Final Project Part 3 - Predicting NBA2K22 Ratings

Hongyuan (Steven) Liu, 1007077976

1. Introduction

NBA2K series is one of the most trending basketball games worldwide. Besides the game itself being fascinating, the NBA2K rating has also become the most objective and authorized measure of NBA players' ability. This study aims to build a linear regression model to determine whether we can predict one's newest NBA2K rating based on his current NBA season stats (per game), including minutes, points, rebounds, blocks, rebounds, assists, fouls, and turnover. Currently, the NBA2K ratings are manually assigned based on people's impressions of players. By building such a model, we can obtain a statistically significant prediction to minimize bias while assigning the ratings. Besides, these predictions are useful for audiences to get a better understanding of players' newest ability score; and as a reference in awards (e.g. MVP, MIP) selections. Some studies have shown that MVP winners are more likely to have more points per minute, points per game, and field goal percentage (*Chen*). As an extension, our model wishes to explore more pattern in data that reflects players' overall ability.

2. Methods

2.1 variable selections

According to the Introduction, the response variable is the NBA2K rating, and the initial predictors consist of the player's points, rebounds, blocks, assists, fouls, and turnover[1]. An full model will be built based on the response variable and all of the initial predictors using linear regression.

Based on the initial model, we will build new models by reducing variables. We will propose to eject any variables that are insignificant in the full models (with p-value> 0.05); and use a Partial ANOVA F-test ($\alpha = 0.05$) to determine whether this ejection is eligible. If eligible, the reduced model will be a candidate for our final model. We can repeat this process to keep reducing variables if any new variables in the reduced model seem insignificant.

Once we have obtained several candidates, we should decide which model works best. The principle here is, we want to have a model that has fewer predictors and predicts well. To assess the wellness of predictions, we will choose the models with the highest adjusted R^2 with the least AIC and BIC. If several candidates are similarly good, we can choose the most validated model as the best model (Section 2.2).

[1] All statistics are measured per game.

2.2 Model Validation

To test the model validation, we will randomly split the dataset into a training dataset and a testing dataset, with a ratio of 70/30; and only use the training dataset during variable selection. When testing each model's validation, we will fit another linear regression model with exactly the same predictors but using the testing dataset. A model is validated if both the training model and testing model have similar coefficients, adjusted R^2, numbers of problematic points, and the same number of predictors.

2.3 Model Violations and Diagnostics

For each model we fitted, we want to verify whether it satisfies the two conditions and the four assumptions of linear regression models. Condition 1 is satisfied when points are randomly scattered around an identity function on the response versus the fitted values plot. Condition 2 is satisfied if there is no evidence of a non-linear relationship between predictors in the predictors' pair plots. Whenever any conditions fail, we should consider applying transformation on predictors or reselecting variables from the original dataset.

The assumptions are checked by looking at the residuals against predictors plots, the residuals against fitted values plot, and a normal QQ-Plot. Linearity will be satisfied if there is no systematic pattern, uncorrelated errors will be satisfied if there are no clusters of residuals, and the constant variance will be satisfied if there is no pattern such as a fanning pattern. Lastly, We will verify the normality by checking whether the QQ-Plot contains a straight diagonal line with minimal deviations at the end. If any of these assumptions are violated, we should consider applying transformation on response or predictors. Box-Cox transformation is the tool we used to decide the transformation.

Besides, our model should avoid multi-colinearity. We will check the VIF table of models and attempt to exclude the variables with VIF values greater than 5. A Partial ANOVA F-test ($\alpha = 0.05$) is used during the model reduction. We should consider respecifying the model if a variable is both significant and highly muli-linearly.

Once all the steps above are processed, we will check the existence of problematic points. We will identify leverage points, and outliers by their cutoffs[2], and identify influential points by looking at the cook's distance and DFFITS values. We will not modify these problematic points, but we will analyze their existence. We could expect a better model with fewer problematic points.

[2] We will use [-2,2] as the cutoff for outliers since the dataset is small.

