Modeling NBA Point Totals Using Multiple Linear Regression

Colin MacPherson

2025-03-12

Contents

Introduction

Context In recent years, data has become increasingly important within sports. This is no different in the NBA, where the difference in winning or losing a championship can come down to a single missed shot. Understanding what metrics lead to higher points totals and having the predictive power to estimate a team's points based on these metrics can give teams a greater insight into what tactics they should pursue and what parts of their game need improvement.

Objective The aim of this project is to find a multiple linear regression model capable of producing robust and accurate predictions using our selected metrics.

Descriptive Analysis

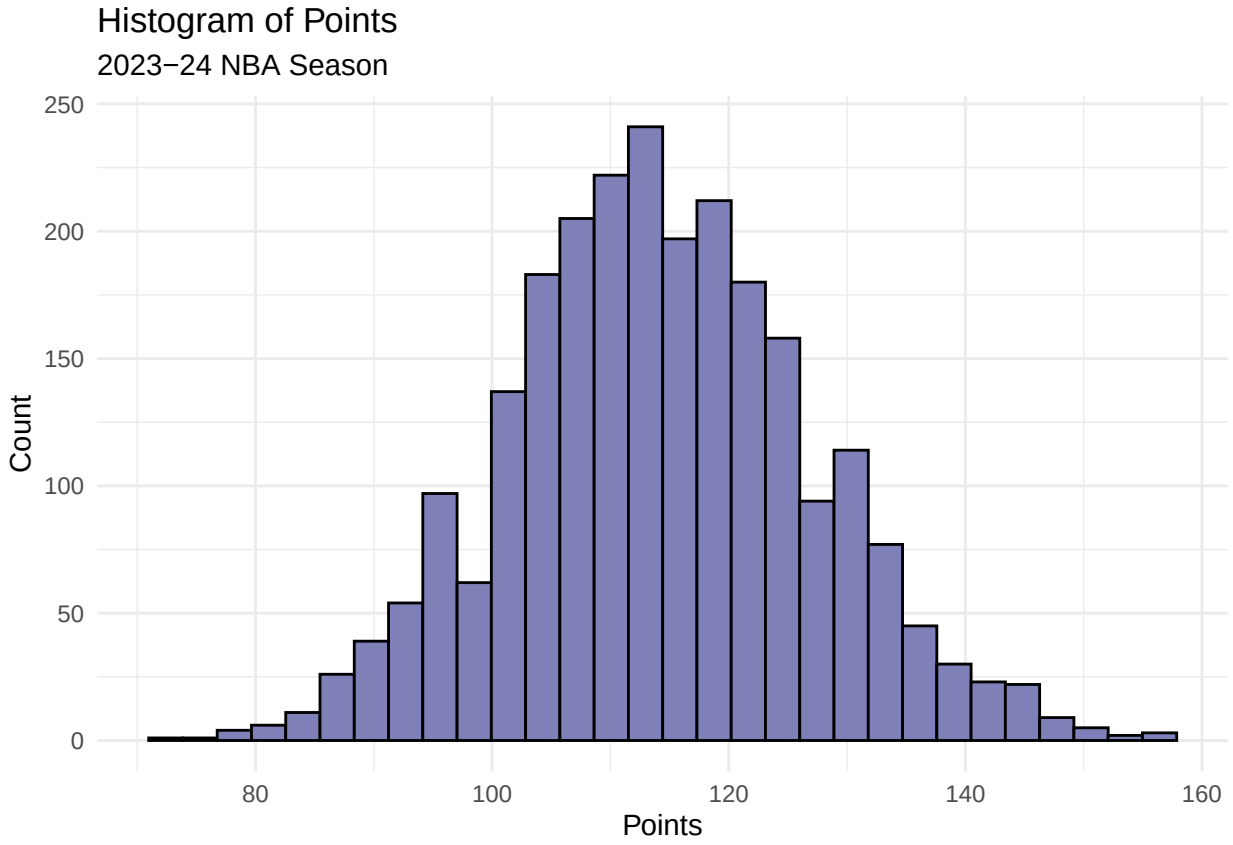
The following section summarizes and highlights the characteristics of the variables used in the model.

