Making Informed Bracket Decisions During March Madness: Examining the Relationship Between Adjusted Efficiency Margins and Performance in the NCAA Men's Basketball Tournament

Josh Kraus

Study

The NCAA Men's Basketball Tournament, known as "March Madness", is a single-elimination tournament consisting of 64 of the best college men's basketball teams in the country. Its popularity and "madness" come from trying to correctly predict the outcomes of each game, an extremely difficult task which involves accurately selecting the participants and winner of 63 total games. The total number of possible brackets are 2^{63}, or 9.2 quintillion, and these highly improbable odds show why no one has ever predicted a perfect bracket.

There are a number of sports analysts who have tried to increase these odds of selecting a perfect bracket by offering advanced analytics which describe the performance of any given team throughout the season. Most notably, Ken Pomeroy provides men's college basketball ratings on his website KenPom which use advanced statistical methods (read about how he determines adjusted efficiencies here) to depict the how efficient a team has played throughout the year. His KenPom ratings are often referred to as some of most accurate college basketball ratings available.

Understanding the relationship between how efficient a team has been throughout the year and what round they advanced to in the NCAA tournament could be beneficial for making informed decisions when filling out a bracket. If advanced metrics like KenPom ratings provide a good measure of what round a team can be expected to advance to in a tournament, then using these results could improve one's chances of correctly filling out a bracket.

The purpose of this study was to use KenPom ratings to predict how far a team should be expected to advance in the NCAA Men's Basketball Tournament. The main research question was:

- How can adjusted efficiency margins be used to predict how far a team should be expected to advance in the NCAA Men's Basketball Tournament?

Dataset

The dataset that was used for this project contained data from every team that has competed in the NCAA Men's Basketball Tournament from 2002 to 2021, excluding the 2020 tournament which was cancelled due to COVID-19. This data was collected through a web scraper designed for KenPom's website and a website titled "Sports Reference: College Basketball" (see code here). 2002 was chosen as a starting point for this range because it was the first year that data was available on KenPom's website. Most other sports analytics websites such as "Sports Reference" only offer ratings data from 2011 onward, which is why KenPom was chosen for this project due to its thoroughness. This data represented 52.7% of the NCAA tournament data available (19 of 36 years) since the NCAA transitioned to a 64-team men's tournament in 1985.

The sample for this study was each observation of a team, categorized by the round they advanced to in the tournament. Since the focus of this study was on what round a team with a certain adjusted efficiency margin could be expected to advance to in the tournament, the rounds in question were Round 2 through the Champion. The reason the 1st Round was not included was because it was fundamentally different from the other rounds, since if the highest round a team advanced to was the 1st Round this meant that they did not advance at all in the tournament. Since 32 teams made it to the 2nd Round over the 19 years of data collected, this meant there were 608 different observations of teams that constituted the sample of this study. Each observation contained the year a team competed in the tournament, the round the team advanced to in the tournament, as well as their adjusted efficiency margin.

This final variable, adjusted efficiency margin, (or E.M., for short) was retrieved from KenPom's ratings. It was calculated by subtracting a team's adjusted offensive efficiency minus their adjusted defensive efficiency schedule to give a piece of data that summarized how efficient a team had been throughout the year. These ratings were altered using KenPom's strength of schedule rating, which was determined by measuring the efficiency of the opponents that a team had faced throughout the year, compared to the NCAA Division-I average efficiency. This factor was what made these efficiency margins "adjusted" as they considered not only how efficient a team has been throughout a year, but also how difficult their opponents were. Typically, teams with larger adjusted efficiency margins could be expected to make it to further rounds in the NCAA tournament. A brief view of the dataset can be seen below:

```
head(tourney)
```

```
##   Year Round  E.M.
## 1 2021     2 10.11
## 2 2002     2 18.73
## 3 2004     4 16.50
## 4 2006     2 12.48
## 5 2018     2 13.41
## 6 2021     3 25.09
```

