Measures of Players' Performance and Factors in College Basketball during the 2015 Season

Rohith Bachina, Liam Donovan, Molly Murtagh, Olivia Sousa, Pius Bonjui

Research Question:

Are the following variables conducive to success in college basketball?

Background

Our data set contained data from division I (NCAA) college basketball teams, over the years 2015-2019. Because there was data from five years, we decided to use data from one season, 2015. We chose the number of wins as the response variable. The options for the predictor variables are as follows:

Team
Conference
Games Played
Adjusted Offensive Efficiency
Adjusted Defensive Efficiency
Power Rating
Effective Field Goal Percentage Shot
Effective Field Goal Percentage Allowed
Turnover Percentage Allowed (Turnover Rate)
Turnover Percentage Committed
Offensive Rebound Rate
Offensive Rebound Rate Allowed
Free Throw Rate
Free Throw Rate Allowed
Two-Point Shooting Percentage
Two-Point Shooting Percentage Allowed
Three-Point Shooting Percentage
Three Point Shooting Percentage Allowed
Adjusted Tempo
Wins Above Bubble

Choice of Predictors

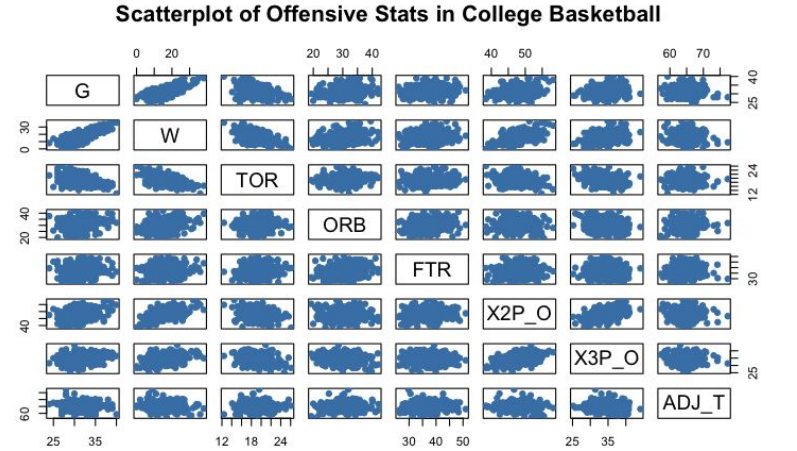
We elected to choose variables that we believed to be most important to the winning percentage of a given team. We then fit a model to check if our choice of variables was appropriate. We chose:

- Adjusted Offensive Efficiency (number of points scored per 100 possessions; measure of how well a team scores points); **ADJOE**
- Adjusted Defensive Efficiency (number of points allowed per 100 possessions; measure of how well a team defends); **ADJDE**
- Turnover Percentage Committed (measurement of the amount of turnovers per 100 possessions; measure of offensive proficiency); **TORD**
- Offensive Rebound Rate Allowed (percent of rebounds allowed; other team recovering possession after a missed shot); **DRB**
- Free Throw Rate Allowed (amount of other team's points are made up of free throws; a measurement of how often the team fouls); **FTRB**
- Two-Point Shooting Rate Allowed (percentage of two-point shots made against the team); **X2P_D**
- Three-Point Shooting Rate Allowed (percentage of three-point shots made against the team); **X3P_D**

Scatterplot and Correlation Matrix

- One moderately strong positive relationship that offense has with winning is Two-Point Shooting Percentage. This stat is fairly straight forward, and is simply calculated by taking the number of Two-Point Field Goals made divided by the total number of Two-Point Field Goals attempted. Hence the better a team shoots from the the inside the three point line the higher the chances a team has to win the game.
- One moderately strong negative relationship that offense has with winning is with Turnover Percentage Allowed aka Turnover rate. This stat is the estimated number of turnover a team commits per 100 plays and is calculated by (TOV ÷ (FGA + (0.44 x FTA) + TOV)) x 100%, with "**TOV**" is the number of turnovers allowed in a season."**FGA**" is the number of field goals attempted (do not include free throws). "**FTA**" is the number of free throws attempted. So naturally a team that has a higher percentage will have less potential possessions to score, hence negatively impacting there chances of winning

