NBA Wins Analysis

Chapter 0: Abstract

The purpose of this project is to examine the factors that affect total wins for teams in the National Basketball Association (NBA). A linear regression model with 14 independent variables is evaluated through OLS estimation. The Dataset is cross-sectional, containing observations from 30 teams during the 2018-2019 season.

Chapter 1: Introduction

The ultimate goal of any professional sports team is to consistently win. This makes sense, as winning games, as well as championships can have a startling impact on profitability. One such example was Joe Lacob and Peter Guber, who purchased the Golden State Warriors for 450 million dollars in 2010. Now, after three championship titles in the 2010s, that amount has skyrocketed to 4.3 billion dollars. More than a 1000 percent increase in value. To find similar financial success, millions of dollars are allocated to analytical departments in NBA franchises. All of whom are trying to discover the factors that foster a competitive team. Economists and statisticians have conducted various studies and tests towards that end as well, which has led to a substantive amount of relevant literature on the topic developing over time. Most analysts, however, try to evaluate what makes an impactful player rather than an impactful team. This difference is subtle but significant. Famous statistician John Hollinger created PER, a one-number metric that attempts to encapsulate a player's entire value. Similar player evaluation metrics such as adjusted box-plus minus and WOWY are the most common forms of NBA statistical analysis. This study will be considerably different as it will be utilizing fourteen different independent variables in a linear regression model and evaluating their impact on winning percentage from a team perspective. With a cross-sectional dataset, containing observations from 30 teams during the 2018-2019 season. Our data comes from basketball reference and will discussed in greater depth in the Data Source section. The aim is to look at a larger sample size of players rather than individuals, in hopes that the findings will be more easily replicable. For example, if it's found that shooting more three-point shots is advantageous, then a coach can automatically make that change by creating plays that focus on shots outside the arc. While individual player metrics are not useful if there is no clear avenue to get the highest-rated players on your roster in the immediate future. In order to achieve this goal, first the model – including its exogenous variables – will be detailed and then the results from the model estimation will be depicted. We expect the variable of average three point percentage to be the highest positive related variable and for turnovers to be the highest negatively related variable.

Chapter 2: Methodology

Part 1: Data Source

As aforementioned the data comes from basketball reference, a website that stores data on relevant college, European and American basketball statistics. Our data was originally structured into two tables: regular statistics and advanced statistics. Each table contained cross-sectional aggregated data from the 82 games played during the 2018-2019 season and had 30 rows. Each row represents a specific team during that season and the variables reflect statistics (mostly per-game averages) regarding that season. While they had the same number of rows, their column amounts differed as the advanced statistics table had 14 columns while the regular statistics table had 25 columns. The data was in excel format, and so we loaded it onto R-Studio using the 'readxl' package. After loading in the relevant tables we merged them together through the common column 'team'. We labeled this dataset total and it had 38 columns and 30 rows. Now that we had the main dataset we changed the names of relevant independent variables since the 'lm' function in R can't work with variables that have percentage symbols, numbers or periods in their name.

Part 2: The Variables in the Model

Our dependent variable will be total wins and we have a total of 30 observations to evaluate. Total wins represents the number of wins each team received during the 2018-2019 season. We chose the 2018-2019 season because it was the most recent season that wasn't affected by the Covid-19 pandemic which would of been unreliable data to work with. The pandemic brought upon a lot of irregularities such as not all teams playing the same amount of games, periods where there were no fans in the arenas and numerous other changes that would of made the data an erroneous reflection of previous seasons. Out of all the independent variables available we chose 14 that we believed may have some influence on the dependent variable. We will now list all of our variables:

1. W - Total Number of Wins (Dependent Variable/Discrete)
2. BLK - Average number of Blocks per Game (Independent Variable/Continuous)
3. Age - Average Age of the Roster (Independent Variable/Continuous)
4. Pace - Average number of Possessions per game (Independent Variable/Continuous)
5. TOV - Average number of Turnovers per game (Independent Variable/Continuous)
6. TRB - Average number of Total Rebounds per game (Independent Variable/Continuous)
7. FGA - Average number of Field Goal Attempts per game (Independent Variable/Continuous)
8. STL - Average number of Steals per game (Independent Variable/Continuous)
9. FTA - Average number of Free Throw Attempts per game (Independent Variable/Continuous)
10. ThreePA - Average number of Three Point Attempts per game (Independent Variable/Continuous)
11. TwoPA - Average number of Two Point Attempts per game (Independent Variable/Continuous)
12. AttendPerGame - Average Attendance per game (Independent Variable/Continuous)
13. FTr - Average Free Throw Rate per game (Independent Variable/Continuous)
14. ThreePPercentage - Average Three Point Percentage per game (Independent Variable/Continuous)
15. TwoPPercentage - Average Two Point Percentage per game (Independent Variable/Continuous)

Part 3: Modeling

First, we will create an additive model where all the relevant variables are included. Then we will check this full-model for any signs of multicollinearity between our independent variables. Before even creating the full-model we started omitting variables that were considered redundant or would have obvious multicollinearity issues. One such example, were the variables Offensive Rebounding and Defensive Rebounding which are redundant, especially with the inclusion of Total Rebounds (simply the combination of ORB and DRB). Consequently, we only included Total Rebounding in our full-model. We will utilize the 'imcdiag' and 'Vif' functions on our full-model to evaluate their variance inflation factor ratings. Redundant values will be removed. Following the reduction of the model we will conduct a global F-Test to make sure that at least one our independent variables is significantly related to total wins. Once we have confirmed that at least one independent variable is significant, we will do three distinct stepwise procedures to create the best model. These three will be regular stepwise selection using the function 'ols_step_both_p', backward stepwise selection using the function 'ols_step_backward_p' and finally forward stepwise selection utilizing the function 'ols_step_forward_p'. We will evaluate each of these models, along with their R-squared Adj. values and their Root Mean Square Error (RMSE) in order to find the combination of independent variables that create the best-fit model. Once we have our best combination of independent variables we will add interaction terms into the model and evaluate the significance of the interaction terms through individual t-tests. If there are no significant interaction terms then we will continue with the original best fit model. However, if any interaction terms are significant then they will be included along with the involved variables unless its inclusion worsens the overall fit of the model. The final part in the model creation process is checking whether a higher-order-model is necessary. Any use of a higher order model will be limited because we don't want to over-fit the model and the higher power variables are limited in terms of real-world explanation. Finally, once our model has passed these various checks and tests we will make sure they match five crucial assumptions:

1. Linearity Assumption - Evaluating the Residual vs Fitted Plot
2. Normality Assumption - Evaluating the Normal Q-Q plots; Conducting Shapiro-Wilk normality test
3. Equal Variance Assumption - Evaluating the Residual vs Fitted and Scale-Location Plots; Conducting Breusch-Pagan test
4. Multicollinearity - Evaluating GGpairs plot; Calculating and Evaluating the variance inflation factors (VIF)
5. Outliers - Evaluate Cook's distance and leverage

Notice that we do not have a test for the independence assumption since our data is not time-series. If any of these assumptions are not met we can conduct various tests and procedures to make the model better match the assumption in question. Finally, once all the assumptions have been tested the coefficients and the final R-Squared adjusted value will be interpreted.

