“To what extent do key performance criteria and playing time metrics influence an NBA player’s points per game?”

Introduction

Points per game (PTS) is a vital metric in basketball, reflecting a player's scoring ability and offensive contribution. This study investigates how key performance metrics, including minutes played (MP), games started (GS), games played (G), field goal percentage (FG%), assists (AST), and All-Star appearance (Play), influence an NBA player’s scoring performance. The analysis focuses on NBA players from recent seasons (2015–2023), aligning with modern basketball trends.

Previous research supports the relevance of these metrics. Mikołajec et al. (2013) identified that shooting efficiency, assists, and minutes played are critical for determining game outcomes. Casals and Martinez (2013) demonstrated that players with more minutes played and higher usage percentages tend to score more. Similarly, Winston (2014) emphasized efficiency metrics, such as field goal percentage, as key factors in maximizing scoring opportunities. These studies form the foundation for this research, emphasizing the strong link between game-related metrics and player performance. Linear regression is the ideal statistical tool for this analysis, as it quantifies the relationship on average between multiple predictors and a continuous response variable, PTS. This method not only measures the relative importance of each predictor but also provides insight into how these metrics contribute to scoring. Our focus is on predictive power and overall model fit, allowing for actionable insights into player performance.

By identifying the most influential factors on scoring, this study provides value for NBA coaches, analysts, and recruiters. It highlights the interplay of efficiency, playing time, and game involvement in shaping a player’s scoring success, contributing to performance evaluation in the modern NBA.

Methods

Data Preprocessing

The dataset with 1408 rows and 54 columns was imported and subset to include relevant predictors (PTS, MP, GS, G, FG., AST, Play). Missing data rows were removed. Summary statistics were reviewed for initial understanding.

Initial Model and Diagnostics

A linear regression model was fitted using PTS (points per game) as the response variable and predictors (MP, GS, G, FG., AST, Play). Residuals vs. fitted and Q-Q plots were checked for constant variance, normality, linearity, and uncorrelated error violations, which are indicated by fanning patterns, large deviations from the Q-Q line, non linear curves, and clustering.

Transformations and Model Refinement

The response variable was square root-transformed to address violations, and a new model was fitted. Power transformations were applied to predictors (G, GS, MP, FG., AST) based on power estimates. The updated model with transformed predictors and response variables was assessed through residual diagnostics, including residuals vs. fitted and residuals vs. predictors plots. Multicollinearity was checked using correlation matrices and scatterplot pairs for unexpected patterns and clusters as indicators.

ANOVA & Multicollinearity

Anova test was then used followed by the partial t tests for the coefficients to deduce whether any significant linear relationship between at least one of the predictors and the response, as well as determining which specific predictors are of high significance to explain the variance in mean response. These conclusions are made by looking at the p values and t values which should be smaller than the significance level. Moreover, variance inflation factor(VIF) was used to conclude whether severe correlation exists between predictors which can be concluded from VIF values above 5.

Problematic Observations & Automated Selection Methods

To address leverage points, outliers, and other influential observations, appropriate cutoffs associated with each type of points were used to filter the problematic observations using the calculated cutoffs. Finally, automated selection methods were used through AIC, BIC stepwise selection, and all possible subsets methods to conclude which combination of predictors yielded the best model based on corresponding values.

Tools and Techniques

- **Tools**: R programming with libraries like the car for power transformations.
- **Techniques**: Linear regression modelling, residual diagnostics, response and predictor transformations, multivariate correlation analysis, automated selection methods

Results

Residual Analysis: Response Variable

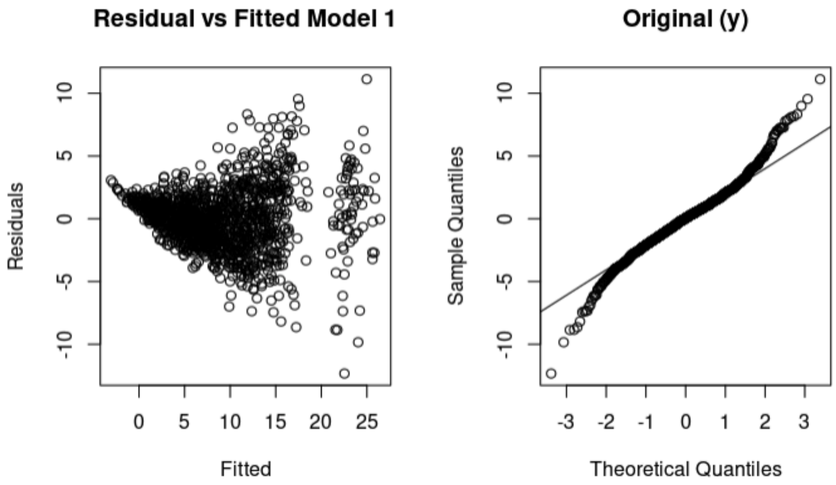

Figure 1

We began by fitting the initial model PTS~MP+GS+G+FG.+AST+Play (Model1). A residual vs. fitted plot for the response variable showed some homoskedasticity, seen in figure 1 above. We decided to apply a square root transformation to the response variable, resulting in **Model 2**. This transformation improved the residual vs. fitted plot, reducing nonlinearity and making residuals appear more evenly distributed. Additionally, we checked the QQ Normal Plot and decided to keep the transformation as it minimized the deviation of points from the QQ line compared to figure 1.