	G	W	TOR	ORB	FTR	X2P_O	X3P_O	ADJ_T
G	1.00	0.76	-0.37	0.26	0.13	0.32	0.24	-0.06
W	0.76	1.00	-0.56	0.32	0.20	0.59	0.42	-0.08
TOR	-0.37	-0.56	1.00	0.05	0.16	-0.31	-0.37	-0.02
ORB	0.26	0.32	0.05	1.00	0.21	-0.06	-0.21	0.06
FTR	0.13	0.20	0.16	0.21	1.00	0.12	-0.02	0.09
X2P_O	0.32	0.59	-0.31	-0.06	0.12	1.00	0.47	0.01
X3P_O	0.24	0.42	-0.37	-0.21	-0.02	0.47	1.00	-0.05
ADJ_T	-0.06	-0.08	-0.02	0.06	0.09	0.01	-0.05	1.00



Initial Model

Without any transformation, the following model was fit:

$$\hat{Y} = 0.58904(\mathbf{ADJE_o}) + 0.37834(\mathbf{ADJE_d}) + 1.35377(\mathbf{TORD}) - 0.64697(\mathbf{DRB}) - 0.19533(\mathbf{FTRD}) \\ -0.80899(\mathbf{X2P_D}) - 0.89094(\mathbf{X3P_D}) - 11.91397$$

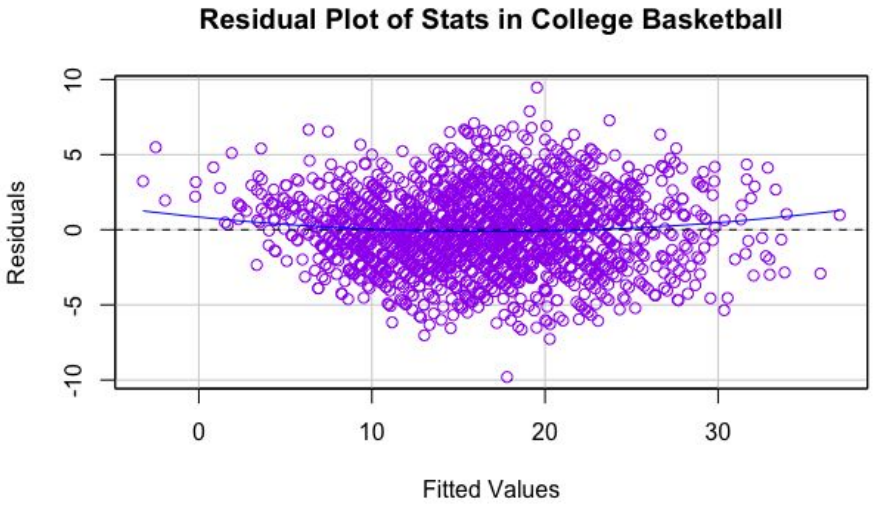
(*where ADJOE = Adjusted Offensive Efficiency, ADJDE = Adjusted Defensive Efficiency, TORD = Turnover Percentage Committed, DRB = Offensive Rebound Rate Allowed, FTRD = Free Throw Rate Allowed, X2P_D = Two-Point Shooting Percentage Allowed, X3P_O = Three-Point Shooting Percentage*)

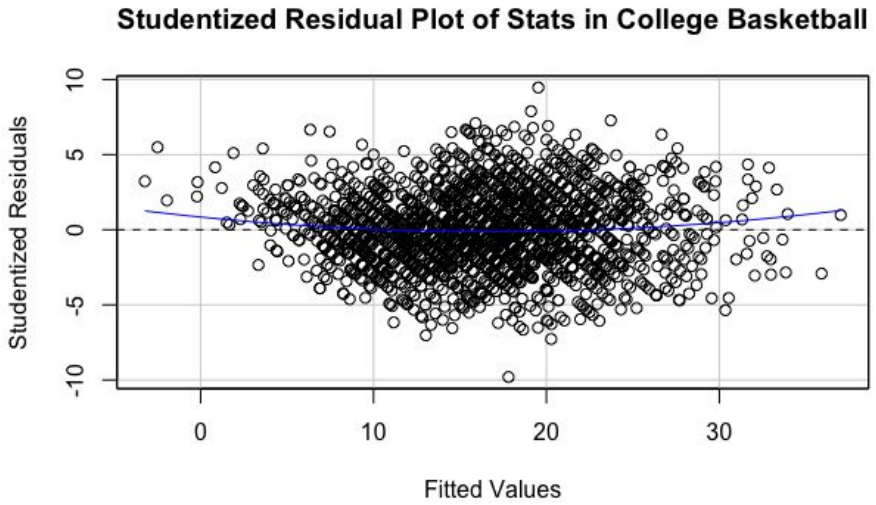
In order to see if the initial model is linear, we checked for four different assumptions within our data. We checked for independence, linearity of the regression function, constant variance, and normality of error terms.

