Analysis of Win Percentage of NBA Teams in the 2015-2016 season.

Mohammad Khan

Rodrigo Rogel-Perez

Anil Timbil

Meixi Chen

Yves Pierre

1. **Introduction**

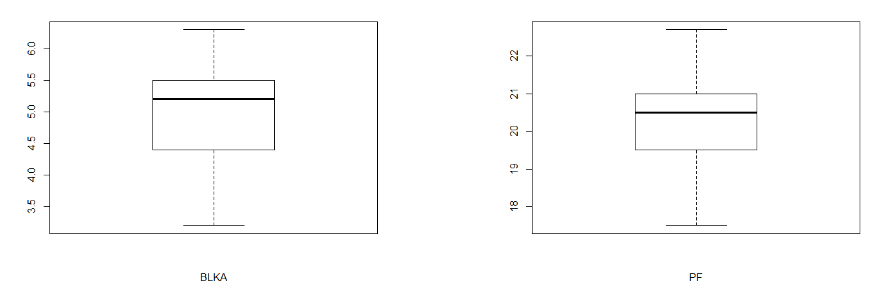
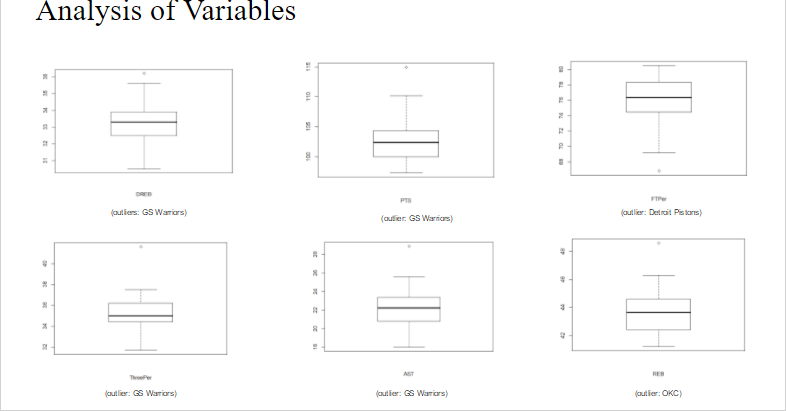
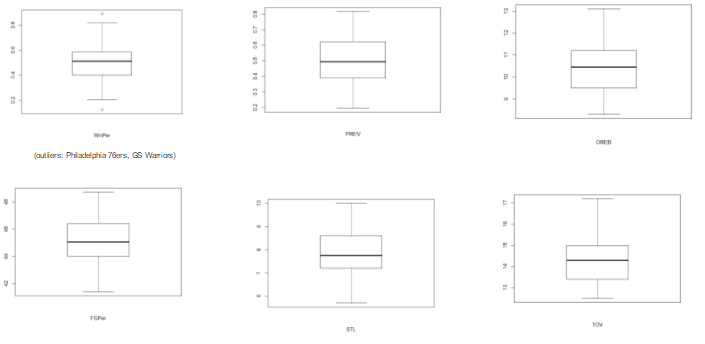
In this paper, we aim to predict the win percentage of NBA teams in 2015-2016 season. We obtained this data from the NBA Statistics webpage[[1]](#footnote-0). Within this approach, our theory leads us to consider the predictor variables such as: Points per game (PTS), Field Goal Percentage (FGPer), Three Point Percentage (ThreePer), Free Throw Percentage (FTPer), Steals per game (STL), Assists per game (AST), Rebounds per game (REB), Blocks Against Team per game (BLKA), Personal Fouls per game (PF), Turnovers per game (TOV), Defensive Rebound per game (DREB) and Offensive Rebounds per game (OREB). We also augmented this season’s data by adding two other columns: the Previous Year’s Win Percentage (PREV) and the team’s Region (REG) signified as a qualitative variable[[2]](#footnote-1). We would like to check if there is a linear relationship between our response variable, Win Percentage (WinPer), and all the predictor variables. Throughout this paper, we will conduct a pre analysis of our variables, examine our full and reduced models, test the 4 regression assumptions after establishing a final model, and create a confidence and prediction interval using this model.

It is important to establish a firm understanding of the impact each predictor variable has on the Win Percentage in regards to how basketball is played. We expect a team to have a better chance at winning if it has more PTS. A high FGPer suggests that a team is highly accurate in terms of making shots, and this allows the team to score more points. As we know scoring more points allows for a better chance at winning. We can make a similar argument for ThreePer and FTPer. Having more STL allows the team to not only stop the other team from scoring, but to also have another opportunity to score. When a team has a high OREB, it means that it has extra opportunities to score - these are also called “second chance” opportunities. Having more opportunities to score suggests that the team can score more points. When a team has a high DREB, it means that the team is able to secure many shots that the opposing team missed. This suggests that the opposing team is missing opportunities to score. If the opposing team scores less, the defensive team has a better chance at winning. When a team gets blocked (BLKA), it misses opportunities to score. This variable should decrease the team’s winning percentage. Late in games, some team’s can be in a penalty time frame. This means that if the team fouls, the opposing team will get Free Throw attempts. Thus, with more PF, we can hypothesize that the opposing team has a chance to score more points and hence a team’s WinPer will decrease.