Table 1: Variable Descriptions

Variable.Name	Role	Type	Description
PTS	Response	Numeric	Total points scored
FTP	Predictor	Numeric	Free throw percentage (made/free throw attempts)
OREB	Predictor	Numeric	Offensive rebounds (rebounds collected on offense)
DREB	Predictor	Numeric	Defensive rebounds (rebounds collected on defense)
AST	Predictor	Numeric	Assists (passes leading directly to a made basket)
STL	Predictor	Numeric	Steals (forcing an opponent turnover by taking the ball)
BLK	Predictor	Numeric	Blocks (deflecting an opponent's shot attempt)
TOV	Predictor	Numeric	Turnovers (losing possession of the ball)
PF	Predictor	Numeric	Personal fouls committed
DIFF	Predictor	Numeric	Plus-minus (point differential in a game)

Table 2: Summary Statistics

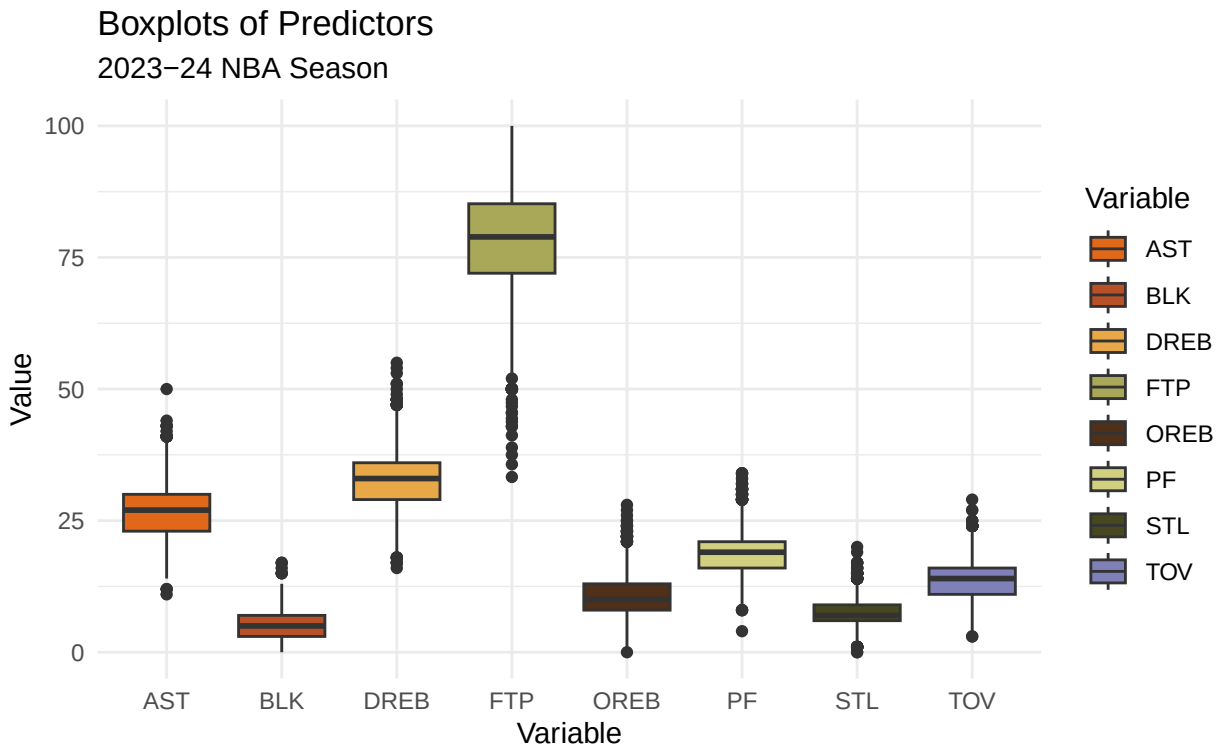
	PTS	FTP	OREB	DREB	AST	STL	BLK	TOV	PF	DIFF
Min.	73.00	33.30	0.00	16.00	11.00	0.00	0.00	3.0	4.00	-62
1st Q.	105.00	72.00	8.00	29.00	23.00	6.00	3.00	11.0	16.00	-10
Median	114.00	78.90	10.00	33.00	27.00	7.00	5.00	14.0	19.00	0
Mean	114.21	78.33	10.55	32.99	26.67	7.47	5.14	13.6	18.73	0
3rd Q.	123.00	85.20	13.00	36.00	30.00	9.00	7.00	16.0	21.00	10
Max.	157.00	100.00	28.00	55.00	50.00	20.00	17.00	29.0	34.00	62



