Joe Coyne

Dr. Bernhardt

STAT 8406: Regression Methods

7 May 2025

Regression Methods Final Report

Introduction

It's often said that defense wins championships, but how true is this really? My project aims to investigate this claim, and more broadly, to determine what factors best predict how successful a college basketball team will be in a particular season. As will be explained further in this report, my project aims to predict the total wins a college basketball team will have during a season. It's important for a team to know what types of rosters to build or what kind of schemes and strategies to employ, because at the end of the day, the goal is to win. And since rosters change so much from year to year, it becomes even more helpful to know what type of play best sets a team up to win.

The main goal of this project is to explore what factors are associated with the total number of wins a college basketball team has in a given season. Through this analysis, I also want to explore which of these factors have the most impact on the expected number of wins, either positively or negatively. One of the factors I will be exploring is the conference a team plays in. Part of this analysis will also be regarding whether the conference a team plays in has an impact on the total number of wins. While intuitive, the main hypothesis is that having a higher scoring percentage than scoring percentage allowed will be the most important factor in predicting wins. Another initial hypothesis is that the percentage a team scores on 3-point attempts will be important in the final model as basketball continues to emphasize 3-point shooting more than in the past. And finally, the interaction between a team's tempo, or how many possessions a team has per 40 minutes, and their shooting percentage will be significant. Teams that play at a higher pace tend to take more shot attempts, thus altering their percentage more than a team that takes fewer shots on average.

Data Gathering and Details

The dataset used for this project comes from the BartTorvik website and scrapes data from 2013-2014 season through the 2023-2024 season. This college basketball dataset includes 3,885 observations for each team and year combination. The variables included in this data are as follows: **W** (number of games won), which is the response variable in this analysis, **ADJOE** (adjusted offensive efficiency), **ADJDE** (adjusted defensive efficiency), **EFG_O** (effective field goal percentage shot), whose formula can be found in Formula 1 of the Appendix, **EFG_D** (effective field goal percentage allowed), **TOR** (turnover percentage allowed / turnover rate), **TORD** (turnover percentage committed / steal rate), **ORB** (offensive rebound rate), **DRB** (offensive rebound rate allowed), **FTR** (free throw rate, which is how often a team shoots free throws), **FTRD** (free throw rate allowed), **2P_O** (2 point shooting percentage), **2P_D** (2 point shooting percentage allowed), **3P_O** (3 point shooting percentage), **3P_D** (3 point shooting percentage allowed), **ADJ_T** (adjusted tempo, which is the number of possessions per 40 minutes), and **WAB** (Wins Above Bubble, which is how many more or fewer games a team has won against its schedule than a bubble-quality team would be expected to win). The data also has variables denoting the season the observation originated from, called **YEAR**, as well as how far the team made it in the NCAA Tournament, if they did, denoted **POSTSEASON**, what seed, if any, the team was in the NCAA Tournament, denoted **SEED**, and **BARTHAG,** a power rating metric. None of the Year, Postseason, Seed, or BARTHAG variables were used in this particular analysis, but they could be used in future analyses.

The largest cause for concern in terms of biases or cofounders is in the fact that the teams in the dataset will play each other during the season, violating the independence assumption. For example, if UConn typically allows a 3-point shooting percentage of 25% and Villanova shoots 27% from 3, but Villanova shoots 30% against them, both Villanova's 3-point shooting percentage and UConn's 3-point shooting percentage allowed percentages will increase.

Another potential bias lies in the fact that players will stay on a team for multiple years, so the year-to-year data for a particular team may also not be independent. However, there might not be as much cause for concern because nowadays in college basketball, there tends to be a lot of roster turnover and less players that stay on a team for multiple years.

Statistical Summaries

First, the response variable Wins was analyzed and diagnostic plots were made. As seen in Figure 1 of the Appendix, the histogram plot of Wins shows a normal distribution with no real skew. Looking at the boxplot in Figure 2 of the Appendix, we have further evidence of a normal distribution of the Wins variable. Wins has a minimum value of 0 and a maximum value of 38, with a mean of 16.08 and a median of 16. There are only two outliers at 38 wins (2015 Kentucky) and 37 wins (2017 Gonzaga), but there isn't any reason for concern, as the next highest value of Wins is 36.

Looking at the correlogram in Figure 1 of the Appendix, we see that out of all the numeric variables, ADJOE (Adjusted Offense), ADJDE (Adjusted Defense), BARTHAG (Power Rating), EFG_O (Effective Field Goal Percentage Shot), and EFG_D (Effective Field Goal Percentage Allowed) are most correlated with Wins. However, these are complex variables that are also highly correlated with the other predictors in the dataset. We see this further when calculating the VIF for a linear model with all of the variables in the dataset, as EFG_D has a VIF of 238.4 and EFG_O has a VIF of 154.8. ADJOE, ADJDE, and BARTHAG also have VIFs above 20, indicating that there is a lot of multicollinearity between these variables and the more standard statistics. WAB (Wins Above Bubble) also has a high VIF and is another win-type variable, so we won't use this in our models either.

Referring back to the correlogram in Figure 1 of the Appendix, we see that X2P_O, X2P_D, X3P_O, and X3P_D (2 or 3 point shooting percentage shot or allowed) are all highly correlated with Wins, so we have early indications that they will be significant predictors in a linear regression. Contrary to my initial hypothesis, ADJ_T (Adjusted Tempo) has little to no correlation with total wins, per this correlogram.