Chapter 3: Results

#PACKAGES

```r
library(olsrr)
```

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers
```

```r
library(leaps)
library(readr)
library(readxl)
library(mctest)
library(car)
```

```
## Loading required package: carData
```

```r
library(ggplot2)
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v purrr   0.3.4     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::recode() masks car::recode()
## x purrr::some()   masks car::some()
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

#Part 0. Data Wrangling

Here we uploaded the datasets onto R-Studio, combined them by utilizing the ‘Team’ column and renamed essential variables for easier use of the ‘lm’ function.

```
Advanved_Stats <- read_excel("~/Desktop/Data 603 Project Data.xls")
Regular_Stats <- read_excel("~/Desktop/Data603.xls")
total <- merge(Advanved_Stats,Regular_Stats,by=c("Team","Team"))

#Changing Names of columns
setnames(total, old = c('3PAr','TS%','Attend.','Attend./G','FG%','3P','3PA'),
         new = c('ThreePAr','TSPercent','Attend','AttendPerGame','FGPercent',
                 'ThreeP','ThreePA'))
setnames(total, old = c('3P%','2P','2PA','2P%','FT%'),
         new = c('ThreePPercentage','TwoP','TwoPA','TwoPPercentage',
                 'FreeThrowPercentage'))
```

Part 1. Variable Selection Procedures

We built a full model that consisted of every relevant explanatory variable as a base comparison as shown below.

```r
#Combine potentially significant Variables into Multi-Linear Model
fullmodel <- lm(data=total,formula=W~BLK+Age+Pace+TOV+TRB+FGA+STL+FTA+ThreePA+
                  TwoPA+AttendPerGame+FTr+ThreePPercentage+TwoPPercentage)
summary(fullmodel)
```

```
## 
## Call:
## lm(formula = W ~ BLK + Age + Pace + TOV + TRB + FGA + STL + FTA + 
##     ThreePA + TwoPA + AttendPerGame + FTr + ThreePPercentage + 
##     TwoPPercentage, data = total)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.5885 -1.9873 -0.1481  1.4557  6.2342 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       5.860e+01  3.993e+02   0.147 0.885280    
## BLK               1.430e+00  1.829e+00   0.782 0.446375    
## Age              -2.905e-01  1.049e+00  -0.277 0.785591    
## Pace             -3.779e+00  1.209e+00  -3.126 0.006939 ** 
## TOV              -1.634e+00  1.350e+00  -1.210 0.244936    
## TRB               3.361e+00  7.870e-01   4.271 0.000669 ***
## FGA              -1.902e+01  1.546e+01  -1.230 0.237620    
## STL               4.268e+00  1.341e+00   3.182 0.006190 ** 
## FTA               1.082e+01  1.779e+01   0.608 0.551997    
## ThreePA           1.763e+01  1.546e+01   1.140 0.271975    
## TwoPA             1.747e+01  1.542e+01   1.133 0.274963    
## AttendPerGame    -2.507e-04  5.977e-04  -0.419 0.680835    
## FTr              -8.250e+02  1.561e+03  -0.529 0.604882    
## ThreePPercentage  3.959e+02  7.537e+01   5.253 9.74e-05 ***
## TwoPPercentage    3.086e+02  7.445e+01   4.145 0.000864 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.989 on 15 degrees of freedom
## Multiple R-squared:  0.9431, Adjusted R-squared:  0.8901 
## F-statistic: 17.77 on 14 and 15 DF,  p-value: 7.703e-07
```

The above result is helpful to identify the independent variables that are significant and should be included in our further analysis.

But to begin with the variable selection procedure, we decided to do some preliminary multicollinearity testing and eliminate the predictor variables that have higher VIF values, because if we directly perform a stepwise regression procedure on our full model, it can result in removing the important predictors because of multicollinearity.

#Part 1a. MULTI-COLLINEARITY TEST:

```
#Utilize Empirical Tests in order to see if there is multi-collinearity
imcdiag(fullmodel, method="VIF")
```

```
##
## Call:
## imcdiag(mod = fullmodel, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                        VIF detection
## BLK                 3.2500         0
## Age                 3.7458         0
## Pace               12.0938         1
## TOV                 3.4868         0
## TRB                 5.0305         0
## FGA              1941.0213         1
## STL                 2.3155         0
## FTA              2722.8614         1
## ThreePA          8054.2340         1
## TwoPA            8901.4010         1
## AttendPerGame       1.7550         0
## FTr              2924.8658         1
## ThreePPercentage    2.4310         0
## TwoPPercentage      4.1185         0
##
## Multicollinearity may be due to Pace FGA FTA ThreePA TwoPA FTr regressors
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

```
vif(fullmodel)
```

```
##              BLK              Age             Pace              TOV
##         3.250024         3.745820        12.093792         3.486765
##              TRB              FGA              STL              FTA
##         5.030516      1941.021314         2.315546      2722.861391
##          ThreePA            TwoPA    AttendPerGame              FTr
##      8054.233991      8901.401030         1.754975      2924.865789
## ThreePPercentage   TwoPPercentage
##         2.431000         4.118453
```

```
Conclusion1 <- "There is multi-collinearity present in pace, FGA, FTA, ThreePA,
TWOPA and FTr."

#Removed FTA and TwoPA in order to remove multi-collinearity by removing
#redundant variables
Newtlm <- lm(data=total,formula=W~BLK+Age+Pace+TOV+TRB+FGA+STL+FTA+
               AttendPerGame+FTr+ThreePPercentage+TwoPPercentage)

imcdiag(Newtlm, method="VIF")
```

```
##
## Call:
## imcdiag(mod = Newtlm, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                         VIF detection
## BLK                  2.9136         0
## Age                  3.6367         0
## Pace                10.7632         1
## TOV                  3.3418         0
## TRB                  4.5181         0
## FGA                163.9076         1
## STL                  2.2960         0
## FTA               2697.8826         1
## AttendPerGame        1.4303         0
## FTr               2890.6687         1
## ThreePPercentage     2.2468         0
## TwoPPercentage       3.2675         0
##
## Multicollinearity may be due to Pace FGA FTA FTr regressors
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

```
vif(Newtlm)
```

```
##              BLK              Age             Pace              TOV
##         2.913606         3.636703        10.763169         3.341804
##              TRB              FGA              STL              FTA
##         4.518148       163.907650         2.296038      2697.882617
##    AttendPerGame              FTr ThreePPercentage   TwoPPercentage
##         1.430294      2890.668661         2.246802         3.267478
```

```
Conclusion2 <- "While the first vif test find no multi-collinearity,
FGA and Pace do seem to have some relationship as they are both
close to 10 on the VIF scale. So I think they should be removed."

#Removed FGA in order to eliminate multi-collinearity by removing redundant
#variables
Newtlm2 <- lm(data=total,formula=W~BLK+Age+TOV+TRB+Pace+STL+FTA+
                AttendPerGame+ThreePPercentage+TwoPPercentage)

imcdiag(Newtlm2, method="VIF")
```

```
##
## Call:
## imcdiag(mod = Newtlm2, method = "VIF")
##
##
```

```
##  VIF Multicollinearity Diagnostics
##
##                       VIF detection
## BLK                1.9482         0
## Age                3.5853         0
## TOV                1.8332         0
## TRB                2.2159         0
## Pace               3.9464         0
## STL                1.7333         0
## FTA                1.8384         0
## AttendPerGame      1.3058         0
## ThreePPercentage 1.9174         0
## TwoPPercentage   2.9854         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

```
vif(Newtlm2)
```

```
##              BLK              Age              TOV              TRB
##         1.948176         3.585294         1.833176         2.215857
##             Pace              STL              FTA    AttendPerGame
##         3.946403         1.733340         1.838383         1.305797
## ThreePPercentage   TwoPPercentage
##         1.917359         2.985414
```

```
Conclusion3 <- "There is no long multi-collinearity present in the model."
```

From the above result it can be seen that, the predictor variables that does not show multicollinearity are: Age,BLK+TOV+TRB+Pace+STL+FTA+AttendPerGame,ThreePPercentage and TwoPPercentage.

Therefore,we will be using these variables for our step-wise,forward and backward regression procedures to select the predictor variables for our best fit model.

##Part 1b. Global F-test##

Significance level $\alpha = 0.05$

Therefore the hypothesis can be defined as-

$H0 : \beta1 = \beta2 = \beta3 = \beta4 = \beta5 = \beta6 = \beta7 = 0$

$Ha : atleast\ one\ \beta_i \neq 0$

```
reg1<-lm(data=total,formula=W~TOV+FTA+TRB+Pace+STL+ThreePPercentage+
           TwoPPercentage)
reg2<-lm(W~1, data=total)
anova(reg2,reg1)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: W ~ 1
## Model 2: W ~ TOV + FTA + TRB + Pace + STL + ThreePPercentage + TwoPPercentage
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     29 4196.0
## 2     22  293.5  7    3902.5 41.792 2.89e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above F- test we can see that the p-value<0.05.Therefore, we reject the Null Hypothesis. In other words, we can say that the F-test suggests that Number of wins depends on at least one of the independent variable variable.

