Slaads – Final Report

Prepared by:	Kevin Gasiorowski Piotr Senkow Srikanth Nanduri Vineet Dcunha ~~Riley Edwards~~ (*dropped class*)
Date (mm/dd/yyyy):	03/15/2020

Introduction:

Betting on NCAA men's basketball is extremely popular. According to Reuters (1), about 47 million people bet a combined $8.5 billion on the NCAA tournament tournament in 2018. But to get into the tournament requires a successful regular season that is gauged by overall winning percentage. In this study, we performed an exhaustive analysis on how well *basic* and *advanced* regular season statistics from 2015 through 2019 predict overall regular season winning percentage for NCAA Division I Men's Basketball teams.

Our goal of this study was to determine which statistics most significantly predict winning percentage. These statistics might then provide assistance and guidance to teams' ownership, management, coaches, and their fans, such as when managers are assembling a team or when fans are betting on games.

The remainder of this document is organized as follows:

- Description of data sources used and preparation performed
- Summary of analysis techniques and metrics used throughout this research
- Description of the individual studies performed against both the *basic* and *advanced* statistics
- Details on the approach and findings of each individual study
- Executive summary and conclusion of our findings

Data Collection and Preparation:

For this study we collected *basic* and *advanced* statistics for all NCAA Division I men's regular season basketball games from 2015 through 2019.

Each NCAA Division I men's basketball team's season statistics from 2015 through 2019 are considered a unique observation. While most teams/universities have observations for all 5 seasons, some universities joined or dropped from Division I during this time, and our collection takes this into account and only collects observations for when universities were officially part of Division I. There are a total of 1752 unique observations used in our study.

We make the distinction between *basic* and *advanced* statistics as follows:

- Basic statistics: Traditional statistics that typically include the summary or average of a single observation typically per game, such as points per game, rebounds per game, etc.
- Advanced statistics: Modern statistics that typically require a formula that combines many of the basic statistics into a single rating where a key tenant is that the statistics are evaluated over possessions, not games.

Basic statistics

All *basic* statistics were retrieved from the following site:

http://web1.ncaa.org/stats/StatsSrv/rankings

There are a total of 22 variables for the *basic* statistics that were considered for this study. The categorical fields of Name and Year uniquely identify the observations, Win Pct. is the response variable, and all remaining 19 numerical variables are considered the independent variables in this study.

The data source also contains many additional totals statistics (total points, total assists, etc.). However, Because all teams did not play the same amount of regular season games each year, we focused on the per-game average and percentage statistics instead of the totals to help mitigate the effects of an uneven regular-season schedule.

The following variables were accumulated for this study of the *basic* statistics:

Name of variable	**Type**	**Description**
Name	Categorical	Name of team (university)
Year	Categorical	Year of the regular season the remaining statistics are associated with (range 2015 - 2019)

Win Pct. (winning percentage)	Numeric	(total wins / total losses) for the team in the associated regular season
PPG (points per game)	Numeric	Total points scored by team / number of games played
OPP PPG (opponent's points per game)	Numeric	Total points scored by opposing teams / number of games played
SCR MAR (scoring margin)	Numeric	(Total Team Points Per Game - total opponent's points per game) / number of games
FG% (field goal percentage)	Numeric	Total field goals made / total field goals attempted
OPP FG% (opponent's field goal percentage)	Numeric	Total field goals made by opponents / total field goals attempted by opponents
3PG (3-point field goals made per game)	Numeric	Total 3-point field goals made by team / number of games
3FG% (3-point field goal percentage)	Numeric	Total number of 3-point field goals made / total number of 3-point field goals attempted
Opp 3P FG% (opponent's 3-point field goal percentage)	Numeric	Total number of 3-point field goals made by opponents / total number of 3-point field goals attempted by opponents
FT% (free throw percentage)	Numeric	Total number of free throws made by team / total number of free throws attempted by team.
RPG (rebounds per game)	Numeric	Total number of rebounds by team / number of games team played
OPP RPT (opponent's rebounds per game)	Numeric	Total number of rebounds by opponents / number of games team played
REB MAR (rebound margin)	Numeric	Number of rebounds team averages per game / number of rebounds opponents average per game
APG (assists per game)	Numeric	Total number of assists by team / number of games team played

Assist TO Ratio (assist to turnover ratio)	Numeric	Total number of assists by team / total number of turnovers by team
BKPG (blocks per game)	Numeric	Total number of blocks by team / number of games team played
StealsPG (steals per game)	Numeric	Total number of steals by team / number of games team played
TOPG (turnover per game)	Numeric	Total number of turnovers made by team / number of games team played
TO Margin (turnover margin)	Numeric	Average number of turnovers by team / average number of turnovers by opponents
Personal Founs PG (personal fouls per game)	Numeric	Total number of personal fouls committed by team / total number of games team played

Data Cleansing/Preparation

The collected data was mostly clean. One difficulty discovered is that for each year, we discovered on average 2 teams did not have a recorded value for TO margin (this data was not collected for a few universities). Therefore, we imputed the TO margin for these missing values using the mean value.

Advanced statistics

All *advanced* statistics were retrieved from the following site:

https://www.sports-reference.com/cbb/seasons/

There are a total of 27 variables for the *advanced* statistics that were considered for this study. The categorical fields of Name and Year uniquely identify the observations, Win Pct. is the response variable, and all remaining 24 numerical variables are considered as the independent variables in this study.

Variable	**Type**	**Description**
Name	Categorical	Name of team (university)
Year	Categorical	Year of the regular season the remaining statistics are associated with (range 2015 - 2019)

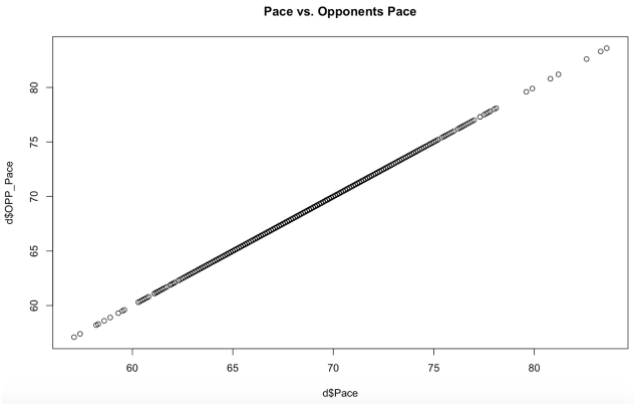
Win Pct. (winning percentage)	Numeric	(total wins / total losses) for the team in the associated regular season
Pace	Numeric	An estimate of possessions per 40 minutes
OPP_Pace	Numeric	An estimate of opponent possessions per 40 minutes
ORtg (Offensive Rating)	Numeric	An estimate of points scored per 100 possessions
FTr (Free Throw Attempt Rate)	Numeric	Number of FT attempts per FG attempt
3Par (3-Point Attempt Rate)	Numeric	Percentage of FG attempts from 3-point range
TrueShootingPer (True Shooting Percentage)	Numeric	A measure of shooting efficiency that takes into account 2-point FG, 3-point FG, and free throws.
TotalReboundPer (Total Rebound Percentage)	Numeric	An estimate of the percentage of available rebounds grabbed by the team
AssistPer (Assist Percentage)	Numeric	An estimate of the percentage of field goals assisted
StealPer (Steal Percentage)	Numeric	An estimate of the percentage of opponent possessions that ended with a steal
BolckPer (Block Percentage)	Numeric	An estimate of the percentage of opponents 2-point field goals that were blocked
EffectiveFGPer (Effective Field Goal Percentage)	Numeric	Field goal percentage adjusted for difference in 3-point shots and 2-point shots
TurnoverPer (Turnover Percentage)	Numeric	An estimate of turnovers per 100 plays
OReboundPer (Offensive Rebounds Percentage)	Numeric	An estimate of the percentage of available offensive rebounds grabbed
FT_Per_FGA	Numeric	Free throws per field goal attempt
OPP_3Par (3-Point Attempt Rate)	Numeric	Opponents percentage of FG attempts from 3-point range

OPP_TrueShootingPer (True Shooting Percentage)	Numeric	A measure of opponents shooting efficiency that takes into account 2-point FG, 3-point FG, and free throws.
OPP_TotalReboundPer (Total Rebound Percentage)	Numeric	An estimate of the opponents percentage of available rebounds grabbed by the team
OPP_AssistPer (Assist Percentage)	Numeric	An estimate of the opponents percentage of field goals assisted
OPP_StealPer (Steal Percentage)	Numeric	An estimate of the opponents percentage of their opponents possessions that ended with a steal
OPP_BlockPer (Block Percentage)	Numeric	An estimate of the opponents percentage of their opponents 2-point field goals that were blocked
OPP_EffectiveFGPer (Effective Field Goal Percentage)	Numeric	Opponents field goal percentage adjusted for difference in 3-point shots and 2-point shots
OPP_TurnoverPer (Turnover Percentage)	Numeric	An estimate of opponents turnovers per 100 plays
OPP_OReboundPer (Offensive Rebounds Percentage)	Numeric	An estimate of the opponents percentage of available offensive rebounds grabbed
OPP_FT_Per_FGA	Numeric	Opponents free throws per field goal attempt

Data Cleansing

The collected data was clean with no missing or incorrect values, therefore no additional effort was required to cleanse the advanced statistics.

We did discover that the independent variables Pace and OPP_Pace have a perfect positive correlation (correlation coefficient of 1) as demonstrated in the plot below. This high degree of collinearity is problematic for a linear regression model, therefore, we have dropped the OPP_Pace variable from this study and kept Pace.



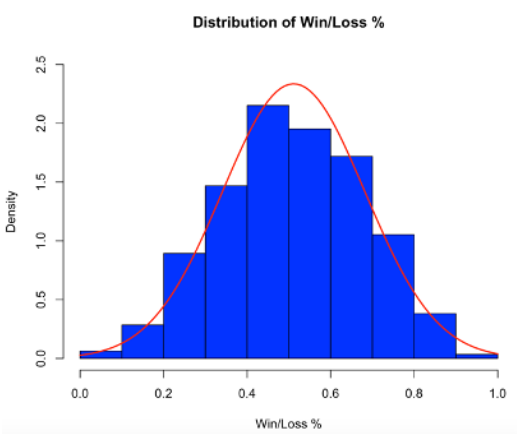

```
> cor(d$Pace, d$OPP_Pace)
[1] 1
```

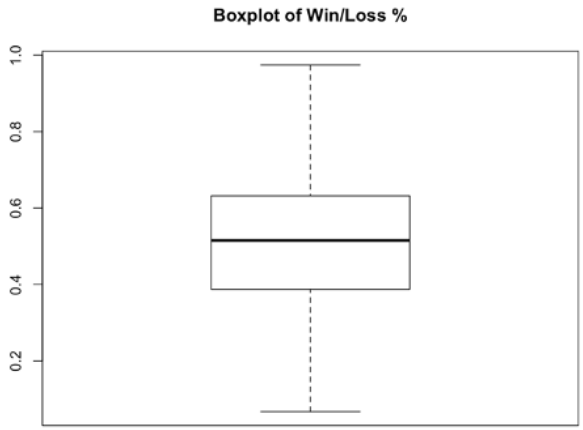
Data Analysis

We have compiled team statistics using domain knowledge that we believe are likely to influence a team's winning percentage (19 basic statistics and 23 advanced statistics). The remainder of this section analyzes the response and independent variables to validate our selection was appropriate for this study.

Response variable: Win/Loss Percentage

We begin this study by first examining the overall distribution of our target variable, **Win/Loss %** across all rows of data for the regular season.





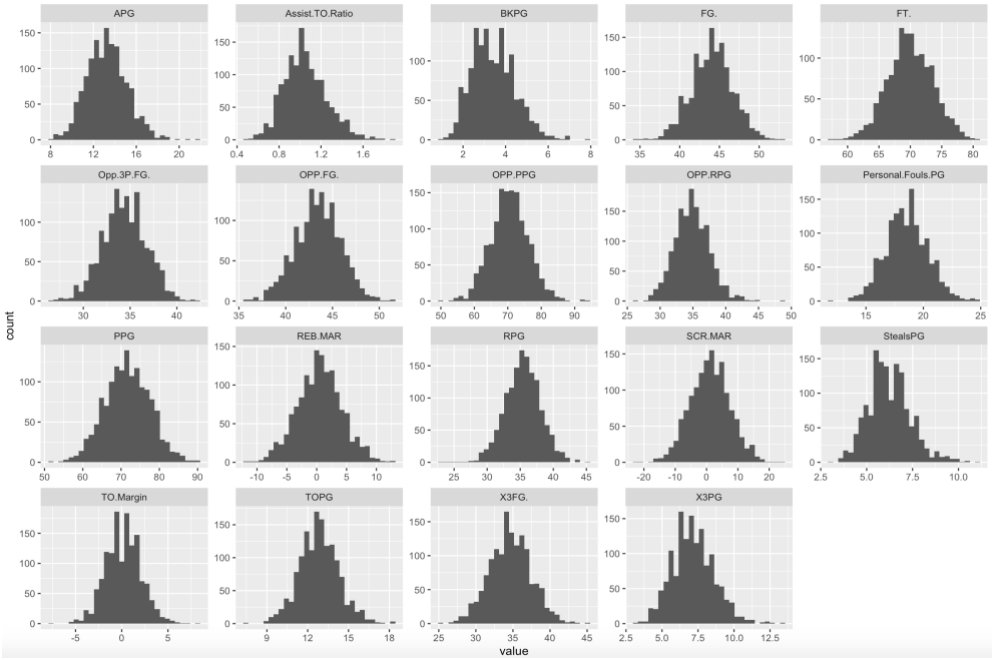

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0670  0.3870  0.5150  0.5114  0.6320  0.9740
```

From the figures above, we can see the overall distribution of Win/Loss % is fairly normal, with a minimum of 6.7%, a maximum of 97.4%, and no outliers present. The standard deviation of Win/Loss % is 0.170794.

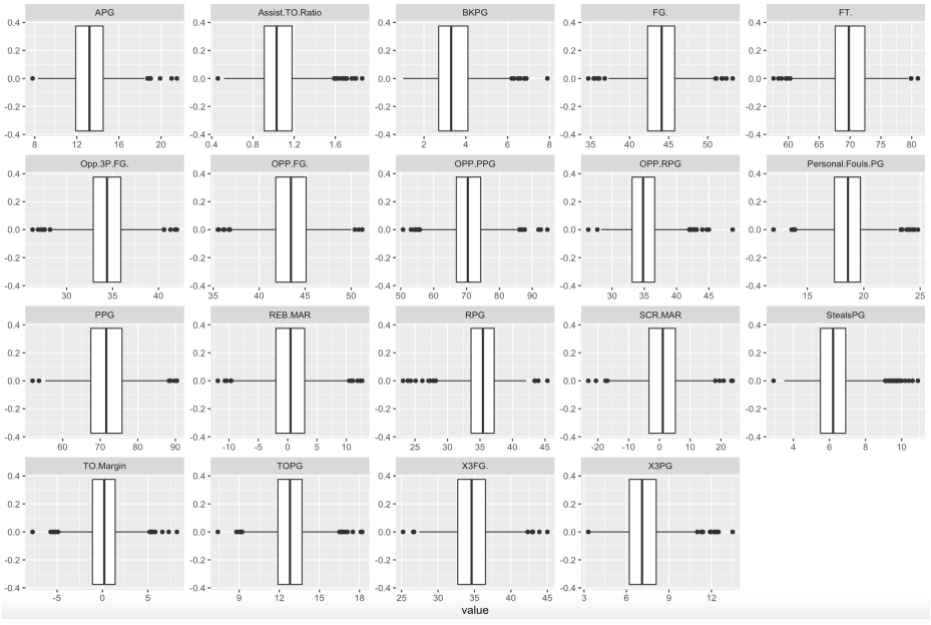
Because the dependent variable is normally distributed, we do not perform any additional transformation of Win/Loss % at this time.

Analysis of Basic Statistics

Evaluating the distribution of all independent *basic* variables, we find that all are normally distributed, therefore no further transformations were attempted to change the distribution prior to analysis.



Through boxplots, we also examined outliers for each independent *basic* variable. We found that outliers do exist for each independent variable. However, as those outliers may represent teams with the higher or lower winning percentage, we do not perform any additional handling of the outliers prior to analysis as they are likely significant to our analysis.

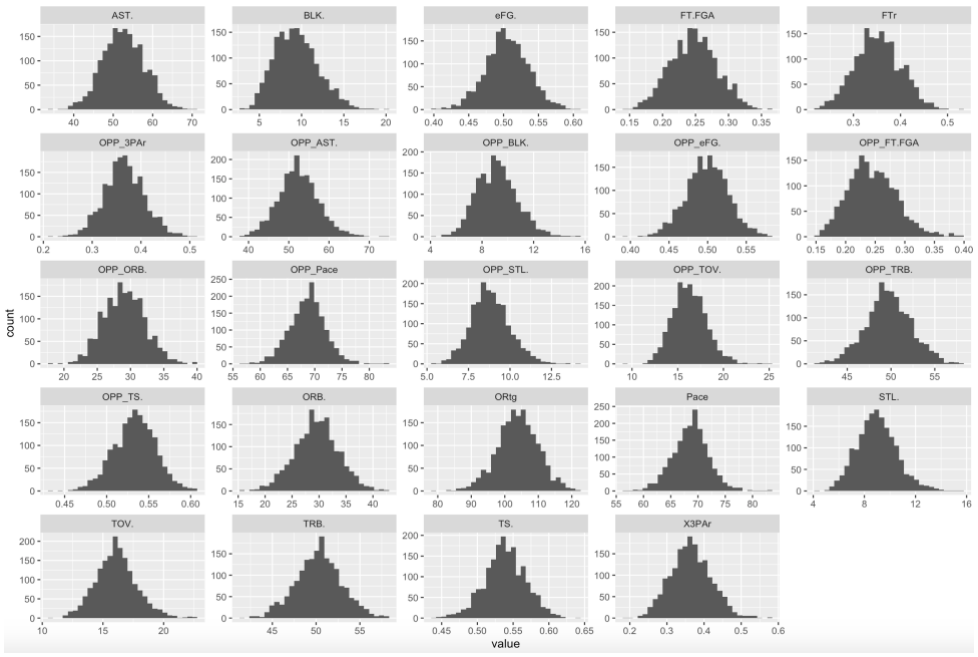


Examining the correlations between the basic independent variables and the target, we find a number of strong and moderate positive and negative relationships, indicating that we have a potentially useful set of variables to analyze.

	Win.Pct.
Win.Pct.	1.00000000
PPG	0.53205738
OPP.PPG	-0.01669702
SCR.MAR	0.12372616
FG.	0.63922270
OPP.FG.	-0.61391376
X3PG	0.22452638
X3FG.	0.43538656
Opp.3P.FG.	-0.47957227
FT.	0.25747807
RPG	0.16760135
OPP.RPG	-0.44815191
REB.MAR	0.59006617
APG	0.49737519
Assist.TO.Ratio	0.61650476
BKPG	0.35625838
StealsPG	0.22219103
TOPG	-0.45512132
Personal.Fouls.PG	-0.23476714

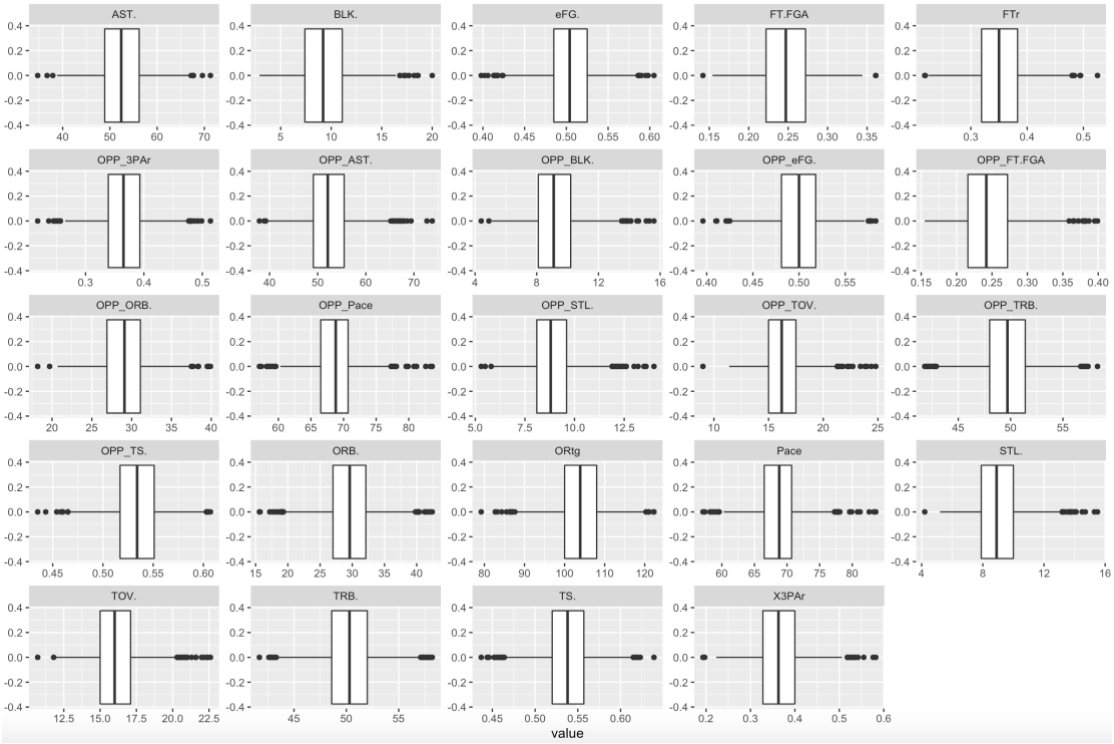
Analysis of Advanced Statistics

Evaluating the distribution of all independent *advanced* variables, we find that all are normally distributed, therefore no further transformations were attempted to change the distribution prior to analysis.



Through boxplots, we also examined outliers for each independent *advanced* variable. We found that outliers do exist for each independent variable. However, as those outliers may represent teams with the higher or lower winning percentage, we do not perform any

additional handling of the outliers prior to analysis as they are likely important to our analysis.



Examining the correlations between the *advanced* independent variables and the target, we find a number of strong and moderate positive and negative relationships, indicating that we have a potentially useful set of variables to analyze.

	[,1]
Pace	-0.06429262
ORtg	0.78248802
FTr	0.14810170
X3PAr	0.04090476
TrueShootingPer	0.62685167
TotalReboundPer	0.57706399
AssistPer	0.25458589
StealPer	0.22939252
BlockPer	0.35610224
EffectiveFGPer	0.60451834
TurnoverPer	-0.48977648
OReboundPer	0.28881899
FT_Per_FGA	0.23542248
OPP_3PAr	-0.07014372
OPP_TrueShootingPer	-0.63767314
OPP_TotalReboundPer	-0.57708802
OPP_AssistPer	-0.33175728
OPP_StealPer	-0.38232828
OPP_BlockPer	-0.24473852
OPP_EffectiveFGPer	-0.61494163
OPP_TurnoverPer	0.14991099
OPP_OReboundPer	-0.34665452
OPP_FT_Per_FGA	-0.29636836

NOTE: Due to the large numbers of variables being considered, we perform a further analysis of the relationship between independent variables in the Individual Studies section.

Model Building

We aim to find a way to accurately predict the highest winning percentage of all NCAA Division I teams in the regular season that are therefore likely to make the tournament.

This project attempts to determine which basic and advanced regular season statistics most strongly predict overall regular season winning percentage for NCAA Division I Men's Basketball teams. We considered independent variables featuring team offensive statistics, defensive statistics, and combination statistics.

The models were created using the Win Percentage as the response variable.

The models were built using the first order, second order and interaction terms as defined below:

- **Linear regression:** Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y.
- **Second order regression:** The model is simply a general linear regression model with k predictors raised to the power of i where i=1 to k. A second order (k=2) polynomial forms a quadratic expression (parabolic curve), a third order (k=3) polynomial forms a cubic expression and a fourth order (k=4) polynomial forms a quartic expression.
- **Interaction terms:** In regression, an interaction effect exists when the effect of an independent variable on a dependent variable changes, depending on the value(s) of one or more other independent variables.

The final models were validated against the forward selection and backward elimination methods, defined as follows:

- **Forward selection:** Forward selection begins with an empty equation. Predictors are added one at a time beginning with the predictor with the highest correlation with the dependent variable. Variables of greater theoretical importance are entered first. Once in the equation, the variable remains there.
- **Backward elimination:** Backward elimination (or backward deletion) is the reverse process. All the independent variables are entered into the equation first and each one is deleted one at a time if they do not contribute to the regression equation.

The independent variables were checked for multicollinearity using the VIF function, as defined:

- **Multicollinearity:** Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly related. It occurs when independent variables in a regression model are correlated.

The histogram graph was plotted for the residuals to find any unconventional data which cannot be determined using the regression models.

- **Residual:** A residual is the difference between the measured value and the predicted value of a regression model.

The QQ plot and PP plot determined 95% of the data was plotted on the line.

- **QQ Plot:** A graph of the residuals versus the expected order statistics of the standard normal distribution. The graph is also called a Q-Q Plot because it plots quantiles of the data versus quantiles of a distribution.

The stabilization transformation was applied based on the residuals plots.

We determine the best fit regression model on the basis of adjusted R-squared while also measuring the score of the model at predicting results from test data.

Individual Milestones

Our analysis divides the independent variables using the following criterias.

- Basic statistics: Typically include the summary or average of a single observation typically per game (e.g. points per game, rebounds per game, etc.).
- Advanced statistics: Modern statistics that typically require a formula that combines many of the basic statistics into a single rating where a key tenant is that the statistics are evaluated over possessions, not games.

We then further divide the statistics into the following sub-categories:

- Offensive statistics: Includes stats that are collected while the team is attacking.
- Defensive statistics: Includes stats that are collected while the team is defending.
- Neutral statistics: Attributes within the data which are neither considered defensive or offensive fall into this category

Basic Statistics Categories

Offensive statistics	Defensive statistics	Neutral statistics
Points per game	Opponent points per game	Scoring margin
Field goal percentage	Opponent field goal percentage	Rebound margin
3 point field goals per game	Opponent 3-point field goal percentage	Assist-turnover ratio
Free throw percentage	Opponent rebounds per game	Turnovers per game
Rebounds per game	Blocked shots per game	Personal fouls per game
Assists per game	Steals per game	

Advanced Statistics Categories

Offensive statistics	Defensive statistics	Neutral statistics
Offensive rating	Steal %	Total Rebound %
Pace	Block %	Turnover %
Free throw attempt rate	Opponent 3-point attempt rate	Opponent total rebound %
True shooting %	Opponent true shooting %	Opponent turnover %
Assist %	Opponent assist %	
Effective field goal %	Opponent steal %	
Offensive rebound %	Opponent block %	
FT per FGA	Opponent effective field goal %	
3-point attempt rate	Opponent offensive rebound %	
	Opponent FT per FGA	

Individual studies then were performed on the individual categories as follows:

1. Predict Winning Percentage from the *basic* offensive statistics
2. Predict Winning Percentage from the *basic* defensive statistics
3. Predict Winning Percentage from the *basic* neutral statistics
4. Predict Winning Percentage from the *advanced* statistics

Offensive Data Analysis (Srikanth Nanduri)

The purpose of the below is to separate the independent statistics related to offense. This excludes the following variable categories: defensive, neutral, and advanced statistics.

Problem

In recent years, a 32-year old man named Stephen Curry changed the game of basketball forever. The Golden State Warriors built a powerhouse offensive system based on Curry's three point shooting ability. The theories tested by the GSW basketball team were widely adopted both in the NBA and NCAA. Teams are shooting more threes and scoring more points. Does more points or offensive characteristic mean a higher win percentage?

Goal

The question that we are trying to answer is as follows: "How much of an impact do offensive statistics have on a team's winning percentage?" The purpose of this experiment was to divide the wide range of statistics available measuring a team's performance and determining the individual impact in affecting the overall win percentage of the model. The offensive data is composed of the following independent variables:

1. Points Per Game
2. Field Goal Percentage
3. Three Point Goals made
4. Three Point Field Goal Percentage
5. Free Throw Percentage
6. Rebound per Game
7. Assists Per Game

Note: The neutral and advanced variables are usually datasets that are calculated through the use of other basic recorded statistics. For example, Rebounding Margin involves a computation based on both offensive and defensive rebounds collected in a game. As these calculations involve both offensive and defensive combinations, the below analysis is primarily raw offensive statistics.

Data

See the *Data* section above for further details of the defense statistics data set, variables, and preparation used by this study.

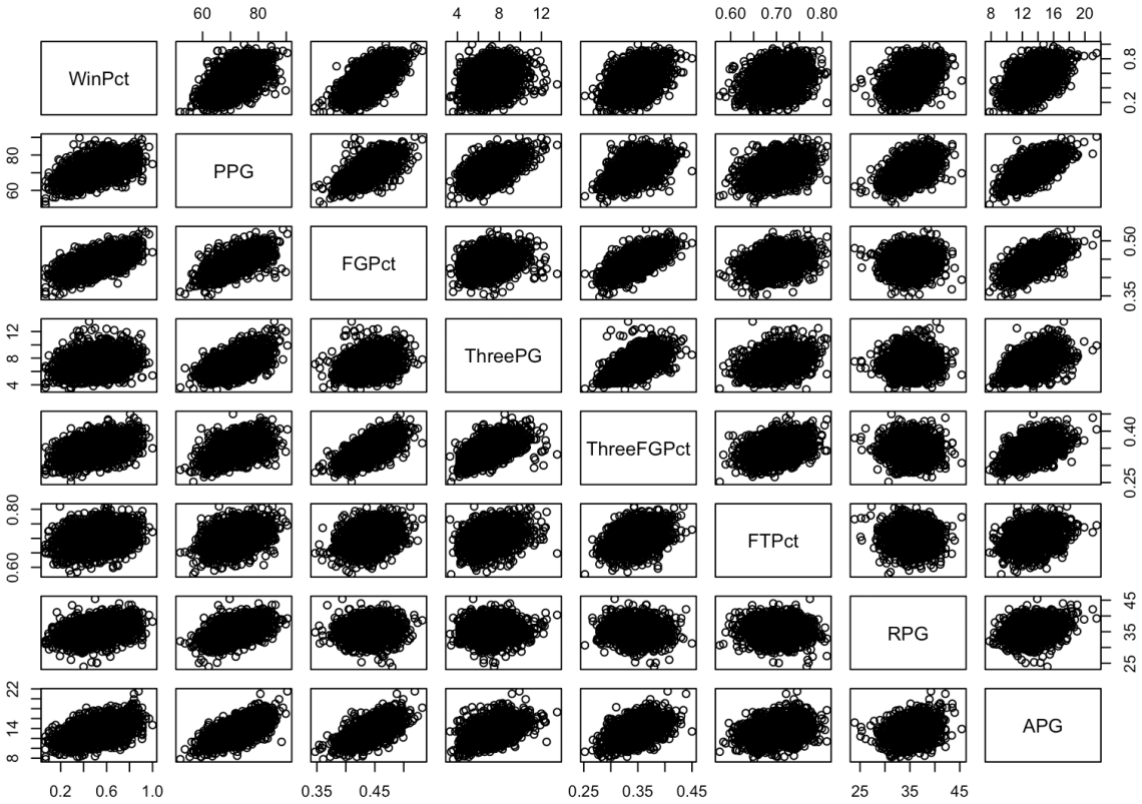
To begin, I reviewed the basic summary statistics associated with the offensive data as shown below.