Testing Independence Between Predictors

We elect to test the independence of the observation terms by using two methods.

First, we observe a scatter plot of the residual (error) terms as well as the studentized residuals.





Testing Independence Between Predictors Cont.

Since neither plot shows a significant trend among the residuals, we can be reasonably sure that the predictors are independent. We also can test by use of VIF:

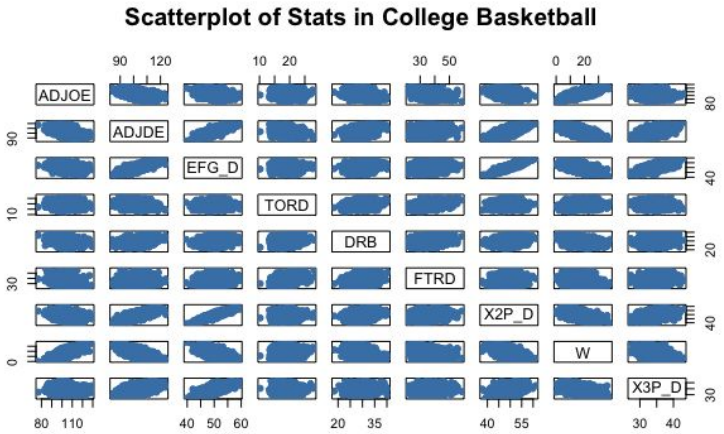
```{r}
vif(linmod)
```

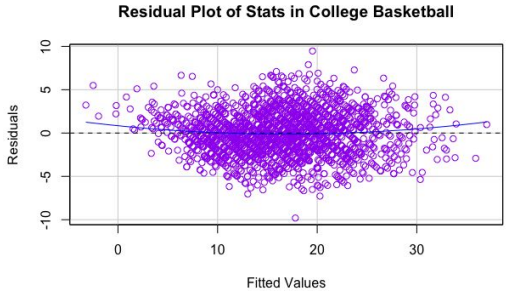
```
   ADJOE    ADJDE     TORD      DRB     FTRD    X2P_D    X3P_D
1.722997 5.986890 2.064020 1.544368 1.446517 2.821404 1.597051
```

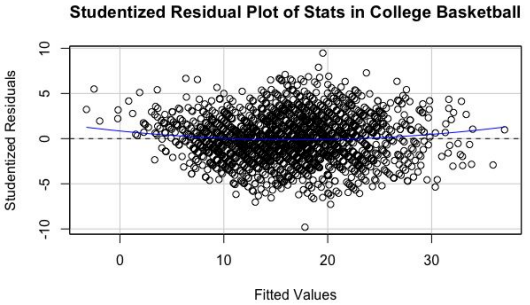
Since none of the VIF scores exceed 10, we can, again, be reasonably confident that the terms are independent.

Testing for Linearity of the Regression Function

To test for linearity, we checked to see if the relationship between the predictors and wins were linear. Hence, we checked the scatter plots and residual plots for linearity.