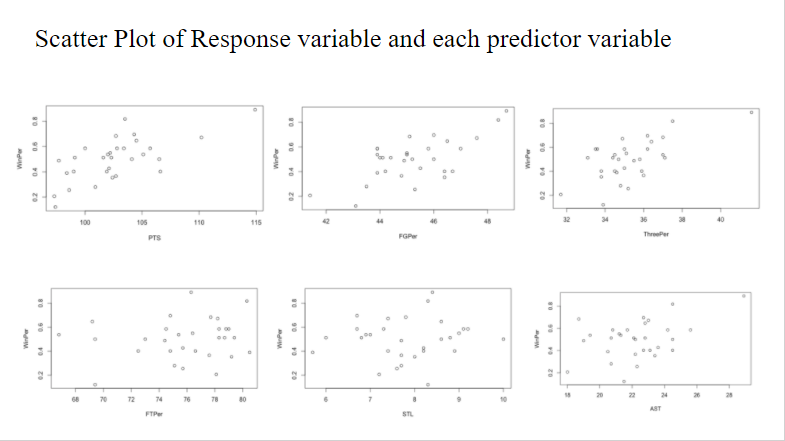
We decided to include the previous year’s winning percentage, PREV, after collecting our initial data with the assumption that performance in a previous year may provide extra information on the ability of a team to win games. We added REG to our data because, from a first glance, it seems as if teams in the Western conference are stronger than teams in the Eastern conference in regards to WinPer. We interpret this variable in a qualitative way, where Eastern conference teams are labeled with a 1 and Western conference teams with a 0.

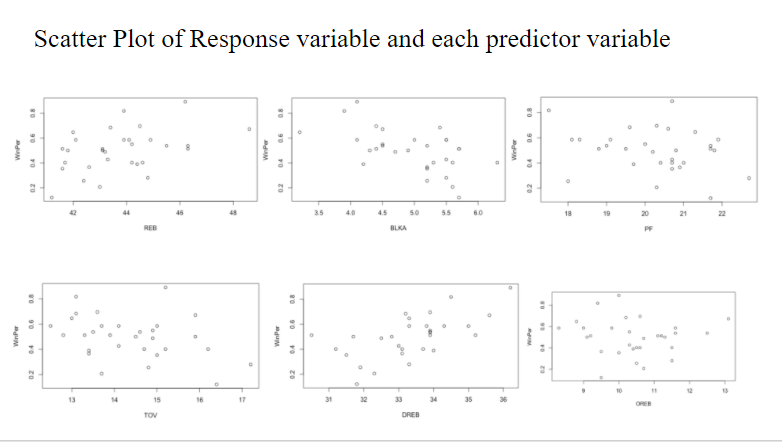
**II. Pre Analysis of the Variables**

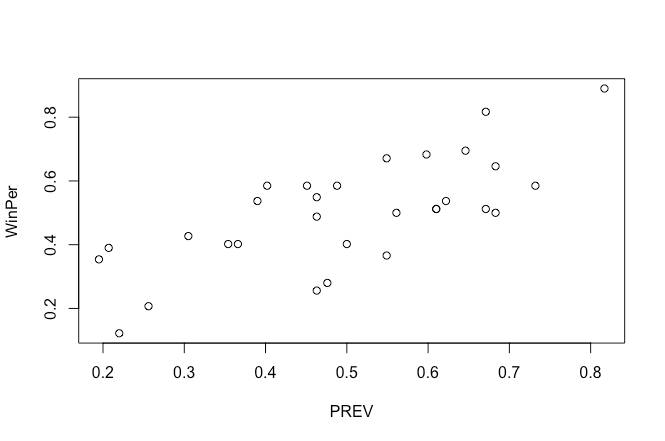
In this part, we used box plots and scatter plots to check if there are any observations that may cause a problem in our future model. For each boxplot, we tried to identify the data points that were away from the rest of the observations. We also wanted to see a linear relationship in the scatter plots of the response and each predictor variable.

**Figure 1.** Box Plot representations of the variables.

By looking at Figure 1, we can notice that some of the box plots have observations that are away from the mean of the data, that we can call outliers. For instance in WinPer graph, we see that there are two outliers Warriors at the top and Philadelphia at the bottom. Even though, they are away from the mean, we think removing them may not be a good idea for now. They might be statistically useful. Furthermore, we mostly identify Warriors in the other boxplots as an outlier. In addition, FTPer and REB graphs show that Detroit and Oklahoma might cause us a problem in our future model. The rest of the box plots show the data do not have any outstanding observations that might be problematic. Before we create our full model, we also looked at the scatter plot of each predictor variable against the response variable.