```
> summary(o1)
     WinPct            PPG            FGPct          ThreePG         ThreeFGPct        FTPct            RPG
 Min.   :  6.70   Min.   :52.00   Min.   :34.70   Min.   : 3.300   Min.   :25.20   Min.   :57.60   Min.   :18.78
 1st Qu.: 38.70   1st Qu.:68.10   1st Qu.:42.40   1st Qu.: 6.300   1st Qu.:32.90   1st Qu.:67.90   1st Qu.:31.83
 Median : 51.50   Median :72.00   Median :44.20   Median : 7.200   Median :34.60   Median :70.20   Median :34.90
 Mean   : 51.56   Mean   :72.03   Mean   :44.17   Mean   : 7.274   Mean   :34.73   Mean   :70.22   Mean   :33.25
 3rd Qu.: 63.60   3rd Qu.:76.00   3rd Qu.:45.90   3rd Qu.: 8.200   3rd Qu.:36.52   3rd Qu.:72.70   3rd Qu.:36.90
 Max.   :100.00   Max.   :90.40   Max.   :53.20   Max.   :13.500   Max.   :45.00   Max.   :81.00   Max.   :43.97
      APG
 Min.   : 7.80
 1st Qu.:12.10
 Median :13.30
 Mean   :13.34
 3rd Qu.:14.50
 Max.   :21.50
```

Plot (*Win Percentage vs. all dependent variables*)

Subsequently, I plotted the independent variables in order to visually identify linear relationships between win percentage and the other variables. Below is a list of observations from the data plot:

1. From a general review of Figure 1, Win Percentage has the strongest linear relationship with Points Per Game, FG Percentage, Three Point Percentage, and assists Per Game.
2. PPG has a linear relationship with FGPct, ThreePG, RPG, and APG.
3. APG has strong linear relationship with PPG, FGPct, ThreePG, and ThreeFGPct.
4. Moreover, the plots showed primarily a linear relationship, as such, there was no significant need to attempt to create second order models.
5. There is a potential for interaction terms for PPG and FGPct, PPG and ThreePG, APG and PPG, and APG and FGPct.



Correlations with target

My observation regarding Win Percentage was supported by the review of the correlation matrix which showed that the strongest relationship for Win Percentage is with PPG, FGPct, ThreeFGPct, and APG. The highest correlation is between Win Percentage and FG percentage.

```
              WinPct       PPG      FGPct    ThreePG  ThreeFGPct       FTPct         RPG       APG
WinPct     1.0000000 0.5333452 0.64118008 0.22938891  0.44167086  0.25254996  0.36222145 0.4992204
PPG        0.5333452 1.0000000 0.65265574 0.55813063  0.49266452  0.37876004  0.47051313 0.6638909
FGPct      0.6411801 0.6526557 1.00000000 0.27721040  0.65326049  0.28642304  0.06657069 0.6375979
ThreePG    0.2293889 0.5581306 0.27721040 1.00000000  0.56059842  0.33470029  0.01414809 0.4483067
ThreeFGPct 0.4416709 0.4926645 0.65326049 0.56059842  1.00000000  0.35969412 -0.06488716 0.4778716
FTPct      0.2525500 0.3787600 0.28642304 0.33470029  0.35969412  1.00000000 -0.06801255 0.2441864
RPG        0.3622214 0.4705131 0.06657069 0.01414809 -0.06488716 -0.06801255  1.00000000 0.2606469
APG        0.4992204 0.6638909 0.63759790 0.44830666  0.47787157  0.24418642  0.26064692 1.0000000
```

Complete First-Order Model (inclusive of potential interaction terms)

We began with a basic model with all the variables and reviewed the summary statistics related to that model. Below is the summary:

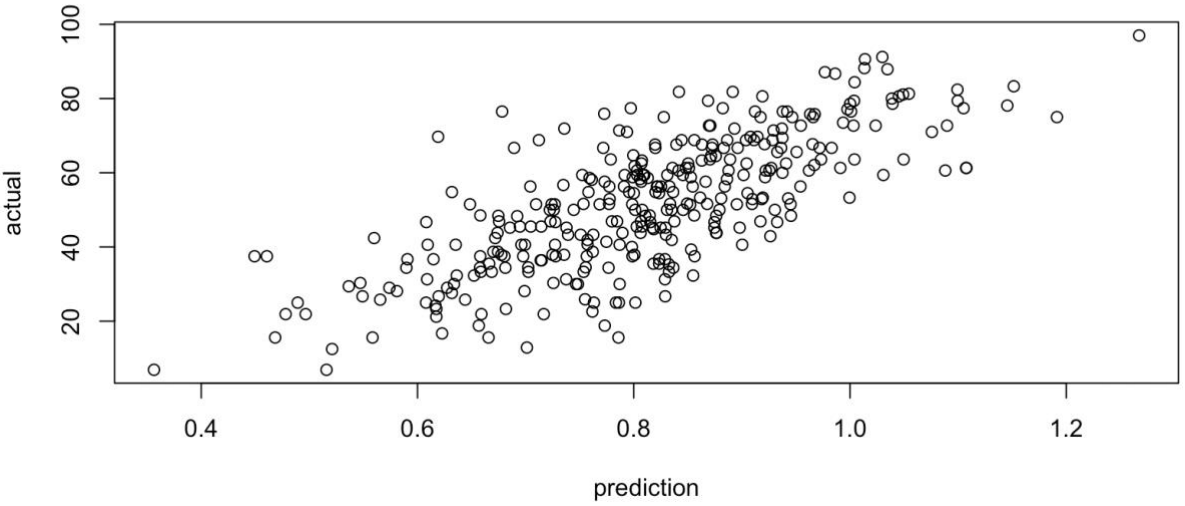
```
Call:
lm(formula = asin(sqrt(WinPct/100)) ~ ., data = trainOF)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4476 -0.0821 -0.0018  0.0835  0.5815

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.55e+00   6.72e-01   -3.80  0.00015 ***
PPG           -1.23e-02   1.46e-02   -0.84  0.39944
FGPct          4.83e-02   2.00e-02    2.42  0.01567 *
ThreePG        5.50e-03   4.86e-02    0.11  0.90995
ThreeFGPct     2.51e-03   2.01e-03    1.25  0.21269
FTPct          6.74e-03   1.08e-03    6.22  6.7e-10 ***
RPG            3.15e-02   1.82e-03   17.29  < 2e-16 ***
APG            3.13e-02   4.48e-02    0.70  0.48443
FGPct_ThreePG  4.12e-04   1.26e-03    0.33  0.74468
PPG_FGPct      9.13e-05   3.34e-04    0.27  0.78440
PPG_ThreePG   -1.46e-04   5.19e-04   -0.28  0.77864
APG_PPG        6.80e-05   5.27e-04    0.13  0.89737
APG_FGPct     -7.43e-04   1.16e-03   -0.64  0.52059
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.128 on 1390 degrees of freedom
Multiple R-squared:  0.533,     Adjusted R-squared:  0.529
F-statistic:  132 on 12 and 1390 DF,  p-value: <2e-16
```

From the information, I determined that the p-value for the F-test significant, therefore we can reject the null hypothesis that all the beta values of the model are equal to 0 and accept the alternative that at least 1 beta value is not equal to 0. However, in review of the T-test for each of the variables in the model, the values are insignificant deeming the variables to be potentially unnecessary for the model. So beginning with the interaction terms, I removed each variable and reviewed the key statistics to determine if they were required for the model to be effective.

Below is the summary of the final model. The summary shows that the model (tested on the training data) has 5 variables: PPG, FGPct, ThreePG, FTPct and RPG. Both the F-statistic and T-statistics for the model pass and the adjusted R-squared value for the model shows that 53.4% of the variability is explained by this model.

```
Call:
lm(formula = asin(sqrt(WinPct/100)) ~ ., data = trainOF)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4362 -0.0839 -0.0035  0.0849  0.5793

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.55287    0.10075  -25.34  < 2e-16 ***
PPG         -0.00751    0.00124   -6.07  1.6e-09 ***
FGPct        0.05034    0.00192   26.28  < 2e-16 ***
ThreePG      0.01446    0.00310    4.67  3.3e-06 ***
FTPct        0.00719    0.00105    6.87  9.3e-12 ***
RPG          0.03008    0.00173   17.34  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.127 on 1413 degrees of freedom
Multiple R-squared:  0.536,     Adjusted R-squared:  0.534
F-statistic:  326 on 5 and 1413 DF,  p-value: <2e-16
```

Based on the model comparison against the test values, the below is the chart showing the predicted versus actual values. The correlation value between the predicted and actual is 0.73 which indicates that model predicted values and actual values in the test data have a linear relationship.



Model Validation and Review

Subsequently, we performed a few tests to understand whether the model was aptly built. The first was to review whether the variables in the model were multicollinear. Then we tested the model to see if the same results could be reached using the backward and forward selection process.

Multicollinearity

```
> vif(ofm1)
    PPG   FGPct ThreePG   FTPct     RPG
   4.47    2.13    1.78    1.27    1.95
```

To evaluate whether multicollinearity existed within the model, we evaluated whether any of the variables in the model were more related with each other than the response variables. Per the analysis, we concluded that the VIF value for all the independent variables was less than 10 evidencing that there is no multicollinearity in the model.

Backward Elimination

The backward elimination process removed one variable from the system at a time to determine the key variables required to reduce the amount of information lost from the data. Based on the data, the final predicted model is outlined below. The model confirmed our analysis that the interaction terms are unnecessary. However, all the remaining terms were required.

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
asin(sqrt(WinPct/100)) ~ PPG + FGPct + ThreePG + ThreeFGPct +
    FTPct + RPG + APG + FGPct_ThreePG + PPG_FGPct + PPG_ThreePG +
    APG_PPG + APG_FGPct

Final Model:
asin(sqrt(WinPct/100)) ~ PPG + FGPct + ThreePG + ThreeFGPct +
    FTPct + RPG + APG


             Step Df Deviance Resid. Df Resid. Dev   AIC
1                                  1739       28.1 -7214
2 - FGPct_ThreePG  1 9.63e-06      1740       28.1 -7216
3       - APG_PPG  1 1.97e-04      1741       28.1 -7218
4   - PPG_ThreePG  1 2.77e-03      1742       28.1 -7220
5     - PPG_FGPct  1 6.98e-03      1743       28.1 -7221
6     - APG_FGPct  1 7.17e-03      1744       28.1 -7223
```

In review of the model summary against the complete dataset, the variables selected through the backward elimination model had a few issues. The first issue was that for a few select variables the T-test failed. This means that we were unable to reject the null hypothesis that the beta value for that variable was 0. The second issue is that the overall variability explained through the R-squared value for the backward elimination model did not significantly improve 53.4% (validated model) versus 53.9% (backward elimination model).

```
Call:
lm(formula = WinPct ~ PPG + FGPct + ThreePG + ThreeFGPct + FTPct +
    RPG + APG, data = o1)

Residuals:
   Min     1Q Median     3Q    Max
-40.04  -7.78   0.03   8.12  36.55

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -260.7863     8.8640  -29.42  < 2e-16 ***
PPG           -0.7705     0.1077   -7.15  1.3e-12 ***
FGPct          4.4899     0.2134   21.04  < 2e-16 ***
ThreePG        1.1706     0.3106    3.77  0.00017 ***
ThreeFGPct     0.2825     0.1650    1.71  0.08710 .
FTPct          0.6381     0.0899    7.10  1.9e-12 ***
RPG            2.8846     0.1491   19.34  < 2e-16 ***
APG            0.3228     0.2258    1.43  0.15301
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.9 on 1744 degrees of freedom
Multiple R-squared:  0.541,     Adjusted R-squared:  0.539
F-statistic:  293 on 7 and 1744 DF,  p-value: <2e-16
```

Forward Selection

Similarly, for the forward selection process, the below was the final model showing the required variables for the least amount of information lost by not including the variable.

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
WinPct ~ 1

Final Model:
WinPct ~ FGPct + RPG + FTPct + PPG + ThreePG + ThreeFGPct + APG


          Step Df Deviance Resid. Df Resid. Dev   AIC
1                                1751     535934 10029
2      + FGPct  1   220329      1750     315605  9103
3        + RPG  1    54965      1749     260641  8770
4      + FTPct  1     5522      1748     255118  8735
5        + PPG  1     3490      1747     251628  8713
6    + ThreePG  1     4730      1746     246899  8681
7 + ThreeFGPct  1      387      1745     246512  8681
8        + APG  1      289      1744     246224  8680
```

In evaluation of the summary of the model created to the forward selection process, the forward selection process found the same variable requirement as the backward elimination process. However, similar to the backward elimination process, the model is inclusive of multiple variables that fail the T-test and the R-squared value is 53.9% which is marginally greater than the initial cross validation model.

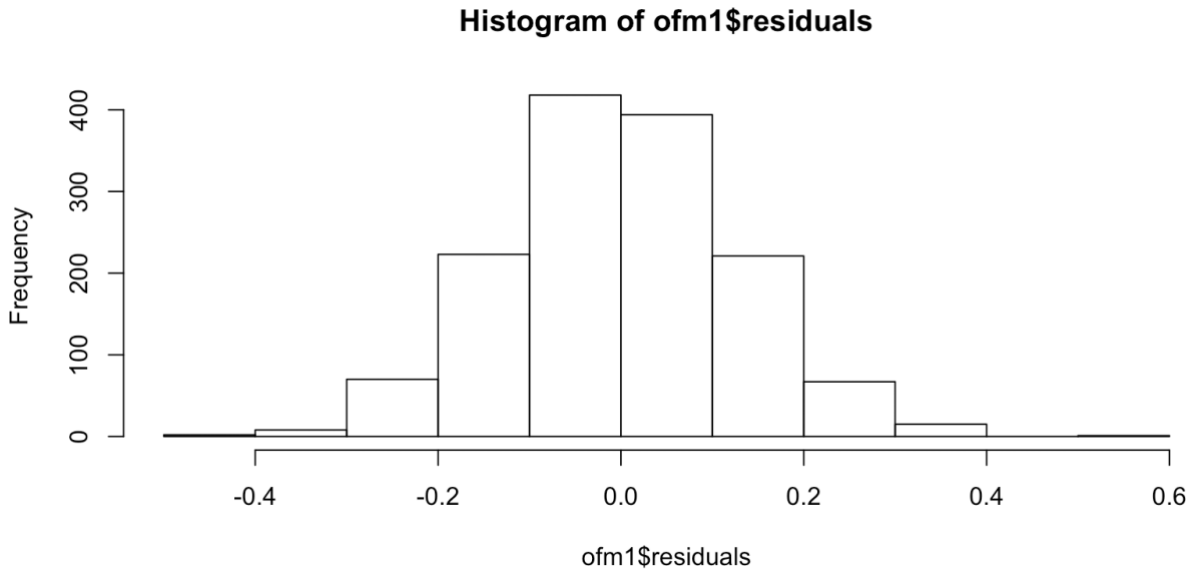
```
Call:
lm(formula = WinPct ~ FGPct + RPG + FTPct + PPG + ThreePG + ThreeFGPct +
    APG, data = o1)

Residuals:
   Min     1Q Median     3Q    Max
-40.04  -7.78   0.03   8.12  36.55

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -260.7863     8.8640  -29.42  < 2e-16 ***
FGPct          4.4899     0.2134   21.04  < 2e-16 ***
RPG            2.8846     0.1491   19.34  < 2e-16 ***
FTPct          0.6381     0.0899    7.10  1.9e-12 ***
PPG           -0.7705     0.1077   -7.15  1.3e-12 ***
ThreePG        1.1706     0.3106    3.77  0.00017 ***
ThreeFGPct     0.2825     0.1650    1.71  0.08710 .
APG            0.3228     0.2258    1.43  0.15301
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.9 on 1744 degrees of freedom
Multiple R-squared:  0.541,     Adjusted R-squared:  0.539
F-statistic:  293 on 7 and 1744 DF,  p-value: <2e-16
```

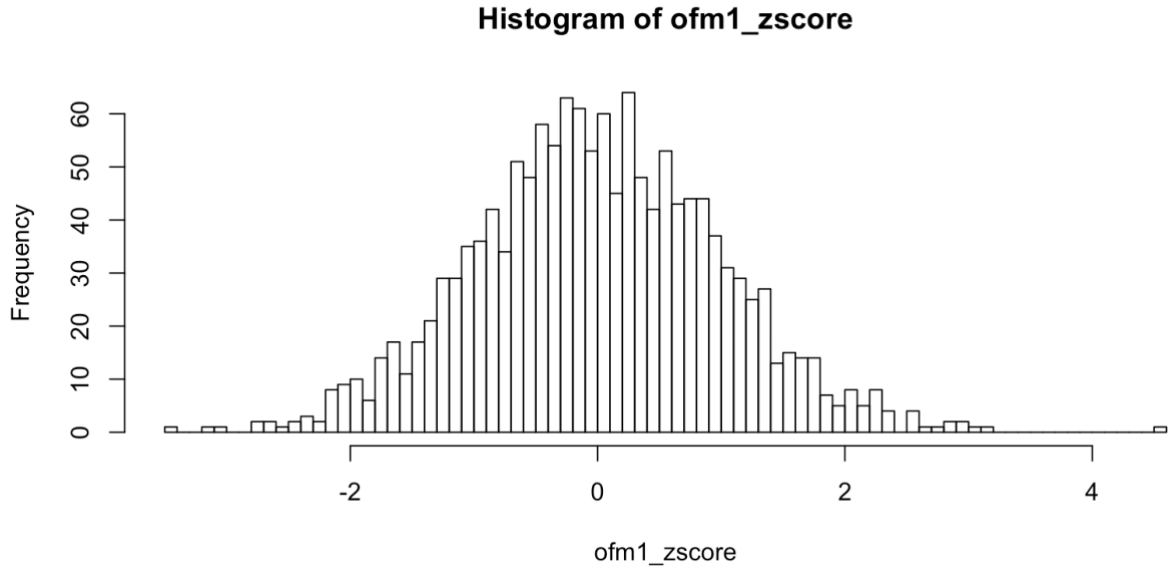
Review of the 4 Assumptions

Next, we reviewed and evaluated whether the 4 assumptions of linear models were met. The four assumptions are as follows:

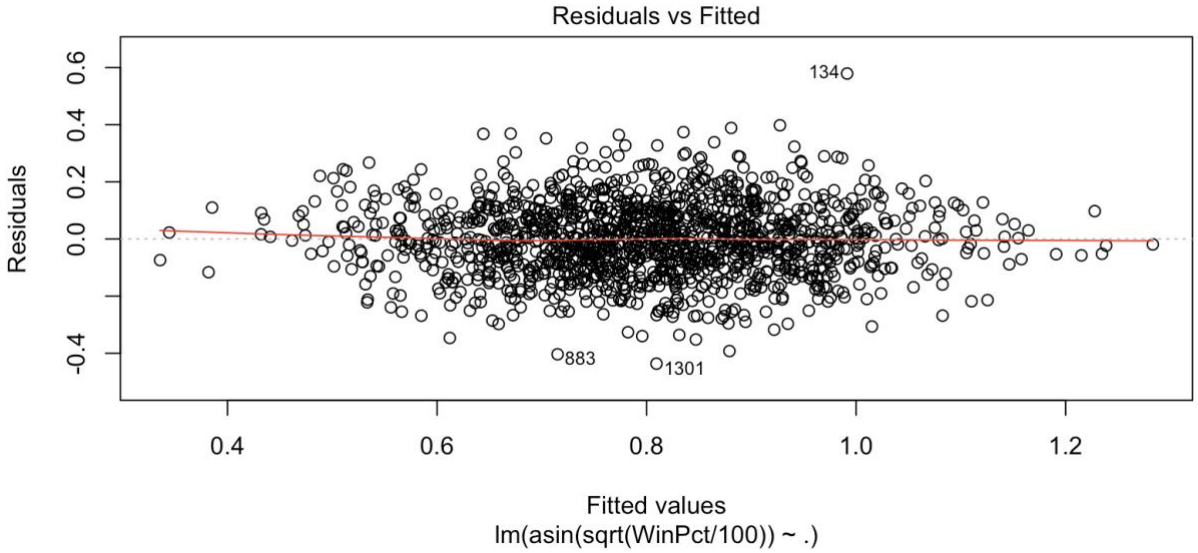
1. Normal Distribution of Residuals
2. Residuals are Homoscedastic
3. Mean sum of residuals is 0
4. Residuals are independent from one another

Beginning with the normal distribution, the histogram below shows that the residuals are normally distributed. This is shown through the plot of each residual in the first histogram and plotting the z-scores of the residuals.

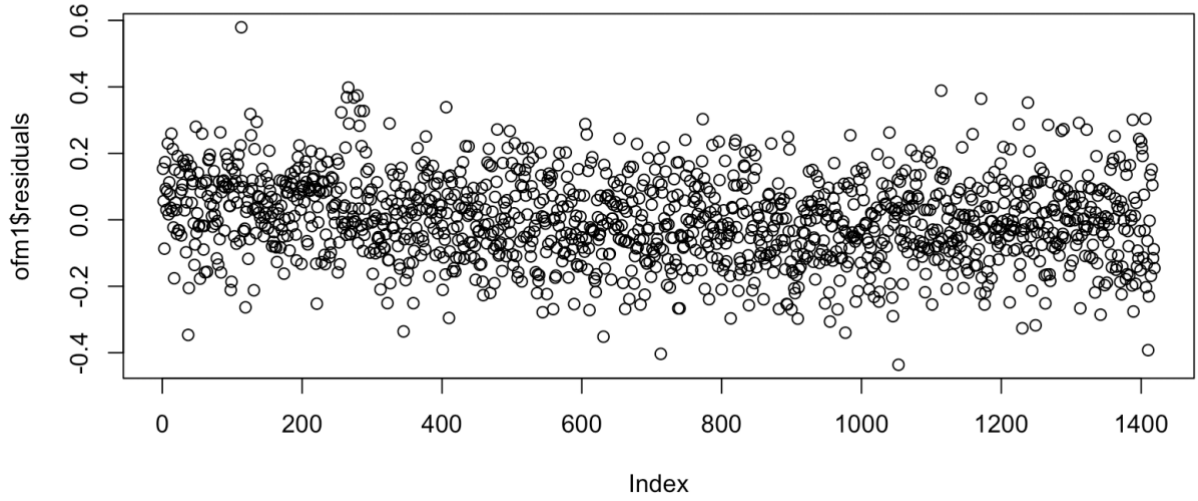




Per review of the sum of the mean of the residuals, the value was very close to 0 (7.4e-19). Although the plot of the residuals versus fitted shows a slight unequal variance, the response variable has been already transformed to counteract this. As such, the model is acceptable and does not require additional transformations.



Per review of the residuals plot on the index, we concluded that residuals plot appears to be random. As such, we concluded that the residuals are independent.



Lack of Fit

Based on the Durbin Watson Test, the value of the statistic was between 0 - 2.5. As such, the residuals are independent of one another.

