Does three point shooting determine NCAA Tournament game outcomes more than any other factor?

Jon Michael Stroh

2022-03-16

Data Exploration (2021 NCAA Tournament)

New Variables

More3s Frequency of the winning team making more 3 point shots than losing team:

```
## # A tibble: 3 x 3
##   more3s   cnt percent
##   <fct>  <int>   <dbl>
## 1 Winner    42    63.6
## 2 Loser     15    22.7
## 3 Tie        9    13.6
```

In the 2021 tournament, the winning team made more three pointers in 63.6% (42/66) of games, the teams tied in 13.6%, and the losing team made more three pointers in 22.7%.

Better3s Frequency of the winning team out shooting the losing team on 3 point shooting percentage:

```
## # A tibble: 2 x 3
##   better3s   cnt percent
##   <fct>    <int>   <dbl>
## 1 Winner      51    77.3
## 2 Loser       15    22.7
```

In the 2021 tournament, the winning team made a higher percentage of three pointers in 77.3% (51/66) of games. Wow, that's fairly significant. Let's check both together.

More & Better 3s

```
## # A tibble: 6 x 4
## # Groups:   better3s [2]
##   better3s more3s count percent
##   <fct>    <fct>  <int>   <dbl>
## 1 Winner   Winner    38   57.6
## 2 Loser    Loser      9   13.6
## 3 Winner   Tie        7   10.6
## 4 Winner   Loser      6    9.09
## 5 Loser    Winner     4    6.06
## 6 Loser    Tie        2    3.03
```

In 2021, in only 13.6% of games did the losing team make more 3s and a higher percentage of them. In contrast, in 68.2% (45/66) of games did the winning team either make more 3s and shoot a higher percentage or tie on the number of makes and shoot a higher percentage.

Alright, so in fairly conclusive fashion, we have determined that teams that shoot better on threes win games at a much higher percentage in this NCAA tournament. Now, we must contrast it with winning percentage of teams that shoot better on 2s, or get and make more free throws, or grab more rebounds, etc.

More 2s First, we will look at the frequency of winning teams making more 2 point shots.

```
## # A tibble: 3 x 3
##   more2s   cnt percent
##   <fct>  <int>   <dbl>
## 1 Winner    42   63.6
## 2 Loser     20   30.3
## 3 Tie        4    6.06
```

The winning teams made more 2 point shots at exactly the same frequency as three point shots (~64%). Interesting to note that losing teams made more 2 point shots in 30% of game compared to 23% for threes.

It is important to remember, both of these frequencies (more2s and more3s) are measuring make totals which is heavily influenced by the total number of attempts. Total number of attempts is further influenced by other factors such as offensive rebounds and turnovers, etc. We can investigate these later.

Better 2s Frequency of winning team out shooting losing team on 2 point shot percentage:

```
## # A tibble: 3 x 3
##   better2s   cnt percent
##   <fct>    <int>   <dbl>
## 1 Winner      44   66.7
## 2 Loser       21   31.8
## 3 Tie          1    1.52
```

WOW... in only 67% of NCAA tourney games did the winning team shoot a higher percentage on 2s. This was 77% for three point percentage.

moreFgs Just a quick glimpse at percent of games winning teams made more field goals in general:

```
## # A tibble: 3 x 3
##   morefgs   cnt percent
##   <fct>   <int>   <dbl>
## 1 Winner     56   84.8
## 2 Loser       7   10.6
## 3 Tie         3    4.54
```

Well haha, this should make sense. Scoring points is in fact how one wins the game. Actually, ~85% seems almost lowish. The only other option to win is by making more 3s as a portion of those field goals or more free throws or both. There were 3 ties.

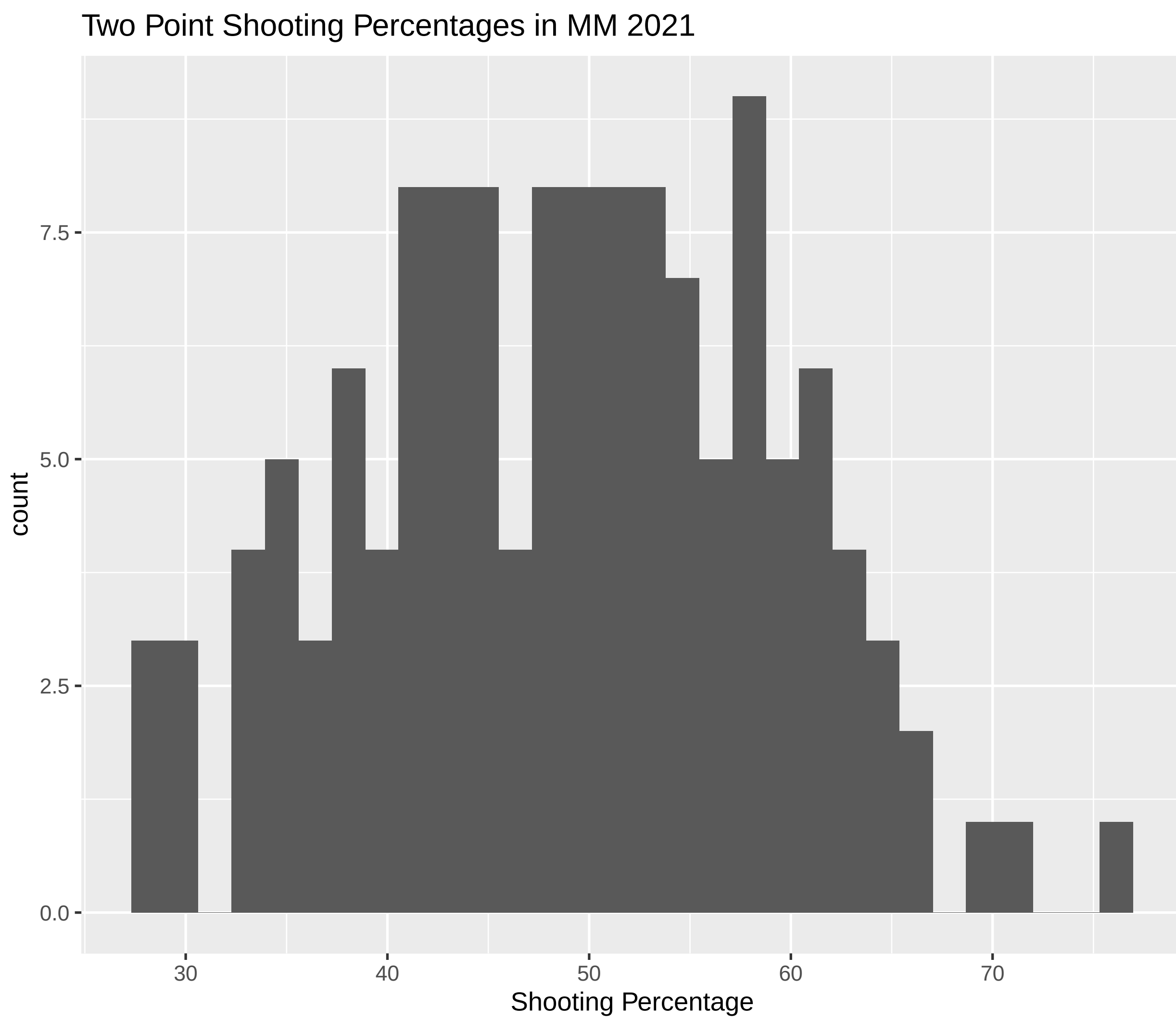
betterfgs We'll finely look at overall shooting percent as it relates to winners and losers:

```
## # A tibble: 3 x 3
##   betterfgs   cnt percent
##   <fct>     <int>   <dbl>
## 1 Winner       47   71.2
## 2 Loser        18   27.3
## 3 Tie           1    1.52
```

