Predicting NBA Win Percentages With Linear Regression

Eric Liu

December 17, 2021

Introduction

The question of research in this report is to determine what factors attribute to the winning games in the NBA. The goal of this research is to explore new possible combinations of factors that could contribute to determining how games are won. The significant outcome of this report, if there is one, will hope to provide the most definitive answer on which factor or factors attributes most towards winning. A blog (Kotzias, 2018) referenced an interesting insight dubbed called the "Four Factors of Basketball Success". These four factors simply weights the effectiveness of each contributor which are, Shooting (40%), Turnover Rate (25%), Offensive Rebound Rate (20%), and Free Throw Rate (15%). The plan for this report will be to keep these kind variables in mind and perhaps find new factors that contribute to winning.

Methods

Variable Selection

When choosing variables for a model we generally want to choose variables that are linearly significant to the response. In other words, we want variables that will yield the most significant p-values. Moreover, it also has to have significant AIC, BIC, and adjusted R-squared values as well. There are numerous ways to choose variables and one could even arbitrarily pick variables that turn out to be significant in all areas, but we will be using a couple of selection methods that will conveniently give us some optimal models including forward and backward step-wise selection, BIC backwards selection, and all best subset selection. All these selection methods will choose variables based on optimal AIC and BIC to yield optimal p-values and adjusted R-squared values.

Model Validation

To validate the model we will divide the original data into a training set and a testing set. We will divide it 50/50 of 180 observations, so we will have 90 observations in our training data and 90 in our testing data. We then perform the model selection process as described above on our training data set to find the most optimal models. Once we have obtained these models in the training set, we use these models on the testing set and see how it compares to the training set. We will be comparing many characteristics that both models will yield. Some which include if the p-values of predictors become significant or insignificant, if there exists more or less multicollinearity, and if the adjusted R-squared changes significantly. Although the listed are some of the more important areas to note, we generally want consistency in both data sets and nothing that will deviate too much from the other in which we can then conclude and validate the model.

Model Violations & Diagnostics

After choosing the some of the best models through the model selection process, we need to perform some diagnostics on these models to see if they have any violations. Perhaps the first thing we would check is the multicollinearity of the predictors, which can be determined by checking the VIFs of the predictors in the model. If any variables have a VIF value greater than 5, then we need to consider other optimal models and if all of them have high VIFs, then this would be a limitation due to the nature of our data. Afterwards we check for influential points that may skew the data. We can do this by using Cook's distance and DFFITS and hope that the models don't exhibit too many influential points. Finally, a model diagnostic to perform would be checking the assumptions of linearity, constant variance, uncorrelated errors, which can be checked through a combination of residual plots, pairwise predictor plots, response and fitted value plot, and QQ-plots. If there are violated assumptions, we would need to perform a Box-cox or a power transformation on some predictors to satisfy assumptions.

Results

Data Description

The entire data set contains 180 observations taken the seasons during 2013-2019 and it is divided evenly into a training and testing set with 90 observations in both. In each set I have arbitrarily divided the points to avoid any biases that might come up when comparing models between the two sets. There are a total of 20 predictor variables and 1 response variable being the win percentage (Win_per). Table 1 below gives some basic numerical summaries on the mean and standard deviation on all 21 variables in both the training and testing sets. We can see that the means in both sets are relatively close with nothing too differing. One thing to mention is that some of the means are slightly below zero, but these variables are mostly net rating systems which allows negative ratings and this makes sense intuitively since the net usually has an average of zero. However, the standard deviation differs in a couple of variables particularly in MOV, SRS, DRtg, and NRtg. The reason for this is largely be due by chance and the effects of these differences may effect how we validate our models.

Table 1: Summary statistics in training and test data set, each with 90 observations.

Variable	mean (s.d.) in training	mean (s.d.) in test
Age	26.346 (1.644)	26.733 (1.793)
MOV	-0.179 (4.337)	0.18 (5.008)
SOS	0.013 (0.398)	-0.016 (0.396)
SRS	-0.167 (4.211)	0.164 (4.891)
ORtg	107.917 (3.61)	107.587 (3.745)
DRtg	108.112 (2.972)	107.38 (3.38)
NRtg	-0.196 (4.476)	0.207 (5.136)
Pace	96.211 (2.996)	96.252 (2.975)
FTr	0.268 (0.031)	0.271 (0.029)
ThreePAr	0.307 (0.062)	0.301 (0.055)
TSper	0.547 (0.02)	0.548 (0.02)
OFF_eFG	0.51 (0.021)	0.51 (0.022)
OFF_TOV	12.98 (0.991)	13.063 (0.965)
OFF_ORB	24.063 (2.548)	23.582 (2.485)
OFF_FT_FGA	0.204 (0.024)	0.206 (0.021)
DEF_eFG	0.512 (0.017)	0.508 (0.019)
DEF_TOV	12.99 (1.027)	13.058 (1.119)
DEF_ORB	76.316 (1.98)	76.012 (2.057)

Variable	mean (s.d.) in training	mean (s.d.) in test
DEF_FT_FGA	0.206 (0.02)	0.204 (0.021)
Attend_per_game	1.7792311×10^4 (1878.303)	1.7804644×10^4 (1756.447)
Win_per	0.495 (0.145)	0.505 (0.159)

Model Selection Process & Final Model

Before performing the model selection process, it is important to check assumptions on the response vs fitted values shown in Figure 1 and it clearly shows that it satisfies all assumptions. Therefore any transformation on the response is not needed.

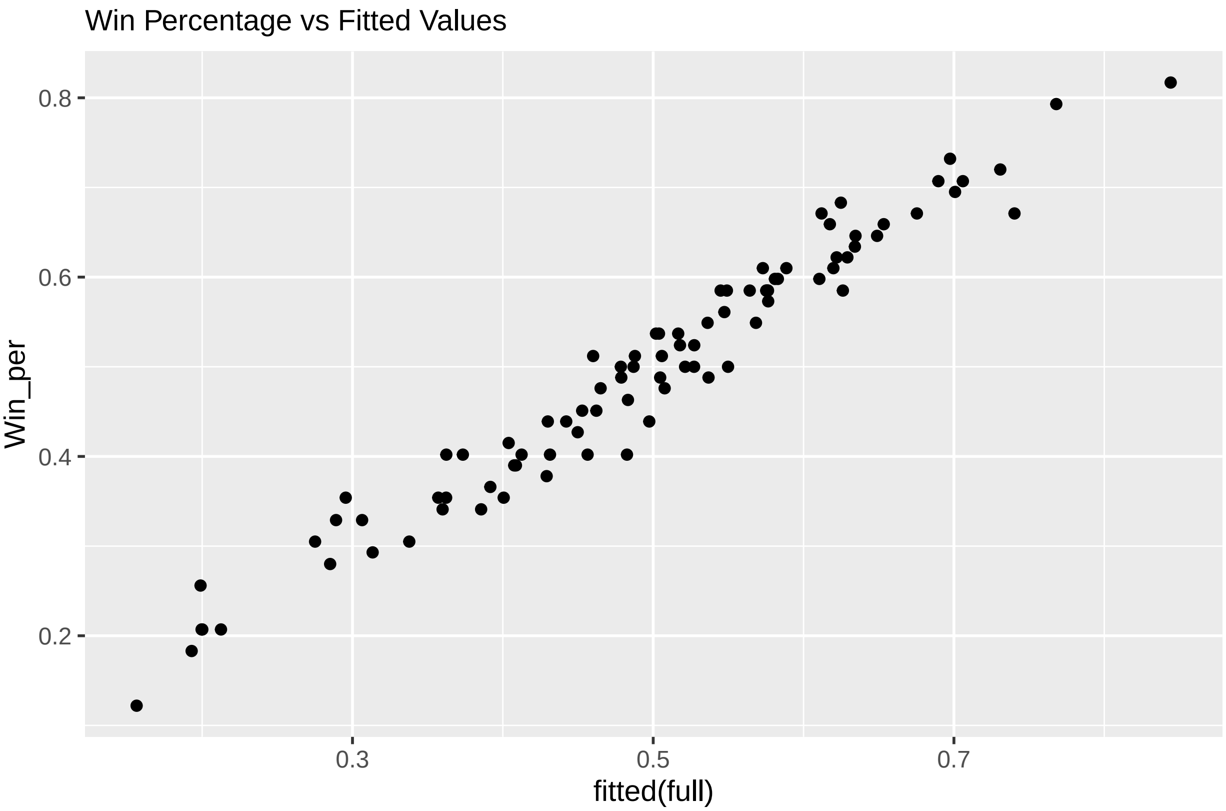

Figure 1

It is also important to check some of the model assumptions. Figure 2 below plots the predictors by the residuals and most of these plots do not violate any assumptions. I later discovered that the Offensive Turnover (OFF_TOV) variable had some minor violations, but it was fixed by a simple log transformation given the Box-cox transformation information. Not all predictors were plotted and the rest can be found in Figure 6 of the appendix.

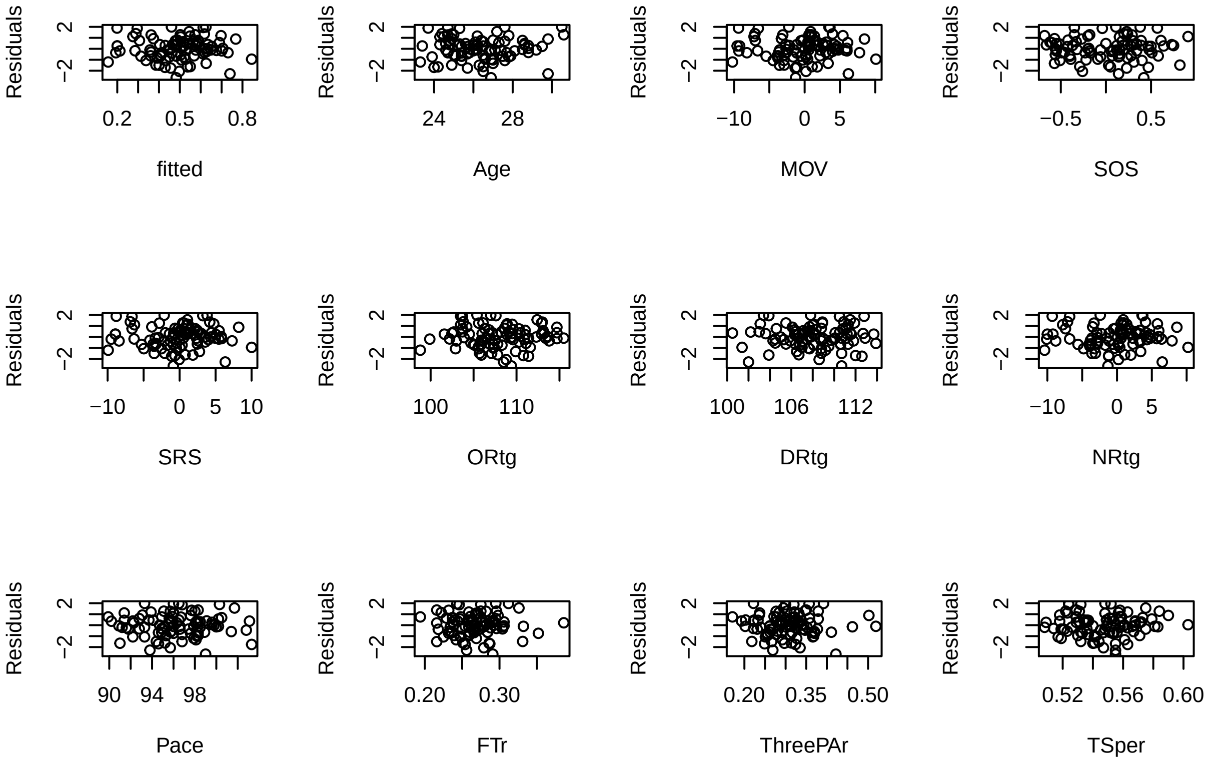

Figure 2

Afterwards the model selection began. I first used the best subset method to obtain three models with the best subset of 2-4 predictors. The models with 5 or more had extremely high VIF values so I had to only consider the models with 2-4 predictors. I obtained one model using forward selection and another model using BIC based backward selection, which gave a total of five optimal models to choose from. I then checked for influential points that may alter the variability of the model. The Cook's distance determined that all five models did not have any observations that were influential, but the DFFITs determined there were about 5-8 observations that were influential own its own predicted values.

The validation of these models were not satisfactory as all of them had some complications when comparing it with the testing set. In fact, all the models tried on the testing set showed that their VIFs increased, had differing intercepts, and at least one predictor became insignificant. In the end I was ultimately limited by the nature of the data and so I chose the model that had the smallest differences in these areas while also having the most satisfactory residual plots and QQ-plots. The model of the best subset of four predictors ended up being the final chosen model.

These four predictors in the final model were the average age (Age), the Net rating (NRtg), the average attendance (Attend_per_game), and free throw rate (OFF_FT_FGA). I then performed some additional model diagnostics shown in both figure 3 and 4 below and figure 5 in the appendix. In all cases we can see that all assumptions are satisfied.

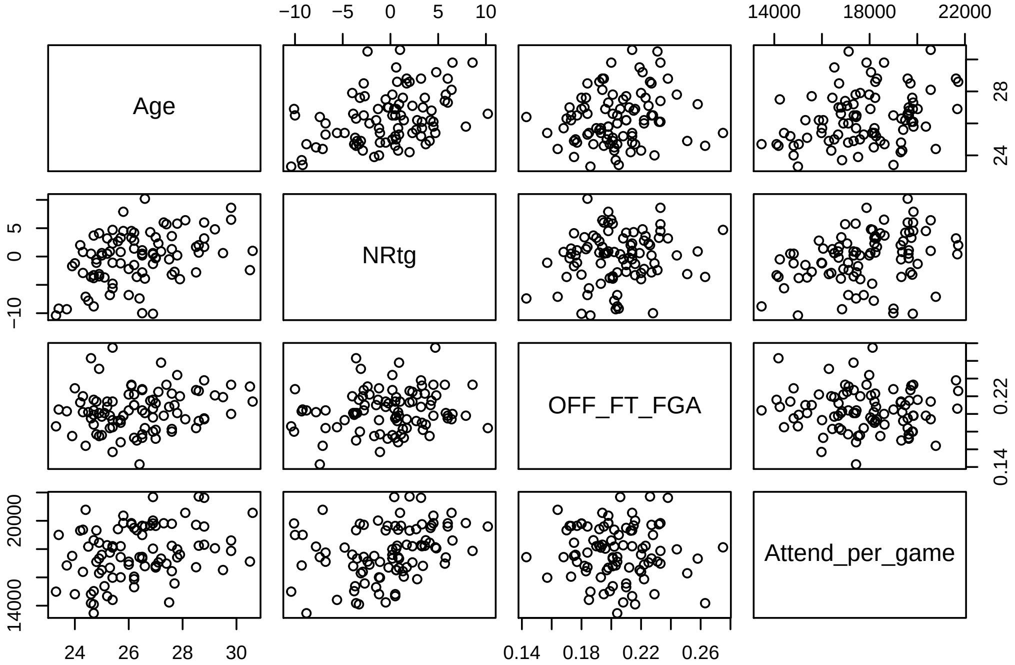

Figure 3

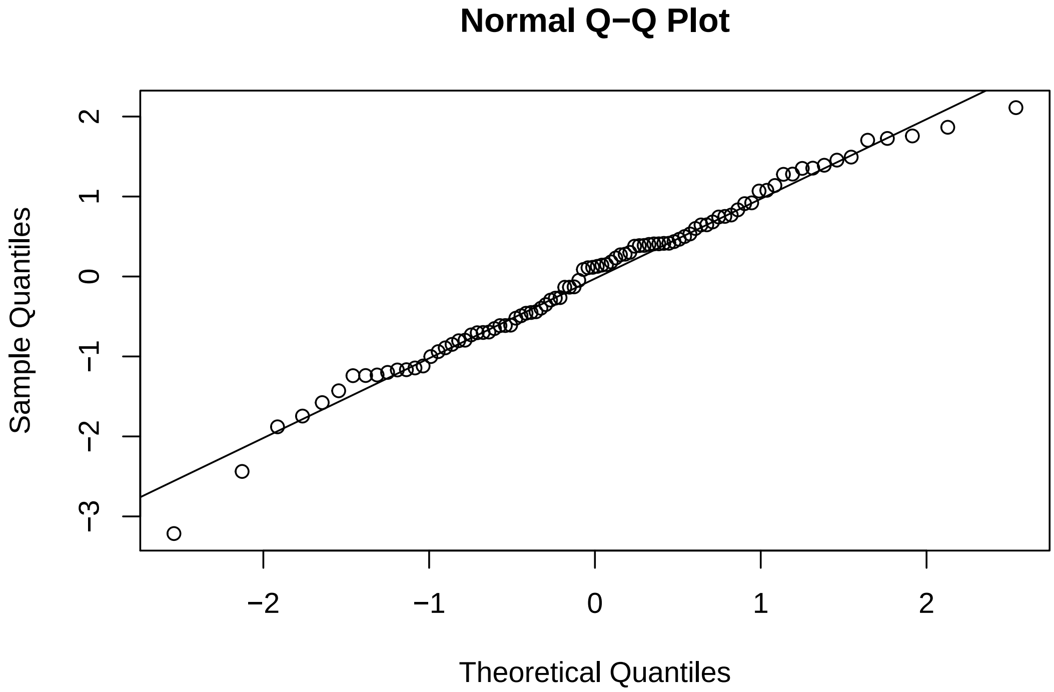

Figure 4

Discussion

Final Model & Interpretation

The final model is Win Percentage ~ Average Age + Net Rating + Average Attendance + Average Free Throws per Field Goal Attempt. The Age variable makes sense contextually the older the team, the more experienced they are. The Net Rating is on average the difference in amount of points scored and points allowed, which essentially describes how a team performs on both ends of the floor. The amount of free throws per shot attempt is an underrated statistic as it essentially details the teams ability to basically get easy points. Finally the most surprising variable that made it in this model is the average attendance variable and perhaps a larger audience does in fact play a role in a teams performance as drives more motivation for the players to play better. The main takeaway from this model is that it contained only one variable in the previous studies mentioned in the beginning of this report (free throws), which begs to question if this model lacks some other variables I have not considered or if it really is the best model.

Limitations

There were several limitations that arose in this project. One being that my training set models couldn't have been truly validated in which the testing data had insignificant p-values and increased multicollinear variables. If we can't entirely validate our model, then we can't completely confirm the model yields the accurate predictions for all cases. On the other hand, there were some influential points calculated from the DFFITs and this can affect the overall slope of the regression. In both cases the VIF increased by a constant amount in all models and the same variables became insignificant, and so these limitations could not be corrected by any form of transformation and are in fact due to the nature of the original data.

Appendix

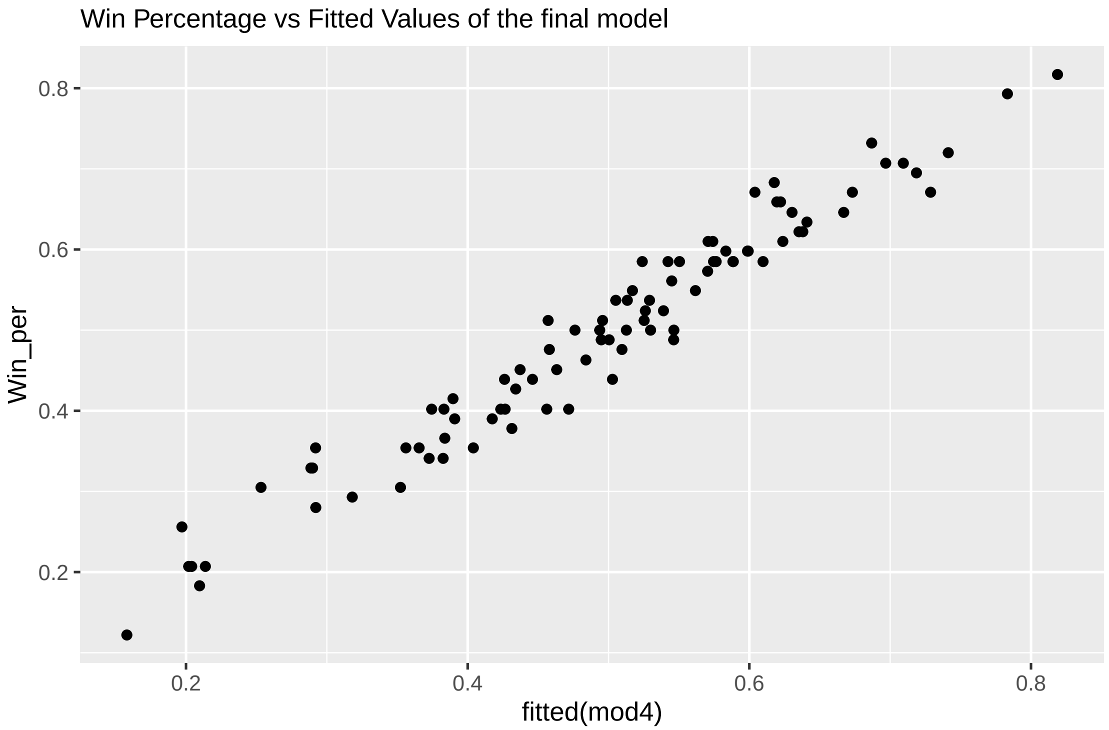

Figure 5

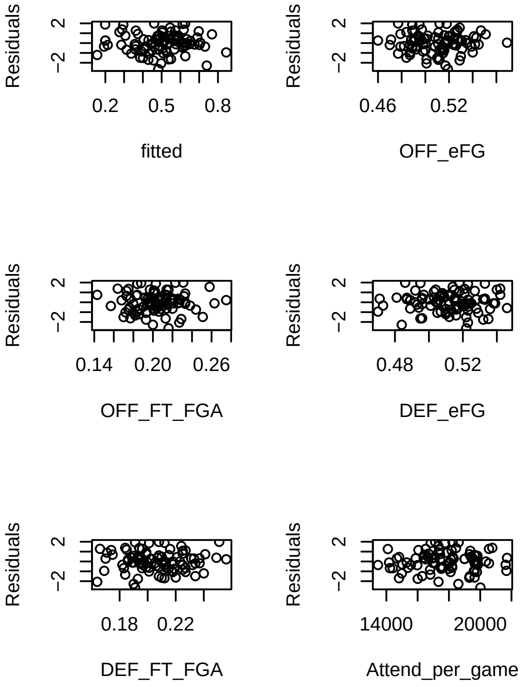

Figure 6

References

Kotzias, K., 2018. The Four Factors of Basketball as a Measure of Success - Statathlon. [online] Statathlon: Intelligence as a Service. Available at: https://statathlon.com/four-factors-basketball-success/ [Accessed 22 October 2021].

Basketball-Reference.com. NBA Season Summary | Basketball-Reference.com. [online] Available at: https://www.basketball-reference.com/leagues/NBA_2019.html#all_shooting_team-opponent [Accessed 22 October 2021].