STA302 Fall 2023 Methods of Data Analysis 1
Final Project (Part 3)

Word Count: 1997

Names of Group Members	Contribution to Proposal
Alex Lugard	Introduction, Methods, Ethics Discussion
James Li	Discussion, R Code
Tristan Gauntley	Results, R Code, Editing
Ziana Suleman	Methods, Results, R Code, Editing

INTRODUCTION

The purpose of this report is to explore the relationship between several predictor/categorical variables and a response variable using STA302 content. Specifically, we aim to answer the following research question: *What effect do Number of Games Started, Turnovers per Game, Field Goals Attempted per Game, Age, and a player's Primary Position have on an NBA player's average Points per Game?* In this report, Primary Position is a categorical variable. While inspiration was taken from three sources of background literature (Kalén [1], Papadaki [2], Page [3]), our analysis greatly differs from each of these as they only explore the relationship between one or two predictor variables and the response variable, or analyze data using different forms of regression such as Gaussian Process or non-linear regression. We aim to answer the research question by obtaining coefficients to the predictor and categorical variables. These coefficients will help deduce how PPG is affected by different predictors, which are measurable statistics that we hypothesized would have the greatest impact on PPG.

METHODS

The dataset found on Kaggle contained many more variables than what was necessary for this project. Based on the relevance of certain variables over others, we converged to 5 predictor variables (Number of Games Started, Turnovers per Game, Field Goals Attempted per Game, Age, and a player's Primary Position), and one response variable (average Points per Game). We created a linear regression model to visualize the strength of the relationship between each individual predictor variable and our response variable. The original purpose of the dataset, explained by the author, was to create a model that predicted the score of each NBA player in the All-Star Game. Our research question differs because we are observing how different predictors will impact the average PPG of a player.

A preliminary multiple linear regression model was created, where we derived initial relationships between the predictor and response variables. Additionally, several plots were produced, including residual, response, QQ, and pairwise plots. From these graphs, we were able to identify violated assumptions (Constant Variance and Normality) from visual clues (fanning pattern, data trailing off from the linear pattern at smaller and larger quantiles).

Following this model, a flow chart was created to mitigate any significant errors that violated our assumptions. The first step in our flow chart is to use a combination of Box Cox transformations and variance stabilizing transformations to correct the violations. After each transformation is applied, we test the model against conditions 1 and 2. If these conditions don't hold, we cannot trust the residual plots and we must use a model where they do. If these conditions hold, we can trust our residuals to check violated assumptions and we use our transformations again to correct any violations. We stop applying transformations once we have a model that corrects all of our violated assumptions.

Using R's summary output, we conclude the ANOVA test by looking at the p value: if this is less than the critical value, we have a significant linear relationship for at least one predictor. If Pr(>|t|) from the R summary is greater than the critical value, we remove the predictor associated with the value because it is insignificant. We then conduct a partial F test using our reduced model and our full model. If the p value is greater than the critical value then we can confidently use our reduced model as our preferred model. We check our reduced model to see if conditions 1 and 2 hold and apply any necessary transformations for violated assumptions.

Based on our conclusions from the partial F test we either use the full model or the reduced model in the following steps. We check the VIF for all predictors and if any of the values are greater than 5 we have serious multicollinearity and we need to remove predictors with the highest multicollinearity and recheck the VIF. Once our VIF values are less than 5, we can check for problematic observations by calculating each associated measure and seeing which observations do not meet the cutoff. If we have specific observations that are in all of our categories of problematic points then we will discuss if removing them will change our research question and if it does then we won't remove them.

To ensure we are checking all models, we use the method of finding all possible subsets to find the best SLR model to the best 5 predictor model. Based on the models that include our significant predictors, we compare the models to find the lowest AIC, BIC, and AICc, and highest R^2 values. If the best model is not our reduced model, we recheck our assumptions, conduct the partial F test, find our variance inflation factor, and identify any problematic observations. If our assumptions are not violated and our VIF is less than 5, we can choose this as our best model. If there are some violations, this will not be the best model and we may want to look at one of our other models from the all possible subsets method that includes our significant predictors.