Next, the predictor variables were analyzed. Most of the variables had similar distributions to TOR (Turnover Rate) and ORB (Offensive Rebound Rate), whose histogram plots can be seen in Figure 4 of the Appendix. These variables all seem to have normal distributions and there is no cause for concern in using them to predict Wins. However, a select few predictor variables do have slight deviations from normal distributions, like FTR (Free Throw Rate), whose histogram plot can also be seen in Figure 4 of the Appendix. FTR, FTRD, and WAB all seem to have slightly right skewed data, but the skewness is so minimal that there is no reason for concern. The only variable that has more serious cause for concern is the Games variable. During the 2020-2021 college basketball season, many teams had a lot of their games cancelled due to COVID, so some teams only played a handful of games. Thus, the histogram of the Games variable seen in Figure 4 is extremely left skewed. To combat this, we removed any observations where the team didn't play at least 20 games. This removed 87 observations, with 86 of them coming from the 2020-2021 season. The only exception was Incarnate Word during their 2013-2014 season, where they 27 total games, but only 15 were against Division I opponents and the other 12 weren't recorded on the BartTorvik website. Looking at Figure 5, the updated Games histogram, we see a much more normal distribution of values, and there is no longer a cause for concern.

Finally, we look at the marginal relationships between the predictors and the response variable, Wins. Running simple linear regressions to predict Wins with each of the predictors, we found that the best 5 variables in predicting total wins were: **2P_O** (2 Point Shooting Percentage) with an Adjusted R^2 of .313, **2P_D** (2 Point Shooting Percentage Allowed) with an Adjusted R^2 of .2693, **3P_D** (3 Point Shooting Percentage Allowed) with an Adjusted R^2 of .2010, **TOR** (Turnover Rate) with an Adjusted R^2 of .196, and **CONF** (Conference) with an Adjusted R^2 of .1172. Figures 6-9 in the Appendix visualize these marginal relationships with total wins. We see that 2 Point Shooting Percentage has a positive linear relationship with Wins, whereas 2 Point Shooting Percentage Allowed, 3 Point Shooting Percentage Allowed, and Turnover Rate all have negative, linear relationships, with Turnover Rate having the steepest slope of all 4 predictors. Since Conference is a factor variable and can't be visualized in a scatterplot, Figure 10 shows a bar chart ranking the conferences by Mean Wins, with the Big 12, Big Ten, and ACC having the most and the SWAC, GWC, and Independent having the least.

Multiple Regression Model(s)

After doing some statistical summaries and basic analyses on the predictors, we put the following variables into an initial multiple linear regression model: CONF, TOR, TORD, ORB, DRB, FTR, FTRD, X2P_O, X2P_D, X3P_O, X3P_D, and ADJ_T. This resulted in an Adjusted R^2 of 0.8253 and a Global Hypothesis Test p-value < $2.2 * 10^{-16}$. All of the predictors were found the be statistically significant at the .05 level. We then ran best subset selection, forward selection, backward selection, and stepwise selection processes on this model both with alpha levels of .05 and trying to minimize the AIC, but this resulted in the same model, as seen below, where I(CONF) indicates the 33 indicator variables for the 34 conferences:

$$\boxed{\begin{aligned} \widehat{\boldsymbol{Total\,Wins}} = \mathbf{13.819} + \boldsymbol{I(CONF)} - \mathbf{0.938}\ (\boldsymbol{TOR}) + \mathbf{0.761}\ (\boldsymbol{TORD}) + \mathbf{0.372}\ (\boldsymbol{ORB}) \\ - \mathbf{0.329}\ (\boldsymbol{DRB}) + \mathbf{0.141}\ (\boldsymbol{FTR}) - \mathbf{0.146}\ (\boldsymbol{FTRD}) + \mathbf{0.549}\ (\boldsymbol{2P_O}) \\ - \mathbf{0.\ 533}(\boldsymbol{2P_D}) + \mathbf{0.540}\ (\boldsymbol{3P_O}) - \mathbf{0.547}\ (\boldsymbol{3P_D}) + \mathbf{0.058}\ (\boldsymbol{ADJ_T}) \end{aligned}}$$

The model performs well looking at the R^2 values and p-values, but looking at the Residuals vs Fitted plot from Figure 11 of the Appendix, there is a clear lack of constant variance, as there is a funnel-type shape that grows as the fitted values increase. The Normal Q-Q plot in Figure 12 of the Appendix has a left tail that trails a little from the line, but is not a huge concern compared to the constant variance issue. These suspicions are confirmed when the Studentized Breush-Pagan Test and Shapiro-Wilks Test are run, as we get p-values of $1.835 * 10^{-15}$ and $1.926 * 10^{-6}$, respectively. These results claim that both the constant variance assumption and normal residual assumptions, respectively, are not reasonable.

Before transforming the data, outlier checks are performed using Standardized Residuals, Studentized Residuals, PRESS statistics, R-student values, and Hat Values. When running these diagnostics, we see that there is one observation (2014-2015 NJIT) with a Hat Value of exactly 1, meaning that this is a perfect prediction and an extremely influential point. This observation was removed and the model was refit to see if the outlier removal helped any of the assumption checks, however the plots and tests returned almost identical results.