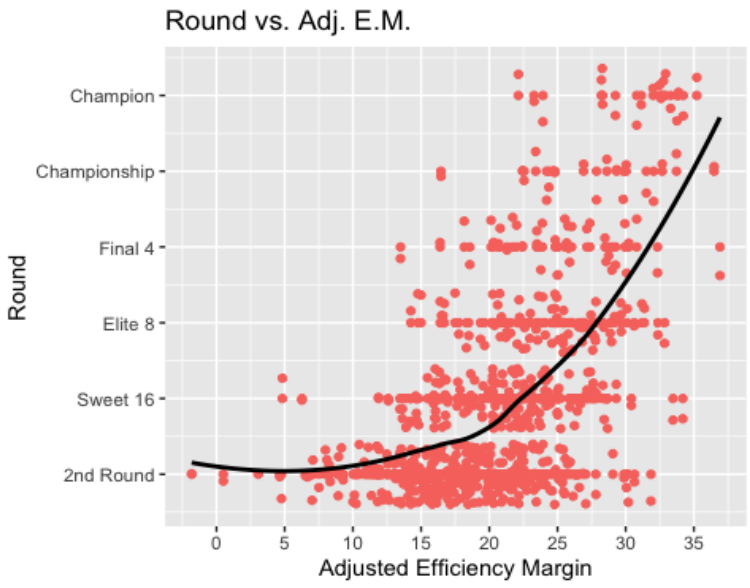
Methodology

The main analysis technique used in the study was multiple regression, which was chosen to quantify the relationship between a team's adjusted efficiency margin and the farthest round they advanced to in the tournament. For this model, a team's adjusted efficiency margin was the explanatory variable and the round a team advanced to in the tournament was the response variable. Multiple regression was chosen for this analysis to understand how adjusted efficiency margins could be used to predict how far a team should be expected to advance in the NCAA Men's Basketball Tournament. The reasoning behind this analysis being multiple regression instead of simple linear regression was due to the unique relationship between the explanatory and response variables, which will be examined later.

Right away there was a notable limitation of this study. The round a team advanced to in the tournament was an interval variable ranging between 1 (2nd Round) and 6 (the champion, though not technically its own round in the tournament). For the purposes of fitting a multiple regression model, this variable had to be treated as a continuous variable. This meant that predictions given from any model fit would most likely not provide definitive information for what round a team could be expected to make it to in the NCAA Men's Basketball Tournament. Predictions would most likely fall between rounds and would be limited at 1 (the 2nd round) and 6 (the champion), though it was assumed that providing prediction intervals could still provide meaningful information regarding what round(s) a team with a certain adjusted efficiency margin could be expected to make it to in the tournament.

Initial Plot Examination

An initial plot was created to examine the relationship between a team's adjusted efficiency margin and the round they advanced to in the tournament was, which can be seen below: (note: jitter was used in this plot for ease of viewing of the interval variable "Round")



From this plot, the most notable feature was that there was a clear nonlinear relationship between a team's adjusted efficiency margin and the round they advanced to in the tournament. The rate of round advancement for teams with the lowest adjusted efficiency margins did not change at the same rate of that for teams with the highest adjusted efficiency margins. This association was noted before analysis began so the appropriate transformation could be used to fit a model.