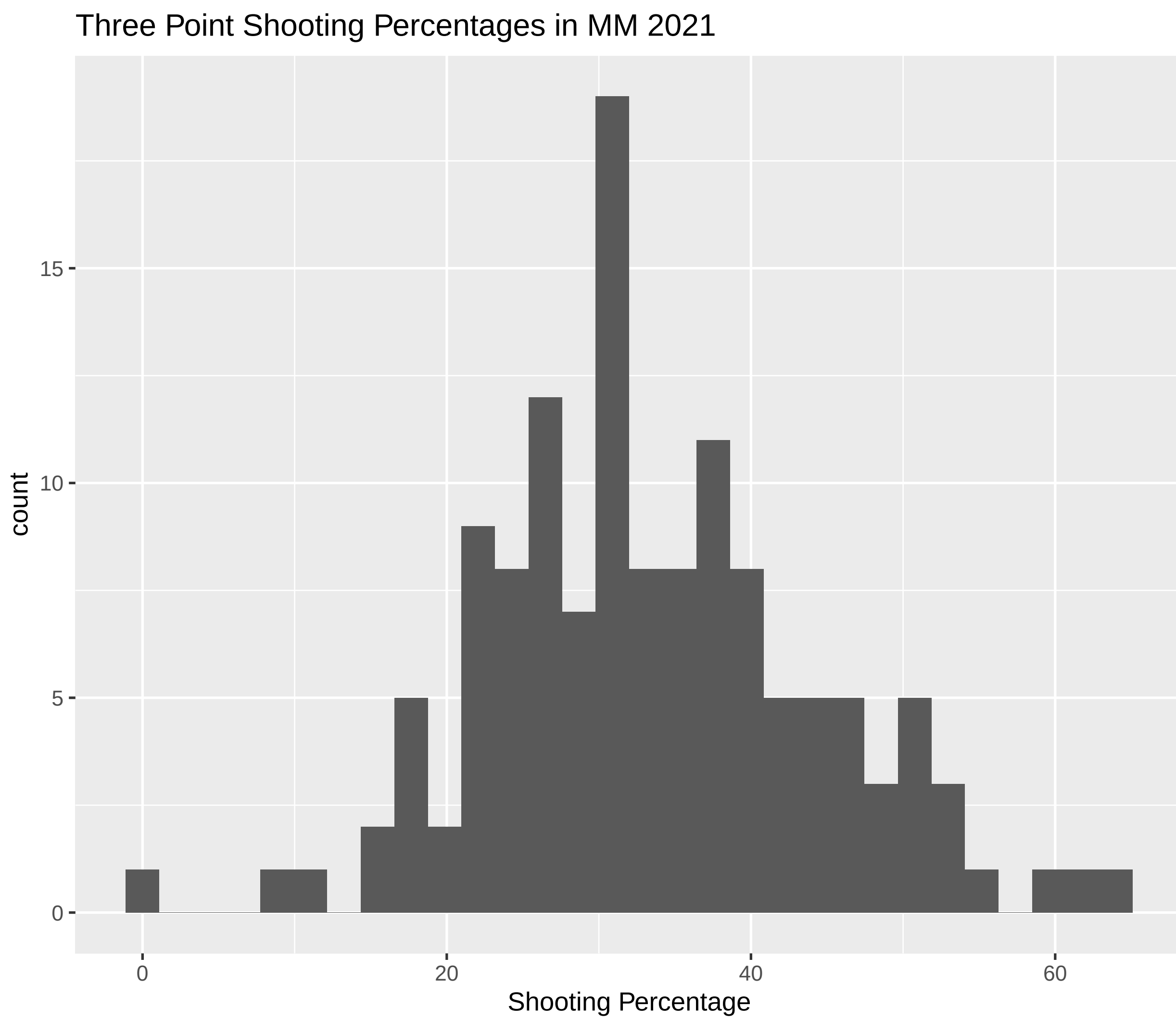
However, interesting enough, in only 71.2% of games (47/66) did the winning team shoot a higher percentage on all field goals. This just reaffirms that clearly other factors can influence a win, the likely ones for investigation is free throw shooting (although there tends to be a small difference in this between teams in a single game), three point shooting (our variable of primary interests), and factors that can get a team more shots (offensive rebounding and forcing turnovers while limiting those thingsagainst and for themselves).

Distribution of Shooting

Two point shooting twoPercent is the two point shooting percentage by team as a percentage, rather than a decimal, for graphing.



Three point shooting threePercent is the three point shooting percentage by team as a percentage, rather than a decimal, for graphing.