RESULTS

Our original model is presented as:

$$PPG = -0.178819 + 0.010921(AGE) + 0.366318(TOV) + 0.004283(GS) + 1.241(FGA) \\ -0.645645\,[I(PF)] - 1.142013\,[I(PG)] - 0.773393\,[I(SF)] - 1.082634[I(SG)]$$

The predictor variables chosen from the larger dataset were the ones most aligned with our research question. Number of games started serves as an indicator of a player's importance and is preferred over minutes played. Turnovers were selected over steals for their direct impact on scoring chances. Field goal attempts were favoured over field goal percentage as they have a more significant influence on points per game. Age was chosen as a predictor variable based on skill development trends. Player position (categorical 1-5) reflects diverse player roles and influences points per game more directly than team affiliation. These choices optimize the analysis for a comprehensive understanding of player contributions, and allow for the best analysis of our research question.

From the residual plots associated with the original model [A2], clear violations of Normality and Constant Variance are notable. Thus, appropriate transformations need to be performed to arrive at a more appropriate model. Based on the structure of our flow chart, we used a Box Cox transformation on our response variable and transformed this variable based on the maximum lambda in our Profile Log-likelihood graph (λ=0.8). After applying this transformation, we created a new model which included our transformed responses variable and checked our new model to identify violated assumptions. Condition 1 did not hold for this new model, meaning that our residual plots were unreliable (Figure 1).

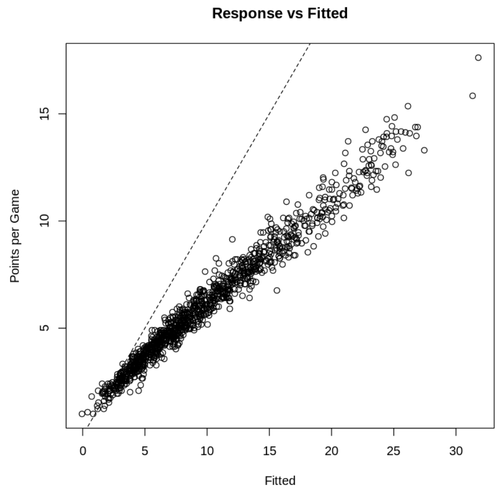


Figure 1: Response vs. Fitted plot, with stark deviation from the plotted line, indicative of a violation

We decided to apply a power transformation to the Turnovers and Field Goals Attempted predictors. Using the R output, we transformed the predictors to: Field Goals Attempted^⅓ and Turnovers^¼ . We created a new model called boxCoxXY_model with the transformed responses and predictors. Conditions 1 and 2 hold, but our new model had quite a few violated assumptions in the residuals plots for both the response and the predictors. Our next step was to plot a histogram for the transformed response where we observed a skew (Figure 2).

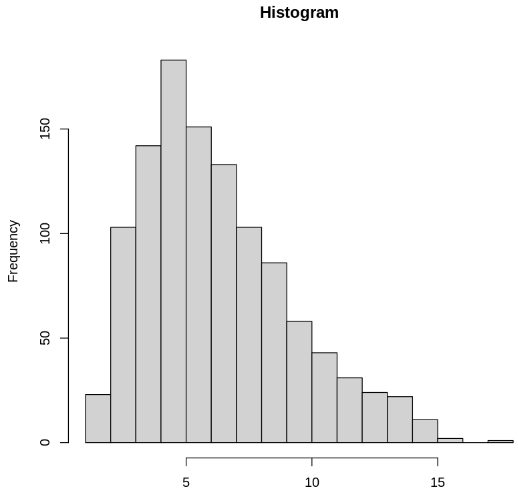


Figure 2: A histogram for Points per Game, with evidence of a right skew

Building off of our boxCoxXY_model, we applied a square root transformation to our previously transformed response. In this new model (sqrt_model) conditions 1 and 2 held and no assumptions were violated, rendering our transformations successful in correcting the Constant Variance and Normality violations.