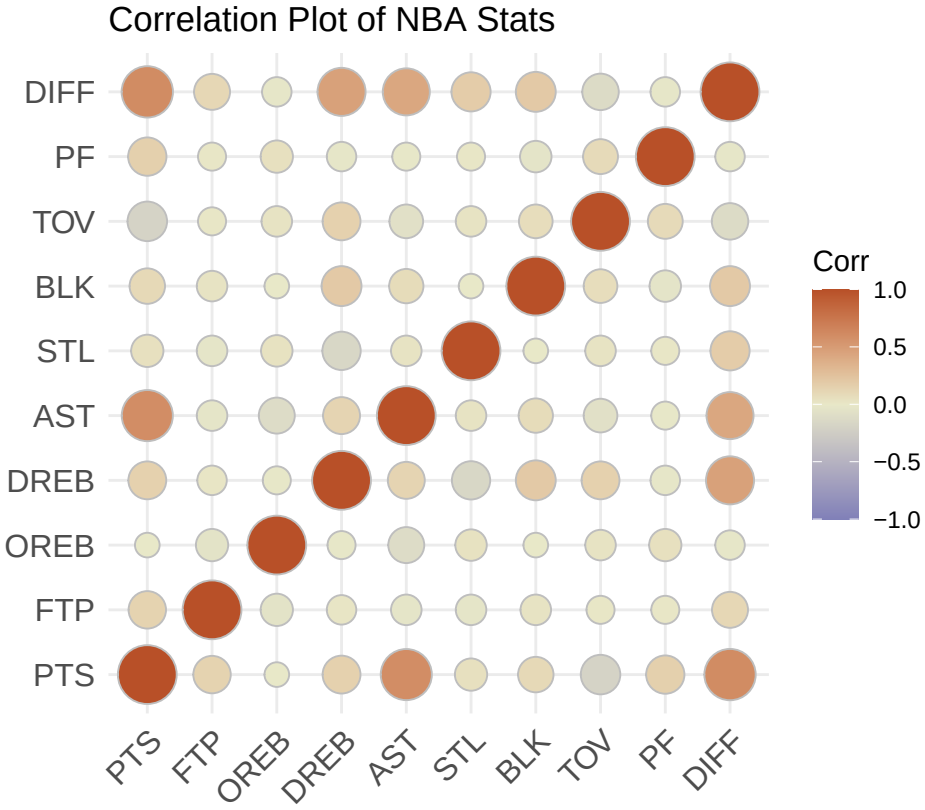
The histogram of our PTS variable shows that it is approximately normally distributed.

Method	df	Statistic	P-value
Jarque Bera Test	2	5.190428	0.0746299

The Jarque-Bera test confirms that the variable is normally distributed. While this is not a required assumption in multiple linear regression, it increases the likelihood that our residuals are normally distributed.



The graph above does excludes the DIFF variable solely for aesthetic purposes.



From the correlation plot, the DIFF and AST variables appear to have the highest correlation with PTS. There does not appear to be too much correlation between our predictors suggesting multicollinearity will not be an issue in our model.

Model Selection

To identify the most relevant variables for the model, we employ both the Boruta algorithm and stepwise regression. This ensures only the most relevant features are retained for further analysis.

Table 4: Boruta Algorithm Results

	meanImp	medianImp	minImp	maxImp	normHits	decision
AST	77.701625	78.549427	70.6559436	83.847729	1.0000000	Confirmed
DIFF	76.331712	76.531520	68.0203505	82.912958	1.0000000	Confirmed
PF	18.556437	18.480340	15.3974839	21.176177	1.0000000	Confirmed
DREB	13.298848	13.367837	11.1435722	15.683498	1.0000000	Confirmed
TOV	12.805387	12.718576	9.6016445	15.636515	1.0000000	Confirmed
FTP	12.306137	12.216829	10.1141502	14.887859	1.0000000	Confirmed
STL	3.112249	3.069727	0.0919647	5.575289	0.8513514	Confirmed
OREB	2.085513	2.072503	-0.5202150	5.051219	0.5945946	Tentative
BLK	1.184175	1.013014	-1.0797476	3.353691	0.1081081	Rejected

The boruta algorithm results are in line with our preliminary takeaways from our correlation plot. AST and DIFF are deemed to be the most important variables. BLK is the only variable rejected— thus, we will likely not include it in the model.

Table 5: Forward Stepwise Regression Results

Iteration	FTP	OREB	DREB	AST	STL	BLK	TOV	PF	DIFF
1									*
2				*					*
3				*				*	*
4			*	*				*	*
5	*		*	*				*	*
6	*		*	*			*	*	*
7	*		*	*	*		*	*	*
8	*	*	*	*	*		*	*	*
9	*	*	*	*	*		*	*	*

Stepwise Regression confirms that BLK should not be included in the model.

