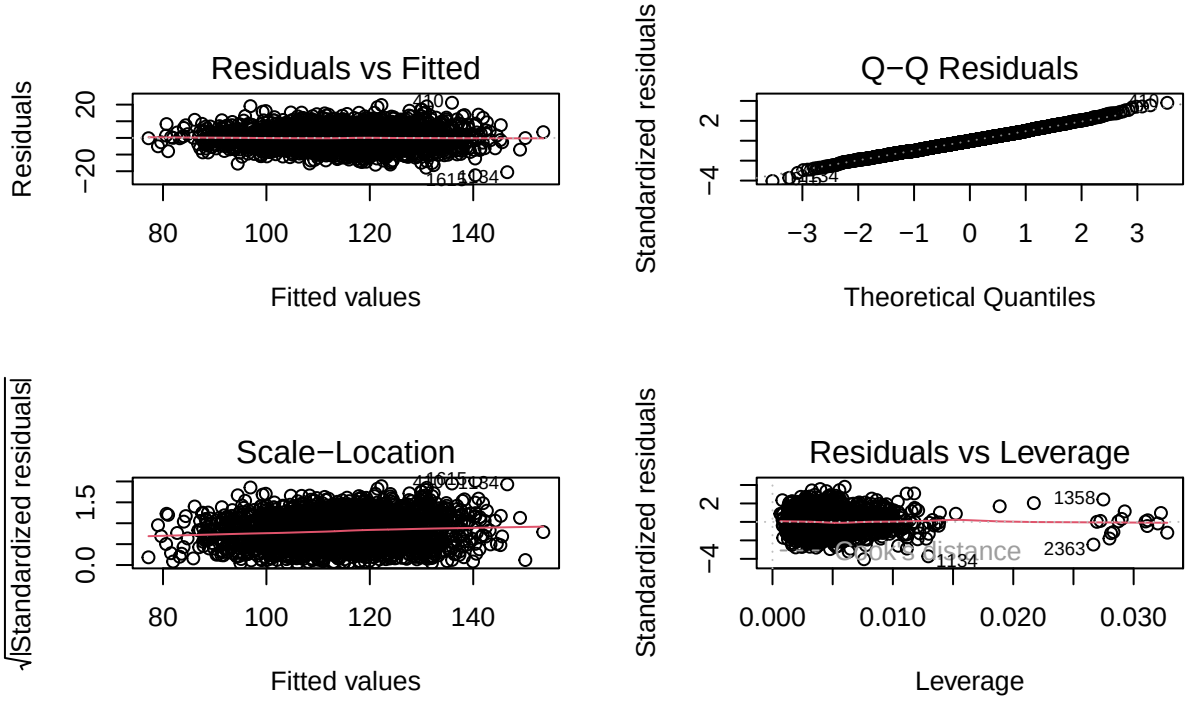
Although the initial multiple linear regression model did relatively well in predicting the total points scored in each NBA game, we still want to find the subset of predictor variables that can perform at optimal rates. Therefore, we use stepwise regression on the dataset in order to derive the best predictive model for total points scored. Below, we see the step by step results of the forward stepwise regression algorithm.

Table 4: Stepwise Regression Results

Iteration	Model.Equation	AIC
Iteration 1	PTS ~ ‘FG%‘+‘3P%‘+‘FT%‘+OREB+DREB+AST+STL+BLK+TOV+PF+MIN	8450.36
Iteration 2	PTS ~ ‘FG%‘+‘3P%‘+‘FT%‘+OREB+DREB+AST+STL+TOV+PF+MIN	8450.12

New MLR Model Assumption Checks

However, we must check if the new model satisfies all the model assumptions for multiple linear regression models.



Linearity

Looking at the Residuals vs Fitted plot, we see that the line is horizontally straight and the average of the residuals stayed close to 0. Therefore the model fulfills the linearity assumption.

Normality

Looking at the Normal Q-Q plot, we see that the data points continue to follow a straight line with that has a slope of 1. This means that the error terms are normally distributed.