#Part 1c. STEPWISE,FORWARD AND BACKWARD REGRESSION MODEL:

```
stepmod = ols_step_both_p(Newtlm2,pent = 0.05, prem = 0.1, details=FALSE)
summary(stepmod$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9807 -2.1463  0.1678  1.5304  5.6577
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -148.6837    37.0107  -4.017 0.000578 ***
## TRB                 3.7100     0.4374   8.482 2.20e-08 ***
## STL                 4.8999     0.9912   4.943 6.05e-05 ***
## ThreePPercentage  342.7352    50.0063   6.854 6.95e-07 ***
## Pace               -2.9395     0.5034  -5.840 7.09e-06 ***
## TwoPPercentage    307.5280    38.6313   7.961 6.41e-08 ***
## FTA                 1.0456     0.3548   2.947 0.007452 **
## TOV                -1.9284     0.8858  -2.177 0.040498 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.652 on 22 degrees of freedom
## Multiple R-squared:  0.9301, Adjusted R-squared:  0.9078
## F-statistic: 41.79 on 7 and 22 DF,  p-value: 2.89e-11
```

```
Backmodel = ols_step_backward_p(Newtlm2,pent = 0.05, prem = 0.1, details=FALSE)
summary(Backmodel$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9807 -2.1463  0.1678  1.5304  5.6577
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -148.6837    37.0107  -4.017 0.000578 ***
## TOV                -1.9284     0.8858  -2.177 0.040498 *
## TRB                 3.7100     0.4374   8.482 2.20e-08 ***
## Pace               -2.9395     0.5034  -5.840 7.09e-06 ***
## STL                 4.8999     0.9912   4.943 6.05e-05 ***
## FTA                 1.0456     0.3548   2.947 0.007452 **
## ThreePPercentage  342.7352    50.0063   6.854 6.95e-07 ***
## TwoPPercentage    307.5280    38.6313   7.961 6.41e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.652 on 22 degrees of freedom
## Multiple R-squared:  0.9301, Adjusted R-squared:  0.9078
## F-statistic: 41.79 on 7 and 22 DF,  p-value: 2.89e-11
```

```
Forwardmodel = ols_step_forward_p(Newtlm2,pent = 0.05, prem = 0.1,
                                  details=FALSE)
summary(Forwardmodel$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1052 -1.9395  0.4687  2.6005  6.0609
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -175.2798    46.1721  -3.796 0.000932 ***
## Age                 1.1574     0.8883   1.303 0.205485
## TRB                 3.8220     0.5070   7.539 1.17e-07 ***
## STL                 4.7758     1.1457   4.168 0.000370 ***
## ThreePPercentage  329.5274    61.4946   5.359 1.92e-05 ***
## Pace               -2.6700     0.5557  -4.804 7.58e-05 ***
## TwoPPercentage    243.6792    58.9776   4.132 0.000405 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.254 on 23 degrees of freedom
## Multiple R-squared:  0.9008, Adjusted R-squared:  0.8749
## F-statistic:  34.8 on 6 and 23 DF,  p-value: 2.013e-10
```

```
#first order model
Newtlm3 <- lm(data=total,formula=W~TOV+TRB+Pace+STL+FTA+ThreePPercentage+
                TwoPPercentage)
```

From the output it can be seen that, the forward regression procedure contains a insignificant term “Age” whose p-value is greater than 0.05, whereas in stepwise and backward regression procedure all the variables are significant and it also have the highest Adjacent R-Squared value and the lowest RMSE value,which clearly indicates that the model containing the predictor variables that are obtained using the backward and stepwise regression can be considered as a good fit model.

Therefore,the predictor variables for our first order model are: TOV,Pace,TRB,STL,FTA,ThreePPercentge and TwoPPercentage.

Our first order model is given below:

$$\hat{Y}_{wins} = \hat{\beta}_0 + \hat{\beta}_1 TOV + \hat{\beta}_2 PACE + \hat{\beta}_3 TRB + \hat{\beta}_4 STL + \hat{\beta}_5 FTA + \hat{\beta}_6 ThreePPercentage + \hat{\beta}_7 TwoPPercentage$$

#Part 1d. INTERACTION-MODEL

We are also interested in identifying the significant interactions terms that can be included in our first order model. Therefore,the interaction model with all possible interaction terms is shown shown below:

```
interaction_model <- lm(data=total,formula=W~(TOV+TRB+Pace+STL+FTA+
                                            ThreePPercentage+TwoPPercentage)^2)
summary(interaction_model)
```

```
##
## Call:
## lm(formula = W ~ (TOV + TRB + Pace + STL + FTA + ThreePPercentage +
##     TwoPPercentage)^2, data = total)
##
## Residuals:
##        1        2        3        4        5        6        7        8
##  0.07351 -0.05129  0.38193 -0.02229 -0.34867  0.23479  0.22375 -0.32419
##        9       10       11       12       13       14       15       16
## -0.27734 -0.28679 -0.08545 -0.30503 -0.09162 -0.28129 -0.03242  0.22617
##       17       18       19       20       21       22       23       24
## -0.01995  0.78093  0.26366 -0.35494 -0.23379  0.27040  0.03584  0.15166
##       25       26       27       28       29       30
## -0.39265 -0.17831  0.02479  1.26999 -0.22637 -0.42503
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -1.458e+04  3.165e+03  -4.608   0.1360
## TOV                              1.288e+02  7.922e+01   1.626   0.3510
## TRB                              1.408e+02  1.180e+02   1.193   0.4441
## Pace                             3.500e+01  5.002e+01   0.700   0.6114
## STL                             -5.700e+02  1.476e+02  -3.863   0.1613
## FTA                              1.371e+02  4.578e+01   2.994   0.2052
## ThreePPercentage                -1.170e+03  5.397e+03  -0.217   0.8641
## TwoPPercentage                   3.749e+04  6.739e+03   5.563   0.1132
## TOV:TRB                          8.473e+00  2.594e+00   3.267   0.1891
## TOV:Pace                        -3.629e+00  1.382e+00  -2.626   0.2316
## TOV:STL                          9.612e+00  2.417e+00   3.977   0.1568
## TOV:FTA                          9.810e-01  6.376e-01   1.539   0.3669
## TOV:ThreePPercentage            -6.822e+01  1.183e+02  -0.577   0.6669
## TOV:TwoPPercentage              -4.204e+02  1.176e+02  -3.575   0.1737
## TRB:Pace                        -2.005e+00  5.338e-01  -3.756   0.1657
## TRB:STL                         -9.077e+00  3.975e+00  -2.283   0.2628
## TRB:FTA                          6.308e+00  1.261e+00   5.002   0.1256
## TRB:ThreePPercentage            -7.386e+02  3.672e+02  -2.012   0.2937
## TRB:TwoPPercentage               2.384e+02  9.186e+01   2.595   0.2341
## Pace:STL                          4.653e+00  3.026e+00   1.537   0.3671
## Pace:FTA                         -1.500e+00  5.171e-01  -2.900   0.2114
## Pace:ThreePPercentage             4.886e+02  2.127e+02   2.298   0.2613
## Pace:TwoPPercentage              -1.377e+02  7.272e+01  -1.893   0.3094
## STL:FTA                           1.049e+01  1.600e+00   6.561   0.0963 .
## STL:ThreePPercentage             -7.138e+01  3.606e+02  -0.198   0.8756
## STL:TwoPPercentage                3.192e+02  1.497e+02   2.133   0.2791
## FTA:ThreePPercentage              1.426e+02  5.521e+01   2.583   0.2351
## FTA:TwoPPercentage               -8.051e+02  1.432e+02  -5.621   0.1121
## ThreePPercentage:TwoPPercentage -3.254e+04  8.698e+03  -3.741   0.1663
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.968 on 1 degrees of freedom
## Multiple R-squared:  0.9991, Adjusted R-squared:  0.9732
## F-statistic: 38.64 on 28 and 1 DF,  p-value: 0.1267
```

We evaluate each result by utilizing the individual t-test method. From the above result, it can be clearly seen that there are no significant interaction terms with p-value less than 0.05, which indicates that our best fit model should not contain any interaction term.