Residual Analysis: Predictors

To further improve the model, we applied power transformations to the predictors. Specifically, GS,G and MP were transformed, as these showed significant improvement in residual vs. predictor plots, while FG% and AST did not and thus were kept the same. A boxplot of residuals by the categorical variable Play (All-Star appearance) revealed no strong evidence of heteroskedasticity related to this factor. This yielded our **Model 3,** square root of response variable and power transformations on GS, G, MP, with FG and AST the same, which after rechecking assumptions minimized potential violations as shown below in figure 2, with an even spread, no clustering or curves, and points closer to the normal qq line.

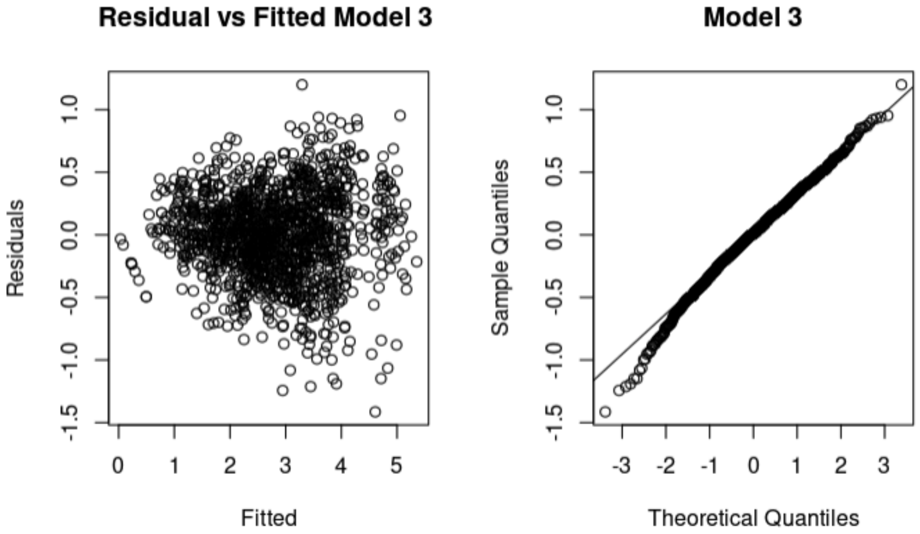

Figure 2

Additional Conditions

We further checked key regression conditions: **Conditional Mean Response with a** plot of sqrt PTS vs. fitted values indicated that the conditional mean response was linear with homoscedastic residuals. **Conditional Mean Predictors,** correlation and pair plots revealed multicollinearity among predictors, particularly for Games Played (G). Based on these diagnostics, we excluded G from the model. Our updated **Model 4** was sqrtPT ~ GS transformed + MG transformed + FG. + Ast + Play

Hypothesis Testing

The ANOVA test showed the p value less than 0.5, shown below in table 3 cell 1, indicating a linear relationship exists between at least one of the predictors, and the t tests for each predictors showed all p values being less than 0.05 and confidence intervals for each predictor not including 0, thus these findings as shown in the table further supported the use and statistical significance of model 4's predictors. Partial F test was used to conduct anova analysis on model 4 with model 3. The p value being less than 0.5 indicated that the predictor G which we excluded for model4 is statistically significant to explain the variance, despite there being obviously violated assumptions in the conditional mean predictors test mentioned above for multicollinearity.

p value = 2.2x10^-16	**Estimate**	**Pr(>|t|)**	**Confidence Interval**
Intercept	0.160620	0.0127	(0.0340667, 0.28693289)
GS_transformed	-0.437680	6.99×10^-12	(-0.56180361, -0.31355639)
MP_transformed	0.200321	2x10^-16	(0.19177058, 0.20887183)
FG.	1.377101	2x10^-16	(1.17273551, 1.58146687)
AST	0.040410	3.37x10^-0.8	(0.02613307, 0.05468774)
Play = Yes	0.647590	2x10^-16	(0.55696568, 0.73821359)

Table 3

Assessing Multicollinearity

Multicollinearity was further assessed using VIF, results showing the full model having one predictors value above 5, opposed to the reduced model. There were conflicting results with the anova analysis yielding significance in the removed G predictor, but issues with multicollinearity exist when including this predictor. As aforementioned, the purpose is predictive power of the model and thus excluding a predictor deemed statistically significant by the partial F-test\would reduce predictive power. The issue we found in the full model was only one predictor having a value of 5.27 which is a marginal increase above 5, but still exists. Thus while this is a limitation that there is high correlation with the MP transformed predictor that may negatively influence variance interpretation, it was concluded that it was small enough to not hold more weight than G being statistically significant. Therefore we moved forward going back to model 3.