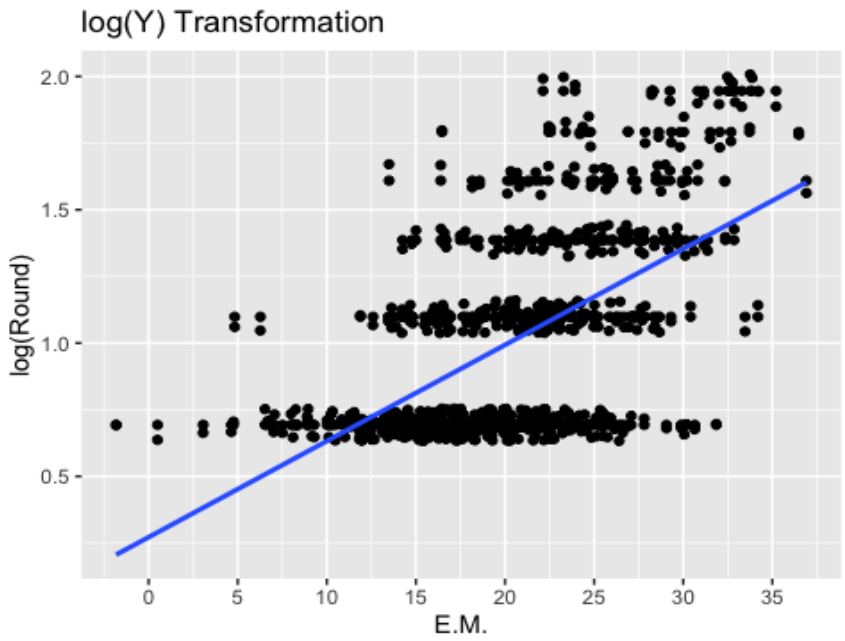
Transforming Y

To determine the best transformation for addressing the nonlinear relationship between a team's adjusted efficiency margin and the round they advanced to in the tournament, a log transformation on the response variable was tested. The log(Y) model was:
$\mu(log(Round)|Adj.EM) = \beta_0+\beta_1 Adj.EM$

```
fit.y.transform = lm(log(Round)~E.M.,data=tourney)
```

Fitted Plot: log(Y) vs. X

A plot of X versus log(Y) values was created with the regression line for the model superimposed onto the plot. This can be seen below:



From this plot, it can be seen that the log transformation on the response variable did not appear to adjust for much of the nonlinear association between the explanatory and response variables. There is still a long flat association with the lowest adjusted efficiency margins which the log(Y) model does not account for. Additionally, the rate of round advancement does not appear to be strong enough at higher adjusted efficiency margins, as the model underpredicts the performance of teams with the highest adjusted efficiency

margins. Both of these issues would lead to large residuals if this model was used to make further inferences, which showed that this model was not a good fit.

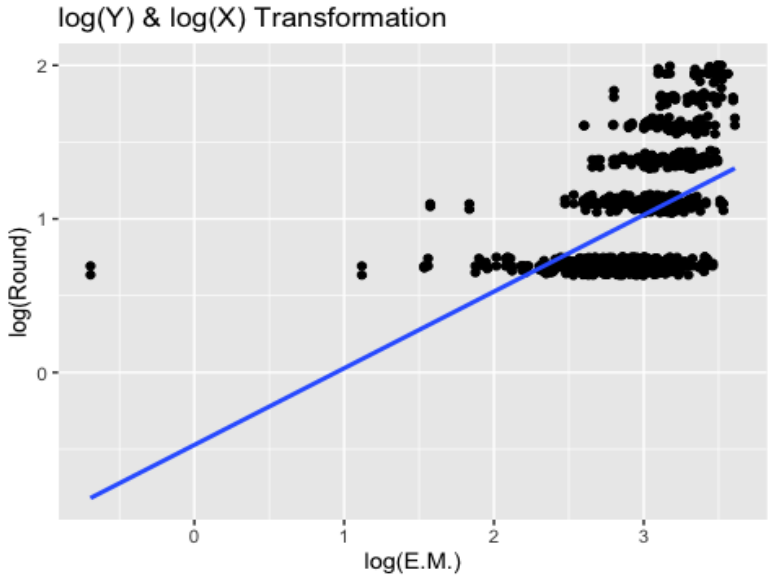
Transforming Y & X

Since the log transformation on the response variable alone was not a good fit, a log transformation on both the response and explanatory variables was tested. The model was: $\mu(log(Round)|log(Adj.EM)) = \beta_0+\beta_1 log(Adj.EM)$

```
fit.yandx.transform = lm(log(Round)~log(E.M.),data=tourney)
```

Fitted Plot: log(Y) vs. log(X)

A plot of log(X) versus log(Y) values was created with the regression line for the model superimposed onto the plot. This can be seen below:



From this plot, it can be seen that the log transformation on both variables still did not appear to adjust for much of the nonlinear association. The log(Y) versus log(X) model had the same issues as the log(Y) versus X model, with the long flat association at the lowest adjusted efficiency margins accentuated in the log(Y) versus log(X) model. Similarly to the last transformed model, these issues would lead to large residuals if this model was to make further inferences, which showed that this model was not a good fit.