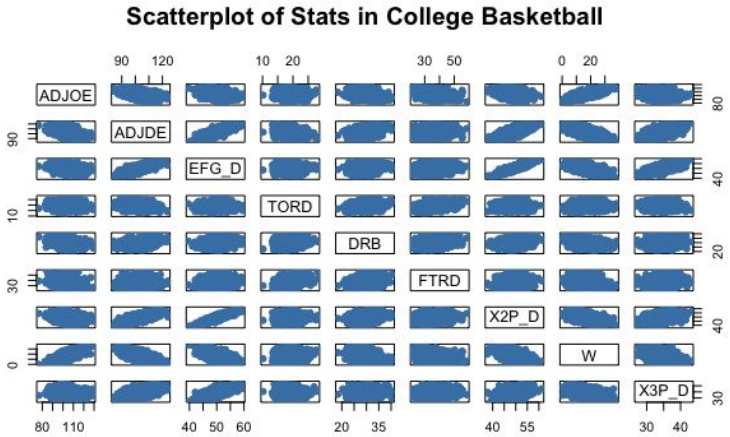


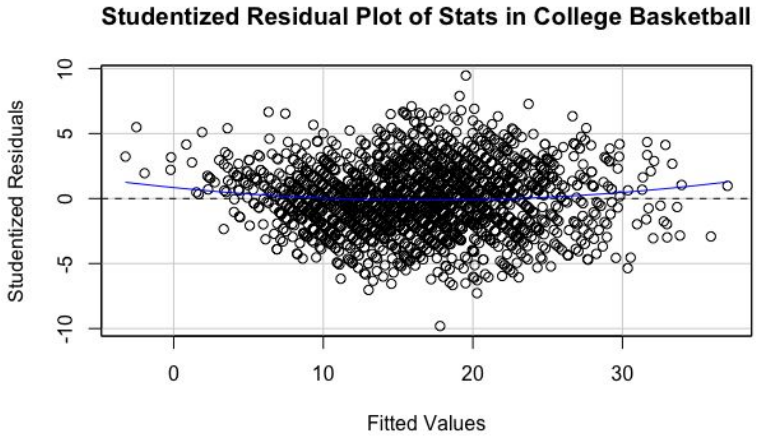




Testing for Linearity of the Regression Function Cont.

The scatter plot display linearity for each of the corresponding predictors. In addition, the residual plot contains scatter data points above and below the 0 residual line. Therefore, they pass the linear model assumption for the linearity of the regression function.





Testing for Constant Variance

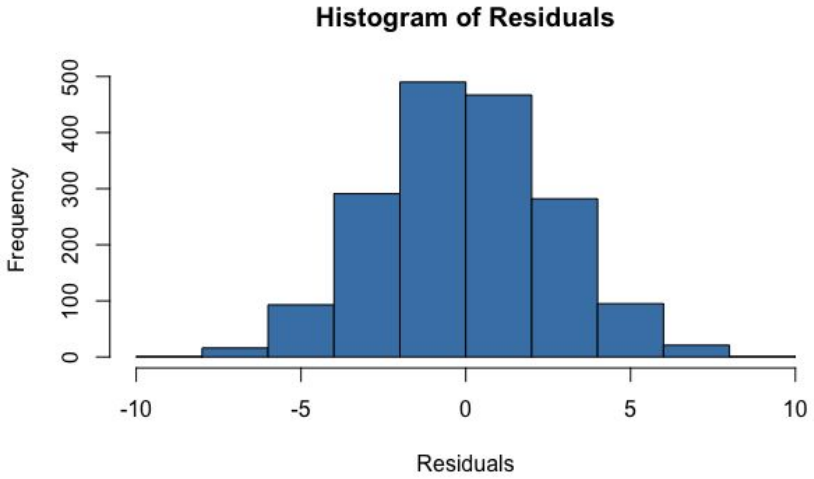
We checked for homoscedasticity by utilizing the Levene's Test. The test yielded a p-value of 0.88. This p-value is a significantly larger value than the alpha, hence, similar to the residual plots, we can conclude that the variances are approximately equal.