```
        Durbin-Watson test

data:  ofm1
DW = 2, p-value = 0.01
alternative hypothesis: true autocorrelation is greater than 0
```

Summary

In summary, the model details are the following:

1. arcsin(sqrt(WinPct)) = PPG + FGPct + ThreePG + FTPct + RPG

```
Call:
lm(formula = asin(sqrt(WinPct/100)) ~ ., data = trainOF)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4362 -0.0839 -0.0035  0.0849  0.5793

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.55287    0.10075  -25.34  < 2e-16 ***
PPG         -0.00751    0.00124   -6.07  1.6e-09 ***
FGPct        0.05034    0.00192   26.28  < 2e-16 ***
ThreePG      0.01446    0.00310    4.67  3.3e-06 ***
FTPct        0.00719    0.00105    6.87  9.3e-12 ***
RPG          0.03008    0.00173   17.34  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.127 on 1413 degrees of freedom
Multiple R-squared:  0.536,     Adjusted R-squared:  0.534
F-statistic:  326 on 5 and 1413 DF,  p-value: <2e-16
```

2. The model passes the F-test and the T-test for the variables are significant, therefore we can reject the null hypothesis and accept the alternative beta values identified through the model.
3. The histogram of the residuals from the model shows the residuals are normally distributed.
4. The sum of the mean of the residuals, the value was very close to 0 (7.4e-19).
5. The plot of the residuals versus the fitted model shows a slight unequal variance, however, this was reduced through transformation of the response variable to arcsin(sqrt(WinPct)). As such, the residuals are homoscedastic.
6. The plot of the residuals versus index showed no pattern which shows that the residuals are independent.

Defensive Statistics Analysis (Vineet Dcunha)

Problem

As the saying goes, “Defence wins games,” I will put to test this state and understand whether a strong defensive regular season directly impacts the percentage of games won.

Goal

I will be performing the regression model focusing on defensive variables. I will use a complete second order regression model in my approach to find the model that explains the most variability in the response variable (win percentage).

My analysis will also consider, if any multicollinearity exists between any of the independent variables. I will validate my second order regression model against the forward or backward selection model to identify the difference in my final model.

Specifically, the following data categories (independent variables) are included in my analysis:

Independent Variables:

- Opponents Field Goal Percentage
- Opponent 3 Point Field Goal Percentage
- Opponent Average 3 Pointers Made/Game
- Opponents Points per Game
- Steals Per Game
- Blocked Shots Per Game

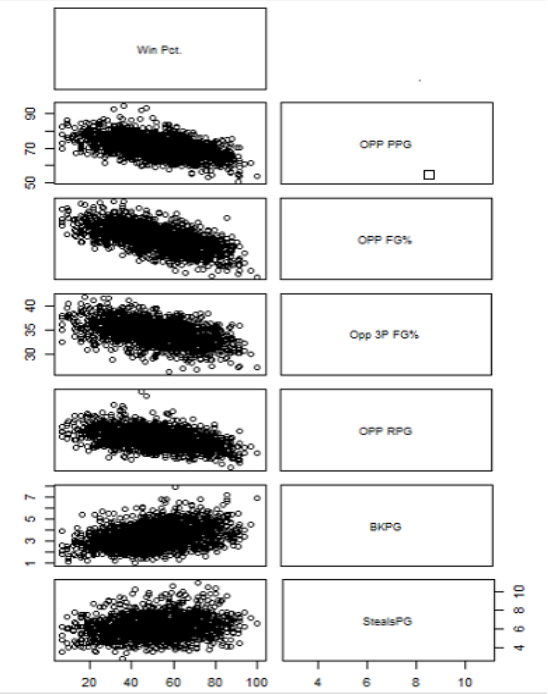
Dependent Variable: Win Percentage

Data

See the *Data* section above for further details of the defense statistics data set, variables, and preparation used by this study.

Plot (*Win Percentage vs. all dependent variables*)

The plot (Win Percentage vs. all independent variables) shows a moderate to good linear relationship between Win Percentage and various independent variables.



Correlations with target

> cor(NCAA)

	Win Pct.	OPP PPG	OPP FG%	Opp 3P FG%	OPP RPG	BKPG	StealsPG
Win Pct.	1.0000000	-0.50798460	-0.60714290	-0.48212603	-0.45345773	0.35103942	0.21114144
OPP PPG	-0.5079846	1.00000000	0.72040033	0.51861724	0.64482686	-0.19422561	-0.00940435
OPP FG%	-0.6071429	0.72040033	1.00000000	0.62052323	0.30747845	-0.49794944	0.01276254
Opp 3P FG%	-0.4821260	0.51861724	0.62052323	1.00000000	0.14678155	-0.21934756	-0.08482734
OPP RPG	-0.4534577	0.64482686	0.30747845	0.14678155	1.00000000	0.03229654	0.15405215
BKPG	0.3510394	-0.19422561	-0.49794944	-0.21934756	0.03229654	1.00000000	0.11763493
StealsPG	0.2111414	-0.00940435	0.01276254	-0.08482734	0.15405215	0.11763493	1.00000000

As analyzed in the above correlation stats, there exists no multicollinearity amongst various independent variables. Multicollinearity only exists, if any of the above value is greater than 0.90

The maximum value we see between the various variables is 0.72 which exists between Opponents Field Goal Percentage and Opponent 3 Point Field Goal Percentage.

Multicollinearity

```
> vif(m1)
   NCAA$`OPP PPG`     NCAA$`OPP FG%` NCAA$`Opp 3P FG%`     NCAA$`OPP RPG`
         3.908824          3.620221          1.754668          2.042682
      NCAA$BKPG     NCAA$StealsPG
         1.499392          1.088329
```

All the variables are showing a VIF value of less than 10. Therefore we can confirm no multicollinearity exists between any independent variable.

First-Order Model

```
> m1 <- lm(NCAA$`Win Pct.` ~ NCAA$`OPP PPG` + NCAA$`OPP FG%` + NCAA$`Opp 3P FG%` +
NCAA$`OPP RPG` + NCAA$BKPG + NCAA$StealsPG)

> summary(m1)

Call:
lm(formula = NCAA$`Win Pct.` ~ NCAA$`OPP PPG` + NCAA$`OPP FG%` +
    NCAA$`Opp 3P FG%` + NCAA$`OPP RPG` + NCAA$BKPG + NCAA$StealsPG)

Residuals:
    Min      1Q  Median      3Q     Max
-41.118  -7.630  -0.047   7.887  37.159

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      279.67657    7.09492  39.419  < 2e-16 ***
NCAA$`OPP PPG`     0.99224    0.09642  10.291  < 2e-16 ***
NCAA$`OPP FG%`    -3.52894    0.21416 -16.478  < 2e-16 ***
NCAA$`Opp 3P FG%` -1.64983    0.15523 -10.628  < 2e-16 ***
NCAA$`OPP RPG`    -3.44003    0.14788 -23.263  < 2e-16 ***
NCAA$BKPG          1.84795    0.33654   5.491 4.58e-08 ***
NCAA$StealsPG      4.15359    0.25170  16.502  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.46 on 1745 degrees of freedom
Multiple R-squared:  0.5723,      Adjusted R-squared:  0.5709
F-statistic: 389.2 on 6 and 1745 DF,  p-value: < 2.2e-16
```

As shown in the above summary report for linear regression model, the p value for overall model is below the level of significance (a = 0.5). The value of **F test at 389.2** supports this assumption.

Therefore we reject the null hypothesis and accept the alternative that at least one of the betas is not equal to zero.

The value of **R squared** looks decent at **0.5723**. Considering the number of beta variables, the value of **adjusted R squared** is **0.5709**.

This concludes the **~ 57%** of variability in dependent variables can be predicted by this model.

The p value for all the defensive independent variables is below the level of significance (a = 0.5). The null hypothesis can be rejected and accepting the alternative hypothesis confirming the value of beta is not equal to zero.

Therefore all the dependent variables can be included in this model.

Second Order Model

The above plot diagram (*Win Percentage vs. all dependent variables*) doesn't show any curvatures. Therefore we can conclude there are no second order terms.

Interaction Model

```
> m2 <- lm(NCAA$`Win Pct.` ~ NCAA$`OPP PPG` + NCAA$`OPP FG%` + NCAA$`Opp 3P FG%` +
NCAA$`OPP RPG` + NCAA$BKPG + NCAA$StealsPG + NCAA$`OPP FG%`*NCAA$`OPP RPG` +
NCAA$`OPP FG%`*NCAA$BKPG + NCAA$`OPP RPG`*NCAA$StealsPG )

> summary(m2)

Call:
lm(formula = NCAA$`Win Pct.` ~ NCAA$`OPP PPG` + NCAA$`OPP FG%` +
    NCAA$`Opp 3P FG%` + NCAA$`OPP RPG` + NCAA$BKPG + NCAA$StealsPG +
    NCAA$`OPP FG%` * NCAA$`OPP RPG` + NCAA$`OPP FG%` * NCAA$BKPG +
    NCAA$`OPP RPG` * NCAA$StealsPG)

Residuals:
    Min      1Q  Median      3Q     Max
-41.956  -7.430   0.071   7.862  37.672

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   530.80147   63.26196   8.391  < 2e-16 ***
NCAA$`OPP PPG`                  0.94987    0.09629   9.864  < 2e-16 ***
NCAA$`OPP FG%`                 -8.16213    1.48450  -5.498 4.40e-08 ***
NCAA$`Opp 3P FG%`              -1.63544    0.15461 -10.578  < 2e-16 ***
NCAA$`OPP RPG`                 -9.36351    1.69574  -5.522 3.86e-08 ***
NCAA$BKPG                     -10.86510    4.42350  -2.456  0.01414 *
NCAA$StealsPG                  -3.24358    2.99960  -1.081  0.27970
NCAA$`OPP FG%`:NCAA$`OPP RPG`   0.10537    0.03994   2.639  0.00840 **
NCAA$`OPP FG%`:NCAA$BKPG        0.29602    0.10280   2.880  0.00403 **
NCAA$`OPP RPG`:NCAA$StealsPG    0.20817    0.08492   2.451  0.01433 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.39 on 1742 degrees of freedom
Multiple R-squared:  0.578,    Adjusted R-squared:  0.5758
F-statistic: 265.1 on 9 and 1742 DF,  p-value: < 2.2e-16
```

Following the analysis and trial of multiple interaction terms, we can conclude the interaction terms (marked in bold) are the best fit.

As shown in the above summary report for interaction model, the p value for overall model is below the level of significance (a = 0.05). The value of **F test at 265.1** supports this assumption.

Therefore we reject the null hypothesis and accept the alternative that at least one of the betas is not equal to zero.

The value of **R squared** looks decent at **0.578**. Considering the number of beta variables, the value of **adjusted R squared** is **0.5758**.

This concludes the **~ 57%** of variability in dependent variables can be predicted by this model.

The p value for all the interaction variables is below the level of significance (a = 0.05). The null hypothesis can be rejected and accepting the alternative hypothesis confirming the value of beta is not equal to zero.

Therefore all the dependent and interaction variables can be included in this model.

Even though the p value of NCAA$StealsPG is above the threshold (level of significance a = 0.05), we will include this variable in the model (since the interaction term is included in the model).

Backward selection

Backward elimination (or backward deletion) is the reverse process. All the independent variables are entered into the equation first and each one is deleted one at a time if they do not contribute to the regression equation.

```
> step <- stepAIC(m2,direction="backward")
```

```
> step$anova
```

```
> step$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
NCAA$`Win Pct.` ~ NCAA$`OPP PPG` + NCAA$`OPP FG%` + NCAA$`Opp 3P FG%` +
    NCAA$`OPP RPG` + NCAA$BKPG + NCAA$StealsPG + NCAA$`OPP FG%` *
    NCAA$`OPP RPG` + NCAA$`OPP FG%` * NCAA$BKPG + NCAA$`OPP RPG` *
    NCAA$StealsPG

Final Model:
NCAA$`Win Pct.` ~ NCAA$`OPP PPG` + NCAA$`OPP FG%` + NCAA$`Opp 3P FG%` +
    NCAA$`OPP RPG` + NCAA$BKPG + NCAA$StealsPG + NCAA$`OPP FG%` *
    NCAA$`OPP RPG` + NCAA$`OPP FG%` * NCAA$BKPG + NCAA$`OPP RPG` *
    NCAA$StealsPG


  Step Df Deviance Resid. Df Resid. Dev      AIC
1                       1742     226175 8535.685
```

The backward selection model is producing the same set of variables and model as the interaction model. The backward selection model is just producing a step 1 without dropping any of the variables.

Forward selection

Forward selection begins with an empty equation. Predictors are added one at a time beginning with the predictor with the highest correlation with the dependent variable. Variables of greater theoretical importance are entered first. Once in the equation, the variable remains there.

```
> m3 <- lm(NCAA$`Win Pct.` ~ 1)
```

```
> step <- stepAIC(m3,direction="forward",scope=list(upper=m2,lower=m3))
```

```
Start:  AIC=10029.14
NCAA$`Win Pct.` ~ 1

                   Df Sum of Sq    RSS     AIC
+ NCAA$`OPP FG%`    1    197557 338377  9225.5
+ NCAA$`OPP PPG`    1    138297 397637  9508.2
+ NCAA$`Opp 3P FG%` 1    124576 411359  9567.7
+ NCAA$`OPP RPG`    1    110201 425733  9627.8
+ NCAA$BKPG         1     66042 469892  9800.7
+ NCAA$StealsPG     1     23892 512042  9951.2
<none>                         535934 10029.1

Step:  AIC=9225.48
NCAA$`Win Pct.` ~ NCAA$`OPP FG%`

                   Df Sum of Sq    RSS    AIC
+ NCAA$`OPP RPG`    1     42124 296253 8994.6
+ NCAA$StealsPG     1     25682 312695 9089.2
+ NCAA$`Opp 3P FG%` 1      9678 328699 9176.6
+ NCAA$`OPP PPG`    1      5553 332824 9198.5
+ NCAA$BKPG         1      1691 336686 9218.7
<none>                         338377 9225.5

Step:  AIC=8994.56
NCAA$`Win Pct.` ~ NCAA$`OPP FG%` + NCAA$`OPP RPG`

                              Df Sum of Sq    RSS    AIC
+ NCAA$StealsPG                1     38058 258195 8755.7
+ NCAA$`Opp 3P FG%`            1     12249 284003 8922.6
+ NCAA$BKPG                    1      8015 288238 8948.5
+ NCAA$`OPP PPG`               1      5545 290707 8963.5
+ NCAA$`OPP FG%`:NCAA$`OPP RPG` 1      2144 294109 8983.8
<none>                                    296253 8994.6

Step:  AIC=8755.66
NCAA$`Win Pct.` ~ NCAA$`OPP FG%` + NCAA$`OPP RPG` + NCAA$StealsPG

                              Df Sum of Sq    RSS    AIC
+ NCAA$`OPP PPG`               1   11881.7 246313 8675.1
+ NCAA$`Opp 3P FG%`            1    8038.8 250156 8702.2
+ NCAA$BKPG                    1    4644.4 253550 8725.9
+ NCAA$`OPP RPG`:NCAA$StealsPG 1    2490.3 255704 8740.7
+ NCAA$`OPP FG%`:NCAA$`OPP RPG` 1    1461.5 256733 8747.7
<none>                                    258195 8755.7


Step:  AIC=8675.12
NCAA$`Win Pct.` ~ NCAA$`OPP FG%` + NCAA$`OPP RPG` + NCAA$StealsPG +
    NCAA$`OPP PPG`

                              Df Sum of Sq    RSS    AIC
+ NCAA$`Opp 3P FG%`            1   13152.3 233161 8581.0
+ NCAA$BKPG                    1    2276.4 244036 8660.9
+ NCAA$`OPP RPG`:NCAA$StealsPG 1    1848.1 244465 8663.9
+ NCAA$`OPP FG%`:NCAA$`OPP RPG` 1    1146.3 245167 8668.9
<none>                                    246313 8675.1

Step:  AIC=8580.98
NCAA$`Win Pct.` ~ NCAA$`OPP FG%` + NCAA$`OPP RPG` + NCAA$StealsPG +
    NCAA$`OPP PPG` + NCAA$`Opp 3P FG%`

                              Df Sum of Sq    RSS    AIC
+ NCAA$BKPG                    1    3960.3 229200 8553.0
+ NCAA$`OPP RPG`:NCAA$StealsPG 1    1363.3 231797 8572.7
+ NCAA$`OPP FG%`:NCAA$`OPP RPG` 1    1168.7 231992 8574.2
<none>                                    233161 8581.0
```

```
Step:  AIC=8552.96
NCAA$`Win Pct` ~ NCAA$`OPP FG%` + NCAA$`OPP RPG` + NCAA$StealsPG +
    NCAA$`OPP PPG` + NCAA$`Opp 3P FG%` + NCAA$BKPG

                                Df Sum of Sq    RSS    AIC
+ NCAA$`OPP RPG`:NCAA$StealsPG   1   1286.84 227913 8545.1
+ NCAA$`OPP FG%`:NCAA$`OPP RPG`  1   1183.78 228016 8545.9
+ NCAA$`OPP FG%`:NCAA$BKPG       1    736.88 228463 8549.3
<none>                                       229200 8553.0

Step:  AIC=8545.1
NCAA$`Win Pct` ~ NCAA$`OPP FG%` + NCAA$`OPP RPG` + NCAA$StealsPG +
    NCAA$`OPP PPG` + NCAA$`Opp 3P FG%` + NCAA$BKPG + NCAA$`OPP
RPG`:NCAA$StealsPG

                                Df Sum of Sq    RSS    AIC
+ NCAA$`OPP FG%`:NCAA$BKPG       1    834.50 227079 8540.7
+ NCAA$`OPP FG%`:NCAA$`OPP RPG`  1    661.77 227252 8542.0
<none>                                       227913 8545.1

Step:  AIC=8540.67
NCAA$`Win Pct` ~ NCAA$`OPP FG%` + NCAA$`OPP RPG` + NCAA$StealsPG +
    NCAA$`OPP PPG` + NCAA$`Opp 3P FG%` + NCAA$BKPG + NCAA$`OPP
RPG`:NCAA$StealsPG +
    NCAA$`OPP FG%`:NCAA$BKPG

                                Df Sum of Sq    RSS    AIC
+ NCAA$`OPP FG%`:NCAA$`OPP RPG`  1    903.93 226175 8535.7
<none>                                       227079 8540.7

Step:  AIC=8535.69
NCAA$`Win Pct` ~ NCAA$`OPP FG%` + NCAA$`OPP RPG` + NCAA$StealsPG +
    NCAA$`OPP PPG` + NCAA$`Opp 3P FG%` + NCAA$BKPG + NCAA$`OPP
RPG`:NCAA$StealsPG +
    NCAA$`OPP FG%`:NCAA$BKPG + NCAA$`OPP FG%`:NCAA$`OPP RPG`
```

The forward selection model is producing the same variables and model as the interaction model. The forward selection model is iterating 10 times to produce the same regression model.

We do not see a drastic difference for the AIC value between step 1 and step 10.

Residual analysis

This section performs an analysis to confirm the four assumptions about the errors:

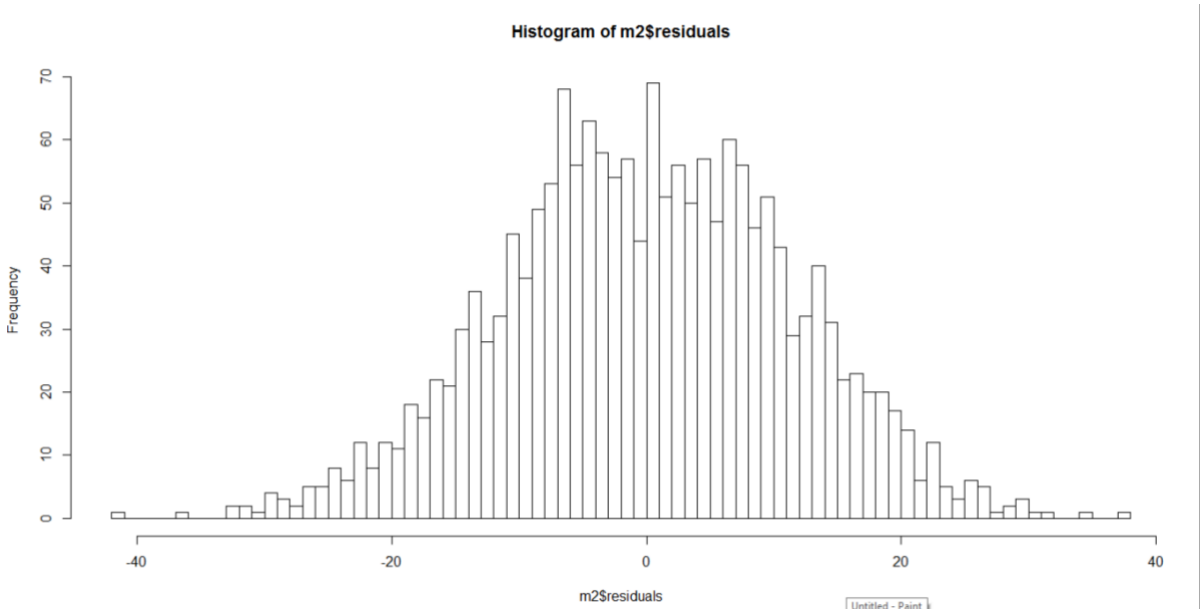
Mean of errors is 0

The sum of mean of the errors is zero.

```
> sum(m2$residuals)
[1] -9.828249e-13
```

The histogram graph for the residuals shows a normal distribution pattern.

```
> hist(m2$residuals,breaks = 100)
```



Following are the general stats of the residuals.

Mean:

```
> mean(m2$residuals)
[1] -5.648921e-16
```

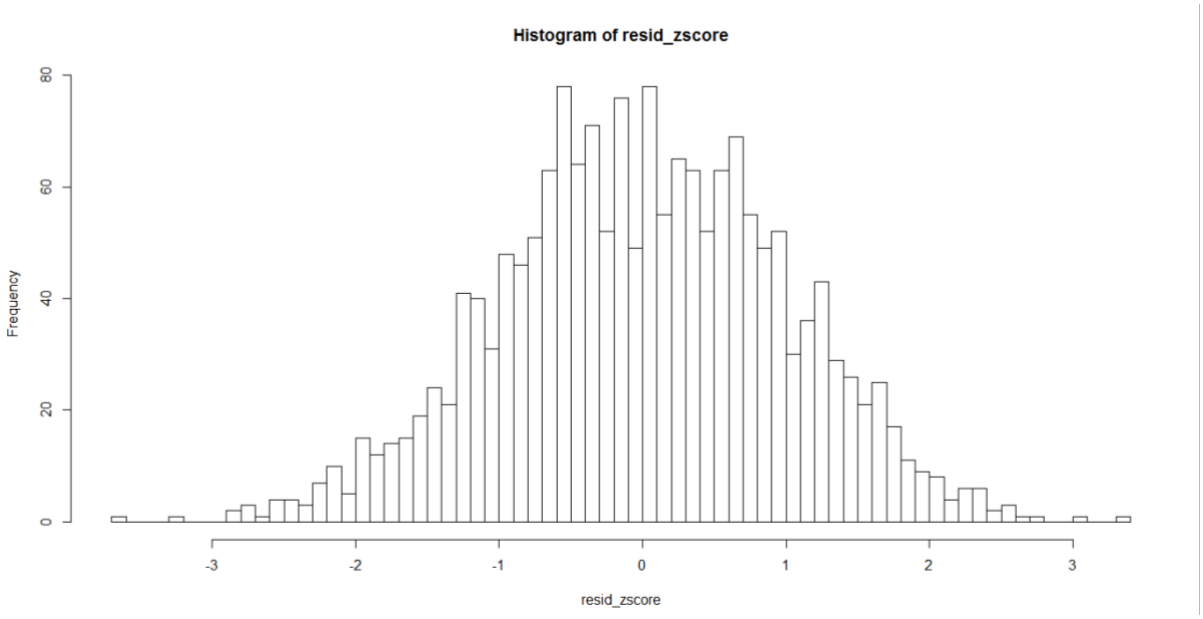
Standard Deviation:

```
> sd(m2$residuals)
[1] 11.36526
```

Histogram graph for z score normalisation.