Final Model

Based on our results from Boruta and Stepwise regression, we arrive at our final model:

$$PTS = 82.306 + 0.126FTP + 0.184OREB - 0.324DREB + 1.033AST - 0.342STL - 0.312TOV + 0.534PF + 0.405DIFF$$

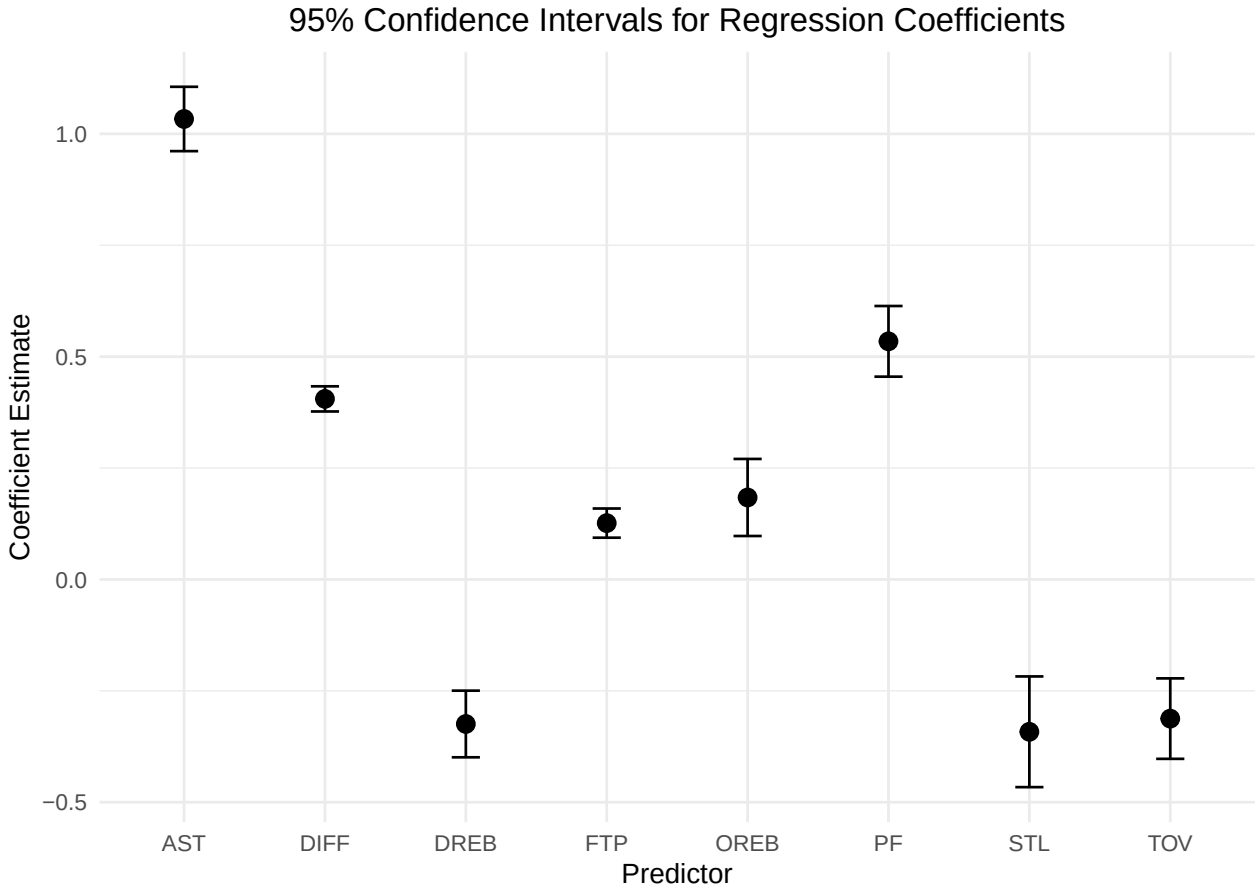
Table 6: Regression Model Results

	Estimate	SE	t.value	p.value	Significance
(Intercept)	82.3055443	2.5475781	32.307368	4.473e-191	***
FTP	0.1264090	0.0166928	7.572667	5.146e-14	***
OREB	0.1839252	0.0441014	4.170510	3.145e-05	***
DREB	-0.3244114	0.0381329	-8.507397	3.048e-17	***

	Estimate	SE	t.value	p.value	Significance
AST	1.0334346	0.0368114	28.073742	1.327e-150	***
STL	-0.3419516	0.0633453	-5.398219	7.377e-08	***
TOV	-0.3124024	0.0460461	-6.784562	1.455e-11	***
PF	0.5343455	0.0404395	13.213449	1.479e-38	***
DIFF	0.4051912	0.0144040	28.130394	3.976e-151	***

	Value
Adjusted R^2	0.5857428

All of our selected variables are statistically significant as indicated by their extremely small p-values. In addition, the $R^2_{adj} = 0.5857428$ indicates the model explains ~58.6% of the variation in PTS. This value indicates a good fit and suggests the model may be useful for prediction.