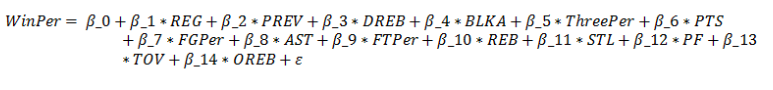


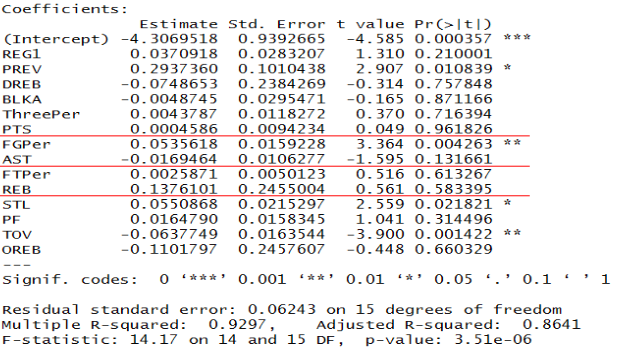


**Figure 2.** Scatter Plots representations of the response variable against each predictor variable

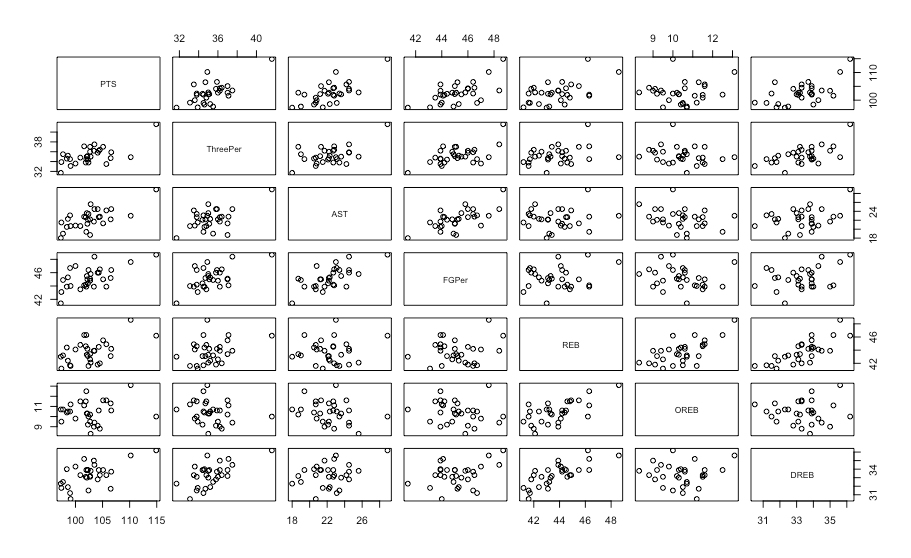
In Figure 2, we see that PREV, BLKA and DREB had very clear linear relationships with our response variable. On the other hand, PTS, FGPer, ThreePer, AST, REB, STL, OREB and TOV had somewhat linear relationships with the response variable. However, PF and FTPer both are randomly distributed, and we cannot observe a clear linear relationship. To get a more linear trend, we tried to perform common transformations such as square, natural log, and square root, but we could not get any results that were closer to a linear relationship. The fact that we did not see any parabolic curves or curves that are representative of log and square root functions in our scatter plots reinforced our decision to move forward without transformations.

**III. Initial Approach to our Full Model**

After a thorough analysis of the response and predictor variables, we saw that most of the predictor variables suggest a clear or somewhat clear linear relationship when regressed against the response variable. Keeping in mind the observations that might cause a model failure, we decided to create our first model. We chose a direct approach and ran all the predictor variables against WinPer our response variable. Our reasoning for doing so is simple: We thought it would be easier and more efficient to start with everything (the full model), and little by little pick out what did not “work” until we reached the “best” model. Also, we found this approach to be much more informative in terms of seeing the effects that different variables have on the model when they are removed rather than added. Figure 3 presents our first model which we refer to as the *Full Model* and its corresponding regression table. We analyze the results.

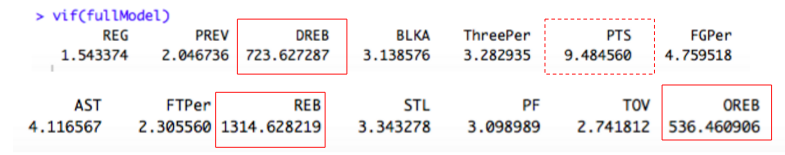
**Figure 3.** The linear regression equation that we started with and its summary table. 

After fitting our linear regression equation relating WinPer to all the other predictor variables, we encountered a problem that caused a model failure. We suspected collinearity was present because of inconsistent coefficient estimates accompanied by a high R2 value and insignificant predictor variables. R2 was approximately 0.93 which suggested a very strong linear relationship between the response and predictor variables. However, most of the predictor variables shown in Figure 3, were not very significant in predicting the response variable. Specifically, the predictor variables DREB, BLKA, ThreePer, PTS, REB and OREB had very high p-values suggesting a model failure. This is a very clear sign of collinearity along with the inconsistent parameter coefficients. Looking at the slope parameters, we realized that some of the coefficients did not make sense in the prediction. For example, we intuitively expected to see an increase in WinPer as DREB, OREB and AST increase individually. Offensive rebounds, defensive rebounds, and assists are all things that keep the prospect of scoring alive, and in fact have led to teams to score in the past. However, a unit increase in these variables suggested a decrease in WinPer. Assuming all of the other predictor variables constant, a unit increase in DREB (Defensive Rebound) caused a decrease of 0.0748% in WinPer; a unit increase in OREB (Offensive Rebound) caused a decrease of 0.1101% in WinPer; a unit increase in AST (Assists) caused a decrease of 0.0169% in WinPer. Because of their negative parameter coefficients, these three variables suggested a discrepancy when compared to our real world expectations.