Therefore, our best fit model is our first order model with no interaction terms.

#Part 1e. HIGH ORDER REGRESSION MODEL

We are also interested in identifying the predictor variables that show a curvature in the relationship with the response variable, therefore our higher order regression model to identify those variables is given below:

```
NewtlmHigh_Order <- lm(data=total,formula=W~I(Age^2)+Age+I(TRB^2)+TRB+
            I(Pace^2)+Pace+I(STL^2)+STL+I(ThreePPercentage^2)+ThreePPercentage)
summary(NewtlmHigh_Order)
```

```
##
## Call:
## lm(formula = W ~ I(Age^2) + Age + I(TRB^2) + TRB + I(Pace^2) +
##     Pace + I(STL^2) + STL + I(ThreePPercentage^2) + ThreePPercentage,
##     data = total)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3968  -2.8819  -0.5989   2.5340  10.8438
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -5047.0905  2505.6347  -2.014   0.0584 .
## I(Age^2)                 -0.1893     0.4777  -0.396   0.6964
## Age                      13.4309    25.0071   0.537   0.5974
## I(TRB^2)                  0.3495     0.2074   1.685   0.1083
## TRB                     -27.5844    18.7647  -1.470   0.1579
## I(Pace^2)                -0.5107     0.2382  -2.144   0.0452 *
## Pace                    100.6722    47.7988   2.106   0.0487 *
## I(STL^2)                 -1.0509     1.5950  -0.659   0.5179
## STL                      21.1463    24.6540   0.858   0.4017
## I(ThreePPercentage^2) -1817.9575  4740.3099  -0.384   0.7056
## ThreePPercentage       1604.1234  3416.6748   0.469   0.6441
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.316 on 19 degrees of freedom
## Multiple R-squared:  0.8721, Adjusted R-squared:  0.8047
## F-statistic: 12.95 on 10 and 19 DF,  p-value: 1.708e-06
```

```
NewtlmHigh_Order2 <-lm(data=total,formula=W~Age+TRB+Pace+I(Pace^2)+STL+
                         ThreePPercentage)
summary(NewtlmHigh_Order2)
```

```
##
## Call:
## lm(formula = W ~ Age + TRB + Pace + I(Pace^2) + STL + ThreePPercentage,
##     data = total)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0734  -3.0448   0.3614   3.2085  10.2408
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -4559.2772  2304.3662  -1.979 0.059965 .
## Age                  3.6540     0.8294   4.406 0.000205 ***
## TRB                  4.2140     0.6147   6.855 5.45e-07 ***
## Pace                85.2919    46.0069   1.854 0.076624 .
## I(Pace^2)           -0.4341     0.2294  -1.893 0.071059 .
## STL                  5.9526     1.3618   4.371 0.000223 ***
## ThreePPercentage   232.5034    73.3367   3.170 0.004270 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.224 on 23 degrees of freedom
## Multiple R-squared:  0.8504, Adjusted R-squared:  0.8114
## F-statistic:  21.8 on 6 and 23 DF,  p-value: 2.033e-08
```

From the above result it can be seen that there are no significant higher order regression terms,therefore it can be concluded that no independent variable in our first order model follows a curvature in the relationship with the response variable.

After performing all the variable selection procedures,and identification of interaction terms and higher regression terms,our best fit regression model is as follows:

$$\hat{Y}_{wins} = \hat{\beta}_0 + \hat{\beta}_1 TOV + \hat{\beta}_2 PACE + \hat{\beta}_3 TRB + \hat{\beta}_4 STL + \hat{\beta}_5 FTA + \hat{\beta}_6 ThreePPercentage + \hat{\beta}_7 TwoPPercentage$$

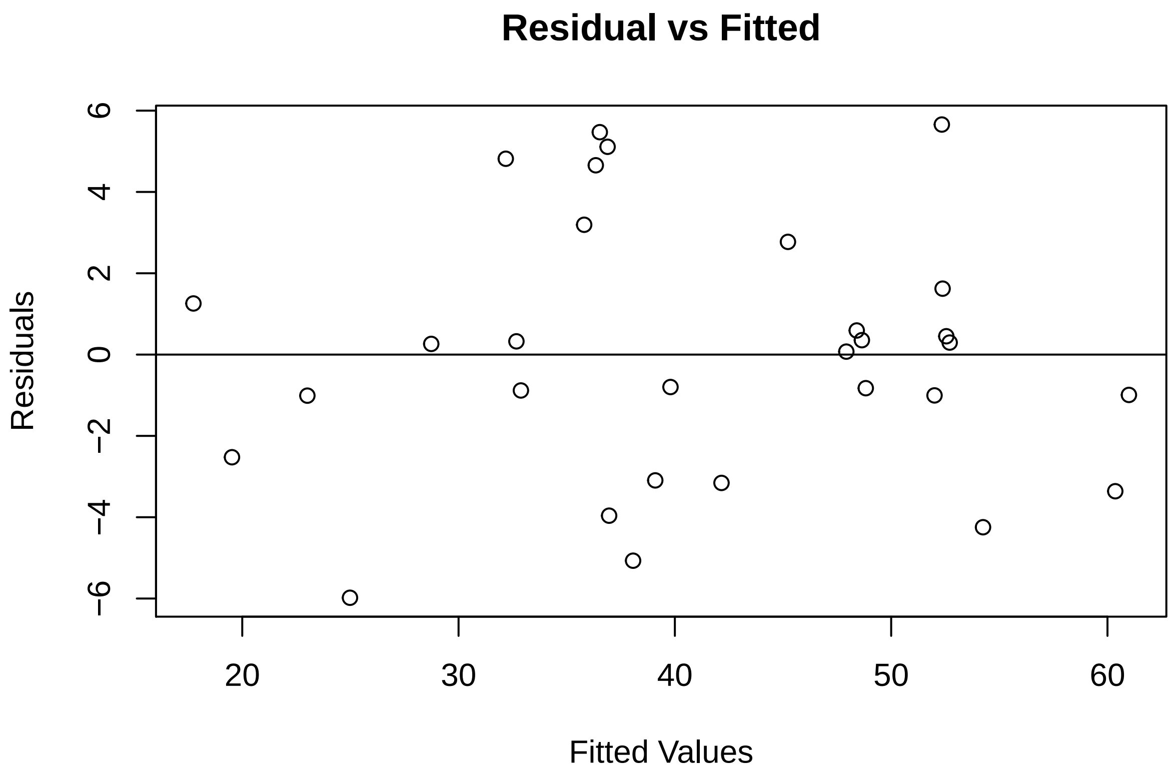
2. Multiple Regression Assumptions:

#Part 2a. Linearity Assumption

Our model assumes that there is straight-line relationship between the predictor variables and the response variable.To test whether the linearity assumption holds true, we will be plotting residual vs fitted plot.The plot for testing the assumption is shown below:

```
Newtlm3 <- lm(data=total,formula=W~TOV+TRB+Pace+STL+FTA+ThreePPercentage+
                TwoPPercentage)
```

```r
plot(fitted(Newtlm3), residuals(Newtlm3),xlab="Fitted Values",
     ylab="Residuals")
abline(h=0,lty=1)
title("Residual vs Fitted")
```



From the graph, it can be seen that the residual vs fitted plot for our best fit model does not shows any pattern and is almost linear, which clearly indicates that the linearity assumption holds true for our predicted model.