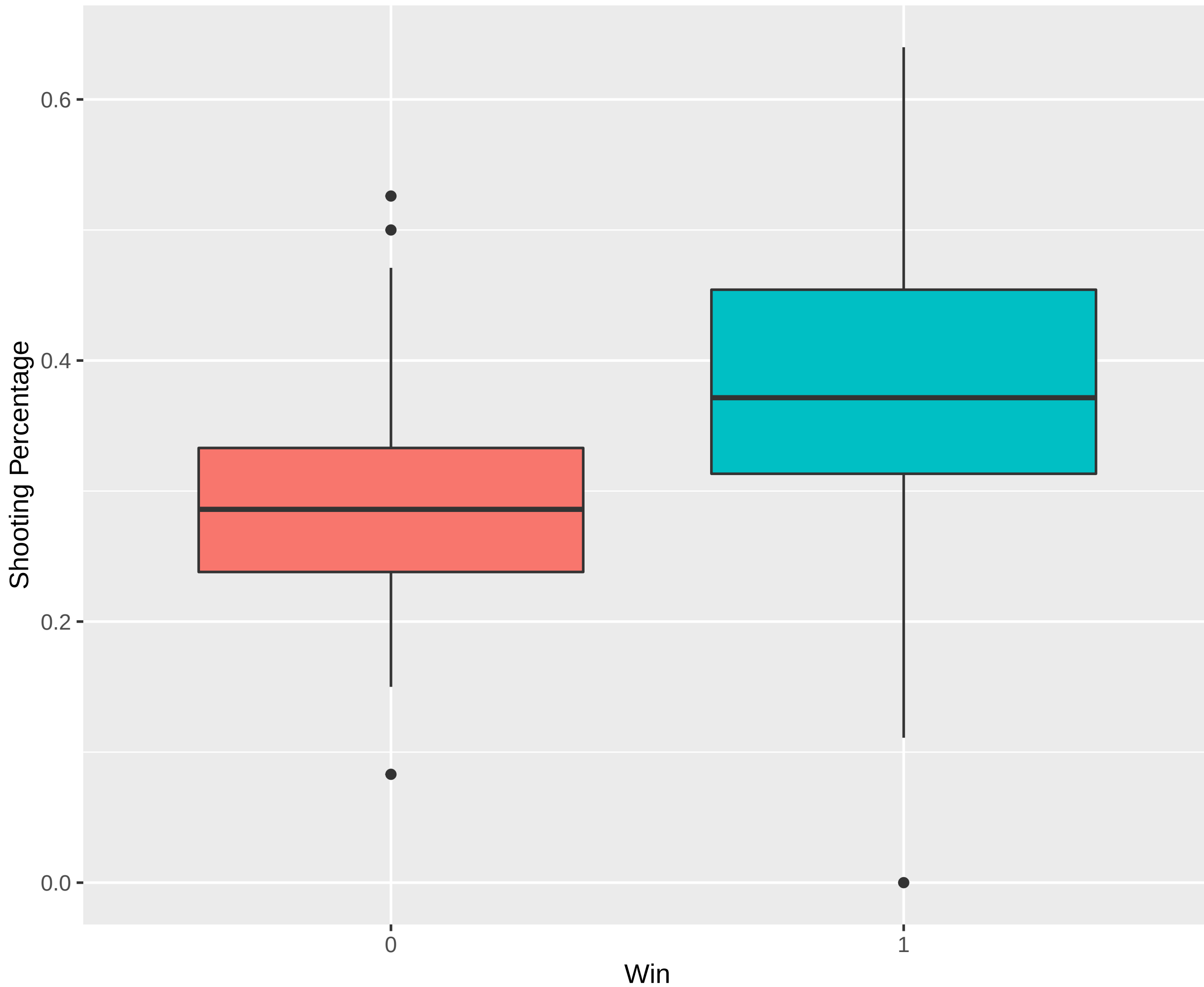
Comparing Shooting *Note: twoPercent_mean is not the shooting percent for all two point shots in the march madness tournament. It is the mean shooting two point shooting percentage for each team each game. Overall, it can not represent the percentage for total two point shooting percent because it is not adjusted for attempts by game. The same thing goes for threePercent_mean*

```
## # A tibble: 1 x 4
##   twoPercent_Mean threePercent_Mean twoPercent_SD threePercent_SD
##             <dbl>             <dbl>         <dbl>           <dbl>
## 1            49.0              33.4          10.1            11.0
```

I hypothesized that there much exist a larger standard deviation in three point shooting by game throughout the tourney. That is, three point shooting performance varies more game by game. In this tournament, this did not occur. The standard deviations are nearly identical (less than a percentage point greater for three point shooting).

Three Point Percentage by Outcome

Three Point Shooting Percentages by Outcome in MM 2021




```
##   Win    3P%
## 1   0 0.2860
## 2   1 0.3715
```

The median three point percentage for winning teams is 37.2% compared to only 28.6% for losing teams.

Based on data exploration for all 66 games in the 2021 NCAA tournament, there seems to be a strong association between a team's three point shooting percentage for a single game and the probability that they win that game. Thus, we intend to explore models with three point percentage as a predictor variable for whether a team wins an NCAA tournament game or not.

Model Building

According to a 2004 study, 4 factors were indicated to most affect the outcome of basketball games: "shooting efficiency, number of turnovers, offensive rebounds and free throws made" (Oliver). Thus, I will fit a logistic regression model that includes two and three point percentage, total turnovers, total offensive rebounds, and total free throws attempted as predictor variables for whether or not a team wins a game. Additionally, through my own observation of college basketball, assisted shots tend to be better shots, and thus more likely to go in. Therefore, I will add total assists to my model as a predictor. Finally, teams that generate many steals tend to generate easy, fast break scoring opportunities. Also, teams that limit fouls often limit the amount of times a team goes to the free throw line. Thus, I will add total steals and total fouls to the model. Finally, I want to mean center all the variables in the model so that the intercept can be interpreted.

term	estimate	std.error	statistic	p.value
(Intercept)	0.014	0.243	0.058	0.954
twoPercentCent	0.110	0.034	3.261	0.001
threePercentCent	0.100	0.030	3.367	0.001
ORBCent	0.094	0.075	1.242	0.214
ASTCent	-0.019	0.076	-0.247	0.805
TOVCent	-0.218	0.073	-3.005	0.003
STLCent	0.260	0.098	2.659	0.008
PFCent	-0.230	0.082	-2.808	0.005
FTACent	0.135	0.046	2.922	0.003

According to the model, the predictor variables with statistical significance for predicting whether a team won or loss is two point percentage, three point percentage, total turnovers, total steals, total fouls, and free throws attempted.

