Applying a Multiple Linear Regression to Predict NBA Points

Matthew Dick

2025-03-22

Introduction

Context

In the context of NBA analytics, understanding which game statistics are most predictive of a team's total points scored can provide valuable insights for coaches, analysts, and fans. Points are the primary outcome that determines the result of a basketball game, and identifying the key factors that drive scoring can support data-driven decisions in strategy and player evaluation.

Dataset Introduction

The dataset used in this project was obtained from NBA.com and contains a wide range of game-level statistics for professional basketball teams. The table below provides a description of each variable in the dataset including a descritpion of removed or unused

```
library(knitr)
library(kableExtra)

vars <- data.frame(
  Variable = c("Team", "Match Up", "Game Date", "W/L",
               "MIN", "PTS", "FGM", "FGA", "FG%",
               "3PM", "3PA", "3P%",
               "FTM", "FTA", "FT%",
               "OREB", "DREB", "REB",
               "AST", "STL", "BLK",
               "TOV", "PF", "+/-"),

  Role = c("Not Used", "Not Used", "Not Used", "Not Used",
           "Predictor", "Response", "Not Used", "Not Used", "Predictor",
           "Not Used", "Not Used", "Predictor",
           "Not Used", "Not Used", "Predictor",
           "Predictor", "Predictor", "Removed",
           "Predictor", "Predictor", "Predictor",
           "Predictor", "Predictor", "Predictor"),

  Description = c(
    "Name of the team; excluded due to being a categorical label" ,
    "Game matchup (team vs opponent); dropped as a string variable",
    "Date of the game; removed due to lack of predictive ability",
    "Win or loss outcome; removed to avoid response redundancy",

    "Total minutes played by the team during the game.",
    "Total points scored by the team - this is the target outcome.",
```

```
    "Total field goals made by the team; excluded due to collinearity with FG%",
    "Total field goals attempted; excluded due to collinearity with FG%",
    "Field goal shooting percentage (FGM/FGA * 100).",

    "Three-point field goals made; excluded due to collinearity with 3P%",
    "Three-point field goals attempted; excluded due to collinearity with 3P%",
    "Three-point shooting accuracy (3PM/3PA * 100).",

    "Free throws made; excluded due to collinearity with FT%",
    "Free throws attempted; excluded due to collinearity with FGT%",
    "Free throw percentage (FTM/FTA * 100).",

    "Offensive rebounds recorded",
    "Defensive rebounds recorded",
    "Total rebounds (OREB + DREB); excluded due to collinearity with OREB and DREB",

    "Assists made during the game",
    "Steals made by the team",
    "Blocked shots",

    "Turnovers committed.",
    "Personal fouls committed.",
    "Final score differential (team score minus opponent score)"
  )
)


vars %>%
  kable("latex", booktabs = TRUE, caption = "Overview of Dataset Variables") %>%
  kable_styling(
    latex_options = c("striped", "hold_position"),
    font_size = 9
  )
```

Table 1: Overview of Dataset Variables

Variable	Role	Description
Team	Not Used	Name of the team; excluded due to being a categorical label
Match Up	Not Used	Game matchup (team vs opponent); dropped as a string variable
Game Date	Not Used	Date of the game; removed due to lack of predictive ability
W/L	Not Used	Win or loss outcome; removed to avoid response redundancy
MIN	Predictor	Total minutes played by the team during the game.
PTS	Response	Total points scored by the team — this is the target outcome.
FGM	Not Used	Total field goals made by the team; excluded due to collinearity with FG%
FGA	Not Used	Total field goals attempted; excluded due to collinearity with FG%
FG%	Predictor	Field goal shooting percentage (FGM/FGA * 100).
3PM	Not Used	Three-point field goals made; excluded due to collinearity with 3P%
3PA	Not Used	Three-point field goals attempted; excluded due to collinearity with 3P%
3P%	Predictor	Three-point shooting accuracy (3PM/3PA * 100).
FTM	Not Used	Free throws made; excluded due to collinearity with FT%
FTA	Not Used	Free throws attempted; excluded due to collinearity with FGT%
FT%	Predictor	Free throw percentage (FTM/FTA * 100).
OREB	Predictor	Offensive rebounds recorded
DREB	Predictor	Defensive rebounds recorded
REB	Removed	Total rebounds (OREB + DREB); excluded due to collinearity with OREB and DREB
AST	Predictor	Assists made during the game
STL	Predictor	Steals made by the team
BLK	Predictor	Blocked shots
TOV	Predictor	Turnovers committed.
PF	Predictor	Personal fouls committed.
+/-	Predictor	Final score differential (team score minus opponent score)