#Part 2b. Equal Variance Assumption

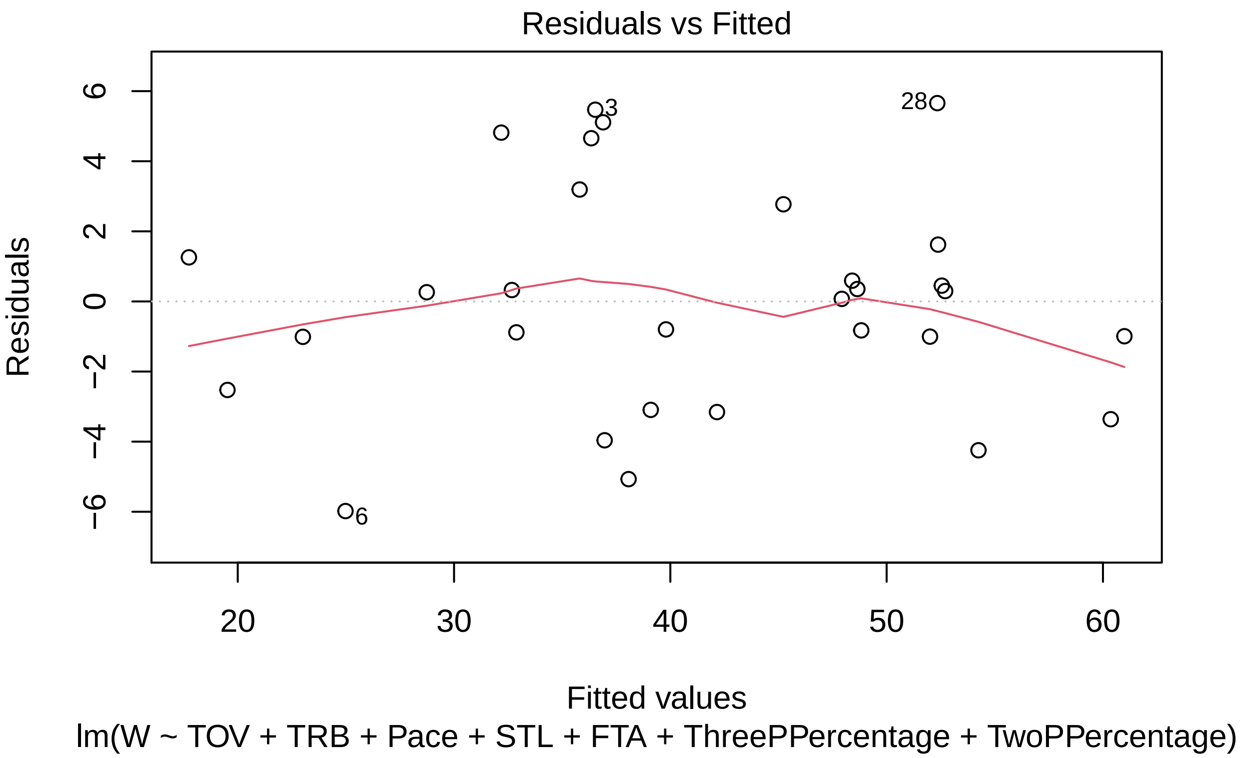
One of the key assumptions is that there is equal variance among the residuals. In order to test that we can look at the Residuals vs Fitted and Scale-Location plots. Furthermore, we can conduct a Bruesch-Pagan Test.

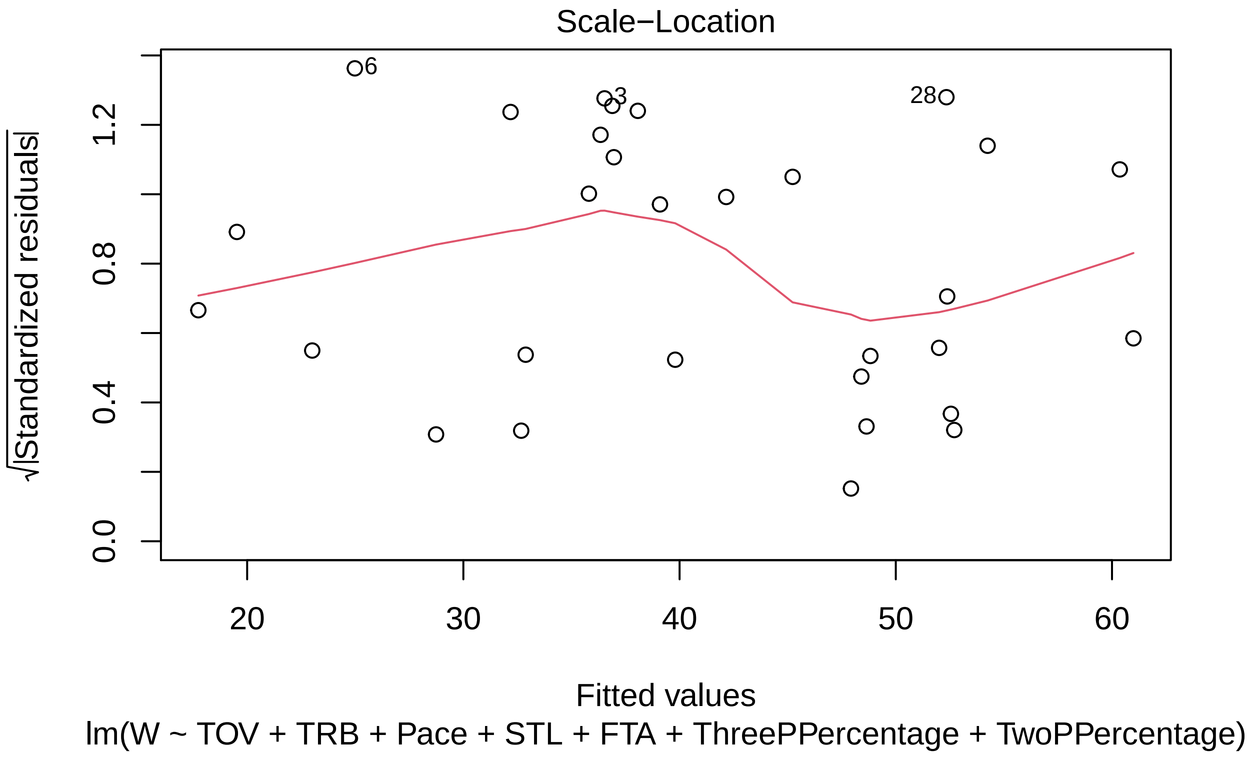
Significance level $\alpha = 0.05$

Therefore the hypothesis can be defined as-

$H0 : Heteroscedasticityisnotpresent(homoscedasticity)$ $Ha : Heteroscedasticityispresent$

```
plot(Newtlm3, which=c(1,3))
```





```
bptest(Newtlm3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Newtlm3
## BP = 7.7445, df = 7, p-value = 0.3557
```