Diagnostics

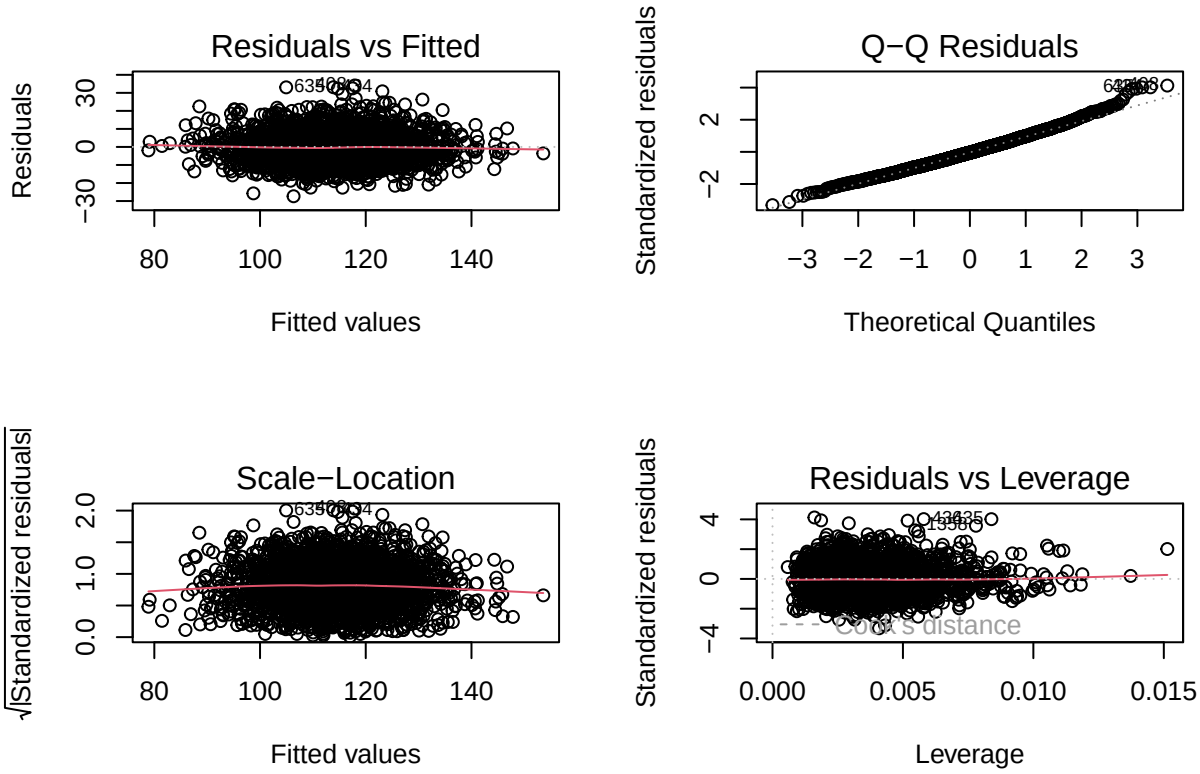
Model Assumptions

1. There is a linear relationship between the predicted variable and the predictors

2. Error terms are normally distributed with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2 \ \epsilon \ N(0, \sigma^2)$

3. Outliers, Leverage points, and Influential points are addressed

4. There is no multicollinearity



Residuals vs Fitted: The line appears to be quite horizontal at 0, indicating the linearity assumption holds and $E(\epsilon) = 0$

Q-Q Residuals: The majority of points fall on the line indicating the normality assumption holds.

Standardized Residuals vs Fitted Values: There is no discernible pattern from the plot indicating the constant variance assumptions holds.

Standardized Residuals vs Leverage: There are points outside of the [-2,2] range for the Standardized Residuals indicating the existence of outliers. However, all of the points lie within the bounds for Cook's distance indicating there are no influential points.