To further validate our suspicion that collinearity was present, we analyzed the correlation matrix and scatter plots of the predictor variables. According to the 3rd assumption, all of the predictor variables are supposed to be linearly independent of each other. However, some of the correlation values between the predictor values were significantly high again suggesting a model failure. The approximate correlation coefficients between variables of interest are as follows: 0.64 for *PTS* & *ThreePer*; 0.68 for *PTS* & *AST*; 0.60 for *PTS* & *FGPer*; 0.66 for *AST* & *FGPer*; 0.77 for *REB* & *DREB*; 0.65 for *REB* & *OREB*. Additional inspection of the corr. matrix revealed that the predictor variables, PTS, REB and AST are highly correlated with multiple predictor variables. Believing these variables were the root of our problem, we considered removing them from our model. Before doing so however, we analyzed the scatter plots of the predictor variables which reinforced the linear relationship between some of the predictor variables. We can see some examples of this linearity in Figure 4. Most of these scatter plots do not suggest a random distribution of observations between predictor variables. 

**Figure 4.** The scatter plots of some predictor variables that illustrate the dependence between the predictor variables

After the graphical analysis of our predictor variables, we ran a VIF (Variance Inflation Factor) test to see which predictor variables had high VIF values.

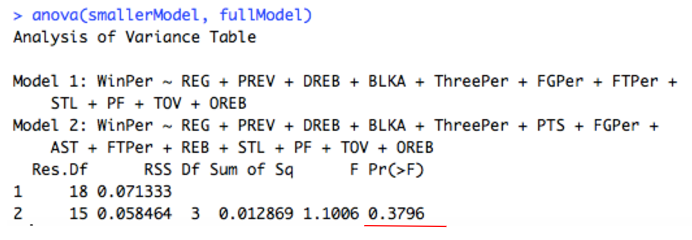


**Figure 5.** VIF Test of the full model

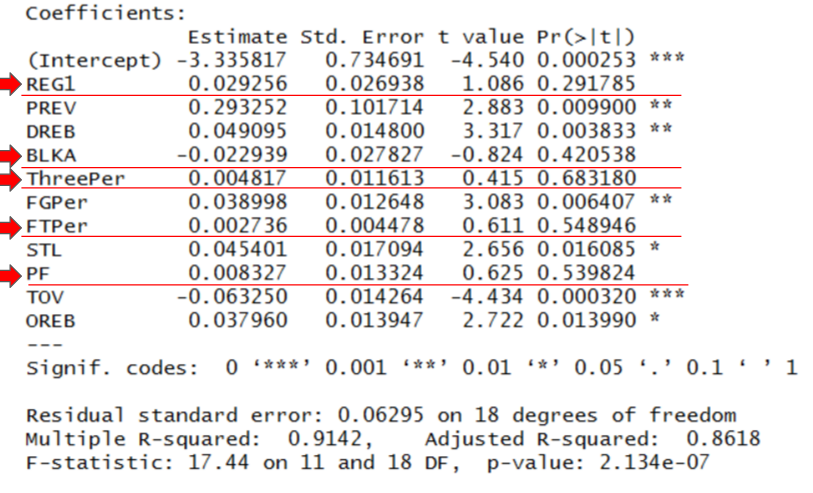
As we can see in Figure 5, DREB, REB and OREB have very high vif values whereas PTS have a vif value that is very close to 10. Once again, we saw that our full model violated the 3rd regression assumption and we decided to remove some of the variables.

Our decision to remove some of the variables were based on everything we analyzed. We considered inconsistent parameter coefficients, scatter plots and correlation matrix of predictor variables and the VIF Test. VIF Test revealed the interrelation between REB, DREB and OREB. We leaned towards keeping OREB and DREB because we knew REB is the sum of OREB and DREB. We thought removing only REB would solve the collinearity issue between these variables. The scatter plots of these two variables also suggested a clear linearity between other variables. Also, we thought that removing REB would fix the slope parameters of OREB and DREB. However, we could not find anything that would fix the parameter coefficient of AST. Therefore, the best way to eliminate the collinearity was to remove AST from the model. For the reasons above, we decided to remove PTS, REB and AST from our full model and run a reduced model with the remaining variables.

**IV. Reduced Model**

We removed variables PTS, AST, and REB to produce a new and simpler model. We call this model the *Reduced Model*. As a reminder, we chose to omit these variables because their presence caused collinearity in our model, and the third assumption of regression analysis specifically says that predictor variables are independent. We found that the presence of one or more of these variables caused other variables to have inconsistent coefficient estimates i.e a negative or positive coefficient where we expected to see otherwise. A specific example of this is that the coefficient estimate for the variable OREB was negative. Recall that we expected this coefficient to be positive since recovering the ball after a missed shot should still offer the prospect of scoring. Afterwards, we did an ANOVA test (shown in Figure 6) to see if our new model is adequate compared to the full model. We found that the p-value was so high (p-val = .3796). This means that we failed to reject the null hypothesis, which claims that the slope parameters of PTS, REB and AST equal to 0. So, our interpretation was that the reduced model is adequate. Convinced that our *Reduced Model* was better than our previous model, we decided to move forward with it and ran a regression. Figure 7 shows the regression (summary) table. 