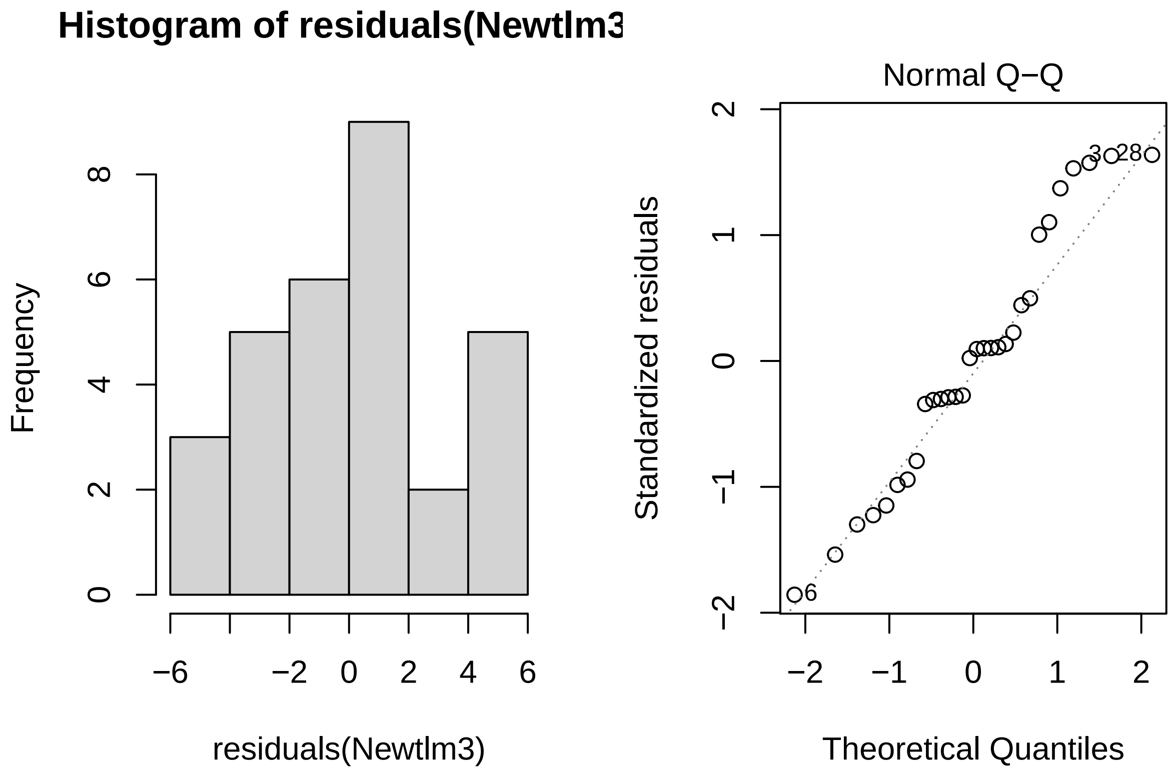
```
pvalue <- 0.3557
```

While our lines in both graphs aren't perfectly straight they do closely resemble what we would expect if the equal variance assumption was met. The studentized Breusch-Pagan test confirmed that the Equal Variance Assumption was met as we fail to reject the null with our $p - value = 0.3557$. Consequently, Heteroscedasticity is not present in the model.

#Part 2c. Normality Assumption For our best fit model to folow the normality assumption the error between the observed and the residuals of the regression model should be normally distributed.To test this assumption a histogram and a Q-Q plot has been shown below:

```
par(mfrow=c(1,2))

#histogram
hist(residuals(Newtlm3))
#normal q-q plot
plot(Newtlm3, which=2)
```



From the above graph it can be seen that the majority of the points fall close to the diagonal reference line, which suggest that our model follows a normal distribution.

To further verify our results, we can also use shapiro-wilk test.The Hypothesis for the test is defined as follows:

$H0 : Sampledataissignificantlydistributed\ Ha : Sampledataisnotsignificantlydistributed$

let the significance level for our test be $\alpha = 0.05$

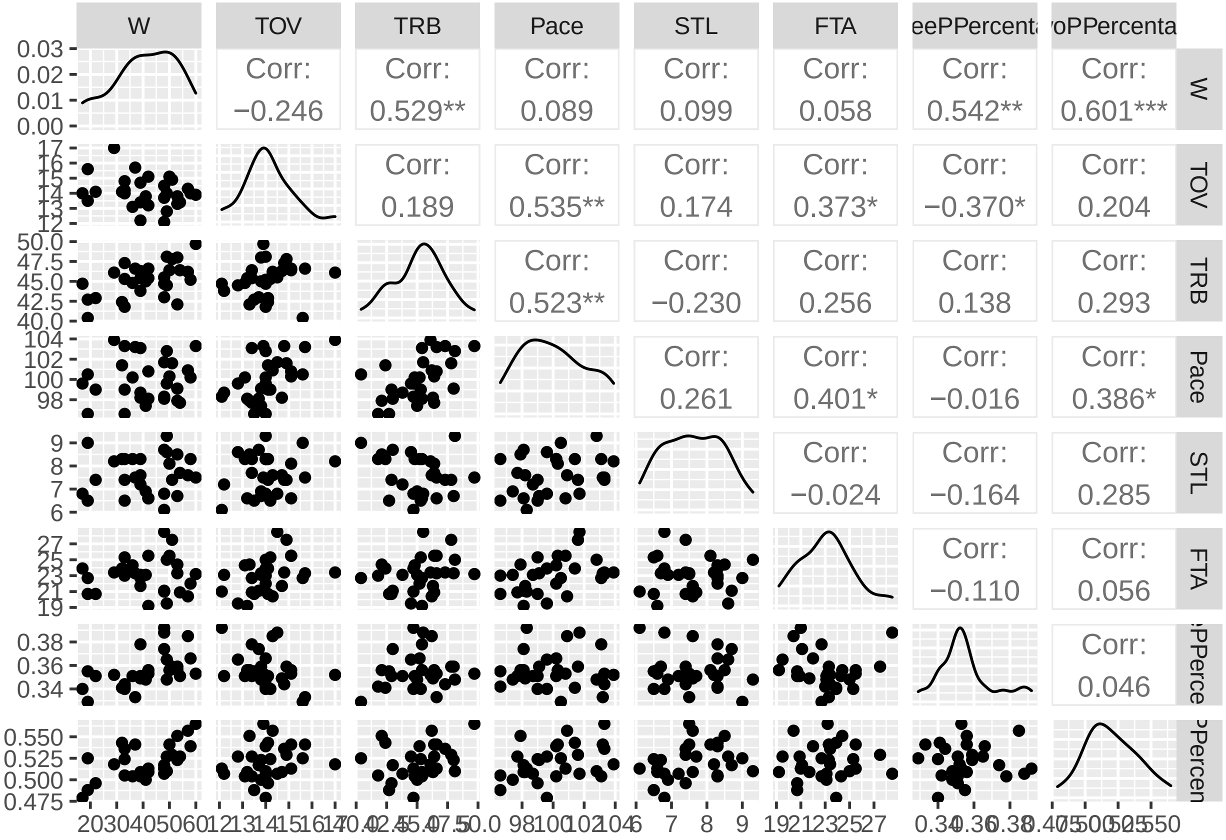
```
#shapiro-wilk test
shapiro.test(residuals(Newtlm3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(Newtlm3)
## W = 0.95918, p-value = 0.2951
```

From the output it can be seen that the $p - value = 0.2951$ is greater than the significance level.Thus we fail to reject the Null Hypothesis which indicates that the errors terms are normality distributed.Hence,our model follows follows the Normality assumptions. #Part 3c. Further Multi-Collinearity Testing

We know that none of the variables have exceedingly high VIF values from our earlier testing. So here, we will focus on looking at ggpairs to make sure none of the indpendent variables have a correlation coefficient greater than r > .80

```r
ggpairs(data= select(total,c(W,TOV,TRB,Pace,STL,FTA,ThreePPercentage,
                             TwoPPercentage)))
```