After analyzing the new model's summary, a significant linear relationship for at least one predictor was evident, due to a small p value of <2.2e-16. Using the hypothesis tests, Pr(>|t|), we were able to remove our age predictor because the value of Pr(>|t|) for age was 0.17279 which is greater than our critical value of 0.05. We created a reduced model by removing the age predictor. We then conducted a partial F-test between our new model and the reduced version of our new model. Using the F-statistic = 1.86 and Pr(>F)= 0.17, we see that since our p-value was greater than our critical value of 0.05, we can confidently remove the age predictor.

Figure 3 shows that our reduced model was checked and since there were not any violated assumptions present, we were able to conclude that this could be the best model.

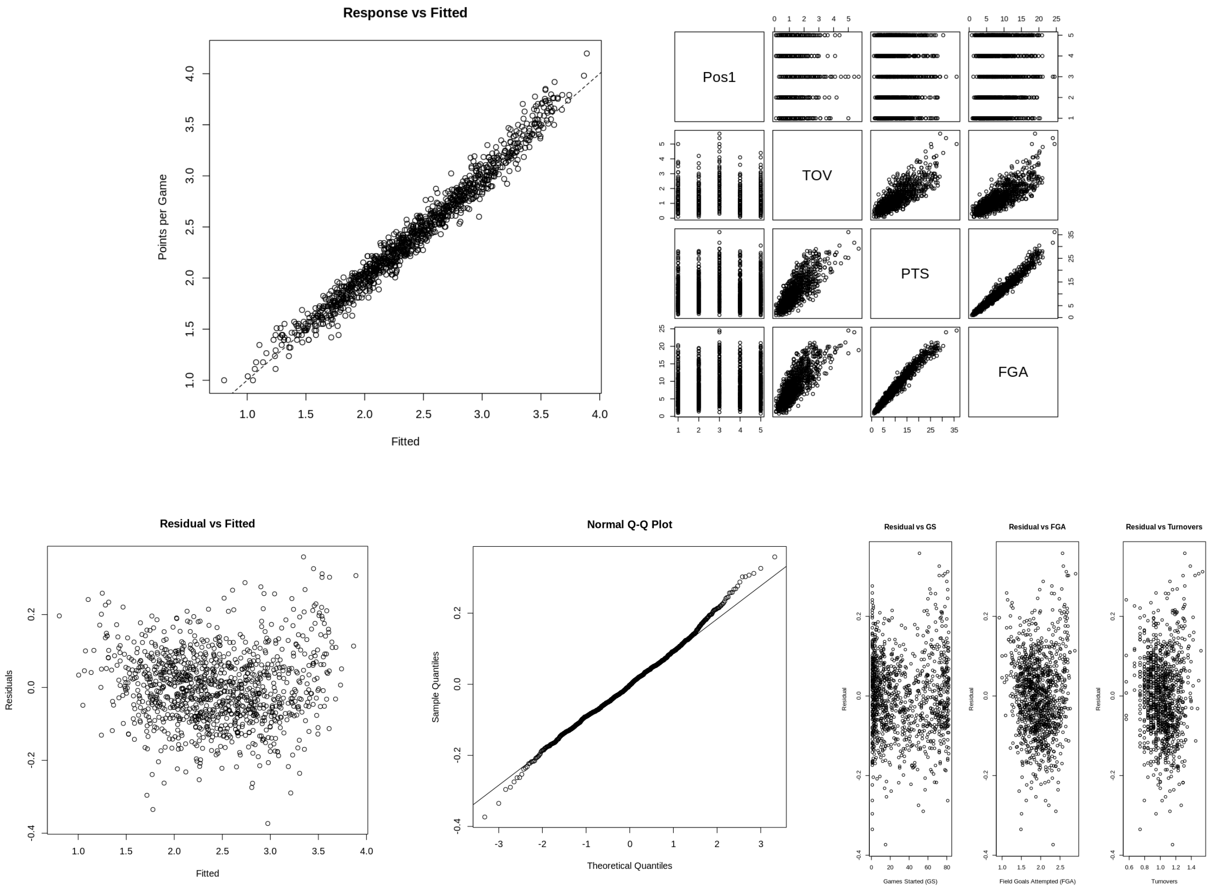


Figure 3: Conditions 1 and 2 are met as there is not an easily identifiable non-linear trend or pattern which means our residuals will be reliable for this model. There is random scatter with no visible fanning pattern, systematic pattern, or large clusters of points and our normal Q-Q plot has minimal deviations from the diagonal line

To determine if our reduced model is the best, we checked the variance inflation factor and since all VIF values are less than 5, we conclude that we don't have serious multicollinearity present between all predictors.

We checked our model for any problematic observations by identifying leverage points, outlier points, and influential points. Based on the criteria given in class, we output the problematic observations in R and since there weren't any observations that fall under each category of problematic observations, we conclude no observations need to be removed.

Finally, to ensure that we have chosen the best model, we use the process of finding all possible subsets. We use this method on our sqrt_model as it does not violate any assumptions.