According to the model, the odds of a team winning the game with the mean predictor values – 48.9666667% two point shooting, 33.4469697% three point shooting, 8.469697 offensive rebounds, 13.7575758 assists, 10.3560606 turnovers, 5.3409091 steals, and 15.9621212 fouls – is 1.014. This value makes sense because a team with the mean statistics in this value has an almost 1 to 1 odds of winning a game or roughly a 50% win percentage.

According to the model, for every 1 percentage point increase in two point shooting percentage the odds the team wins the game is multiplied by a factor of 1.116, holding all else constant. Similarly, according to the model, for every 1 percentage point increase in three point shooting percentage the odds the team wins the game is multiplied by a factor of 1.105, holding all else constant. Thus, the coefficients for both two and three point shooting percentages are close, but the factor in which the odds of winning are multiplied by are slightly larger for two point shooting than three.

Now, to fit a better model, I will conduct backwards model selection using AIC as well as a drop in deviance test to attempt to find the strongest model:

Backwards Selection

term	estimate	std.error	statistic	p.value
(Intercept)	0.021	0.241	0.088	0.930
twoPercentCent	0.095	0.027	3.479	0.001
threePercentCent	0.090	0.024	3.713	0.000
TOVCent	-0.214	0.071	-3.021	0.003
STLCent	0.277	0.095	2.907	0.004
PFCent	-0.229	0.080	-2.867	0.004
FTACent	0.138	0.045	3.060	0.002

Drop-in-deviance test Hypotheses for Drop-in-deviance test:

Null Hypothesis: $H_0 : \beta_{ORBCent} = \beta_{ASTCent} = 0$ (These variables don't add information to the model after accounting for two & three point shooting percentage, steals, turnovers, fouls, and free throws attempted)

Alternative Hypothesis: $H_a :$ at least one β_j is not equat to 0

```
## # A tibble: 2 x 5
##   Resid..Df Resid..Dev    df Deviance p.value
##       <dbl>      <dbl> <dbl>    <dbl>   <dbl>
## 1       125       111.    NA    NA     NA
## 2       123       109.     2     1.55   0.461
```

Because the p-value is much larger than 0.05, we fail to reject the null hypothesis. That data does not provide sufficient evidence that the coefficients for ORBCent and ASTCent are different from 0. The best model includes only two & three point shooting, turnovers, steals, fouls, and free throw attempts.

Model Interpretation

term	estimate	std.error	statistic	p.value
(Intercept)	0.021	0.241	0.088	0.930
twoPercentCent	0.095	0.027	3.479	0.001
threePercentCent	0.090	0.024	3.713	0.000
TOVCent	-0.214	0.071	-3.021	0.003
STLCent	0.277	0.095	2.907	0.004
PFCent	-0.229	0.080	-2.867	0.004
FTACent	0.138	0.045	3.060	0.002

Again, according to the model, the factor in which the odds of winning are multiplied by for each 1 percentage point increase in shooting percentages of twos and threes is still greater for an increase in two point percentage (1.1 versus 1.094). Although, the coefficients are very close.

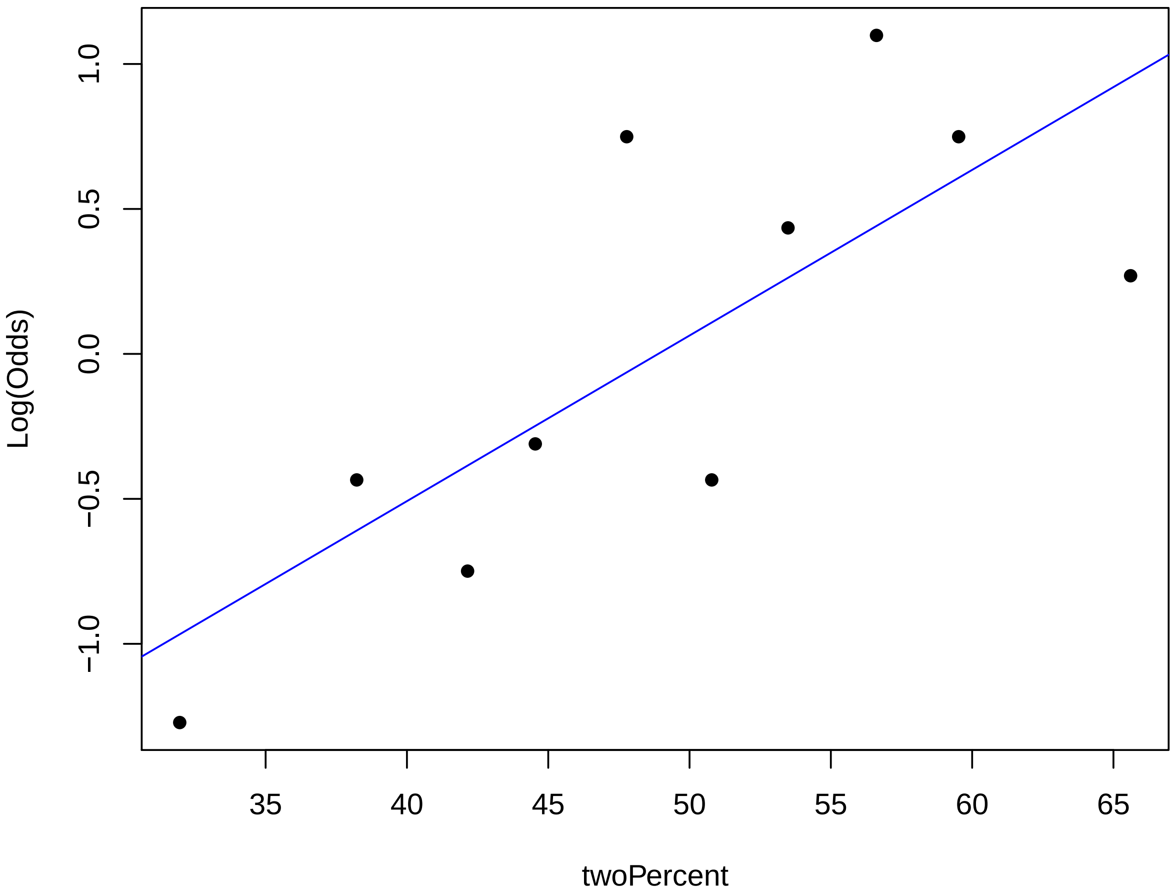
To make predictions for team's game statics, we can refit the model without mean centering the variables.