The next step in trying to fix the assumption checks was to transform the response variable, Wins. Since the Residuals vs Fitted plot was a funnel shape, both a log and square root transformation were conducted. First, using the log transform, we ran the same predictors from the original model through stepwise selection, and all of the variables returned significant. However, when running the full model predicting the log of total wins, the Adjusted R^2 decreased to 0.786, and both the assumption check plots got worse. The Residuals vs Fitted plot in Figure 13 of the Appendix is now a negative quadratic shape that still has a funnel shape, but one that gets smaller as the fitted values increase. The Normal Q-Q plot in Figure 14 of the Appendix looks a lot worse as both tails fall below the line, and the left tail strays a lot more. Both the Studentized Breusch-Pagan Test and the Shapiro-Wilks Test have p-values less than $2.2 * 10^{-16}$, indicating stronger evidence against constant variance and normality of errors compared to the original model.

The square root transformation was then performed on the response variable, Wins. All of the predictors from the first two models remained significant through the stepwise selection when predicting the square root of Wins. This model performed the best of the three models that had been run, returning the largest Adjusted R^2 of 0.8327 and best assumption check plots. Looking at the Residuals vs Fitted Plot in Figure 15, there is finally constant spread about 0 and no real trend or concerning shapes. This is also shown by the Studentized Breusch-Pagan Test, as the p-value is only 0.0027, much larger than before and much closer to retaining the null hypothesis that there is constant variance. The Normal Q-Q plot in Figure 16 is still worse than the plot from the untransformed model, confirmed with a Shapiro-Wilks Test p-value of less than $2.2 * 10^{-16}$, but it is better than the log transformed plot, since while both tails fall below the line, they fall closer than before.

In order to further improve the constant variance and rectify the normality assumption, the residuals from the square root model were plotted against the predictors to see if any quadratic terms were needed. Of all the predictors currently in the model, **2P_O** (2 Point Shooting Percentage), **3P_O** (3 Point Shooting Percentage), and **3P_D** (3 Point Shooting Percentage Allowed) all had quadratic trends with relation to the square root residuals, as seen in Figures 17, 18, and 19, respectively. To test if these quadratic terms ($2P_O^2$, $3P_O^2$, $3P_D^2$) were significant in predicting the square root of Wins, along with the predictors already in the model, another stepwise selection process was run, and all predictors including the quadratic terms were statistically significant. This updated square root model resulted in an Adjusted R^2 of 0.8361 and similar Residuals vs Fitted and Normal Q-Q plots, as seen in Figures 20 and 21, respectively. Some interaction terms were also tested to see if there were any significant interactions between predictors, such as 2 Point Shooting Percentage vs 3 Point Shooting Percentage, Tempo vs 2 Point Shooting Percentage Allowed, and Tempo vs 3 Point Shooting Percentage Allowed. However, these added terms made some of the basic statistics insignificant during the stepwise selection and didn't improve the Adjusted R^2 by much, if at all in some cases. There was also no major improvement to the Residuals vs Fitted plots or Normal Q-Q plots, so the final model used for this project is the original square root model, since it is the easiest to interpret. The final model formula is listed below:

$$\boxed{\begin{aligned}\widehat{\boldsymbol{Total\,Wins}} = [\ & \mathbf{3.69} + \boldsymbol{I(CONF)} - \mathbf{0.127}\ \boldsymbol{(TOR)} + \mathbf{0.101}\ \boldsymbol{(TORD)} + \mathbf{0.047}\ \boldsymbol{(ORB)} \\ & - \mathbf{0.046}\ \boldsymbol{(DRB)} + \mathbf{0.019}\ \boldsymbol{(FTR)} - \mathbf{0.018}\ \boldsymbol{(FTRD)} + \mathbf{0.073}\ \boldsymbol{(2P_O)} \\ & - \mathbf{0.\ 069}\boldsymbol{(2P_D)} + \mathbf{0.069}\ \boldsymbol{(3P_O)} - \mathbf{0.072}\ \boldsymbol{(3P_D)} + \mathbf{0.008}\ \boldsymbol{(ADJ_T)}\ ]^2\end{aligned}}$$

The indicator variables for the different conferences can be found in Figure 22 of the Appendix.

Finally, to validate the final model, we predict the total number of wins for the 2024-2025 Villanova team, who had the following statistics: **TOR** = 15.3, **TORD** = 15.8, **ORB** = 30.3, **DRB** = 27.7, **FTR** = 29.2, **FTRD** = 30.4, **2P_O** = 51.3, **2P_D** = 50.5, **3P_O** = 39.0, **3P_D** = 34.5, and **ADJ_T** = 63.5. The model above results in a predicted 20.344 wins for the 2024-2025 Villanova team. Villanova actually won 21 games this season, so the residual was only 0.6564, indicating that the model did well in predicting the 2024-2025 Villanova win total.

Conclusion and Discussion

Overall, this analysis found that of all the standard basketball statistics, turnover and steal rate changed the expected number of wins for a team the most. Conversely, adjusted tempo has a very minimal impact on total expected wins. And finally, while the conference a team plays in has a small impact on the total expected wins, being in a power conference is associated with a decrease in total expected wins. This is in contrast to the initial simple linear regressions we ran with only one predictor at a time, where the power conferences such as the ACC, Big Ten, Big 12, etc. had positive Win coefficients. This might be due to the fact that better conferences tended to have more wins overall, but within these conferences, each team is expected to have less total wins because they are playing tougher opponents than those in lower conferences.