Homoscedasticity (Constant Variance)

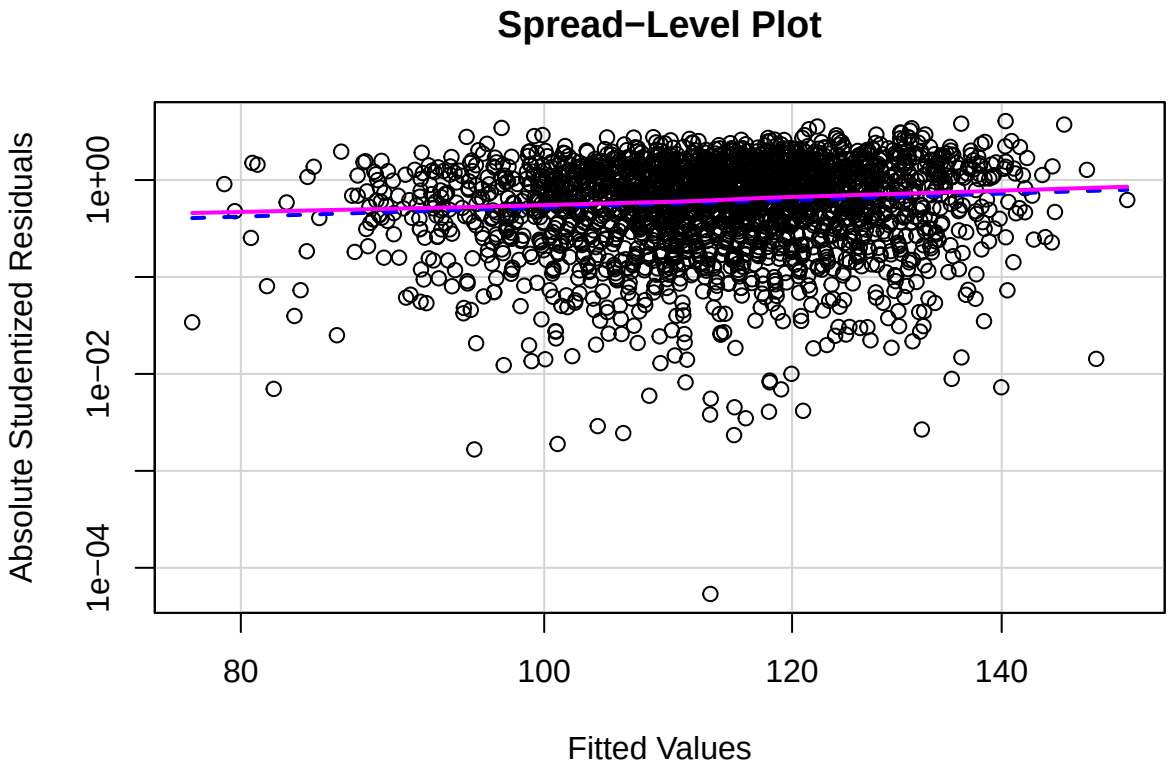
Looking at the Scale-Location plot, we see that the plot still does not show any apparent patterns, so the model fulfills the constant variance assumption. To do a more in-depth analysis, we analyze the results from

the Breusch-Pagan test and use the Spread-Level Plot. In the Breusch-Pagan test, we see that our p-value actually increases, but it is still insignificant at a value of 4.5419e-14. This is an improvement on our initial model, which had a lower p-value result. But if we look at the spread-level plot, we see that there are no signs of a pronounced funnel shape, and since the slope is nearly flat, it shows that the residuals maintain constant variance.

```
ncvTest(step)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 56.91877, Df = 1, p = 4.5419e-14
```

```
spreadLevelPlot(step, main = "Spread-Level Plot")
```




```
##
## Suggested power transformation:  0.02322289
```