term	estimate	std.error	statistic	p.value
(Intercept)	-5.473	1.982	-2.761	0.006
twoPercent	0.095	0.027	3.479	0.001
threePercent	0.090	0.024	3.713	0.000
TOV	-0.214	0.071	-3.021	0.003
STL	0.277	0.095	2.907	0.004
PF	-0.229	0.080	-2.867	0.004
FTA	0.138	0.045	3.060	0.002

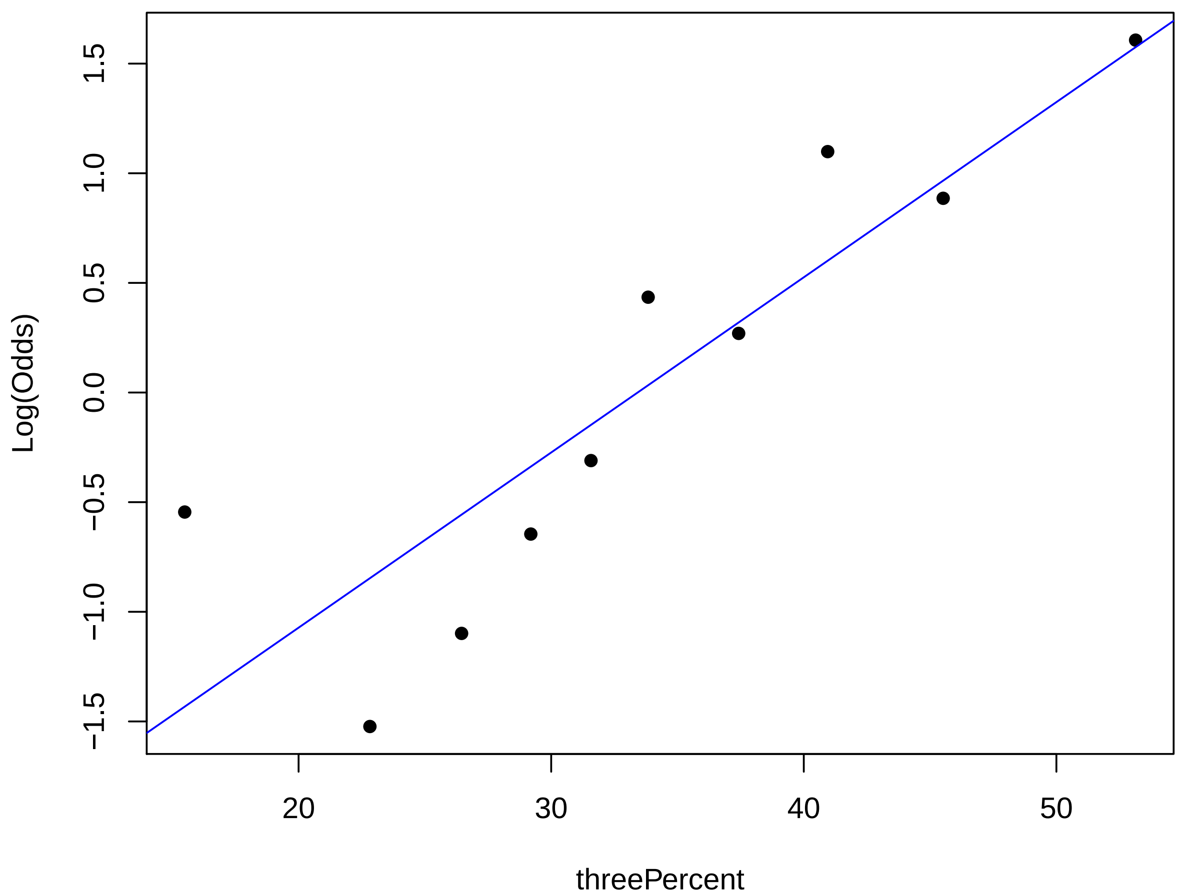
For example, in the first round of the NCAA tournament, 1 seed Baylor beat 16 seed Hartford 79 to 55. According to the model, the odds that Baylor would win the game given its game statistics are expected to be 5.912 ($exp^{-5.473+0.095(47.6)+0.090(33.3)-0.214(10)+0.277(15)-0.229(16)+0.138(10)}$). This indicates that given its game statistics, Baylor is expected to win 5.912 games for every one loss. Similarly, given the same formula, according to the model, the odds that Hartford would win the game given its game statistics are expected to be 0.012. Thus, Hartford would be expected to win 0.012 games for every loss, or in a better interpretation, 1 win the team is expected to lose roughly 83.3 games. This may seem extreme but Hartford did have a very poor performance (24 turnovers!), and so it is unlikely a team will win many games with 24 turnovers and such poor shooting.