After analyzing our models, taking our research question into account, as well as our previous analysis, we did not want to consider any model that removed predictors besides the age predictor because the rest of our predictors are statistically significant to our response. This means that they will be relevant in addressing our research question. This left us to compare the 5 predictor and 4 predictor model, which we compared earlier using the F-partial test. We further compare these models by summarizing their AIC, BIC, AICc, and R^2 values.

Table 1: Summary of the AIC, BIC, AICc, and R^2 values for the best 4 and 5 predictor models

Model	R^2	AIC	BIC	AICc
4 predictor model	0.968	-5056.613	-5031.525	-5056.537
5 predictor model	0.968	-5056.203	-5026.098	-5056.103

Table 1 shows that the AIC, BIC, AICc values are lower for the best 4 predictor model and the R^2 is the same for the best 4 predictor model. Using this information as well as the information from the F-partial test, test for multicollinearity, and problematic observations, we conclude our 4 predictor model is our best model.

Table 2: A summary of the summary statistics of our final model

Variables	Numerical Summary			
	Estimate	Std. Error	T value	Pr(>|t|)
Intercept	-0.5456784	0.0242735	-22.480	< 2e-16
Games Started	0.0005988	0.0001486	4.030	5.96e-05
Turnovers	0.1297082	0.0359164	3.611	0.000318
Field Goal Attempts	1.4957174	0.0165374	90.444	< 2e-16
Position - Power Forward	-0.0962740	0.0095144	-10.119	< 2e-16
Position - Point Guard	-0.1486165	0.0098348	-15.111	< 2e-16
Position - Small Forward	-0.1113649	0.0099837	-11.155	< 2e-16
Position - Shooting Guard	-0.1359615	0.0095821	-14.189	< 2e-16
RSS	0.09854			
R^2	0.9712			
R^2_{adj}	0.971			
F- Statistic	5337			
p-value	< 2.2e-16			

DISCUSSION

After applying all necessary transformations and conducting important tests on the regression model, we have obtained our finalized model that can best relate how points per game (PPG) is affected by different predictors, which are measurable statistics that we hypothesized would have the greatest impact on PPG. The final model follows, where the I(...) predictors are indicator functions representing each of the 5 positions as an individual predictor, with the center position as the intercept.

$$\sqrt{(PPG)^{\frac{4}{5}}} = -0.5456784 + 0.129708(TOV^{\frac{1}{4}}) + 1.495714(FGA^{\frac{1}{3}}) + 0.0005988(GS) \\ - 0.0962740(I(PF) - 0.1486165(I(PG)) - 0.1113649(I(SF)) - 0.1359615(I(SG))$$