After an initial review, one row was removed due to a missing value ("-") in the `FT%` column. Rather than impute a potentially inaccurate value, the row was excluded due to the robustness of the overall sample size.

Initial MLR Model and Significance

The initial multiple linear regression (MLR) model was fitted using all selected predictors. The variable `BLK` (blocked shots) was removed from the model due to its lack of statistical significance, as indicated by both the model summary and ANOVA output ($p = 0.1085$ in `summary()`, $p = 0.735$ in `anova()`). All remaining predictors were statistically significant at the 0.05 level.

To assess whether the model met the assumption of homoscedasticity, a Breusch–Pagan (BP) test was conducted. The test returned a significant result, indicating that heteroscedasticity was present. To address this violation, a weighted least squares (WLS) transformation was applied, using the inverse of the fitted values' variance as weights.

The final WLS model is specified below:

$$\begin{aligned}\text{PTS} = \; & -33029.36 + 73.47 \cdot \text{MIN} + 346.01 \cdot \text{FG\%} + 72.18 \cdot \text{3P\%} + 41.44 \cdot \text{FT\%} \\ & + 199.26 \cdot \text{OREB} + 68.48 \cdot \text{DREB} + 75.99 \cdot \text{AST} + 90.02 \cdot \text{STL} \\ & - 185.50 \cdot \text{TOV} + 86.90 \cdot \text{PF} - 10.17 \cdot (\pm)\end{aligned}$$

```
library(readxl)
df <- read_excel("Dataset1.xlsx", col_types = "text")
```

```r
df$`FT%` <- as.numeric(ifelse(df$`FT%` == "-", NA, df$`FT%`))
df <- df %>% filter(!is.na(`FT%`))
```

```r
library(kableExtra)

final_summary <- data.frame(
  Predictor = c("(Intercept)", "MIN", "FG%", "3P%", "FT%", "OREB", "DREB", "AST", "STL", "TOV", "PF", "+
  B         = c(-33029.361, 73.465, 346.014, 72.178, 41.439, 199.262, 68.483, 75.986, 90.022, -185.496,
  SE        = c(1225.895, 4.852, 7.624, 3.823, 2.455, 7.288, 6.428, 6.371, 10.000, 6.972, 6.017, 2.829)
  t_value   = c(-26.943, 15.141, 45.387, 18.880, 16.878, 27.340, 10.653, 11.928, 9.003, -26.607, 14.441
  Pr_t = c(rep("< 2e-16", 11), "0.00033"),
  VIF       = c(NA, 1.063, 3.054, 1.683, 1.056, 1.277, 2.009, 1.676, 1.340, 1.228, 1.032, 3.627)
)

colnames(final_summary) <- c("Predictor", "B", "SE", "t-value", "Pr(>|t|)", "VIF")

kable(final_summary, "latex", booktabs = TRUE,
      caption = "Final WLS Model Summary") %>%
  kable_styling(latex_options = c("striped", "hold_position"))
```