Quadratic Model

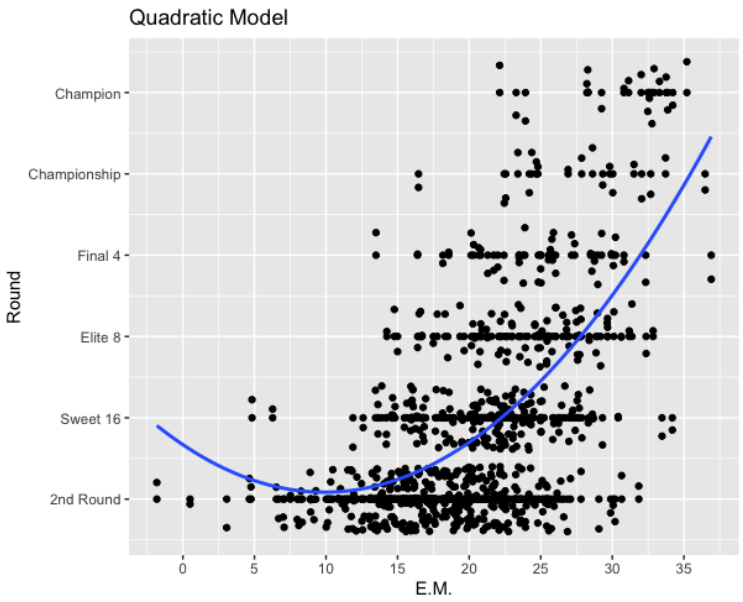
Since there was a clear nonlinear relationship between a team's adjusted efficiency margin and the round they advanced to in the tournament and since both previous transformations were not appropriate, a polynomial term was added to the model. First, a

quadratic model was fit by adding a polynomial term to the explanatory variable. The quadratic model was: $\mu(\text{Round}|\text{Adj.EM}) = \beta_0 + \beta_1 Adj.EM + \beta_2 Adj.EM^2$

```
fit.tourney.quadratic = lm(Round~poly(E.M.,2),data=tourney)
```

Fitted Plot: Quadtratic Model

The X versus Y plot that was used during the initial examination was recreated with the regression line for the quadratic model superimposed onto the plot. This can be seen below:



From this plot, it was seen that the quadratic model accounted for the nonlinear higher rate of increase in round advancement with higher adjusted efficiency margins. The model did appear to be affected by cases with low adjusted efficiency margins which advanced to the Sweet 16, however, which caused the model to fit a higher rate of increase in round advancement with the lowest adjusted efficiency margins. If this model was used, this discrepancy could lead to large residuals at lower fitted values.

Quadratic Model: Summary

A summary of the quadratic model can be seen below:

```
summary(fit.tourney.quadratic)

##
## Call:
## lm(formula = Round ~ poly(E.M., 2), data = tourney)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9910 -0.5828 -0.2001  0.4978  4.0263
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.97035    0.04123   72.05  < 2e-16 ***
## poly(E.M., 2)1 18.01419    1.01568   17.74  < 2e-16 ***
## poly(E.M., 2)2  7.60701    1.01568    7.49 2.46e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.016 on 604 degrees of freedom
## Multiple R-squared:  0.3803, Adjusted R-squared:  0.3782
## F-statistic: 185.3 on 2 and 604 DF,  p-value: < 2.2e-16
```

The quadratic model accounted for 38.03% of the variation in the round that a team advanced to. The test of H_0: β_2=0 had a p-value of < 0.0001, which provided convincing evidence that the quadratic term was nonzero.