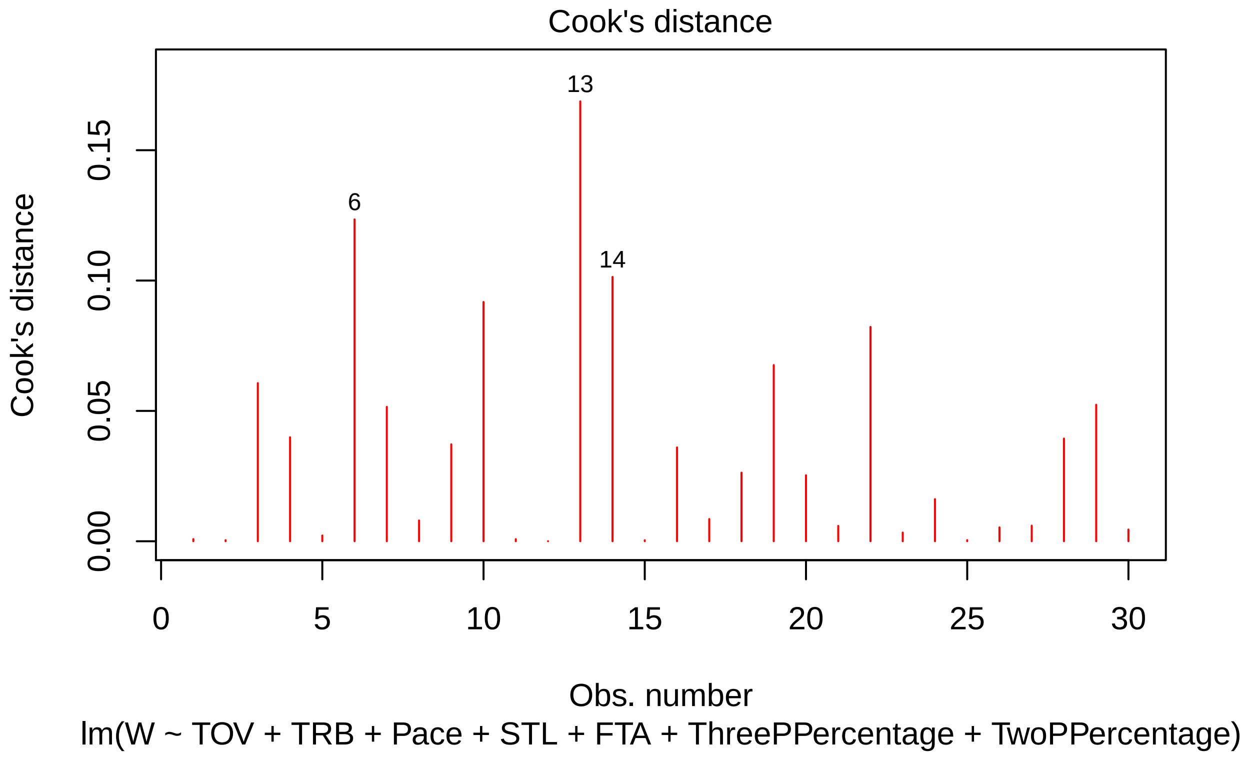


As none of the values are near the .8 correlation amount and we know that the VIF values are low, we can assume that there is no multicollinearity in the model.

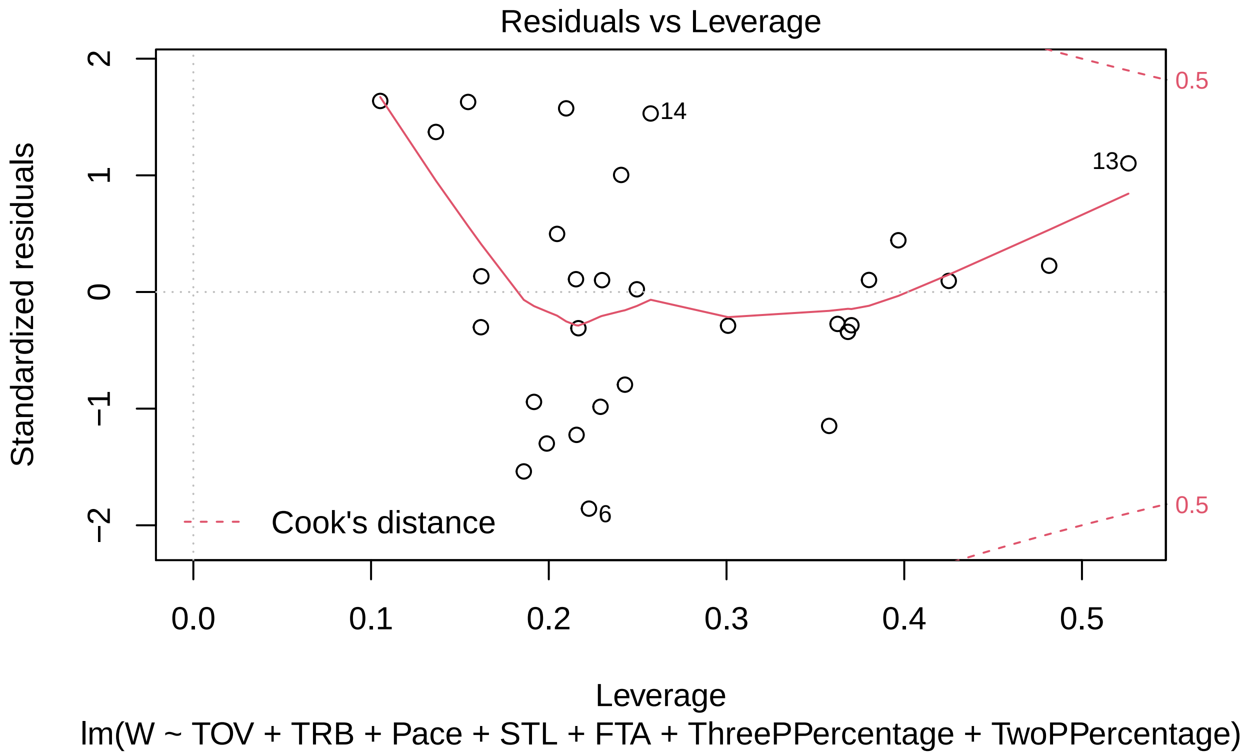
#Part 3c. Significant Outliers

In order to evaluate any influential points and outliers in the model we have to look at the Cook's Distance and the Residuals vs Leverage Plot.

```r
#Cooks Distance
plot(Newtlm3,pch=18,col="red",which=c(4))
```



```
plot(Newtlm3,which=5)
```



The first graph shows the Di values, where any value over 0.5 should be scrutinized and any value over 1.0 is definitely considered an influential outlier. However, we find no value near 0.5 and so there is no reason for concern when looking at the Cook’s Distance plot. Similarly, when looking at the Residual Vs. Leverage plot there are no points past the 0.5 mark. Consequently, when it comes to outliers, there is no reason to change or manipulate our model further.

Part 3: Interpertation

The final Theoretical Model:

’W = -B0 + B1*TRB + B2*STL + B3*ThreePPercentage - B4*Pace + B5*TwoPPercentage + B6*FTA - B7*TOV + Ei’

The final Model:

’W = -148.6837 + 3.7100*TRB + 4.8999*STL + 342.7352*ThreePPercentage - 2.9395*Pace + 307.5280*TwoPPercentage + 1.0456*FTA - 1.9284*TOV’

R-Squared Adj: 0.9078 RMSE: 3.652

Interpertation: **For every indpendent variable, it assumed for their interpertations that all other independent variables are held constant.

B0: This represents the value of wins when all the indpendent variables are equal to zero. Its interpretation does not have much real-world application.

B1: Indicates that one additional average Total Rebounds per game leads to an additional 3.7100 wins.

B2: Indicates that one additional average Steals per game leads to an additional 4.8999 wins.