Table 2: Final WLS Model Summary

Predictor	B	SE	t-value	Pr(>|t|)	VIF
(Intercept)	-33029.361	1225.895	-26.943	< 2e-16	NA
MIN	73.465	4.852	15.141	< 2e-16	1.063
FG%	346.014	7.624	45.387	< 2e-16	3.054
3P%	72.178	3.823	18.880	< 2e-16	1.683
FT%	41.439	2.455	16.878	< 2e-16	1.056
OREB	199.262	7.288	27.340	< 2e-16	1.277
DREB	68.483	6.428	10.653	< 2e-16	2.009
AST	75.986	6.371	11.928	< 2e-16	1.676
STL	90.022	10.000	9.003	< 2e-16	1.340
TOV	-185.496	6.972	-26.607	< 2e-16	1.228
PF	86.895	6.017	14.441	< 2e-16	1.032
+/-	-10.171	2.829	-3.596	0.00033	3.627

Results Explanation

The adjusted R^2 of this model is 0.8262, indicating that approximately 82.62% of the variance in points scored (PTS) is explained by the included predictors. This suggests a strong overall fit. An F-test was conducted to evaluate the model's overall significance, returning a large F-statistic of 1062.649 with a p-value less than 0.001, confirming that the model is statistically significant.

Each individual predictor also passed the significance threshold ($p < 0.05$) based on their t-tests, indicating that all variables meaningfully contribute to explaining PTS. Additionally, all variance inflation factor (VIF) values were below 5, suggesting that multicollinearity is not a concern in this model.

MLR Diagnostics & Assumptions

The model assumptions for multiple linear regression are as follows:

1. Response variable and predictors are linearly related (**linearity**)
2. Error terms are normally distributed (**normality**)
3. Error terms have constant varinace (**normality**)
4. Outliers, leverage points, and influential points
5. Predictors don't demonstrate high correlation with each other (**multicollinearity**)

These assumptions were evaluated using residual and diagnostic plots, as shown in Figure 1:

```
df$PTS <- as.numeric(df$PTS)

df$PTS_transformed <- df$PTS^2

df_reduced1 <- df[, !(names(df) %in% c("PTS", "Team", "Match Up", "Game Date",
                                       "W/L", "FGM", "FGA", "3PM", "3PA",
                                       "FTM", "FTA", "BLK", "REB"))]

model <- lm(PTS_transformed ~ ., data = df_reduced1)

weights <- 1 / fitted(model)^2
weights <- weights[1:nrow(df_reduced1)]

wls_model1 <- lm(PTS_transformed ~ ., data = df_reduced1, weights = weights)
```

```
par(mfrow = c(2, 2))
plot(wls_model1, which = 1)
plot(wls_model1, which = 2)
plot(wls_model1, which = 3)
plot(wls_model1, which = 5)
```