**Figure 6.** The output of the Anova table comparing Full and Reduced Model

**Figure 7.** The summary table of the reduced model is shown.

The first thing we noticed was the increase in significance for some of the remaining variables. PREV now is more statistically significant having a lower p-value, a similar story follows for DREB and TOV. We interpreted this to be a good sign for the more statistically significant a coeff. estimate is the more likely it is that that variable adds value to the model. We did see a fall in the coeff. of determination, R^2 (it was previously .93), but we did not agonize over this for the difference is dismal and the ANOVA test confirmed that our new model is better than the previous. However, one thing that did concern us was the presence of some insignificant coeff. estimates. This was so because we had just eliminated collinearity and this could only mean the remaining insignificant variables were adding no value to our model. In other words, it was an indication that our model could still improve removing insignificant variables. As a sanity check, our group ran another VIF test and found no traces of collinearity with all VIF values smaller than 10. Having confirmed that the presence of the remaining insignificant variables (reg1, BLKA, ThreePer, FTPer and PF) was not an indication of collinearity, we decided to omit their corresponding variables from our model as they added no value to it.

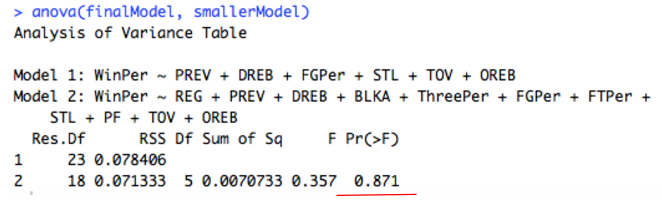
**V. Final Model**

We removed variables REG, BLKA, ThreePer, FTPer and PF to further simplify and improve our model. Our new model which we call *Final Model* is shown in Figure 8.

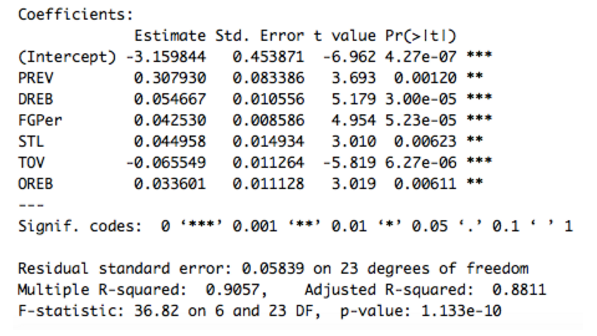
FinalM_report.PNG

**Figure 8.** The linear regression equation of the Final Model

We proceeded with an ANOVA test (as shown in Figure 9) to again test whether or not our new model is more adequate than the old one. The ANOVA table calculated a p-val of .871 which is pretty high. Therefore, we failed to reject the null. The results suggest that the the *Final Model* is better than the previous one and the variables that we removed had no significant values in predicting the WinPer. We decided to move forward with this model and do a regression analysis on it. The results of the regression table are shown in Figure 10.



**Figure 9.** The summary output of the ANOVA test that shows some of the predictor variables are insignificant.



**Figure 10.** The summary table of the reduced model is shown.

Solely looking at the regression table, our first impression was that this model was as good as it was going to get. All coeff. estimates are statistically significant and there’s still no trace of collinearity. In addition to the really low p-vals, we noticed that our coeff. of determination R2 is still relatively high at .90 which is good. Having eliminated collinearity and all useless variables, we were ready to check for check for the other assumptions.

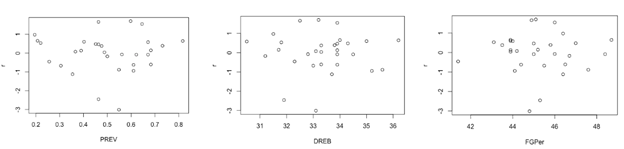
**VI. Testing Assumptions of Final Model:**

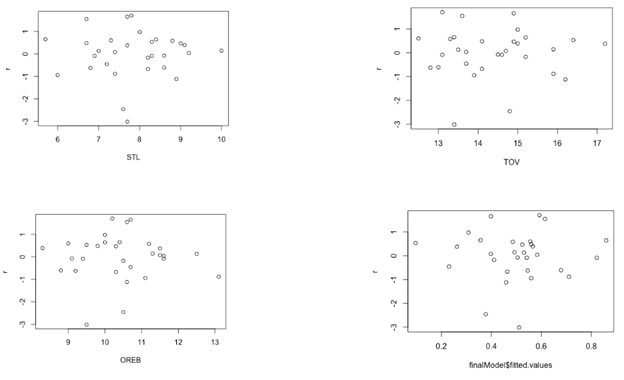
After we established a final model to run a regression on, we seek to test if the assumptions hold. The Variance Inflation Factors were calculated again for this model to test assumption number 3:



**Figure 11.** The output of running the VIF function for the final model is provided

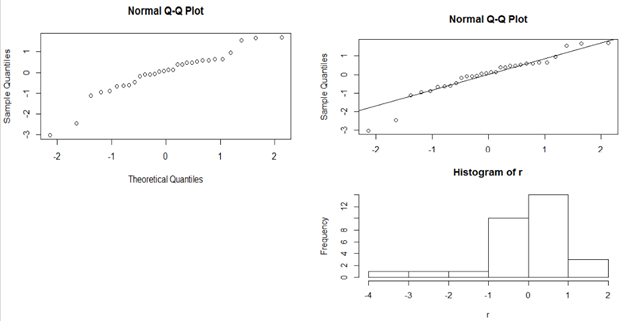
We can immediately see that none of these values are anywhere close to the value of 10. This provides sufficient evidence to make the claim that collinearity is not an issue for this model. In other words, the predictor variables are linearly independent. We can also complete an analysis of the relationship between the predictor variables and the response variables to check for assumption 1. After running the model, we graph the residuals against each predictor variable in order to come to conclusions about linearity. The graphs of the residuals against the predictor variables are provided in Figure 12. In general, we hope to see a “shotgun” type of distribution in the points on these graphs. For the most part, these points look distributed as such. However, the bottom two points seem to skew the randomness of the residuals. Assuming that these points are outliers (residual values very close to 3), we can conclude that the model holds the linearity assumption.





**Figure 12.** The graphs of the residuals from the final model plotted against each predictor variable along with a graph of the residuals with the fitted values

We can use these same graphs to come to conclusions about the assumption number 2 in regards to the errors. An argument could be made to note that there seems to be a nonconstant variance along the length of the x-axis. However, we note that in most of these graphs, the two bottommost points are clear outliers. There is not enough evidence to suggest a cone shaped pattern, which is one giveaway of heteroscedasticity. We hold that this model has constant variance (homoscedascticity), although it can be improved. These points cause the graphs to be compressed in the top portions. We will consider these two points for further analysis in the next steps we take to remedy the model. In general, we can also see that the mean of the residuals lies around 0. Another section of the second assumption deals with normality. We have provided the graphs in Figure 13 to supplement this discussion. In the Normal Q-Q Plot we have a fairly good graph because we ideally want all of our points to fall on the straight line. While this graph gives us some confidence that our errors are normally distributed, the tails of the graph diverse slightly. We hypothesize again that this might be due to the outliers in the graphs we saw for each predictor against the residuals. The histogram for the residuals doesn’t come off as a perfect bell curve; however we can clearly see normality around 0. Again, removal of some outliers may help remedy this.



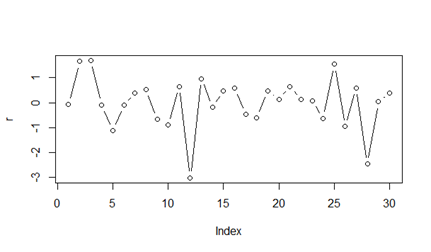
**Figure 13.** Graphs provided to visualize the normality of errors

The last part of the second assumption deals with the concept of autocorrelation. From the index plot in Figure 14, we can clearly see that there seems to be a decent amount of runs which suggest the absence of order based autocorrelation. Again, the data points of row 12 and 28 seem to be outliers. To confirm our conjecture, we run a Durbin-Watson test on this model. Our returned DW test statistic is 2.397 with a p-value of .8414. Our hypothesis test is set up as such:

Ho: p = 0

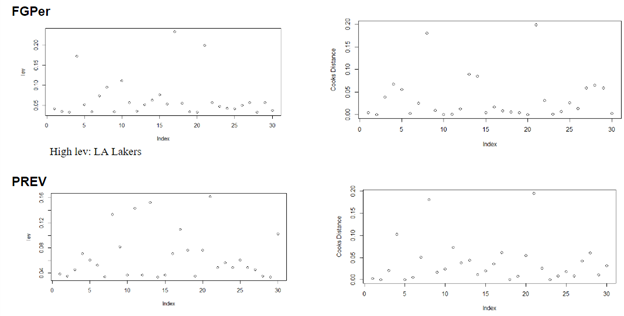
Ha: p > 0

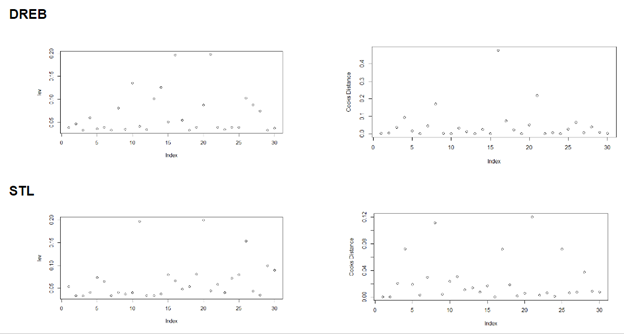
With our p-value, we fail to reject the null scenario. We note that we don’t have enough evidence to suggest that there is first order autocorrelation in this model. It is also useful to note that we did not encounter any evidences of higher order autocorrelation.