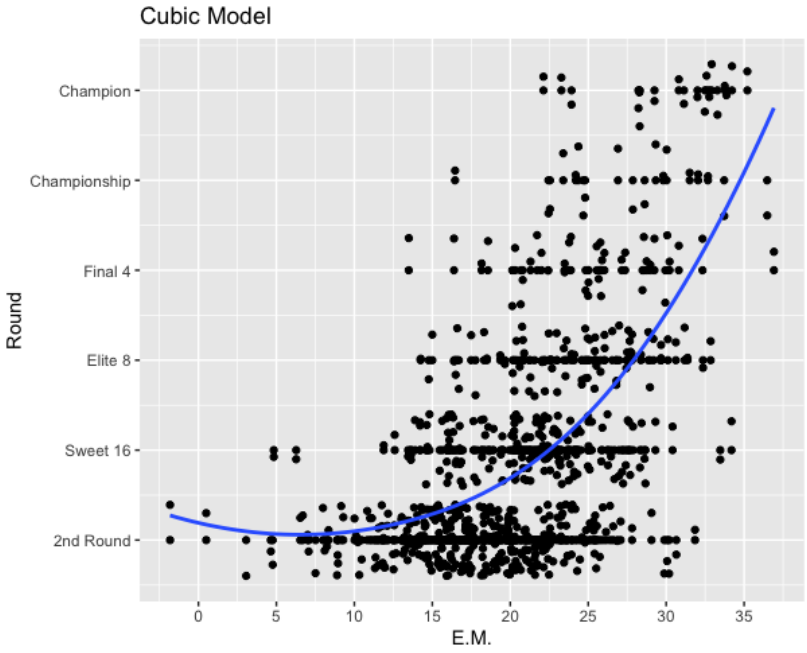
Cubic Model

To determine if a higher degree polynomial would provide a better fit for the data, a cubic model was fit. The cubic model was: μ(Round|Adj.EM) = $\beta_0+\beta_1 Adj.EM+\beta_2 Adj.EM^2+\beta_3 Adj.EM^3$

```
fit.tourney.cubic = lm(Round~poly(E.M.,3),data=tourney)
```

Fitted Plot: Cubic Model

The X versus Y plot was recreated again with the regression line for the cubic model superimposed onto the plot. This can be seen below:



From this plot, it appeared that the cubic model did not account for the nonlinear higher rate of increase in round advancement with higher adjusted efficiency margins much differently than the quadratic model did. Additionally, the cubic model still adjusted for a higher rate of increase in round advancement with the lowest adjusted efficiency margins, but not as drastically as the quadratic model did. These conclusions were supported by the regression summary of the cubic model which accounted for less than 1% of additional variation compared to the quadratic model. Additionally, a test of H_0: β_3= 0 had a p-value of 0.237 which provided no evidence that the cubic term was nonzero. A summary of the cubic model can be seen below:

```
summary(fit.tourney.cubic)

##
## Call:
## lm(formula = Round ~ poly(E.M., 3), data = tourney)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0389 -0.5881 -0.2182  0.5149  4.0495
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.97035    0.04121  72.076  < 2e-16 ***
## poly(E.M., 3)1 18.01419    1.01534  17.742  < 2e-16 ***
## poly(E.M., 3)2  7.60701    1.01534   7.492 2.43e-13 ***
## poly(E.M., 3)3  1.20219    1.01534   1.184    0.237
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.015 on 603 degrees of freedom
## Multiple R-squared:  0.3817, Adjusted R-squared:  0.3787
## F-statistic: 124.1 on 3 and 603 DF,  p-value: < 2.2e-16
```

Since the quadratic model provided a decent fit, the simpler model was retained.

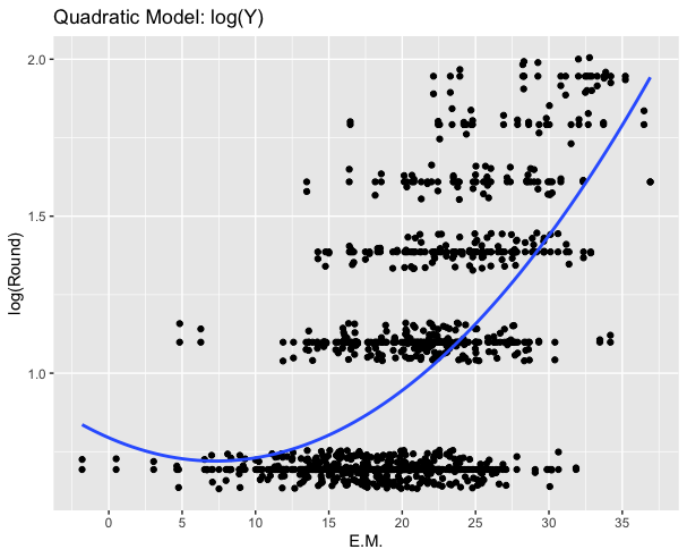
Quadratic Model: log(Y)