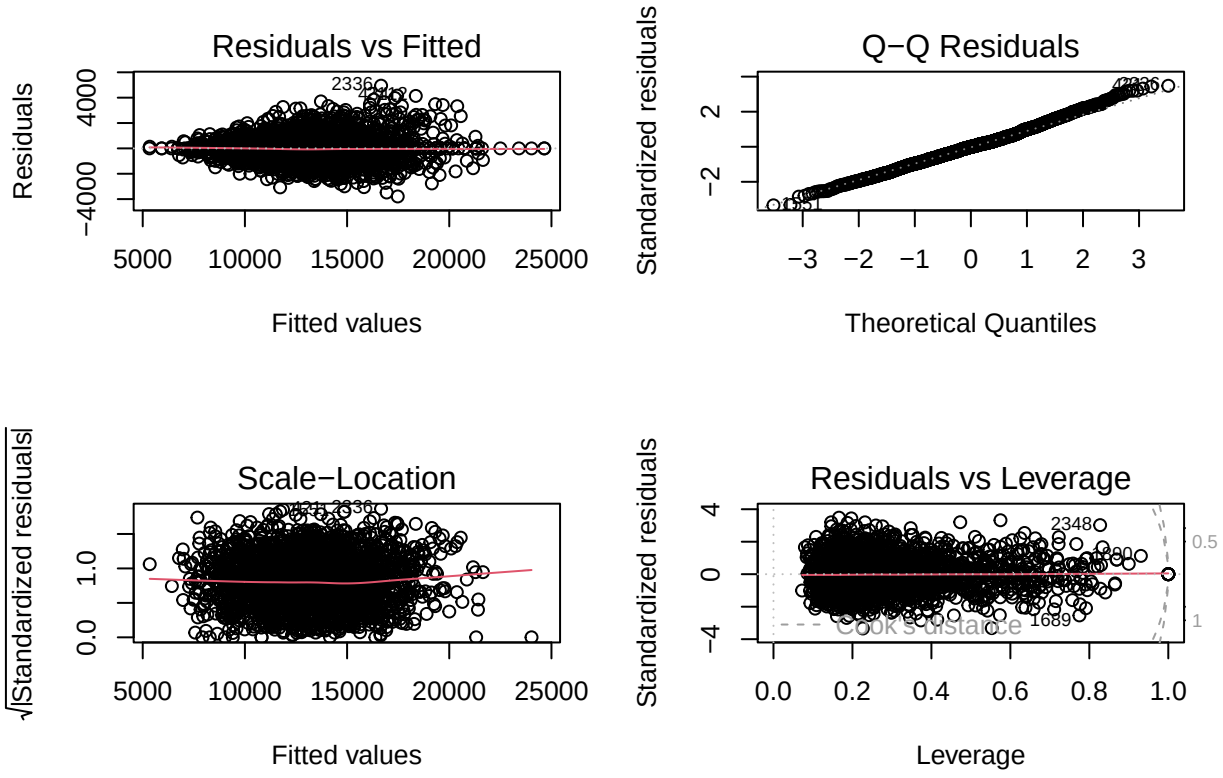


Figure 1: Diagnostic Plots

Linearity: The residuals vs fitted values plot displays a relatively horizontal and random scatter around zero, which suggests a linear relationship between the predictors and the response variable.

Normality of Errors: The Q-Q plot shows that most standardized residuals align closely with the theoretical quantile line, indicating that the residuals are approximately normally distributed.

Homoscedasticity (Constant Variance): The scale-location plot reveals that the spread of standardized residuals remains roughly consistent across fitted values, supporting the assumption of constant variance.

Multicollinearity: All predictors had VIF values below 5 (shown in Table 2), suggesting no concerning multicollinearity among them.

Residual Analysis: To assess model assumptions and fit, standardized residual plots were generated for each predictor, shown in Figure 2:

```
standardized_res <- rstandard(wls_model1)

model_data <- model.frame(wls_model1)

selected_predictors <- c("MIN", "FG%", "3P%", "FT%", "OREB", "DREB", "AST", "STL", "TOV", "PF", "+/-")

par(mfrow = c(3, 4), mar = c(4, 4, 2, 2))

for (var in selected_predictors) {
  x_vals <- model_data[[var]]
  plot(x_vals, standardized_res,
       xlab = var,
       ylab = "Standardized Residuals",
       main = paste("Residuals vs", var),
```

```
       pch = 1)
  abline(h = 0, col = "black")
  abline(h = c(-2, 2), col = "red", lty = 2)
}
```