```
> hist(resid_zscore,breaks = 100)
```

The histogram graph for the Z score shows a normal distribution pattern.

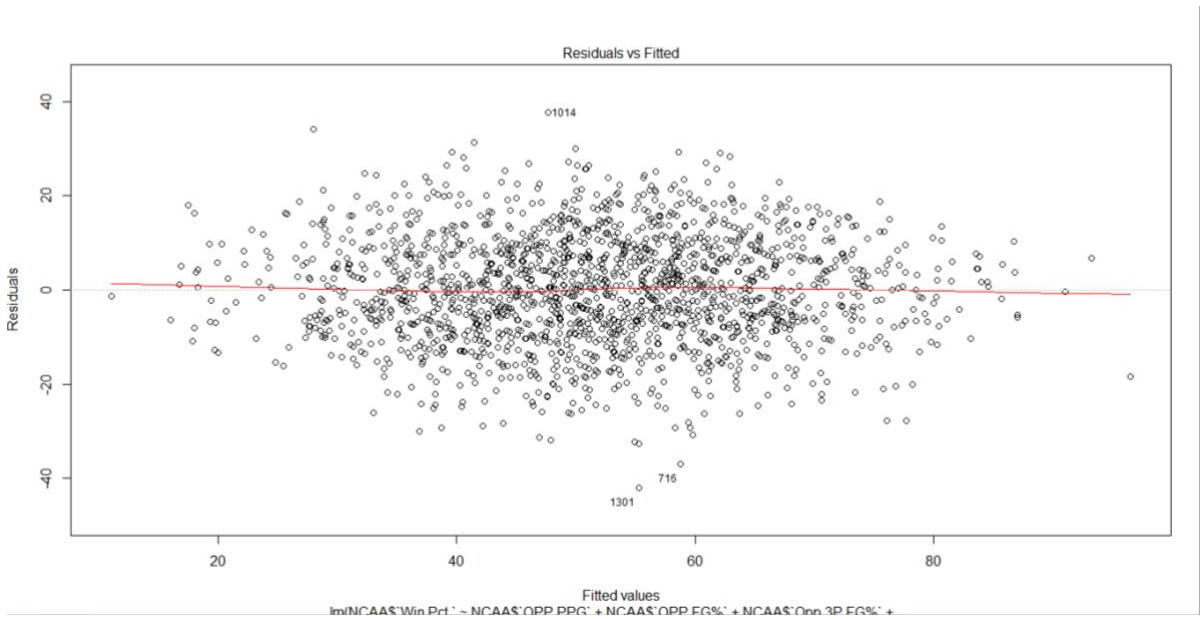


Homoscedasticity

Residuals vs Fitted Plot

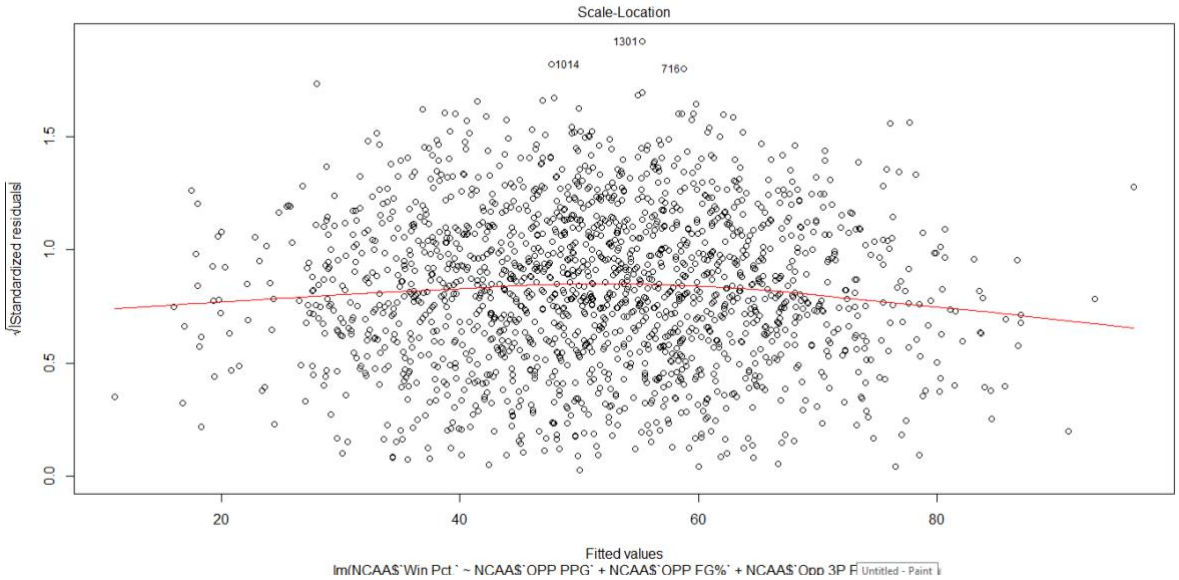
The residuals vs fitted graph shows binomial formation.

We would have to apply a stabilization transformation on the dependent variable (Win Percentage)



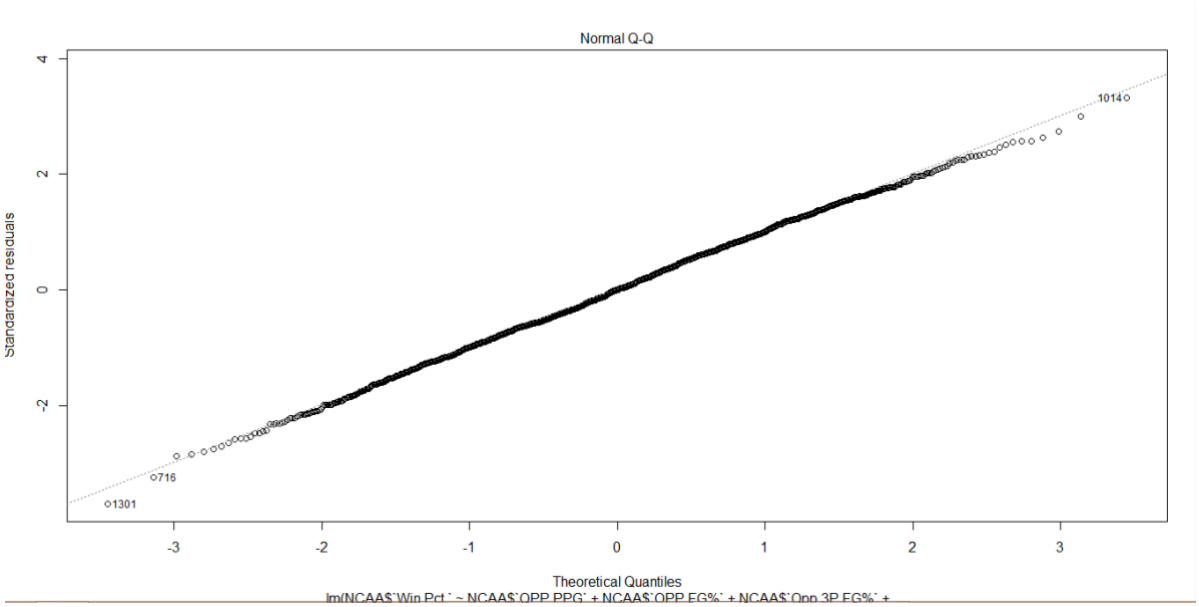
Square root of standardised residuals vs fitted

The graph square root of standardised residuals vs fitted values shows a normal variance.



Normality

The qq line plots all the data points across the line. 95% of the data is plotted on the line.



Independence

Durbin-Watson tests for autocorrelation in residuals from a regression analysis.

```
> durbinWatsonTest(m2)
```

```
 lag Autocorrelation D-W Statistic p-value
   1      0.01219258      1.974841   0.598
 Alternative hypothesis: rho != 0
```

The test statistic ranges in between 0 to 4. A value of 2 indicates that there is no autocorrelation. Value nearing 0 (i.e., below 2) indicates positive autocorrelation and value towards 4 (i.e., over 2) indicates negative autocorrelation.

Stabilization transformation

Since the residuals vs fitted graph shows a binomial formation, we can apply the stabilization transformation on the dependent variable (Win Percentage) and add it to the model.

```
> NCAA$Y_WIN <- asin(sqrt(NCAA$`Win Pct.`/100))
```

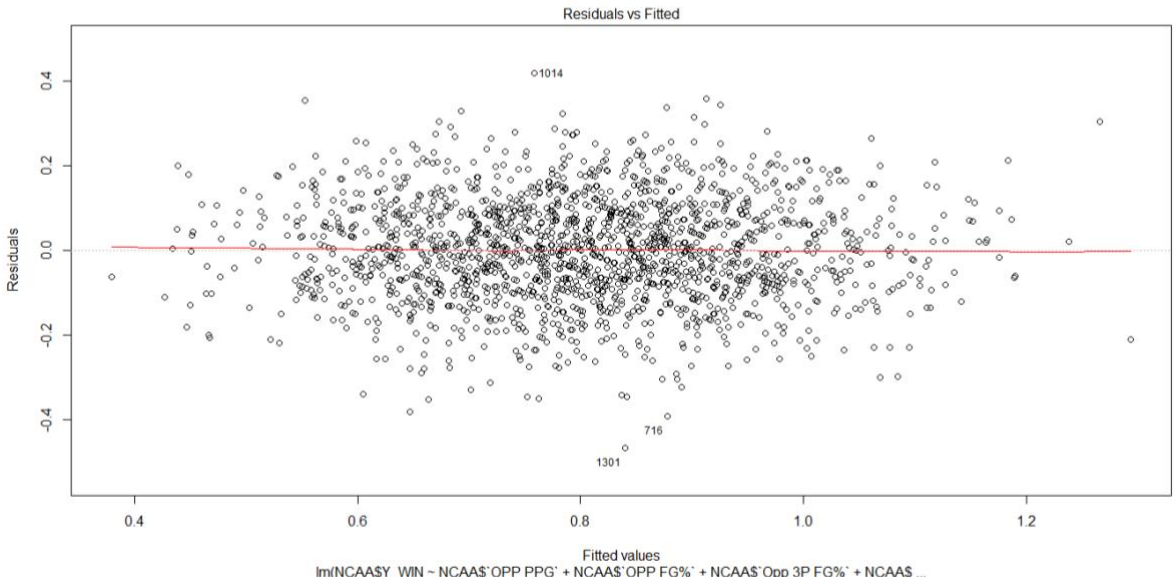
```
> m3 <- lm(NCAA$Y_WIN ~ NCAA$`OPP PPG` + NCAA$`OPP FG%` + NCAA$`Opp 3P FG%` + NCAA$`OPP RPG` + NCAA$BKPG + NCAA$StealsPG + NCAA$`OPP FG%`*NCAA$`OPP RPG` + NCAA$`OPP FG%`*NCAA$BKPG + NCAA$`OPP RPG`*NCAA$StealsPG )
```

The model shows the same stats and summary with the transformed variable (Y_WIN)

Plotting the residuals with the new dependent variable (Y_WIN).

```
plot(m3)
```

With the introduction of the transformed variable $\sin^{-1}$(sqrt(Win Pct), the residuals vs fitted graph shows an equal and a better variance plot.



Summary

Final Model:

```
m2 <- lm(NCAA$`Win Pct.` ~ NCAA$`OPP PPG` + NCAA$`OPP FG%` + NCAA$`Opp 3P FG%` +
NCAA$`OPP RPG` + NCAA$BKPG + NCAA$StealsPG + NCAA$`OPP FG%`*NCAA$`OPP RPG` +
NCAA$`OPP FG%`*NCAA$BKPG + NCAA$`OPP RPG`*NCAA$StealsPG )
```

All the defensive independent variables considered under NCAA dataset are showing a moderate to strong linear relationship with 'win percentage'.

The regression model is producing an average adjusted R squared value of 0.55 along with an acceptable value of p test for all the independent variables and interaction terms.

Transformation:

```
NCAA$Y_WIN <- asin(sqrt(NCAA$`Win Pct.`/100))
```

The residual plot is showing a binomial formation and hence the transformation was applied on the dependent variable (Win Percentage).

The summary of the model with the transformed dependent variable is producing the same stats as the selected final model.

The backward selection process and the regression model are producing the same set of variables.

Correlation and VIF was applied on the model to determine no multicollinearity exists between any of the independent variables.

To conclude all the variables (independent and interaction) can be included in the model and can be perfectly used to predict the value of dependent variable (Win Percentage) for NCAA tournament.

Neutral Statistics Analysis (Piotr Senkow)

Problem:

My primary focus was on determining which of the neutral statistics had the greatest predicting power of the target variable *Winning Percentage*. These were statistics that did not belong in either the offensive or defensive categories.

Goal:

Create the least complex linear model that best explains the variance in the target variable *Winning Percentage* using neutral statistics as my independent variables. I will eliminate any independent variables that display multicollinearity or decrease the performance of my model. I will also attempt to transform any variables as the need arises. Lastly,I will evaluate the residuals of my final model and make sure they are healthy.

Data:

See the *Data* section above for further details of the advanced statistics data set, variables, and preparation used by this study.

Correlation / Multicollinearity:

I created a correlation matrix of all these variables against each other to see which variables are correlated with *Winning Percentage* and to begin my investigation of any multicollinearity occurrences between the independent variables. Here is the resulting correlation matrix.

	Winning %	Scoring Margin	Rebound Margin	Assist TO Ratio	Turnovers per Game	Personal Fouls per Game
Winning %	1	0.941970	0.569439	0.663986	-0.536923	-0.329206
Scoring Margin	0.941970	1	0.630104	0.683094	-0.544132	-0.271252
Rebound Margin	0.569439	0.630104	1	0.259118	-0.132544	-0.129904
Assist TO Ratio	0.663986	0.683094	0.259118	1	-0.747399	-0.355016
Turnovers per Game	-0.536923	-0.544132	-0.132544	-0.747399	1	0.423891
Personal Fouls per	-0.329206	-0.271252	-0.129904	-0.355016	0.423891	1

Game						

With this correlation matrix it became apparent to me that *Scoring Margin* is strongly correlated with *Winning Percentage* and that variables *Rebound Margin* and *Assist Turnover Ratio* were moderately correlated with the target variable. There were slight correlations between some of the independent variables, but not enough for me to suspect the presence of multicollinearity amongst the independent variables (<0.9).

I then ran a Variance Inflation Factor on all the variables and I concluded that there exists no multicollinearity since all the values are >= 10. All the variables are independent of each other. The VIF results can be seen below:

```
  SCR.MAR           REB.MAR   Assist.TO.Ratio              TOPG Personal.Fouls.PG
 3.337968          2.058256          2.793088          2.162658          1.112572
```

First-order linear regression model:

Using the programming language R, I constructed a first-order linear model consisting of all the neutral statistics as independent variables predicting the target variable *Winning Percentage*. The summary of the model can be seen below:

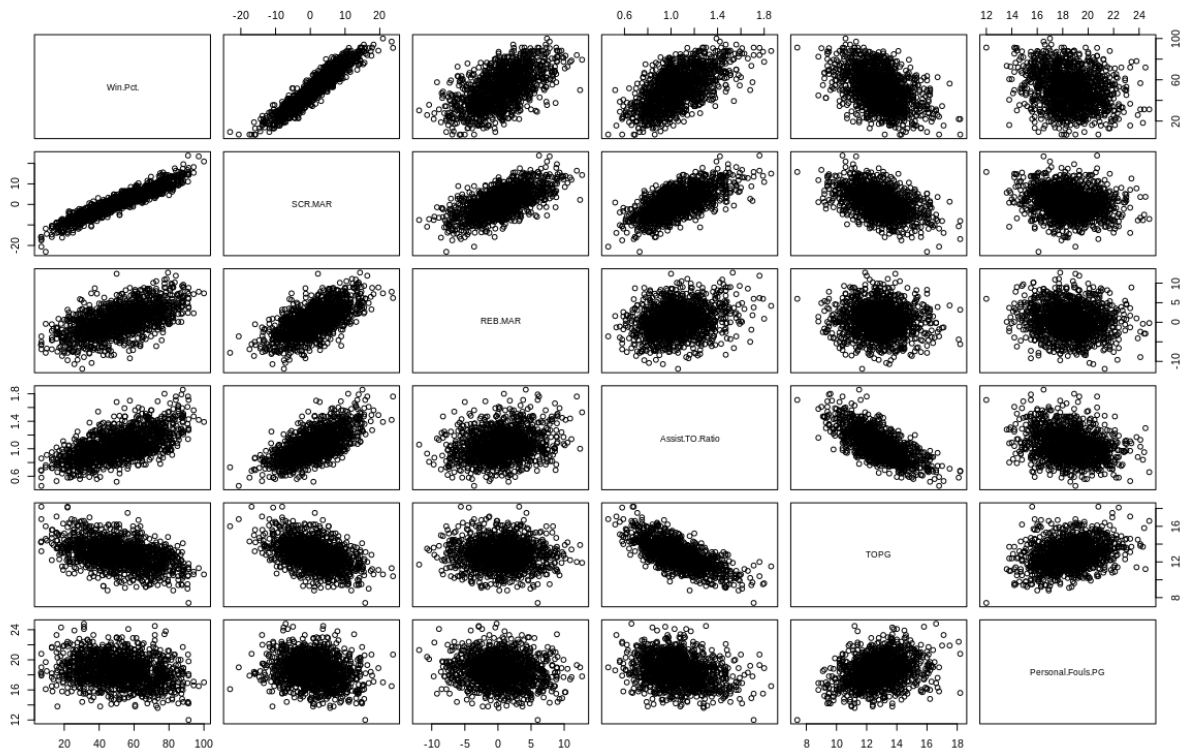
```
Call:
lm(formula = Win.Pct. ~ ., data = newestdata)

Residuals:
     Min       1Q   Median       3Q      Max
-18.5510  -4.0825   0.0536   3.9788  20.1252

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       59.17960    3.29504  17.960  < 2e-16 ***
SCR.MAR            2.68914    0.04561  58.954  < 2e-16 ***
REB.MAR           -0.16038    0.06110  -2.625  0.00876 **
Assist.TO.Ratio   -2.48394    1.28308  -1.936  0.05308 .
TOPG              -0.09076    0.16212  -0.560  0.57569
Personal.Fouls.PG -0.35846    0.09139  -3.922  9.2e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.952 on 1400 degrees of freedom
Multiple R-squared:  0.8864,    Adjusted R-squared:  0.886
F-statistic:  2186 on 5 and 1400 DF,  p-value: < 2.2e-16
```

We can reject the null hypothesis that none of the independent variables contribute to the variance in *Winning Percentage* since the F-stat for this model was very low (2.2e-16), proving significance at a 95% confidence level. We can also accept the null hypothesis that one or more of the independent variables contributes to the variance of the target variable. Variables *Scoring Margin, Rebound Margin, Assist Turnover Ratio,* and *Personal Fouls per Game* are all significant at a 95% confidence level so we can accept the alternate hypothesis that they are related to *Winning Percentage*. Removing the insignificant independent variable *Turnovers per Game* (Pr>|t| = 0.57) did not increase the adjusted R-squared value so we decided to keep it in the neutral statistics first order linear models. Here is a summary of the model without *Turnovers per Game*:

```
Call:
lm(formula = Win.Pct. ~ SCR.MAR + REB.MAR + Assist.TO.Ratio +
    Personal.Fouls.PG, data = newestdata)

Residuals:
     Min       1Q   Median       3Q      Max
-18.6257  -4.0749   0.0232   4.0266  19.9708

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       57.79428    2.17521  26.569  < 2e-16 ***
SCR.MAR            2.69349    0.04494  59.939  < 2e-16 ***
REB.MAR           -0.16857    0.05931  -2.842  0.00454 **
Assist.TO.Ratio   -2.11403    1.09959  -1.923  0.05474 .
Personal.Fouls.PG -0.36752    0.08992  -4.087 4.62e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.951 on 1401 degrees of freedom
Multiple R-squared:  0.8864,    Adjusted R-squared:  0.8861
F-statistic:  2733 on 4 and 1401 DF,  p-value: < 2.2e-16
```

Variable transformation:

In terms of variable transformation, I plotted all the neutral statistics against the target variable to see if there was a need for any second-order terms or interaction terms. The only trend I noticed was a slight curve when comparing *Winning Percentage* vs. *Assist Turnovers Ratio*. The plot can be seen below:



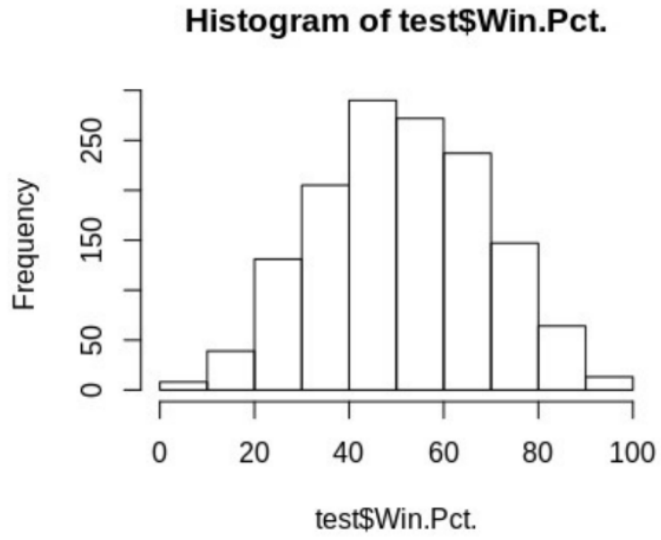
As a result, I created a second order term of *Assist Turnovers Ratio,* which I named *Assistsq,* and I created a new linear model including it. The summary can be found below:

```
Call:
lm(formula = Win.Pct. ~ SCR.MAR + REB.MAR + Assist.TO.Ratio +
    Personal.Fouls.PG + Assistsq, data = newestdata)

Residuals:
     Min       1Q   Median       3Q      Max
-18.2882  -4.1085   0.0568   3.9926  20.1173

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       55.14010    3.67050  15.022  < 2e-16 ***
SCR.MAR            2.68902    0.04521  59.472  < 2e-16 ***
REB.MAR           -0.16469    0.05947  -2.769  0.00569 **
Assist.TO.Ratio    2.92611    5.72074   0.511  0.60909
Personal.Fouls.PG -0.37000    0.08997  -4.112 4.14e-05 ***
Assistsq          -2.26047    2.51788  -0.898  0.36946
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.951 on 1400 degrees of freedom
Multiple R-squared:  0.8865,    Adjusted R-squared:  0.8861
F-statistic:  2187 on 5 and 1400 DF,  p-value: < 2.2e-16
```

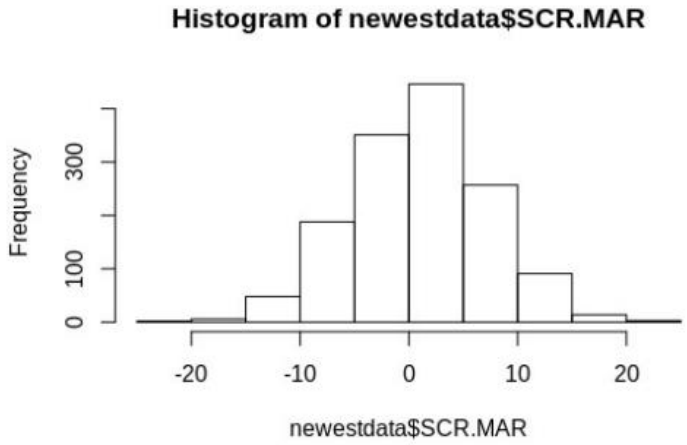
Due to its insignificance on a 95% confidence interval and since the model barely increased, I decided to omit the second order term *Assistsq* to reduce the complexity of the model.

Next, I investigated how normal the data was for each term by creating histograms to see if there were any log transformations needed to be made. It seemed to me that each term had a very normal distribution, eliminating the need for any log transformations of independent variables.

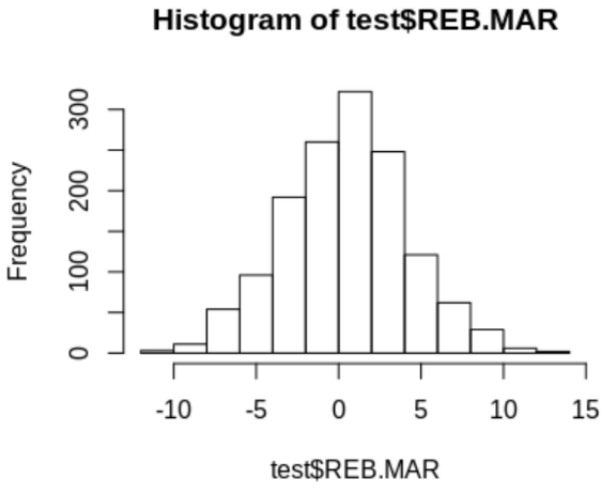
Histogram of *Winning Percentage:*



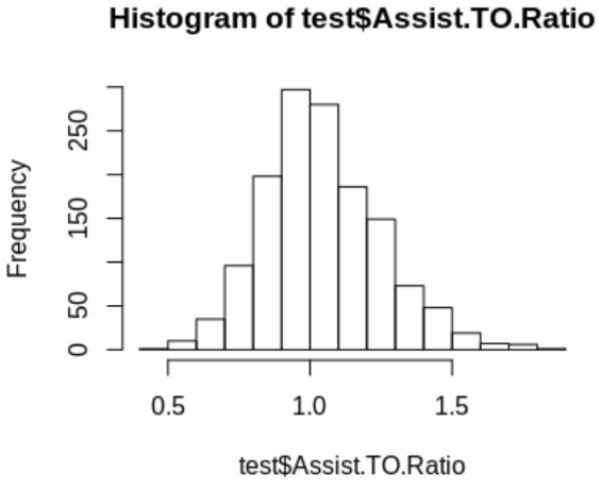
Histogram of *Scoring Margin:*



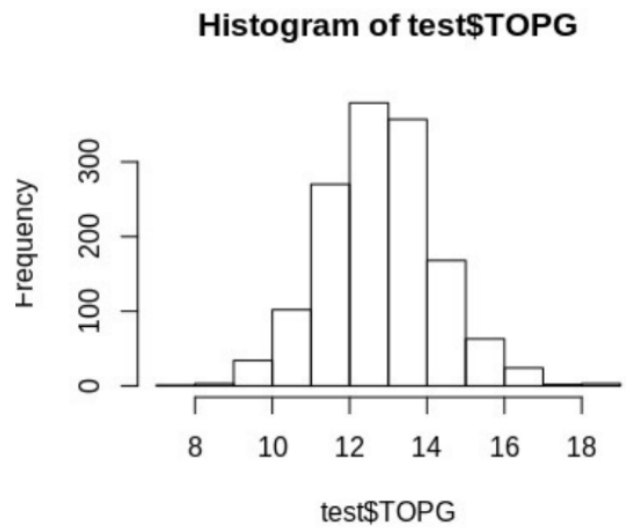
Histogram of *Rebound Margin:*



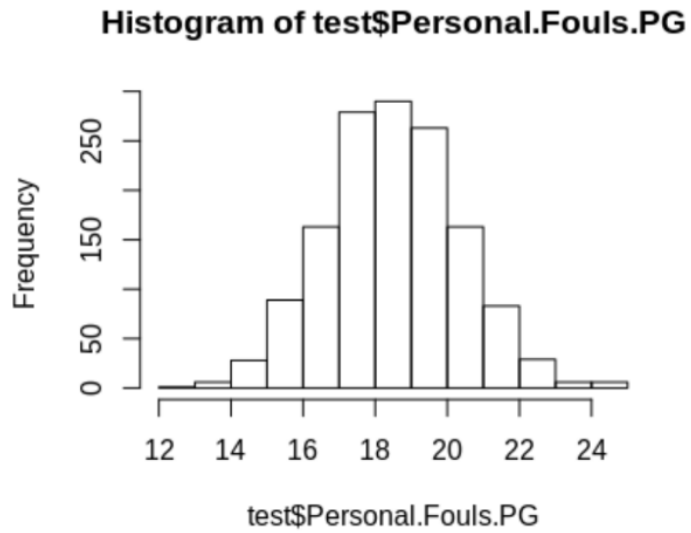
Histogram of *Assist Turnover Ratio*



Histogram of *Turnovers per Game:*



Histogram of *Personal Fouls per Game:*

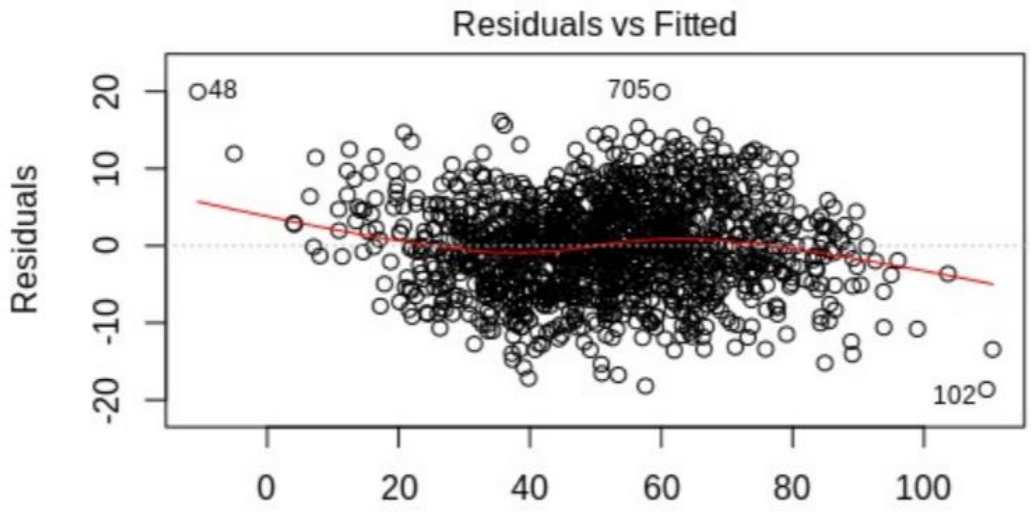


Residual Analysis:

There are four I was looking for when evaluating the residuals of my first-order linear regression model. Homoscedasticity, normality, a mean of 0, and independence of other variables.

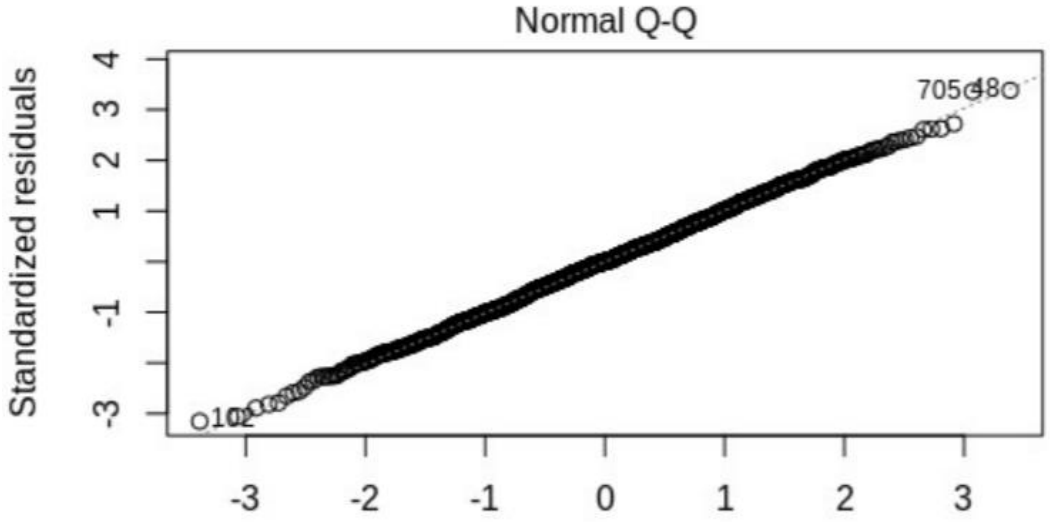
Homoscedasticity:

Plotting the Residuals against the fitted values, it was apparent that the errors are homoscedastic as they tend to be equal in width relative to the red line on both sides. This can be witnessed in the *Residuals vs Fitted* graph below:



Normality:

Creating a QQ plot of the residuals reveals that there is a normal distribution of the errors.



Mean of Zero:

After rounding, the residuals mean was found to be 0.