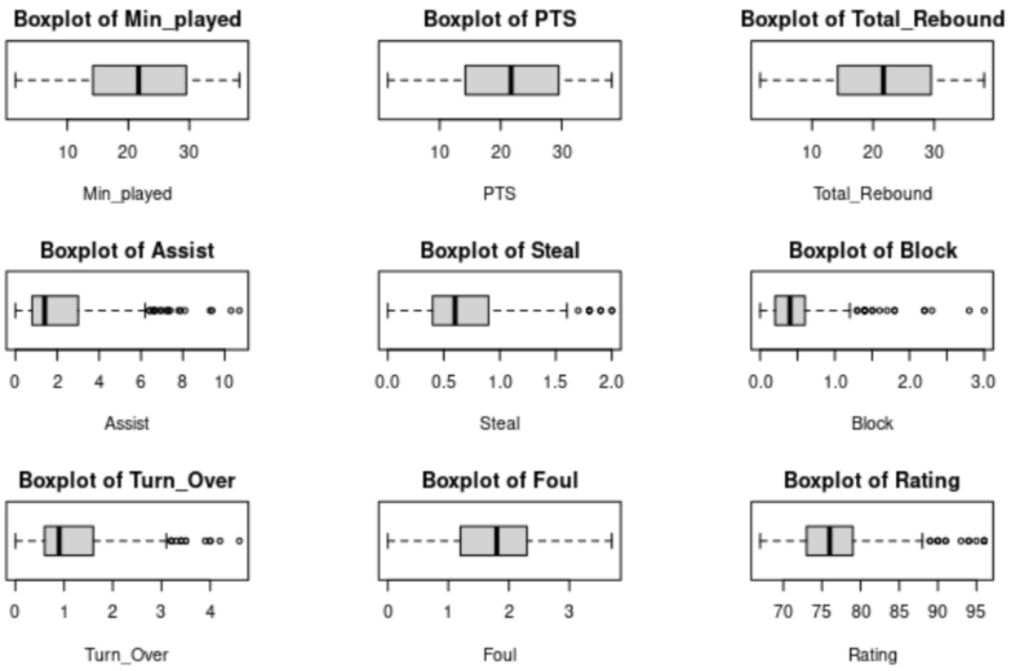
3. Result

3.1 Description of Data

Variable	mean (s.d.)	Correlation with response
Min_played	21.607 (9.255)	0.747
PTS	9.835 (6.478)	0.848
Total_Rebound	3.981 (2.348)	0.641
Assist	2.18 (1.946)	0.699
Steal	0.679 (0.407)	0.559
Block	0.454 (0.419)	0.347
Turn_Over	1.19 (0.864)	0.773
Foul	1.764 (0.737)	0.509
Rating	76.459 (5.483)	NA

Table 1 - Variables in the dataset and their mean, standard error, and correlation with response

Our data comes from two datasets, the first dataset contains 594 NBA players' stats across 30 variables in the regular season 2021-2022 (*David*), and the second dataset contains 464 NBA players' names and NBA2K22 ratings (*HoopsHype*). After merging two datasets and selecting desired variables, we obtained 377 observations over 9 variables.


Boxplot of Min_played

Boxplot of PTS

Boxplot of Total_Rebound

Boxplot of Assist

Boxplot of Steal

Boxplot of Block

Boxplot of Turn_Over

Boxplot of Foul

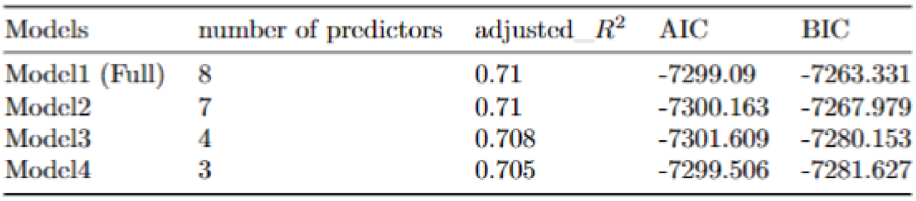
Boxplot of Rating

Figure 1 - Boxplots of all variables

From Figure 1, we notice that most predictors are right-skewed, which may lead to a violation of linearity. Response (Rating) seems normally distributed, where normality should be satisfied.