Standardized Residuals, Leverage Points, and Cook's Distance

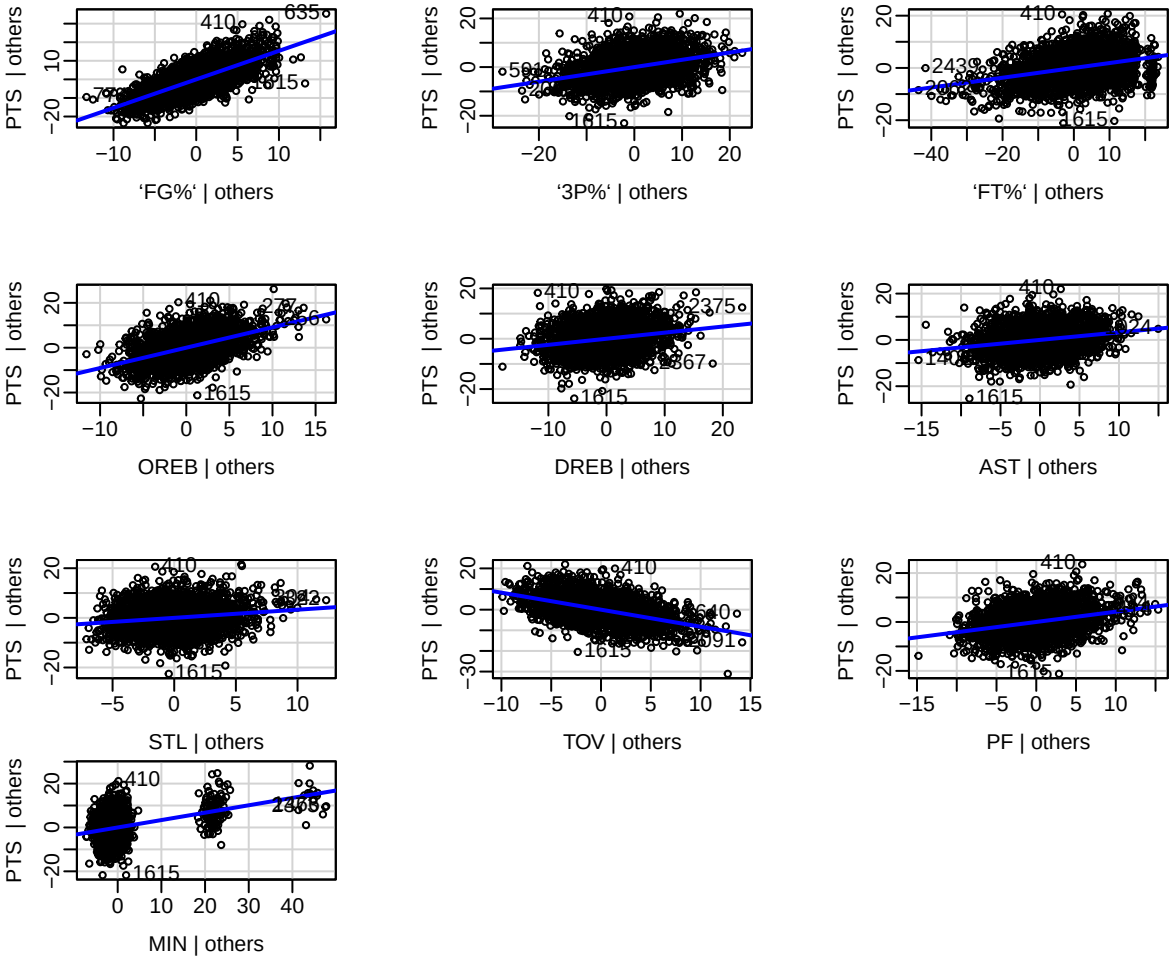
Looking at the Residuals vs Leverage plot, we see that there still remains some data points that have standardized residuals greater than 3 and -3. Also, the plot still has a few influential points outside of Cook's distance, but most data points still remain within Cook's distance. Therefore, there are still some outliers are evident in the model. As for leverage points, there is a similar number of points with leverage above 0.015, but the majority of the data points have a smaller leverage.

Independence

Using the Durbin Watson test, we see that the model has a Durbin-Watson statistic of 1.938689 and a p-value of 0.102. With a decently high p-value, it suggests that there still exists little to no autocorrelation in the residuals. Therefore, the model fulfills the independence assumption.

Multicollinearity

In order to detect if multicollinearity is evident in the model, we plot an added variable plot for each predictor variable and check if they exhibit weak slopes, high variance, or strange patterns that indicate that the predictor is redundant.



After analyzing the added variable plots, we see the partial relationship between each predictor and the PTS response variable, while holding all other predictors constant. In the plots, they all have non-horizontal slopes passing through the origin, and all the data points are randomly scattered. Although there are a few outliers, we see that no single data point is dominating the slope. Also, all the predictors have a VIF score below 5. Taking all these into account, that shows that multicollinearity remains nonexistent in our model.