B3: A one percentage point increase in average three-point percentage leads to a 343 percent increase in wins. For example, if a team would win 10 wins at 39 percent three-point shooting per game, if their average three-point shooting increased to 40 percent then they would win around 44 games.

B4: Indicates that one additional possession per game leads to 2.9395 less wins.

B5: A one percentage point increase in two-point percentage leads to a 308 percent increase in wins. For example, if a team would win 10 wins at 39 percent two-point shooting per game, if their average two-point shooting increased to 40 percent then they would win around 41 games.

B6: Indicates that one additional average free-throw attempted per game leads to 1.0456 more wins.

B7: Indicates that one additional average turnover per game leads to 1.9284 less wins.

R-Squared Adj: According to our R-Squared Adj. the model accounts for 90.78% of the variation of the dependent variable. This is an extremely high R-Squared Adj and indicates that it has a great fit. However, there is one limitation in our model, in that it reflects a single-season. So, while its prediction prowess and Adj R-Squared may not necessarily be as high in predicting other season (90 percent of variation is extremely high) it is undoubtedly still a very well fit model and predictor of wins in general.

Chapter 4: Conclusion and Discussion #Conclusion

#Part 1: Summary of Findings

The goal of this project was to predict wins in the NBA by utilizing various relevant statistics. We also wanted to evaluate whether certain variables were more important than others in evaluating total wins and to see the relationship (if any) between all of our variables.

Firstly, during our checks for multi-collinearity we used empirical tests to calculate the VIF's of the independent variables. After the test we found out there is multi-collinearity present in Pace, FGA,FTA,ThreePA,TWOPA, and FTr. We eliminated the variables from this group that were redundant to improve the model.

Then, we used stepwise regression model comparison along with forward selection and backward elimination model.The stepwise model and backward regression procedure had really high Adjusted R-squared values, showing that almost 90% of the variance in the data is explained by the model and the lowest RMSE Value.

None of the relationships in interaction model were significant. Therefore, we didn't use the interaction model. Even though higher order term were significant, they were borderline significant. After we combined the main effects with higher order term, and the interaction together, we found out that all the terms in our model were not significant through the partial F-test. To fix this we removed higher order term from our model and only kept the significant terms. Overall our best fit model includes (turnovers) TOV, (Total Rebounds) TRB, Pace, (Steals) STL, (Free Throw Attempts) FTA, ThreePPercentage, TwoPPercentage.

#Part 1b: Final model and Interpertation of variables

The final Theoretical Model:

'W = -B0 + B1*TRB + B2*STL + B3*ThreePPercentage - B4*Pace + B5*TwoPPercentage + B6*FTA - B7*TOV + Ei'

The final Model:

'W = -148.6837 + 3.7100*TRB + 4.8999*STL + 342.7352*ThreePPercentage - 2.9395*Pace + 307.5280*TwoPPercentage + 1.0456*FTA - 1.9284*TOV'

R-Squared Adj: 0.9078 RMSE: 3.652

Interpretation: **For every independent variable, it assumed for their interpretations that all other independent variables are held constant.

B0: This represents the value of wins when all the independent variables are equal to zero. Its interpretation does not have much real-world application.

B1: Indicates that one additional average Total Rebounds per game leads to an additional 3.7100 wins.

B2: Indicates that one additional average Steals per game leads to an additional 4.8999 wins.

B3: A one percentage point increase in average three-point percentage leads to a 343 percent increase in wins. For example, if a team would win 10 wins at 39 percent three-point shooting per game, if their average three-point shooting increased to 40 percent then they would win around 44 games.

B4: Indicates that one additional possession per game leads to 2.9395 less wins.

B5: A one percentage point increase in two-point percentage leads to a 308 percent increase in wins. For example, if a team would win 10 wins at 39 percent two-point shooting per game, if their average two-point shooting increased to 40 percent then they would win around 41 games.

B6: Indicates that one additional average free-throw attempted per game leads to 1.0456 more wins.

B7: Indicates that one additional average turnover per game leads to 1.9284 less wins.

R-Squared Adj: According to our R-Squared Adj. the model accounts for 90.78% of the variation of the dependent variable. This is an extremely high R-Squared Adj and indicates that it has a great fit. However, there is one limitation in our model, in that it reflects a single-season. So, while its prediction prowess and Adj R-Squared may not necessarily be as high in predicting other season (90 percent of variation is extremely high) it is undoubtedly still a very well fit model and predictor of wins in general.

RMSE: The standard deviation of the residuals is 3.652. The lower the value the better and this was among the lowest of any model we created.

Overall, this model was the best because it underwent tremendous testing and met all the assumptions. Furthermore, the high Adjusted R-Squared value indicates that the model is already a tremendous fit and any further attempts could lead to the over-fitting of the model.

#Discussion

#Part 1:

As avid NBA fans, we felt that creating a model to predict the outcome of NBA games would be an enjoyable project. We were able to utilize many of the concepts learned in our DATA 603 class for this project — including multiple linear regression, higher-order models, interaction models, checking for assumptions of linearity, heteroscedasticity, normality, multicollinearity, etc, — and want to thank Professor Dr. Thuntida Ngamkham for her fantastic work in teaching throughout the semester.

One of the main challenges in our modeling was to deal with multicollinearity. There was a high correlation between Pace, FGA, FTA, ThreePA, TWOPA, and FTr. After eliminating five variables FGA, FTA, ThreePA, TWOPA, and FTr, multicollinearity was removed. There were no unexpected relations between the variables.

As we can see from the outputs of the model, there is undoubtedly a strong linear correlation between (turnovers) TOV, (Total Rebounds) TRB, Pace, (Steals) STL, (Free Throw Attempts) FTA, ThreePPercentage, TwoPPercentage, and WINS. It makes logical sense because teams with more turnovers should win less games and teams with more Total Rebounds, Steals, Free Throw Attempts, as well as higher ThreePPercentage and TwoPPercentage are supposed to win more games. The interesting value was pace, where it was unclear from the beginning which direction it would be related (if related at all) to wins and it turned out that it was negatively related. That was one of the most interesting findings for the project. As aforementioned earlier, the relatively high R^2 values (0.9078), low Residual standard error (3.652) on 22 degrees of freedom, Multiple R-squared(0.9301), F-statistic (41.79), and significant F-statistics confirm the goodness of fit of the model. There, however, are some complications about the model that we would like to address. We would like to point out some scenarios where the model could potentially fail to work.

1) When teams make major changes in their roster during the season. Trades between teams can happen and since no two players can have the same skillset, it may negatively affect team play which means it can affect the rate at which the teams win.

2) Another reason for concern is injuries that can happen to players, especially the major ones. If, for example, Jordan gets an injury. His team, Chicago Bulls, might have lesser win ratios until he comes

back. Injuries can also have an enormous impact on team morale. These kinds of factors can't be accounted for in our model.

3) There is somewhat of a limit on how good of predictor this would be on wins for different seasons. This is especially true for seasons that took place a long time ago. For example, in the 60's there wasn't even an NBA three point line and steals (among other statistics) weren't recorded, so some variables would become useless.

Overall, while our model is good (and we had fun making it) there can be some interesting ways to improve or expand it in the future. One way would be the inclusion of more seasons, as this could build stronger results for the game of basketball as a whole. Another improvement could be more tests done to evaluate the ability of the model to predict the dependent variable. Lastly, deeper research on variables that should of been included but weren't could be informative.

#References

"2020-21 NBA Season Summary." Basketball-Reference.Com, https://www.basketball-reference.com/leagues/NBA_2021.html. Accessed 8 Dec. 2021.

Statista. 2003-2020. "Golden State Warriors franchise value from 2003 to 2020 (in millions U.S.dollars)." https://www.statista.com/statistics/194654/franchise-value-of-the-golden-state-warriors-of-the-nba-since-2006/