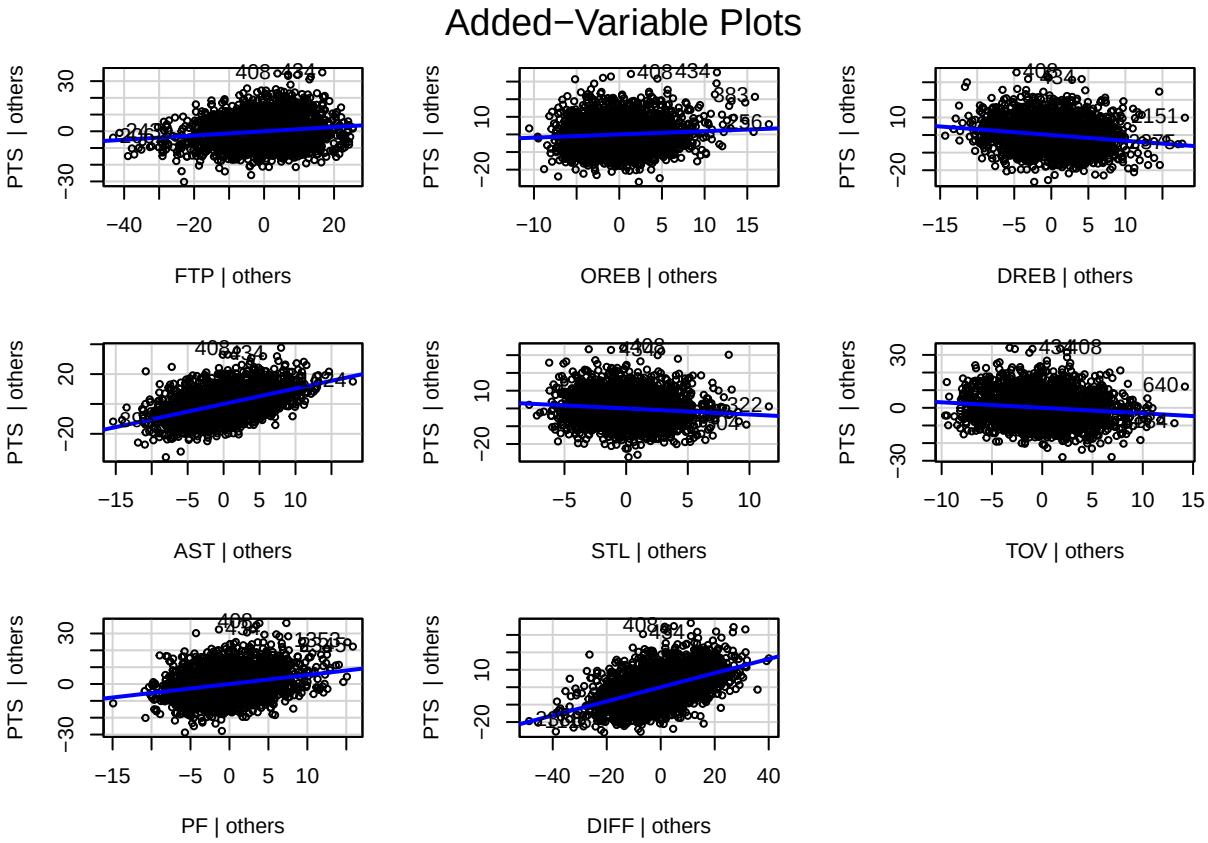
Multicollinearity

Table 8: VIF

	x
FTP	1.033236
OREB	1.019417
DREB	1.530571
AST	1.268374
STL	1.149406

	x
TOV	1.108017
PF	1.012553
DIFF	1.859741

The VIFs are all close to 1 which indicates little to no correlation between the variables. This means the model is not influenced by collinearity.



The Added Variable Plots reinforce the absence of multicollinearity. All of the variables are shown to have a significant effect on the model.

Conclusion

Based on these results, we conclude that Points from an NBA game can be modeled by the following key statistics: assists, point differential, rebounds(offensive and defensive), turnovers, steals, and personal fouls. The model is able to explain ~58.6% of the variation in Points, which is quite high given the randomness and variability of NBA games.

The RMSE of the model is 8.25 which means our predicted values for points are, on average, 8.25 points off.

The model passes all the diagnostics for a multiple linear regression model indicating our estimates are valid and accurate. The next step would be to test the model on games from the 2024-25 season to determine its predictive accuracy.