6 Final Model Fit and Interpretation

Based on the results of the new model, we can conclude that it is an improvement from our initial model. After having performed stepwise regression on our model, we obtained the most optimal predictor variables to predict the total score of each NBA game, which contains FG%, 3P%, FT%, OREB, DREB, AST, STL, TOV, PF, and MIN. After removing the BLK statistic, we saw a slight decrease in the AIC score from 8450.36 to 8450.12, which shows that our model is better fitting to the data and that it is explaining the data without unnecessary complexity. By removing the BLK statistic, the model is able to generalize better to new observations, thus reducing any possibilities of overfitting and helping to improve interpretability. Although we still have the same R^2 and adjusted R^2 values, we now have that all predictors are statistically significant based on their t-test and F-test results.

As for the interpretation of the final model coefficients, we have the following:

1. The 1.535(`FG%`) term shows a one unit increase in FG% leads to a 1.535 point increase in the total scoring output in each NBA game. Higher shot making percentage means that more points are being scored, which explains why the term has a high weight coefficient.
2. The 0.299(`3P%`) term shows a one unit increase in 3P% leads to a 0.299 point increase in the total scoring output in each NBA game.
3. The 0.187(`FT%`) term shows a one unit increase in FT% leads to a 0.187 point increase in the total scoring output in each NBA game.
4. The 0.908(OREB) term shows a one unit increase in OREB leads to a 0.908 point increase in the total scoring output in each NBA game.This leads to more possessions/opportunities for the team to score more points.
5. The 0.242(DREB) term shows a one unit increase in DREB leads to a 0.242 point increase in the total scoring output in each NBA game. This leads to more possessions/opportunities for the team to score more points.
6. The 0.325(AST) term shows a one unit increase in AST leads to a 0.325 point increase in the total scoring output in each NBA game.
7. The 0.328(STL) term shows a one unit increase in STL leads to a 0.328 point increase in the total scoring output in each NBA game. This makes sense since steals lead to more possessions and thus more scoring.
8. The -0.825(TOV) term shows a one unit increase in TOV leads to a 0.825 point decrease in the total scoring output in each NBA game. This makes sense since turnover lead to less possessions and thus less scoring.
9. The 0.420(PF) term shows a one unit increase in PF leads to a 0.420 point increase in the total scoring output in each NBA game.
10. The 0.337(MIN) term shows a one unit increase in MIN leads to a 0.337 point increase in the total scoring output in each NBA game. The higher amount of minutes played, the more potential there is to score more points.

Thought the coefficients are relatively close to the coefficients of the initial model, it is much more clear to understand the impact that each predictor has on the total score. By removing the BLK statistic, which does not neccessarily correlate to higher points scored, we get a better set of coefficients that attribute higher weight to predictors that actually make a different when it comes to predicting total output.

7 Limitations

However, we should note that this model faces a variety of limitations. For one, 6 predictor variables were omitted from being used in our model, which were FGM, FGA, FTM, FTA, 3PM, and 3PA. These predictors are very crucial when it comes to predicting the total scoring output, as these statistics directly tell how many points were scored in a game. Adding these predictors would result in a high R^2 and adjusted R^2, which can be beneficial in helping to provide better predictions of total scoring output. Although a high R^2 can imply possible overfitting or multicollinearity, we can check for these issues and reassure any critics as to why it is beneficial to add these 6 crucial predictors. Also, with the R^2 value being approximately 0.81, that means that 19% of the variance in PTS is missing, which could be due to factors that are not included in the model. This could include other key statistics such as opponent ranking, game pace, as well as input about the conditioning of the players (i.e. fatigue or injury).

8 Conclusion

Based on the results from the event, we can conclude that there exists a positive linear relationship between the response variable PTS and the predictors FG%, 3P%, FT%, OREB, DREB, AST, STL, PF, and MIN. As for the TOV statistic, there exists a negative linear relationship between TOV and PTS. Through our

analysis, we see that our model is able to explain approximately 81% of the variance in PTS, which implies that our model has very strong predictive power, and that the features that were chosen to be included in our model are very relevant when it comes to predicting the amount of points scored in an NBA game. We see that the most influential predictor when it comes to predicting the total points scored is FG%, which makes complete sense. Though this model is high performing when it comes to predicting total points scored, improvements can always be made, as it can play a key part in sports analytics when it comes to helping NBA teams perform to their highest potential.