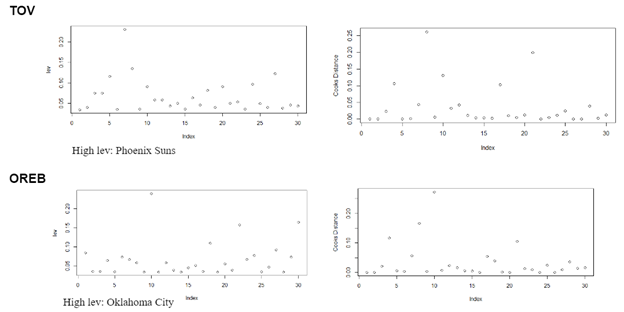


**Figure 14.** Index plot of residuals to check for error independence

In order to address the assumption 4, we consider points of leverage and influence. Points of high leverage are points that occur in the predictor variables, the x-space. These points tend to have small residual values as they pull the regression equation toward them. Points of high influence are points that cause significant shifts in the model should they be removed. Graphs to measure both influence and leverage are provided in Figure 15.

****

****

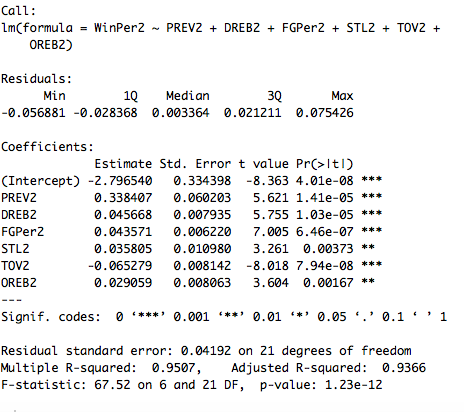
****

**Figure 15.** We outline both leverage plot and Cook’s distance plots to highlight any points of high leverage and any points of high influence.

We identify points of high leverage as any points that are above .2 in the leverage plots, and we identify points of high influence as any points that are above 1 in the Cooke’s distance plots. From our graphs, we can see that we have no outstanding points of high influence. In terms of high leverage, the Los Angeles Lakers stand out in FGPer graph, the Phoenix Suns stand out in TOV graph, and Oklahoma City Thunder stand out in OREB graph. We note these teams for future reference should we identify other problems. However, we decided that since there are no points of high influence in our analysis, that we will not remove these teams from the dataset. However, we do note that rows 12 and 28 from our dataset were consistently outliers in our standardized residual graphs. To test the impact of these teams, we next decided to see how our model changes when we remove them.

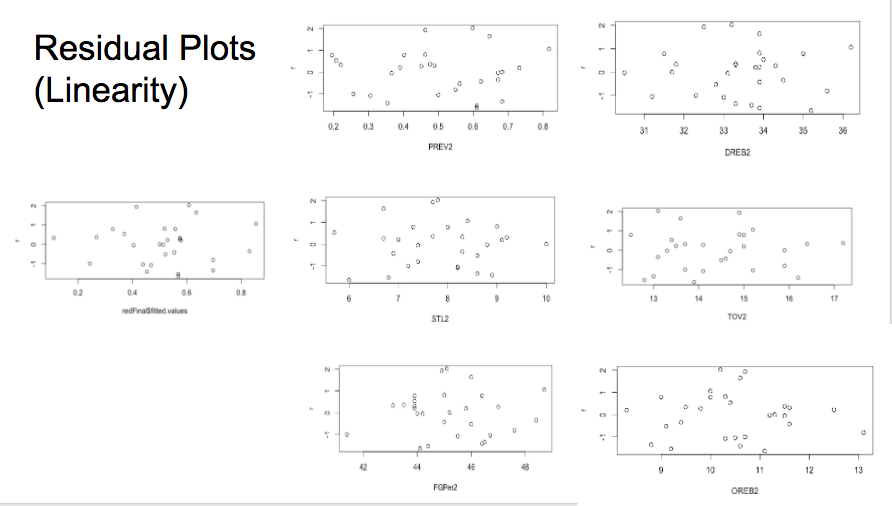
**VII. Final Model After Two Data Points Removed**

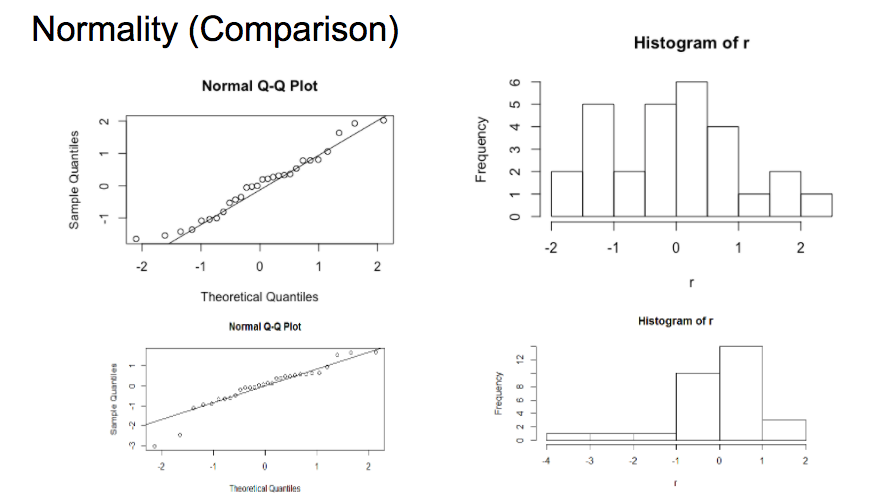
We removed rows 12 and 28 since they were the two major outliers found in our assumption testing and to see if our model could do a better job at predicting the response variable.