From the final model, we note that there is a significant correlation between the selected predictors and the measured response, with the predictor that contributes most to the points scored per game is the number of field goals taken, which is an expected and measurable outcome; generally, if a player is taking more field goals in a game, their total points scored will increase. Interestingly, the categorical variables each have a negative coefficient, which means relative to our reference predictor (the center position), every other position on average scores less points than players at the center position. Lastly, the turnover predictor having a positive correlation with points scored was unexpected since we had originally expected that an increase in turnovers would lead to a lower point total, yet our model expects the contrary. In addition, studies conducted by NBA analysts using simple machine learning models have found that one of the statistics most correlated to a player's points scored is their field goals attempted season average [4][5].

The limitation that we encountered when developing the final model includes the prohibition of zero entries in our dataset, especially when applying transformations to the response and predictors. This limitation was relieved simply by removing players with a "0" value, which only removed around 10% of the players in the dataset. This decision is justified by noting that these players removed would not have had enough production in the NBA to have significant contribution to what we aim to measure in our research question. Another limitation lies in the size of the dataset, only capturing players from 2016-2019, which can lead to biases towards the current era's basketball style and does not capture exactly how basketball was played over the entire existence of the NBA.

ETHICS

There were several reasons as to why we chose to implement a manual selection tool. The primary reason was because it was recommended for a model with few predictors. Despite its simplicity, using the hypothesis test, F partial test, ANOVA test, as well as a comparison of all possible subsets, we were able to get a proper overview of the best model to choose. Secondly, we refrained from using automated selection as it would have resulted in only an idea of the preferred model and not necessarily a concrete answer for the best model. Furthermore, the methods of automated selection will not always agree on the preferred model.The last reason was for simplicity purposes. The automated selection tool does not take into account the transformations already applied to the data set, and will run even in the presence of issues/violations. This implies we would maybe be retransforming a model that may not have been the best one to begin with.

The methods are indeed not ethically the same. It is clear from lecture content that the automated selection tool is ignorant and does not consider any context of the data or question in its decision making process. Furthermore, it actually created bias in the model as it searches for data based on significance rather than collecting data to test a specific hypothesis.

WORKS CITED

[1] Anton Kalén, Alexandra Pérez-Ferreirós, Pablo B. Costa & Ezequiel Rey (2021). “Effects of age on physical and technical performance in National Basketball Association (NBA) players”DOI: 10.1080/15438627.2020.1809411

[2] Papadaki, Ioanna & Tsagris, Michail. (2022). “Are NBA Players’ Salaries in Accordance with Their Performance on Court?”. 10.1007/978-3-030-85254-2_25. https://www.researchgate.net/publication/357883185_Are_NBA_Players%27_Salaries_in_Accordance_with_Their_Performance_on_Court

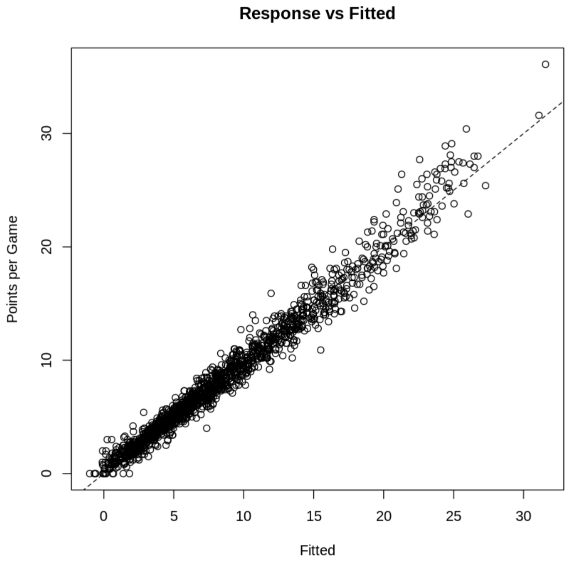
[3] Page, Garritt & Barney, Bradley & McGuire, Aaron. (2013). “Effect of position, usage rate, and per game minutes played on NBA player production curves”. Journal of Quantitative Analysis in Sports. 1-9. 10.1515/jqas-2012-0023.