If I were able to run this analysis again, I would have liked to include some individual player data too. Teams with star players that can take over games during crunch time will tend to win more games, but this isn't accessible just from the team statistics. I would have also liked to compare yearly models to see what variables remain important from year to year in terms of predicting Wins and which variables change. The game of basketball is always changing, so generalizing trends from the last decade plus may not give the most accurate representation of the current models and trends.

In an effort to further improve the constant variance and normality of errors assumptions, I would have liked to run both a Ridge Regression and a LASSO model on the data. Some further questions I would like to answer with more analyses are how this data does in predicting if a team will make the NCAA Tournament, and if so, how far they will advance. I would also like to see how much what conference a team plays in affects how far they advance in the NCAA Tournament. Overall, BartTorvik does a very good job in displaying very relevant data and variables in terms of determining the success of a team.

Appendix

Formula 1:

$$eFG\% = \frac{(FGM + (0.5 X 3PTM))}{FGA}$$

FGM = Field Goals Made (2PT and 3PT)

3PTM = Three Point Goals Made

FGA = Field Goals Attempted (2PT and 3PT)

Figure 1:

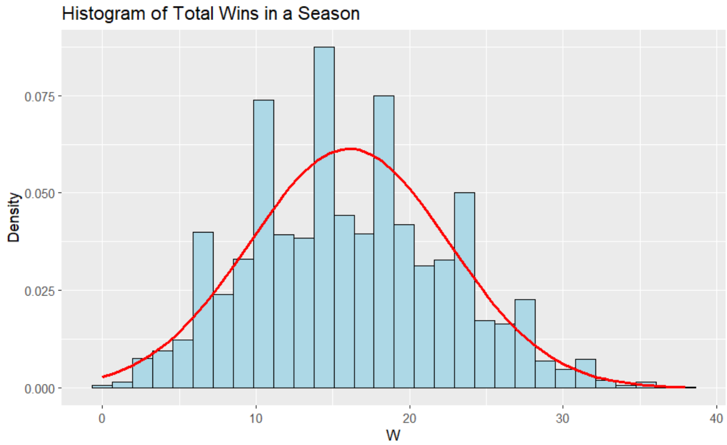


Figure 2:

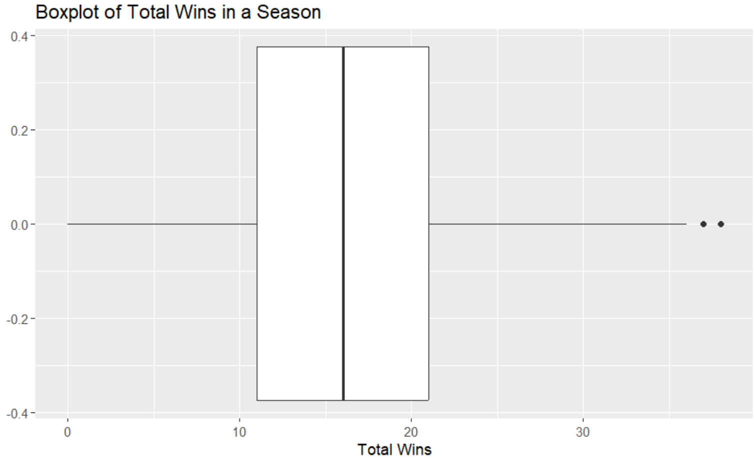


Figure 3:

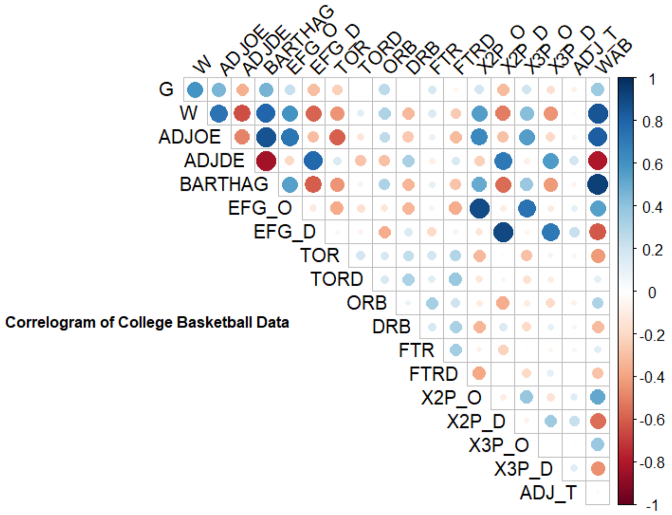


Figure 4:

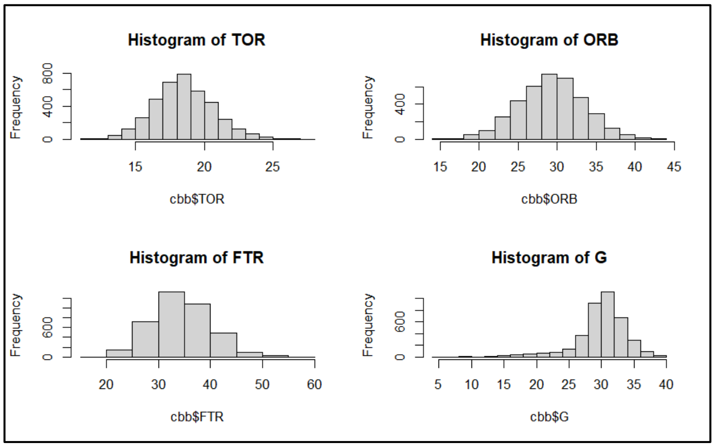


Figure 5:

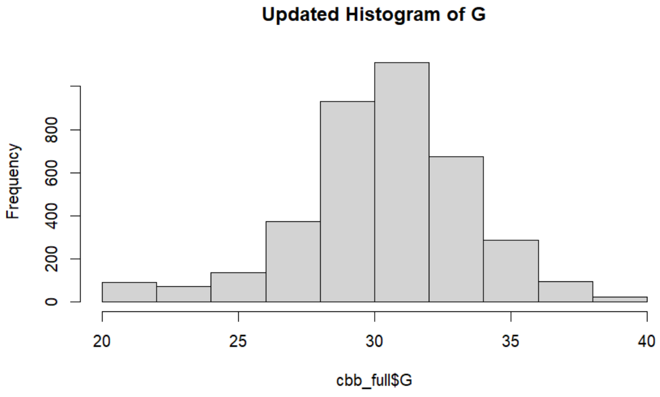


Figure 6:

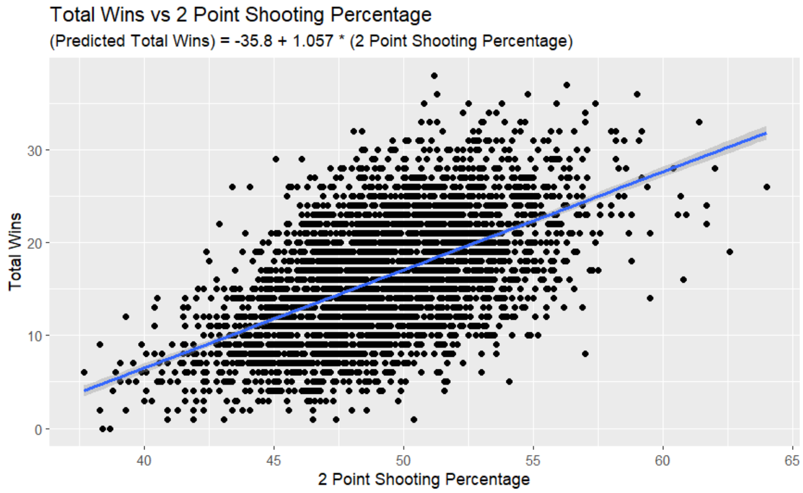


Figure 7:

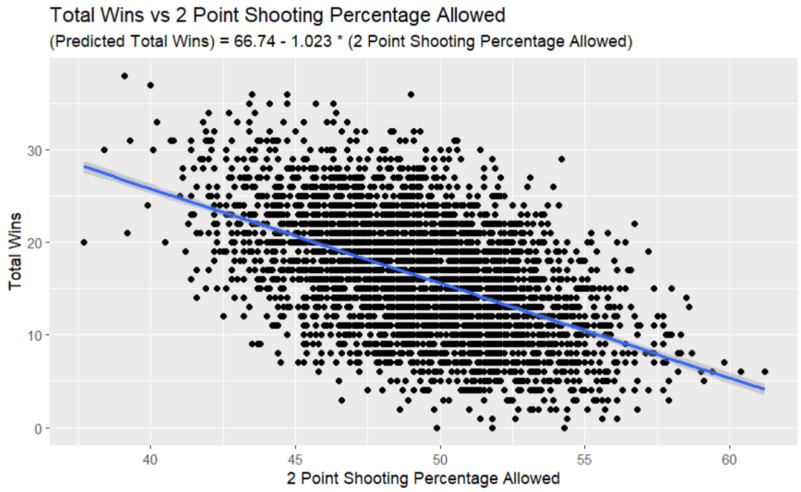


Figure 8:

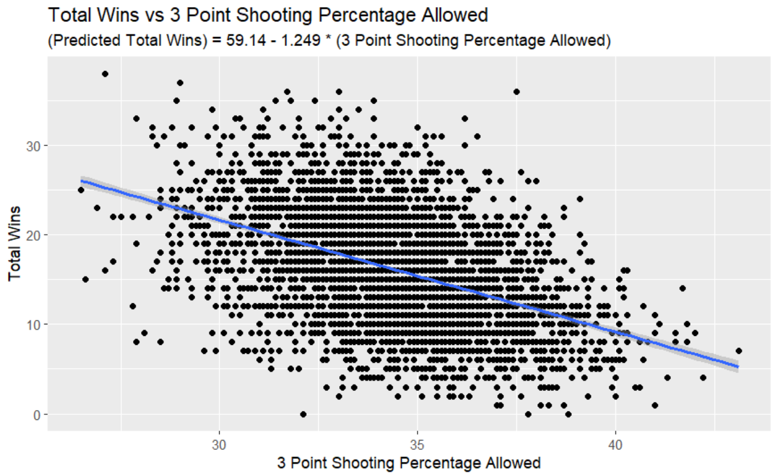


Figure 9:

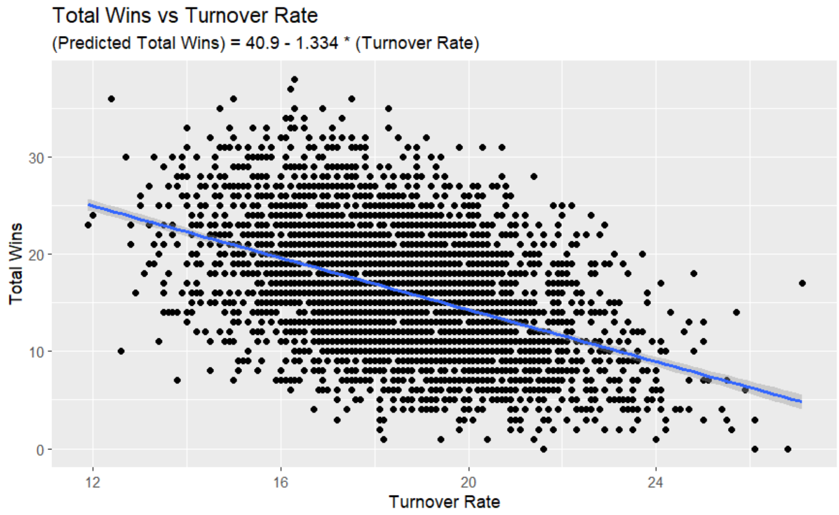


Figure 10:

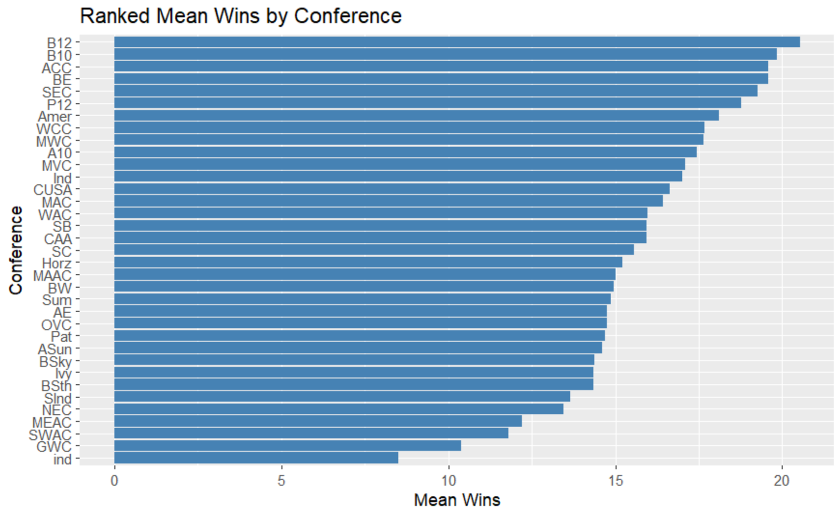


Figure 11:

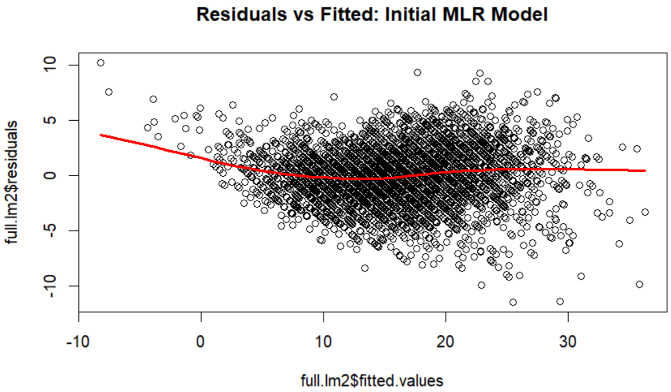


Figure 12:

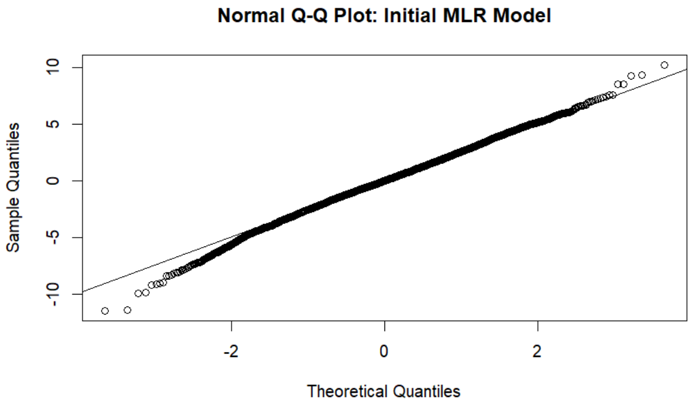


Figure 13:

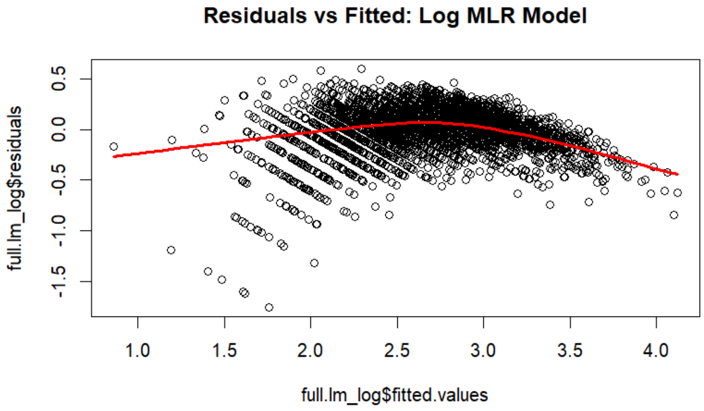


Figure 14:

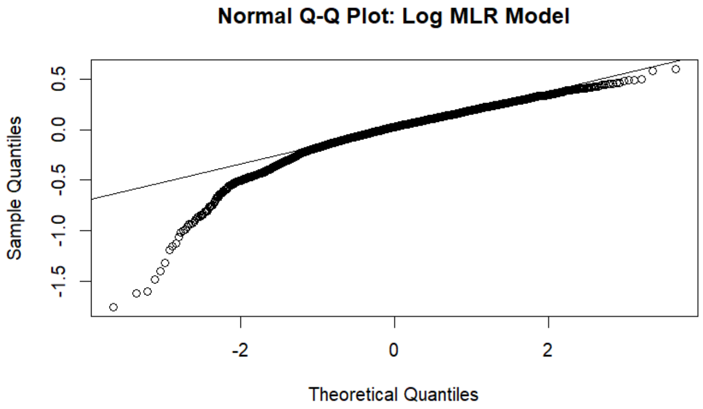


Figure 15:

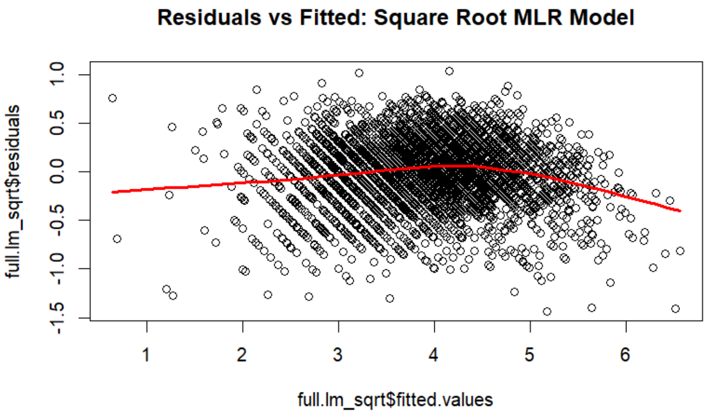


Figure 16:

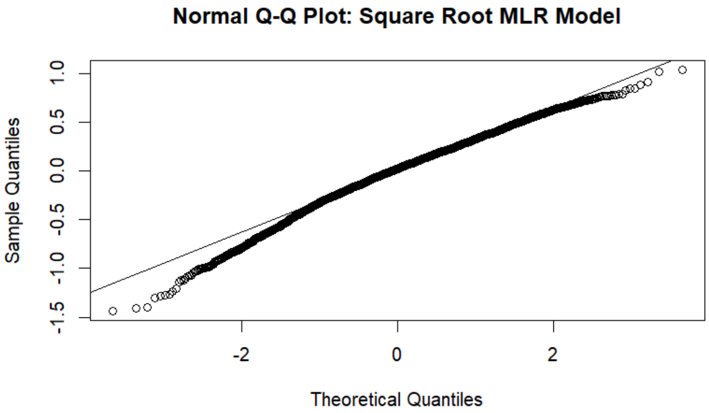


Figure 17:

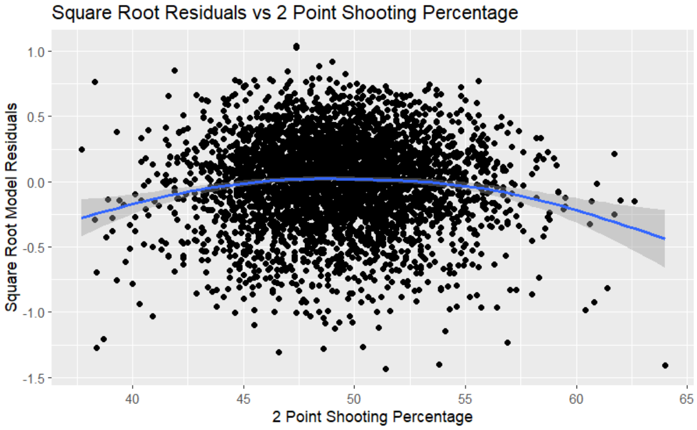


Figure 18:

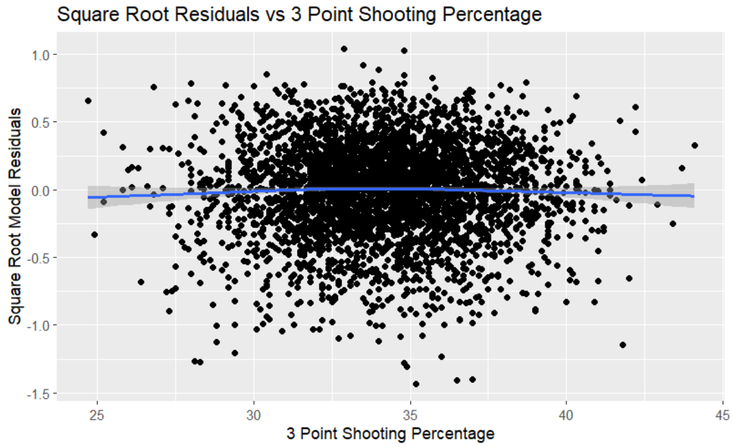


Figure 19:

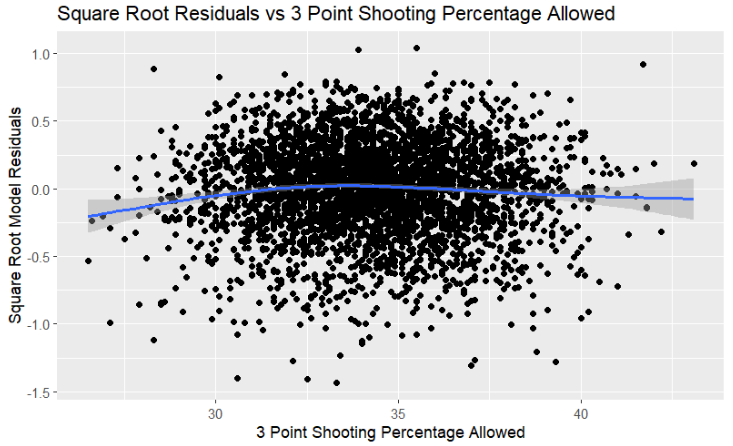


Figure 20:

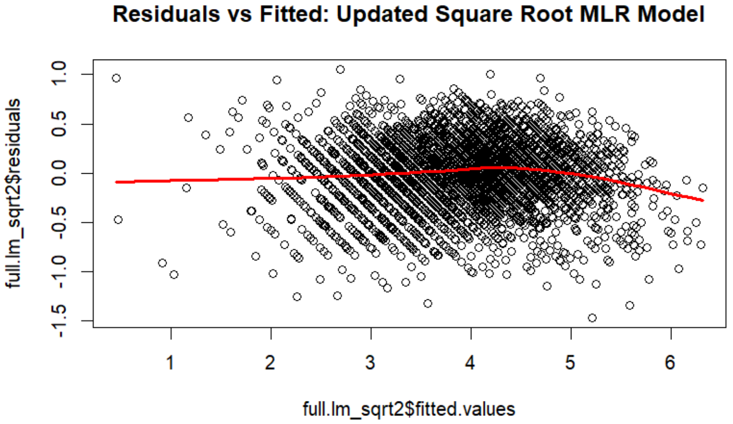


Figure 21:

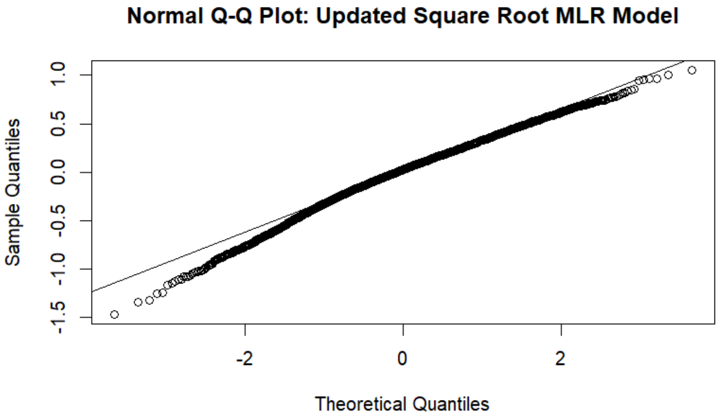
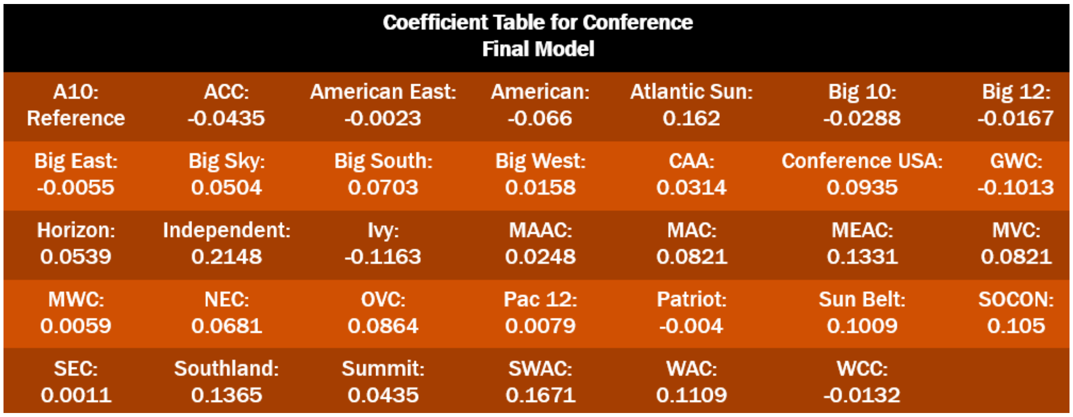


Figure 22:



Coefficient Table for Conference
Final Model

A10: Reference	ACC: -0.0435	American East: -0.0023	American: -0.066	Atlantic Sun: 0.162	Big 10: -0.0288	Big 12: -0.0167
Big East: -0.0055	Big Sky: 0.0504	Big South: 0.0703	Big West: 0.0158	CAA: 0.0314	Conference USA: 0.0935	GWC: -0.1013
Horizon: 0.0539	Independent: 0.2148	Ivy: -0.1163	MAAC: 0.0248	MAC: 0.0821	MEAC: 0.1331	MVC: 0.0821
MWC: 0.0059	NEC: 0.0681	OVC: 0.0864	Pac 12: 0.0079	Patriot: -0.004	Sun Belt: 0.1009	SOCON: 0.105
SEC: 0.0011	Southland: 0.1365	Summit: 0.0435	SWAC: 0.1671	WAC: 0.1109	WCC: -0.0132	