**Figure 16.** A summary of the final model with data points removed is shown above

As shown by the summary in our console, the significance level of estimate PREV2 has gone up to three stars from its original two stars which can be seen as an improvement. Value of Multiplied R - squared also gone up from 0.90 to 0.95, indicating that most of our data points are fitted closer to our fitted regression line.

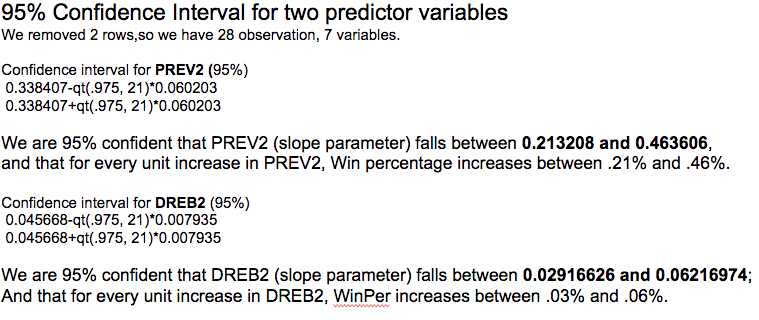
**Figure 17.** A collection of the Scatter Plots of the residuals against the predictors and the fitted values

**Figure 18.** Graphs to show the change in the normality of the data

After the removal of the rows 12 and 28, we do not see the two outliers in our residual plots any longer as shown in Figure 17. All data points are very much randomly distributed on the residual plots. Points are approximately centered around 0, having a mean of 0. There is no discernable patterns and the distribution is random and good. We no longer have points that have absolute values greater than three. The residual plots we have now is coherent in proving that there is good linearity between our predictor variables and our response variable, and that now the linearity is even stronger.

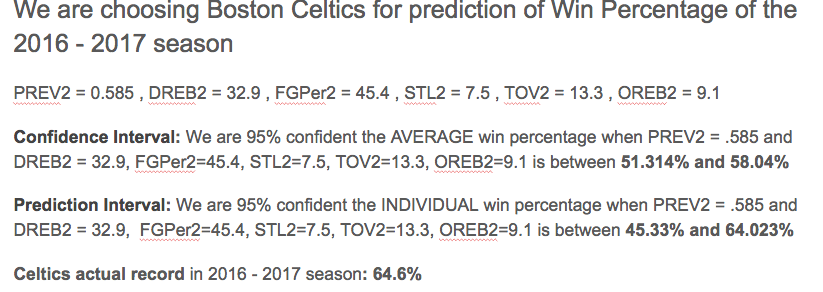
Afterwards, we did a normal Q-Q plot to check if normality has improved or not. We compared the new plot with the original one before removal. The improvement is quite significant, as the two points and the left side that was relatively far away from our fitted line. We can see that our response variable is more normally distributed than before. To assure that our normality of the response variable has improved, we also did a histogram of the r residuals and compared that to our old histogram of r. As shown in Figure 18, we barely have a bell- like curve in the old one. The graph was quite crooked with more hiner frequency of distribution on the right side and it is comparatively hard to tell if the residuals are really centered around 0. However, with our new histogram of r after data point removal, the distribution became more even on both sides above and below 0. The graph is comparatively more bell shaped and more symmetric, which indicates that the normality assumption for residuals hold.

After checking the plots that we felt most concerned about, we felt confident to call the Final Model our best model after the removals. Therefore, we decided to go with this model with two less observations and do both confidence and prediction intervals on this model.



**Figure 19.** A summary of both the confidence intervals conducted

For the confidence intervals, we found the sample means for the two variables, the z value of 95% confidence level and also variances of the two samples. Therefore we were able to construct the confidence intervals for the two variables. The set up and the analysis for both the intervals are shown in Figure 19.



**Figure 20.** A summary of both the confidence intervals conducted

As shown in Figure 20, we chose the Boston Celtics and compared our confidence interval and prediction interval of its win rates based on the statistics of the predictor variables of our model. After constructing the intervals, we compared the actual record to our intervals. From Figure 20 above, we can see that the value of the actual record is beyond the confidence interval and is 0.5% above the prediction interval. Since 0.5% is a fairly small percentage, we can say that the prediction interval of our finest model still offers some degree of confidence in predicting the win percentage of Boston Celtics in the next season.

1. <http://stats.nba.com/> [↑](#footnote-ref-0)
2. These two variables can also be found at <http://stats.nba.com/> [↑](#footnote-ref-1)