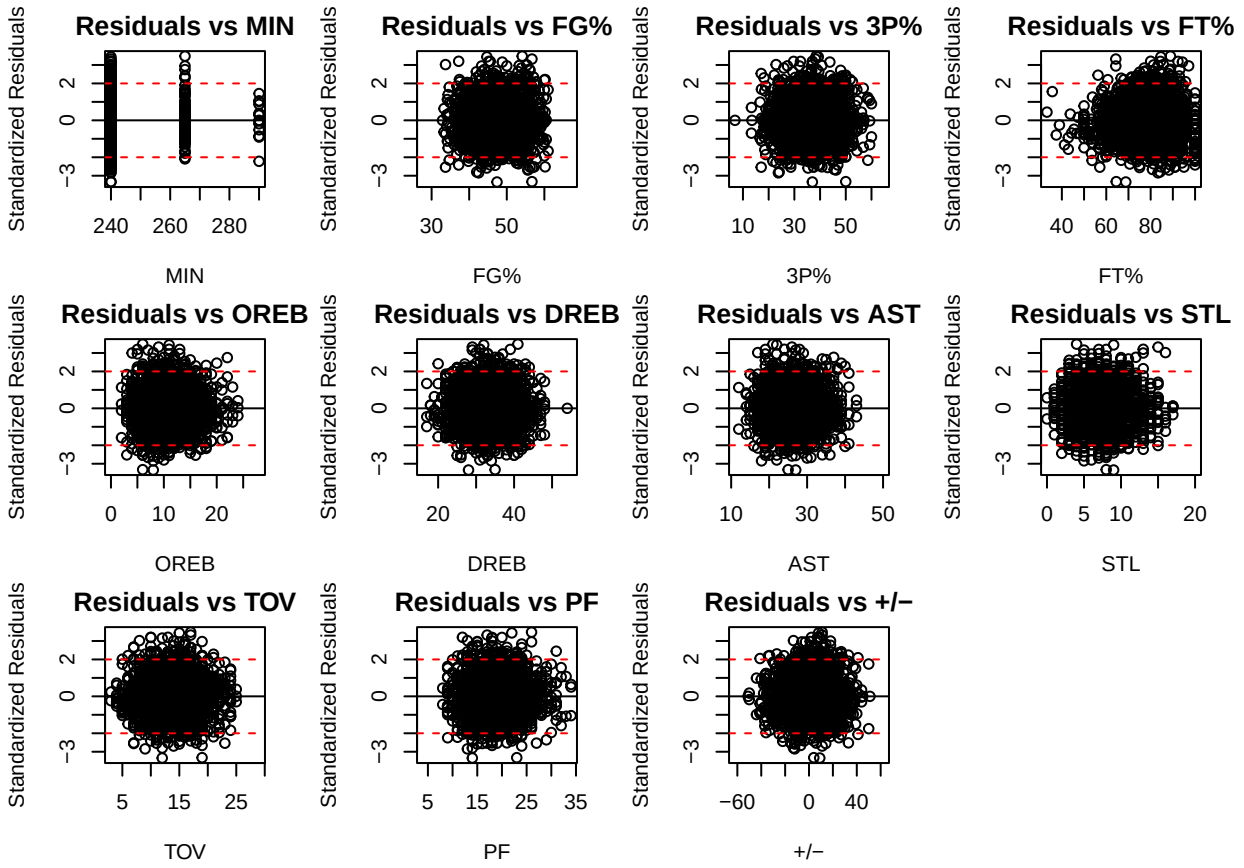


Figure 2: Standardized Residual vs Predictor Plots

Figure 2 displays standardized residuals plotted against each predictor to identify potential outliers in the model. Standardized residuals measure how far each observation's actual value is from the predicted value, in units of standard deviation.

In these plots, any point falling outside the range of $[-2, 2]$ is considered an outlier. Observations beyond this threshold indicate residuals that deviate substantially from what the model expects.

Outliers and Influential Points:
To investigate whether removing high-leverage, high-residual, and influential data points would improve model performance, standard thresholds were applied:

- Leverage: $h_i > \frac{4}{n}$
- Cook's Distance: $D_i > \frac{4}{n-2}$
- Standardized Residuals: $|r_i| > 2$

After removing the rows meeting these conditions, the weighted least squared model was fit. However, the resulting model had a **lower Adjusted $\mathbf{R^2}$** of 0.6196 (compared to 0.8262 before removal), suggesting that the new model explained less of the variation in the response variable.

Additionally, several predictors that were previously statistically significant were not longer significant in the

cleaned model. These finding indicate that the previous model provided a better fit, even with the presence of some outliers.