3.2 Presenting the Analysis Process and the Results

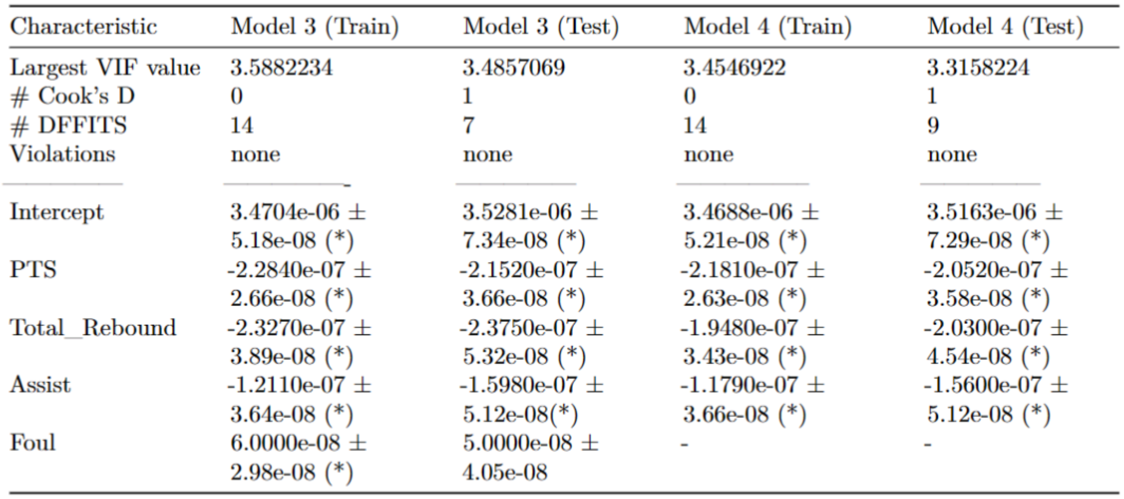
Initially, we fit the full model using all variables in Table 1 without applying any transformation. The clusters and fanning pattern in the residuals plots (*Appendix 1*) indicate a transformation is necessary. Referring to the Box-cox transformation power, we decide to apply a power of -3 transformation to Rating (Box-cox power -3.34), and a square root transformation to PTS, Total_Rebound, Assist, Steal, Block, and Turn_Over (All with Box-cox power around 0.5). We have fit model 1 using all the transformed variables. Applying partial ANOVA F-test, we have successfully excluded Min_Played (p-value=0.3443) in model 2; and further excluded Steal, Block, and Turn_Over (p-value=0.2564) in model 3; and further excluded Foul in model 4 (p-value=0.0973).



Models	number of predictors	adjusted_R^2	AIC	BIC
Model1 (Full)	8	0.71	-7299.09	-7263.331
Model2	7	0.71	-7300.163	-7267.979
Model3	4	0.708	-7301.609	-7280.153
Model4	3	0.705	-7299.506	-7281.627

Table 2 - Model Comparison with Number of Predictors, adjusted R², AIC, and BIC

Notice that Model 3 has the smallest AIC and Model 4 has the smallest BIC, and both models contain fewer predictors than Model 1&2.



Characteristic	Model 3 (Train)	Model 3 (Test)	Model 4 (Train)	Model 4 (Test)
Largest VIF value	3.5882234	3.4857069	3.4546922	3.3158224
# Cook's D	0	1	0	1
# DFFITS	14	7	14	9
Violations	none	none	none	none
Intercept	3.4704e-06 ± 5.18e-08 (*)	3.5281e-06 ± 7.34e-08 (*)	3.4688e-06 ± 5.21e-08 (*)	3.5163e-06 ± 7.29e-08 (*)
PTS	-2.2840e-07 ± 2.66e-08 (*)	-2.1520e-07 ± 3.66e-08 (*)	-2.1810e-07 ± 2.63e-08 (*)	-2.0520e-07 ± 3.58e-08 (*)
Total_Rebound	-2.3270e-07 ± 3.89e-08 (*)	-2.3750e-07 ± 5.32e-08 (*)	-1.9480e-07 ± 3.43e-08 (*)	-2.0300e-07 ± 4.54e-08 (*)
Assist	-1.2110e-07 ± 3.64e-08 (*)	-1.5980e-07 ± 5.12e-08(*)	-1.1790e-07 ± 3.66e-08 (*)	-1.5600e-07 ± 5.12e-08 (*)
Foul	6.0000e-08 ± 2.98e-08 (*)	5.0000e-08 ± 4.05e-08	-	-

Table 3 - Model Comparison, Model Validations, and the summary of Model 3 and 4

When validating these two models, we found that the VIF, Cook's Distance, and DFFITS for Model 4 are more consistent between training and testing datasets. Besides, Model 4 contains the same number of significant predictors. Therefore, Model 4 will be the final model.

3.3 Goodness of the Final Model

As discussed in section 3.2, the model initially violates the assumptions of the linear regression model, so we have applied proper transformation on both predictors and the response. Therefore, we will now be checking the conditions and assumptions for the final model (with variables after transformation).