```
> sum(newestmodel$residuals)
[1] 2.448736e-13
> mean(newestmodel$residuals)
[1] 1.732685e-16
```

Independence:

A D-W Statistics value of 1.99307 shows residual independence as we can reject to fail the null hypothesis that the residuals aren't correlated.

```
> durbinWatsonTest(newestmodel)
 lag Autocorrelation D-W Statistic p-value
   1     0.003149388       1.99307   0.924
 Alternative hypothesis: rho != 0
```

Final Model / Summary of Findings:

The final model built using the neutral statistics ended up being the initial first-order linear regression model that was built as it appeared that the performance of the model did not increase when removing or transforming any of the independent variables. All but one, *Turnovers per Game*, independent variables from the neutral statistics were significant on a 95% confidence level in predicting the target variable *Winning Percentage*. *Turnovers per Game* were left in the final linear regression model simply because removing them did not improve the models performance (increase in 0.01 for adjusted R-squared). Likewise, the F-stat for this model was very low (2.2e-16), proving significance at a 95% confidence level that allowed us to accept the null hypothesis that one or more of the independent variables contributes to the variance of the target variable. The summary of the final model can be found below:

```
Call:
lm(formula = Win.Pct. ~ ., data = newestdata)

Residuals:
     Min       1Q   Median       3Q      Max
-18.5510  -4.0825   0.0536   3.9788  20.1252

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       59.17960    3.29504  17.960  < 2e-16 ***
SCR.MAR            2.68914    0.04561  58.954  < 2e-16 ***
REB.MAR           -0.16038    0.06110  -2.625  0.00876 **
Assist.TO.Ratio   -2.48394    1.28308  -1.936  0.05308 .
TOPG              -0.09076    0.16212  -0.560  0.57569
Personal.Fouls.PG -0.35846    0.09139  -3.922  9.2e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.952 on 1400 degrees of freedom
Multiple R-squared:  0.8864,    Adjusted R-squared:  0.886
F-statistic:  2186 on 5 and 1400 DF,  p-value: < 2.2e-16
```

All Basic Statistics Data Analysis (Srikanth Nanduri (previously Riley Edwards))

Problem

Up until now, we categorized and analyzed the data in silos: Offensive, Defensive and Neutral. While this helps understand the independent category specific variables on Win Percentage, a basketball game is composed and impacted by both offensive and defensive characteristics.

Goal

The goal of this assessment is to determine the best model and predictor of the win percentage based on all the basic statistics which is inclusive of offensive, defensive and neutral variables.
Note: This analysis does not incorporate any of the advanced statistics in the modeling and evaluation.

Data

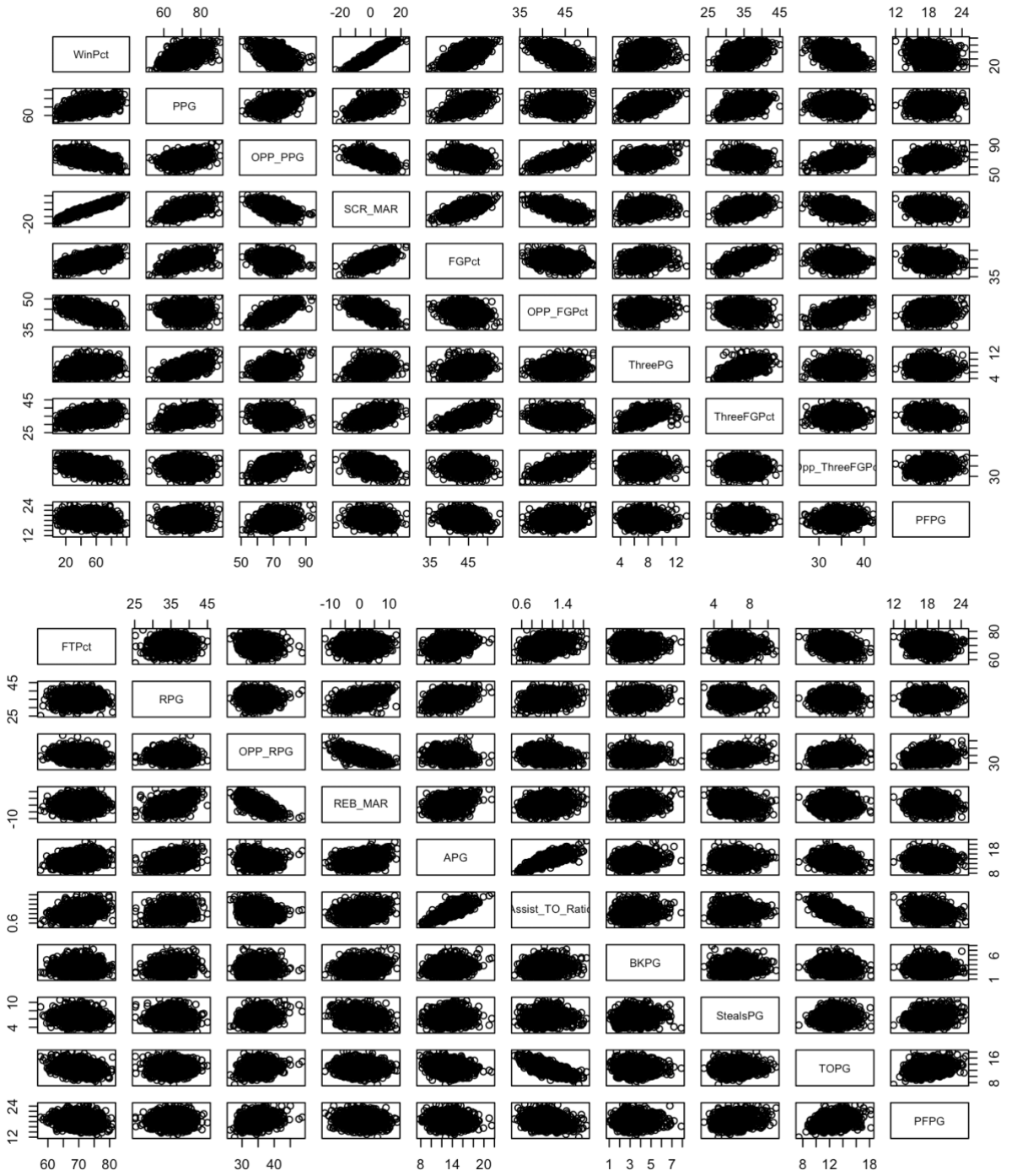
See the *Data* section above for further details of the defense statistics data set, variables, and preparation used by this study.

To begin, I reviewed the basic summary statistics associated with the offensive data as shown below.

```
     WinPct            PPG          OPP_PPG         SCR_MAR           FGPct        OPP_FGPct       ThreePG
 Min.   :  6.7   Min.   :52.0   Min.   :50.7   Min.   :-23.10   Min.   :34.7   Min.   :35.5   Min.   : 3.30
 1st Qu.: 38.7   1st Qu.:68.1   1st Qu.:67.4   1st Qu.: -3.23   1st Qu.:42.4   1st Qu.:42.0   1st Qu.: 6.30
 Median : 51.5   Median :72.0   Median :70.8   Median :  1.10   Median :44.2   Median :43.5   Median : 7.20
 Mean   : 51.6   Mean   :72.0   Mean   :70.9   Mean   :  1.10   Mean   :44.2   Mean   :43.6   Mean   : 7.27
 3rd Qu.: 63.6   3rd Qu.:76.0   3rd Qu.:74.6   3rd Qu.:  5.30   3rd Qu.:45.9   3rd Qu.:45.2   3rd Qu.: 8.20
 Max.   :100.0   Max.   :90.4   Max.   :94.6   Max.   : 23.80   Max.   :53.2   Max.   :51.2   Max.   :13.50
   ThreeFGPct   Opp_ThreeFGPct     FTPct            RPG          OPP_RPG        REB_MAR            APG
 Min.   :25.2   Min.   :26.3   Min.   :57.6   Min.   :25.2   Min.   :26.3   Min.   :-11.90   Min.   : 7.8
 1st Qu.:32.9   1st Qu.:33.0   1st Qu.:67.9   1st Qu.:33.6   1st Qu.:33.1   1st Qu.: -2.00   1st Qu.:12.1
 Median :34.6   Median :34.5   Median :70.2   Median :35.5   Median :34.8   Median :  0.50   Median :13.3
 Mean   :34.7   Mean   :34.5   Mean   :70.2   Mean   :35.4   Mean   :34.9   Mean   :  0.47   Mean   :13.3
 3rd Qu.:36.5   3rd Qu.:36.0   3rd Qu.:72.7   3rd Qu.:37.3   3rd Qu.:36.6   3rd Qu.:  2.80   3rd Qu.:14.5
 Max.   :45.0   Max.   :42.0   Max.   :81.0   Max.   :45.0   Max.   :48.7   Max.   : 12.70   Max.   :21.5
 Assist_TO_Ratio      BKPG          StealsPG          TOPG           PFPG
 Min.   :0.46    Min.   :1.00   Min.   : 2.90   Min.   : 7.4   Min.   :12.0
 1st Qu.:0.91    1st Qu.:2.70   1st Qu.: 5.50   1st Qu.:11.9   1st Qu.:17.3
 Median :1.03    Median :3.30   Median : 6.20   Median :12.8   Median :18.4
 Mean   :1.05    Mean   :3.41   Mean   : 6.24   Mean   :12.8   Mean   :18.5
 3rd Qu.:1.18    3rd Qu.:4.10   3rd Qu.: 6.90   3rd Qu.:13.7   3rd Qu.:19.7
 Max.   :1.86    Max.   :7.90   Max.   :10.90   Max.   :18.2   Max.   :24.8
```

Plot (*Win Percentage vs. all dependent variables*)

Subsequently, I plotted the independent variables in order to visually identify linear relationships between win percentage and the other variables.



Based on the analysis of the data, we noticed that the strongest linear relationship was between WinPct and Scr_Mar. Further review of the data shows additional areas where variables are correlated:

1. APG versus Assist_TO_Ratio
2. Turnover Per Game versus Assist_TO_Ratio
3. Scoring Margin versus FG Percentage
4. Opponent Points Per Game versus Opponent Field Goal Percentage
5. Field Goal Percentage versus Win Percentage

This is supported by the correlation matrix as shown below.

Correlations with target

There is strong correlation between the following variables as identified in the correlation matrix.

1. APG vs Assist TO Ratio: 0.8113
2. Turnover Per Game versus Assist TO Ration: -0.7065
3. Scoring Margin Versus Win Percentage: 0.940
4. Field Goal Percentage versus Win Percentage: 0.641
5. Rebound Margin versus Opponent RPG
6. Rebound Margin versus RPG

	WinPct	PPG	OPP_PPG	SCR_MAR	FGPct	OPP_FGPct	ThreePG	ThreeFGPct	Opp_ThreeFGPct	FTPct
WinPct	1.000	0.5333	-0.5080	0.940	0.6412	-0.6071	0.22939	0.44167	-0.48213	0.2525
PPG	0.533	1.0000	0.3862	0.572	0.6527	0.0103	0.55813	0.49266	-0.02057	0.3788
OPP_PPG	-0.508	0.3862	1.0000	-0.535	-0.1015	0.7204	0.28589	-0.02289	0.51862	0.0862
SCR_MAR	0.940	0.5721	-0.5353	1.000	0.6877	-0.6314	0.25634	0.47097	-0.48056	0.2703
FGPct	0.641	0.6527	-0.1015	0.688	1.0000	-0.1876	0.27721	0.65326	-0.12900	0.2864
OPP_FGPct	-0.607	0.0103	0.7204	-0.631	-0.1876	1.0000	0.15066	-0.05260	0.62052	0.0403
ThreePG	0.229	0.5581	0.2859	0.256	0.2772	0.1507	1.00000	0.56060	0.06147	0.3347
ThreeFGPct	0.442	0.4927	-0.0229	0.471	0.6533	-0.0526	0.56060	1.00000	-0.00707	0.3597
Opp_ThreeFGPct	-0.482	-0.0206	0.5186	-0.481	-0.1290	0.6205	0.06147	-0.00707	1.00000	0.0555
FTPct	0.253	0.3788	0.0862	0.270	0.2864	0.0403	0.33470	0.35969	0.05547	1.0000
RPG	0.380	0.4788	0.0410	0.402	0.2233	-0.3406	0.14716	0.20823	-0.19557	0.0181
OPP_RPG	-0.453	0.0741	0.6448	-0.505	-0.4665	0.3075	0.17406	-0.29057	0.14678	-0.1670
REB_MAR	0.586	0.2864	-0.4303	0.645	0.3820	-0.5374	-0.11420	0.16137	-0.29967	0.0711
APG	0.499	0.6639	0.0733	0.542	0.6376	-0.1107	0.44831	0.47787	-0.09089	0.2442
Assist_TO_Ratio	0.618	0.5654	-0.1598	0.659	0.5998	-0.1940	0.43790	0.51523	-0.13526	0.3139
BKPG	0.351	0.1990	-0.1942	0.356	0.1574	-0.4979	-0.08412	0.00832	-0.21935	-0.0157
StealsPG	0.211	0.2247	-0.0094	0.214	0.0799	0.0128	-0.00794	-0.08512	-0.08483	-0.0875
TOPG	-0.456	-0.1848	0.3442	-0.474	-0.2517	0.1928	-0.21613	-0.30637	0.11150	-0.2476
PFPG	-0.230	0.0351	0.2763	-0.214	-0.2189	0.1529	-0.10220	-0.15765	0.12130	-0.1488

	RPG	OPP_RPG	REB_MAR	APG	Assist_TO_Ratio	BKPG	StealsPG	TOPG	PFPG
WinPct	0.3795	-0.4535	0.5857	0.4992	0.6184	0.35104	0.21114	-0.4559	-0.2302
PPG	0.4788	0.0741	0.2864	0.6639	0.5654	0.19904	0.22473	-0.1848	0.0351
OPP_PPG	0.0410	0.6448	-0.4303	0.0733	-0.1598	-0.19423	-0.00940	0.3442	0.2763
SCR_MAR	0.4021	-0.5052	0.6447	0.5422	0.6587	0.35593	0.21440	-0.4743	-0.2138
FGPct	0.2233	-0.4665	0.3820	0.6376	0.5998	0.15735	0.07994	-0.2517	-0.2189
OPP_FGPct	-0.3406	0.3075	-0.5374	-0.1107	-0.1940	-0.49795	0.01276	0.1928	0.1529
ThreePG	0.1472	0.1741	-0.1142	0.4483	0.4379	-0.08412	-0.00794	-0.2161	-0.1022
ThreeFGPct	0.2082	-0.2906	0.1614	0.4779	0.5152	0.00832	-0.08512	-0.3064	-0.1576
Opp_ThreeFGPct	-0.1956	0.1468	-0.2997	-0.0909	-0.1353	-0.21935	-0.08483	0.1115	0.1213
FTPct	0.0181	-0.1670	0.0711	0.2442	0.3139	-0.01573	-0.08747	-0.2476	-0.1488
RPG	1.0000	-0.0396	0.5639	0.3220	0.2104	0.24961	-0.05771	0.0134	0.0210
OPP_RPG	-0.0396	1.0000	-0.6935	-0.1332	-0.2015	0.03230	0.15405	0.1671	0.2337
REB_MAR	0.5639	-0.6935	1.0000	0.2834	0.2290	0.25191	-0.13720	-0.0484	-0.1111
APG	0.3220	-0.1332	0.2834	1.0000	0.8113	0.16389	0.14269	-0.1865	-0.1323
Assist_TO_Ratio	0.2104	-0.2015	0.2290	0.8113	1.0000	0.13975	0.06020	-0.7065	-0.2894
BKPG	0.2496	0.0323	0.2519	0.1639	0.1397	1.00000	0.11763	-0.0464	-0.0450
StealsPG	-0.0577	0.1541	-0.1372	0.1427	0.0602	0.11763	1.00000	0.0675	0.1893
TOPG	0.0134	0.1671	-0.0484	-0.1865	-0.7065	-0.04636	0.06752	1.0000	0.3168
PFPG	0.0210	0.2337	-0.1111	-0.1323	-0.2894	-0.04500	0.18933	0.3168	1.0000

1st Complete First-Order Model (inclusive of potential interaction terms)

We began with a basic model with all the variables and reviewed the summary statistics related to that model. Below is the summary:

```
Call:
lm(formula = asin(sqrt(WinPct/100)) ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17509 -0.04151  0.00115  0.04077  0.19976

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.02e+00   2.24e-01    4.56  5.5e-06 ***
PPG             1.34e-02   1.09e-02    1.22  0.22191
OPP_PPG        -1.01e-02   1.09e-02   -0.92  0.35579
SCR_MAR         1.64e-02   1.09e-02    1.50  0.13323
FGPct          -3.70e-03   2.26e-03   -1.63  0.10231
OPP_FGPct      -4.06e-03   2.03e-03   -2.00  0.04548 *
ThreePG        -5.87e-03   2.05e-03   -2.85  0.00437 **
ThreeFGPct      1.52e-03   1.01e-03    1.51  0.13115
Opp_ThreeFGPct -2.44e-03   1.04e-03   -2.36  0.01859 *
FTPct          -4.84e-04   6.64e-04   -0.73  0.46589
RPG             2.00e-03   1.02e-03    1.97  0.04923 *
OPP_RPG        -5.46e-03   3.21e-03   -1.70  0.08853 .
REB_MAR         8.01e-03   7.75e-03    1.03  0.30121
APG            -3.21e-03   1.80e-02   -0.18  0.85823
Assist_TO_Ratio 1.55e-01   9.54e-02    1.63  0.10434
BKPG            2.20e-03   2.11e-03    1.04  0.29685
StealsPG        5.95e-04   2.26e-03    0.26  0.79232
TOPG            5.74e-03   5.47e-03    1.05  0.29395
PFPG           -4.18e-03   1.09e-03   -3.84  0.00013 ***
APG_ATOR       -6.17e-03   3.61e-03   -1.71  0.08772 .
TO_ATOR         3.22e-03   1.85e-02    0.17  0.86205
RM_ORPG        -4.90e-04   1.36e-04   -3.61  0.00031 ***
RM_RPG          7.13e-05   1.38e-04    0.52  0.60440
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0617 on 1389 degrees of freedom
Multiple R-squared:  0.893,     Adjusted R-squared:  0.891
F-statistic:  525 on 22 and 1389 DF,  p-value: <2e-16
```

Based on the information from the initial model, the F-test passed while the T-test for a significant number of the variables was insignificant. In order to isolate the variables required for the model, we removed each variable and reviewed the key statistics to determine if they were required for the model to be effective.

```
Call:
lm(formula = asin(sqrt(WinPct/100)) ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17971 -0.04067 -0.00126  0.04180  0.20048

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.953622   0.032168   29.64  < 2e-16 ***
SCR_MAR        0.027926   0.000371   75.36  < 2e-16 ***
Opp_ThreeFGPct -0.003511   0.000793   -4.43  1.0e-05 ***
REB_MAR        -0.002227   0.000570   -3.91  9.8e-05 ***
PFPG           -0.003193   0.000909   -3.51  0.00046 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0607 on 1420 degrees of freedom
Multiple R-squared:  0.895,     Adjusted R-squared:  0.895
F-statistic: 3.02e+03 on 4 and 1420 DF,  p-value: <2e-16
```

This makes sense as Scoring Margin is very strongly correlated with winning percentage. At this point, we questioned whether Scoring Margin was an apt number to be used in predictive models.

Scoring margin is calculated as the average difference between the respective team's score minus the opponents score divided by the number of games played. This is measured to determine the overall margin either teams are defeating their opponents or getting defeated by their opponents. While this is helpful in getting the most predictive model, it is not helpful or insightful in helping a team manager identify which aspect of a game the team needs to improve their win percentage in their overall evaluation of the team.

As such, we decided to remove the Scoring Margin from the evaluation of the win percentage and identify if there is any level of predictive power in the variable to accurately calculate a team's win percentage.

2nd Complete First-Order Model (inclusive of potential interaction terms)

Again, beginning with all the independent variables (except Scoring Margin), the below is the excerpt of the overall model before removing the variables.

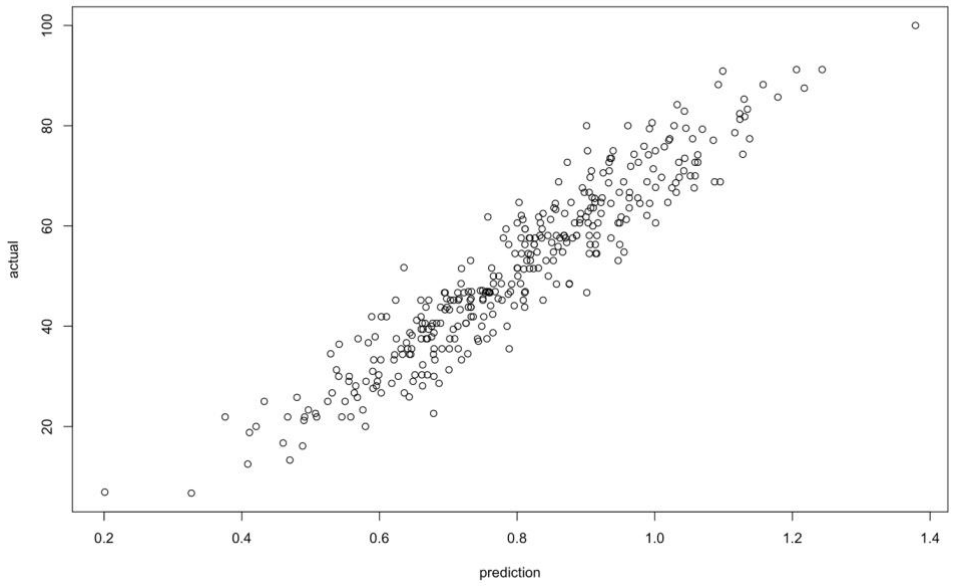
```
Call:
lm(formula = asin(sqrt(WinPct/100)) ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17717 -0.04276 -0.00089  0.04131  0.17125

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.96e-01   2.27e-01    4.39  1.2e-05 ***
PPG              3.07e-02   1.42e-03   21.58  < 2e-16 ***
OPP_PPG         -2.70e-02   1.32e-03  -20.39  < 2e-16 ***
FGPct           -5.01e-03   2.28e-03   -2.19  0.02841 *
OPP_FGPct       -3.63e-03   2.06e-03   -1.76  0.07848 .
ThreePG         -7.13e-03   2.09e-03   -3.41  0.00066 ***
ThreeFGPct       1.36e-03   1.03e-03    1.31  0.18944
Opp_ThreeFGPct  -2.27e-03   1.03e-03   -2.21  0.02734 *
FTPct           -7.78e-04   6.66e-04   -1.17  0.24266
RPG              1.51e-03   1.00e-03    1.51  0.13090
OPP_RPG         -6.54e-03   3.20e-03   -2.05  0.04099 *
REB_MAR          6.22e-03   8.02e-03    0.78  0.43817
APG              5.55e-03   1.91e-02    0.29  0.77218
Assist_TO_Ratio  2.67e-01   9.93e-02    2.69  0.00732 **
BKPG             3.14e-03   2.12e-03    1.48  0.13816
StealsPG        -5.76e-04   2.28e-03   -0.25  0.80045
TOPG             9.22e-03   5.47e-03    1.68  0.09240 .
PFPG            -4.91e-03   1.11e-03   -4.41  1.1e-05 ***
APG_ATOR        -1.19e-02   4.05e-03   -2.92  0.00352 **
TO_ATOR         -1.81e-03   1.95e-02   -0.09  0.92604
RM_ORPG         -4.40e-04   1.43e-04   -3.08  0.00211 **
RM_RPG           6.83e-05   1.42e-04    0.48  0.63139
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0623 on 1375 degrees of freedom
Multiple R-squared:  0.89,      Adjusted R-squared:  0.888
F-statistic:  530 on 21 and 1375 DF,  p-value: <2e-16
```

The model shows that the P-test passed however, we need to reevaluate each of the variables to see if they are significantly required for the model. The R-squared value from the model shows an 88% predictability, so scoring margin might not be a requirement for a strong predictive model.

Below is the summary of the final model. The summary shows that the model (tested on the training data) has 5 variables: PPG, OPP_PPG, ThreePG, Opp_RPG, REB_MAR, Assist_TO_Ratio, PFPG, APG_ATOR and RM_ORPG. Both the F-statistic and T-statistics for the model pass and the adjusted R-squared value for the model shows that 89% of the variability is explained by this model.