Model Selection

To identify the optimal subset of predictors, stepwise regression using the Akaike Information Criterion (AIC) was conducted in both directions (forward and backward). The objective was to minimize the AIC value while retaining the strongest explanatory model. Based on the stepwise selection results shown in Table 3, the model including **all predictors** had the lowest AIC value (AIC = -11370.22), meaning no predictors were dropped during the process.

```
library(kableExtra)
library(stringr)

SP_AIC <- data.frame(
  Step = str_pad(c("None", "- +/-", "- STL", "- DREB", "- AST", "- PF", "- MIN",
                   "- FT%", "- 3P%", "- TOV", "- OREB", "- FG%"),
                 width = 8, side = "left"),
  AIC_vals = c(-11370.2, -11359.3, -11292.1, -11260.7, -11233.3, -11171.1,
               -11152.0, -11101.4, -11037.8, -10747.4, -10716.8, -9870.5)
)

colnames(SP_AIC) <- c("Step", "AIC Values")

kable(SP_AIC, "latex", booktabs = TRUE, caption = "Backwards Stepwise Regression Results") %>%
  kable_styling(latex_options = c("striped", "hold_position"), full_width = FALSE)
```

Table 3: Backwards Stepwise Regression Results

Step	AIC Values
None	-11370.2
- +/-	-11359.3
- STL	-11292.1
- DREB	-11260.7
- AST	-11233.3
- PF	-11171.1
- MIN	-11152.0
- FT%	-11101.4
- 3P%	-11037.8
- TOV	-10747.4
- OREB	-10716.8
- FG%	-9870.5

Therefore, all predictors remained in the final model, which supports the earlier results from the t-tests and F-test indicating that each predictor is statistically significant.

Final Selected Predictors: `MIN`, `FG%`, `3P%`, `FT%`, `OREB`, `DREB`, `AST`, `STL`, `TOV`, `PF`, and `+/-`

Model Training

To evaluate model generalizability, the dataset was split into training and test sets (70/30 split). The final weighted least squares model was trained on 70% of the data and used to predict points (PTS) on the

remaining 30%.

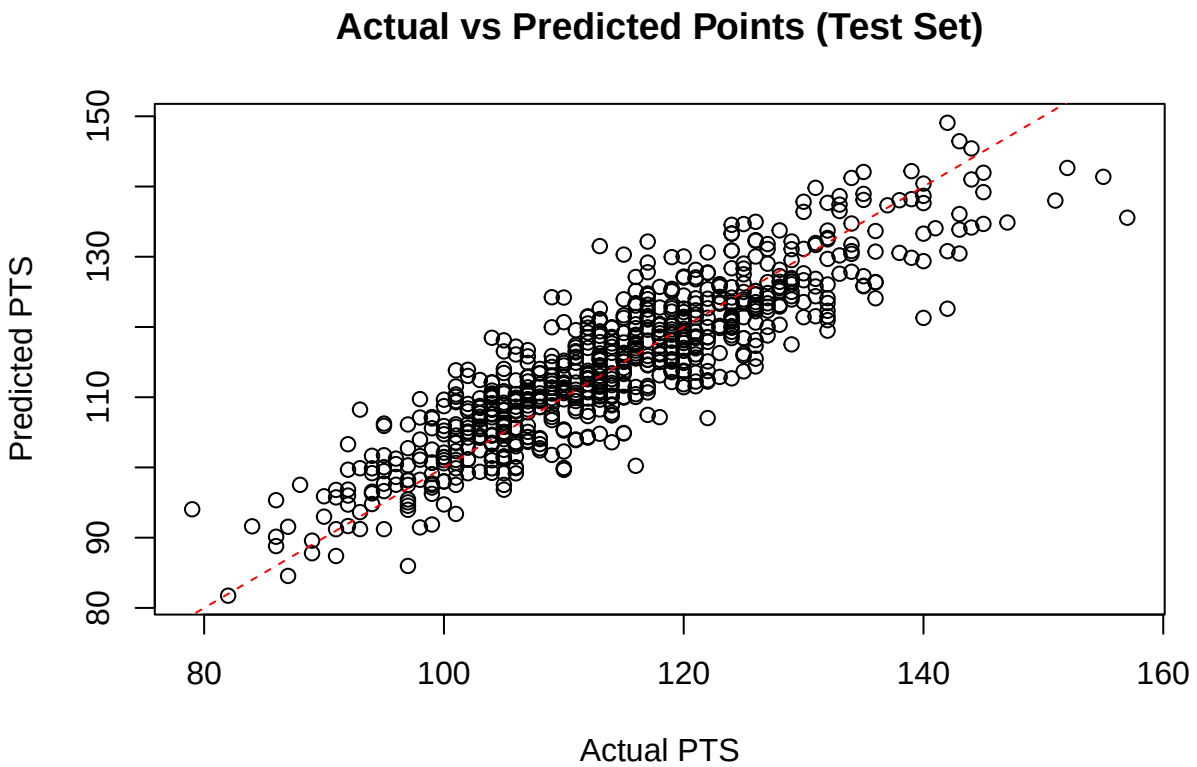
```
exclude_cols <- c("Team", "Match Up", "Game Date", "W/L")
df[] <- lapply(names(df), function(col) {
  if (col %in% exclude_cols || !is.character(df[[col]])) {
    df[[col]]
  } else {
    as.numeric(df[[col]])
  }
})
```

```
set.seed(123)
n <- nrow(df)
df[] <- lapply(df, function(x) if (is.factor(x)) as.numeric(as.character(x)) else x)
train_indices <- sample(1:n, size = 0.7 * n)
train_data <- df[train_indices, ]
test_data <- df[-train_indices, ]

model_train <- lm(PTS ~ MIN + `FG%` + `3P%` + `FT%` + OREB + DREB + AST + STL + TOV + PF + `+/-`, data =

preds <- predict(model_train, newdata = test_data)

plot(test_data$PTS, preds,
     xlab = "Actual PTS", ylab = "Predicted PTS",
     main = "Actual vs Predicted Points (Test Set)")
abline(0, 1, col = "red", lty = 2)
```