From Figure 2, we can verify both conditions by saying points are scattered around an identity function in the response versus fitted values plot, and no evidence of a non-linear relationship between predictors (*Appendix 2*). Normality is also satisfied as a straight diagonal line appears in the Normal QQ-Plot. Besides, there is no evidence of a systematic pattern, clusters of residuals, and fanning pattern in the residuals versus fitted values/predictors plots indicating linearity, uncorrelated errors, and constant variance are satisfied. Thus, the final model satisfies all the conditions and assumptions of linear regression.

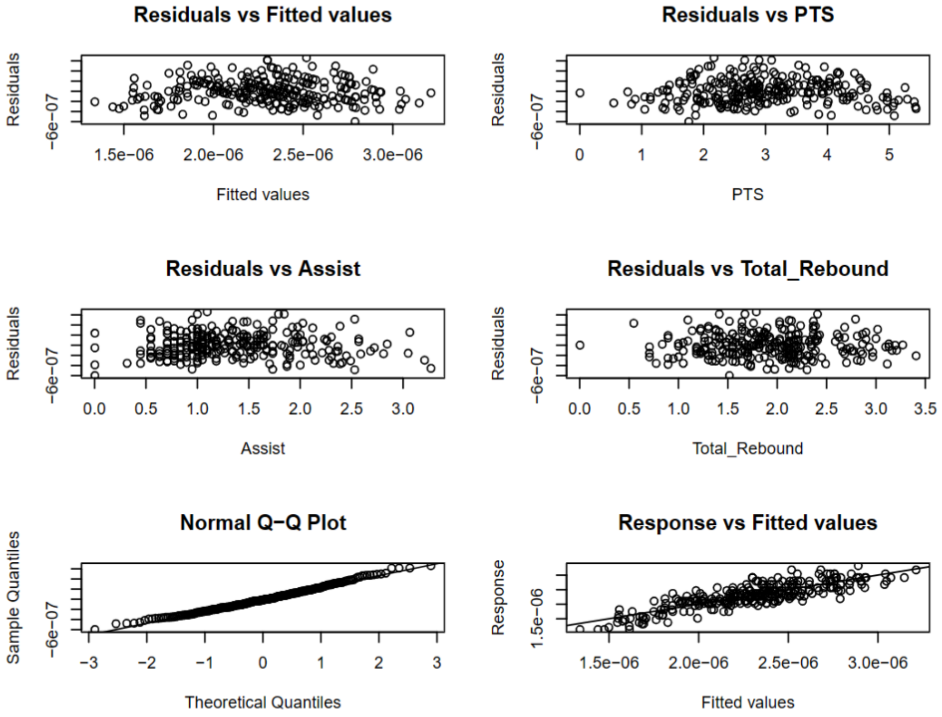


Figure 2 - Residuals vs. Fitted Values/Predictors Plots, Normal QQ-Plot, and Response vs. Fitted values for the Model 4

When validating Model 4, the adjusted R^2 are 0.71 and 0.75 for training and testing dataset, which appears to be similar. Furthermore, as discussed in section 3.2 and Table 3, we can conclude Model 4 is validated as both training and testing have similar performance.

4. Discussion

4.1 Final Model Interpretation and Importance

The coefficients of the final model (Model 4 Train) are summarized in Table 3. The coefficients tell us an increase in the square root of PTS or Total_Rebound will decrease the value of $\frac{1}{Rating^3}$ by approximately 2.0e-7. That being said, an increase in PTS and Total_Rebound will approximately increase the rating by the same factor. The coefficient of the square root of Assist is approximately 1.0e-7, which implies a higher assist will also lead to a higher rating. Overall, the model illustrates PTS and Total_Rebound are the tied top significant variables that affect NBA2K rating, and the assist is another factor but less significant. These three variables are all positively correlated to the response. As an answer to the research question, our model indicates we can predict the NBA2K ratings based on PTS, Total_Rebound, and assist.

4.2 Limitations of the Analysis

There are several limitations to the analysis. From Appendix 2, we can tell that many variables are correlated with each other. This will increases the bias of the model. One reason this is not the main concern is we do not have too many predictors in the final model. Another limitation is, the rating is correlated with some other variables such as age with a non-parametric relationship, which can not be specified using linear regression. These variables are not selected during the EDA. So we are taking risk of misspecifying the model, but the model still provides us with useful interpretation among other predictors. Lastly, the number of influential points is relevantly large compared with the size of the dataset. Some parameters might be highly affected by these points, which makes it difficult to validate models. We are not supposed to remove these points because of their special meanings, so when splitting the dataset, we have split these influential points evenly to minimize this effect.