Model Conditions

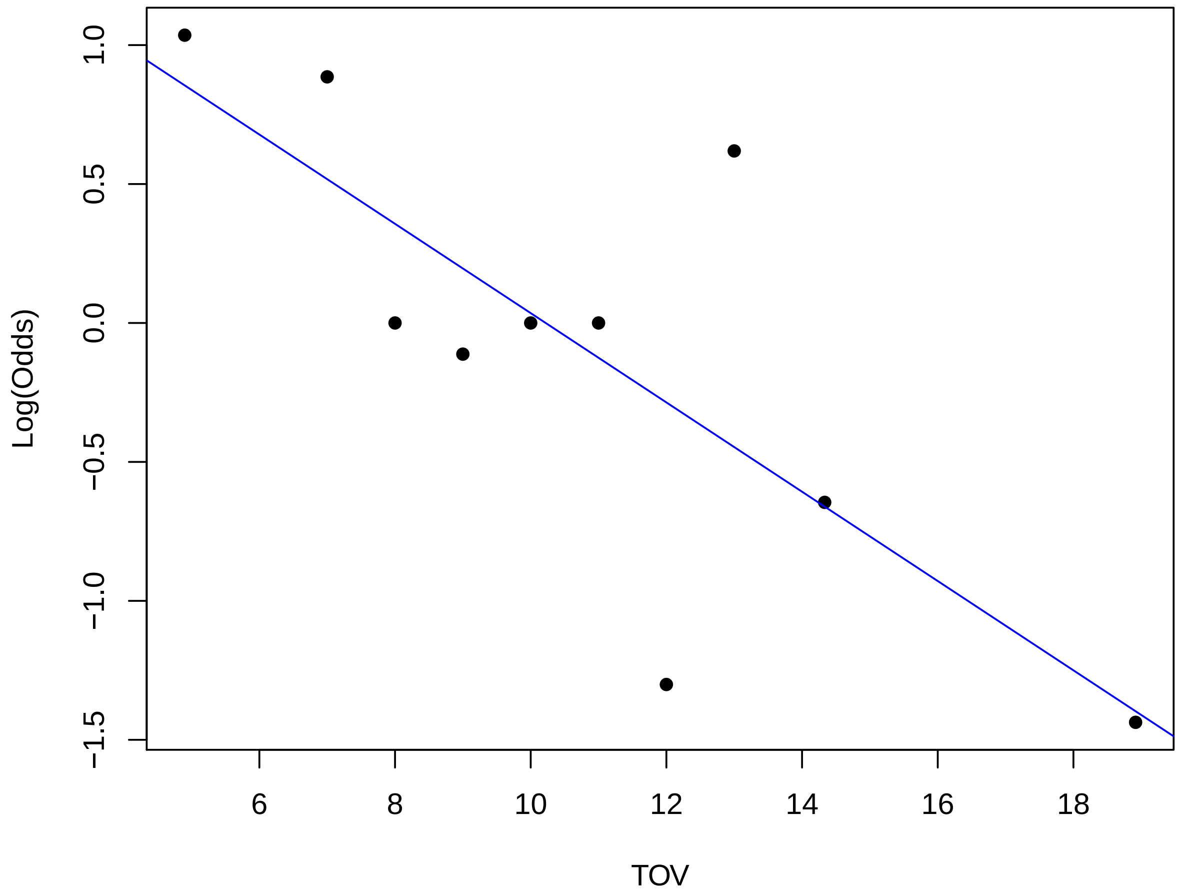
Linearity To check if linearity is satisfied, we plot the predictor variable against the empirical logit and are looking for a linear relationship.



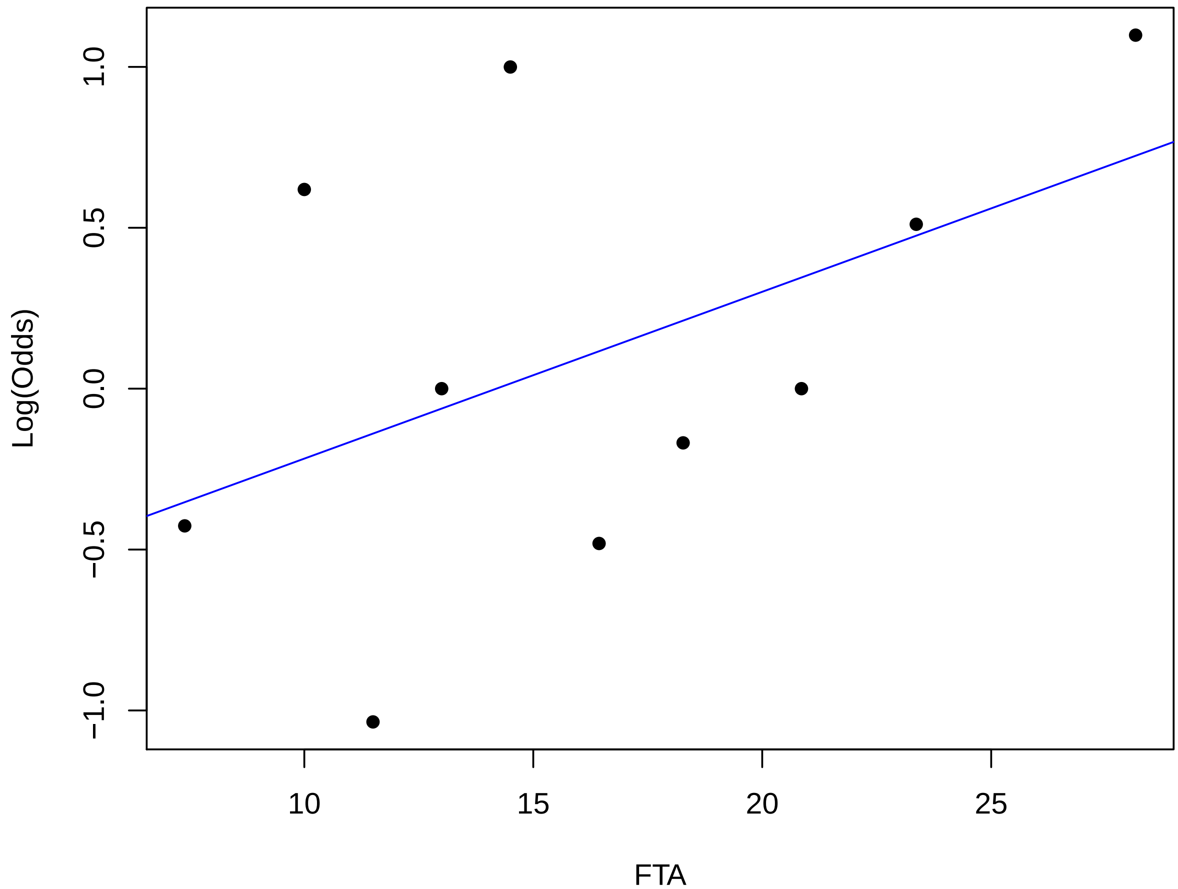
Log(Odds)
1.0
0.5
0.0
−0.5
−1.0
35
40
45
50
55
60
65
twoPercent