To determine if a further transformation of the quadratic model would provide a better fit for the data, a log(Y) transformation was performed. The log(Y) quadratic model was: $\mu(\log(\text{Round})|\text{Adj.EM}) = \beta_0+\beta_1 Adj.EM+\beta_2 Adj.EM^2$

```
fit.tourney.quad.transform = lm(log(Round)~poly(E.M.,2),data=tourney)
```

Fitted Plot: log(Y) Quadratic Model

A plot of X versus log(Y) values was created with the regression line for the transformed quadratic model superimposed onto the plot. This can be seen below:



From this plot, it does not appear that the log(Y) transformation made the fit of the quadratic model much better. This model did not account for the nonlinear higher rate of increase in round advancement with higher adjusted efficiency margins much differently than the non-transformed quadratic model did. Additionally, the transformed model still adjusted for a higher rate of increase in round advancement with the lowest adjusted efficiency margins. These conclusions were supported by the regression summary for the transformed quadratic model, which showed that this model accounted for less variation than the non-transformed quadratic model did. A summary of the transformed quadratic model can be seen below:

```
summary(fit.tourney.quad.transform)

##
## Call:
## lm(formula = log(Round) ~ poly(E.M., 2), data = tourney)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86998 -0.19645 -0.05934  0.20887  0.95460
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.01243    0.01212  83.568  < 2e-16 ***
## poly(E.M., 2)1  5.23646    0.29848  17.543  < 2e-16 ***
## poly(E.M., 2)2  1.74565    0.29848   5.848 8.13e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2985 on 604 degrees of freedom
## Multiple R-squared:  0.3615, Adjusted R-squared:  0.3594
## F-statistic:   171 on 2 and 604 DF,  p-value: < 2.2e-16
```

Since the non-transformed quadratic model provided a better fit, it was retained over the transformed model.

Quadratic Model: Outliers

To determine if there were any influential outliers present which may have affected the fit of the quadratic model, a set of case-influence statistics were examined.

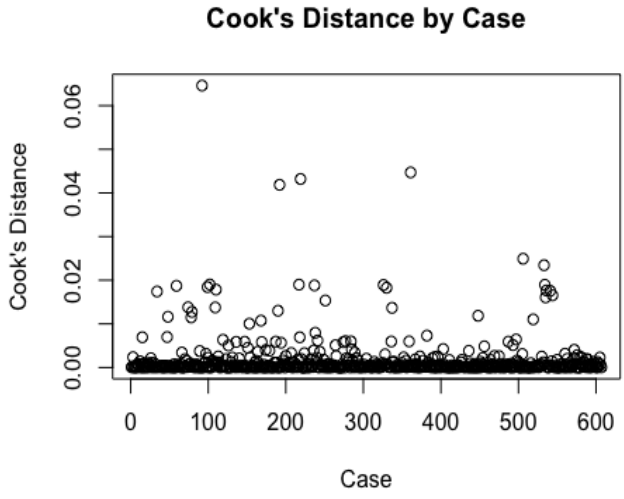
```
tourney$studres = rstudent(fit.tourney.quadratic)
tourney$hat = hatvalues(fit.tourney.quadratic)
tourney$cooks = cooks.distance(fit.tourney.quadratic)
```

Cook's distance, leverage, and studentized residuals were calculated for all 608 cases in this study. Next, cases were filtered to find those with residuals greater than 2 or less than -2, and also with leverage greater than 2*p/n.

```
which((tourney$studres > 2 | tourney$studres < -2) & tourney$hat > (2*2/608))
```

```
## [1]  59  92 110 192 251 330 361 533 535 541
```

This showed that cases 59, 92, 110, 192, 330, 361, 533, 535, and 541 were possibly influential. To determine if they actually were, a plot of Cook's distance for each case was examined:



This plot showed four cases with a Cook's distances noticeably larger than the rest of the cases, though notably none of these values were that large. Cases 92, 192, and 361 were three of the these, however, and each also had a significant leverage and studentized residual. Because of this, these three cases were determined to be influential outliers and were removed before refitting the quadratic model.