```
Call:
lm(formula = asin(sqrt(WinPct/100)) ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.18034 -0.04243  0.00025  0.04241  0.16932

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.843082   0.051754   16.29  < 2e-16 ***
PPG              0.028773   0.000609   47.27  < 2e-16 ***
OPP_PPG         -0.027831   0.000673  -41.35  < 2e-16 ***
ThreePG         -0.003203   0.001513   -2.12  0.03449 *
Opp_ThreeFGPct  -0.002631   0.000987   -2.67  0.00778 **
OPP_RPG          0.000940   0.001165    0.81  0.41994
REB_MAR          0.010660   0.004721    2.26  0.02409 *
Assist_TO_Ratio  0.054442   0.035282    1.54  0.12304
PFPG            -0.003931   0.001018   -3.86  0.00012 ***
APG_ATOR        -0.003110   0.001522   -2.04  0.04120 *
RM_ORPG         -0.000353   0.000134   -2.63  0.00856 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.062 on 1390 degrees of freedom
Multiple R-squared:  0.891,     Adjusted R-squared:  0.89
F-statistic: 1.13e+03 on 10 and 1390 DF,  p-value: <2e-16
```

Based on the model comparison against the test values, the below is the chart showing the predicted versus actual values. The correlation value between the predicted and actual is 0.942 which indicates that model predicted values and actual values in the test data have a strong linear relationship. As we can see from the plot, we can understand that a model is relatively accurate when compared to the test data.



Model Validation and Review

Subsequently, we performed a few tests to understand whether the model was aptly built. The first was to review whether the variables in the model were multicollinear. Then we tested the model to see if the same results could be reached using the backward and forward selection process.

Multicollinearity

```
> vif(as1)
          PPG       OPP_PPG       ThreePG  Opp_ThreeFGPct       OPP_RPG       REB_MAR Assist_TO_Ratio          PFPG       APG_ATOR
         4.55          5.31          1.79          1.92          3.48        109.52         18.78          1.21         17.59
      RM_ORPG
       109.30
```

To evaluate whether multicollinearity existed within the model, we evaluated whether any of the variables in the model were more related with each other than the response variables. The model shows 4 variables with high VIF values which indicates multicollinearity. However, per further inspection and observation, the high VIF values are due to terms that were kept in the model because they are a part of the interaction term. As such, we determined that multicollinearity is not a problem within the model.

Backward Elimination

The backward elimination process removed one variable from the system at a time to determine the key variables required to reduce the amount of information lost from the data. Based on the data, the final predicted model is outlined below. The model determined that the interaction terms were necessary for the highest amount of predictive value.

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
WinPct ~ PPG + OPP_PPG + FGPct + OPP_FGPct + ThreePG + ThreeFGPct +
    Opp_ThreeFGPct + FTPct + RPG + OPP_RPG + REB_MAR + APG +
    Assist_TO_Ratio + BKPG + StealsPG + TOPG + PFPG + APG_ATOR +
    TO_ATOR + RM_ORPG + RM_RPG

Final Model:
WinPct ~ PPG + OPP_PPG + FGPct + ThreePG + ThreeFGPct + Opp_ThreeFGPct +
    Assist_TO_Ratio + TOPG + PFPG + APG_ATOR + RM_ORPG + RM_RPG


          Step Df Deviance Resid. Df Resid. Dev  AIC
1                              1730      59472 6219
2   - StealsPG  1    0.734      1731      59472 6217
3    - TO_ATOR  1    0.946      1732      59473 6215
4    - REB_MAR  1    5.702      1733      59479 6214
5        - APG  1   10.460      1734      59489 6212
6      - FTPct  1   45.184      1735      59535 6211
7       - BKPG  1   58.774      1736      59593 6211
8        - RPG  1   66.093      1737      59659 6211
9    - OPP_RPG  1   44.934      1738      59704 6210
10 - OPP_FGPct  1   60.181      1739      59765 6210
```

In review of the model summary against the complete dataset, the variables selected through the backward elimination model had a few issues. The first issue was that for a few select variables the T-test failed. This is a strong model that explains 89% of the variability in the data.

```
Call:
lm(formula = asin(sqrt(WinPct/100)) ~ PPG + OPP_PPG + FGPct +
    ThreePG + ThreeFGPct + Opp_ThreeFGPct + Assist_TO_Ratio +
    TOPG + PFPG + APG_ATOR + RM_ORPG + RM_RPG, data = n1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17606 -0.04197 -0.00013  0.04105  0.20237

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.18e-01   9.81e-02    7.32  3.8e-13 ***
PPG              2.97e-02   7.00e-04   42.43  < 2e-16 ***
OPP_PPG         -2.82e-02   6.16e-04  -45.81  < 2e-16 ***
FGPct           -2.40e-03   1.18e-03   -2.04  0.04191 *
ThreePG         -6.71e-03   1.65e-03   -4.07  4.9e-05 ***
ThreeFGPct       1.89e-03   8.70e-04    2.18  0.02960 *
Opp_ThreeFGPct  -2.87e-03   8.33e-04   -3.45  0.00057 ***
Assist_TO_Ratio  2.06e-01   8.27e-02    2.49  0.01303 *
TOPG             8.47e-03   3.84e-03    2.20  0.02758 *
PFPG            -4.08e-03   9.31e-04   -4.38  1.3e-05 ***
APG_ATOR        -8.19e-03   3.03e-03   -2.71  0.00687 **
RM_ORPG         -3.21e-04   7.86e-05   -4.08  4.6e-05 ***
RM_RPG           2.41e-04   7.67e-05    3.14  0.00171 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.062 on 1739 degrees of freedom
Multiple R-squared:  0.891,     Adjusted R-squared:  0.89
F-statistic: 1.18e+03 on 12 and 1739 DF,  p-value: <2e-16
```

Forward Selection

Similarly, for the forward selection process, the below was the final model showing the required variables for the least amount of information lost by not including the variable.

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
WinPct ~ 1

Final Model:
WinPct ~ FGPct + OPP_FGPct + PPG + OPP_PPG + RM_ORPG + PFPG +
    ThreePG + Opp_ThreeFGPct + ThreeFGPct + REB_MAR + APG_ATOR


              Step Df Deviance Resid. Df Resid. Dev   AIC
1                                 1751     535934 10029
2           + FGPct  1 220329.0      1750     315605  9103
3       + OPP_FGPct  1 131669.7      1749     183936  8160
4             + PPG  1  31882.4      1748     152053  7828
5         + OPP_PPG  1  89800.0      1747      62253  6265
6         + RM_ORPG  1    621.2      1746      61632  6250
7            + PFPG  1    520.1      1745      61112  6237
8         + ThreePG  1    444.5      1744      60667  6226
9  + Opp_ThreeFGPct  1    335.5      1743      60332  6219
10     + ThreeFGPct  1    191.6      1742      60140  6215
11        + REB_MAR  1    150.3      1741      59990  6213
12       + APG_ATOR  1     81.5      1740      59909  6212
>
```

In evaluation of the summary of the model created to the forward selection process, the forward selection process found the same variable requirement as the backward elimination process. However, similar to the backward elimination process, the model is inclusive of multiple variables that fail the T-test and the R-squared value is 89% which is the same value as the model initially created.

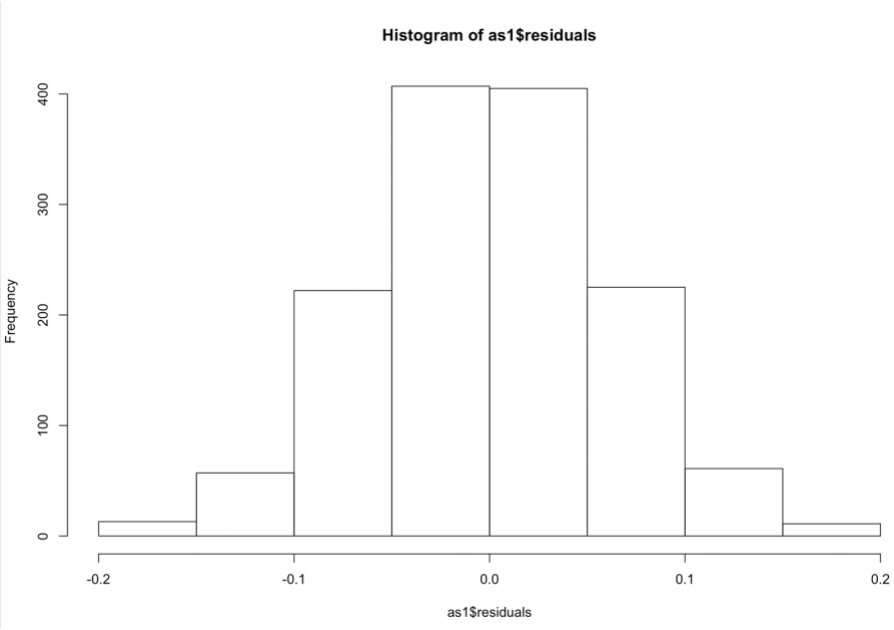
```
Call:
lm(formula = asin(sqrt(WinPct/100)) ~ FGPct + OPP_FGPct + PPG +
    OPP_PPG + RM_ORPG + PFPG + ThreePG + Opp_ThreeFGPct + ThreeFGPct +
    REB_MAR + APG_ATOR, data = n1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17834 -0.04152  0.00035  0.04066  0.20134

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.954693   0.047111   20.26  < 2e-16 ***
FGPct          -0.002011   0.001132   -1.78  0.07582 .
OPP_FGPct      -0.001308   0.001084   -1.21  0.22793
PPG             0.029419   0.000674   43.66  < 2e-16 ***
OPP_PPG        -0.027840   0.000619  -44.97  < 2e-16 ***
RM_ORPG        -0.000435   0.000120   -3.63  0.00029 ***
PFPG           -0.004204   0.000904   -4.65  3.6e-06 ***
ThreePG        -0.006296   0.001634   -3.85  0.00012 ***
Opp_ThreeFGPct -0.002539   0.000846   -3.00  0.00274 **
ThreeFGPct      0.001961   0.000860    2.28  0.02271 *
REB_MAR         0.012483   0.004214    2.96  0.00310 **
APG_ATOR       -0.000682   0.000499   -1.37  0.17233
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.062 on 1740 degrees of freedom
Multiple R-squared:  0.891,     Adjusted R-squared:  0.89
F-statistic: 1.29e+03 on 11 and 1740 DF,  p-value: <2e-16
```

Review of the 4 Assumptions

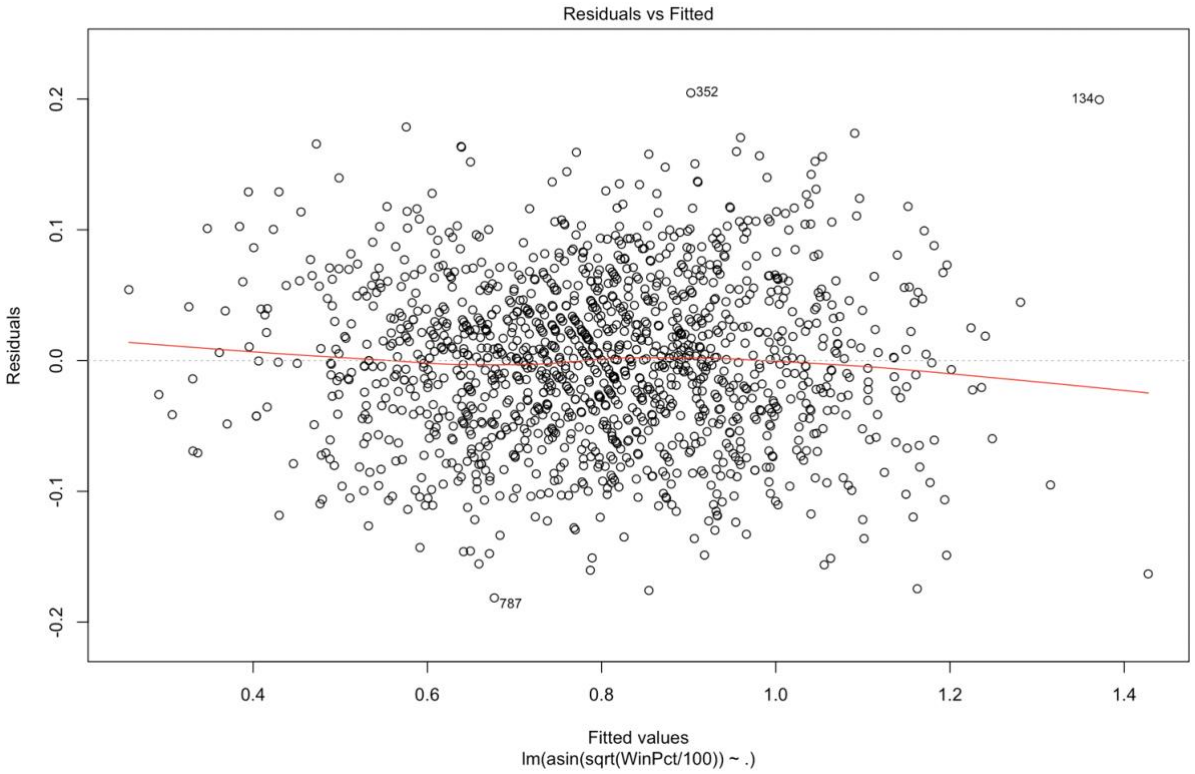
Next, we reviewed and evaluated whether the 4 assumptions of linear models were met. The four assumptions are as follows:

5. Normal Distribution of Residuals
6. Residuals are Homoscedastic
7. Mean sum of residuals is 0
8. Residuals are independent from one another

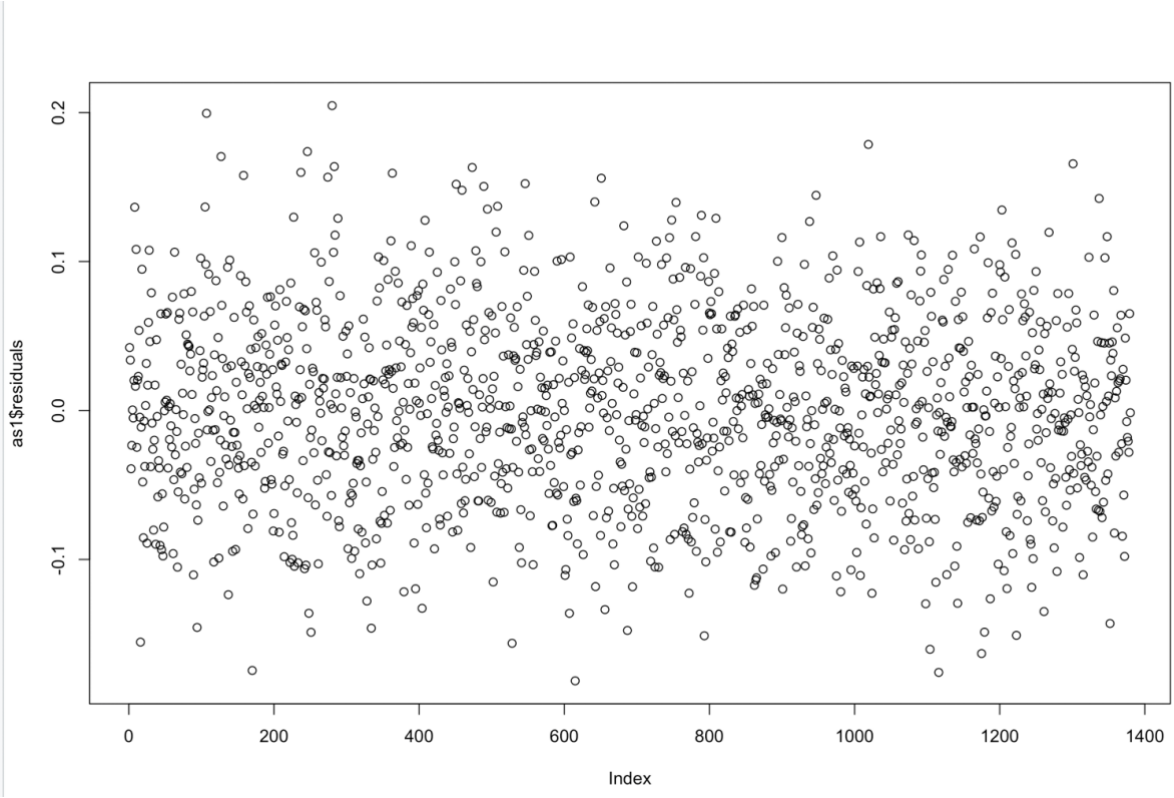
Beginning with the normal distribution, the histogram below shows that the residuals are normally distributed. This is shown through the plot of each residual in the first histogram and plotting the z-scores of the residuals.



Per review of the sum of the mean of the residuals, the value was very close to 0 (5.48e-19). Although the plot of the residuals versus fitted shows a slight unequal variance, the response variable has been already transformed to counteract this. As such, the model is acceptable and does not require additional transformations.



Per review of the residuals plot on the index, we concluded that residuals plot appears to be random. As such, we concluded that the residuals are independent.



Based on the Durbin Watson Test, the value of the statistic was between 0 - 2.5. As such, the residuals are independent of one another.

```
        Durbin-Watson test

data:  as1
DW = 2, p-value = 0.8
alternative hypothesis: true autocorrelation is greater than 0
```

Lastly, we measured the models predictive value against the test data set. The model was 94% accurate in predicting the test values.

```
> cor(prediction, actual)
[1] 0.938
```

Summary

In summary, the model details are the following:

1. Below is the final model. It predicts 89% of the variability in Win Percentage.

```
Call:
lm(formula = asin(sqrt(WinPct/100)) ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.18034 -0.04243  0.00025  0.04241  0.16932

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.843082   0.051754   16.29  < 2e-16 ***
PPG              0.028773   0.000609   47.27  < 2e-16 ***
OPP_PPG         -0.027831   0.000673  -41.35  < 2e-16 ***
ThreePG         -0.003203   0.001513   -2.12  0.03449 *
Opp_ThreeFGPct  -0.002631   0.000987   -2.67  0.00778 **
OPP_RPG          0.000940   0.001165    0.81  0.41994
REB_MAR          0.010660   0.004721    2.26  0.02409 *
Assist_TO_Ratio  0.054442   0.035282    1.54  0.12304
PFPG            -0.003931   0.001018   -3.86  0.00012 ***
APG_ATOR        -0.003110   0.001522   -2.04  0.04120 *
RM_ORPG         -0.000353   0.000134   -2.63  0.00856 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.062 on 1390 degrees of freedom
Multiple R-squared:  0.891,     Adjusted R-squared:  0.89
F-statistic: 1.13e+03 on 10 and 1390 DF,  p-value: <2e-16
```

2. The model passes the F-test and the T-test for the variables are significant, therefore we can reject the null hypothesis and accept the alternative beta values identified through the model.
3. The histogram of the residuals from the model shows the residuals are normally distributed.
4. The sum of the mean of the residuals, the value was very close to 0 (5.48e-19).
5. The plot of the residuals versus the fitted model shows a slight unequal variance, however, this was reduced through transformation of the response variable to arcsin(sqrt(WinPct)). As such, the residuals are homoscedastic.
6. The plot of the residuals versus index showed no pattern which shows that the residuals are independent. This is supported by the DurbinWatson Test.

Advanced Statistics Analysis (Kevin Gasiorowski)

Problem

In recent years, advanced statistics have gained in popularity among sports teams' ownership, management, coaches, and their fans. As an example, management uses these statistics when assembling a team and fans use them when betting on games. These advanced statistics go beyond the traditional box scores and use advanced formulas to rank teams and players. For example, the advanced basketball statistic of Offensive Rating created by statistician Dean Oliver is calculated using the following formula (3):

- Offensive Rating = (Points Produced / Individual Possessions) x OAPOW × PPG + FTM/FT * 3pt% + FG%

But just how well do these advanced statistics actually predict the overall winning percentage of a team for a season? This is the question we will look to address in this study.

Goal

While baseball was one of the first American sports to aggressively adopt advanced statistics as a core component to team management, basketball has recently also fully embarrassed the use of advanced statistics (5).

In this study, we examine how well advanced basketball statistics predict overall winning percentage of NCAA Division I men's basketball teams by creating the best possible first-order regression model.

We then expand on the best fit first-order regression model by introducing interaction and second-order terms along with feature engineering to predict regular season winning percentage from only advanced statistics.

We make the distinction between "basic" and "advanced" statistics as follows:

- Basic statistics: Traditional statistics that typically include the summary or average of a single observation typically per game, such as points per game, rebounds per game, etc.
- Advanced statistics: Modern statistics that typically require a formula that combines many of the basic statistics into a single rating where a key tenant is that the statistics are evaluated over possessions, not games.

Data

See the *Data* section above for further details of the advanced statistics data set, variables, and preparation used by this study.

Correlations with target

One assumption we make on the data based on our knowledge of the sport, is that offensive statistics should be positively correlated with Win/Loss % and defensive (or offensive statistics of the opponent) should be negatively correlated with Win/Loss %.

Next we examine the correlation between our target variable Win/Loss % against all other variables in the data set (all other advanced statistics).

```
                          [,1]
Pace                -0.06429262
ORtg                 0.78248802
FTr                  0.14810170
X3PAr                0.04090476
TrueShootingPer      0.62685167
TotalReboundPer      0.57706399
AssistPer            0.25458589
StealPer             0.22939252
BlockPer             0.35610224
EffectiveFGPer       0.60451834
TurnoverPer         -0.48977648
OReboundPer          0.28881899
FT_Per_FGA           0.23542248
OPP_3PAr            -0.07014372
OPP_TrueShootingPer -0.63767314
OPP_TotalReboundPer -0.57708802
OPP_AssistPer       -0.33175728
OPP_StealPer        -0.38232828
OPP_BlockPer        -0.24473852
OPP_EffectiveFGPer  -0.61494163
OPP_TurnoverPer      0.14991099
OPP_OReboundPer     -0.34665452
OPP_FT_Per_FGA      -0.29636836
```

From this analysis, we can validate our assumption that advanced offensive statistics are positively correlated with Win/Loss % and advanced defensive statistics (opponents statistics) are negatively correlated with Win/Loss %. In addition, we've identified the advanced offensive and defensive statistics that are most highly correlated with Win/Loss % as follows.

From this result, we find that the top 3 positive correlations are:

Variable	Correlation Coefficient
ORtg (Offensive Rating)	.78247702
TrueShootingPer (True Shooting Percentage)	.62685167
EffectiveFGPer (Effective Field Goal Percentage)	.604551834

The top 3 negative correlations are:

Variable	Correlation Coefficient
OPP_TrueShootingPer (True Shooting Percentage)	-.63767314
OPP_EffectiveFGPer (Effective Field Goal Percentage)	-.61494163
OPP_TotalReboundPer (Total Rebound Percentage)	-.57708802

NOTE: For this analysis, we have not validated that the distribution of all independent variables is normal, but just use this as a guide into subsequent analysis and model building.

Multicollinearity

Next we check for a high degree multicollinearity among the independent variables (indicating that are not independent but are correlated). We started from a correlation matrix among the independent variables.

	W.L.	Pace	ORtg	FTr	X3PAr	TrueShootingPer	TotalReboundPer	AssistPer	StealPer	BlockPer	EffectiveFGPer
W.L.	1.00000000	-0.06429262	0.78248802	0.14810170	0.04090476	0.62685167	0.57706399	0.25458589	0.229392517	0.35610224	0.604518335
Pace	-0.06429262	1.00000000	0.09995122	0.03409895	0.13022053	0.13324071	-0.12083620	-0.10198255	-0.010434854	-0.08792173	0.123038170
ORtg	0.78248802	0.09995122	1.00000000	0.04742284	0.17118491	0.85153107	0.49385930	0.25776028	0.048226593	0.17419401	0.825597998
FTr	0.14810170	0.03409895	0.04742284	1.00000000	-0.35586396	0.02382136	0.19800165	-0.07022314	0.061717337	0.10575377	-0.107904757
X3PAr	0.04090476	0.13022053	0.17118491	-0.35586396	1.00000000	0.31817101	-0.25707097	0.21610059	-0.111598348	-0.20183446	0.353567715
TrueShootingPer	0.62685167	0.13324071	0.85153107	0.02382136	0.31817101	1.00000000	0.27369300	0.33040496	-0.020517233	0.03533696	0.975827621
TotalReboundPer	0.57706399	-0.12083620	0.49385930	0.19800165	-0.25707097	0.27369300	1.00000000	0.11334806	-0.119606206	0.27799257	0.251797401
AssistPer	0.25458589	-0.10198255	0.25776028	-0.07022314	0.21610059	0.33040496	0.11334806	1.00000000	0.025077656	0.02921738	0.341989267
StealPer	0.22939252	-0.01043485	0.04822659	0.06171734	-0.11159835	-0.02051723	-0.11960621	0.02507766	1.000000000	0.20657742	-0.008786804
BlockPer	0.35610224	-0.08792173	0.17419401	0.10575377	-0.20183446	0.03533696	0.27799257	0.02921738	0.206577420	1.00000000	0.030344360
EffectiveFGPer	0.60451834	0.12303817	0.82559800	-0.10790476	0.35356772	0.97582762	0.25179740	0.34198927	-0.008786804	0.03034436	1.000000000
TurnoverPer	-0.48977648	-0.06362581	-0.62641900	0.13766541	-0.09608702	-0.28044083	-0.11379825	-0.01773485	-0.043216907	-0.09900014	-0.275235393
OReboundPer	0.28881899	-0.09609671	0.22907983	0.27903737	-0.43189417	-0.15070384	0.71019825	-0.04525979	0.146795922	0.33150849	-0.172535506
FT_Per_FGA	0.23542248	0.06412058	0.20466264	0.93345905	-0.26159879	0.20661030	0.21664742	-0.01972630	0.014981934	0.08783734	0.027229116
OPP_3PAr	-0.07014372	0.04168198	-0.03956525	-0.16646940	0.08096299	0.02052395	0.02225612	-0.01654030	0.054485465	-0.03960503	0.034722700
OPP_TrueShootingPer	-0.63767314	0.29432535	-0.27065662	-0.15687446	0.13158171	-0.14204879	-0.50126589	-0.15584639	-0.018014995	-0.51602290	-0.138678099
OPP_TotalReboundPer	-0.57708802	0.12079158	-0.49388941	-0.19797711	0.25703048	-0.27372004	-0.99999958	-0.11340065	0.119634034	-0.27801365	-0.251830594
OPP_AssistPer	-0.33175728	-0.07696438	-0.30755203	-0.01879723	-0.10383200	-0.28848310	-0.21945190	0.09963618	0.218168808	0.04387637	-0.284553160
OPP_StealPer	-0.38232828	-0.09126336	-0.46994807	0.08583555	-0.13624790	-0.25602126	-0.04502851	0.09589970	-0.007346036	-0.01740843	-0.251066576
OPP_BlockPer	-0.24473852	-0.01731344	-0.23253135	0.10406208	0.21623318	-0.18638490	-0.16944497	0.03149717	-0.008086089	-0.19638696	-0.203416517
OPP_EffectiveFGPer	-0.61494163	0.29622399	-0.23394421	-0.19430812	0.16139850	-0.10043201	-0.50092451	-0.13126700	-0.059625572	-0.54150231	-0.091674337
OPP_TurnoverPer	0.14991099	-0.01107577	-0.09010637	0.07774119	-0.07798996	-0.11658744	-0.13909848	0.01269432	0.806004228	0.06642041	-0.110798874
OPP_OReboundPer	-0.34665452	0.03028998	-0.29885477	0.10685816	-0.12250055	-0.31827421	-0.60330447	-0.11405149	0.278332498	0.10628735	-0.324402434
OPP_FT_Per_FGA	-0.29636836	0.04481785	-0.31881962	0.19364338	-0.15149361	-0.33403676	-0.15329281	-0.16600709	0.225999621	-0.03247677	-0.356594716

	TurnoverPer	OReboundPer	FT_Per_FGA	OPP_3PAr	OPP_TrueShootingPer	OPP_TotalReboundPer	OPP_AssistPer	OPP_StealPer	OPP_BlockPer	OPP_EffectiveFGPer
W.L.	-0.48977648	0.28881899	0.235422482	-0.07014372	-0.63767314	-0.57708802	-0.33175728	-0.382328283	-0.244738516	-0.61494163
Pace	-0.06362581	-0.09609671	0.064120584	0.04168198	0.29432535	0.12079158	-0.07696438	-0.091263364	-0.017313436	0.29622399
ORtg	-0.62641900	0.22907983	0.204662640	-0.03956525	-0.27065662	-0.49388941	-0.30755203	-0.469948068	-0.232531349	-0.23394421
FTr	0.13766541	0.27903737	0.933459052	-0.16646940	-0.15687446	-0.19797711	-0.01879723	0.085835549	0.104062079	-0.19430812
X3PAr	-0.09608702	-0.43189417	-0.261598793	0.08096299	0.13158171	0.25703048	-0.10383200	-0.136247901	0.216233181	0.16139850
TrueShootingPer	-0.28044083	-0.15070384	0.206610302	0.02052395	-0.14204879	-0.27372004	-0.28848310	-0.256021258	-0.186384903	-0.10043201
TotalReboundPer	-0.11379825	0.71019825	0.216647425	0.02225612	-0.50126589	-0.99999958	-0.21945190	-0.045028511	-0.169444968	-0.50092451
AssistPer	-0.01773485	-0.04525979	-0.019726297	-0.01654030	-0.15584639	-0.11340065	0.09963618	0.095899700	0.031497167	-0.13126700
StealPer	-0.04321691	0.14679592	0.014981934	0.05448547	-0.01801499	0.11963403	0.21816881	-0.007346036	-0.008086089	-0.05962557
BlockPer	-0.09900014	0.33150849	0.087837340	-0.03960503	-0.51602290	-0.27801365	0.04387637	-0.017408428	-0.196386955	-0.54150231
EffectiveFGPer	-0.27523539	-0.17253551	0.027229116	0.03472270	-0.13867810	-0.25183059	-0.28455316	-0.251066576	-0.203416517	-0.09167434
TurnoverPer	1.00000000	-0.03803163	0.042686660	0.05481298	0.17216272	0.11382243	0.22360066	0.755110391	0.204503542	0.13412515
OReboundPer	-0.03803163	1.00000000	0.209217719	-0.10082676	-0.28490602	-0.71019620	0.02534125	0.075361320	-0.035569642	-0.31856337
FT_Per_FGA	0.04268666	0.20921772	1.000000000	-0.13781072	-0.13751246	-0.21662401	-0.06322675	0.008459466	0.072954899	-0.16582175
OPP_3PAr	0.05481298	-0.10082676	-0.137810715	1.00000000	0.18573397	-0.02225990	0.44696620	-0.027642118	-0.040036435	0.21280336
OPP_TrueShootingPer	0.17216272	-0.28490602	-0.137512459	0.18573397	1.00000000	0.50128752	0.25713803	0.100226752	0.177197159	0.97363643
OPP_TotalReboundPer	0.11382243	-0.71019620	-0.216624013	-0.02225990	0.50128752	1.00000000	0.21947841	0.045045017	0.169439899	0.50093918
OPP_AssistPer	0.22360066	0.02534125	-0.063226749	0.44696620	0.25713803	0.21947841	1.00000000	0.192466851	0.121047761	0.24651911
OPP_StealPer	0.75511039	0.07536132	0.008459466	-0.02764212	0.10022675	0.04504502	0.19246685	1.000000000	0.173606510	0.09203647
OPP_BlockPer	0.20450354	-0.03556964	0.072954899	-0.04003644	0.17719716	0.16943990	0.12104776	0.173606510	1.000000000	0.16737464
OPP_EffectiveFGPer	0.13412515	-0.31856337	-0.165821748	0.21280336	0.97363643	0.50093918	0.24651911	0.092036470	0.167374644	1.00000000
OPP_TurnoverPer	0.07802600	0.10225386	0.023075054	0.09684195	0.09926818	0.13913250	0.17191676	-0.013178426	0.041604349	0.02787500
OPP_OReboundPer	0.16266667	0.04420596	0.037688555	-0.16348066	0.16758118	0.60329013	0.30019785	0.193311563	0.132367328	0.15597748
OPP_FT_Per_FGA	0.26190149	0.13800208	0.122727020	-0.12750072	0.31552441	0.15331917	0.14223168	0.108847241	0.116548613	0.11667518

	OPP_TurnoverPer	OPP_OReboundPer	OPP_FT_Per_FGA
W.L.	0.14991099	-0.34665452	-0.29636836
Pace	-0.01107577	0.03028998	0.04481785
ORtg	-0.09010637	-0.29885477	-0.31881962
FTr	0.07774119	0.10685816	0.19364338
X3PAr	-0.07798996	-0.12250055	-0.15149361
TrueShootingPer	-0.11658744	-0.31827421	-0.33403676
TotalReboundPer	-0.13909848	-0.60330447	-0.15329281
AssistPer	0.01269432	-0.11405149	-0.16600709
StealPer	0.80600423	0.27833250	0.22599962
BlockPer	0.06642041	0.10628735	-0.03247677
EffectiveFGPer	-0.11079987	-0.32440243	-0.35659472
TurnoverPer	0.07802600	0.16266667	0.26190149
OReboundPer	0.10225386	0.04420596	0.13800208
FT_Per_FGA	0.02307505	0.03768855	0.12272702
OPP_3PAr	0.09684195	-0.16348066	-0.12750072
OPP_TrueShootingPer	0.09926818	0.16758118	0.31552441
OPP_TotalReboundPer	0.13913250	0.60329013	0.15331917
OPP_AssistPer	0.17191676	0.30019785	0.14223168
OPP_StealPer	-0.01317843	0.19331156	0.10884724
OPP_BlockPer	0.04160435	0.13236733	0.11654861
OPP_EffectiveFGPer	0.02787500	0.15597748	0.11667518
OPP_TurnoverPer	1.00000000	0.16028227	0.39603882
OPP_OReboundPer	0.16028227	1.00000000	0.19972571
OPP_FT_Per_FGA	0.39603882	0.19972571	1.00000000

Given the large number of independent variables (23), we've also attempted to highlight below the fields with a high (>.90) absolute value correlation coefficient that is a strong indicator of multicollinearity.

Field 1	Field 2	Correlation Coefficient

TrueShootingPer	EffectiveFGPer	0.97582762
TotalReboundPer	OPP_TotalReboundPer	-0.99999958
OPP_TrueShootingPer	OPP_EffectiveFGPer	0.97363643

Using our domain knowledge, these correlations among independent variables are expected. For example, the number of rebounds per game get completely distributed among the home (target) team and the opponent. If there are 50 available rebounds per game, and the home team gets 30, then the opponent team must get the remaining 20).

In our subsequent model building, we will be sure to eliminate one of each of these field pairs from our final model.

Complete First-Order Model

Next, we build a complete first-order linear regression model for Win/Loss % from all other 23 independent variables as follows:

```
Call:
lm(formula = W.L. ~ ., data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.169676 -0.037971  0.001378  0.038402  0.174962

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        41.2773809 56.8928982   0.726 0.468226
Pace                0.0004023  0.0004469   0.900 0.368142
ORtg                0.0037865  0.0042189   0.898 0.369573
FTr                -0.0222548  0.3082207  -0.072 0.942448
X3PAr               0.0470352  0.0421841   1.115 0.265006
TrueShootingPer     1.7065086  1.5972782   1.068 0.285496
TotalReboundPer    -0.4147422  0.5689704  -0.729 0.466140
AssistPer           0.0003583  0.0003060   1.171 0.241881
StealPer            0.0020329  0.0017008   1.195 0.232157
BlockPer            0.0002784  0.0006907   0.403 0.686962
EffectiveFGPer      0.5720208  1.1852911   0.483 0.629441
TurnoverPer        -0.0185569  0.0060779  -3.053 0.002299 **
OReboundPer         0.0122398  0.0046451   2.635 0.008488 **
FT_Per_FGA          0.4181772  0.5916434   0.707 0.479782
OPP_3PAr            0.0618693  0.0479128   1.291 0.196776
OPP_TrueShootingPer -2.4101139  0.5993727  -4.021 6.04e-05 ***
OPP_TotalReboundPer -0.4020998  0.5689338  -0.707 0.479810
OPP_AssistPer       -0.0012230  0.0003803  -3.216 0.001323 **
OPP_StealPer         0.0003478  0.0019428   0.179 0.857928
OPP_BlockPer        -0.0017983  0.0009757  -1.843 0.065501 .
OPP_EffectiveFGPer  -0.7563046  0.4964591  -1.523 0.127842
OPP_TurnoverPer      0.0235682  0.0018086  13.031  < 2e-16 ***
OPP_OReboundPer     -0.0121578  0.0037384  -3.252 0.001167 **
OPP_FT_Per_FGA      -0.3129570  0.0874551  -3.578 0.000355 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05642 on 1733 degrees of freedom
Multiple R-squared:  0.8923,    Adjusted R-squared:  0.8909
F-statistic: 624.3 on 23 and 1733 DF,  p-value: < 2.2e-16
```

From the regression model above we can:

- Reject the null hypothesis that none of the independent variables contribute to the variance in Win/Loss % because the p-value for the F-statistic (2.2e-16) is significant at the 95% confidence level. We can also accept the alternate hypothesis that one or more of the independent variables does contribute to the variance in the dependent variable.
- For variables TurnoverPer, OReboundPer, Opp_TrueShootingPer, OPP_AssistPer, OPP_TurnoverPer, OPP_OReboundPer, and OPP_FT_Per_FGA, we can also reject the null hypothesis that each individual variable contributes no information for the prediction of Win/Loss % because their associated p-values for the test statistic is significant at the 95% confidence level. Therefore, we can also accept the alternate hypothesis that these variables are linearly related to Win/Loss % with a slope differing from 0.

  NOTE: As expected these variables are also exhibited the strongest correlation with our target variable as identified in the previous section.

- For all other independent variables (not listed in the point above, such as Pace), we cannot reject the null hypothesis that each individual variable contributes no information for the prediction of Win/Loss % because their associated p-values for

the test statistic is **not significant** at the 95% confidence level. Therefore, more data is needed, and we should consider removing these variables from the model.

- Finally, the overall model has an Adjusted R-squared (coefficient of determination) value of 0.8909 indicating that ~89% of the variance in Win/Loss % can be explained by the model.

But because this model contains many independent variables that have insignificant p-values for their test statistic, we cannot accept this model as our final solution.

Reduced First-Order Model (removing independent variables with insignificant p-values)

From the model built in the previous section, we need to eliminate the independent variables that had an insignificant p-value (at the 95% confidence level) for their associated test statistic.

Note, that we've done this iteratively. That is, we eliminated the independent variable with the largest insignificant p-value in the model (such as FTr in the initial complete first order model). We then rebuilt the model and looked for and eliminated the independent variable with the highest insignificant p-value in the resultant model. We continued this approach until we were left with a model that contained independent variables all with a significant p-value. But due to the large number of independent variables, we did not document all of these intermittent steps:

Step 1 (remove FTr):

```
Call:
lm(formula = W.L. ~ d$Pace + d$ORtg + d$X3PAr + d$TrueShootingPer +
    d$TotalReboundPer + d$AssistPer + d$StealPer + d$BlockPer +
    d$EffectiveFGPer + d$TurnoverPer + d$OReboundPer + d$FT_Per_FGA +
    d$OPP_3PAr + d$OPP_TrueShootingPer + d$OPP_TotalReboundPer +
    d$OPP_AssistPer + d$OPP_StealPer + d$OPP_BlockPer + d$OPP_EffectiveFGPer +
    d$OPP_TurnoverPer + d$OPP_OReboundPer + d$OPP_FT_Per_FGA,
    data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.169616 -0.037890  0.001321  0.038425  0.174937

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          41.2508341 56.8753887   0.725 0.468375
d$Pace                0.0004028  0.0004468   0.902 0.367355
d$ORtg                0.0037653  0.0042074   0.895 0.370961
d$X3PAr               0.0473431  0.0419561   1.128 0.259309
d$TrueShootingPer     1.8045054  0.8419264   2.143 0.032227 *
d$TotalReboundPer    -0.4145335  0.5687998  -0.729 0.466231
d$AssistPer           0.0003591  0.0003058   1.174 0.240442
d$StealPer            0.0020283  0.0016992   1.194 0.232751
d$BlockPer            0.0002799  0.0006902   0.406 0.685110
d$EffectiveFGPer      0.4906694  0.3679819   1.333 0.182574
d$TurnoverPer        -0.0185923  0.0060564  -3.070 0.002175 **
d$OReboundPer         0.0122395  0.0046437   2.636 0.008471 **
d$FT_Per_FGA          0.3758349  0.0783784   4.795 1.76e-06 ***
d$OPP_3PAr            0.0618586  0.0478989   1.291 0.196723
d$OPP_TrueShootingPer -2.4123117  0.5984276  -4.031 5.79e-05 ***
d$OPP_TotalReboundPer -0.4019144  0.5687648  -0.707 0.479882
d$OPP_AssistPer       -0.0012230  0.0003802  -3.217 0.001320 **
d$OPP_StealPer         0.0003413  0.0019401   0.176 0.860379
d$OPP_BlockPer        -0.0018022  0.0009739  -1.850 0.064431 .
d$OPP_EffectiveFGPer  -0.7537746  0.4950790  -1.523 0.128058
d$OPP_TurnoverPer      0.0235698  0.0018079  13.037  < 2e-16 ***
d$OPP_OReboundPer     -0.0121474  0.0037345  -3.253 0.001165 **
d$OPP_FT_Per_FGA      -0.3126724  0.0873412  -3.580 0.000353 ***
---
```

.

.

.

Step 13 (remove OPP_BlockPer):

```
Call:
lm(formula = W.L. ~ d$X3PAr + d$TrueShootingPer + d$TurnoverPer +
    d$OReboundPer + d$FT_Per_FGA + d$OPP_TrueShootingPer + d$OPP_AssistPer +
    d$OPP_TurnoverPer + d$OPP_OReboundPer + d$OPP_FT_Per_FGA,
    data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.169765 -0.039707  0.001092  0.038233  0.176589

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.5454591  0.0550750   9.904  < 2e-16 ***
d$X3PAr               0.0750067  0.0297000   2.525  0.01164 *
d$TrueShootingPer     2.8292122  0.0597322  47.365  < 2e-16 ***
d$TurnoverPer        -0.0256810  0.0008607 -29.837  < 2e-16 ***
d$OReboundPer         0.0085894  0.0003904  22.004  < 2e-16 ***
d$FT_Per_FGA          0.2444235  0.0434523   5.625 2.16e-08 ***
d$OPP_TrueShootingPer -3.0637163  0.0629481 -48.671  < 2e-16 ***
d$OPP_AssistPer       -0.0010166  0.0003024  -3.362  0.00079 ***
d$OPP_TurnoverPer      0.0266672  0.0007987  33.390  < 2e-16 ***
d$OPP_OReboundPer     -0.0063441  0.0004834 -13.125  < 2e-16 ***
d$OPP_FT_Per_FGA      -0.1972405  0.0416272  -4.738 2.33e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05649 on 1746 degrees of freedom
Multiple R-squared:  0.8912,    Adjusted R-squared:  0.8906
F-statistic:  1430 on 10 and 1746 DF,  p-value: < 2.2e-16
```

The end result of this iterative process removing the independent variables with insignificant p-values is a model of 10 independent variables, all with significant p-values for their associated t-tests. The p-value of the F-statistic is also significant at the 95% confidence level.

The associated adjusted r-squared value of this model (.8906) is almost identical to our original complete first-order model (.8909).

Therefore, this model is a potential candidate solution for our overall analysis.

Forward selection attempt

Starting from an empty model, we attempted forward selection. We based our selection of the field added during each step based on the **Akaike information criterion** (**AIC**). AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model (4). For each step of our forward elimination process, we added the field with the smallest AIC value. The resultant model of our forward selection was:

```
Call:
lm(formula = W.L. ~ ORtg + OPP_EffectiveFGPer + OPP_TurnoverPer +
    OPP_FT_Per_FGA + OPP_OReboundPer + FTr + OPP_AssistPer +
    OPP_TrueShootingPer + TurnoverPer + TrueShootingPer + OReboundPer +
    X3PAr + OPP_BlockPer, data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.169350 -0.038447  0.001183  0.038214  0.171156

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.3704480  0.1041003   3.559 0.000383 ***
ORtg                0.0045966  0.0040521   1.134 0.256795
OPP_EffectiveFGPer -0.8704630  0.4806839  -1.811 0.070331 .
OPP_TurnoverPer     0.0266781  0.0007996  33.365  < 2e-16 ***
OPP_FT_Per_FGA     -0.3378839  0.0848482  -3.982 7.11e-05 ***
OPP_OReboundPer    -0.0062576  0.0004899 -12.772  < 2e-16 ***
FTr                 0.2078922  0.0331186   6.277 4.34e-10 ***
OPP_AssistPer      -0.0009809  0.0003021  -3.247 0.001189 **
OPP_TrueShootingPer -2.0489650  0.5472013  -3.744 0.000187 ***
TurnoverPer        -0.0192727  0.0057534  -3.350 0.000826 ***
TrueShootingPer     2.0630127  0.6784019   3.041 0.002393 **
OReboundPer         0.0062119  0.0021515   2.887 0.003933 **
X3PAr               0.0875902  0.0354124   2.473 0.013477 *
OPP_BlockPer       -0.0018370  0.0009590  -1.916 0.055584 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05638 on 1743 degrees of freedom
Multiple R-squared:  0.8918,    Adjusted R-squared:  0.891
F-statistic:  1105 on 13 and 1743 DF,  p-value: < 2.2e-16
```

We then iteratively eliminated the independent variables that had an insignificant p-value for their test static or had high multicollinearity, resulting in:

```
Call:
lm(formula = W.L. ~ OPP_TurnoverPer + OPP_FT_Per_FGA + OPP_OReboundPer +
    FTr + OPP_AssistPer + OPP_TrueShootingPer + TurnoverPer +
    TrueShootingPer + OReboundPer + X3PAr, data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.168023 -0.039528  0.001219  0.037626  0.175814

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.5148573  0.0553398   9.304  < 2e-16 ***
OPP_TurnoverPer     0.0265137  0.0007971  33.262  < 2e-16 ***
OPP_FT_Per_FGA     -0.1980539  0.0415905  -4.762 2.08e-06 ***
OPP_OReboundPer    -0.0063914  0.0004839 -13.208  < 2e-16 ***
FTr                 0.1853068  0.0320500   5.782 8.74e-09 ***
OPP_AssistPer      -0.0009878  0.0003023  -3.267  0.00111 **
OPP_TrueShootingPer -3.0471414  0.0632481 -48.178  < 2e-16 ***
TurnoverPer        -0.0259929  0.0008666 -29.994  < 2e-16 ***
TrueShootingPer     2.8747270  0.0570870  50.357  < 2e-16 ***
OReboundPer         0.0085295  0.0003911  21.809  < 2e-16 ***
X3PAr               0.0799451  0.0298672   2.677  0.00751 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05647 on 1746 degrees of freedom
Multiple R-squared:  0.8913,    Adjusted R-squared:  0.8907
F-statistic:  1432 on 10 and 1746 DF,  p-value: < 2.2e-16
```

The result of forward selection is almost an identical model to our reduced complete first order model from the previous selection, with the same number of independent variables (10), slight improvement in adjusted r-squared (.8907 vs. .8906), and standard error (.05647 vs. .05649).

Backward elimination attempt

Starting from the complete first-order model, we attempted backward selection. We again based our selection of the field removed during each step based on the **Akaike information criterion** (**AIC**). For each step of our backward elimination process, we removed the field with the largest AIC value. The resultant model of our backward selection was:

```
Call:
lm(formula = W.L. ~ X3PAr + TrueShootingPer + EffectiveFGPer +
    TurnoverPer + OReboundPer + FT_Per_FGA + OPP_TrueShootingPer +
    OPP_AssistPer + OPP_BlockPer + OPP_EffectiveFGPer + OPP_TurnoverPer +
    OPP_OReboundPer + OPP_FT_Per_FGA, data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.169476 -0.039054  0.001059  0.037909  0.173062

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.5249667  0.0645331   8.135 7.75e-16 ***
X3PAr               0.1039999  0.0319166   3.258 0.001142 **
TrueShootingPer     2.0931155  0.4064414   5.150 2.90e-07 ***
EffectiveFGPer      0.6202859  0.3615095   1.716 0.086372 .
TurnoverPer        -0.0256326  0.0008800 -29.128  < 2e-16 ***
OReboundPer         0.0086270  0.0003923  21.990  < 2e-16 ***
FT_Per_FGA          0.3771663  0.0771431   4.889 1.11e-06 ***
OPP_TrueShootingPer -2.0766364  0.5481408  -3.789 0.000157 ***
OPP_AssistPer       -0.0009897  0.0003023  -3.274 0.001082 **
OPP_BlockPer        -0.0018617  0.0009584  -1.942 0.052246 .
OPP_EffectiveFGPer  -0.8476525  0.4810391  -1.762 0.078223 .
OPP_TurnoverPer      0.0266961  0.0008019  33.291  < 2e-16 ***
OPP_OReboundPer     -0.0062218  0.0004894 -12.714  < 2e-16 ***
OPP_FT_Per_FGA      -0.3321781  0.0846944  -3.922 9.12e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0564 on 1743 degrees of freedom
Multiple R-squared:  0.8918,    Adjusted R-squared:  0.891
F-statistic:  1105 on 13 and 1743 DF,  p-value: < 2.2e-16
```

We then again iteratively eliminated the independent variables that had an insignificant p-value for their test static or had high multicollinearity, resulting in:

```
Call:
lm(formula = W.L. ~ X3PAr + TrueShootingPer + TurnoverPer + OReboundPer +
    FT_Per_FGA + OPP_TrueShootingPer + OPP_AssistPer + OPP_TurnoverPer +
    OPP_OReboundPer + OPP_FT_Per_FGA, data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.169765 -0.039707  0.001092  0.038233  0.176589

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.5454591  0.0550750   9.904  < 2e-16 ***
X3PAr               0.0750067  0.0297000   2.525  0.01164 *
TrueShootingPer     2.8292122  0.0597322  47.365  < 2e-16 ***
TurnoverPer        -0.0256810  0.0008607 -29.837  < 2e-16 ***
OReboundPer         0.0085894  0.0003904  22.004  < 2e-16 ***
FT_Per_FGA          0.2444235  0.0434523   5.625 2.16e-08 ***
OPP_TrueShootingPer -3.0637163  0.0629481 -48.671  < 2e-16 ***
OPP_AssistPer       -0.0010166  0.0003024  -3.362  0.00079 ***
OPP_TurnoverPer      0.0266672  0.0007987  33.390  < 2e-16 ***
OPP_OReboundPer     -0.0063441  0.0004834 -13.125  < 2e-16 ***
OPP_FT_Per_FGA      -0.1972405  0.0416272  -4.738 2.33e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05649 on 1746 degrees of freedom
Multiple R-squared:  0.8912,    Adjusted R-squared:  0.8906
F-statistic:  1430 on 10 and 1746 DF,  p-value: < 2.2e-16
```

The result of backward selection is the exact same model as our reduced complete first order model from a previous selection.

Simplified forward selection model – Solution Model

From the previous sections, we found that our forward selection process built a model with the highest adjusted r-squared value and an equivalent number of independent variables (10) as our other approaches.

We then attempted to further simplify the forward selection model by iteratively eliminating independent variables that only provided very slight contributions to the overall adjusted r-squared value or had high multicollinearity. The result of this process was the following model.

```
Call:
lm(formula = W.L. ~ OPP_TurnoverPer + OPP_OReboundPer + OPP_TrueShootingPer +
    TurnoverPer + TrueShootingPer + OReboundPer, data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.167332 -0.040783  0.001201  0.038956  0.193919

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.5354395  0.0548579    9.76   <2e-16 ***
OPP_TurnoverPer     0.0250913  0.0007498   33.46   <2e-16 ***
OPP_OReboundPer    -0.0064021  0.0004785  -13.38   <2e-16 ***
OPP_TrueShootingPer -3.2019960  0.0590446  -54.23   <2e-16 ***
TurnoverPer        -0.0261355  0.0008500  -30.75   <2e-16 ***
TrueShootingPer     3.0370166  0.0527151   57.61   <2e-16 ***
OReboundPer         0.0083392  0.0003610   23.10   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05733 on 1750 degrees of freedom
Multiple R-squared:  0.8877,    Adjusted R-squared:  0.8873
F-statistic:  2306 on 6 and 1750 DF,  p-value: < 2.2e-16
```

The adjusted r-squared of this model (.8873) is very slightly less (.0034) than our previous best model (.8907). However, this model has the following advantages:

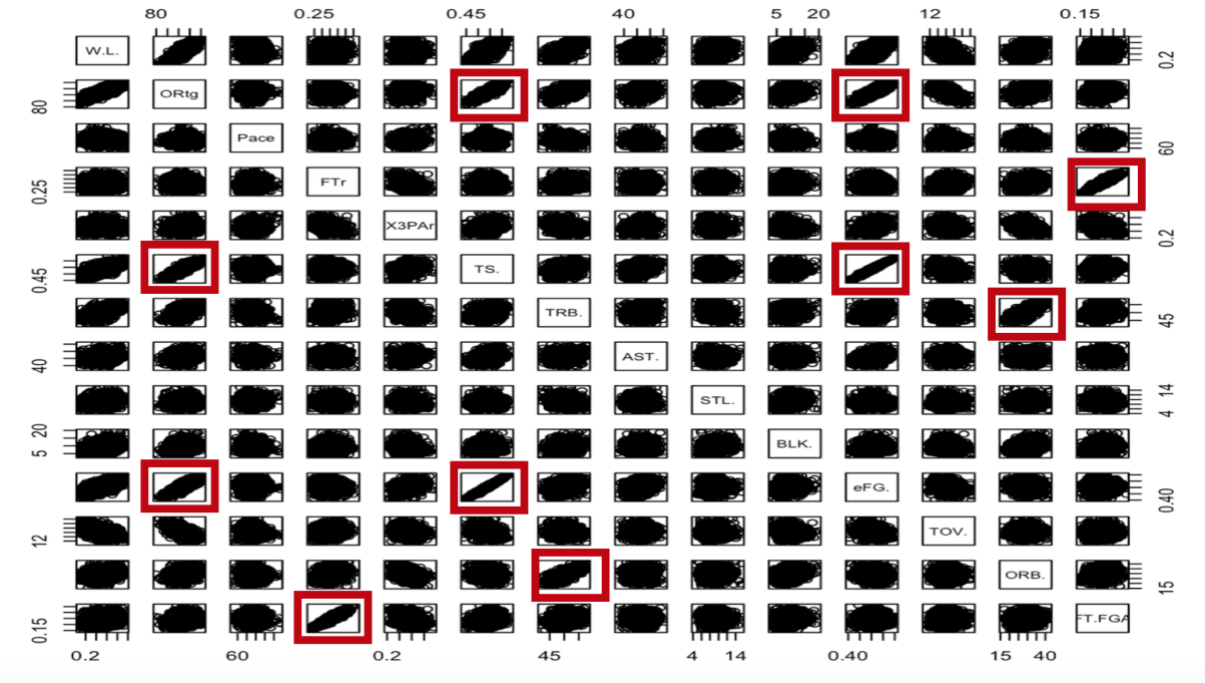
- Less independent variables (6 vs. 10)
- All independent variables have a very significant p-value for their associated test-statistics (2e-16)
- The degree of multicollinearity between independent variables is low

Therefore, due to the improved simplicity of this model, and the nearly negligible difference in adjusted r-squared value, **we selected this model as our best solution to Milestone 1.**

Interaction terms

Building on the best fit first-order regression model from the previous section, I attempt to identify and add interaction terms to help improve the overall accuracy of the model.

I do this by first examining a plot of all offensive advanced statistics. From this plot, there are multiple interactions between independent variables as highlighted below in red.



Of these interactions between independent variables, they all mostly appear to be a linear relationship. Therefore, I iteratively attempted to add an interaction term for each pair highlighted above.

Interaction Term	Result
ORtg*TrueShootingPer*EffectiveFGPer	Improved adjusted r-squared by .0014
FT*FT.FGA	Insignificant p-value of 0.309237
OffensiveReboundPer*TotalReboundPer	Insignificant p-value of 0.819873

For the interaction terms that improved the model with significant p-values, I also added the corresponding defensive interaction term to the model.

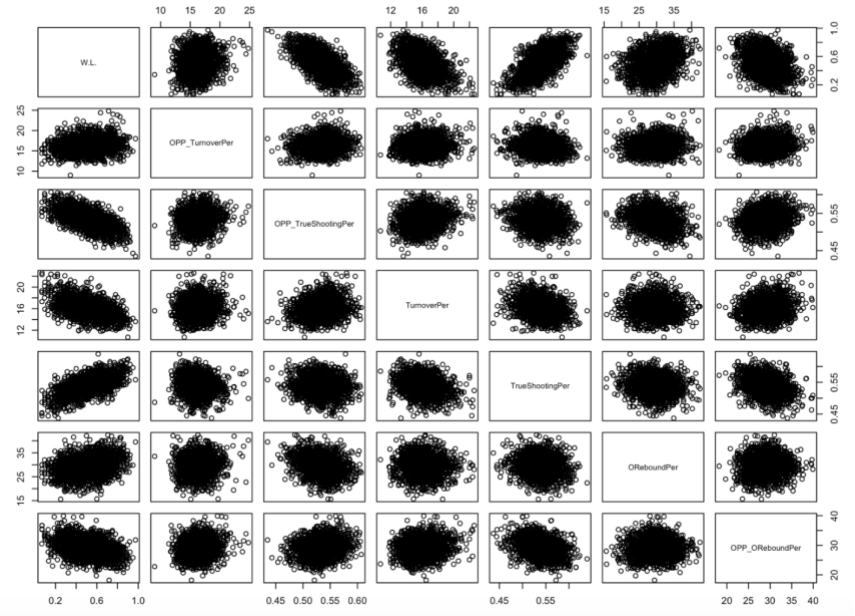
As a result, only the interaction terms that had a significant p-value remained in my candidate solution model as shown below.

