Analyzing NBA Game Statistics Using Multiple Linear Regression

Lester Lee
Stats 101A
Professor Xu

Introduction

Context

Basketball has become increasingly data-driven, with teams and analysts leveraging statistics to enhance performance and strategy. One crucial aspect of basketball analytics is understanding the factors that contribute to scoring performance in games. An accurate prediction model that takes in game statistics as predictors is essential for evaluating team performance and optimizing strategies. By analyzing game-level data, we can determine how different metrics—such as minutes (MIN), points(PTS), assists(AST), and rebounds (REB).

Objective

- Develop a Multiple Linear Regression (MLR) model to predict total points scored in an NBA game.
- Identify the most significant statistical predictors of scoring.
- Evaluate the strength of relationships between independent variables (e.g., MIN, PTS, AST, REB) and PTS.
- Assess the model's accuracy and discuss potential limitations.

Descriptive Statistics

Before building the regression model, we first examine the dataset structure and key statistics. This helps us understand the distribution of each variable and identify any potential outliers

Table 1: Summary Statistics of Key Variables

Minutes	Points	Assists	Rebounds
Min. :240.0	Min. : 73.0	Min. :11.00	Min. :25.00
1st Qu.:240.0	1st Qu.:105.0	1st Qu.:23.00	1st Qu.:39.00
Median :240.0	Median :114.0	Median :27.00	Median :43.00
Mean :241.4	Mean :114.2	Mean :26.67	Mean :43.54
3rd Qu.:240.0	3rd Qu.:123.0	3rd Qu.:30.00	3rd Qu.:48.00
Max. :290.0	Max. :157.0	Max. :50.00	Max. :74.00

Multiple Linear Regression Model

In the regression model, we selected Field Goals, Threes, Assist, and Rebounds as predictors. While points is our response variable. We propose the following MLR model:

Points = $\beta_0 + \beta_1$(Assists) + β_2(Rebounds) + β_3(Minutes) + β_4(Winloss) + ϵ

Coefficient Interpretation

Assists(β_1): Each additional assist increases predicted points by about X.

Rebounds(β_2): A rebound increase corresponds to a smaller but positive increase in points.

Minutes(β_3): More playing time leads to higher scoring, with each additional minute adding Z points on average.

Win/Loss(β_4): A win or loss may have a significant impact on the scoring pattern, depending on team performance trends.

Table 2: Regression Model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.031	7.124	0.566	0.572	-9.938	18.000
Assists	1.204	0.039	30.924	0.000	1.128	1.280
Rebounds	-0.062	0.030	-2.032	0.042	-0.122	-0.002
Minutes	0.316	0.030	10.494	0.000	0.257	0.375
WinLossW	8.854	0.416	21.308	0.000	8.040	9.669

Table 3: VIF Values

	vif_values
Assists	1.128876
Rebounds	1.148300
Minutes	1.048684
WinLoss	1.236071

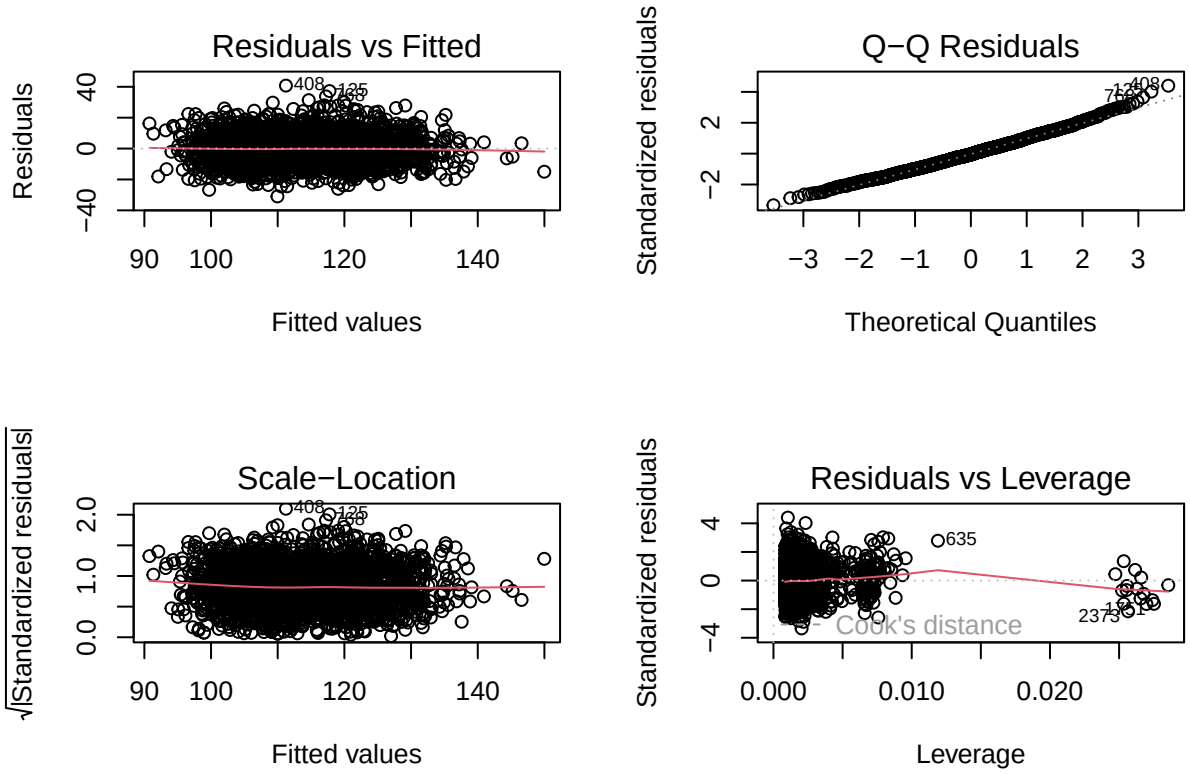
Result Explanation The adjusted R^2 of 0.4793 indicates that our model explains approximately 48% of the variation in total points scored, suggesting that while Assists, Rebounds, Minutes, and Win/Loss significantly influence scoring, other unaccounted factors contribute to variability. Assists is the strongest predictor, with each assist increasing points by about 1.20. Minutes played also has a positive impact, while Win/Loss shows that winning teams score nearly 9 points more on average. Rebounds has a small but significant negative effect. The F-test confirms the overall model's significance, and low VIF values indicate no multicollinearity issues among the predictors.