The plot in Figure 3 shows the predicted team points against the actual team points for the test set. Each dot represents a single game. The red dashed line represents the ideal scenario where the predicted values match the actual values perfectly. Most points lie close to this reference line, indicating that the model performs well in estimating team points.

To quantify prediction accuracy, the root mean squared error (RMSE) on the test set was calculated to be approximately **5.65**, indicating that, on average, the model's predicted team points deviate from the actual values by about 5.65 points. The test set R^2 was about **79%**, meaning the model explains roughly 79% of the of the variability in team points for new, unseen data.

Interpretation & Discussion

The final weighted least squares (WLS) regression model provides strong explanatory power, with an adjusted R^2 of 0.8262. This means that approximately 82.62% of the variance in NBA team points (PTS) can be explained by the selected predictors, suggesting that the model is highly effective in capturing the underlying patterns in the data.

Each of the included predictors was found to be statistically significant based on t-tests and the overall F-test, reinforcing their relevance in predicting PTS. Notably, variables like **FG%**, **OREB**, and **TOV** exhibited particularly large coefficients, indicating they have substantial influence on point totals. For instance, an increase in **FG%** or **OREB** tends to boost scoring, while higher **TOV** (turnovers) reduces it.

From a real-world perspective, this model suggests that shooting efficiency, rebounding (especially offensive), and maintaining possession (avoiding turnovers) are key drivers of team success in terms of scoring. These insights can be valuable for coaches looking to optimize strategies, as well as analysts assessing player and team performance.

It's also worth noting that multicollinearity was not a concern (VIF values were all below 5) and assumption diagnostics showed no major violations. However, one limitation of the final model is that it was selected using backwards stepwise regression. While this approach is computationally efficient, it does not guarantee identification of the optimal model. Stepwise selection relies on a heuristic process that may overlook better-fitting models involving alternative combinations of predictors. A more comprehensive approach would involve comparing all possible subsets of predictors using information criteria such as AIC or BIC. Although computationally intensive, such analysis provides a more robust evaluation of model fit and selection.

In conclusion, the WLS model with all selected predictors provides a statistically sound and practically useful framework for understanding the factors that most influence NBA scoring outcomes.