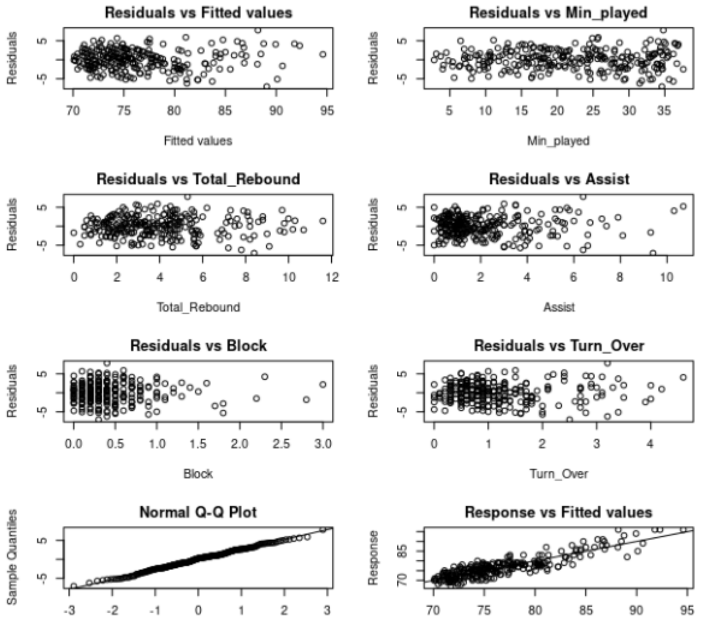
5. Bibliography

1. *Chen. (October 25-26, 2017). Predict NBA Regular Season MVP Winner. ieomsociety.org. from https://ieomsociety.org*

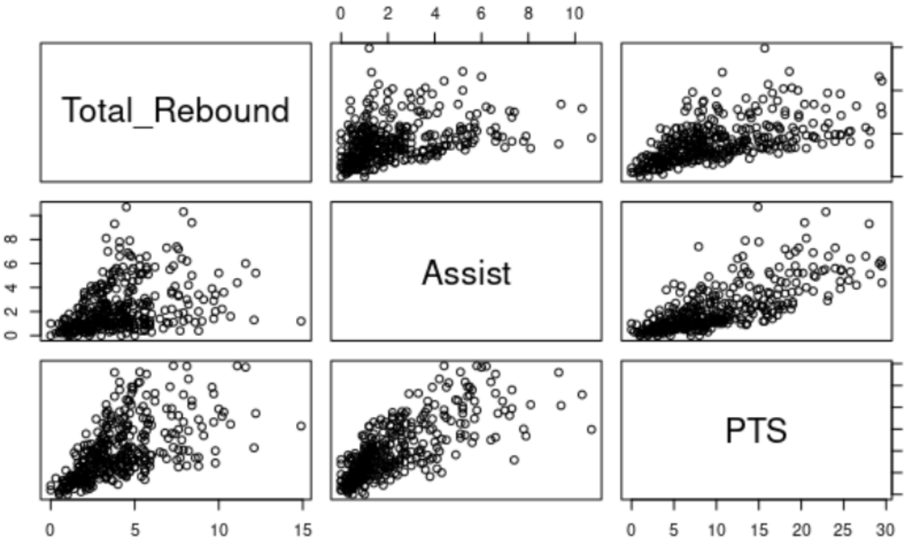
2. *David. (2022, March 8). NBA-stats-2021-2022/2021-2022 NBA player Stats.csv at main · 119502173David/NBA-stats-2021-2022. GitHub. Retrieved October 20, 2022, https://github.com/119502173David/nba-stats-2021-2022/blob/main/2021-2022%20NBA%20Player%20Stats.csv*

3. *HoopsHype. (n.d.). These are the NBA 2K22 ratings of all NBA players. HoopsHype. Retrieved October 20, 2022, from https://hoopshype.com/nba2k/2021-2022/*

6. Appendix



Appendix 1 - *Residuals vs. Fitted Values/Predictors Plots, Normal QQ-Plot, and Response vs. Fitted values of Full Model before transformation*



Appendix 2 - Pair plots for predictors before transformation