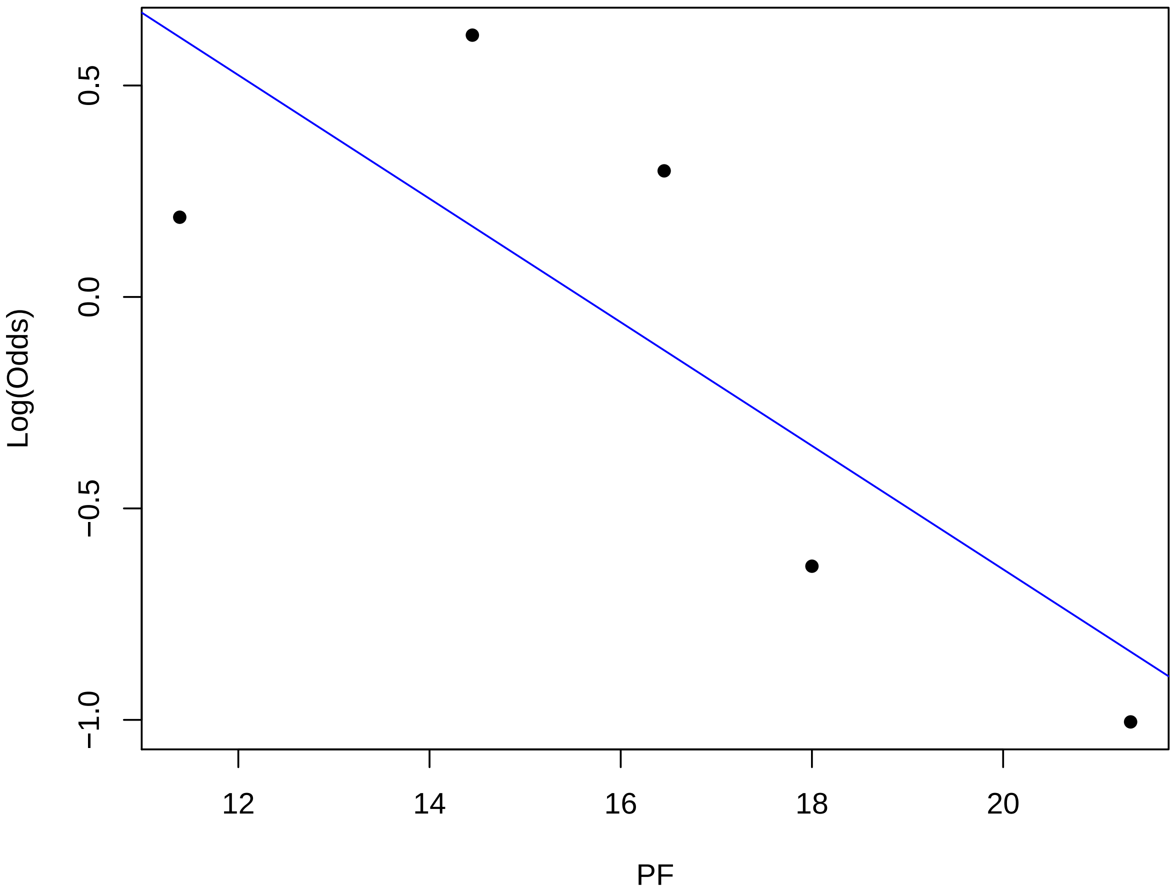
Log(Odds)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
20
30
40
50
threePercent

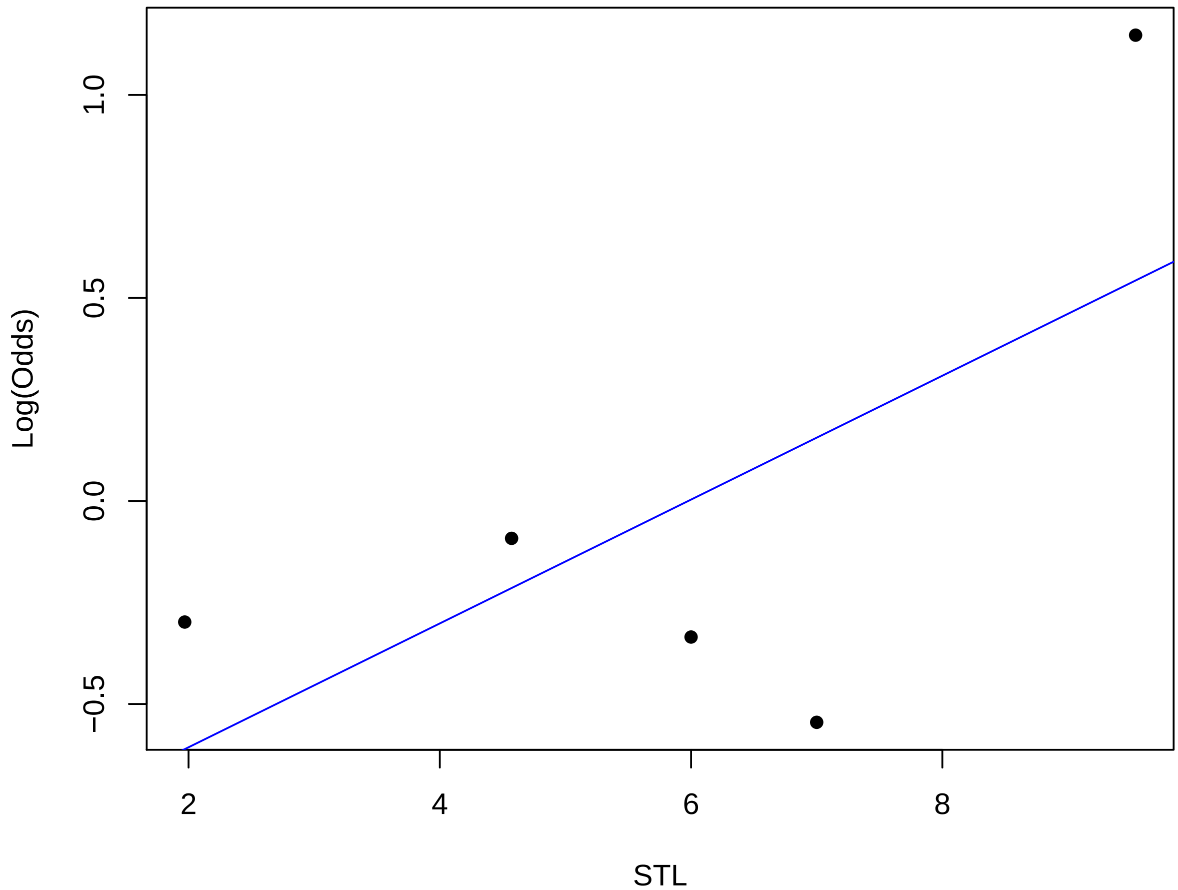


1.0
0.5
0.0
−0.5
−1.0
−1.5
Log(Odds)
6
8
10
12
14
16
18
TOV



Log(Odds)
1.0
0.5
0.0
−0.5
−1.0
10
15
20
25
FTA


0.5
0.0
−0.5
−1.0
Log(Odds)
12
14
16
18
20
PF



Based on the empirical logit plots, the linearity condition is met for each predictor variable.