Problematic Observations

All types of problematic observations were assessed using the appropriate cutoff values for each type of point, which showed that no points were influential on all values. However, there were a moderate amount of leverage points and outlier points, and a small amount of observations that were influential on coefficients or its own fitted values. These may pose limitations to the model, but are vital to be included because there is a strong probability the leverage points or outliers can be attributed to high level caliber players whose attributes are extraordinary compared to the average player, and the same can be said on the negative scale as well.

Automated Selection Methods

Lastly, automated selection methods were used with the stepwise AIC, BIC, and all possible subsets being used with conflicting results. Both AIC and all possible subsets indicated the best model being model 3 with all predictors(including G), while BIC indicated the best model would be model 4 which omitted G. Ultimately, given the reasons earlier for the anova test coupled with the fact that stepwise AIC and all possible subsets supported the use of model3(full model), it was concluded to use model3 as the final model, which is: sqrtPT ~ G transformed + GS transformed + MG transformed + FG. + Ast + Play.

Conclusions & Limitations

Conclusion:

This study explored the relationship between key performance metrics and Points per Game (PTS) using a linear regression model. The results explicitly answer the research question by identifying which metrics significantly influence scoring performance. Notably, the coefficient for *MP_transformed* (minutes played) is 0.204, indicating that a one-unit increase in playing time corresponds to a 0.204-unit increase in the square root of PTS, holding all other variables constant. This aligns with prior studies, such as Casals and Martinez (2013), which emphasize the importance of increased playing time on scoring.

The coefficient for *FG.* (field goal percentage) is 1.408, demonstrating that shooting efficiency is one of the strongest predictors of scoring performance. This finding is consistent with Winston's (2014) assertion that field goal percentage maximizes scoring opportunities. Additionally, the coefficient for *AST* (assists) is 0.039, highlighting the positive impact of playmaking on a player's scoring ability, as previously noted by Mikołajec et al. (2013). Lastly, the indicator variable *PlayYes* (All-Star appearance) has a coefficient of 0.645, suggesting that All-Star players score significantly more, likely due to their elite skills and greater role in offensive schemes.

However, the negative coefficients for *G_transformed* (-0.0004) and *GS_transformed* (-0.413) suggest that consistently starting or participating in more games may inversely relate to scoring, possibly reflecting the fatigue or reduced role specialization of regular starters.

Limitations:

1. **Extreme Observations:** Outliers such as high-scoring stars (e.g., MVP-level players) or low-scoring bench players could influence the model. Their impact was retained to preserve data integrity but may skew results.
2. **Violated Assumptions:** Residual plots indicate potential heteroscedasticity, which may affect model validity.
3. **Multicollinearity:** High correlations among variables like G, GS and MP complicate the interpretation of individual coefficients, despite variance inflation factor (VIF) adjustments.
4. **Data Scope:** The dataset spans 2015–2023, limiting generalizability to other basketball eras or styles.

The findings provide robust insights into player scoring performance while aligning with existing literature. Addressing these limitations in future studies could enhance model accuracy and broader applicability.

Ethics Discussion

In terms of automated selection and manual selection methods, we chose to value both methods when ultimately deciding on a model to fit. What made this decision a bit easier for us than some other cases was that both methods actually produced similar models as mentioned above in the results section. Using manual methods, the ANOVA test and partial f test indicated that model 3 using all predictors instead of removing G was the best model to fit, despite tests for multicollinearity having one coefficient slightly above 5. When using AIC stepwise selection and all possible subsets method, they agreed with these findings and chose to include all predictors just like model 3. However, we were aware that AIC and BIC produced conflicting models, and that in the presence of model violations and multicollinearity, results will still be produced. Thus we ensured to assess and correct model assumptions and multicollinearity as much as possible, and we also ensured to use manual selection methods because they use context that are missed in automated methods, to check for conflicting results. In the ethics module we discussed using a calculator to calculate a complex equation over calculating by hand. While the calculator is easier, there are chances it could malfunction or inputs may not be in correct order. Looking at automated methods as the calculator and manual methods as by hand calculations, we ensured to use both to have full perspective, and luckily both methods yielded similar results which strengthened our model choice.

Contributions:

Nicole: methods, ethics, results
Domenica: Introduction, Results, Conclusion & Limitations
Josh: Methods, Results, Ethics Discussion
All: R code, editing, poster

Bibliography

Mikołajec, K., Maszczyk, A., & Zając, T. (2013). Game indicators determining sports performance in the NBA. Journal of Human Kinetics, 37(1), 145-151. https://doi.org/10.2478/hukin-2013-0035

Sampaio, J., Gonçalves, D., Rentero, T. M., & Ribeiro, N. (2013). Modelling player performance in basketball through mixed models. International Journal of Performance Analysis in Sport, 13(1), 64-82

Winston, W. L. (2014). Mathletics: How gamblers, managers, and sports enthusiasts use mathematics in baseball, basketball, and football. Princeton University Press.