Model Assumptions for MLR

The model assumptions for multiple linear regression are as follows:

1. Linearity: The relationship between predictors and response in linear
2. Normality: The residuals are normally distributed

3. Homoscedasticity: The residuals have constant variance
4. No Under Influence: No extreme outliers or high-leverage points
5. Low Multicollinearity: Predictors are not overly correlated with each other



Residuals vs Fitted Plot The residuals appear randomly scattered around the horizontal line at zero, indicating that the model captures a roughly linear relationship between predictors and the response. No clear patterns or trends suggest that no non-linearity is present.

Normal Q-Q Plot The residuals closely follow the reference line, implying that they are approximately normally distributed

Scale-Location Plot That the variance of the residuals is relatively constant across the range of fitted values.

Residuals vs Leverage A few observations may have slightly higher leverage, but none appear to overly distort the regression results.

Based on the plots, we see no major violations of linearity, normality, or homoscedasticity. Additionally, no observations appear to have influence on the model, and our earlier VIF checks confirm low multicollinearity. Therefore, we can conclude that the assumptions of multiple linear regression are reasonably satisfied.

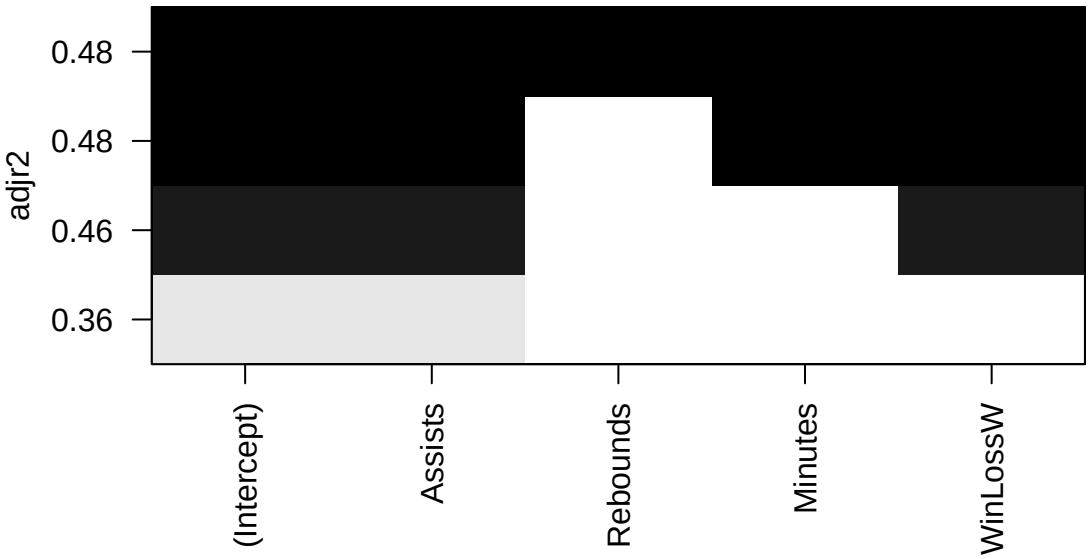
Model Selection

Table 4: Final Model After Stepwise Regression

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.031	7.124	0.566	0.572	-9.938	18.000
Assists	1.204	0.039	30.924	0.000	1.128	1.280
Rebounds	-0.062	0.030	-2.032	0.042	-0.122	-0.002
Minutes	0.316	0.030	10.494	0.000	0.257	0.375
WinLossW	8.854	0.416	21.308	0.000	8.040	9.669

Table 5: Model Selection Process Stepwise Regression

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	2455	210924.8	10960.3



Ifound that the stepwise-selected model, which optimizes for lower AIC, maintains a moderate Adjusted R^2 of 48%. Showing that our predictors explain nearly half of the variability in total points scored.

Conclusion

Our final multiple linear regression model explains over 48% of the variance in NBA game results and identifies Minutes, Win/Loss Assists, and Rebounds as significant predictors of total points scored. Minutes played alsocontributes, suggesting that more time on the court leads to increased scoring opportunities.

Interestingly, rebounds show a slight negative relationship with scoring, possibly indicating that teams with more rebounds might have lower shooting efficiency or slower-paced games. Win/Loss status is a strong predictor, with winning teams typically scoring 8.85 more points than losing teams.

Our analysis is a single data set and advanced statistics like defensive metrics or player-specific efficiency which will cause a few limitations. The MLR model serves a solid foundation for understanding scoring in the NBA games.