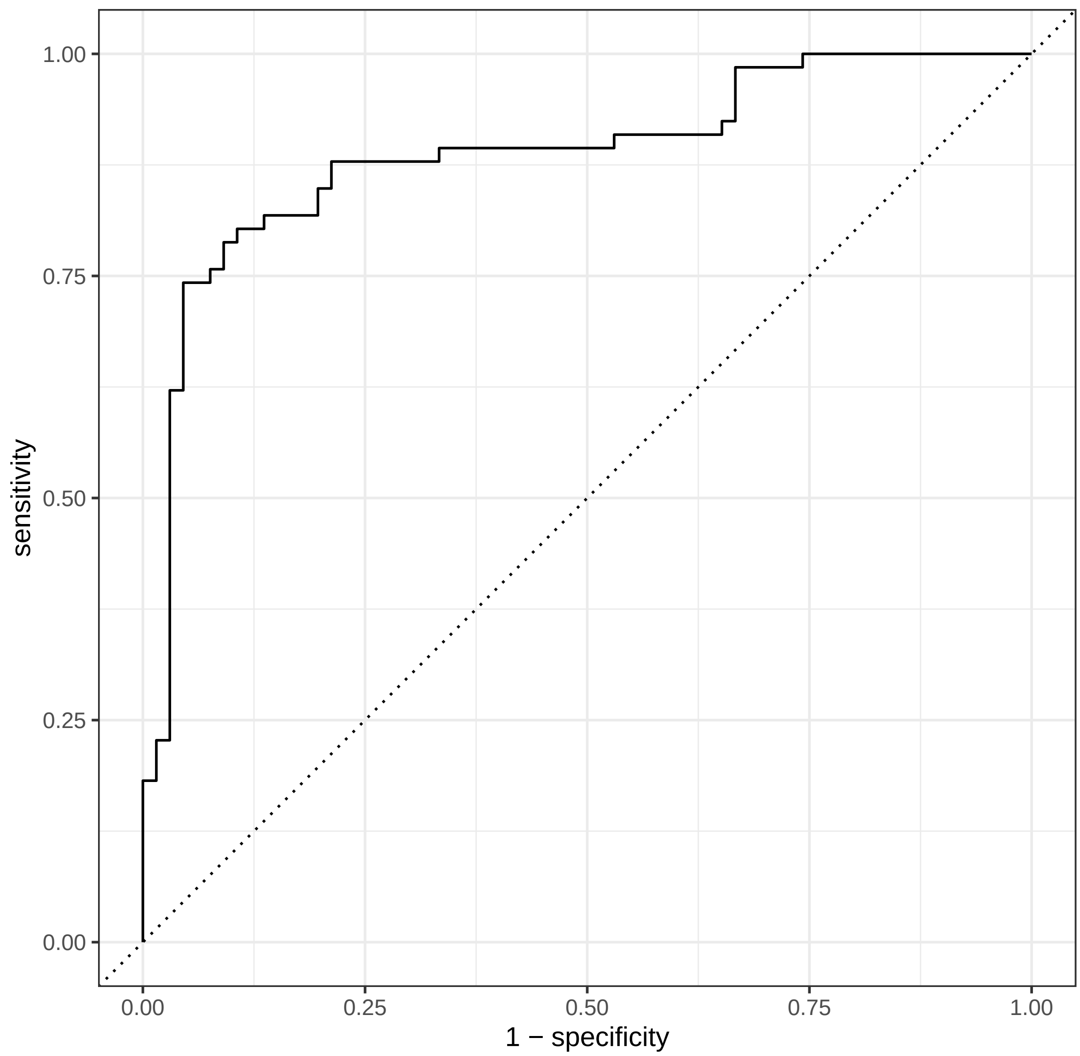
Randomness The randomness condition is murky. Each individual basketball is game is a random event. However, teams are inherently unequal, and therefore each game is not a 50-50 random event.

Indepedence Again, murky. The outcome of one game does not directly affect the outcome of another game. However, it does effect the teams that play, with better teams typically advancing.

Assessing Model Fit

```
## # A tibble: 4 x 3
##   Win   pred_Win           n
##   <fct> <chr>          <int>
## 1 0     Predicted Loss    55
## 2 0     Predicted Win     11
## 3 1     Predicted Loss    12
## 4 1     Predicted Win     54
```

According to the confusion matrix for the model with a threshold of 0.55, the misclassification rate is 17.4%, with a sensitivity of 81.8% and a specificity of 83.3%. The threshold 0.55 was observed as the best to minimize misclassification as well as relatively equal sensativity and specificity.




```
## [1] 0.8872819
```

According to the ROC Curve and the AUC of 0.887, the model is a good fit for the data. Considering there are only six predictor variables (the model being fairly simple), AUC values closer to 1 and farther from 0.5 indicate better model fit for the data.

Benefits and Drawbacks to this Model Certain benefits to this model include: its applicability to predicting games in the NCAA tournament. To predict future NCAA tournament games, one can apply the average statistics for a team to predict whether or not they will win the game. Additionally, its accuracy is based upon the previous NCAA tournament and what current trends/developments can be more applicable than the previous tournament. Furthermore, the model is very simple.

However, the drawbacks: first, the 20% error rate. But, more importantly, the quantitative statistcs other than shooting percentages – turnovers, steals, fouls, and free throw attempted – require pace adjusting that the model can not take into account. Teams that play at a faster pace will have a greater total number of raw statistics– shots, rebounds, and yes, turnovers, steals, fouls, and free throw attempted. Thus, it is important if making predictions to take into account team pace.

What this model does not successfully do, which was the purpose of this study, is indicate whether three point shooting has a greater affect than any other factor in determining the outcome of NCAA tournament games. However, assigning value to different factors in the importance for wins is difficult. This model did indicate that of the four factors, shooting percentages, offensive rebounds, turnovers, and free throws attempted, offensive rebounds was not statistically significant and left out of the model.