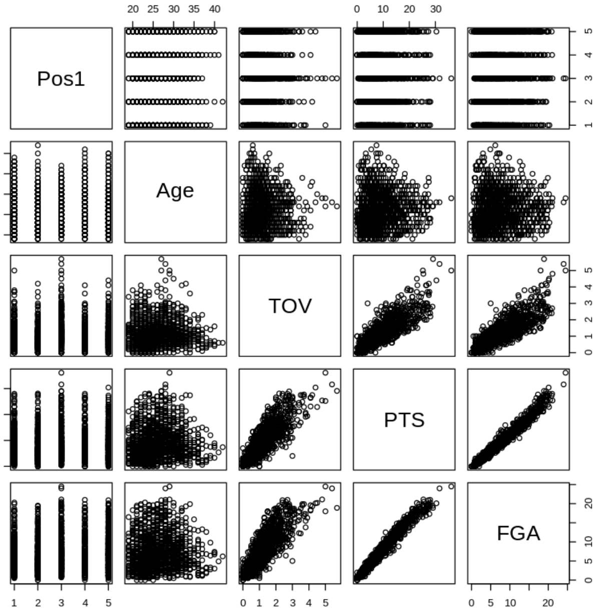
[4] T. Zovak, A. Šarčević, M. Vranić and D. Pintar, "Game-to-Game Prediction of NBA Players’ Points in Relation to Their Season Average," 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2019, pp. 1266-1270, doi: 10.23919/MIPRO.2019.8756733.

[5] Kevin Wheeler, Predicting NBA Player Performance, https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=c63034d4918b1cef04658d8250a2d0c96add2689 [4.12.2018.]

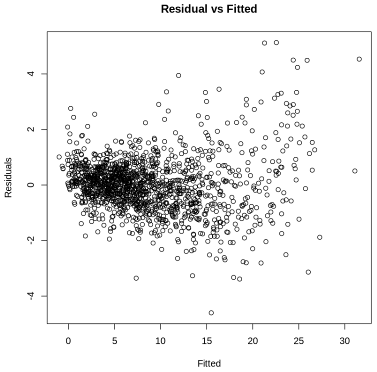
APPENDIX

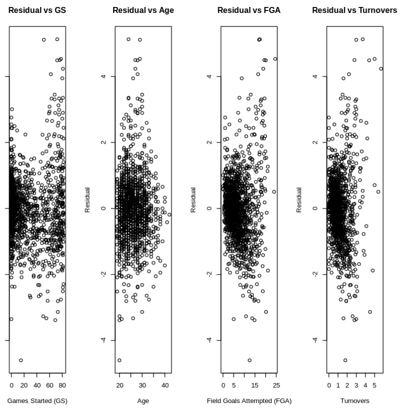
[A1] Conditions 1 and 2 from Original Model

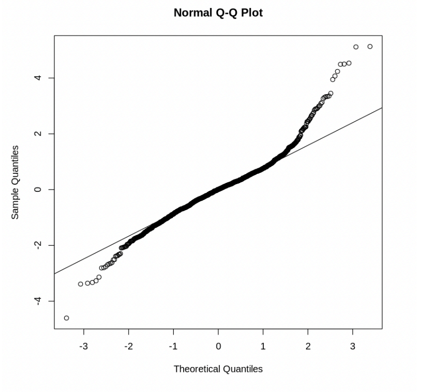




[A2] Residuals vs. fitted for response and predictors and Normal Q-Q Plot from Original Model


Residual vs Fitted
Residuals
Fitted


Residual vs GS
Residual vs Age
Residual vs FGA
Residual vs Turnovers
Residual
Games Started (GS)
Age
Field Goals Attempted (FGA)
Turnovers


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles