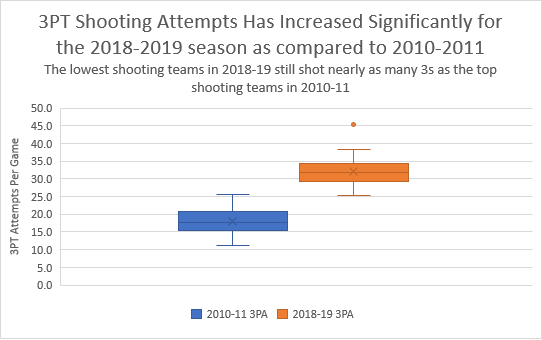
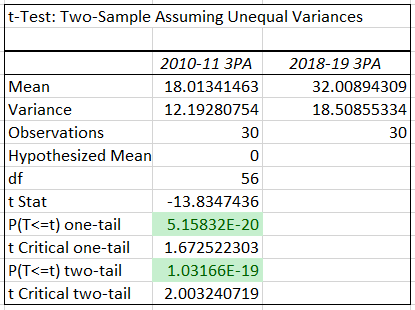
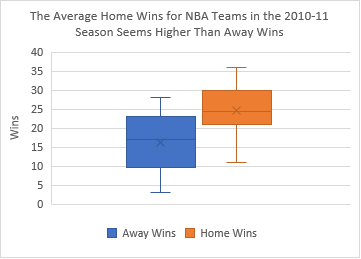
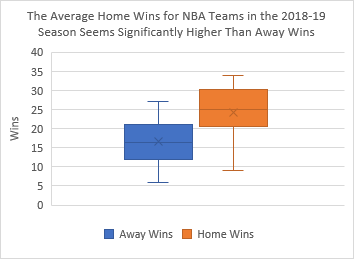
Dear Commissioner Silver,

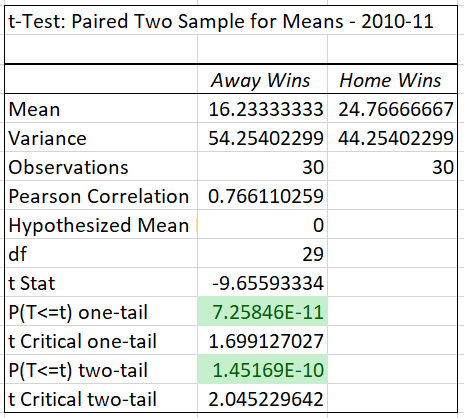
Thank you for assigning this project to investigate the changing trends of the game over the past decade. Using the shooting statistics and home/road splits data I was provided, as well as some additional variables, I am confident I have compiled an analysis that helps explain some of the changes in NBA basketball between the 2010-11 season and the 2018-19 season.

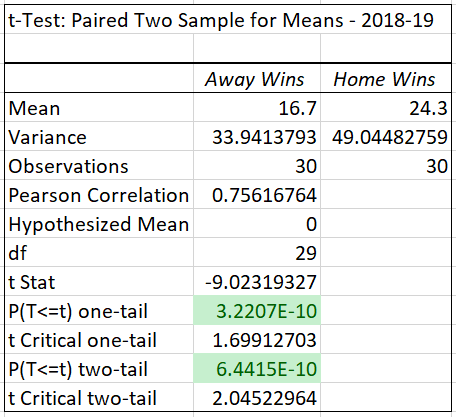
As background to my analysis, I will first provide you an overview of the data I worked with. My first exploratory analysis was to look at changes in 3pt shooting across the league between the two seasons. As the boxplot below indicates, there seems to have been a significant increase in 3pt shooting for the 2018-19 season as compared to 2010-11. It is important to note that there is one significant outlier in the 2018-19 data, as the Houston Rockets attempted 45.4 3-pointers per game. As the leaders of the 3-point revolution, Houston plays a style that is indicative of the shift in philosophy among NBA teams, relying on analytics to influence their offensive strategies to take the most efficient shots on the court.



In addition to looking at this trend visually, I conducted a two-sample t-test assuming unequal variances to confirm that the changes in 3-point shooting is statistically significant. In simple terms, a t-test determines if there is a significant difference between the means of two sets of normally distributed data. The t-test confirmed that difference between the average 3PA for the 2010-11 season and the 2018-19 season are statistically significant.

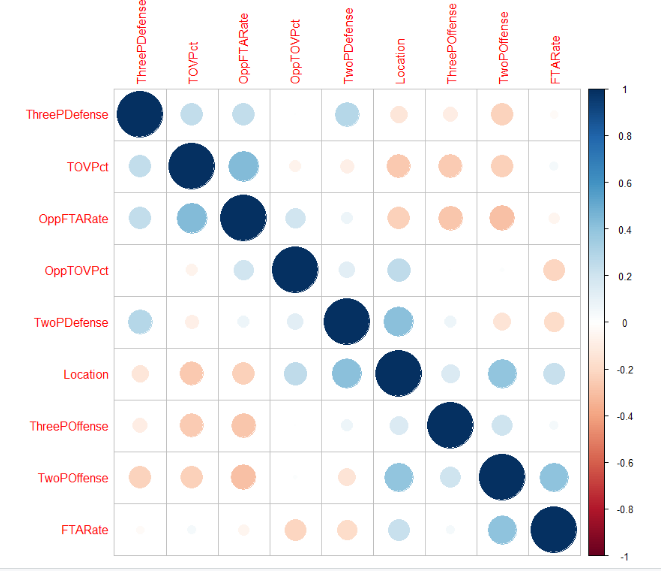
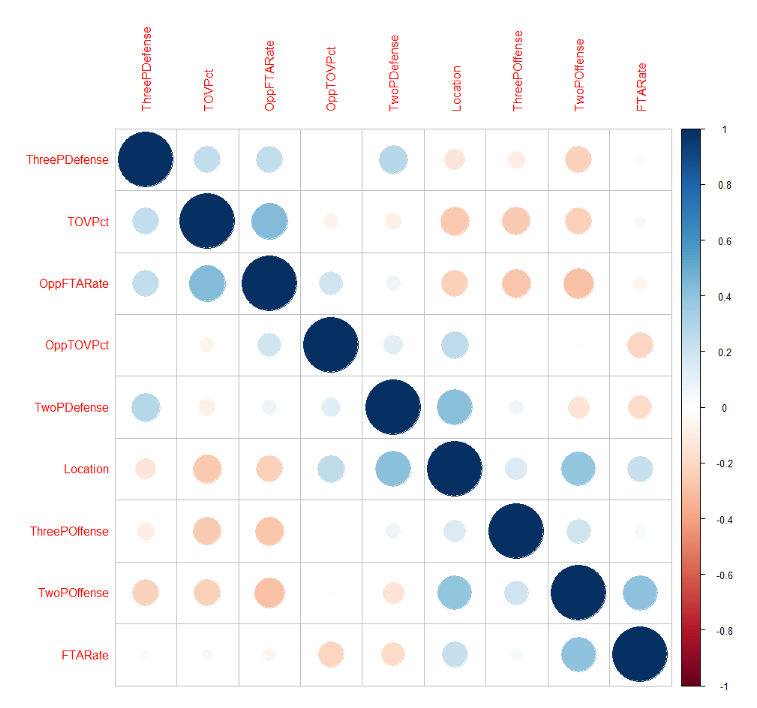
I was also provided data on the home/away wins for NBA teams in these seasons. I performed a similar exploratory analysis, first creating boxplots to look at the difference between home and away wins visually. Both boxplots seem to indicate that home wins are higher on average than away wins. There is slightly more overlap for the 2010-11 boxplot than the 2018-19.



I also conducted two t-tests for the home/away data. In this case, I used a paired t-test to test if the difference between the means in the data sets were statistically significant. A paired t-test was appropriate because the home and away wins data outlines the number of wins for each team under two separate scenarios (home or away). This differs from the shooting data, which looked at two completely separate datasets, as the teams in the 2010-11 NBA had no impact on the 2018-19 teams. For the wins data, the home and away wins are correlated, as the same team is taking the court each time.

The paired t-tests reveal that the difference between home and away wins is statistically significant for both data sets. This indicates that “home court advantage” is a statistically proven phenomenon and has an impact on winning in the NBA.

Following this exploratory analysis, I sought out to develop regression models that determine the relationship between wins, shooting performance, location of games, and other variables that are crucial to winning in basketball. I elected to add turnover percentage (offensive & opponent), offensive rebounding percentage (offensive & opponent), and free throw rate (offensive & opponent) to the models. With the addition of these variables, the models incorporate statistics from each of the major aspects of the game of basketball. While no model is perfect, I am confident that these models give a general sense of the factors that have contributed to how the game of basketball has changed over the past decade.

Prior to running my models, I needed to ensure that the variables themselves were not correlated. Multicollinearity reduces the precision of the individual coefficients, which weakens the overall usefulness of the model. For this reason, I ran a correlation analysis to ensure that there was no multicollinearity between my variables. As the correlation plots visualize, there is no evidence of any variables being strongly correlated with one another.

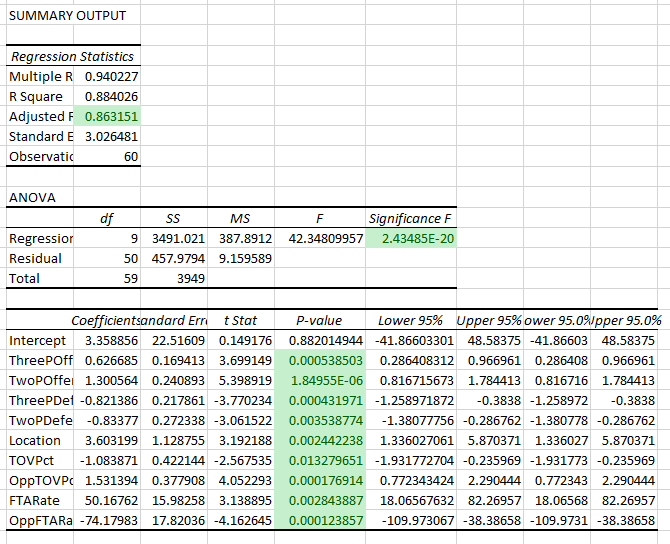
2010 Data Correlation Plot

2018 Data Correlation Plot

It is important to clarify the assumptions I am making with this model. First off, I assume that the models satisfy the tests for linearity, independence of errors, normality of errors, and equal variance. A residual analysis can be found in Appendix A and B for each of the models, which explains how these assumptions are met. I am also assuming that teams have control over the variables that I have included in the models. For example, I assume that a team is able to control their turnover percentage and it is not a superfluous variable out of a team’s control.

For the 2010 model, I ended up removing the rebounding variables, as they were not statistically significant for the 2010 data. For the 2018 model, I removed the free throw rate variables, as those variables were not statistically significant for the 2018 data. For future analysis on other seasons of data, I would advise that those variables are included for the initial testing of the model, as they are logically important aspects of the game of basketball.

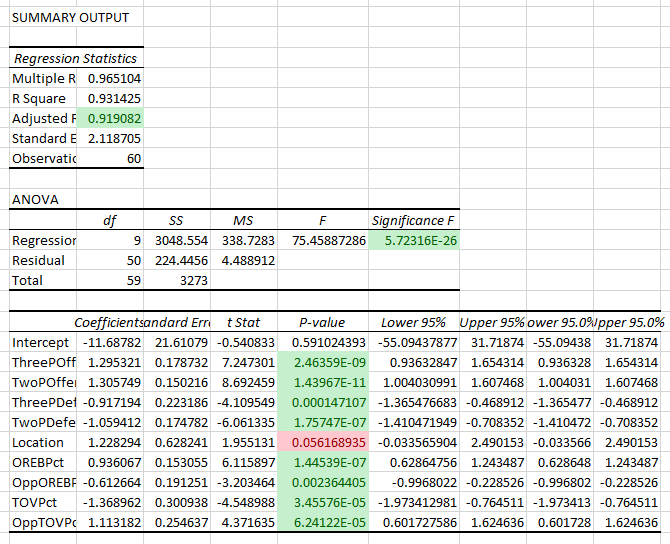
**NBA 2010-11 Model:**



With an adjusted r-squared of .863, this model explains 86.3% of the variation in predicting wins. The r-squared measures correlation and indicates that these 9 variables account for 86.3% of the variation in wins. The standard error is approximately 3.03, as the observed win totals variate slightly from the regression line. The significance F is extremely low, so it is safe to assume that this model is statistically significant. Also, on an individual variable level, each of the 9 variables is statistically significant at the 95% confidence level, as each of the p-values is below .05.

In the general sense, this seems to be a useful, reliable model for predicting wins for the 2010-11 NBA season. The model is statistically significant and each of the variables seems to contribute to the model. Refer to Appendix C for the predictions of wins on a team-by-team basis during the 2010-11 season using the model.

**NBA 2018-19 Model:**

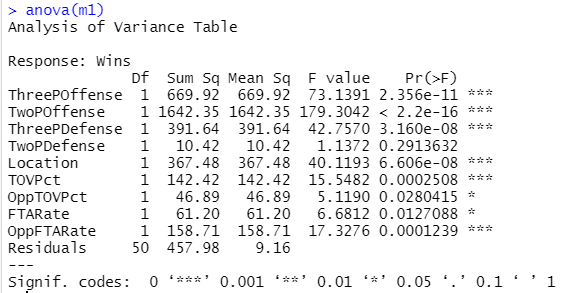
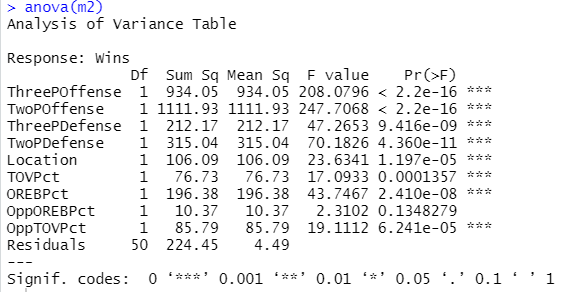


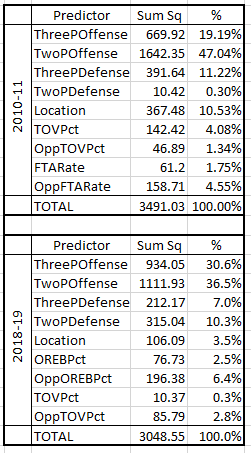
With an adjusted r-squared of .919, this model explains 91.9% of the variation in predicting wins. The standard error is approximately 2.12, as the observed win totals variate slightly from the regression line. The significance F is extremely low, so it is safe to assume that this model is statistically significant. Also, on an individual variable level, 8 of the 9 variables are statistically significant at the 95% confidence level, as each of those p-values is below .05. The only variable that has a p-value above .05 is location, with a p-value of .056. While this variable is not statistically significant at the 95% confidence level, I have elected to keep it in the model because location does seem to be an important factor for predicting wins in the NBA given the previous exploratory analysis. It may be slightly superfluous for explaining wins in this given data set, but it did not seem to be worth removing from the model.

In the general sense, this seems to be a useful, reliable model for predicting wins for the 2018-19 NBA season. The model is statistically significant and each of the variables, despite location’s statistical significance, seem to contribute to the model. Refer to Appendix D for the predictions of wins on a team-by-team basis during the 2018-19 season using the model.

**Comparison Using ANOVA:**

To compare these two models, I used an analysis of variance (ANOVA) to identify the most impactful variables in each of the models and how they differ between the two seasons studied. The outputs for the ANOVA analysis are below, with 2010-11 on the left:





Using this output, I was able to analyze the individual importance of each predictor to predicting wins, assuming no collinearity and that the other predictors are held constant. In the table to the right, the % column gives the relative importance for each predictor.

For the 2010-11 data, the model indicates that 47.04% of the change in Wins is due to Two Point Offense, holding the other variables constant and assuming no collinearity. Three Point Offense was the next most influential predictor, as it explained 19.19% of the change in wins. Three Point Defense and Location were the next most influential predictors.

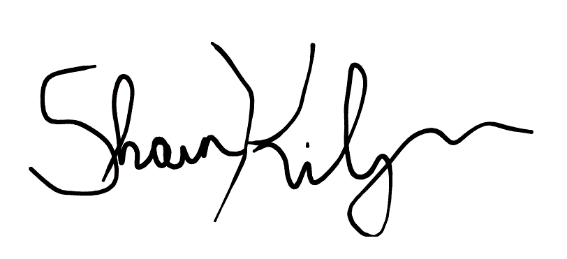
For the 2018-19 data, the model indicates that Two Point Offense was still the most influential predictor at 36.5%. Three Point Offense explained 30.6% of the change in wins in the 2018-19 data, holding the other variables constant. Two Point Defense was the next most influential predictor.

**Conclusion:**

Based on these models, the rise in 3-point shooting has undoubtedly changed how a team is successful in winning games in the NBA. While 2-point offense is still one of the most important indictors of winning, 3-point offense was significantly more important to winning in the 2018-19 season than it was in the 2010-11 season. While other factors, such as rebounding, turnovers, and free throws, do contribute to a team’s winning formula, it seems clear that offensive output, in the form of 2-point and 3-point scoring, is most important for a team to win. That is evident in both seasons, but the way offenses distribute those shots between 3s and 2s seems to have significantly changed.

It was a pleasure to complete this analysis. Please feel free to stop by my office with any questions.

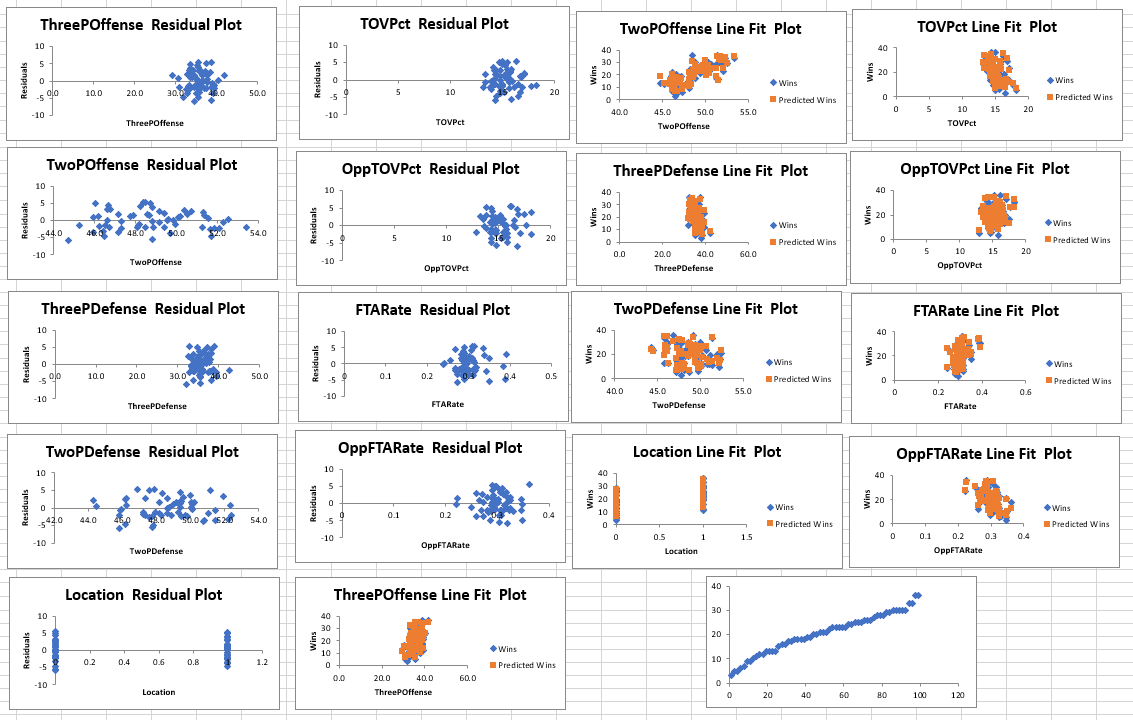
Best,

 Shawn Kilcoyne

Director of Player and Team Analysis,

National Basketball Association

Appendix A: Residual Analysis for NBA 2010-11 Model



Linearity:

There is no evidence of a curvilinear pattern in the residual plots, so this assumption is met.

Independence of Errors:

There is no evidence of a cyclical pattern in the plots, so the data seems to be independent.

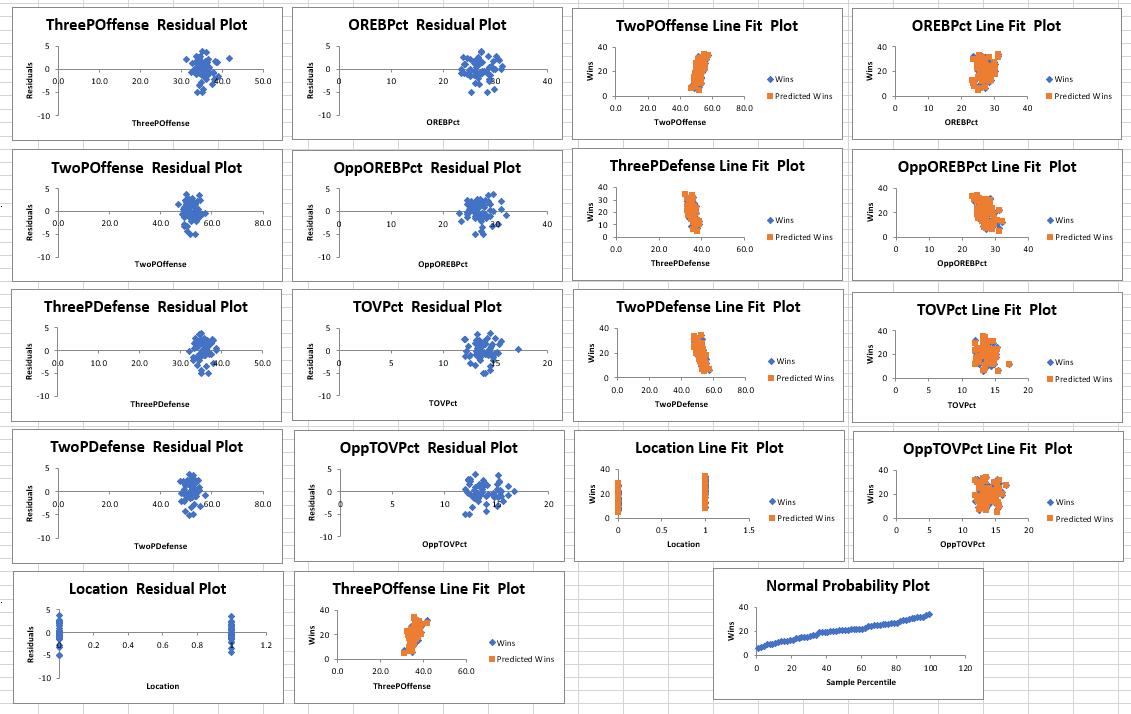
Normality of errors:

The normal probability plot features a linear line, indicating that the data seems to be normally distributed. The tail does slightly deviate from the linear line, but it does not seem to be significant enough to affect the normality of the errors.

Equal variance:

Each of the residual plots seems to have constant variance, so this assumption is met.

Appendix B: Residual Analysis for NBA 2018-19 Model



Linearity:

There is no evidence of a curvilinear pattern in the residual plots, so this assumption is met.

Independence of Errors:

There is no evidence of a cyclical pattern in the plots, so the data seems to be independent.

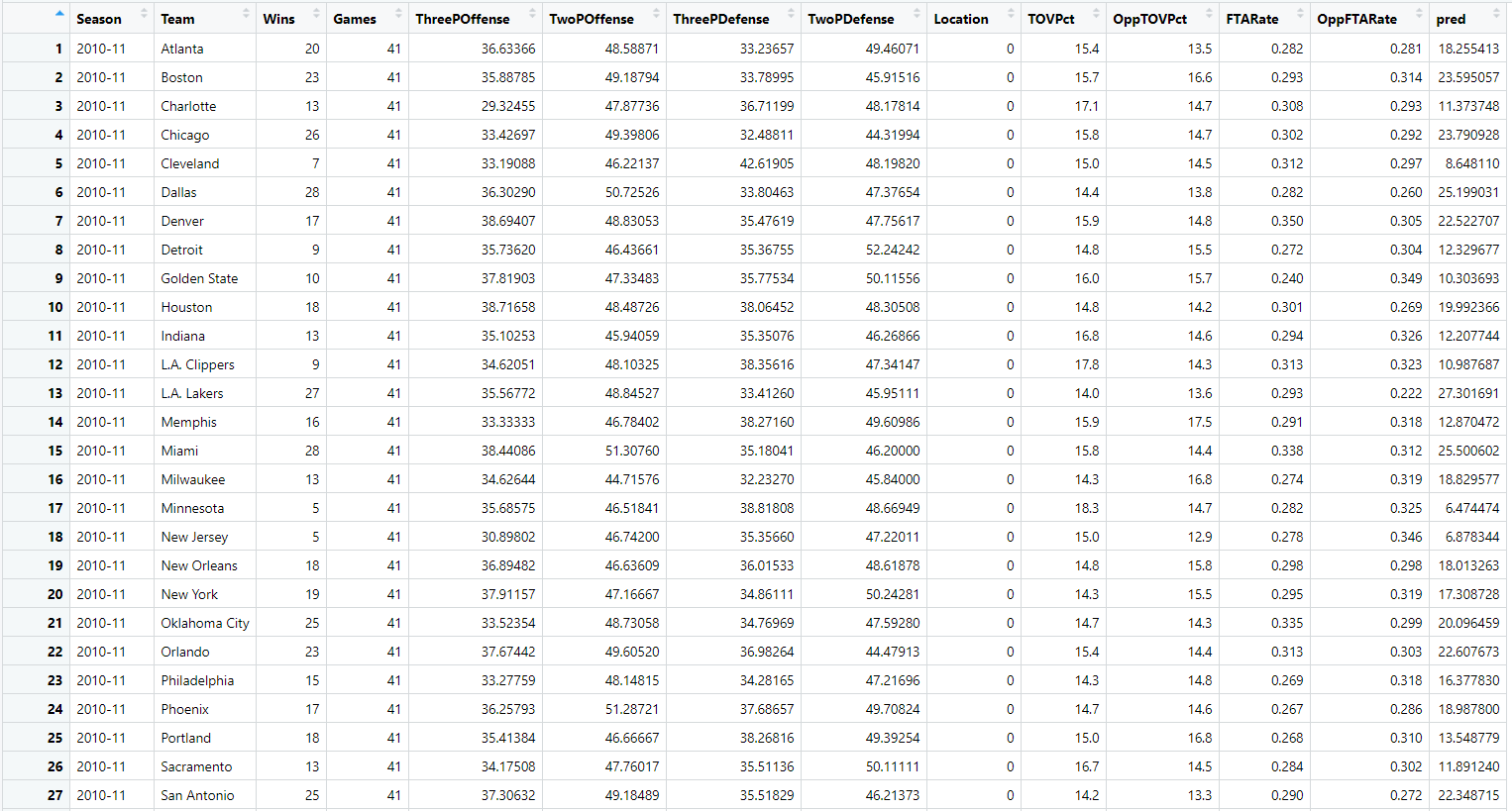
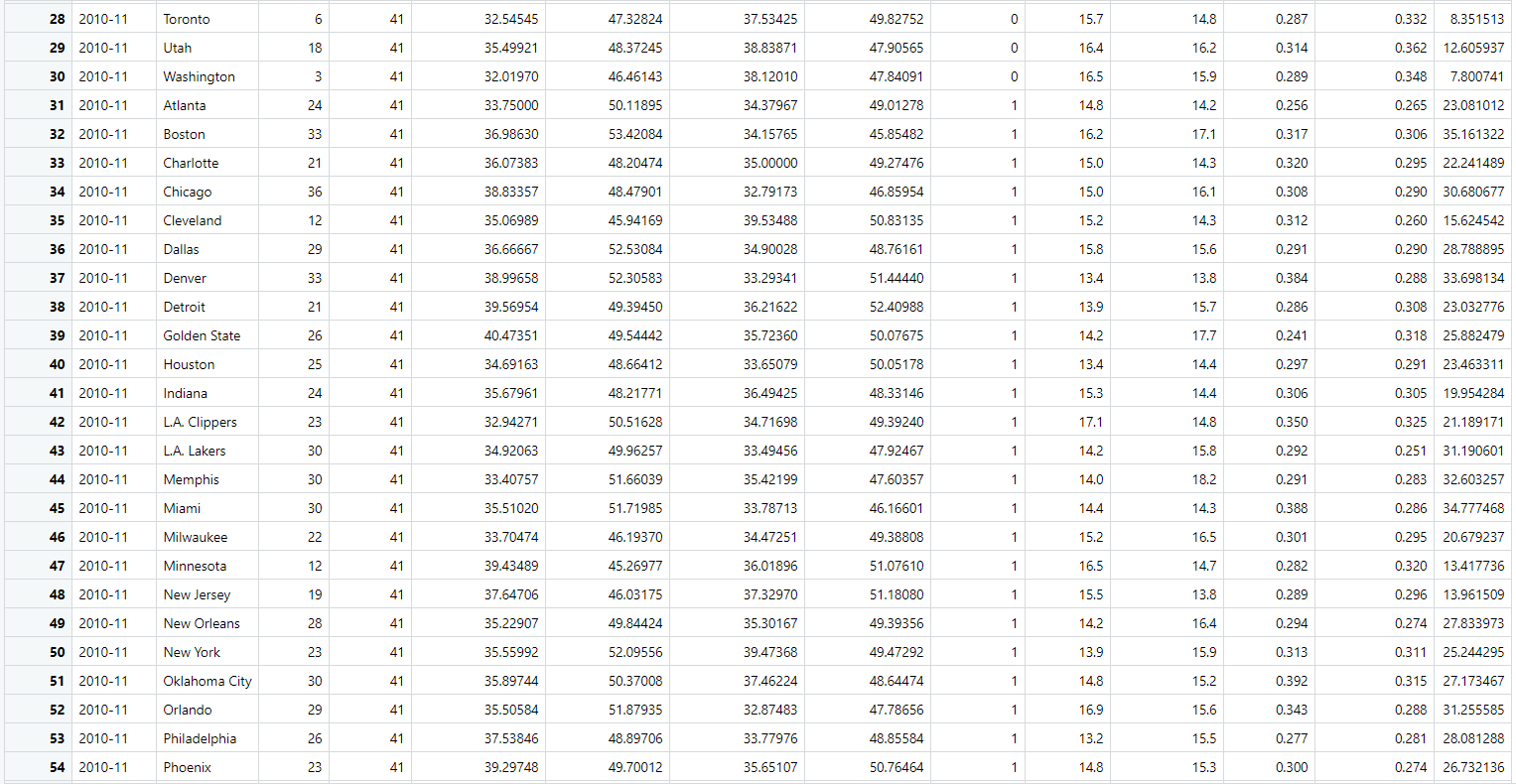
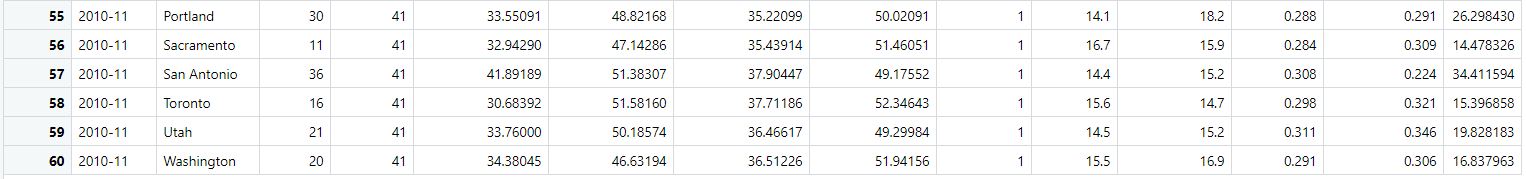
Normality of errors:

The normal probability plot features a consistent, smooth linear line.

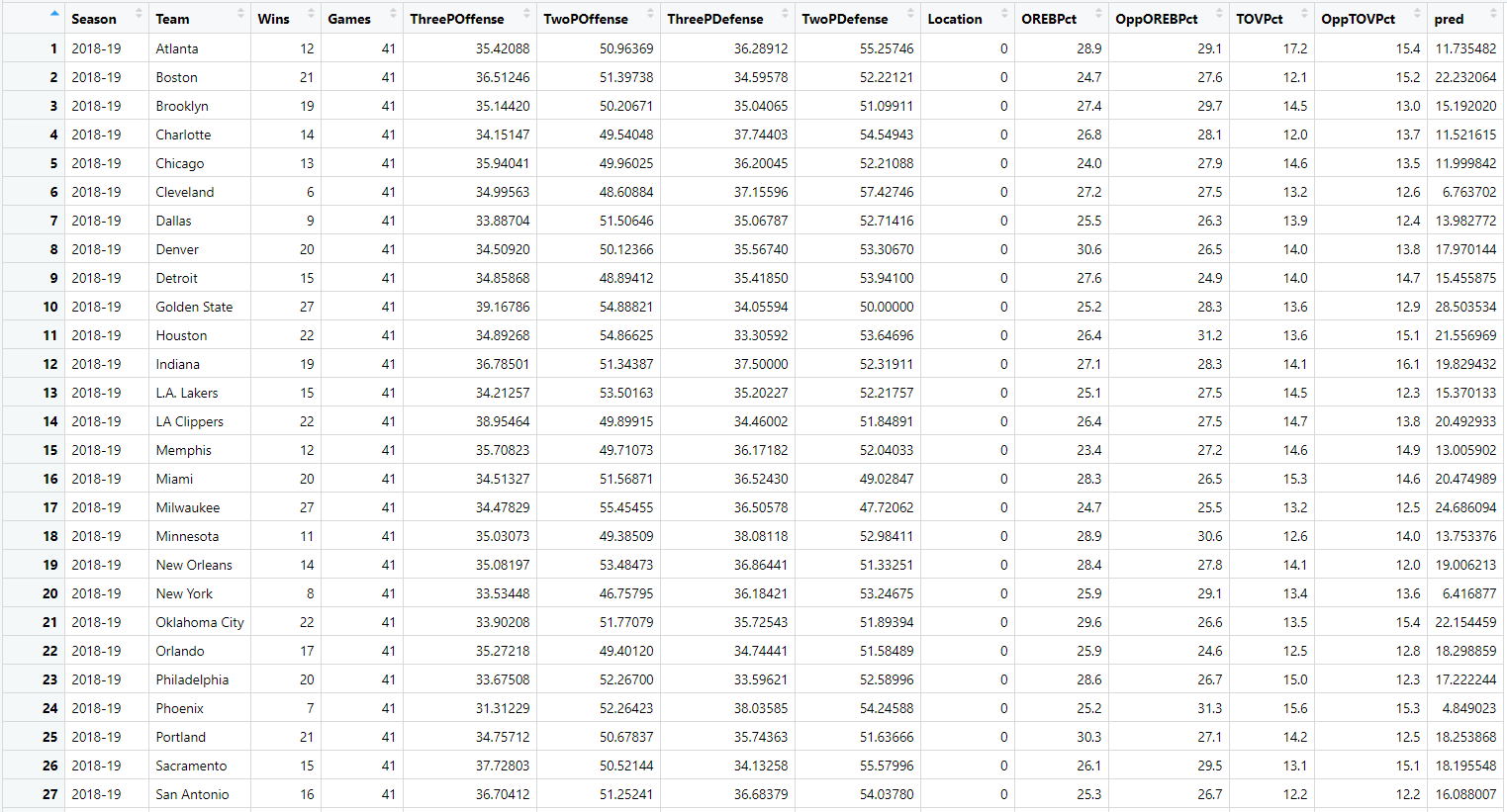
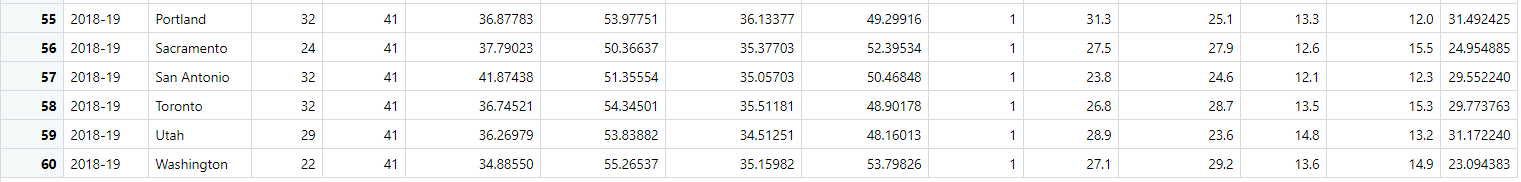
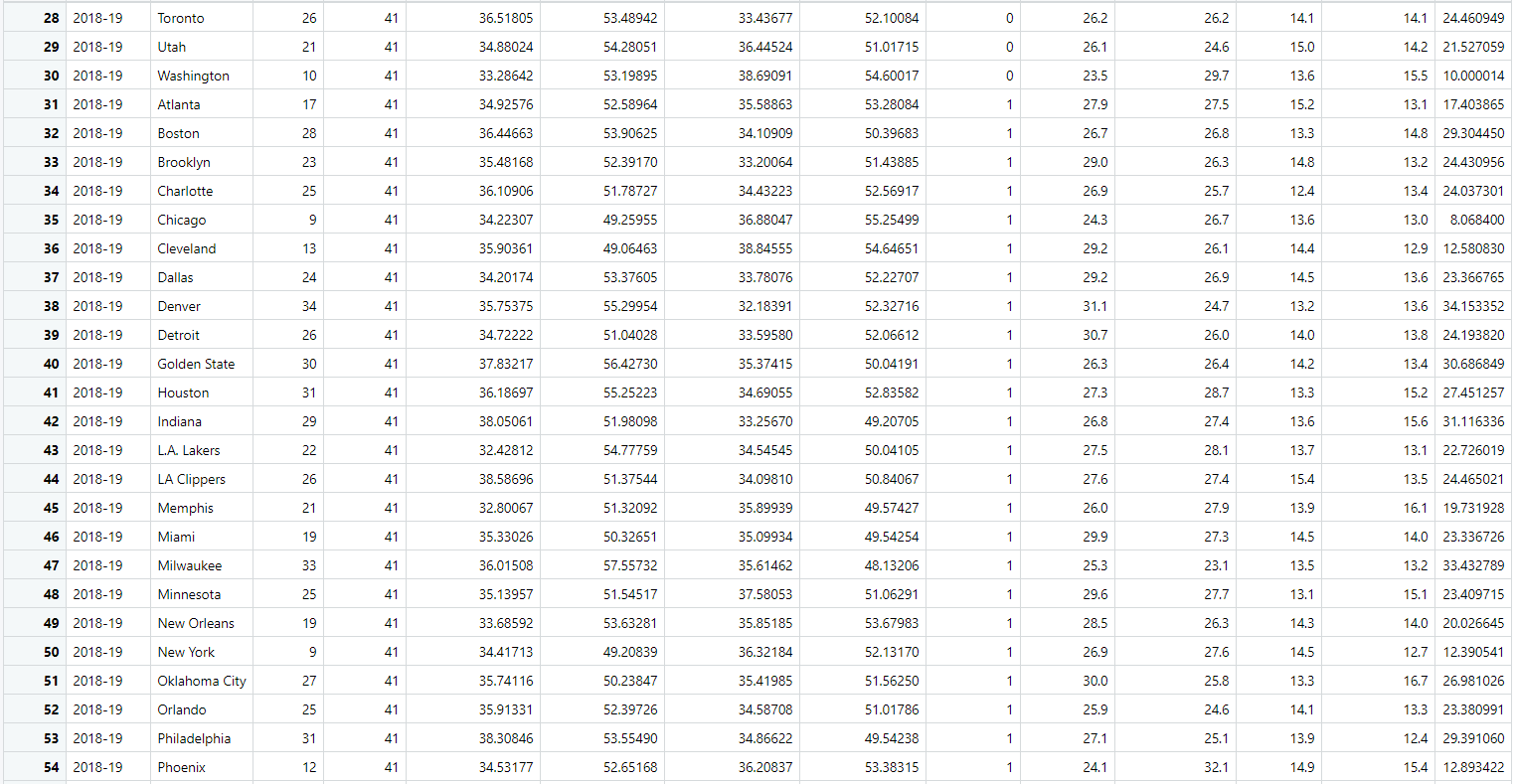
Equal variance:

Each of the residual plots seems to have constant variance, so this assumption is met. The Opponent Turnover Percentage plot is the only residual plot that could possibly show any evidence of non-constant variance, but even that does not seem to be clearly non-constant.

Appendix C: Predictions Using NBA 2010-11 NBA Model

Appendix D: Predictions Using 2018-19 NBA Model

Appendix E: Project Code