```
Call:
lm(formula = W.L. ~ OPP_TurnoverPer + OPP_TrueShootingPer + TurnoverPer +
    TrueShootingPer + OReboundPer + ORtg + EffectiveFGPer + OPP_ORtg +
    OPP_EffectiveFGPer + ORtg * TrueShootingPer * EffectiveFGPer +
    OPP_ORtg * OPP_TrueShootingPer * OPP_EffectiveFGPer, data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.157661 -0.038817  0.001354  0.037924  0.192463

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 -7.849594   5.497920  -1.428 0.153547
OPP_TurnoverPer                              0.007847   0.001415   5.546 3.37e-08 ***
OPP_TrueShootingPer                         41.297819   9.400665   4.393 1.19e-05 ***
TurnoverPer                                 -0.023077   0.005262  -4.386 1.22e-05 ***
TrueShootingPer                            -19.284593   6.872732  -2.806 0.005073 **
OReboundPer                                  0.007300   0.001894   3.855 0.000120 ***
ORtg                                        -0.104026   0.031059  -3.349 0.000827 ***
EffectiveFGPer                              -14.963367   7.311848  -2.046 0.040862 *
OPP_ORtg                                     0.151822   0.046291   3.280 0.001060 **
OPP_EffectiveFGPer                          31.870303  10.803510   2.950 0.003220 **
TrueShootingPer:ORtg                         0.231326   0.065012   3.558 0.000383 ***
ORtg:EffectiveFGPer                          0.151257   0.071362   2.120 0.034184 *
TrueShootingPer:EffectiveFGPer              33.575882  11.960488   2.807 0.005053 **
OPP_TrueShootingPer:OPP_ORtg                -0.384042   0.093133  -4.124 3.91e-05 ***
OPP_ORtg:OPP_EffectiveFGPer                 -0.262958   0.105096  -2.502 0.012438 *
OPP_TrueShootingPer:OPP_EffectiveFGPer      -73.646726  18.001771  -4.091 4.49e-05 ***
TrueShootingPer:ORtg:EffectiveFGPer          -0.347798   0.112419  -3.094 0.002008 **
OPP_TrueShootingPer:OPP_ORtg:OPP_EffectiveFGPer  0.641911   0.172532   3.721 0.000205 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05649 on 1739 degrees of freedom
Multiple R-squared:  0.8917,    Adjusted R-squared:  0.8906
F-statistic: 842.1 on 17 and 1739 DF,  p-value: < 2.2e-16
```

However, while these interaction terms did improve the adjusted r-squared value of the model, it was only by .0033. Therefore, the benefit (improved explanation of variance) is not nearly significant enough to justify adding this complexity to our solution model. Therefore, I do not propose these added interaction terms are kept in the solution model.

Second-order terms

Again examining a plot between all independent variables and the dependent variable, we don't find any pattern other than a linear relationship between the variables. Therefore, I don't find any justification for adding a second-order term to the model.



Feature engineering ("Four Factors" model)

As an alternative to the previous solution model, I attempted to create a model that reflects the "Four Factors" to winning basketball games as attributed to statistician Dean Oliver (2) .

According to Mr. Oliver the "Four Factors of Basketball Success" are:

1. Shooting (40%)
2. Turnovers (25%)
3. Rebounding (20%)
4. Free Throws (15%)

The number in parentheses is the approximate weight Mr. Oliver assigned each factor. Therefore, I attempted to first create a model with the statistics that corresponded to each of these factors:

```
Call:
lm(formula = W.L. ~ EffectiveFGPer + TurnoverPer + OReboundPer +
    FT.FGA, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34325 -0.07387  0.00247  0.07408  0.31238

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.1580194  0.0610076  -18.98   <2e-16 ***
EffectiveFGPer  3.1356869  0.0839496   37.35   <2e-16 ***
TurnoverPer    -0.0328812  0.0015070  -21.82   <2e-16 ***
OReboundPer     0.0142828  0.0006282   22.73   <2e-16 ***
FT.FGA          0.7852022  0.0710338   11.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1025 on 1752 degrees of freedom
Multiple R-squared:  0.6407,    Adjusted R-squared:  0.6399
F-statistic: 781.1 on 4 and 1752 DF,  p-value: < 2.2e-16
```

This "four-factors" model resulted in an adjusted r-squared value of 0.6399 which is significantly lower than my previous best fit model of 0.8873.

However, I then attempted to also include the defensive (or opponents) statistics in the factors, by engineering the following new variables:

```
# Engineer 4 factors that considers both offensive and defensive statistics
d$shooting = (d$EffectiveFGPer)/(d$OPP_EffectiveFGPer)
d$turnovers = (d$TurnoverPer/d$OPP_TurnoverPer)
d$rebounding = (d$OReboundPer/d$OPP_OReboundPer)
d$freethrows = (d$FT.FGA/d$OPP_FT.FGA)
```

Using these engineered features in my model resulted in:

```
Call:
lm(formula = W.L. ~ shooting + turnovers + rebounding + freethrows,
    data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.211700 -0.040844  0.000672  0.039738  0.202417

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.720341   0.020322  -35.45   <2e-16 ***
shooting     1.257663   0.017924   70.17   <2e-16 ***
turnovers   -0.426320   0.009440  -45.16   <2e-16 ***
rebounding   0.219298   0.008261   26.55   <2e-16 ***
freethrows   0.152112   0.007098   21.43   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05869 on 1752 degrees of freedom
Multiple R-squared:  0.8822,    Adjusted R-squared:  0.8819
F-statistic:  3280 on 4 and 1752 DF,  p-value: < 2.2e-16
```

With a resultant adjusted r-squared value of 0.8819, and significant p-values for the F-statistic and test-statistics, this model using the engineered independent variables comes very close to my previously proposed best fit model (0.8873).

Using a similar 80:20 train/test split against this Four Factors engineered model, the resultant
prediction to observed correlation is:

```
> cor(prediction, actual)
[1] 0.948923
```

This rivals the correlation of the previous best fit model of:

```
> cor(prediction, actual)
[1] 0.9413907
```

Therefore, this “Four Factors” model using the engineered features appears equally as robust as the Milestone 1 first-order model and might be even a bit easier to understand.

But what’s most interesting when comparing the models, is that while both strongly consider rebounds, shooting, and turnovers, I was previously able to build a model that didn’t include free throws but explains the variance equally well. Therefore, perhaps free throws aren’t as important to winning as Mr. Oliver proposes.

Solution model summary

As described in the previous section, considering all 23 independent variables at predicting Win/Loss %, our solution model to this study is:

```
Call:
lm(formula = W.L. ~ OPP_TurnoverPer + OPP_OReboundPer + OPP_TrueShootingPer +
    TurnoverPer + TrueShootingPer + OReboundPer, data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.167332 -0.040783  0.001201  0.038956  0.193919

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.5354395  0.0548579    9.76   <2e-16 ***
OPP_TurnoverPer     0.0250913  0.0007498   33.46   <2e-16 ***
OPP_OReboundPer    -0.0064021  0.0004785  -13.38   <2e-16 ***
OPP_TrueShootingPer -3.2019960  0.0590446  -54.23   <2e-16 ***
TurnoverPer        -0.0261355  0.0008500  -30.75   <2e-16 ***
TrueShootingPer     3.0370166  0.0527151   57.61   <2e-16 ***
OReboundPer         0.0083392  0.0003610   23.10   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05733 on 1750 degrees of freedom
Multiple R-squared:  0.8877,    Adjusted R-squared:  0.8873
F-statistic:  2306 on 6 and 1750 DF,  p-value: < 2.2e-16
```

From the proposed solution regression model above we find:

- We can reject the null hypothesis that none of the independent variables contribute to the variance in Win/Loss % because the p-value for the F-statistic (2.2e-16) is significant at the 95% confidence level. We can also accept the alternate hypothesis that one or more of the independent variables does contribute to the variance in the dependent variable.
- For all independent variables, we can also reject the null hypothesis that each individual variable contributes no information for the prediction of Win/Loss % because their associated p-values for the test statistic is significant at the 95% confidence level. Therefore, we can also accept the alternate hypothesis that these variables are linearly related to Win/Loss % with a slope differing from 0.
- The Mean Squared error of the model is:

```
> mse <- mean(sm$residuals^2)
> mse
[1] 5439.281
```

- Using the Variance Inflation Factor, we conclude that no one predictor is strongly related to another (no value is >= 10):

```
> vif(partial_forward_model)
    OPP_TurnoverPer     OPP_OReboundPer OPP_TrueShootingPer        TurnoverPer     TrueShootingPer         OReboundPer
           1.051432           1.153894           1.177892           1.115039           1.232215           1.154275
```

- The overall model has an Adjusted R-squared (coefficient of determination) value of 0.8873 indicating that ~88.73% of the variance in Win/Loss % can be explained by the model.

Residual analysis

This section performs an analysis to confirm the four assumptions about the errors in linear regression against the proposed solution model:

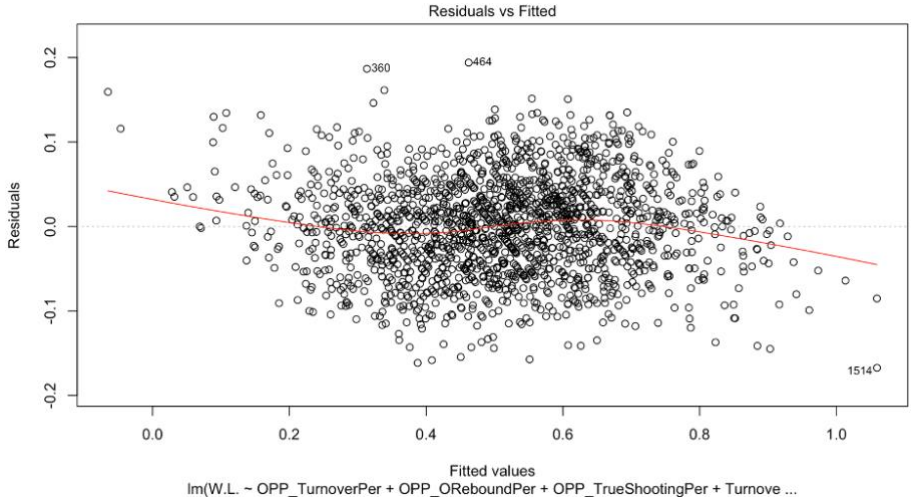
Mean of errors is 0

To validate the mean of the errors is 0, I examine the sum and mean of the residuals, and it should be close to 0 (accounting for possible rounding/precision errors)

```
> # Mean of errors is 0
> model <- lm(formula = W.L. ~ OPP_TurnoverPer + OPP_OReboundPer + OPP_TrueShootingPer + TurnoverPer + TrueShootingPer + OReboundPer,data = d)
> sm <- summary(model)
> sum(model$residuals)
[1] -2.120157e-15
> mean = mean(model$residuals)
> mean
[1] -1.215616e-18
```

Homoscedastic (constant variance)

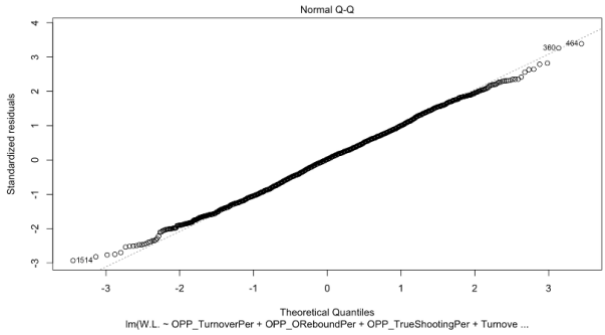
To validate the errors are homoscedastic, I plot the residuals against the predicted values:



In this plot, the variance of the residuals appears mostly constant for all fitted values. Therefore, I conclude the errors are homoscedastic.

Normal

I validate the errors have a normal distribution by creating a q-q plot and then verify the points roughly form a straight line.



In this plot, except for a few leading/trailing points, most all form a straight line, concluding that the errors are normal.

Independence

The assumption that errors are independent means that there is little to no correlation between the errors. I verify independence through the Durbin-Watson test.

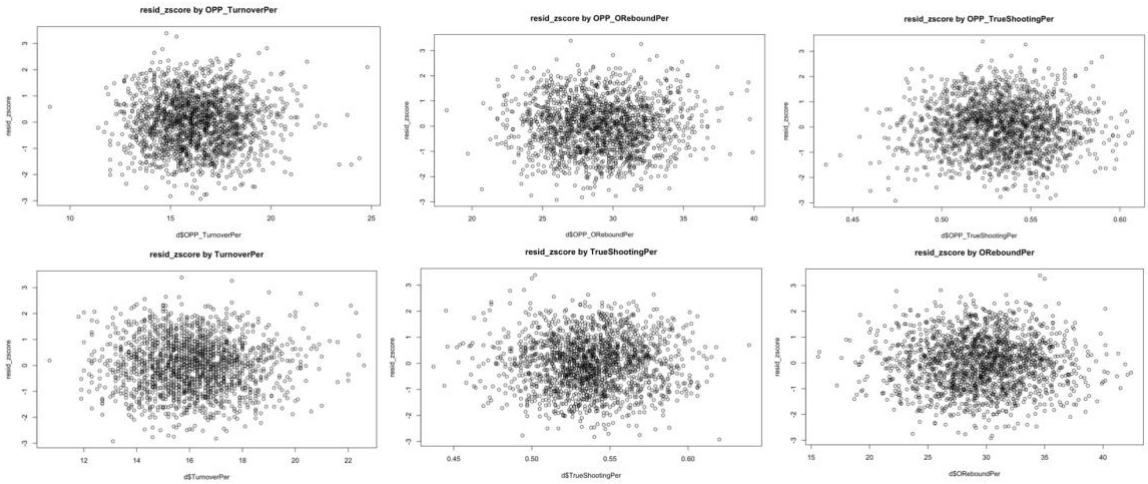
```
> durbinWatsonTest(model)
 lag Autocorrelation D-W Statistic p-value
   1     -0.01861209      2.037106   0.436
 Alternative hypothesis: rho != 0
```

If independence is true, the Durbin-Watson statistic should be between 1.5 and 2.5. In this model, the statistic of 2.037106 prevents me from rejecting the null hypothesis that there is no residual correlation, proving independence.

Lack of fit

Finally, I plot the standardized residuals against each of the independent variables looking for

- Trends
- Changes in variability
- More than 5% of points outside 2 standard deviations



In the six plots above, there is no discernable pattern, most points fall within 2 standard deviations, and there is not noticeable change in variability. Therefore, I conclude there is no lack of fit in the model.

Summary of findings

From this study we find it interesting that of the 23 available advanced statistics, only 6 are needed to strongly predict regular season Win/Loss %. And of these 6 advanced statistics, they are actually 3 distinct statistics, mirroring offense vs. defensive productivity.

- TurnoverPer <-> OPP_TurnoverPer
- TrueShootingPer <-> OPP_TrueShootingPer
- OReboundPer <-> OPP_OReboundPer

We are also quite pleased that the solution is able to explain ~89% of the variance in Win/Loss % with these variables.

From this study, Managers/Coaches may already be able to consider improving the following team's skills to have an impact on improving their overall Winning Percentage:

- Rebounding
- Protecting the ball (less turnovers)
- Stealing the ball
- Shooting efficiently

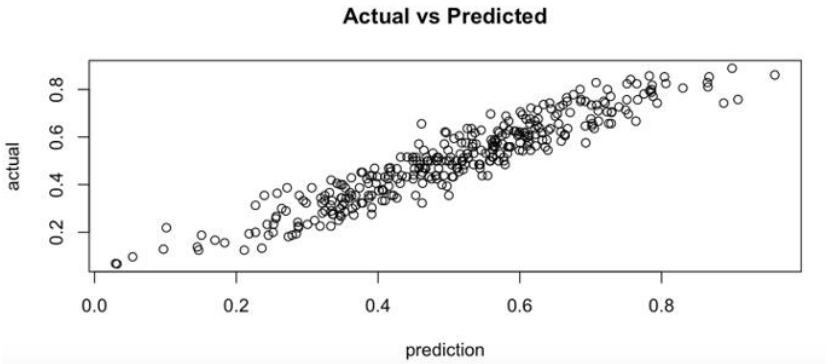
In addition, we built an alternative model using Dean Oliver's "Four Factors of Basketball Success" (shooting, turnovers, rebounds, and free throws), but it was not better than our best fit model that excluded any specific free throw statistics. Therefore, one may wish to further analyze if free throws remain as critical an aspect for NCAA basketball winning success as Mr. Oliver proposed.

Using this model to predict Winning Percentage, we broke all data from seasons 2015 through 2019 randomly into an 80:20 train/test split.

The resultant correlation coefficient of predicted vs. observed identifies a strong, positive relationship:

```
> cor(prediction, actual)
[1] 0.9413907
```

And the plot of Actual vs. Predicted again shows a strong, positive linear relationship with no apparent outliers.



Executive Summary

In this study, we performed an analysis on how well *basic* and *advanced* regular season statistics from 2015 through 2019 predict overall regular season winning percentage for NCAA Division I Men's Basketball teams.

The models were created using the Win Percentage as the response variable. Our analysis was divided into multiple sections for the following criterias.

- Basic statistics: Typically include the summary or average of a single observation typically per game (e.g. points per game, rebounds per game, etc.).
- Advanced statistics: Modern statistics that typically require a formula that combines many of the basic statistics into a single rating where a key tenant is that the statistics are evaluated over possessions, not games.

We then further divided the statistics into the following sub-categories:

- Offensive statistics: Includes stats that are collected while the team is attacking.
- Defensive statistics: Includes stats that are collected while the team is defending.
- Neutral statistics: Attributes within the data which are neither considered defensive or offensive fall into this category

The statistics from the above criterias were used as the independent variables in our models. The models were built using the first order, second order and interaction terms. The final models were validated against the forward selection and backward elimination methods.

The independent variables were checked for multicollinearity using the VIF function.
The histogram graph was plotted for the residuals to find any unconventional data which cannot be determined using the regression models. The QQ plot and PP plot determined 95% of the data was plotted on the line. The stabilization transformation was applied based on the residuals plots.

Overview

Our results suggest several key findings, including:

- Scoring-Margin is the best single basic statistic at predicting winning percentage
- Offensive Rating is the best single advanced statistic at predicting winning percentage

- Models built with only offensive statistics explained only a 54% variability in the overall win percentage.
- Models built with defensive statistics explained 58% of the variability in win percentage of a team.
- Models built with neutral statistics explained 89% of the variability in win percentage of a team.
- Of the 19 independent basic variables (game-play statistics) evaluated, 4 variables are to describe 88.8% of the variability in a team's winning percentage. The variables are as follows:
    - Scoring Margin (Stat Category: Neutral)
    - Opponents Three point Field Goal percentage (Stat Category: Defense)
    - Rebounding Margin (Stat Category: Neutral)
    - Personal Fouls Per Game (Neutral)
- Models built with the advanced statistics explained 88.7 percent of the variability in the win percentage which is .07% lower than primarily using basic variables.

Further details on each of these findings are provided below.

Best Single Basic Statistic

For the data analysis, the basic statistics measures were extracted and broken up into several categories: offensive, defensive and neutral. In order to identify the appropriate independent variables that explain the highest degree of variability in a team's winning percentage. We utilized the Akaike Information Criteria measure calculated with the forward selection linear modeling process to determine the best and most valuable single statistic for the linear model.

From our analysis, Scoring Margin was the independent variable without which the linear model would lose significant information and thus loses effective predictive value. The basic statistic scoring margin is calculated as the average difference between a team's score and the opponent's score. The statement is additionally supported by measuring the correlation value between scoring margin and win percentage. The data shows that scoring margin has 0.94 correlation value with win percentage which shows a strong, positive linear dependence of win percentage with scoring margin.

Additionally, a linear model built with only the Scoring Margin as the independent variable explains 88.6% of the variability in win percentage. After breaking up the dataset into training and test data, the linear model with only scoring margin was able to predict a team's win percentage with 94% accuracy.

Additional Models Evaluated

While scoring margin is a strong statistical measure, in an effort to dig deeper and conduct additional analysis, we reviewed offensive, defensive and neutral variables independently.

To begin with the offensive variables, the four significant variables for predicted win percentage were Points per Game, FG percentage, Three pointers made, Free throw percentage and rebounds per game. These four variables were able to explain 53% of the variability in win percentage. The model was accurate 75% of the time in predicting the values in the test data.

Similarly, for the defensive variables, the important variables for predicting win percentage were Opponents Field Goal Percentage, Opponent 3 Point Field Goal Percentage, Opponent Average 3 Pointers Made/Game, Opponents Points per Game, Steals Per Game and Blocked Shots Per Game. ~ 58% of variability in dependent variables can be predicted by these variables.

Best single advanced statistic

Utilizing the advanced statistics, the independent variable ***Offensive Rating*** (ORtg) performed better than any other single advanced statistic in predicting winning percentage. Offensive rating is one of the most popular and commonly referenced advanced statistics, as we found it highlighted on every single site that we researched. This advanced statistic is calculated by normalizing the average number of points scored per 100 possessions. When performing forward selection, it was found to result in the lowest amount of information lost for a given model. On its own, it is able to explain 61.21% of variance in Winning Percentage.

However, what we found interesting, is that by combining other advanced statistics that considered turnovers, rebounds, and field goal efficiency, we were able to produce a model that was much more accurate and robust, explaining 88.73% of the variance in winning percentage and correctly predicting 95% of our test data. This robust model **did not** include this Offensive Rating statistic (when we tried to add it, the p-value for its test statistic was insignificant). Therefore, one might conclude that while scoring a lot of points does somewhat predict winning, strong defense, rebounding, and efficient scoring is a much stronger predictor for winning.

Are the "Four Factors" to winning basketball still relevant?

According to statistician Dean Oliver's 2004 book "Basketball on Paper: Rules and Tools for Performance Analysis" (6), the "Four Factors of Basketball Success" are:

1. Shooting (40%)
2. Turnovers (25%)
3. Rebounding (20%)
4. Free Throws (15%)

The number in parentheses is the approximate weight Mr. Oliver assigned each factor

Using our data from NCAA men's regular season games from 2015 through 2019, we engineered variables corresponding to Mr. Oliver's "four factors" and did find we were able to build a model that explained 88.19% of variance in winning percentage.

However, we were also able to build a model, again using shooting, turnovers, and rebounds, but replacing free throws with TrueShootingPercentage which gives additional consideration for 3-pointers and less consideration for free throw percentage. This model performed slightly better, explaining 88.78% of the variance in winning percentage.

Therefore, one may wish to further analyze if free throws remain as critical an aspect for NCAA basketball winning success as Mr. Oliver proposed in 2004, or if perhaps the game has evolved and more importance should now be given to 3-pointers over free throws.

Conclusions

Splitting up the data into 4 logical partitions of offensive, defensive, neutral, and advanced statistics brought a lot of insights to our team. The first one that became quickly apparent to us was that offensive and defensive statistics alone were not enough to predict a team's winning percentage. Predicting a team's winning percentage using those stats alone produced models that only explained 58% of the data's variability for the defensive statistics and 53% for the offensive. Neutral and advanced statistics, however, produced models with adjusted R-squared scores of 89%.

Based on our analysis, there are a multitude of ways to get to the "right" answer. However, the value derived from the models depends on the user's purpose as the two best models (All Basic Statistics model and Advanced Statistics Model) report equivalent predictive value. Data evaluation is generally easier for an individual to collect and consume when they are raw basic statistics. This method would require the least amount of computational power and is an efficient way to get to the end result. However, this is not to say the advanced statistics are not necessary. They also help coach or team manager get a better understanding of a team's progress and compare it between years and identify specific areas for improvement. Moreover, the advanced statistics are especially useful in the duration of a game since they are evaluated in a 'per possession' manner as opposed to 'per game'. This is especially useful for teams to understand how the team is faring during the game. At the end the user decides what they want to pursue and can choose the model that best fits their purpose.

Citations:

1. https://www.reuters.com/article/us-basketball-ncaa-gambling/americans-to-bet-85-billion-on-ncaas-march-madness-basketball-tournament-report-idUSKCN1QZ0YH

2. https://www.basketball-reference.com/about/factors.html

3. https://en.wikipedia.org/wiki/Offensive_rating

4. https://en.wikipedia.org/wiki/Akaike_information_criterion

5. https://qz.com/1104922/data-analytics-have-revolutionized-the-nba/

6. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwiClL-b753oAhUGKKwKHXGmAe8QFjAAegQIAhAB&url=https%3A%2F%2Fwww.amazon.com%2FBasketball-Paper-Rules-Performance-Analysis%2Fdp%2F1574886886&usg=AOvVaw3Z9xv9-uZjz_vSK3WaZP-a