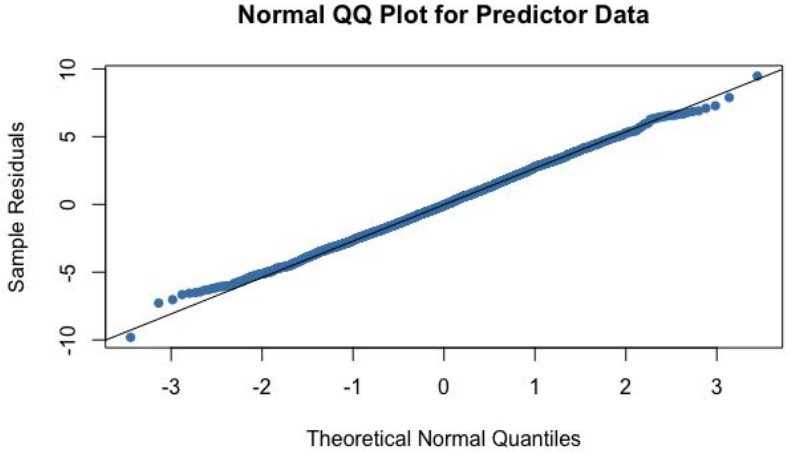
```r
leveneTest(modified_filtered_data_offense$W, modified_filtered_data_offense$ADJOE +
modified_filtered_data_offense$ADJDE + modified_filtered_data_offense$TORD +
modified_filtered_data_offense$DRB + modified_filtered_data_offense$FTRD +
modified_filtered_data_offense$X2P_D + modified_filtered_data_offense$X3P_D)
```

```
Warning: modified_filtered_data_offense$ADJOE + modified_filtered_data_offense$ADJDE +
modified_filtered_data_offense$TORD + modified_filtered_data_offense$DRB +
modified_filtered_data_offense$FTRD + modified_filtered_data_offense$X2P_D +
modified_filtered_data_offense$X3P_D coerced to factor.Levene's Test for Homogeneity of
Variance (center = median)
       Df F value Pr(>F)
group  502  0.9143 0.8813
      1254
```

Testing for Normality

We created histogram of the residuals as well as a normality probability plot, QQ plot, from the scatter plot. After creating these two graphs, we checked to see if the histogram displayed a normal distribution and compared if the quantile points appear to fall in a straight line.





Testing for Normality Cont.

We also can make use of the Shapiro-Wilk Test, to check normality:

```r
shapiro.test(linmod$residuals)
```

```
        Shapiro-Wilk normality test

data:  linmod$residuals
W = 0.99892, p-value = 0.3685
```

Since W is approximately 1 and the p-value is significantly larger than the alpha value (.05), we can conclude that the chosen data is about normal.

Plausible Transformations and Model Fitting

Because the assumptions are so well fit and contain normal distributions, we found the initial model to be sufficient to answer the research question. Therefore, no transformations were selected. As a result, there are no polynomial terms. Also, due to the absence of qualitative variables, use of interaction terms are also not needed.

Model Utility

We use a t-test to assess the utility of the model. The null hypothesis is that the beta variables are equal to 0, thus making the utility of the model low.

```{r}
SSE <- sum((fitted(linmod)-modified_filtered_data_offense$W)^2)
SSR <- sum((fitted(linmod)-mean(modified_filtered_data_offense$W))^2)
#compute df for MSE and MSR
n <- nrow(modified_filtered_data_offense)
p <- ncol(modified_filtered_data_offense)
df1 <- p-1
df2 <- n-p
#compute MSR, MSE
MSR <- SSR/df1
MSE <- SSE/df2
MSE
MSR
#test stat
#alpha
alpha <- .05
f.test <- MSR/MSE
f.test
```

```
#critical value
f.crit <- qf(1-alpha, df1, df2)
f.crit
#test if f.test > f.crit if so
#conclude alternate
if(f.test > f.crit){
  print("conclude alternate")
}else{
  print("conclude null")
}
```

```
[1] 6.930261
[1] 9016.273
[1] 1301
[1] 2.014806
[1] "conclude alternate"
```

Model Utility Cont.

Thus, since the test concluded the alternate hypothesis, we find that the utility of the model is high.

Confidence and Prediction Intervals

Let alpha = 0.05, therefore, we will create a 95% confidence interval.

```
#confidence interval
alpha <- .05
confint(linmod, level = 1-alpha)
```

```
##                   2.5 %     97.5 %
## (Intercept) -16.6662233 -7.1617148
## ADJOE         0.5668969  0.6111776
## ADJDE         0.3317650  0.4249204
## TORD          1.2698313  1.4377033
## DRB          -0.6969838 -0.5969626
## FTRD         -0.2204421 -0.1702156
## X2P_D        -0.8719283 -0.7460482
## X3P_D        -0.9566451 -0.8252279
```

Answering the Research Question

Due to the model’s strong positive linear relationship, it was found that our predicted variables were very conducive to winning, for college basketball.