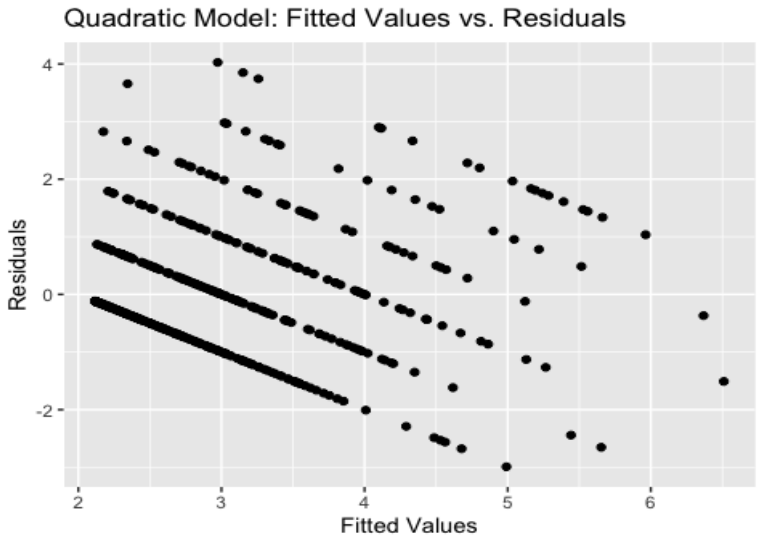
```
tourney = tourney[-c(92,192,361)]
fit.tourney.quadratic = lm(Round~poly(E.M.,2),data=tourney)
```

After the quadratic model was refit, diagnostic plots were examined to determine the appropriateness of this model.

Diagnostics

Quadratic Model: Residuals Plot

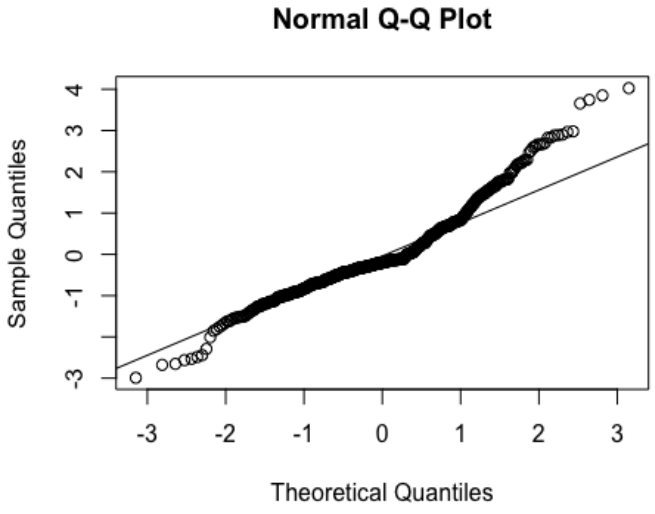
First, a plot was created of the residuals from the quadratic model versus the fitted values from this same model. This plot can be seen below:



From this plot, the limitation of the round a team advanced to being treated as a continuous variable was seen as points grouped in 6 different lines for the six different rounds surveyed. Additionally, it appeared that there was a linear trend across the 6 lines as fitted values increased, with a few notable outliers in quadrant four of the plot. From this, it did not appear that the equal variance assumption was met.

Quadratic Model: Normal Q-Q Plot

Next, a normal probability plot was created to assess the normality of sub-populations. Which can be seen below:

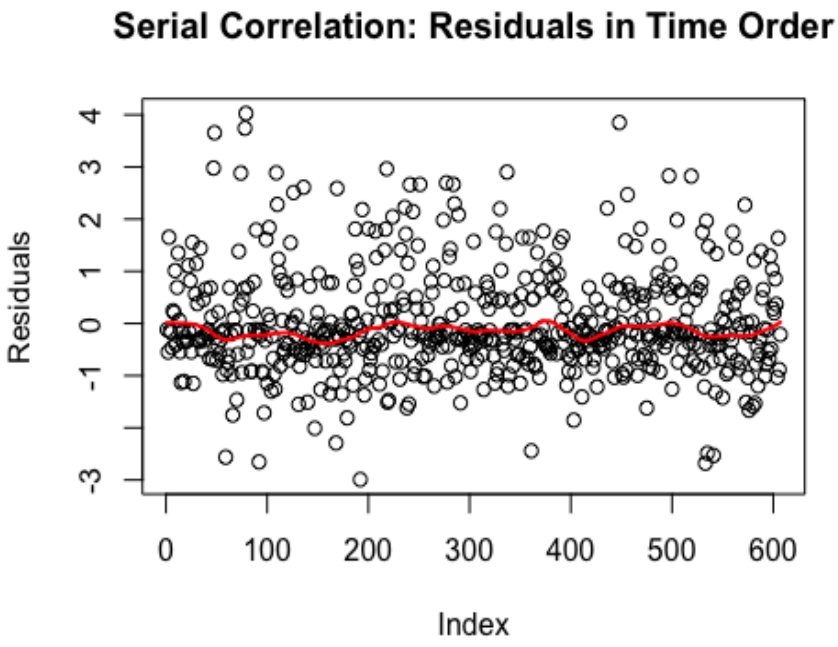


This plot showed that sub-populations may be somewhat long-tailed, but not skewed in either direction. From this, it appeared that the normality assumption was met, though calculating prediction intervals would be unwise.

Given that observations were independent within each tournament and that observations were independent between each tournament, it appeared that the independence assumption was met as well.

Quadratic Model: Serial Correlation

Since data were collected over time, it was necessary to check for a serial correlation in the data. To determine if any serial correlation was present, a plot of residuals in time order from the quadratic model was created with a LOWESS line superimposed on the plot. This can be seen below:



From this plot, it appeared that there was relatively normal noise of residuals in time order. From this it can be concluded that there is no serial correlation present in the data.

After examining various diagnostic plots associated with the quadratic model, inferences and conclusions regarding this study were ready to be discussed.

Limitations

The limitations of this study must be noted before discussing its results. The round a team advanced to in the tournament had to be treated as a continuous variable when in fact it is an interval variable which was restricted at 1 (2nd Round) and 6 (Champion). This meant that if predictions were to be drawn they would most likely fall in between rounds or would extend beyond the rounds possible in the tournament, meaning results would have

to be rounded to drawn any real conclusions. Additionally, there was a clear linear trend between residuals from the model and fitted values, showing that the equal variances assumption was not met. Furthermore, sub-populations were long-tailed and while the normality assumption may have been met this meant that prediction intervals could not be calculated using this model. A transformation or model which met all the assumptions of multiple regression could not be fit, which is a notable limitation for drawing any inferences from this study. Even with what was determined to be the best fit model for these data, these limitations showed issues with quantifying the relationship between a team's adjusted efficiency margin and the round they advanced to in the tournament.

Conclusions

These data provided convincing evidence that the relationship between adjusted efficiency margin and round advancement was not linear (p-value < 0.0001). The rate of round advancement was higher for higher adjusted efficiency margins than it was for lower ones. This lower rate of round advancement for lower adjusted efficiency margins was contradicted by some cases, however, as some cases with the lowest adjusted efficiency margins advanced further in the tournament. Data were fit using a quadratic model which accounted for 38.03% of the variation in the round that a team advanced to. Findings from this study provided a possible regression equation for how far a team could be expected to advance in the NCAA Men's Basketball Tournament using advanced metrics that described a team's performance over a given season. The limitations of this study should be noted, however, as they may impact the accuracy of decisions made using this model. This study demonstrated the complexity inherent with attempting to analyze team performance in the NCAA Men's Basketball Tournament. Team's do not always perform as expected, hence the "madness" associated with this tournament. This variability makes it difficult to predict team success even with the use of advanced analytical tools. This partially explains why no one has ever correctly predicted the NCAA Men's Basketball Tournament, but nonetheless the difficulty of this task is part of what makes creating a bracket a popular activity among sports fans.